Information Retrieval Models

by

MOHAMMAD ALI NOROZI

THESIS

for the degree of

MASTER OF SCIENCE

(Master i Anvendt matematikk og mekanikk)

Faculty of Mathematics and Natural Sciences
University of Oslo

May 2008

Det matematisk- naturvitenskapelige fakultet
Universitetet i Oslo

To the memory of my parents, my family and Fast Search & Transfer

Acknowledgements

Hard work is thought to be difficult but once you get used to work hard you enjoy — the hard work. Because its first rewarding than challenging.

To keep a young and energetic soul alive and young is not an individual task. It is the whole that makes it possible to see the end. The road has been rather long — not to mention somewhat winding.

To dive into the sea of concepts and ideas and to come out and bring about the precious *treasures* of new and peculiar ideas, that make a difference in the area, is a really challenging task — but if done with interest and enthusiasm.

At times you feel that the work that you are doing have no value, but there are few times that you feel your work is going to be worth. Overall its a balance between what you do and what you think is worth, and the overall work in the thesis is to find that balance where you feel contended with work. I think I have found my satisfaction to a certain level before I hand in the thesis.

The one year long thesis has been a motivating and has been my good fortune to encounter many people who have given me more of their time, companionship, professional and personal help, and above all: patience than was perhaps warranted by my seeming determination to indefinitely position the deadline for finishing this thesis at "next year".

I would first of all like to thank my supervisor at FAST, Torbjørn Helvik. He not only gave me the scientific support and supervision that a graduate student can expect from his supervisor, but he also allowed and encouraged me to remain up to date and energetic. Thanks to him, I have never been without a desk, a computer, a friendly ear, and the occasional job at the Fast Search & Transfer. Without those I would never have made it this far.

Geir Dahl, of University of Oslo, is my primary supervisor at University. His ideas, his research, and especially his unique brand of enthusiasm form the bedrock on which much of this thesis was built.

Torgeir Hovden, of Fast Search & Transfer, is my second supervisor at Fast Search & Transfer. I could not have wished for a more thorough discussion partner and sounding board (on *any* subject under the sun), while his excellent long term views (particularly the managerial perspective and the broader and prospective views) has been instrumental in helping to be within the boundaries of thesis and remain dynamic.

A list that, alas, has far too many names on it to mention separately is that of all the co-workers, group members, and roommates that I have worked, talked, and lunched with over the year. My gratitude goes out to all these colleagues and fellows at the Fast Search & Transfer; the Computational Scicence (CMA) group at

University of Oslo.

Moving towards more personal acknowledgements, I would like to execute a big `blocksend()` of aggregated thanks towards all my family and friends — with a special shout-out to my dearest friends *Sultana Ali* and *Sani e Zahra* — for their help, friendship and patience, and for the fact that they never gave in to the temptation to make fun of my thesis.

I am, of course, particularly indebted to my parents and family for their monumental, unwavering support and encouragement on all fronts. They have truly always been there for me, and without them none of this would have been even remotely possible.

Oslo, *Mohammad Ali Norozi*
May 2008

Abstract

In "Information Retrieval", *relevance* is a numerical score assigned to a search result, representing how well the results meet the information need of the user that issued the search query. In many cases, a result's relevance determines the order in which it is presented to the user. In this thesis we have explored the information retrieval models in general and relevancy ranking within information retrieval in particular.

Several mathematical tools have been used in research for improving the relevancy ranking models. A simple yet useful type of relevancy models are based on viewing each document and each query as elements in a high dimensional vector space, and using the angle between the document and the query as a measure of similarity. More advanced concepts in *linear algebra*, such as the *Singular Value Decomposition*, and theory of *Markov chains* have also been employed for innovating relevancy ranking. Some of researches have also suggested and which is also true to certain extent that probability theoretic based models, such as *inference and neural networks* are the best theoretical foundation for relevancy ranking models.

A particularly important question is how to assess the "goodness" of a relevancy model. There is also a greater need to focus on effective and optimized implementations, such as *query latency* times should be in the sub-second domain. Theoretically "recall" and "precision" are used as measures for analyzing the effectiveness of a relevancy ranking models. But with the advent of new and sophisticated models there is a need to have a better framework for evaluation.

In collaboration with *Fast Search and Transfer ASA (FAST)*, I have conducted a study in the area of information retrieval and relevancy ranking models. After an initial literature study, I have partially looked into FAST's query evaluation framework to perform experiments and investigations. But due to unavailability of the required structures that I needed for the study, I had to perform an independent and standalone evaluations.

I have looked into different relevancy models applied in numerous contexts and based on them formulated the focus area, i.e., the *Link Analysis Ranking (LAR)*. The initial investigations therefore made it possible to suggest a case which could be targeted during the study. I have analyzed in light of the theory different concepts and models.

A related area which also has been considered is the automated optimization of the relevancy models. Obtaining good query relevancy is also related to other parts of the system, such as document pre-processing, indexing approaches, the availability of statistics about term distribution, and various query optimization techniques. Thus, a study of such aspects of the system may also be required, in this study we have discussed them, but they are not the focus of this thesis. Instead I looked purely into relevancy models and discussed other important elements only tentatively.

I have found a novel improvement in algorithms like HITS, SALSA and their descendants (e.g., Exponentiated and Randomized HITS) using the *Extrapolation techniques*. Through which I was able to accelerate the algorithms in terms of reducing the number of iterations and therefore uncovered a much faster convergence. In the experiments I even got much better results than theoretically predicted results, a speedup of order $3 - 19$ times better.

The contribution of Extrapolation is unique primarily because I have extensively read through minute details through empirical and theoretical analyses. And it is first time that Extrapolation techniques are used and evaluated in query-dependent algorithms. Previously same kind of techniques has been applied to PageRank, but the improvements were not as promising as in the case of the query-dependent algorithms.

Later, I have theoretically presented numerous personalization models both in query-independent and query-dependent LAR algorithms. And empirically analyzed a smaller subset of the ideas theoretically discussed earlier. A generic framework for personalization has also been discussed which might be considered as an initial step towards personalization.

In the end, extensive experimental evaluations have been performed for the ideas explored earlier in extrapolation and personalization for query-dependent algorithms.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

"Search is going to be the *interaction paradigm*", John M. Lervik, CEO, Fast Search & Transfer [fas].

1.1 Introduction to Search Engines & Preliminaries

1.1.1 A Brief taxonomy of Search

Search engines today are totally different than maybe 5 years ago. Search is not just the search field; it is a platform and infrastructure. Search is equal to advertising, but advertising is driving just one side of the search. The other crucial side is the *relevancy*, match the user needs (informational or recreational), match people to people and connect the people.

The boom of *World Wide Web (WWW)* have made it easier to have access to various sources of information, reaching a wider audience than ever possible, and all kinds of digital communication provided greater access to networks. Information sources are available even if they are far away, they can be accessed quickly and effectively. Users can generously post whatever information that is in some ways considered useful by them. Users on the web are thus expanding from being an active *information consumer* to becoming an active *information producer*. With such a boom in information retrieval and information explosion, there is an ever increasing demand for access, quick responses, and relevant results. Thus there has been a much greater focus on a lot of different techniques that yield highly plausible results.

In this chapter we will broadly introduce the challenges and fundamentals of information retrieval (IR).

1.1.2 Search Engines and its Challenges

Search Engines are considered to be a vital part of the today's fast growing era. According to the anticipation of key search engine companies everything now should come across the retrieval/filtering process – from information need to entertainment and foresightedly to any need. Everything must have to be searched or filtered first in order to have a better match to the needs. Varieties are growing rapidly, and to choose between those myriad varieties the seeker (user) needs to have an opportunity to materialize (specify) their requirements, which the system can recognize. The avalanches of varieties have baffled the seeker today.

Users of the information found themselves floundering, how to retrieve the relevant information. Today's user formulates their needs and expects to have a match to their needs. That is, instead of finding it themselves

by brute force, they need a solution which can bring such facilitation. The users of the information systems hope to retrieve *relevant* information in an effective (best match) and efficient (fast) way.

We consider the user need as *user query* and information collection as the *information*. Most IR systems require user to formulate a query by transforming a complex representation of an information need into a short list of *keywords*. User query must therefore easily be formulated by the user, and efficiently recognized by the retrieval processes. There is an active research on the topic of user query formulation and language modelling, to better facilitate or make it easier for the users to formulate their queries [Met04].

The users want to save time by using an IR system, and expect to have a much better match of their requirements. The size of information is increasing at an incredible pace (exponentially, by many order of magnitude), with such an increase rate, the users are more concerned with the "time" it takes to retrieve a good match to their needs. Thus finding the required information from huge collections is an error-prune and frustrating task for the users. But today information retrieval or filtering is not just a facility – it's an inevitable need, without retrieval by computational systems, retrieval is simply not possible. So, IR systems have become an essential part of today's life.

The increase in the magnitude of information not only troubled the users but also the IR entities (such as, search engines, libraries, etc). It has been increasingly hard and challenging to satisfy the users' needs from the huge and dynamic repositories of information. The explosive nature of information on the Web has further complicated the already difficult problem of identifying usable information from a large collection. There has been huge amount of ongoing research on characterizing and classifying information on these huge reservoirs, in order to make them more manageable for retrieval. Hence the intentions are to come up with some strategies or models that can bring the information closer to its user.

The information collections have been evolving in fascinating ways, apart from just getting larger and larger, it is also dynamically changing. The information sources are no more static, the results which were valid yesterday or maybe an hour ago or maybe a while ago are no more legitimate now. The entries added or updated in the information collections must be accommodated in existing ones. Keeping tracks of the *updates* and *additions* in the information collection and provide the users *up-to-date* information alone is a very gigantic target for research in IR. In addition, the users of IR also have a peculiar nature too – by having dynamically changing needs. So, the challenge is primarily to device a solution which can cope up with the dynamically growing information and dynamically changing needs of the users.

Personalized, context and content sensitive search

There is another interesting and active area in IR systems, i.e., to *personalize* the results according to user's needs. The users of IR systems expect that their behavioural trends are monitored and based on that information personalize the search outcomes according to their preferences (see Chapter 4 for detailed discussion on personalization).

The results from the user query must also take into account the context of the user (from where in the globe the user is requesting for information), and the time at which the request is made. A user query "Christmas" should have different results in other months than in December, and should be different for users in different part of the world. Searches on the topics for which there are opposing communities such as, "death penalty", "religion" or queries that are of interest to different communities, should be appropriately dealt with in IR. Thus the configuration and adaptation of results to different contexts should be possible. IR should be sensitive both to the *content* as well as *context* of retrieval.

Storage Issues

Information repositories are proliferating incredibly and there is also a growing concern about the storage issues. Although today we can afford to have almost indefinite storage, but the price of operating on those indefinite reservoirs are startling. The challenge is how to *represent* and *organize* the huge collections of information, in order to retrieve and process them effectively. The huge collection of information entails a loss of performance and efficiency, because it takes time to process (index, cluster, etc), retrieve (query) and update information in the huge repositories. Thus there is a growing concern about the usability and the interaction time between user and IR systems. A trade-off between the *quality* of the results and *query response time* is mostly considered as an option. High quality and highly relevant results requires more effort in terms of computations hence a delay in the query time and vice versa.

The practical goal of IR systems primarily is to address many of the problems, both in quality and scalability, introduced by scaling search engines technology to such extraordinary numbers (huge information repository).

Relevancy

User query must yield meaningful, manageable and most importantly *relevant* set of results from IR systems. The results should be easily interpretable, and serve to have utmost relevance to what the user is seeking. And if there are many relevant documents to user query, they should be presented in an *ordered* manner, with the most relevant results appearing first. So, the users expect to have ranked results according to relevancy instead of unordered results.

It is commonplace for web search queries to have thousands or millions of results. The impatient user on Web doesn't have the *time* and *patience* to go through all of them to find out the ones that they are interested in. Therefore it's up to the IR system to provide only *minimal* set of results most relevant to the user query.

Relevance scoring and ranking is hence quite a crucial component of an IR system. The purpose of this study is also to explore different relevancy ranking models and their effects on overall search process. This study is primarily an investigation on improving or optimizing the existing relevancy ranking models or proposing any new concept or ideas that could be used to either speedup the interaction time or improve the quality of results.

Relevancy models range from traditional and simple approaches like Boolean based models, to vector based models (see Chapter 2), and more sophisticated approaches like link analysis based models (see Chapter 3). They all make extensive use of concepts from Boolean algebra, geometric representations, vector algebra, probability theory, and significantly the core concepts from Linear Algebra. The focus of this study is to *travel* from a general perspective of relevancy in IR to more specific and classy notion of "importance" in IR by doing *Link Analysis Ranking (LAR).*

Due to the importance of IR in general and large-scale search engines on the Web in particular a large amount of academic research has been done on them. Research range from classical and traditional retrieval models to the sophisticated and modern retrieval models using concepts from a wide range of disciplines.

Stability

In the link analysis based model, *stability* is a desirable property [Ng01b; Ng01a]. If an algorithm is not stable, then slight changes in the *link structure* of the *network* of documents may lead to large changes in the ranking produced by the algorithm. Usually users merely want to see the same results which they have seen before. If the algorithm is not stable enough, the results keep on changing from time to time.

Stability also provides some kind of "protection" against the malicious spammers. Because with stable ranking algorithms they will not be able to easily inflict the ranking by slight changes. *Stability* and *locality* therefore are very important properties of relevancy ranking models, which enables them to prevent the widespread propagation of a change. Just like the *perturbation analysis* in linear algebra, stability and locality properties in relevancy ranking models concerns with the consequences of changes on the search outcomes.

1.1.3 Elements of search

Documents

Documents have quite a central role in IR process. They either *contain* the right information or the *clue* of where the right information can be found, in the form of *citations*.

Document can be a text file, an html file, a pdf book or article, a media file, a blog entry, an image or any other type of object representing information. In short document is a searchable entity in IR and is conceptually more or less equivalent to database table or entries in the table. Documents must be stored and processed in a manner which provides a better result and swift throughput overall. It has become increasingly promising to have more *queriable* documents, with the actual elements of the document being *tagged*, e.g., like XML. Structured or tagged documents (chapters, sections, subsections etc) with optimized representations in index structures directly affect the overall performance of IR.

How IR system views the document is also very crucial, for example, in order to logically or conceptually view the documents, text operations such as *stemming* and *lemmatization* play a vital role. The logical view of the document might be considered as a *shift* smoothly from a full text representation to a higher level representation such as concepts or ideas deduced from the document.

Document structure

Storing information regarding internal structure of the document such as, chapters, sections, subsections etc, helps to provide a much precise understanding of documents and a much defined way of presenting them to users.

There is yet another very important piece of information about the structure of documents, i.e., their *citation structure*. The citation structure of documents provides a valuable source through which the importance of document can be evaluated from its neighbourhood in terms of citations. If a lot of other documents cite one document that means the document in question is important in some particular topic(s). Link analysis ranking models exploits the citation structure of the document to provide popular and important set of documents, relevant to user query or important in general independent of query.

It's quite beneficial and depending on algorithm quite essential to exploit the document's internal structure to provide a more focused and controlled retrieval. The aspects of the DBMS (the fields) in the structured IR, and the schemes of XML retrieval could be exploited in structured document retrieval. There are also places where we would like to shift from structured to non-structured documents and vice versa. Mapping from/to structured to non-structured document is important in such situations.

Natural Language Analysis

The *Natural Language Analysis* techniques can be used to better understand the documents in the corpus, in order to provide a good match between user query and documents. The user formulates their queries in natural language formalism and expects from IR system to understand their needs and bring forth the information

relevant to those needs from the collection. *Natural Language Processing (NLP)* is a discipline in itself, within Artificial Intelligence, but some of the techniques from that discipline could be of utmost relevance when it comes to content-analysis and text operations in IR.

Meta Information

Meta information is the information that can be inferred about a document, but is not contained within the document. Examples of external Meta information include things like *reputation of the source*, *update frequency*, *quality*, *popularity* or *usage*, and *citations*. Meta information have widespread applications, for example, in LAR the excessive exploitation of popularity and citations meta information, and in personalization the application of 'usage' meta information.

Query

Formulating a "good" query is still a *black art*. In particular, many inexperienced and layman users do not have the right skills to formulate a sufficiently precise query which a system can recognize accurately. Most of the times instead of writing query keywords; they ask questions, e.g., 'what is information retrieval?'.

Query subsystem is the other main element of an IR system, which also constitute a considerable influence on the overall search. A misrepresented query will always end up in *topic drift*[1]. The closer the representation of query to the representation of documents the better will be the relevancy of results to the query.

Shifting the burden of enriching their needs in form of a good query that an IR system can understand, the user expects from query subsystem to predict their needs from their sloppy queries. Their expectation is driven by their unawareness of the internal representation of documents in the IR system. It is mostly the query subsystem's responsibility to bridge the representation of query with the internal representation of documents by enriching the user queries with required information needed to decipher them.

The query subsystem takes care of enriching the user query with the required structure and then sends that enriched query to the internal search processes for further processing. Query processing therefore entails, configuring the queries, natural language processing, obtaining personalization information, and thereby supplements the queries with required information.

Users

Users of IR system are responsible foremost to appropriately specify their needs in form of a "good" query. Once the query is submitted, the users may also be required to provide relevance feedback on the search outcomes, which could be achieved in cycles. The explicit user relevance feedback is employed to purify the search outcomes and provide a better match to the query, and consequently keep users' behavioural information for personalization. The implicit user behavioural information can be acquired from their usage logs on webservers. The interpretation and utilization of implicit user feedback is a contemporary mechanism for obtaining users behavioural trends, in order to improve the ordering of top results in search.

1.1.4 World Wide Web (WWW)

WWW is networks of recommendations and it can be related to the *citation structure* of documents. The citation (network) structure of the WWW (the hyperlinked environment) provides a rich source of information, provided

[1] When the documents returned by the IR system is not relevant to the user query, than we say that there is topic drift, i.e., the document does not correspond to the query topic.

we have effective means to understand them. Considerable benefits can be accomplished merely through a pure analysis of the link (citation) structure. By analyses of the link structure we tend to find the *classification* of the documents by figuring out from the link structure what the author or creator of documents intended to classify (implicitly or explicitly). These analyses will help us identify the *proximity* and *relevance* of documents amongst each other. And enables us to find out the social or informational *organization* of the documents.

The work done in LAR models are primarily the successor to the *Bibliometrics*, study of written documents and their citation structure [Lar96]. And it is essentially focused to purely exploit the citation or link structure of the documents.

Hypertext Information (Network structure)

Hypertext information encodes a considerable amount of latent human judgment which can be used to formulate the *notion of importance* from graph theoretic perspective. The hyperlinks therefore define the "context" in which a document appears. We utilize the contextual exposition of documents to deduce the importance or popularity of the documents in the network, by using the core graph theory concepts and techniques.

Web Graph (network)

In the evaluations and analyses of the algorithms in the network or hyperlinked based models, it is fundamentally important to form a legitimate classification and understanding of the graph (network of documents). Through either *clustering* techniques, or applying some other scheme on the *adjacency matrix* of the graph we could achieve a better classification of the network of the documents. These classifications enable to enhance the visualization of the webgraph by grouping together similar documents. Thus it allows us to *inspect* how the top-ranked results of the different algorithms are interconnected with each other and with the rest of the graph. This kind of analyses yields significant insight into behaviour of the algorithms by visually depicting their expositions to the connectivity in the network structure.

Clustering techniques thus involve dissecting a heterogeneous population of documents into subpopulations that are in some way more cohesive or maybe homogeneous [Dri99].

Web search and traditional IR

Web search is *peculiar* in nature in contrast to the traditional IR models. The document corpus is much more dynamic in nature and much more diverse in posture; ranging from structured to non-structured and from media to blogs and many more. The document corpus in web search is a hypertext corpus of enormous complexity, and continues to grow in size and density at a phenomenal rate. *Scalability* therefore is much more important in web search than in conventional IR. How much scalable is the search in comparison to the growing size of the information, is quite a central question in web search. Thus in web search the growing size of the documents collection should be smoothly and transparently dealt with, in order to offer persistent outcomes.

The needs and behaviour of web users and traditional IR users are also different. The web user is hasty in nature, mostly not as serious as in the case of traditional IR.

Notion of similarity in traditional IR perspective means to match the query strings to the documents' strings. Thus relevancy in traditional IR implies that the query strings or query concepts is relevant to the document strings or concepts. While in web search the notion of similarity is largely irrespective to text but subject to the context where documents exist. In essence in web search the focus shifts from *"relevance"* to

"authoritativeness". If the documents are popular or authoritative in the *query topic* in their own contexts, then those documents are deemed to be important and relevant.

Specifically, in the web search there is more focus on the link or citation structure of documents and therefore more focus on behaviour of the network of documents. Analyses of the network structure and its properties thus provide much more farsighted benefits than just pure text analysis. Most of the new models in LAR use the core graph theory concepts like *in-degree*, *out-degree* and *reach-ability* of the network(s) of documents to formulate the notion of similarity.

Ranking purely by in-degree (popularity) can benefit in several cases but have problems also in some other cases. There might be documents with large number of in-degree but lack any thematic unity. Therefore applying different simple *heuristics* might sufficiently help to supplement the ranking based on in-degree. There has been studies in the area where text based techniques (content-analysis) and probabilistic model for modeling the contents are used in conjunction with the link (connectivity) analysis [Bha98; Coh01; Met04; Kai98]. The proposition of basic heuristics by which hyperlinks can enhance the notions of relevance of a document is quite diverse and central to the area of link structure analyses.

Convergence

One of the concerns in the link based ranking methods is the *convergence* to a "good solution" or an *equilibrium state*. A good solution or an equilibrium state is a state where system under certain presumptions can declare the set of good results corresponding to user query. Most of the link analysis based ranking models are *iterative* in nature, they iteratively move towards the required equilibrium state (the good solution). One of the contributions of this study is the speedup of the convergences of the famous LAR algorithms (see Chapter 4).

Convergence is a central phenomenon in *iterative algorithms* [Lay94]. In linear algebra, the iterative methods are employed when direct methods would be prohibitively expensive and in some cases *impossible* even with best possible computing power to find out the actual solution. Essentially the iterative methods provide an *approximation* to the true solution.

Evaluation

For the evaluation of the relevancy models in link based models first we need to have techniques for producing "enriched" samples of the document collection to determine notions of structure and quality that make sense globally. Evaluation of the link analysis based models is therefore a *difficult task*. The difficultly is primarily due to unavailability of a standard framework for evaluation and a representative set of data, which could be confidently used to assess the quality of the algorithms. Tsaparas [Tsa04a], has presented a limited framework for comparing and analyzing LAR algorithms.

A true test of the quality of a search engine would involve an extensive *user study* and results analyses. The *authenticity* of outcomes presented as a result of search is quite important, both for user and as a general ethics in IR. For example search for president name should give first the serious and authentic information than any joke or bizarre information. In results analyses, authenticity of the search results must also be evaluated. In our evaluation we would mainly rely on the user study done in [Tsa04a].

1.2 Contributions

As a result of this study we have made the following contributions which in some ways address some of the challenges broadly discussed in this chapter.

- An extensive exploration of the general concepts and ideas in IR to a certain extent. Several fascinating ideas have been studied which consequently stimulated the next step of the thesis. From the wider perspective of IR got a purposeful motivation towards LAR. LAR models are then explored to certain profundity, in order to form a strong basis for further and deeper analyses.
- In the evaluations and analyses we have incorporated a very effective and novel technique of *Extrapolation* in query-dependent LAR algorithms. This technique offered a very *radical improvement* in convergences of the pioneer algorithms such as, HITS and SALSA. Hence application of Extrapolation to query-dependent LAR algorithms is the major contribution of this study, primarily because this idea was not yet explored in LAR research.
- Studied *personalization* and formulated some of the query-dependent algorithms within framework of the generic personalization model.
- Extensive experimentation of the ideas explored. Both Extrapolation and Personalization are exposed to empirical evaluations to see their behavioural exposition in practical settings.

One of the major contributions of this study is also the exploration of wide range of ideas and the extensive experimental evaluation of a relatively small subset of those ideas. The study is novel because it contains both *theoretical glories* and *empirical evidences*.

In rest of the thesis I have elaborated on the aforementioned contributions. Henceforth, the sketch of contents for rest of thesis is portrayed in the next section.

1.3 Outline

The remainder of the thesis is devoted to look into the challenges that we have discussed quite broadly in this chapter, and also in line with the objectives of the study. To address the objectives I have divided the thesis into *three* main parts.

Following is the brief outline of what we will learn in the rest of the chapters:

- In Chapter 2 we discuss IR models in general, present classical retrieval models and their respective roles in motivating the more sophisticated and newer models. We also review the related literature on IR models in the same chapter.
- In Chapter 3 we introduce the concepts and elements in LAR from literature, the existing models in LAR, their contribution to help resolve the search goals. Both Chapter 2 and 3 form the necessary theoretical background to form a firm basis for the forthcoming analyses and evaluations. Therefore both Chapter 2 and 3 are broadly considered as first part of the thesis, the *theoretical background* part.
- In Chapter 4 we analyze further in depth the models introduced in Chapter 3, specifically with a focus on *query-dependent* LAR models. We elaborate on the important properties of the algorithms and their intrinsic limitations, and subsequently put forward appropriate adjustments which can possibly eliminate

or lessen their side-effects. We also unveil the necessary mathematical implications of the algorithms. Later advocate the novel idea of *extrapolation* to speedup the rate of convergence. Hence we evaluate the core properties of LAR algorithms such as, their convergences, stability, uniqueness and existence of good solution(s). In the end of the chapter we also describe different approaches towards *personalization*, and in the light of theory formulate a generic framework for personalization. In the same section we focus on personalizing the query-dependent algorithms, and explore and exploit numerous ideas for personalization.

- In Chapter 5 we assess the experimental implications of the algorithms and concepts discussed in Chapter 4. The focus of this chapter is therefore to see the relation between the theoretical assertions and practical outcomes. We also present the comparative evaluation of different LAR algorithms, in form of tables and graphs. We have presented all the results of the experimentation in appendices (Appendix A and B). Chapter 4 and 5 constitute the second part of the thesis, the *evaluations, analyses and experiments* part.

- Chapter 6 concludes the thesis with the important results and possible future work, to extend or carry on with the various areas discussed in this study. We also compare the objectives of the thesis with the overall outcomes of the study. Contributions of this study are discussed with more details together with their much broader implications. Ultimately this chapter winds up with the final part of the thesis, the *conclusions, recommendations and future work*.

Part I

Theoretical Background

Chapter 2

Information Retrieval Models and trends

2.1 Information Retrieval (IR) Problem

The IR problem can be broadly viewed from two main perspectives, one as a *computer-centred* and other as a *human-centred*.

The *computer-centred* view of IR problem is essentially concerned with issues like improved crawling and indexing approach, highly efficient ranking algorithms, time efficient replies, efficient representation of the documents, and highly relevant result sets. The utilization of structural information of documents (chapters, sections, subsections, etc) and their representational capabilities is yet another fruitful area in computer-centred approach of IR. This useful source of information provides necessary locality information within the documents, which could be used to assess the relevance of document to a query topic. How the structural information is used; can an IR model function on both the structured and un-structured documents; these are some of the interesting problem statements in this area. There has been a rich amount of research done in only the computer-centred view of IR problem, because of its complexity and dynamics [He05] (see also later sections, and Chapter 3).

The *human-centred* approach on the other hand is primarily concerned with the behaviour, needs and requirements, representations, enrichment and disambiguation of the user queries. The main focus therefore is to understand the users' need, and determine how the involvement of users can benefit the organization and operations of the IR system. The purpose of the human-centred approach is to involve the user more actively in the IR process. For example allowing user first to write their queries in the natural language formalism, the system then fetches the results on this initial query, the user goes through the results and filter out irrelevant results, the system refines the query further on users feedback and then provide much better outcomes. The understanding and implementation of implicit and explicit user relevance feedback can be used in large scale operational environment to improve retrieval. Implicit and explicit user feedback in human-centred approach has been thoroughly studied independently and together with other relevant approaches by different research groups [Agi06a; Bau01; Cha00].

There has been a much greater tendency observed towards human-centred approach of IR problem now [Tan02; Jeh03] (also see Chapter 4). The efforts towards collecting more user relevance feedbacks are primarily destined to *purify* and *personalize* the search outcomes. By doing that, it is expected to have improved retrieval and hypothetically *bridge* the user needs for information with the systems' internal representation of information.

In this chapter our focus is on the IR problem both from computer and human-centred perspective of IR.

We will bring up the candidate point of views in terms of methods and approaches used to solve and deal with the IR problems identified so far. First we define the IR process broadly and then in the later sections elucidate about different IR models with an objective to explore the relevant concepts to the focus of our study.

2.2 The IR Process

The IR process can also be considered from two main viewpoints, the *query-independent* view and the *query-dependent*. Figure 2.1 presents the two viewpoints in graphical form. We will explain the objects and concepts in this figure during the proceedings of this section.

The IR process intrinsically is quite wide and complicated; describing it generally and sufficiently will just serve as a very top level overview. In this section we will therefore have a broad impression of the IR processes, which could be used to establish a basis for realizing the concepts and approaches used in IR later in this study.

2.2.1 Document Corpus

Before anything else in the IR process, there should be a *document corpus* or the document database (e.g., a database of articles, books or webpages). This is the place from where the *contents* or documents get into the retrieval databases. The feeding of the corpuses and collections with the required content is done here. Content feeding is therefore an ongoing process in the life cycle of retrieval, the contents keep on adding, updating and removing. And these are the task-lists of the content feeder, as shown in figure 2.1.

In the case of the World Wide Web (WWW), the *crawler subsystem* collects and stores the documents from all over the web and feed it to the document corpus (permanent storage). In case of traditional IR systems, such as libraries or electronic archives of articles or other materials, the document databases could be populated and feeded in different ways (e.g., manually storing the pdf version of documents or any other automated mechanisms).

2.2.2 Document Manager

After having a repository of the documents the *document manager subsystem* will do the necessary *operations*. The document manager enforces structure to the documents by applying a set of text operations on them. It depends on overall search paradigm, how it sees the documents and what kinds of information are deemed important for retrieval. The text operations broadly transform the original documents by generating a *logical view* of the documents. The logical views of the document are actually the interaction points with the original documents. The logical view contains sufficient information about the document (contents, concepts, structures, etc), which could be used in the later steps in IR process. Efficient and effective logical view (a good representation of the original document) gives a direct boost to the retrieval both in terms of performance and relevance. The logical view of document is formed by applying the natural language operations like, stemming, lemmatization, elimination of stopwords, lexical analysis, recognition of the structure (chapter, section), and other important operations which can help to represent the document accurately. These text operations (transformations) are meant to provide a representative and thematic overview of the documents, depicting the concepts, ideas and structures available in the document. They are primarily used to cope with different IR problems like **Synonymy** and **Polysemy**, words having same meanings or words referring to the same

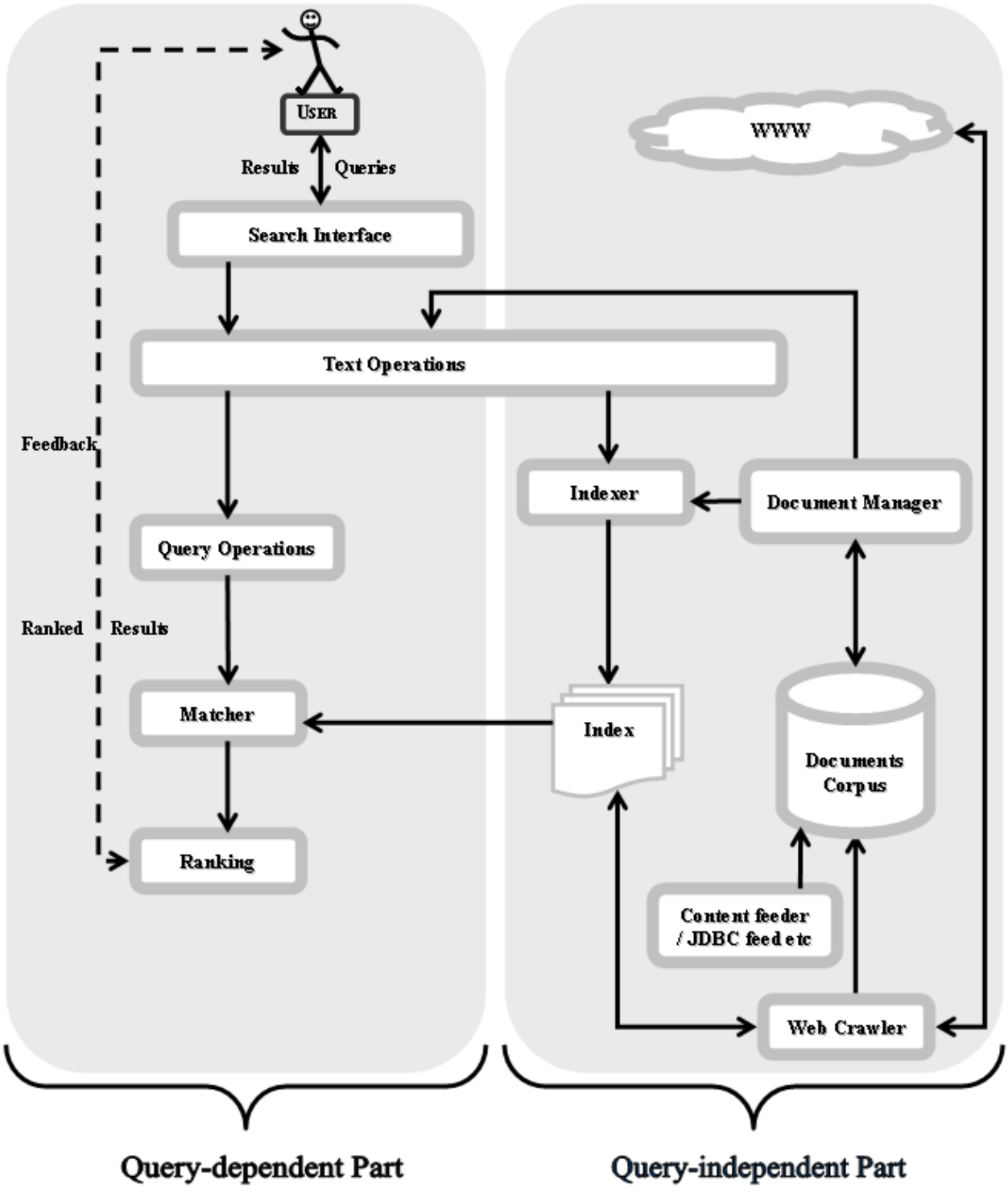


Figure 2.1: *Information Retrieval Process*

concepts respectively. But their main advantage is to provide a representation of the documents which could be effectively used on behalf of the original document, for retrieval processes like matching and ranking.

2.2.3 Indexer

After identifying the logical view of the documents which is done by the document manager, the *indexer subsystem* will take the charge. Based on the logical view, indices are created, which help to efficiently traverse and search for the relevant documents. An index structure in IR model is conceptually similar to the indices in books, which helps to search through huge amount of data swiftly. The primary purpose of the indexing is to provide a *data structure* which can ease the retrieval tasks in IR. Indexing is a wide area in itself in the IR systems. If indexing is not properly taken care of, it incurs considerable performance loss. The data structure produced by the indexing module is used at query time, which means indexing has a direct impact on the search efficiency.

There are a lot of indexing structures available, for example, the most famous ones are inverted files, content index, structure index and other special purposes indexes (image and pdf index, etc). Other valuable information like the *hyperlink structure* of the documents can also be kept as an index structure, which will help the *ranking subsystem* to make use of it. Most IR systems employ numerous index structures at a time depending on the type of the documents, the kind of information available in those documents and essentially on the user/system requirements. The multiple indices are employed either in parallel or synchronously during the *matching* process.

The need for efficiency in index structures mainly depends on the amount of query needed to handle per second (or milliseconds), and the storage required to keep those indices. One of primary considerations when introducing a new index structure is the *space* it occupies. The document databases are usually enormous, and therefore, the index structures corresponding to those enormous databases also occupy huge amount of storage unless properly dealt with. Therefore *compression*, *decompression* and *filtering* techniques have been increasingly used as an integral part of the indexing.

Once the indexing is done the *query-independent* part of the IR process finishes (see figure 2.1), and now the *query-dependent* (interactive) part could be commenced. The information seeker from the IR interface will specify the need for information in form of a *query*; a set of keywords or regular expression with wildcards. The formulation of the user query and its interpretation depends on IR model's specification. After advent of World Wide Web, there has been a consistent effort applied to come up with models which can help ease the users in formulating their queries effectively [Bau01; Met04].

2.2.4 Query processing

The user queries after submission are parsed and then sent to the *query processing subsystem* to enrich them with the necessary and system specific details. The text operations during query processing on the user query will also create a *logical view* of the user query. The text operations are the same natural language operations as in the case of the documents. After the logical view of the query is generated, the system specific representations of the user query will be formulated. The resultant enriched query can be recognized and subsequently processed by the later steps in retrieval.

2.2.5 Matcher

Once the system representation of the query is generated after query operations, the improved query is sent to the core search process, the *matcher subsystem* (see left side of the figure 2.1). The actual matching of query and documents will take place here in order to fetch the set of relevant documents. The matching will be based either on concept matching, or keyword matching, or other matching techniques depending on the IR model's strategies. At this point different IR systems use different heuristics, algorithms and strategies to better predict a match between user query and the documents. As discussed above, depending on how good the indexing is done, the matcher subsystem will perform accordingly. Efficient and comprehensive index structure will yield faster and effective matching between queries and documents.

2.2.6 Ranking

The retrieved documents set from matcher subsystem are usually quite large; presenting them as it is would not be useful for the user. So, the retrieved documents must go through the *ranking subsystem*, to be ranked according to the likelihood of relevance to user query. The output of the ranking is an ordered list of results such that the documents near the top of the list are most likely relevant to user needs.

Ranking is usually done at the query-dependent part of the IR process, but there are IR models which calculate *part* of the ranking scores independent of query (e.g., at crawling time). Some of the models in *Link Analysis Ranking*[1] compute ranking scores independent of any query in order to reduce the overhead in the Ranking subsystem (see Chapter 3 for further details).

The purpose of ranking is to make it manageable for the user to interpret the results from the search. The relevance can be determined on many factors, such as the users' personalization factor, which can be the users' contextual information and users' historical information (browsing history, usage logs, bookmarks, etc). There are many other dynamics independent of users, such as; the frequency of the query terms in the documents ($tf - idf$ factor), where in the documents (heading, abstract, body) user query terms are found, the *popularity* of the documents in their respective contexts. The different ranking strategies or algorithms will be discussed in later sections.

2.2.7 User Feedback

The retrieved and ranked documents are ready to be sent back to the users, who will tend to examine the relevance to their needs. At this point a *user feedback loop* could take place depending on IR model. The user feedback cycle is meant to better understand the users' needs, provide them a more relevant results, and/or *personalize* the results according to their predefined interests. The user feedback could be either *explicitly* gathered (by feedback loops) or *implicitly* obtained through their interaction with the search results (e.g., click-through, time spent on document, document dwell time, etc) [Agi06a; Agi06b].

From the above scenario it seems quite straightforward that users' information needs will be satisfied intuitively. But the reality is in contrary, there has always been petulant users seen, who do not get satisfactory results from querying the IR system. Since most of the users are unaware of the text and query operations, the queries that they usually formulate are mostly insufficient (lacking sufficient expressiveness). And therefore they usually get irrelevant results, because of the fact that they have not formulated their queries appropriately. Hence poorly formulated queries results in unsatisfactory and irrelevant result set, and therefore annoyed users.

[1] PageRank algorithm calculates ranking at a time independent of query.

There are a lot of other factors involved in an efficient IR model; we will go through quite a few of them in next sections, and try to relate them with the focus of this study. Starting with the overview of the *traditional* models in IR in the next section.

2.3 An Overview of the Conventional IR Systems

One of the central components of the IR system is to *rank* the retrieved documents, as identified earlier. Ranking illustrates the *relevancy* of the retrieved documents set to the users query. The issue of predicting which document is relevant and which is not relevant, is quite a central matter in IR systems. It's not far from truth to consider ranking as core of the IR systems. The notion of relevance has been widely explored in order to device efficient algorithms which provide more relevant results to the users. Different IR models determine relevancy in their own peculiar ways, so the question of what is relevant and what is not, depends primarily on the IR model under consideration. There have been a lot of different models proposed over the years. We will discuss a few of the relevant ones here, starting from the three classical and primitive models:

- Boolean model
- Vector model
- Probabilistic model

The implicit assumptions in the classical IR models are first that document corpus is fixed and well structured and manageable in size. Secondly the users are assumed to be trained and cooperative, thus the overall environment in the conventional IR models is much more controlled in comparison to contemporary IR models. Thirdly the initial goals of the classical IR models were not that complex, just to fetch documents relevant to query and ignore context, user specific information and other complex representations. Here we start with these simple classical models and go into much complex models later in the study.

2.3.1 Boolean model

Boolean model is one of the primitive and simplistic IR models, which uses the principle of *exact matching*, in order to match the documents with the user queries. The Boolean IR system require the users to specify their queries using a complex combination of the Boolean logic operators, ANDs, ORs and NOTs. The model is also thought to be a *set theoretic* in nature. A more refined version of Boolean IR model is still in use for example, in electronic libraries. "Boolean" is named because the terms or keywords in the user queries are joined and compared with the documents using the *Boolean algebra*. All the terms in the queries are used as it appears in the query, and matched with the documents, and results are presented without distinction or enumeration.

The user query 'X AND Y' means to fetch all the documents which contains both query terms X and Y. So, the formulation of the query is dependent on the capability of the users to transform their needs into Boolean expression. Considering the fact that any logical statements can be expressed using the Boolean logic, it apparently seems quite promising. The queries are specified as Boolean expressions which have a precise semantics. But the users find it difficult and awkward to express their queries in terms of Boolean expressions.

The query operations in the Boolean model are simplistic usually done by means of regular expressions. Matching only involves to find presence or absence of keywords in the document and relevancies are judged on scale of *relevant* or *irrelevant*, there is no concept of partially relevant. This is one of the main disadvantages of this approach that such type of exact matching produce either too few or too many documents.

But the variants of the Boolean model have been introduced which exploit the inherent advantages of this primitive model and get rid of the inherent limitations, by replacing them with any other germane ideas. For example, the *Fuzzy Boolean model* makes use of the *fuzzy logic* to incorporate the idea of partial matching in Boolean based models. We have discussed a few of the alternative Boolean based models in the later sections.

The Boolean model also fall prey of the two conventional problems in IR, the *synonymy* and *polysemy*. Synonymy refers to words having same meanings such as 'car' and 'automobile', while polysemy refers to words having multiple meanings such as 'bank' (could mean bank of a river or financial centre). These problems together with the intrinsic inabilities of Boolean model may result in a lot of irrelevant outcomes.

The Boolean model also requires the user to have the knowledge of syntax of the query. A user who forgot to put the quotation mark around a phrase might get a lot of irrelevant results.

Boolean model essentially does not allow any kind of relevance ranking of the retrieved document set, although some of the documents are more relevant to the user query than others. Excluding documents which do not precisely match the query terms results a major drawback and ineffectiveness of Boolean model. Thus, Boolean models in reality is much more of data (instead of information) retrieval. And therefore Boolean model is considered weak and ineffective for its general use in IR.

But the variants of the Boolean model do perform equally well. The reasons are that they exploit some of the core advantages of the Boolean model. Creating and programming Boolean model is straightforward and simplistic. Secondly and importantly, queries can be processed quickly, because each query term can be searched in parallel and joined according to the operator used. Thirdly the Boolean model *scales* well to very large document collections, hence accommodating the growing size of collection is much easier [BY99].

Despite the inherent disadvantages of the Boolean model many devotee users still use Boolean systems as they feel more control of the retrieval process [Sin01]. It is also still in use in the traditional data retrieval systems, such as database management systems.

2.3.2 Vector model

The vector based model has identified the inherent shortcoming of Boolean based models that the use of the binary weights is too restrictive, and proposed a solution in which *partial* matching is objectively possible. Developed by Gerard Salton in the early $1960s$, to sidestep some of the problems identified in Boolean based models [Lan06].

Instead of using *binary weighting* (relevant or irrelevant, 1 or 0) the vector model uses non-binary weights to index terms in the queries and documents. Hence both documents and queries are represented as *vectors* in terms space. By doing this, it takes into consideration the documents which are partially similar to the query terms. And therefore calculates the *degree of similarity* between the documents and the query terms. The documents are ranked in decreasing order of degree of similarity. Thus vector based models return documents in an ordered list, sorted according to the relevance to the user query. The result sets are a lot more usable than result sets retrieved by Boolean model [BY99].

Formulation

In essence in vector based model the textual data in the query and document are transformed into *vectors*, and then employs vector algebra techniques to discover the key features and connections between the documents and queries. Therefore a document d_j, and a query q are represented as $n-$dimensional vectors, as shown in figure 2.2. The proposition of the vector based model is to evaluate the degree of similarity between the

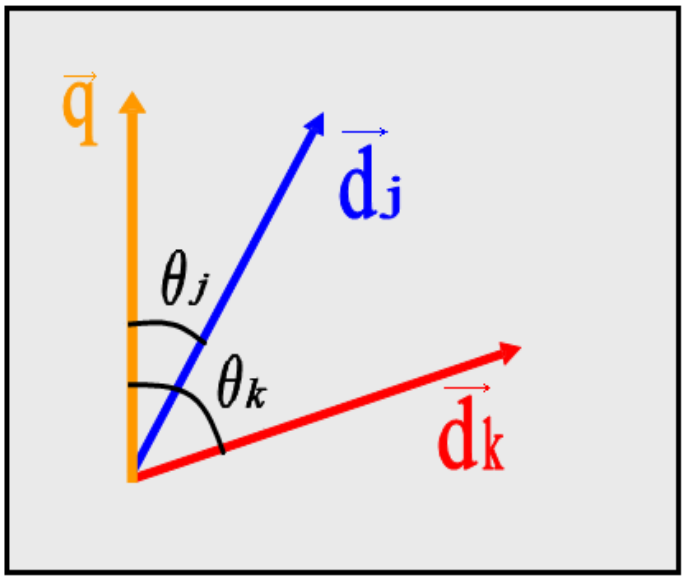


Figure 2.2: *Vector model*

document d_j and the query q, by correlating them to the vector $\vec{d_j}$ and $\vec{q}$ respectively. This correlation can be quantified, for instance, by *cosine* of the angle between the query and document vectors, i.e.,

$$sim(d_j, q) = \frac{\vec{d_j} \bullet \vec{q}}{\left|\vec{d_j}\right| \times |\vec{q}|} \tag{2.1}$$

where $|d_j|$ and $|q|$ are norm (such as *Euclidean norm*) of the document and query vectors. A particular query vector $\vec{q}$ is constant among all the documents, its norm will be same for all the documents, while the norm of document provides a normalization in the space of document.

Typically, the angle between the two vectors (see figure 2.2) is used as a measure of divergence between the vectors, and *cosine* of the angle is used as the numeric similarity (since *cosine* has the nice property that it is $+1$ for identical vectors, i.e., $cos(180^0) = 1$ and 0 for orthogonal vectors, $cos(90^0) = 0$).

Since the $sim(d_j, q)$ varies from 0 to $+1$ (inclusive), the vector model therefore ranks the documents according to their degree of similarity to the query vector. Depending on the *threshold* set for the $sim(d_j, q)$, the documents can be retrieved even they are partially relevant to the query (having value greater than threshold between $[0-1]$).

Clustering

Index term weighting can be calculated differently, depending on requirements and specifics of the model. There have been greater concerns on the ideas to figure out the most *effective* term-weighting techniques for vector based models in specific and generally for wider use in IR.

One of the proposition is for example to support *clustering techniques* [Dri99]. From the entire document collection, the goal of clustering primarily is to separate the document collection into two sets: documents which are related to a set A (a vague set), and documents which are not related to the set A. While more sophisticated clustering might attempt to separate the document collection into various collections, for example clustering of

documents that are syntactically similar [Bro97]. However the decision to assign the query to different clusters is imprecise.

In the scenario of the clustering, the IR problem is reduced to the problem of determining which documents are in the set A and which are not. This means viewing the IR problem as that of clustering problem. But there are two main issues in the clustering problems: *intra-cluster similarity* and *inter-cluster dissimilarity*. Intra-cluster similarity provides quantification of the need to determine, what are the features which better describe the objects in set A. While in inter-cluster dissimilarity one needs to determine what are the features which better distinguish the objects in the set A from the remaining objects in the collection. The most successful clustering algorithms tries to do a balance between these two issues, intra-cluster similarity and inter-cluster dissimilarity [BY99]. Clustering is one of interesting areas in IR; there have been wider and prospective implications of this concept in IR. Clustering is a well-studied subject and there are many notions of how to measure the effectiveness of a cluster. Some of interesting contributions are found in [Dri99; Bro97].

Here in vector model we have a much specific and limited implication of clustering, described in the following subsection.

TF and IDF

Term frequency, tf, the number of occurrences of the query term in the document, is widely used as a weighting measure for documents ranking. In the perspective of vector model, the term frequency factor tf, corresponds to the intra-cluster similarity, and provides a measure of how well the terms describe the document contents. While *inverse document frequency* factor, *idf*, corresponds to the inter-cluster dissimilarity. The motivation for usage of idf factor is that terms which appear in many documents are not very useful for distinguishing a relevant document from the non-relevant ones. However the term frequency tf is dependent on the size of document, i.e., the number of terms present in the document. Thus there is a chance that this measure for term weighting can be misused. As a consequence, tf needs to be normalized by using techniques like term frequency normalization [He05].

For the query term weights, Salton and Buckley [BY99] suggest:

$$w_{i,q} = \left(0.5 + \frac{0.5\ freq_{i,q}}{max_l\ \ freq_{l,q}}\right) \times \log \frac{N}{n_i} \tag{2.2}$$

where $freq_{i,q}$ is the raw frequency of the term k_i in the text of information request q.

There are now several variation of the term weighting identified initially by Salton and Buckley [BY99; Rob04; He05].

Advantages and Disadvantages

Vector based model broadened the perspective of IR models after Boolean based models. There are many *advantages* of the vector based models such as; the term-weighting schemes gives an edge to the retrieval process; the partial matching allows the documents which are partly relevant to the query to appear in the relevant document set; the *cosine* ranking sort them according to the degree of similarity and therefore provide an easily manageable list to the user.

The *disadvantage* of the vector based model is primarily its computational expenses. At query time similarity measures must be computed between each document in the collection and the query. Due to the exponential growth of information collection, the computation becomes further prohibitive. Apparently the vector based model doesn't seem to scale very well to the increasing amount of the information.

The other potential disadvantage of the vector based model theoretically is that the index terms are assumed to be *mutually independent*. And due to the locality of many term dependencies, their arbitrary application to all the documents in the collection might in fact hurt the overall performance.

Despite all these problems, vector based model is widely used and considered to provide a *resilient* ranking strategy [Jon]. Mainly because vector based model is hospitable to other theories. There has been extensive comparison of vector based models with a large variety of other ranking strategies, but the consensus seems to be that, vector based model is either superior or almost as good as the known alternatives [BY99; Sin01].

2.3.3 Probabilistic model

Probabilistic model is based on the *probability ranking principle*, i.e., it tends to rank the documents in the order of their probability of relevance to the query, given that sufficient evidences are available. The probabilistic approach was first presented by Maron and Kuhns (1960). The model was later introduced in 1976 by Robertson and Sparck Jones, which afterwards became known as *Binary Independence Retrieval* (BIR) model [Jon]. The broad intention behind the probabilistic model is to capture or envision IR problem within the framework of theory of probabilities.

Ideal result set

Given a user query, we define an *ideal set* as a set of results which contains only relevant results and no other. Given the properties of the ideal result set we won't have problems to retrieve documents comprising the ideal set. Querying can be thought of as the process of specifying the properties of the ideal result set (which can also be thought of as a clustering problem). The only problem with this newly created IR model is that we do not really know the properties of the ideal result set. We only know the index terms, whose semantics should be used in some ways in order to characterize those properties. Since the properties are not known, an initial guess has to be made, as what these properties could be. These initial guesses will serve as the preliminary probabilistic description of what the ideal result set will be, and they will be used to retrieve the initial set of documents. And then as the probabilistic model recursively operates, it requires that the user take a look at the retrieved documents and decide which ones are relevant and which ones are not, the system successively tend to improve the description of ideal result set. By repeating this iterative cycle many times, it is expected that a suitably well description of the ideal result set can be evolved which will be closer to the original description of the ideal result set.

Formulation

Given a query q and document d_j, the probabilistic model tries to estimate the probability that the user will find the document ${}_jd$ interesting and relevant. As explained above the model assumes that there is a subset of the documents which the user prefers as the ideal result set for the query q. The assumptions are a bit vague because they do not state explicitly how to calculate the probabilities of relevance.

Retrieved documents are ranked by their odds of relevance (the ratio of the probability that the document is relevant to the query divided by the probability that the document is not relevant to the query). Taking the odds of relevance minimizes the probability of an erroneous judgment [BY99; Jon].

$$sim(d_j, q) = \frac{P(R|\vec{d_j})}{P(\overline{R}|\vec{d_j})} \tag{2.3}$$

The index term weight variables are all binary. A query q is subset of the index terms. Let R be the set of documents initially known (guessed) to be relevant, and $\overline{R}$ is complement of R. Let $P(R|\vec{d_j})$ be the probability that document d_j is relevant to the query q and $P(\overline{R}|\vec{d_j})$ be the probability that d_j is non-relevant to q.

Using Bayes' rule [Lay94],

$$sim(d_j, q) = \frac{P(\vec{d_j}|R) \times P(R)}{P(\vec{d_j}|\overline{R}) \times P(\overline{R})} \tag{2.4}$$

where $P(\vec{d_j}|R)$ stands for probability of randomly selecting the document d_j from the set R of relevant documents.

Advantages and Disadvantages

The *advantage* of the probabilistic model in theory is that documents are ranked in decreasing order of their probabilities of being relevant. Secondly, it poses new areas and therefore a family of the IR models (probabilistic or logical retrieval models) continued to grow. One class of probabilistic models which has been used extensively in other application areas consist of those based on networks, such as *inference networks* [Jon] (see Section 2.7.2).

Unfortunately it can be very *hard* to implement the probabilistic model. Their complexity grows quickly, deterring many researchers and limiting their scalability. Because the initial guesses which need to be made about the ideal result set is *vague* and can be very complex in the case if document collection is enormous and heterogeneous. The method also does not take into account the term frequencies (tf) with which an index term occurs inside a document. The assumption of independence of the index terms as well as documents might be prohibitive in practice (see Section 2.6.1). However in practice it is not really obvious that independence of index terms is a bad assumption.

2.4 Future prospects of the classical models

Comparing the three classical models, we see that the primary factor that differentiates between the Boolean model and the rest is the inability to do the *partial matching*. Because the Boolean model is not taking into account the partial matching factor it is considered to be weakest of all the three conventional models. But there are variants of the Boolean model which are addressing this inherent weakness in the primitive Boolean model and therefore they are considered worthy. In the next section we will introduce some of the interesting successors of Boolean model.

There has been a lot of work done in the vector and probabilistic models too. New alternatives and new theories are proposed to address the problems in IR. A series of IR models, the successors to the classical models, have been proposed over the years in the research community [Jon; Met04; Tur91; Rob04]. We will look into a couple of the interesting ones in the next few sections and elaborate briefly how the proposed ideas tend to solve the different problems existing in IR and how they inculcate new ideas in IR. The objective of describing different models here is twofold – first to get to know about the different models and how they solve the IR problem, and secondly to acquire the concepts, ideas and theoretical point of views that each of them advocate. Those concepts and theoretical point of views established in different IR models form a strong basis to explore the more complex problems involved in the IR. They also form the basis for the focus of this study, the *Link Analysis Ranking* models.

2.5 Boolean Based Models

In this section we briefly describe the newer approaches towards the use of Boolean algebra in IR systems.

- Fuzzy Set model
- Extended Boolean IR model

2.5.1 Fuzzy set Model

In the *fuzzy set*[2] model the document and the query terms are represented through a set of keywords, which are only partially related to the real semantic contents of the respective documents and queries. Therefore in this way we are introducing the notion of partial matching in the Boolean based models. This can be done by defining a fuzzy set for the query terms and the documents are considered relevant to the query if they have a degree of membership in this set. The retrieval process is interpreted in terms of the concepts of the fuzzy set theory. The fuzzy set model is introduced by Ogawa, Morita and Kobayashi [BY99].

Fuzzy set theory deals with the representation of the classes whose boundaries are not very well defined. In classical set theory, the membership of elements in a set is assessed in binary terms according to a bivalent condition – an element either belongs to or does not belong to the set. By contrast, fuzzy set theory permits the gradual assessment of the membership of elements in a set; this is described with the aid of a membership function valued in the real unit interval $[0, 1]$. Fuzzy sets generalize the classical sets, since the *indicator functions*[3] of the classical sets are special cases of the *membership functions* of fuzzy sets, if the latter only take values 0 or 1. Thus the membership function and core idea of the fuzzy set theory gives a gradual and partial matching capability to the base Boolean formulation.

Despite of the strong theoretic effectiveness, the fuzzy set models for the IR problem has only been discussed in the context of the fuzzy theory in the literature and therefore not a popular model. For more information about the fuzzy set models see [BY99].

2.5.2 Extended Boolean IR Model

There is another alternative model based on the Boolean logic, known as *Extended Boolean model* for IR problem, introduced by Salton, Fox and Wu [Sal83]. This model is also aimed to relax the strict Boolean logic applied in the primitive Boolean model. It represents a compromise between the strictness of the conventional Boolean system and the lack of structure inherent in the vector-processing system. Hence it preserves the query structure inherent in a Boolean systems and also ranked results of query-document similarity.

The basic idea is that in extended Boolean model the Boolean operations are interpreted in terms of the *algebraic distances*, normalized Euclidean distance and L_p vector norms for $p \in [1, \infty]$. So, the extended Boolean model tends to exploit the properties of both the set theoretic models and the algebraic models. For more information and details on Extended Boolean IR model, see [Sal83].

Nevertheless, despite its inherent weaknesses, Boolean model has been considered a neat framework, which can be the basis for any other enhanced IR model in present and in future too.

[2]Fuzzy sets are sets whose elements have degrees of membership. Fuzzy sets have been introduced by Lotfi A. Zadeh (1965) as an extension of the classical notion of set.

[3]An indicator function or a characteristic function is a function defined on a set X that indicates membership of an element in a subset A of X.

2.6 Vector Based Models

In this section we will study a couple of the fascinating vector based models (successors of classical vector model defined in Section 2.3.2). The purpose is to illustrate them briefly and relate them to the focus of study. We will explore the three famous vector based models:

- Generalized Vector Space model
- Latent semantic indexing model
- Neural network model

2.6.1 Generalized Vector Space model

Independence of index term

The assumption of the *independence* of the index term has been employed in the generalized vector space model. The index terms are expressed in form of *vectors*, like its predecessor. The assumption of the independence of the index term in vector space implies that the set of vectors (expressing the index terms) are *linearly independent* and forms a basis for the *subspace of interest*. The dimension of the subspace of interest is the number of index terms in the overall collection.

Two index term vectors in generalized vector space model might not be orthogonal to each other, which apparently mean that the index term vectors are not seen as the orthogonal vectors which compose the *basis*[4] for subspace of interest. For that we have: if weights associated with the index terms are all binary then all possible term co-occurrences (inside document) can be represented by the set of power of **2** to the number of index terms in the document, the *minterms*[5], given by:

$$m_1 = (1, 0, \ldots, 0), m_2(0, 1, \ldots, 0), \ \ldots \ , m_{2^t}(1, 1, \ldots, 1)$$

where t is the total number of index terms in a collection.

The main theoretical idea that the generalized vector space model is providing is to introduce a set of pairwise orthogonal vectors m_i associated with the set of minterms and to adopt this set as the *basis* for the subspace of interest. And therefore the set of m_i vectors are then taken as the *orthonormal basis* of the generalized vector space model.

The independence of index term is now not seen as orthogonality between index terms vectors itself but rather considered to correlate to the minterms m_i vectors. For example, the vector m_7 is associated with minterm $m_7 = (1, 1, 1, \ldots, 0)$ which corresponds to the document in the collection containing the index terms k_1, k_2, and k_3. And if such a document is found in the collection than we call the minterm m_7 as *active*, and therefore the dependency between the index terms k_1, k_2 and k_3 are induced this way.

The idea of the *dependency* of index term has been independently explored apart from the generalized vector space model (introduced in 1980s). And the basic foundation is the idea that co-occurrence of index terms inside documents in the collection induces dependencies among the index terms. That is why generalized vector space model exploits this basic foundation.

[4] A *basis* is a set of vectors that, in a linear combination, can represent every vector in a given vector space, and such that no element of the set can be represented as a linear combination of the others. In other words, a basis is a linearly independent spanning set [wik].

[5] For a Boolean function of n variables $x_1, \ldots, x_n$, a product term in which each of the n variables appears once (either complemented, or un-complemented) is called a minterm. Thus, a minterm is a logical expression of n variables consisting of only the logical conjunction operator and the complement operator. There are 2^n minterms of n variables - this is true since a variable in the minterm expression can either be in the form of itself or its complement - two choices per n variables [wik].

Practical Implications

Despite its nice formulation and strong theoretic relevance it is still empirically unclear that introduction of the term (in)-dependency in IR model would yield any effective improvement. Therefore it unclear that generalized vector model provides any advantage in practical circumstances. In addition, the cost of computing the ranking in the generalized model can be fairly high with large collections because the number of active minterms might be proportional to the number of documents in the collection. For example, in the case of the WWW, the document collection is enormous, billions of documents and it will be highly unpractical and computationally unattainable to do the operations introduced by generalized vector space model.

2.6.2 Latent Semantic Indexing (LSI) model

Inability of index term

Another interesting problem has been addressed in the *latent semantic indexing* model. The problem concerns the *semantic* of the index term. Because indexing on the index terms at times give poor results. The fact that the index term is present in the document is considered viable for relevancy, but in some documents that doesn't always depict relevancy. After all, the index terms are not what the documents are but the ideas and concept present in text represent the document's identity. It is therefore possible to retrieve a lot of documents that are not quite relevant to the query. And also that there might be quite a few of the relevant documents which are not indexed by the index term, so they might not be retrieved as well. The inherent weaknesses of the index term to represent the content of the document is primarily because of the intrinsic *vagueness* of IR process based on the *keywords*. And also because of the inherent problems in the natural language, i.e., the synonymy and polysemy, discussed previously in Section 2.2.

Matching by concepts

Instead of considering index term as a representative element of the document, we need to explore other interesting features of the documents to match with the query. And therefore the *ideas* in the text more accurately represent which concepts are described in the document. Hence matching under such situations will be, matching the *query concept(s)* with the *documents concept(s)*, rather than matching with the loosely defined index terms. This will allow documents to be retrieved even they don't contain the query terms, which is quite a strong capability of an IR model. With such a strong capability the documents can be *clustered* together according to different concepts and therefore matching will only involve the search across different concepts in the document (concepts) clusters. If a document concept lies under such concept cluster (because it shares the concept), than that document can also be retrieved.

The LSI model is based on the formulation described in the previous paragraph. It tries to solve the issues concerning the index term vagueness. The idea is to map the document and query vectors into a *lower dimensional* space which subsequently corresponds to different concepts. And indirectly this means that we need to map the index term vectors into lower dimensional space. Therefore the expectation is that the reduction of the dimension produces superior retrieval than the conventional way, i.e., by just index terms.

Formulation

The claims above can be achieved by utilizing the nice results from the Linear Algebra. Consider an *association matrix* $\mathbf{M}$, each entry in that matrix defines the relationship between index terms and the documents in terms

of weights, which can be generated using *tf-idf* (see Section 2.3.2) weighting technique. This association matrix is decomposed using the famous *Singular Value Decomposition (SVD)*[6] [Lay94] technique (also see Section 3.4.8 for details about SVD). The reduced SVD of the *terms* $\times$ *documents* matrix $\mathbf{M}$ is given as:

$$\mathbf{M_s} = \mathbf{K_s S_s D_s^T} \tag{2.5}$$

where:
- $\mathbf{M_s}$ is the $t \times d$ (terms $\times$ documents) association matrix. $\mathbf{M_{ij}} = 1$ if document j contains i^{th} index term, and 0 otherwise.
- $\mathbf{K_s}$ is $t \times s$ *unitary matrix*
- $\mathbf{S_s}$ is $t \times d$ diagonal matrix, containing the *singular values* of $\mathbf{M_s}$
- and $\mathbf{D_s^T}$ is $s \times d$ *unitary matrix* in SVD formulation

Only s largest *singular values* of diagonal matrix $\mathbf{S}$ are kept along with their corresponding columns in matrices $\mathbf{K}$ and $\mathbf{D}$. And therefore the resultant matrix $\mathbf{M_s}$ is the matrix of rank s which is closest to original matrix $\mathbf{M}$ in the least square sense [Lay94]. And $s < r$, is the dimensionality of a reduced concept space. The size of s is controversial; it should be large enough to allow fitting all of the structure of the real data and small enough to filtering out irrelevant representational data.

The relationship between documents in dimension s can be obtained from $\mathbf{M_s}^T\ \mathbf{M_s}$ matrix. Because matrix multiplication of $\mathbf{M_s}^T$ $(d \times t)$ with $\mathbf{M_s}$ $(t \times d)$ yields $(d \times d)$ *documents* $\times$ *documents* matrix, hence:

$$\begin{aligned}
\mathbf{M_s}^T\ \mathbf{M_s} &= (\mathbf{K_s S_s D_s}^T)^T\ \mathbf{K_s S_s D_s}^T \\
&= \mathbf{D_s S_s K_s}^T \mathbf{K_s S_s D_s}^T \\
&= \mathbf{D_s S_s S_s D_s}^T, \quad \mathbf{K_s}^T \mathbf{K_s} = \mathbf{I_s}, \text{ since } \mathbf{K_s} \text{ is a unitary matrix} \\
&= (\mathbf{D_s S_s})(\mathbf{D_s S_s})^T
\end{aligned} \tag{2.6}$$

The above matrix quantifies the relationship between each pair of document. Similarly term-term comparisons, i.e., the inner-product of the pairs of term rows, are entries of the matrix $\mathbf{M_s}\ \mathbf{M_s}^T$:

$$\begin{aligned}
\mathbf{M_s}\ \mathbf{M_s}^T &= \mathbf{K_s S_s D_s}^T\ (\mathbf{K_s S_s D_s}^T)^T \\
&= \mathbf{K_s S_s D_s}^T \mathbf{D_s S_s K_s}^T \\
&= \mathbf{K_s S_s S_s K_s}^T, \quad \mathbf{D_s}^T \mathbf{D_s} = \mathbf{I_s}, \text{ since } \mathbf{D_s} \text{ is a unitary matrix} \\
&= (\mathbf{K_s S_s})(\mathbf{K_s S_s})^T
\end{aligned} \tag{2.7}$$

All useful comparisons can be made using the rows of the matrices $\mathbf{K}$ and $\mathbf{D}$ appropriately scaled by the diagonal matrix $\mathbf{S_s}$. And these matrices are much smaller than $\mathbf{M}$ (since we assume s as much smaller than t and d).

LSI computes the left and right *singular vectors* of the matrix $\mathbf{M}$. Equivalently, the eigenvectors of matrices $\mathbf{M\ M^T}$ and $\mathbf{M^T\ M}$ respectively. The simplified expressions in equations (2.7) and (2.6) could be used to do that task efficiently.

The main advantage of this model is that it has introduced a nice application of the singular value decomposition into IR problem. The two matrices $\mathbf{M\ M^T}$ and $\mathbf{M^T\ M}$ are quite useful, we will extensively use them when we study the famous *Link Analysis Ranking* models, such as *HITS*, see Section 3.4. For more details on the LSI model see [Fur88; Coh01].

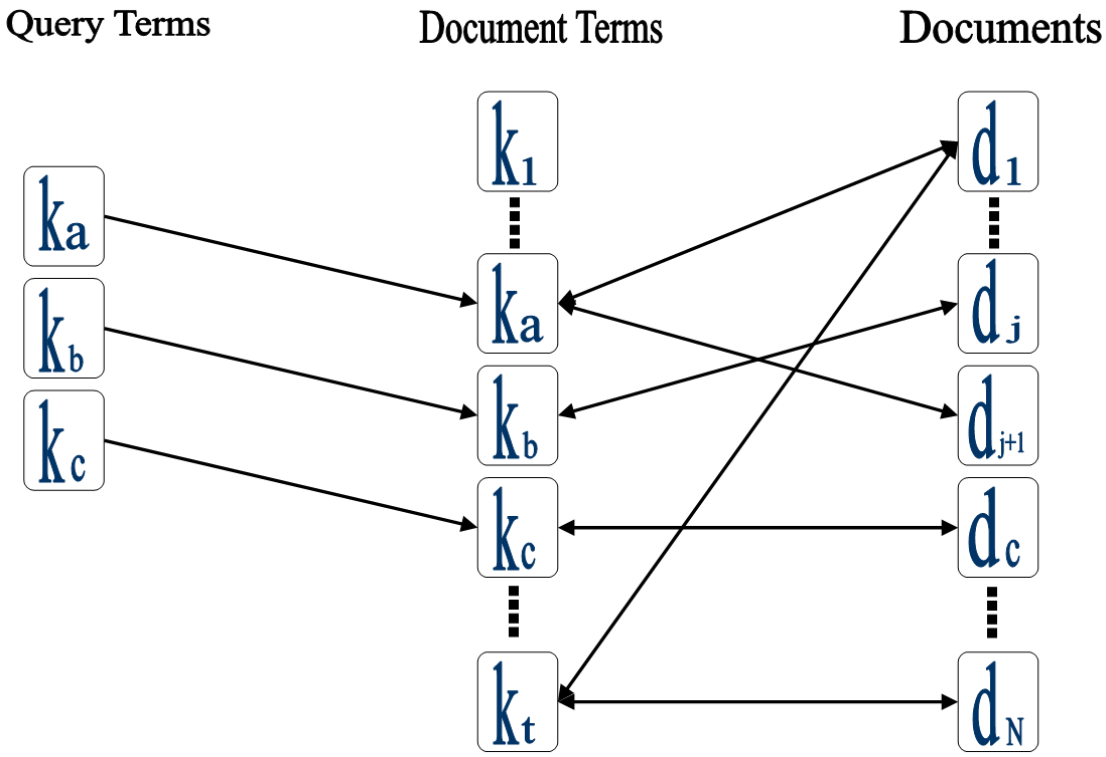


Figure 2.3: *Neural network*

2.6.3 Neural Network Model

The neural network model for IR utilizes the intrinsic capability of the *artificial neural networks*[7] in order to do efficient *pattern matching*. Pattern matching and pattern recognition is quite central topics in artificial intelligence (AI). In the neural network model of retrieval we employ pattern matching techniques from AI. Since the document vectors and the query vector are extensively compared during the matching process, so it is wise to think of some alternative constructs which can help us form a better relationship and much faster computation of the ranking.

Artificial neural network is composed of *neurons* or nodes which can be viewed as the processing unit (like conventional networks). The output or results from one processing unit (neuron) is fed into another neuron for further processing through the arcs or edges connecting them. The process of sending the results back and forth from neuron to neuron can be repeated until the process *converge* to a good solution. This idea of sending results back and forth is quite similar to the idea of *spread of activation* in network models [AND84].

There is a weight assigned to each *edge* in the neural network. Figure 2.3 graphically illustrates a simple neural network model for IR. The neural network in figure 2.3 contains three layers, query terms, document terms and the documents. The activation or results move back and forth among these three layers. Figure 2.3 seems quite similar to the *inference network* described in Section 2.7.2.

In the neural network, figure 2.3, the *signal* or *activation* travels from query neurons to document terms

[6] In linear algebra, the SVD is an important factorization or decomposition of a rectangular real or complex matrix.

[7] *Artificial neural networks* are made up of interconnecting *artificial neurons* (usually simplified neurons) which may share some properties of biological neural networks. Artificial neural networks may either be used to gain an understanding of biological neural networks, or for solving traditional artificial intelligence tasks without necessarily attempting to model a real biological system [wik].

neurons and from document terms to the documents neurons. The query neurons will initiate the signal propagation. However the signal propagation won't stop after first phase of signal propagation. But the document neuron might as well generate signals which are directed back to the document terms, which is why there are bi-directional edges between documents and document terms neurons. Depending on the signal *strength* the document term neuron will send back the signal to the document neuron. The process will repeat unless the signal becomes weak enough to halt. This process might retrieve the document which doesn't even contain the query term, and therefore serve the purpose of being a *thesaurus*. To make it efficient a minimum threshold level is set such that document nodes below this threshold send no signal out.

The neural network model has not been tested as extensively to see how efficiently it solves the IR problem. But it has introduced another important paradigm in the IR models, and therefore the contribution is quite ingenious. It allows retrieving the documents which doesn't even match or contain the query term. Therefore it is quite appealing to use such functionalities together with other potential IR models. In the studies [AND84; Car97; Pir96], the idea of dissemination of signal of neural network model has been largely employed to create a better visualization of information on World Wide Web.

2.7 Probabilistic or network based models

Probabilistic models and its variants have been used since at least the early 1960*s*. Together with the models presented in earlier sections the models in this section form a firm basis for our study. They put forward a lot of interesting concepts and theoretical point of views which are quite necessary to understand the overall IR process in general and *Link analysis ranking* in particular.

Networks have been used to support diverse retrieval functions, including document clustering, neural network formalism, representation of user knowledge or document content, and also to better match the user interest to the document concepts (e.g., LSI) [Tur91]. The significance of the probabilistic model is more entitled to the formalism of the *Bayesian network*, which paved the way for the family of the network based models (which is the focus area of this study as well). Bayesian networks are *directed acyclic graphical models* (DAGs) [Tur91] (see next section).

We will first define the basic prerequisite concepts needed to understand the models. The literature on just probabilistic approaches in IR is by now extensive and at the same time they are highly technical and hard to grasp. In this section we will describe just a few of the relevant ones, in order to capture the new ideas and viewpoints which could help in our study and to better understand and appreciate the problem domain. To comprehend the network topologies in IR and its relevance to the focus area, it is necessary to go a bit more in detail. We first broadly describe the core concepts in network based models and later elucidate on the following model in probabilistic space:

- The inference network model

2.7.1 Basic concepts

Directed acyclic graphs (DAG)

A *directed acyclic graph*, also called as a dag or DAG, is a directed graph with no directed cycles, i.e., for any vertex v, there is no *nonempty* directed path that starts and ends on v. DAGs appear in models where it doesn't make sense for a vertex to have a path to itself, e.g., if an edge $u \rightarrow v$ indicates that v is a part of u, than $u \rightarrow u$ would indicate that u is a part of itself, which is inconsistent. Informally speaking, a DAG "flows" in a

single direction. In the probabilistic or network based models, the networks are usually considered to be DAG, e.g., in case of *Link Analysis ranking*, self reference (self citation) in the documents (nodes in the network) are removed from network, and therefore the webgraph is expected to be a DAG.

Epistemological and Frequentist view

There are two broad categories of probability interpretations which can be called as the *physical* and the *evidential* probabilities.

The *physical* probability or the *frequentist* view takes probability as statistical notion related to law of chance. They are associated with random physical systems, e.g., rolling the dice.

The *evidential* or the *epistemological* view of probability interprets probability as *degree of belief* even when no random process is involved. It is a way of representing the subjective plausibility of a statement, or the degree to which the statement is supported by the available evidence(s). The evidential probability can be assigned to any statement, and in daily life we usually use the evidential probability. The probabilistic model that we describe here takes the epistemological view of the probabilities.

Bayesian networks

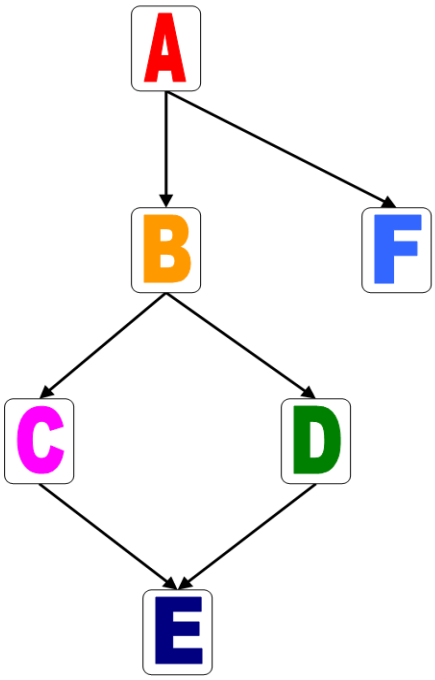


Figure 2.4: *Bayesian Network*

A *Bayesian network* (or a belief network) is a probabilistic graphical model that represents a set of variables and their *probabilistic dependencies or independencies* [wik]. The Bayesian networks (e.g., figure 2.4) are DAG, where the *nodes*, i.e., vertices represent the random variables, and the *edges*, i.e., arcs represent the relationship between the random variables in the network. The parents of a child node are judged to be the direct cause for it. If a proposition represented by the node X_1 causes or implies the proposition represented by node X_2, we draw directed edge from X_1 to X_2 ($X_1 \rightarrow X_2$). The edge between the parent and child nodes characterizes the

dependency relationship between parent and child nodes.

In general, there are *two components* which operate independently in such types of graphs: a *predictive component* in which parent nodes provide *support* for their children (e.g., the degree to which we believe a proposition, depends on the degree to which we believe propositions that might cause it), and a *diagnostic component* in which children provide *support* for their parents (e.g., if our belief in a proposition increases or decreases, so does our belief in its potential causes) [Tur91].

If there is an edge from node X_1 to another node X_2 than X_1 is *parent* of X_2. The set of parent nodes of a node X_i is denoted by $parents(X_i)$. A DAG is a Bayesian Network relative to a set of variables if the *joint distribution*[8] of the node values can be written as the product of the local distributions of each node and its parents:

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | parents(X_i)) \tag{2.8}$$

The network in figure 2.4 shows the Bayesian network for the joint probability distribution $P(A, B, C, D, E, F)$. Notice that node A has no parent, its local probability is said to be *unconditional*. To see the dependencies in the network 2.4, and how those dependencies declared in the network are used to calculate the joint probability distribution, see the following formulation:

$$P(A, B, C, D, E, F) = P(A)\ P(B|A)\ P(C|B)\ P(D|B)\ P(E|C, D)\ P(F|A)$$

The probability $P(A)$ is called the *prior probability* and can be used to model the previous knowledge about the semantics of the application. Given a set of the prior probabilities for the roots of the DAG, these types of networks can be used to compute the probability or *degree of belief* associated with the remaining nodes.

The use of Bayesian or inference networks for IR represent an extension to the probability based retrieval models and has been a key success in the IR research [Tur91].

This finishes the brief description of the basic concepts, for more information about the concepts identified here see [Jon; BY99; Tur91]. Now we are in a stage to embark upon the *inference network model*.

2.7.2 Inference Network Model

Background

The *inference network model* takes on the *evidential probabilistic* view of the IR. It is based on the formalism of the Bayesian network. The essence of this approach is to provide a mathematical rule or formalism to explain how to *change* the existing beliefs in the light of the new evidences. In other words, it allows combining the new data with the *existing* knowledge.

In general, the inference network model is based on the idea to device fundamental techniques for better understanding the contents of the documents and queries. And that understanding is later used to *infer* the expected relationship between documents and queries. Based on the inferred relationship, the retrieved documents are presented.

There has been rich amount of research done on retrieval based on inference or evidential reasoning, and there have been a lot of different competing models proposed and developed in this family [Met04; Tur91; AND84; Coh01]. The techniques employed in inference networks are in some sense comparable to those used in *expert systems*, which are devised to reason from uncertain information. The inference network based techniques have therefore roots leading to the *artificial intelligence* techniques (specifically expert systems).

[8] In the study of probability, given two random variables X_1 and X_2, the *joint distribution* of X_1 and X_2 is the distribution of the intersection of the events X_1 and X_2, that is, when both the events X_1 and X_2 occur together.

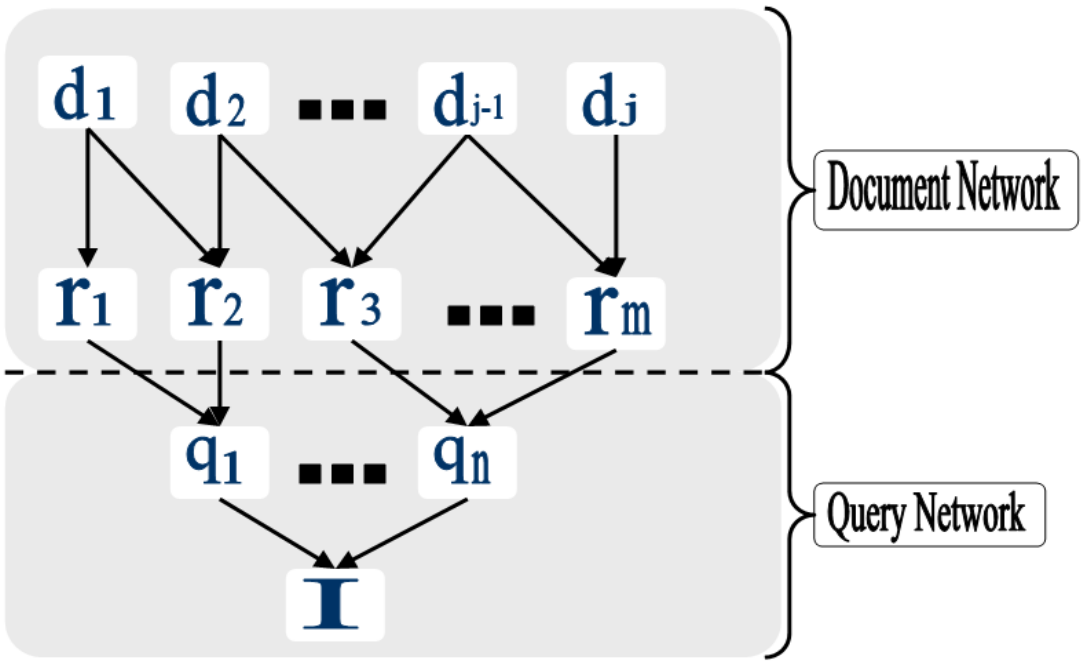


Figure 2.5: *Inference Network*

The Model

Figure 2.5 illustrates an inference network for IR problem. The model broadly comprises two main components, the *document network* and the *query network*. The *document network* can be built from the document collection using different representation scheme (e.g., from text operations), which might depend on the document corpus and area of interest. The *query network* depicts the users' information need and consequently the query representation of that information need. Thus it expresses the user information need based on the system requirement. The document and query networks are joined together by the *concept nodes* through the *links* (directed edges) in the network (see figure 2.5).

Notations

A random variable is associated with each node (e.g., d_j for document nodes), represents the event of observing that node. In the case of documents, d_j corresponds to the fact that the document associated with the random variable d_j is being observed in search of the relevant documents. Therefore the observation of a node is the cause for an increased belief in the variable associated with its child node (in case of the document node it's the *index-term* or *text* node r_m). All the nodes or the random variables in the inference network are *binary* variables.

Documents Network

Document nodes correspond to the *abstract document representation* rather than their physical representation. The number of representations of document in principle are unlimited, in addition to phrase extraction it is expected to have representation based on the *natural language processing* or *automatic keyword extraction*. The choice of different types of representation of the documents depends on the requirements of the user and the

specification IR system. The representation schemes can be enforced at the time of the *text operation* tasks by the *document manager*, see IR process in figure 2.1.

In addition each document node has a prior probability associated with them, as described in Bayesian networks, which expresses the probability of observing the documents when there are no other evidences available. The prior probabilities are usually assumed to be uniform. For instance, in the original work on inference network the prior probability of observing document d_j is set to $1/N$, where N is the total number of documents in the collection.

$$\begin{aligned} P(d_j) &= \frac{1}{N} \\ P(\overline{d_j}) &= 1 - P(d_j) \end{aligned} \tag{2.9}$$

Of course there could be other choices for the prior probabilities of the documents, e.g., in *tf-idf* based ranking strategies, the prior probabilities should be adopted to reflect the knowledge of the importance of document normalization [Tur91; BY99]. In the Link Analysis Ranking models, the prior or initial probabilities are assumed to be in the range of the characteristic polynomial, $det(\mathbf{A} - \lambda\mathbf{I})$ (see Chapter 3). The idea of the prior probability is instructed by the *Bayesian postulate*, that is, previous knowledge of domain should be asserted in the specification of *priors* in the network.

Queries Network

On the other hand the query network is *inverted* DAG, with a single leaf node which corresponds to the event that the user information needs are met (node **I** in the figure 2.5), and with multiple roots which corresponds to the concepts that express the information need (q_n nodes in the figure 2.5).

In general the user's information needs are not specifically known and therefore private to the user. The attempt to express the user's information need in form of multiple *query representation* is truly experimental and judgemental rather than specific or precise. The representation might depict the natural language formalism (e.g., keywords, phrases), user specific information, e.g., history information, or Boolean like formulation of the queries. At this point we could also instil user *personalization* factor into the system, which means to cater user specific information to bias the results according to users need (for more details on personalization see Section 4.4). The representation of information need can be undertaken by the text and query operations graphically shown in the IR process in figure 2.1.

The correspondence between the query representation and information need is not precise, but there are query representations which better characterize the user information need, and several query representations taken together for a single information need maybe considered better than individual ones.

The query network (query representations or concept nodes q_n) must then be connected to the documents network (document representations or concept nodes r_m), depending on some predefined or logical scales of relevance between nodes q_n and r_m.

Network behaviour

The documents network is formed once for the given collection and its structure does not usually change during the query processing. In contrast the query network is built for each information need, and it is modified during the query processing as the query is refined and other additional representations are added in an attempt to better characterize the information need. The attachments of the query concepts (mutable) to the document concepts have no immediate effects on the structure of the documents network (immutable).

The documents network is built independent of the query network and therefore is built once. While the query network characterizing the *user information need*, is built from the available queries, which keeps on refining while the new queries comes in to the system. If the retrieved documents from the available network are inadequate than we need to further elaborate and enrich the query network with more details and specifics in order to better characterize the user information need.

User Information Need

The query nodes q_n, represent distinct query forms or representations and corresponds to the events that the query is satisfied. While the single leaf node **I**, in the query network corresponds to the event that the user information need is met. But generally it is not certain that the user information need will be met, it rather depends more on the representations of the documents and the queries.

The overall inference network provides us the facility to combine the information from the multiple document representations and to combine the multiple query representation to form a single, formally justified estimate of the probability that the user's information need is met [Tur91]. If the characterization of dependencies of the information need on the collection is accurately made, the computed probabilities will provide a *good estimation* and therefore satisfied users.

The iterative propagation of evidences - from known to unknown

The retrieval process in the inference network based model is intended to capture all the significant probabilistic dependencies among all the random variables represented by the nodes in the network. The inference network calculates the overall probabilistic dependencies iteratively – given the *prior probabilities* associated with the document (root) nodes and the *conditional probabilities* associated with the interior nodes, the *posterior probabilities* or belief associated with each node in the network can be computed or inferred from those probabilities. The "evidences" are propagated from the *known nodes* (with prior and conditional probabilities) to the *unknown nodes* (the posterior probabilities) in the network. The network considered this way, represents the dependence of our belief that user's information need is met on the documents in the collection, while the dependence is mediated by the intermediate documents and query representations.

The initial or seed values associated with each node in the network depict the belief associated with them, e.g., the values associated with the information need node **I** represent the probability that the information need is met given no specific documents are observed. If any particular document is observed then the values on the different nodes (beliefs) will be *recomputed* for every node in the network. In particular, the probabilities are recomputed given that a specific document has been observed in the collection. By repeating this process with every document in the collection we can compute the probability that information need is met given that all the documents in the collection have been observed, and therefore *rank* the retrieved documents based on the observations data.

Formulation

There could be different techniques for finding the relevance of the documents to the information need. For example, the simplistic way is that each document is considered in isolation. But considering a subset of the documents might produce higher probability that information need is met. The techniques like clustering, classifying, grouping together related documents and any other sophisticated approach could be used for relevancy. But here we stick to the simplistic way of calculating the relevancy.

The ranking of a document d_j with respect to the query q is a measure of how much evidential support the observation of d_j provides to the query q. In the inference network, the ranking of a document d_j is computed as $P(q \wedge d_j)$ where $q \wedge d_j$ are equal to 1. Let $k = (k_1, k_2, \ldots, k_t)$ is a binary random variable, i.e., $k \in \{0, 1\}$, we have:

$$\begin{aligned} P(q \wedge d_j) &= \sum_{\forall \vec{k}} P(q \wedge d_j | \vec{k}) \times P(\vec{k}) \\ &= \sum_{\forall \vec{k}} P(q \wedge d_j \wedge \vec{k}) \\ &= \sum_{\forall \vec{k}} P(q | d_j \times \vec{k}) \times P(d_j \times \vec{k}) \\ &= \sum_{\forall \vec{k}} P(q | \vec{k}) \times P(\vec{k} | d_j) \times P(d_j) \qquad (2.10) \\ P(\overline{q \wedge d_j}) &= 1 - P(q \wedge d_j) \end{aligned}$$

which is obtained by application of the Bayes' rule [Lay94]. $\overline{q \wedge d_j}$ denote $\neg(q \wedge d_j)$. Going into further details of relevancy calculation is out of scope of this study, for further details see [Tur91; Met04].

Link Matrix

Nodes often connected to multiple parents have to contain functions for computing the probability of their child nodes based on the probabilities of all the parent nodes (which are known). A direct way of encoding the estimated probabilities of all non-root nodes can be done by *link matrix*. Link matrix is also used extensively in Link Analysis Ranking models (see Chapter 3).

This matrix can be of size 2×2^n where n is the number of the parents. They define node probabilities in dependency of parent states that are certain, i.e., parent node probabilities of 0 or 1 not in between. But this encoding is dependent on the number of parents. Encoding on the link matrix is practical only, when the number of the set of the parent nodes is *small*. So, the estimation of the probabilities has two main parts: how to *estimate* the dependence of a node on its set of parents and how to *encode* the estimates in a usable and efficient form.

In literature mostly the link matrix form is used to encode the estimates. There are different canonical link matrix forms for the inference networks to better capture the two main parts described above for the estimation of the probabilities in the network. For more information about the link matrix forms, see [Tur91].

The usage of different link matrices implementing the network's nodes allows implementing features of Boolean based models, vector based model, and probabilistic models. Therefore, inference networks can for example rank documents using Boolean query syntax [Bau01]. *InQuery* is a functional IR system based on the inference network model [Met04].

Computational complexity

The cost of computing the ranks is *linear* on the number of documents in the collection. Therefore the cost of computing an inference network ranking has the same complexity as the cost of computing a *vectorial ranking* [BY99]. The Bayesian network model described above does not impose any significant additional cost for computing ranks. This is due to the structure of the network; it is DAG, which means there aren't any cycles in the network, so there is no unnecessary looping. This implies that the belief propagation can be done in a time proportional to the number of nodes in the network.

The *Bayesian network based models* provide a framework which allows neat combination of *distinct evidential sources* to support a relevance judgement on a given document [BY99].

2.8 Structured and Un-structured document Retrieval Models

2.8.1 Background

The internal structures (sections, subsections, chapters, etc) of the document are very important properties of the documents. Apart from their inherent peculiarities, it is also necessary to take them into consideration when the document corpuses contain structured documents. IR model should therefore exploit the internal structures of the documents in order to offer improved and contextual relevancy ranking. There has been some research done on figuring out whether the document's internal structures do constitute any important source of information for relevancy [Rob04].

It is therefore quite beneficial and sometimes critical to exploit the documents internal structures to enhance relevancy and hence enrich the usability in IR. The IR system built entirely for structured documents is objectively different then IR system built for un-structured documents. At times users are eager to search through the structures of the documents. A query such as, search for word 'information retrieval' in the abstract, subsection, or section of a chapter is highly expected. And even more, a user might require having a visual location of the query string in the document. For example, for the query 'information retrieval', a snapshot of the location where a match has found could suffice. Such requirements cannot be effectively fulfilled by the conventional retrieval models, unless some measures are taken. The visually and structurally manipulated search results can only be achieved in IR systems which are built on the idea of structured document retrieval.

The structured document retrieval might illustrate the appeal of a query language which allows us to *combine* the specification of strings (identifying the visual and structured information) with the specification of structural component of document. Structured document retrieval can therefore be thought of an IR model which tends to combine the information on *contents* with information on document's *internal structures*.

2.8.2 Brief Taxonomy

There has been a series of IR models proposed over the years which tend to focus on structural aspect of documents during retrieval. Various structured retrieval models have appeared throughout 1990s in the literature [BY99]. We will briefly define a few of the relevant ones here, and focus on characterizing and conceptualizing their strong points.

The most ad-hoc retrieval systems apply standard (non-structured) ranking algorithms and tackle the structured documents by combining in some way the scores obtained from different *fields* (abstract, chapter, sections, subsections etc). One of the earliest empirical research in the area of field weighting is the work of Wilkinson [Rob04]. He evaluates different ways to weight and combine the scores obtained on different fields of a document. In practice, many systems exploits structure in an ad-hoc manner, by implementing a *linear combination* of the scores obtained from scoring every field. There have been discussions on how to combine scores efficiently, how are scoring done in terms of simplicity and interpretation, and the computational complexity in terms of accuracy and performance issues involved in different scoring mechanisms. The different methods defines a mapping from structured to non-structured documents, the method can be applied to any ranking function for non-structured documents [Rob04].

2.8.3 BM25

$BM25$ is a method for ranking structured documents. It is based on *2-Poisson model* of term frequencies in documents [Rob04], which could be seen as elementary form of *unigram language model*, with the model parameter for a given term in any document depending on a single binary hidden variable known as 'eliteness'. For each term, the collection of document is split into two classes, *elite* and *non-elite*. It could be seen as, for any given document the *terms* are classified into elite and non-elite. Elite terms are those which appear in the more important part of the document, for example, in the title or abstract. Thus each term occurrence in such a field can be taken as stronger evidence of the *eliteness* of that term in the document than an occurrence in the body. For more information about $BM25$ scheme see [Rob04].

2.8.4 Anchor Text

Anchor text is a text associated with the link in a *source document*, which is assumed to describe the *target document*. Anchor texts are widely used in the documents available on the Web. Apart from its inherent capabilities, the presence of the anchor text in the structured documents also triggers some problems in retrieval. The target that anchor text points to is to be extracted from its source and embed in to the target. One of the problems is that, the source pointing to by the anchor text is not usually written by the same author, and now forms part of the target text. In addition there will be repeated and nested fields in the document (multiple abstract, repeating chapters, etc).

In the study by Chakrabarti et al., [Cha98], an automatic resource compilation uses text in the vicinity of $href$[9] (i.e., the anchor text) in the document as a *descriptive* of the contents present in the target document. Using the weights of terms in the anchor text (the vicinity of $href$) together with the hyperlink information (e.g., by using HITS idea, see Section 3.4) to compute and return a list of web sources are considered to be the most relevant to the query topic.

Discussion about the ways of dealing with the anchor text problems is out of scope of this study, for more information about the anchor text and problems that it brings in see [Rob04; Cha98].

2.8.5 Only Structured Documents

Thinking of the document corpus to have (or consider) entirely structured documents would lead us to approaches that mainly manipulate the structure of the documents. Burkowski [BY99], proposes a model based on *non-overlapping list*. Divide the whole text in each document in non-overlapping text region which are than collected in a list. Since there will be multiple ways to divide a text in non-overlapping regions, so, multiple lists are generated. The text regions in the same list have no overlapping, the text regions from the distinct lists might overlap.

Based on the above model the indexing subsystem must also be accommodated with the structural information and therefore structural component stand as an entry in the indices [BY99].

There could be some efficiencies done to the approaches defined above, for example each of the non-overlapping lists composed of chapters, sections, paragraphs, etc, we could exploit the notion of *nearest neighbour* or *proximity* to do an efficient retrieval and more expressive queries. More complex models for structured retrieval have also been proposed in the literature [BY99].

[9] $href$ is used in a HTML tag. It indicates the URL being linked to and makes the anchor into a link. For example, this tag creates a link to homepage.html: HREF="www.homepage.html"

2.9 Summary and reflections

Traditionally three main entities can benefit from the research in IR models: *library systems*, *specialized retrieval systems* (e.g., ACM, IEEE, etc) and foremost *the Web (WWW)*. But as there is a much higher focus on the retrieval process largely, there is much higher tendency in most sectors where IR models can carry positive contributions. For instance, the growing need for the media and the growing number of varieties led to increasing need of a retrieval system which can facilitate to choose. There has been research going on in the area of *information filtering* which provides recommendations, using content-based and collaborative filtering techniques [Bau01].

Library systems being one of the motivators for the need of IR systems, still provide fuel to the research in IR models and therefore the advent of the digital libraries (with more diverse repositories) have more direct dependencies on efficient IR techniques.

The case of World Wide Web is quite different and unique. The WWW became the ultimate signal for the dominance of the information age. Dealing with the highly dynamic source of information and at the top of it the exponential growth of information has really become a challenge. And having such a huge and dynamic corpus of documents and doing operations like querying and searching on such huge amount of data apparently floundered the user in general and research communities in particular. But there have been remarkable amount of work done throughout the history of IR, for example in 1998 all this thoughts changed, when link analysis hit the IR scene [Bri98; Kle99]. Web search improved dramatically, and web searcher rigorously used the search engines. The addictive and obsessive use of the search engines made them principally a crucial part of the web. The focus of our study is also the *Link Analysis* based IR models (see next chapter). Application of the core concepts of Linear Algebra in Link Analysis Ranking (LAR), especially the most celebrated application of Singular Value Decomposition, as is used in Latent Semantic indexing model, and theory of Markov chains have become the major reasons for the success of LAR.

Chapter 3

Link Analysis Ranking

3.1 Link Analysis Ranking (LAR)

The *exponential* increase in the size of Web is now very eminent. Just the amounts of only textual information are estimated to be in order of hundreds of terabytes. Web is seen as very large, diverse, unstructured but *ubiquitous* databases of information. Subsequently it triggers the need for efficient tools to manage, retrieve and filter information from these databases.

The apparent ease with which the users click from document to documents provides a rich source of information which could be used to understand *what* and *where* to find the important documents. The unstructured and diverse collections of documents are held together by the billions of annotated connections called *hyperlinks*. Analyzing these myriad interconnections between the documents forms the basis for *Link Analysis Ranking*.

Research in Link analysis ranking is objectively derived from *Bibliometrics* research, the analysis of the *citation structure* among academic papers. In Bibliometrics research, the citation structure of a body of document is used to produce numerical measures of the *importance* and *impact* of papers [Lar96; Lan06].

Simple and explicit heuristics and methods might not be able to take advantage of the implicit structure of documents where they reference each other, and therefore, provide a wealth of information. The simple and conventional heuristics rank the pages by the occurrence of query term or any other content based methods. These heuristics are intrinsically susceptible to be misused, e.g., by *spamming*. Spamming is the practice of eliciting favourable rankings, by designing the documents in such a way as to boost their ranking score. One of the main reasons of extensive research in the area of Link Analysis Ranking is to prevent spamming [Lan06].

Not only that the conventional text based methods face problems inherent in the human language, i.e., the problem of Synonymy and Polysemy (as described in Chapter 2). They also need to cater the changing needs of the users by providing them much relevant results from the huge and dynamic sources. With these problems intact an IR system would end up having frustrated users, not satisfied with the search outcomes.

The citation structures of the documents contain a wealth of useful but implicit information. Through citation structure hundreds and millions of documents can be pulled together into a network of knowledge. Foremost such a structure represents the users' behaviour and need. The users discover most relevant and valuable information through recommendations and references from a *good* source of information.

LAR research mostly tends to form a visualization of information, explicitly or implicitly, for making sense of information sets with thousands and millions of objects [Car97]. Through visualization or maybe clustering we can have an overview of the locality and connectivity of information on the Web.

It is the concept of *recommendations* and *endorsements* that exist within the documents in a form of hyperlinks or citations that have been mainly exploited and focused upon in LAR. The *spreading of endorsements* from node to node in the network of documents plays vital role in deciding the importance of documents to the query topic. Different LAR models treat the spread of endorsement in numerous ways. The spread of endorsements in LAR is influenced from the use of *spread of activation*[1] in neural networks–artificial intelligence (see Sectionsubsec:neuralnetwork).

There are studies in LAR which are primarily based only on the connectivity analysis, and there are also other studies that analyze both the content and connectivity information for utilizing the idea of spread of endorsements [Kai98; Car97]. In the study by Pirolli et al., [Pir96], together with content and connectivity analyses, they also consider a graph structure representing the flow of users from node to node, the edges representing the number of users that go along them.

The motivation and wealth of research in LAR is due to the massive size of webgraph and the huge amount of *latent information* available in that graph which could be used in various occasions. Influenced primarily by the usefulness of webgraph and subsequently by other related researches such as [AND84; Pir96; Car97], the seminal work of Brin and Page 1998 [Bri98] and Kleinberg 1998 [Kle99] became the pioneers who formally introduced LAR. The ingenuity was mainly because of the large scale application of core concepts in *linear algebra*.

Hyperlink structures are then used to determine the relative *authority* of the documents and produce improved algorithms for the ranking of search results [Bor05]. Most of the research in LAR is primarily carried out to elaborate the mathematical concepts or improve the ideas initially proposed by Kleinberg, Brin and Page. We will discuss in detail the algorithms developed by them in this chapter and also explore other relevant LAR models. There are also other group of researchers focusing on the LAR problem independent of the two seminal works, who intend to explore any other formulation of LAR, such as analyses of the pure graph structure of web (e.g., *Power-law distribution* formulation of the webgraph) [Bro00; Cha02; Lu04].

In the Chapter 2 we have explored various models in IR in a general setting, but in this chapter the intention is to explore IR models in the more focused settings. Hence we consider documents interconnected with each through hyperlinks or citations. That is, the focus is to independently explore Link Analysis Ranking models. We start with a general model and than illustrate the other focused approaches.

3.2 InDegree Algorithm

One of the general formulations could be that, a *good* document is the one which is pointed to by or cited by many other documents in the network of documents. The number of *in-links* to a page provides a measure of its popularity and quality. The *InDegree* algorithm can be considered as one of the *primitive* algorithms in LAR, that could be used to rank the pages according to their *popularity* or *in-degree*. As the name of the algorithm suggests, it measures the number of documents that link to a document, therefore, it ranks the documents according to their in-degree in the graph. Hence for every node i, we have:

$$a_i = |B(i)| \tag{3.1}$$

This simple heuristic is applied to several IR systems in the early days of the web search [Mar97].

[1] "*Spread activation* is the process by which activation spreads from node to node along network links, making knowledge associated with particular sources of activation (i.e., foci of attention) available for processing" [AND84]. Through spread of activation process one can identify knowledge relevant to current focus of attention.

3.2.1 Relevance to other Models

Kleinberg [Kle99] doesn't seem to be convinced by just the InDegree heuristic and therefore considers *InDegree* algorithm not sophisticated enough to capture the importance or authoritativeness of the documents on Web. And Page et al., [Bri98] extends the idea of InDegree by observing that all the links does not carry the same weight, and therefore proposed a new definition of importance, by saying that a "good" document is the one which is pointed to by many other "good" documents. And when Lempel et al., [Lem00] sashayed into the game, they blend together the ideas presented earlier by Kleinberg, Page and Brin.

If the web graph $\mathbf{G}$ has just one connected component, i.e., the underlying *Markov chain* (see Section 3.3.4) is irreducible, then the algorithm proposed by Lempel et al., the SALSA, reduces to InDegree algorithm [Bor05]. Furthermore, even when the graph $\mathbf{G}$ is not connected, if the starting point of the *random walk* (see Section 3.3.4) is selected with probability proportional to the "popularity" (in-degree) of the node in the graph $\mathbf{G}$, then their algorithm again reduces to the InDegree algorithm. This algorithm was referred to as PSALSA (popularity-SALSA) by Borodin et al., 2001 [Bor05].

The purpose of defining InDegree algorithm is also to see the resemblance of the modern LAR models with this primitive model. Hence in experiments we found that there are some of the new approaches in LAR whose $top-15$ results sufficiently look like the $top-15$ outcomes of this primitive algorithm. Though they differ in the way they calculate the rankings but the outcomes are observed to be similar in the prescribed settings (see Chapter 5).

3.3 PageRank

PageRank algorithm for calculating the relevancy of the pages has been a *landmark* in the field of Link Analysis Ranking. The seminal paper of Brin and Page [Bri98] introduced the initial idea of PageRank, which exploits the hyperlink structure of the Web to determine the relative popularity or importance of each webpage. The hyperlink structure of Web is used to build a *Markov chain* formalism of IR, by iteratively calculating the *stationary distribution* vector until the rank vector is found, *independent* of any queries. The approach is novel primarily because of the relevant and prospective application of Markov chain theory.

"The heart of our [Google.com] software is $PageRank^{TM}$ ..." [goo], PageRank is thus considered as an integral part of the Google's success.

In this section we will explore the PageRank algorithm in terms of core concepts used in the initial scheme and its inherent limitations. We will study the algorithm with the help of suitable examples with necessary details. The purpose is to study different properties and peculiarities of the algorithm and its implication on overall LAR research subsequently its importance and eminence to the focus of our study.

3.3.1 Web Graph

Given the document corpus and index structure containing hyperlink information, the first step is to construct the underlying *graph* from the hyperlink structure in documents. The hyperlink structure of the documents on the Web, form a *massive* directed graph. The node in such a graph represents the webpages (documents), and the edges represent the hyperlinks. The graph is directed because the edge from one node P_A to another node P_B ($P_A \rightarrow P_B$) means that the page P_A contains the hyperlink to page P_B. The webgraph is *simple* in a sense that if there are multiple links from one page to another, only a single edge is placed between them in appropriate direction.

Links within the same webpage are removed (e.g., navigational links, such as, 'back', 'home'). The links into a webpage are called **in-links** while the links out of a webpage are called **out-links**. Every node has a *degree*, similar to the graph theoretic formulation. Degree refers to the number of edges incident to a node (in-links and out-links). Hence the total number of links going out of a node are called **out-degree**, while the total number of links coming into a node are called **in-degree** of a node in the webgraph.

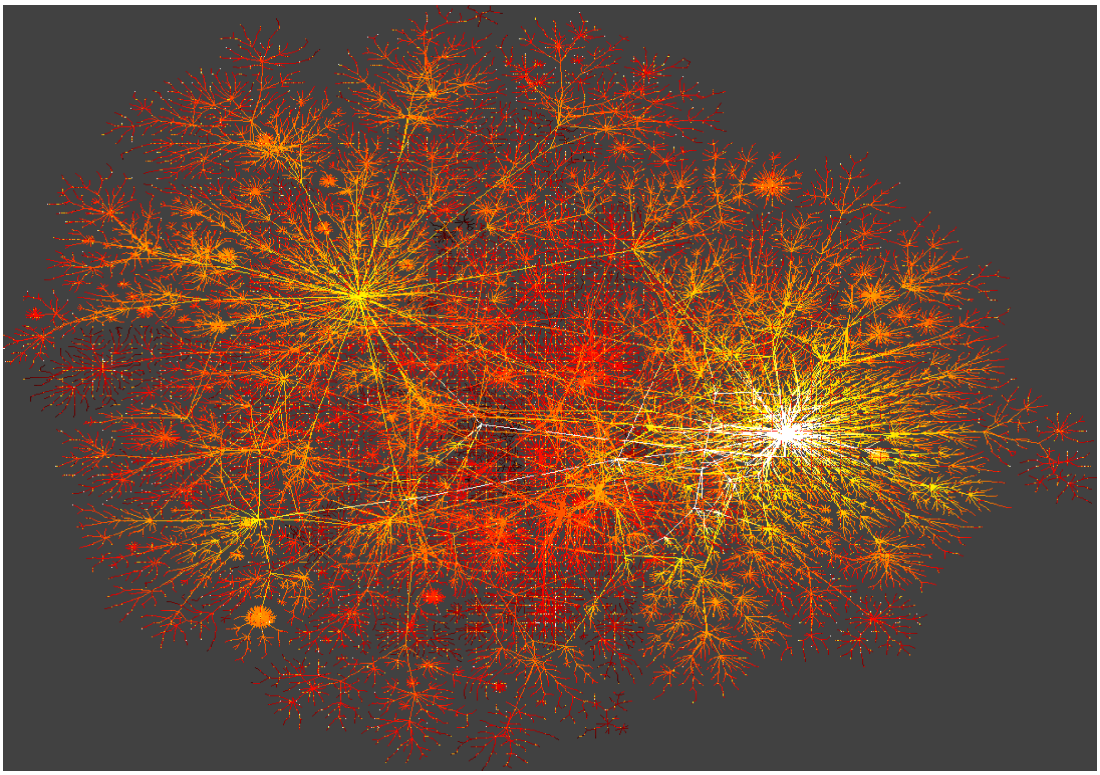


Figure 3.1: *The webgraph or the internet map - Image taken from [com]*

Extracting meaningful information from the massive webgraph is not that straightforward, because the nodes and edges create a very chaotic and dense graph. Drawing the webgraph alone is a Herculestic task, if drawn it will be the largest human artefact ever created. Webgraph the way it is impossible to visualize unless some *clustering* algorithm is used to make it visually friendly. See for example figure 3.1 for the clustered view of the webgraph. It is quite problematic to disentangle such a complex graph in order to even present it in a meaningful way. But there has been some work done such as, the *Atlast of Cyberspace* [Dod01] presents over 300 colourful informative maps of cyberactivities. For more information on structure of web see [Hir00; Bro00; Ara01].

There is also an extensive amount of **graph theory** involved in the Link Analysis Ranking. Because we exploit the massive webgraph, and to extract meaningful and relevant information from such a massive graph requires a much closer understanding of the underlying graph structure and realization of the network topology involved.

The problem of understanding the input graph is of fundamental importance for the study of LAR models. Undoubtedly we need to understand how the structural properties of the graph affect the ranking of the algorithm [Bor05].

One of the interpretation of the hyperlinks in webgraph is to consider a hyperlink from page P_A to page P_B as an *endorsement* given from page P_A to P_B. Academic citations in literature or "bibliometrics" of academic

papers broadly support the idea of endorsements. The "bibliometrics" of the Web hence is largely employed to account for the citations (endorsement) bestowed to each webpage by accumulating their in-links. That is, hyperlink in the documents are considered as a recommendation, an edge from pages P_A to P_B means P_A recommends page P_B, and therefore P_A consider page P_B as an important page, quite similar to academic citations.

3.3.2 Google's PageRank

The *Google's PageRank* considers the endorsement (edges) slightly different, the status of recommender is also considered important here. A recommendation or hyperlink from one page to another page is considered good if the recommender itself is a good source, which means that the recommender has a high PageRank itself. In a nutshell a webpage is important if and only if it is pointed to by other important pages [Bri98].

In some texts, the recommendation is identified as *vote*, and therefore Web becomes a *democracy* where pages vote for the importance of each other by hyperlink towards one another [Bry06].

The idea described above seems *self-referential*. There is some thing more than just vote or recommendation. A single page might illicitly gain influence by recommending a lot of other pages, but we don't want this to happen. Therefore curbing such an effect by evenly dividing the importance or PageRank of the recommender among all of its outgoing links.

In essence the amount of *importance* conferred on P_k by P_j ($P_j \rightarrow P_k$) is directly proportional to the importance of P_j and inversely proportional to the number of pages P_j points to. Thus, if page P_j contains $|P_j|$ links, one of which links to page P_k, then page P_k's score will be boosted by:

$$\pi(P_k) = \sum_{P_j \in B_{P_k}} \frac{\pi(P_j)}{|Pj|} \tag{3.2}$$

where $\pi(P_k)$ is PageRank score of page P_k and B_{P_k} is the set of *back-links* (in-links) of page P_k.

Now the PageRank of P_j is evenly distributed among the pages linked by it (i.e., the set out-links of P_j). But to compute the equation (3.2) we need to know PageRank of P_j, i.e., the PageRank of in-linking pages to page P_k. The PageRank of the in-linking pages can be calculated in an *iterative* fashion. Starting with an initial PageRank values for each page (the seed values) in corpus, e.g., $1/N$, where N is total number of pages in the corpus, the iterative algorithm will compute the new values of the PageRank successively from the previous values. Thus the equation above can be written in iterative form as:

$$\pi^{i+1}(P_k) = \sum_{P_j \in B_{P_k}} \frac{\pi^{i}(P_j)}{|Pj|} \tag{3.3}$$

where $\pi^{i+1}(P_k)$ is the value PageRank of P_k at time step $i+1$.

The process starts with some initial value, and successively compute the new values assuming that with such an approach the PageRank scores will eventually *converge* to some *stable* and desirable final values.

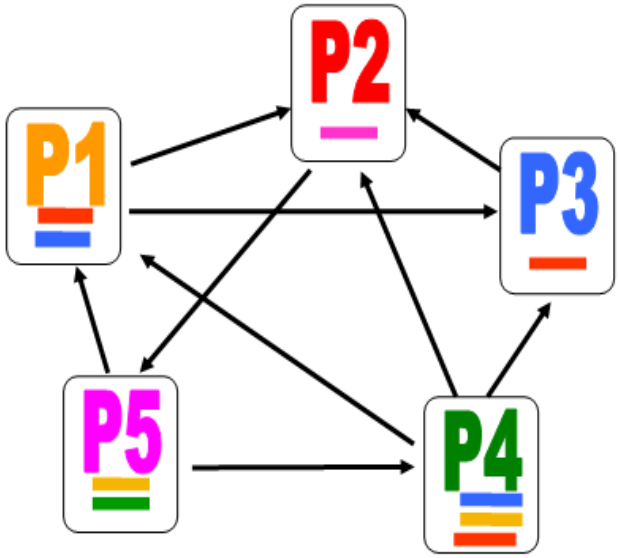


Figure 3.2: *Network of five pages referencing each other via hyperlink*

3.3.3 Link Matrix

Let us consider an example to illustrate the concepts and ideas described above. The figure 3.2 can be represented in a *link matrix* form as follows:

$$\mathbf{A} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \end{pmatrix}$$

The link matrix $\mathbf{A}$ above is a row normalized matrix with, $\mathbf{A_{ij}} = 1/|P_i|$ ($|P_i|$ is the out-degree of page P_i) if there is a link from node i to node j, and 0 otherwise. The matrix $\mathbf{A}$ is similar to the *adjacency matrix* in graph theory; except its nonzero elements represent the endorsements (e.g., through probability). At any iteration t, the next successive values of PageRank $t+1$ for above graph can be written in a system of linear equations as:

$$\begin{array}{lcccccccc} p_1^{t+1} & = & & & & & 1/3p_4^t & + & 1/2p_5^t \\ p_2^{t+1} & = & 1/2p_1^t & & + & p_3^t & + & 1/3p_4^t & \\ p_3^{t+1} & = & 1/2p_1^t & + & & & 1/3p_4^t & & \\ p_4^{t+1} & = & & & & & & & 1/2p_5^t \\ p_5^{t+1} & = & & p_2^t & & & & & \end{array}$$

The above linear system of equations can be written in matrix form as:

$$x = x\ \mathbf{A} \quad where\ x\ =\ [p_1,\ p_2,\ p_3,\ p_4,\ p_5] \tag{3.4}$$

The equation (3.4) seems quite familiar in term of linear algebra formulation. The problem of finding the PageRank is now transformed to "standard" problem of finding the *eigenvectors* of the link matrix $\mathbf{A}$.

Thus equation (3.4) will iteratively compute the PageRank $\mathbf{n}-$ vector x of graph ($\mathbf{n} = 5$ in figure 3.2). Instead of solving the system of linear equations we can solve the matrix formulation. In terms of matrix notation, the PageRank vector can be represented as a row vector π with k^{th} iteration as a superscript as follows:

$$\pi^{(k+1)} = \pi^{(k)}\mathbf{A} \tag{3.5}$$

Considering the structure of the link matrix $\mathbf{A}$, the i^{th} row of $\mathbf{A}$ contains $|P_i|$ (the number of out-links of page P_i) nonzero entries, each equals to $1/|P_i|$, and therefore the row sums to $\mathbf{1}$. This motivates the observation in the next section.

3.3.4 Markov chain and Random Surfer (Walk)

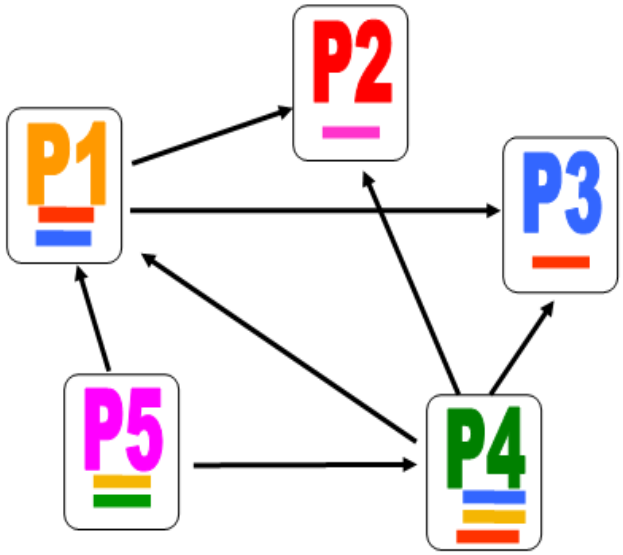


Figure 3.3: *Network of five documents containing "dangling nodes"*

Brin and Page use the notion of a **random surfer** (synonym of random walk in Markov chain theory [Bré99]), to describe the process of calculating the PageRank of the whole web. The random surfer is given a webpage at random and he keeps following the hyperlinks, never clicks 'back'. The surfer hops along randomly chosen hyperlink structure of the web. When he arrives a webpage with a lot of outlinks he chooses one of them randomly, and continues with these random decisions indefinitely unless theoretically all the pages are exhausted (visited).

The relative importance of the webpages (PageRank) can be viewed as the proportion of time the random surfer spends on a webpage. If he spends large proportion of his time on that page, he must have repeatedly found himself returning to that page, when randomly following the hyperlinks. But the random surfer eventually gets *bored* and randomly starts on another random page (e.g., by writing in the browser). PageRank, therefore can be thought of a model depicting user behaviour [Bri98].

It is interesting to note that the behaviour of the random surfer can be depicted through the application of "Markov chain" (i.e., the sequence of random walk). A random surfer will have a random walk on the Web in a form of navigation between pages, each page represents a possible state, and each link represents a possible transition. And therefore the random walks by the surfer is *Markovian* because transition of each step is independent of the previous steps and it only depends on the current state [Bré99] (also see [Mot95, 132]). Hence the PageRank vector π represents the stationary distribution or equilibrium distribution of a random walk on the entire Web.

In the study of Markov chains, a square matrix is called row-stochastic matrix if all of its entries are nonnegative and the entries in each row sum to $\mathbf{1}$. The matrix $\mathbf{A}$ of figure 3.2 is a row-stochastic matrix, while

the matrix corresponding to figure 3.3 is almost a row-stochastic matrix but it contains **dangling nodes**.

$$\mathbf{A} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \end{pmatrix}$$

The nodes with no *outlinks*, which create **0** rows in the matrix are called *dangling nodes*. But all other rows corresponding to non-dangling nodes, therefore creates row-stochastic, thus **A** of figure 3.3 is a *substochastic* matrix.

3.3.5 Limitations and Adjustment

The presence of the dangling nodes in the webgraph create some problems for the random surfer, whenever he enters a dangling node he gets stuck (there are no out-going links). Considering the fact that there will be plenty of dangling nodes on the web, e.g., pages that do not cite other pages, the random surfer must do something to get rid of them. According to theory of Markov chain, the link matrix must be row-stochastic, i.e., the row sum of the link matrix **A** must be equal to **1**. But with existence of dangling node, the row sum of **A** will be **1** but will also be **0** in case of dangling nodes.

In order to fix this problem we need to have an adjustment, the $\mathbf{0^T}$ rows of the link matrix **A** is replaced by $\frac{1}{n}\mathbf{e^T}$. Thus making matrix **A** as row-stochastic. By this rank-one update of the link matrix **A** we will get a new matrix:

$$\mathbf{S} = \mathbf{A} + \mathbf{a}(\frac{1}{n}\mathbf{e}^T)$$

where **a** is a *binary* vector called *dangling node vector* with $\mathbf{a_i} = 1$ if page i is a dangling node (with no out-links) and 0 otherwise.

But still there is another problem; to uniquely compute the PageRank vector the webgraph must be *strongly connected*[2]. Unfortunately it is not always possible that the link matrix **A** will yield a unique PageRank of the webpages. If the enormous webgraph is not connected, containing many *disjoint components* (which is usually the case), it will create difficulty to compute the *unique* ranking of the pages. According to *Ergodic theorem* for Markov chain, the stochastic matrix **A**, should also be *primitive* in order to converge to a *unique stationary distribution vector* (PageRank vector π) [Mot95]. If all states in a Markov chain are *Ergodic* the PageRank vector will be unique. A primitive matrix is both *irreducible* and *aperiodic*. A matrix is irreducible if the graph that this matrix represents is connected, every node is connected to every other node, while aperiodic matrix is the one with *self-loops*.

In order to do another adjustment to the modified link matrix **S**, we need to observe the behaviour of the random surfer (the web user). The random surfer at times get bored, and abandon the hyperlinks surfing, and therefore randomly jumps to another URL e.g., by entering new destination in the browser. This behaviour of the random surfer is recognized as "Teleportation" [Bri98]. When a random surfer teleports, he begins again to surf the hyperlinks, and again gets bored and do another teleportation and so on. This behaviour of the random surfer can be mathematically represented with following matrix:

$$\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\frac{1}{n}\mathbf{e}\mathbf{e}^T \tag{3.6}$$

[2]In graph theory, a directed graph is called *strongly connected* if for every pair of nodes P_A and P_B there is a directed path from P_A to P_B and P_B to P_A.

where $0 \leq \alpha \leq 1$, it's a parameter that controls the proportion of time the random surfer follows the hyperlink as opposed to teleporting. If $\alpha = 0.85$, than 85% of time the random surfer follows the hyperlink structure of the Web, and 15% of time he teleports.

The stochastic matrix $\mathbf{G}$ is the weighted average or more closely the *convex combination* of the link matrix $\mathbf{S}$ and $\mathbf{E} = 1/n\mathbf{e}\mathbf{e}^{\mathbf{T}}$. Notice that matrix $\mathbf{G}$ is now primitive which guarantees a unique and positive PageRank vector π for the entire webgraph. If $\mathbf{G}$ is primitive then Ergodic theorem guarantees that any iterative method applied to the equation (3.6) will converge to a unique and positive PageRank vector (the stationary distribution of the random walk) [Mot95; Hav03b].

But due to the above adjustments now the modified link matrix $\mathbf{G}$ is completely dense, as opposed to the original link matrix $\mathbf{A}$ which was quite sparse. Fortunately we could write $\mathbf{G}$ in terms of the original matrix $\mathbf{A}$, which will yield some computational gains.

$$\begin{aligned} \mathbf{G} &= \alpha\mathbf{S} + (1-\alpha)\frac{1}{n}\mathbf{e}\mathbf{e}^T && (3.7)\\ &= \alpha(\mathbf{A} + \frac{1}{n}\mathbf{a}\mathbf{e}^T) + (1-\alpha)\frac{1}{n}\mathbf{e}\mathbf{e}^T && (3.8)\\ &= \alpha\mathbf{A} + (\alpha\mathbf{a} + (1-\alpha)\mathbf{e})\frac{1}{n}\mathbf{e}^T && (3.9) \end{aligned}$$

The above matrix $\mathbf{G}$ (after two adjustments of matrix $\mathbf{A}$) turns out to resolve the problems identified, and therefore provides a unique PageRank vector π for the entire webgraph. Thus we have

$$\pi^{(i+1)} = \pi^{(i)}\mathbf{G} \tag{3.10}$$

3.3.6 Power Method

The iterative nature of the PageRank computation leads to the choice of different methods to solve it. The available iterative methods could be the famous Gauss-Seidel, Jacobi or GMRES, BICGSTAB, etc [Bar94]. We have the following linear system:

$$\begin{aligned} \pi^{i+1} &= \pi^i\mathbf{G} \\ \pi^T\mathbf{e} &= 1 \end{aligned}$$

The goal is to determine the *dominant* (principal) left eigenvector of $\mathbf{G}$ corresponding to the dominant eigenvalue $\lambda = 1$. The normalization equation $\pi^T\mathbf{e} = 1$ ensures that π is a probability vector as $\mathbf{e}$ is all ones vector. π is stationary vector of Markov chain with transition matrix $\mathbf{G}$, there has been wealth of research on the numerical solution of the Markov chain problem [Bré99; Lan06], they contain over a dozen methods to find the stationary vector π. But the oldest numerical method for computing the principal eigenvector of transition matrix, the *Power method*, conforms more to the peculiar nature of the webgraph's stochastic matrix $\mathbf{G}$. Power method starts with an initial vector π^0, then generate the sequence $\pi^k = \pi^{(k-1)}\mathbf{G}$, ($\pi^k = \pi^0\mathbf{G}^k$) where k approaches ∞ ($lim_{k\to\infty}\pi^{(k)}$), where $\pi^{(k)}$ is the stationary distribution of Markov matrix $\mathbf{G}$.

In linear algebra, the power method is an eigenvalue algorithm, given a matrix $\mathbf{A}$, the algorithm will produce a number λ (the eigenvalue) and a nonzero column vector $\vec{v}$ (the eigenvector), such that $\mathbf{A}\vec{v} = \lambda\vec{v}$.

The intuition behind convergence of Power method is as follows. Assume that the initial vector $\pi^{(0)T}$ lies in the subspace spanned by the eigenvector of $\mathbf{G}$. Now $\pi^{(0)T}$ can be written as a linear combination of the eigenvectors of $\mathbf{G}$:

$$\vec{\pi}^{(0)T} = \vec{u}_1 + \alpha_2\vec{u}_1 + \ldots + \alpha_m\vec{u}_m \tag{3.11}$$

And because the principal or first eigenvalue of the transition matrix of Markov chain $\mathbf{G}$ is $\lambda_1 = 1$

$$\vec{\pi}^{(1)T} = \mathbf{G}\vec{\pi}^{(0)T} = \vec{u}_1 + \alpha_2\lambda_2\vec{u}_1 + \ldots + \alpha_m\lambda_m\vec{u}_m \tag{3.12}$$

Hence,

$$\vec{\pi}^{(k)T} = \mathbf{G}^k \vec{\pi}^{(0)T} = \vec{u}_1 + \alpha_2^k \lambda_2 \vec{u}_1 + \ldots + \alpha_m \lambda_m^k \vec{u}_m \tag{3.13}$$

Since $\lambda_n \leq \ldots \leq \lambda_2 < 1, \mathbf{G}^k \vec{\pi}^{(0)T}$ approaches $\vec{u}_1$ as $k \to \infty$. Thus power method ensures convergence to the principal eigenvector $\vec{u}_1$ of the transition matrix $\mathbf{G}$.

Compared to other iterative methods, power method is generally the *slowest* method. Considering the equation (3.9), where $\mathbf{G}$ is expressed in terms of the sparse matrix $\mathbf{A}$, the power method turns out to be quite appropriate [Lan06]. Power method is at the heart of both the motivation and implementation of PageRank algorithm [Hav03a]. Essentially power method is a *matrix-free*, which means there are less concerns about the handling of the storage issues because of the huge matrix $\mathbf{G}$. There is no extensive manipulation of matrix done in the power method. And we do not store much in each iteration; just the sparse matrix $\mathbf{A}$ and the dangling node vector $\mathbf{a}$ and only one more vector π. Other iterative methods requires storage of at least 10 vectors each iteration [Lan06; Lay94], which in case of the enormous size of the Webgraph is quite expensive. The rate of convergence of the power method is better than other methods, it's hard to find a method that can beat 50 $O(n)$[3] of power iterations [Lan06].

3.3.7 Effects of Random Jump

The idea of *random jump* at the time of teleportation by random surfer provides additional flexibilities to PageRank. Depending on the teleportation probability $(1 - \alpha)$ the surfer visiting any node will jump to a random page. Random jumps offer *multi-faceted* benefits for overall success of PageRank. The random jumps were primarily employed to guarantee the irreducibility of the web's transition matrix. But this model could be used to adapt the PageRank algorithm in numerous situations.

By controlling the random jumps of the surfer we could *bias* the ranking produced by PageRank. In the Section 4.4 we will describe further in detail how we could exploit the random jumps in order to *personalize* the search results according to the user preferences.

The artificial jumps taken by surfer can also reduce the room for *spamming*. A random surfer who jumps oftenly will provide a room for the spammers to inflict the ranking. A high teleport probability means that every page is given a fixed extra "bonus" rank. Link spammers can make use of this bonus to generate local structures to inflate the importance of certain pages. Theoretically the larger values of $\alpha \to \mathbf{1}$ makes PageRank more susceptible to the small perturbation in the link structure of the webgraph. But for larger values of $\alpha \to 1$ the convergence of PageRank also drastically slows down [Bia05].

The random jumps induced by teleportation probability serve to be quite interesting property of PageRank. It could help the PageRank to *personalize* search outcomes (Section 4.4), for *detection* of spamming, to control the rate of *convergence* (depending on the values of α), for *stability* of PageRank to perturbations in link structure, and for the *design* of algorithms to speed up convergence of PageRank (e.g., the second eigenvalue λ_2 of the transition matrix $\mathbf{G}$ is α, see Section 4.3).

3.3.8 The Algorithm

The Algorithm 1 will take as an input the link matrix $\mathbf{A}$, and therefore does the adjustments to make it stochastic and primitive. The power method applied on the modified matrix will then iteratively compute the PageRank vector π_i. The theory of Markov chain (Ergodic Theorem) guarantees that under the conditions

[3]Because each iteration of power method requires $O(n)$ effort, as link matrix $\mathbf{A}$ is quite sparse, and on average it takes about 50 iterations to converge to the principal eigenvector of $\mathbf{G}$.

Algorithm 1 The PageRank Algorithm

1: $\pi_i \Leftarrow 1/n \times ones(1, n)$ {π_i is the PageRank vector}
2: **while** not converged **do**
3: $\pi'_i \Leftarrow \pi_i$ {π'_i is the PageRank vector for previous iteration}
4: $\pi_i \Leftarrow \alpha\ \pi_i\ \mathbf{A} + (\alpha\ (\pi_i\ \mathbf{a}) + (1 - \alpha))\ ((1/n)\ ones(1, n))$
5: {Compute convergence from π'_i & π_i}
6: **end while**

described above the Algorithm 1 is guaranteed to converge to the unique and positive stationary probability distribution vector (dominant eigenvector of Markov matrix).

3.4 HITS (Hypertext Induced Topic Search)

The idea of HITS algorithm was introduced by Kleinberg in 1998[4] [Kle99]. Like PageRank, it works on the hyperlink structure of the web to calculate the relevancy scores of the webpages. Unlike PageRank, HITS first uses the conventional IR system (e.g., text-based search engine) to construct the *sub-network* of pages relevant to a *particular* query, and then computes the relevancy of the pages in the sub-network.

HITS also considers the hyperlinks between the pages as *topical endorsement*, i.e., a link from page u devoted to topic T to another page v is likely to endorse the *authority* of v with respect to topic T [Naj07b]. The central issue that is addressed within this framework is the *distillation* or *characterization* of broad search topic, through the discovery of "authoritative" information sources on such topics [Kle99]. The focus is to extract the important information from the link structure of the web for analyzing the collection of pages relevant to a broad search topic. Hence theoretically, HITS discovers the most "authoritative" pages on those topics, with an expectation that those pages are most central pages for the broad search topics.

3.4.1 Notion of Authority

Given the subnetwork of the documents relevant to a query, how to tell that a given page is *authoritative*? There is no any explicitly endogenous measure of a page that would allow us to properly assess its authoritativeness [Kle99]. Most of the authoritative webpages on some broad topics do not use the *topic-term* in their homepages. For example, for a query "search engines", the most authoritative pages on this query could be Google, AltaVista, Yahoo!, etc., they do not use the term in their main-pages. And there is no reason to expect them to use it.

By closely analyzing the link structure sub-network of the web, we could find potential authorities through the pages that point to them. This is how we could possibly get rid of the problem identified above. The solution lies in the *relationship* that exists between the authorities for a topic and those pages that link to many related authorities – the latter pages are referred to as *hubs* [Kle99].

3.4.2 Authorities and Hubs

Like PageRank, HITS assumes each *link* as an implicit endorsement of the locations that it points to, that is, the links are assumed to confer *authority*. Unlike PageRank, HITS considers another type of pages, the **hubs**, which link to the popular or authoritative pages. Hence there are two types of pages in HITS terminology, *authority* and *hub* pages. HITS therefore produces *two* scores per page, *authority score* and *hub score*.

[4] In almost the same year as that of PageRank was developed (see section 3.3).

It is observed that a certain natural type of *equilibrium* exists between hub and authority weights in the webgraph, which is exploited to develop the HITS algorithm. The algorithm iteratively identifies that equilibrium by identifying both types of the pages (authorities and hubs) simultaneously [Kle99].

3.4.3 Focused Subgraph

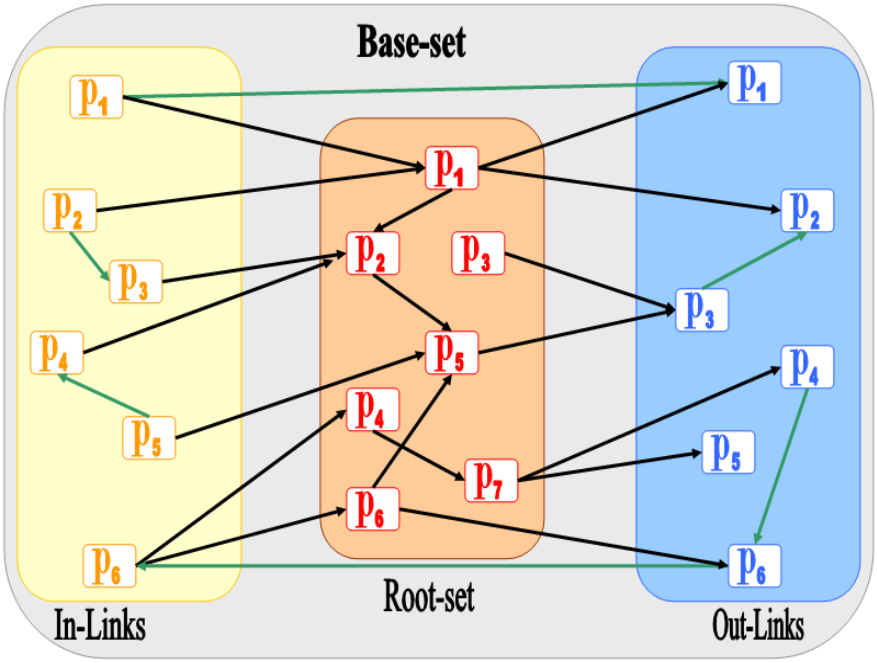


Figure 3.4: *The Base-set*

HITS operates on a *focused subgraph* (the sub-network of Web) instead of the whole webgraph, unlike PageRank. The focused subgraph can be constructed from the output of text-based search engine which can be used to iteratively produce set of pages that are most likely considered to be quite authoritative to the query topic [Kle99].

Suppose a broad-topic query, specified by a query string q, the subgraph corresponding to that broad-topic query will be set Q_q of all pages containing the query term (results of text-based search engine). This set will be relatively large and therefore incurs considerable computational cost. And the set of desired authoritative pages may not belong to this set.

To get a better structured and computational efficient set, k highest ranked pages (typically about 200 pages) will be collected for the query q from the set Q_q. These k highest ranked pages are referred to as the *Root-set* R_q. But still this set R_q is not representative enough to help find the authoritative pages for the query topic. Because there are often extremely few links between pages in R_q, showing that it is essentially "structure-less".

From R_q we can now produce the set S_q (see figure 3.4). The set S_q is referred to as the *Base-set*. To construct the Base-set S_q, we seek a strong authority for the query topic, which may not be in set R_q, but it might be pointed to by some pages in R_q. Hence the set S_q, will be an increase in the number of strong authorities in the Rootset R_q.

The Base-set S_q can be obtained by growing R_q to include any page that pointed to by a page in R_q, and

any page that points to a page in R_q. With a restriction that each page in R_q is allowed to bring in S_q at most d pages (typically $d = 50$) pointing to them [Kle99]. Therefore the resultant set S_q will be computationally beneficial than the Rootset. Typically the Base-set contains $1000 - 5000$ nodes. We will further elaborate on the Base-set when we do experiments, see Section 5.1.1.

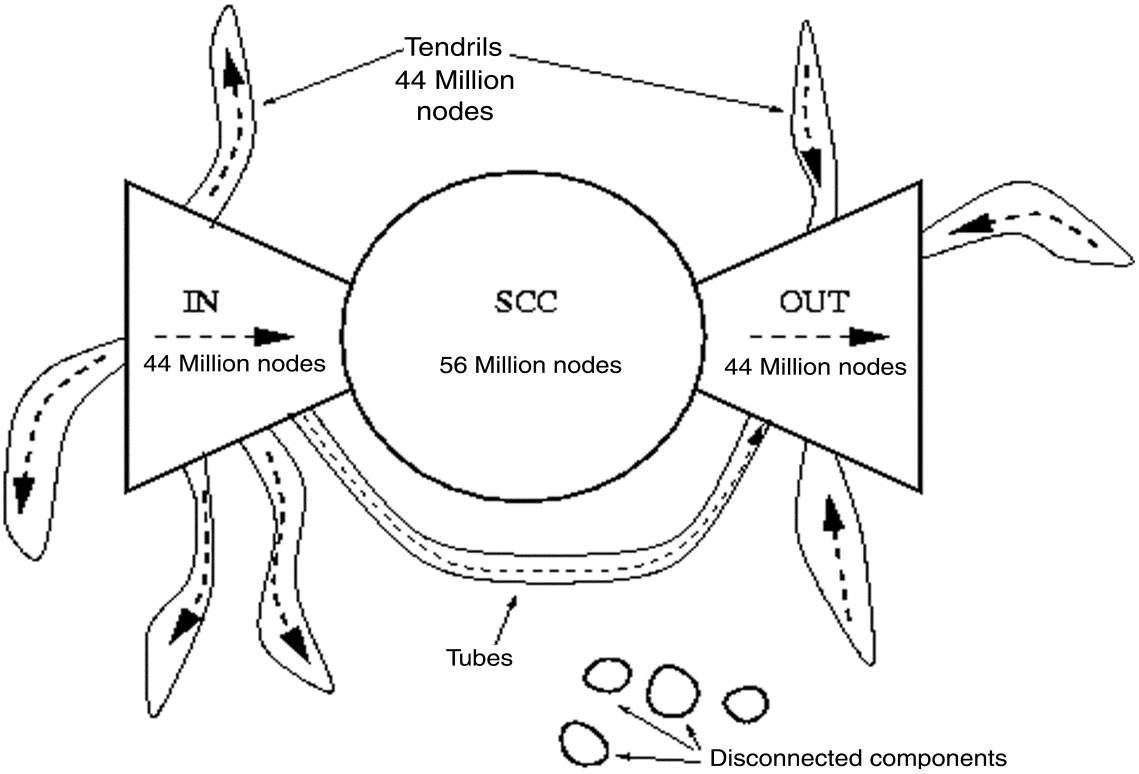


Figure 3.5: *Connectivity of Web, image taken from [Bro00]*

The structure of base-set broadly resembles the overall connectivity structure of the web. The study conducted by Broder et al., [Bro00] was destined to uncover some of the intrinsic properties of the webgraph. They found a macroscopic phenomenon on the entire web, that is, the distribution of degrees on the web (in and out-degrees) are distributed according to *power laws*. The graph structure of the web is found to be characterized by the "bow tie" structure as shown in figure 3.5. Comparing figure 3.5 with figure 3.4, we see that broadly both of them contains three main components. It is therefore quite crucial to observe the behaviour of the Base-set in comparison to the entire webgraph. In the experiments in Chapter 5, we have observed the importance of the Base-set in identifying the set of relevant pages.

3.4.4 Hub and Authority scores

Authoritative pages relevant to the query should not only have high in-degree, they should also have similarities in the sets of pages pointing to them (the hubs). The authorities in the base-set are already authoritative on the common topic (content-wise); we further need to use the neighbourhoods to improve their authoritativeness. Thus pages that point to multiple relevant authoritative pages in the base-set are also important. The hub pages are therefore there to "pull together" authorities on a common topic, and allow us to *disregard* unrelated pages of large in-degree.

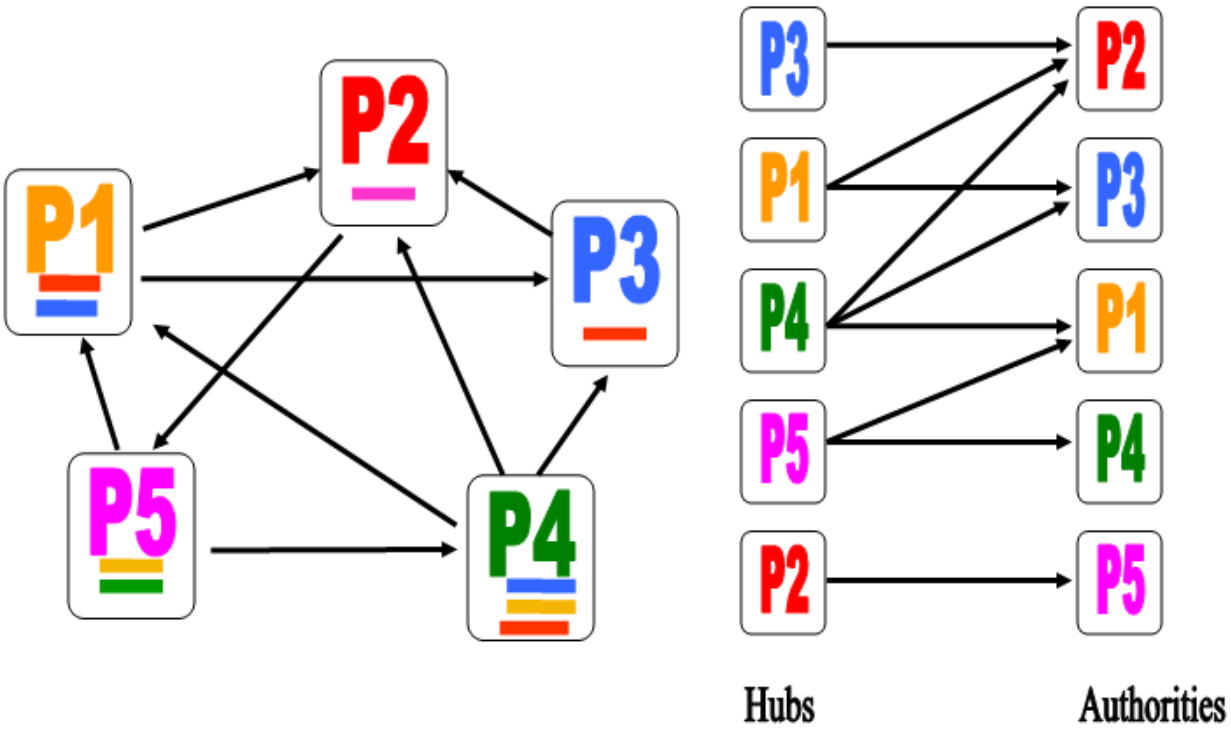


Figure 3.6: *Hubs and Authorities*

Every page therefore has double identity *hub identity* and *authority identity*. Calculating the hub and authority scores of the pages seem circular unless we characterize them as exhibiting a *mutually reinforcing relationship*: "Good authorities are pointed to by good hubs and good hubs points to good authorities" [Kle99].

The hub and authority scores can be computed iteratively for each page p in the base-set, with the help of two operations $\mathbb{I}$ ("in") and $\mathbb{O}$ ("out"), as identified by Kleinberg.

$$\mathbb{I} : a_p \leftarrow \sum_{(q:q\rightarrow p)} h_q \tag{3.14}$$

$$\mathbb{O} : h_p \leftarrow \sum_{(q:q\rightarrow p)} a_q \tag{3.15}$$

$\mathbb{I}$ operation updates and maintains the authority scores while $\mathbb{O}$ operation updates and maintains the hub scores. If page p points to many pages with large hub scores than it should receive a large authority score and vice versa, which signifies the numerical interpretation of the mutually reinforcing relationship.

The operations shown in equations (3.14) and (3.15), will iteratively filter out the top authorities and hubs, and converge from the sequence of vectors $\vec{a}_k$ and $\vec{h}_k$ to the fixed points $\vec{a}^*$ and $\vec{h}^*$.

3.4.5 Principal Eigenvectors

The operations $\mathbb{I}$ and $\mathbb{O}$ in equations (3.14) and (3.15) can be represented in matrix notation with the help of an adjacency matrix $\mathbf{A}$ such that:

$$\mathbf{A_{ij}} = \begin{cases} 1 & \text{if there is a link from page } P_i \text{ to } P_j \\ 0 & \text{otherwise.} \end{cases}$$

For the figure 3.2 the corresponding adjacency matrix $\mathbf{A}$ can be:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

Thus $\mathbb{I}$ and $\mathbb{O}$ in terms of the adjacency matrix can be:

$$\begin{aligned} a &= \mathbf{A^T} h \\ h &= \mathbf{A} a \end{aligned}$$

and also we could further write the above two equations as:

$$\begin{aligned} a &= \mathbf{A^T}\ \mathbf{A}\ a \\ h &= \mathbf{A}\ \mathbf{A^T}\ h \end{aligned}$$

The matrices $\mathbf{A^T}\ \mathbf{A}$ and $\mathbf{A}\ \mathbf{A^T}$ are *symmetric*, *positive semidefinite* and *nonnegative* matrices [Gol96; Lan06]. In *Linear Algebra* the above two equations correspond to the standard problem of finding the **dominant eigenvectors** [5] of the matrices $\mathbf{A^T}\ \mathbf{A}$ and $\mathbf{A}\ \mathbf{A^T}$ respectively. A standard result in linear algebra [Lay94; Gol96] states that if $\mathbf{M}$ is a symmetric matrix and v is a vector not orthogonal to the *principal or dominant eigenvector* of $\mathbf{M}$, then the unit vector in the direction of $\mathbf{M^k}v$ converges to the principal eigenvector of $\mathbf{M}$, as $k \to \infty$. And if $\mathbf{M}$ is nonnegative the principal eigenvector of $\mathbf{M}$ will also be nonnegative. Thus the problem of finding the authority and hub score reduces to the standard problem of finding the principal eigenvectors of matrices $\mathbf{A^T}\ \mathbf{A}$ and $\mathbf{A}\ \mathbf{A^T}$ respectively, thus:

$$a^k = \mathbf{A^T}\ \mathbf{A} a^{k-1} \tag{3.16}$$

$$h^k = \mathbf{A}\ \mathbf{A^T} h^{k-1} \tag{3.17}$$

a^* will be the principal eigenvector of $\mathbf{A^T}\ \mathbf{A}$ and h^* the principal eigenvector of $\mathbf{A}\ \mathbf{A^T}$ as $k \to \infty$ [Kle99]. Thus the symmetric, semidefinite and nonnegative matrices $\mathbf{A^T}\ \mathbf{A}$ and $\mathbf{A}\ \mathbf{A^T}$ assure the *convergence* of HITS algorithm within finite number of iterations of the *Power method*. The rate of convergence is given by the rate at which $(\frac{\lambda_2}{\lambda_1})^k \to 0$ as $k \to \infty$ [Lay94]. For more information on the convergence behaviour, see Section 4.3.2.

Theoretically the convergence of the iterative problem in equations (3.16) and (3.17) is quite rapid, according to [Kle99] one can essentially reach the convergence within $k = 20$. But in the practice the convergence behaviour might vary, from somewhere around $k \in (20 - 160)$ iterations, see for example the results in Table 5.3.

Instead of iterating until convergence, one could have a bounded iterative process, by restricting iterations until some threshold p (number of iterations) and therefore compute scores a^p and h^p instead of a^* and h^*.

The process of finding the principal eigenvectors, equations (3.16) and (3.17), emphasizes the underlying motivation to reinforce the relations $\mathbb{I}$ and $\mathbb{O}$ in equations (3.15).

3.4.6 Non-Unique authority or hub score and Adjustment

Despite the nice formulation of the $\mathbb{I}$ and $\mathbb{O}$ operations and its capability to converge rapidly, there is still one another concern that need to be addressed; that is, the issue of *uniqueness* of authority and hub vectors. Considering the structure of the adjacency matrix $\mathbf{A}$, it might have multiple principal eigenvalues i.e., repeated roots of the *characteristic polynomial, det*$(\mathbf{A^T A} - \lambda I)$. Therefore there will be multiple principal eigenvectors

[5] Because they resemble the standard equation for finding the eigenvalue and eigenvector of a matrix $\mathbf{A}$ i.e. $\lambda x = \mathbf{A}x$

(authority and hub vectors) corresponding to the repeated principal eigenvalues. If we observe more closely the uniqueness problem is actually due to the *reducibility* of the matrix $\mathbf{A}$ [Gol96]. This, in terms of the graph theory means that there are set of states that it's possible to enter, but once entered it's not possible to leave (multiple components in the graph). And that is similar to the problem identified by Page and Brin (see Section 3.3), in the PageRank algorithm.

Irreducible matrices are those in which every *state* is reachable from every other state. Equivalently the matrix $\mathbf{A}$ is irreducible if the graph G is *strongly connected*. And according to the Perron-Frobenius theorem [Kit98; Bré99; Lay94], an irreducible nonnegative matrix possesses a *unique* and *positive* principal eigenvector. A same kind of adjustment as in PageRank, can be applied to the HITS as well, i.e., modify the matrices $\mathbf{A^T}\ \mathbf{A}$ and $\mathbf{A}\ \mathbf{A^T}$, to;

$$\sigma \mathbf{A^T}\ \mathbf{A} + \tfrac{(1-\sigma)}{n}\mathbf{e}\mathbf{e^T}$$

$$\sigma \mathbf{A}\ \mathbf{A^T} + \tfrac{(1-\sigma)}{n}\mathbf{e}\mathbf{e^T}$$

This will ensure that the matrices $\mathbf{A^T}\ \mathbf{A}$ and $\mathbf{A}\ \mathbf{A^T}$ will be primitive and therefore irreducible, based on the same reasons as in PageRank. For more detailed discussion on modifications of HITS, see Section 4.4.

3.4.7 Random Walks and HITS

The natural question to ask here is that whether the output of HITS algorithm can be seen as the *stationary distribution* of certain random walk(s) on the underlying graph (base-set), like PageRank. This has been shown by Borodin et al., [Bor01] that indeed the output of HITS algorithm can also be seen as the stationary distribution of random walk. We present the theorem here without proof.

Theorem 1 *There exist sequences $M_1^a, M_2^a, \ldots, M_n^a, \ldots$, and $M_1^h, M_2^h, \ldots, M_n^h, \ldots$ of Markov Chains, such that, for each $n \geq 1$, the stationary distribution of M_n^a is equal to the authority vector after the n^{th} iteration of Kleinberg's algorithm, and the stationary distribution of M_n^h is equal to the hub vector after the n^{th} iteration of Kleinberg's algorithm.*

According to the fundamental theorem of Markov chains for random walks [Mot95], the stationary distribution of the Markov chain M_n^a, is the same as the authority vector $\vec{a}$ of HITS algorithm. And similarly the hub vector $\vec{h}$ corresponds to the stationary distribution M_n^h. For the proof of the theorem 1 see [Bor01]. Hence the rankings produced by HITS algorithm can be seen as the stationary distribution of two random walks on Base-set.

3.4.8 Singular Value Decomposition

The $\mathbb{I}$ and $\mathbb{O}$ operations can also be viewed in terms of the important factorization technique in linear algebra, the *Singular Value Decomposition (SVD)*. SVD is a powerful technique for data analysis. It reveals a great deal about the structure of the matrix. It is usually applied as a *dimensionality reduction* tool, to obtain a concise and compact representation of a large datasets, mostly applied in image and signal processing and essentially in multimedia IR models. It has also a large scale application in matching the large datasets with each other in IR. In Section 2.6.2, Latent Semantic Indexing model also employed SVD to reduce the dimensionality of the term matrix. Hence it is crucial to have an overview of the use of SVD in LAR models, specifically HITS.

In case of HITS, authority and hub vectors actually represent the *singular vectors* of the adjacency matrix $\mathbf{A}$ [Lay94].

$$\mathbf{A} = \mathbf{U\Sigma V^T} = [\vec{u_1} \;\; \vec{u_2} \;\; \ldots \;\; \vec{u_r}] \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_r \end{bmatrix} \begin{bmatrix} \vec{v_1} \\ \vec{v_2} \\ \vdots \\ \vec{v_r} \end{bmatrix}$$

And the outer product form can be written as;

$$\begin{aligned} \mathbf{A} &= \sigma_1 \vec{u}_1 \vec{v}_1^T + \sigma_2 \vec{u}_2 \vec{v}_2^T + \ldots + \sigma_r \vec{u}_r \vec{v}_r^T \\ \mathbf{A} &= \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^T \end{aligned} \tag{3.18}$$

The matrices $\mathbf{U}$ and $\mathbf{V}$ are *unitary matrices*[6]. The diagonal matrix $\mathbf{\Sigma}$ contains r non-zero singular values $(\sigma_1, \sigma_2, \ldots, \sigma_r)$ which are the square roots of eigenvalues of $\mathbf{A\ A^T}$ and $\mathbf{A^T\ A}$ [Lay94; Gol96]. The vectors corresponding to the columns of matrix $\mathbf{U}$ ($\vec{u_i}$, for $i = 1, 2, \ldots, r$) represent the left singular vectors of matrix $\mathbf{A}$, and they are the eigenvectors of $\mathbf{A\ A^T}$. While vectors corresponding to the rows of matrix $\mathbf{V^T}$ ($\vec{v_i}$, for $i = 1, 2, \ldots, r$) form the right singular vectors of matrix $\mathbf{A}$ and the eigenvectors of $\mathbf{A^T\ A}$. From the above full form of SVD we will have;

$$\begin{aligned} \mathbf{A} &= \mathbf{U\Sigma V^T} \\ \mathbf{A^T} &= \mathbf{V\Sigma}^T\mathbf{U}^T \\ \mathbf{A\ A^T} &= \mathbf{U\Sigma\Sigma}^T\mathbf{U}^T \\ \mathbf{A\ A^T}\vec{u}_i &= \sigma_i^2 \vec{u}_i \quad \text{for } i = 1, 2, \ldots, r \qquad (3.19) \\ \mathbf{A^T\ A} &= \mathbf{V\Sigma}^T\mathbf{\Sigma V}^T \\ \mathbf{A^T\ A}\vec{v}_i &= \sigma_i^2 \vec{v}_i \quad \text{for } i = 1, 2, \ldots, r \qquad (3.20) \end{aligned}$$

As the columns of $\mathbf{U}$ and rows of $\mathbf{V}^T$ are orthogonal (because they are eigenvectors of $\mathbf{A\ A^T}$ and $\mathbf{A^T\ A}$ respectively), therefore any k of them define a basis for $k-$dimensional space. This leads to the observation in the next section.

Linear Trend

It should come as no surprise that extracting the principal eigenvector (component) of matrix $\mathbf{A^T\ A}$ amount to a form of linear trend or factor analysis. Each principal component of $\mathbf{A^T\ A}$ represents an orthogonal direction in which there exist highly-correlated pages. The tightly knit communities therefore correspond exactly to the orthogonal factors of the link matrix $\mathbf{A}$. And hence the most heavily-linked of these tightly knit communities emerge as the principal component or the authoritative sources.

From equations (3.19) and (3.20) we have:

$$\mathbf{AV} = \mathbf{U\Sigma} \tag{3.21}$$

$$\mathbf{A^TU} = \mathbf{V\Sigma} \tag{3.22}$$

To describe the mathematical expressions in equations (3.21) and (3.22), the matrix $\mathbf{A}$ can be thought of as a matrix that associates two different types of entities: *objects* (rows) with *attributes* (columns). Objects

[6]Over the complex field the unitary matrices correspond to the orthogonal matrices. In particular $\mathbf{U} \in C^{m \times n}$ is unitary if $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}_n$ [Lay94; Gol96]

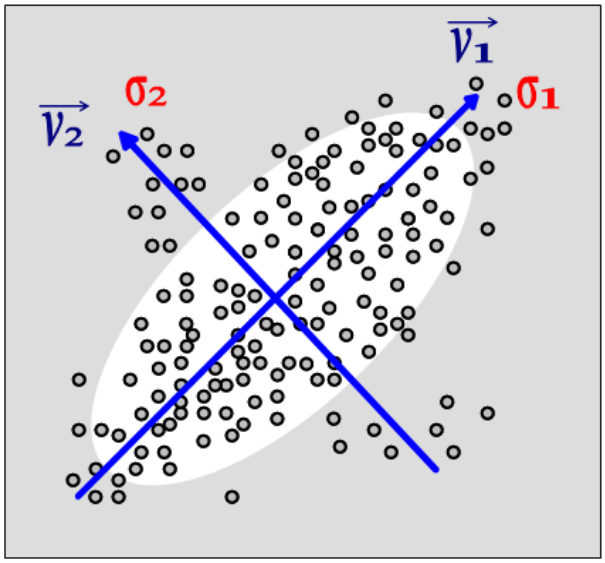


Figure 3.7: *Interpretation of SVD*

are expressed as vectors in attribute space and attributes as vectors in the object space. In equation (3.21) the product $\mathbf{AV}$ defines projection of objects (row vectors) from the attribute space on the *feature space* defined by the column vectors of matrix $\mathbf{V}$. The matrix $\mathbf{V}$ defines the directions on which the object vectors are projected to, and $\mathbf{U\Sigma}$ defines the mapping of the object vectors to the feature space. And similarly the opposite (see equation (3.22)), when $\mathbf{U}$ defines the direction and $\mathbf{V\Sigma}$ defines the position of projection. See figure 3.7 for the graphical representation. In linear algebra, this corresponds to the fact that the columns of matrix $\mathbf{U}$ represent the orthonormal basis for $span(\mathbf{A})$[7]. Similarly equation (3.22) shows that matrix $\mathbf{V}$ correspond to the orthonormal basis for $span(\mathbf{A}^T)$.

From the observations above and equations (3.19) and (3.20), in HITS, the hub vector $\vec{h}$ and the authority vector $\vec{a}$ correspond to the right and left principal singular vectors of $\mathbf{A^T}$ respectively. In the matrix $\mathbf{A^T}$ the $span(\mathbf{A^T})$ correspond to the *hub space*. The $\vec{h}$ vector captures the strongest trend within hub space. The authority vectors are projected on these vectors. The projection lengths are the authority weights and they capture how closely each authority is aligned with the strongest trend $\vec{h}$. Thus authority vector $\vec{a}$ is also the strongest linear trend in the hub space. Similarly in the matrix $\mathbf{A}$ the $span(\mathbf{A})$ correspond to the *authority space*. With the same observation the hub vector $\vec{h}$ is also the strongest linear trend in the authority space.

A vector v is linear trend in matrix $\mathbf{A}$, if it shows the tendency of the row of vectors of $\mathbf{A}$ to align with v. In fact $\vec{u}_i$ and $\vec{v}_i$ are the i^{th} strongest linear trend, and σ_i the strength of i^{th} strongest linear trend [Tsa04a]. Hence SVD is used in IR to discover the linear trends in the data. Relating figure 3.7 and the description above with HITS; it actually tends to discover the strongest linear trend in the *authority space*.

3.4.9 TKC Effect

As described above, we try to map the authority and hub vectors to the hub and vector space of spanned by column and rows of matrix $\mathbf{A}$ respectively. An intriguing question is to understand how the structure of the

[7] $\dim(\text{span}(\mathbf{A}^T)) = \dim(\text{span}(\mathbf{A})) = \text{rank}(\mathbf{A}^T) =$ Number of singular values in $\mathbf{\Sigma}$

graph affects the ranking behaviour of the algorithm. And considering the fact that HITS computes a densely linked collection of pages without regard to their content [Kle99], it is expected to favour pages that belong to most dense component of the graph (either in authority or hub). According to [Bor05; Lem00] in a graph containing multiple communities (densely connected components), the HITS algorithm will only focus on one of them in the top positions of the ranking, the one that contains the hubs and authorities that are most *tightly interconnected*. This phenomenon is therefore termed as *Tightly Knit Community* (TKC) effect [Bor05].

It has been analyzed and stated that HITS favour TKC effect [Kle99; Cha99; Dri99]. In fact the TKC effect is the by-product of the properties of *Singular Value Decomposition*, as we described in the previous section. The tendency of SVD to discover the linear trends in data, and HITS likes the strongest linear trend. Therefore, from the figure 3.7, the resultant authority and hub vectors tend to be in the *densest* component of the graph. Thus the tightly interconnected set of nodes in the graph try to attract HITS algorithm, and HITS gets attracted, therefore possibly causes "topic drift"[8].

3.4.10 Algorithm

Algorithm 2 The HITS Algorithm

1: $\mathbf{A}$: adjacency matrix formed from the base-setS_q
2: $a^{(0)}$: set the seed values of the authority vector
3: $h^{(0)}$) : set the seed values of the hub vector
4: **while** not converged **do**
5: $\mathbb{I}$: $a^k \Leftarrow \mathbf{A^T A} a^{k-1}$
6: $\mathbb{O}$: $h^k \Leftarrow \mathbf{A A^T} h^{k-1}$
7: $a^{k'} \Leftarrow a^k$ Normalize
8: $h^{k'} \Leftarrow h^k$ Normalize
9: $k \Leftarrow k + 1$
10: {Compute the convergence}
11: **end while**

Like PageRank algorithm, in the Algorithm 2 we need the adjacency matrix $\mathbf{A}$ as an input. There is another algorithm employed to construct the base-set S_q, corresponding to a query q, from root-set R_q, and from the set Q_q, see [Kle99]. Constructing the adjacency matrix $\mathbf{A}$ from the base-set S_q will then be straight forward. Also note that the above iterative method is the famous *Power method* as used in PageRank. It also requires the starting or seed vectors $a^{(0)}$ and $h^{(0)}$, a general rule is that it should be in the range of the characteristic polynomial i.e., $det(\mathbf{A^T A} - \lambda I)$. By the construction of the algorithm (the power method), any arbitrarily chosen non-negative seed vectors $a^{(0)}$ and $h^{(0)}$ (such as $h^{(0)} = 1/ne$) also convergences to an eigenvector of the largest eigenvalue [Gol96].

In the Chapter 4 we will further analyze in detail the intriguing properties of HITS and the possibility to improve and / or adjust some of the limitations.

3.5 SALSA (Stochastic Approach for Link-Structure Analysis)

SALSA, the *stochastic* approach towards LAR also exploits mainly the link-structure of the web, was developed by Lempel and Moran, 2000. SALSA makes use of the Kleinberg's mutual reinforcement approach and the *ran-*

[8]When the results are not relevant to the user need or query, than the pages does not correspond to the query topic, i.e. the results drifted the topic therefore topic drift.

dom walk of the Markov chain theory as is used in PageRank. The initial assertion was to employ an algorithm which would computationally be more efficient than the *mutual reinforcement* approach of Kleinberg [Lem00].

Like HITS, SALSA creates both *hub* and *authority* scores for each document on the Web, and like PageRank, the scores are calculated using the theory of Markov chains [Lan06]. However, in contrary to HITS's mutual reinforcement approach, SALSA compute hub and authority scores on the neighbourhood graph (focused subgraph) by performing two *independent* random walks.

3.5.1 Informative and Non-informative links

Considering the link between pages p and q, $(p \rightarrow q)$, p implicitly suggests, or recommends that the page q also contains a relevant topic of interest as in p. Such type of link is called *informative* link. In the wake of Sections 3.3 and 3.4, page p *endorses* page q, and therefore the page q is considered important and hence a candidate to be visited by a surfer visiting page p.

But all the links cannot be categorized as informative links; unfortunately there are many links which confer little or no *authority*. For example, *intra domain links* or *navigational links*, commercial or sponsor links, and links which result from link-exchange agreements [Lem00]. A crucial task prior to applying any link analysis based ranking is to *identify* the informative and non-informative links, and therefore prune the non-informative links from the graph. These pre-processing to the graph create a much realistic link structure between the pages, which could be used to compute a more relevant and pragmatic set of authorities.

3.5.2 Bipartite Graph (Hubs and Authorities)

In the stochastic approach (SALSA), the coupling between the hubs and authorities are less tight then HITS. It actually considers an undirected bipartite graph $\mathbf{G}$[9], whose two parts correspond to hubs and authorities respectively (see figure 3.8). An edge between the hub P_i and the authority P_j means that there is an informative link from P_i to P_j. And therefore pages pertaining to dominant topics in $\mathbf{G}$ should be highly reachable from many pages (nodes in the bipartite graph).

Considering the graph of five pages in figure 3.2, the example graph in figure 3.8 depicts the hubs and authorities corresponding to that graph.

In the stochastic approach of analyzing the link structure, the theory of *random walks* of Markov chain are combined with the notion of hubs and authorities, the two sides of the bipartite graph (figure 3.8). Therefore the algorithm conducts two different and independent random walks corresponding to the hubs and authorities respectively. But unlike the conventional random walks, state transitions in these chains are generated by traversing one link *forward* and one link *backward* [Lem00].

The undirected graph $\mathbf{G} = (V_h, V_a, E)$ (e.g., the figure 3.8), can be constructed from the *site-collection* (the base-set $\mathbf{S_q}$, given a query q, see Section 3.4). We call the base-set here as the site-collection $\mathbf{C}$.

$$\begin{aligned} V_h &= \{s_h \mid s \in \mathbf{C} \; and \; out-degree(s) > 0\} \text{ (hub side of } \mathbf{G}) \\ V_a &= \{s_a \mid s \in \mathbf{C} \; and \; in-degree(s) > 0\} \text{ (the authority side of } \mathbf{G}) \\ E &= \{(s_h, r_a) \mid s \rightarrow r \in \mathbf{C}\} \text{ (the set of directed edges in } \mathbf{C}) \end{aligned}$$

From above formulation, in our example graph in figure 3.8 we have:

$$\begin{aligned} V_h &= \{\ P_3,\ P_1,\ P_4,\ P_5,\ P_2\ \} \\ V_a &= \{\ P_2,\ P_3,\ P_1,\ P_4,\ P_5\ \} \end{aligned}$$

[9]The graph $\mathbf{G}$ can be constructed from the neighbourhood graph made in similar way as in HITS, focused subgraph (see Section 3.4.3)

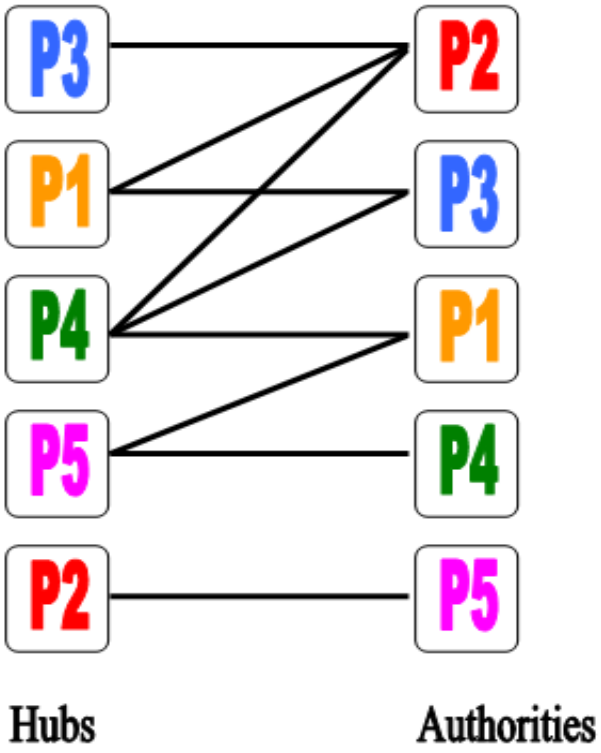


Figure 3.8: *Undirected bipartite graph* **G**

Recall the adjacency matrix **A** of figure 3.2, which can be used in this case too:

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix}$$

And now assume two new matrices $\mathbf{A_r}$ and $\mathbf{A_c}$ derived from matrix **A** by normalizing it's entries in some way. Matrix $\mathbf{A_r}$ is formed by normalizing each *row* of **A** to sum to 1. Similarly $\mathbf{A_c}$ is formed by normalizing the entries of **A** such that for each *column* the sum of entries is 1. For our example figure 3.2 we have:

$$\mathbf{A_r} = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \end{pmatrix}, \quad \mathbf{A_c} = \begin{pmatrix} 0 & 1/3 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1/3 & 0 & 0 & 0 \\ 1/2 & 1/3 & 1/2 & 0 & 0 \\ 1/2 & 0 & 0 & 1 & 0 \end{pmatrix}$$

3.5.3 Two Random Walks

On the bipartite graph in figure 3.8, two *distinct* random walks are conducted. Each walk is meant to traverse the pages on each one of the two sides.

For the first walk, random surfer starts off uniformly at random from any node on authority side. The walk will than proceed alternating between *backward* and *forward* steps. When the surfer is at the authority side of the graph, one of the incoming links is selected uniformly at random and move to the hub side of the graph.

Similarly when at the hub side of the graph, one of the outgoing links is selected uniformly at random and move to the authority side of the graph. And hence the authority weights vector will be the stationary distribution of this random walk, starting from authority side. This first random walk will be conducted on the set V_a (the set of authority nodes), described above.

The second walk will be conducted on the set V_h. Here the surfer starts off from the any node on hub side. And proceeds by alternating between *forward* and *backward* steps. When the surfer is at hub side it will select one of the outgoing links uniformly at random and move to authority side and vice versa. The hub weights vector like authority vector will be the stationary distribution for this second random walk, starting from hub side.

With this settings, we may expect to have t-authorities relevant to topic t amongst the nodes most frequently visited by the random walk on V_a, and similarly we expect to have t-hubs will nodes amongst the most frequently visited by the second random walk on V_h nodes.

3.5.4 Stochastic Matrices

The transition matrices of the two Markov chains representing the random walks on hub and authority side of the bipartite graph can be defined mathematically here.

Refer to the walk starting off from authority side of graph conducted on the set V_a, the Markov chain of this random walk has transition probabilities:

$$\tilde{\mathbf{a}_{ij}} = \sum_{\{k|(i_h,k_a),(j_h,k_a)\in\mathbf{G}\}} \frac{1}{deg(i_h)} \bullet \frac{1}{deg(k_a)} \tag{3.23}$$

In terms of *backward* and *forward* hops of random surfer we can mathematically have:

$$\tilde{\mathbf{a}_{ij}} = \sum_{\{k|k\in B(i)\cap\ B(j)\}} \frac{1}{|B(i)|} \bullet \frac{1}{|F(k)|} \tag{3.24}$$

For some node i we have $B(i) = \{j : \mathbf{A}(j,i) = 1\}$, i.e., $B(i)$ is the set of back-links for node i. And $F(i) = \{j : \mathbf{A}(i,j) = 1\}$, is the set of forward-link for node i.

The random walk conducted on the *authority graph*[10] will have a probability $\tilde{\mathbf{a}_{ij}}$ when the surfer moves from authority i to j.

Similarly for the random walk that starts from hub side, the transition probabilities are:

$$\tilde{\mathbf{h}_{ij}} = \sum_{\{k|(i_h,k_a),(j_h,k_a)\in\mathbf{G}\}} \frac{1}{deg(i_h)} \bullet \frac{1}{deg(k_a)} \tag{3.25}$$

And simplifying to *forward* and *backward* hops of random surfer, we have:

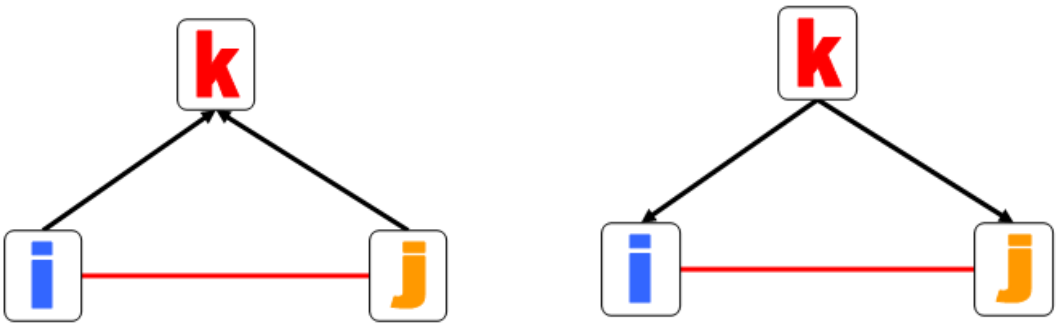
$$\tilde{\mathbf{h}_{ij}} = \sum_{\{k|k\in F(i)\cap\ F(j)\}} \frac{1}{|F(i)|} \bullet \frac{1}{|B(k)|} \tag{3.26}$$

Same interpretation of random walk on *hub graph*[11].

A positive transition probability $h_{ij} > 0$ implies that a certain page k is pointed to by both the pages i and j, and hence page j is reachable from page i by two steps: *forward-link* on $i \to k$ and then *back-link* from

[10] When authority i and j share a hub, we place an undirected edge between authority i and j, the subsequent graph is the authority graph.

[11] When hub i and j share an authority , we place an undirected edge between hub i and j, the underlying graph will be hub graph



(a) i and j has common hub k (b) i and j has common authority k

Figure 3.9: *Hub and Authority graph respectively*

$k \rightarrow j$ [Lem00]. See figure 3.9(a). Similarly $a_{ij} > 0$ implies that a certain page k points to both pages i and j (k is their common authority). Thus i is reachable from j by two steps: *back-link* from $j \rightarrow k$ and then *forward-link* on $k \rightarrow i$. See figure 3.9(b).

Consider now the matrices $\mathbf{A_r}$ and $\mathbf{A_c}$, the stationary distribution vector $\mathbf{a}$ of Markov chain corresponding to random walk on authority graph is actually the principal right eigenvector of matrix $\mathbf{A_c^T}\ \mathbf{A_r}$ i.e.,:

$$\mathbf{a} = \mathbf{A_c^T}\ \mathbf{A_r}\ \mathbf{a} \tag{3.27}$$

where the vector $\mathbf{a}$ is a column vector.

And the stationary distribution vector $\mathbf{h}$ (a column vector) of the Markov chain on hub graph is the principal right eigenvector of matrix $\mathbf{A_r}\ \mathbf{A_c^T}$ i.e.,:

$$\mathbf{h} = \mathbf{A_r}\ \mathbf{A_c^T}\ \mathbf{h} \tag{3.28}$$

SALSA can also be interpreted in terms of the operations in HITS. As in HITS, the $\mathbb{I}$ operation, hubs broadcast their weights to the authorities, and authorities *sum up* the weights of hubs that point to them. SALSA instead of broadcasting, in $\mathbb{I}$ operation, each hub divides its weight equally among its' out-links (the authorities that it points to). And in $\mathbb{O}$ operation of SALSA algorithm each authority equally divides it's weight among its' back-links (the hubs). Thus the modified operations in SALSA are now:

$$\mathbb{I} : a_p \leftarrow \sum_{(q:q\in B(p))} \frac{1}{|F(q)|} h_q \tag{3.29}$$

$$\mathbb{O} : h_p \leftarrow \sum_{(q:q\in F(p))} \frac{1}{|B(q)|} a_q \tag{3.30}$$

Lempel considers the matrix $\tilde{\mathbf{A}}$ to have the non-zero rows and columns of $\mathbf{A_r}{\mathbf{A_c}}^T$ and the matrix $\tilde{\mathbf{H}}$ to have the non-zero rows and columns of ${\mathbf{A_c}}^T{\mathbf{A_r}}^T$. The graph $\mathbf{G}$ is assumed to be connected, which therefore causes the matrices $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{H}}$ to be *irreducible*.

The stochastic matrices $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{H}}$ are therefore *primitive*, because the Markov chains which they represent are aperiodic [Lem00]. Therefore every node on either side of the bipartite graph has a self loop, causing the chain to be aperiodic. The matrices are also *symmetric*, considering matrix $\tilde{\mathbf{A}}$, if $\tilde{a}_{ij} > 0$ then $\tilde{a}_{ji} > 0$, furthermore $a_{ij} > 0$ if and only if $[\mathbf{A}^T\mathbf{A}]_{ij} > 0$. Similarly for matrix $\tilde{\mathbf{H}}$, if $\tilde{h}_{ij} > 0$ then $\tilde{h}_{ji} > 0$. Hence $\tilde{h}_{ij} > 0$

if and only if $[\mathbf{A}\mathbf{A}^T]_{ij} > 0$. Where matrix $\mathbf{A}$ is the original adjacency matrix corresponding the site-collection $\mathbf{C}$ or the base-set.

Like PageRank and HITS, the problem of *reducibility* can cause problems in SALSA too. This means that it is possible to have authority graph which is not connected and similarly the same could happen with hub graph. Or in general the graph $\mathbf{G}$ can be not-connected (it may have disconnected components). Under such circumstances simply multiply each page's authority score by the normalized size of the irreducible component to which it belongs, for details see [Lem00].

But when the graph $\mathbf{G}$ is connected i.e., it contains a single component, the stochastic authority(hub) weights will be proportional to the in(out) degree. SALSA therefore reduces to primitive algorithm, the InDegree, when $\mathbf{G}$ has a single connected component only.

3.5.5 Algorithm

Algorithm 3 The SALSA Algorithm

1: $\mathbf{A}$: adjacency matrix formed from the base-set S_q
2: $\mathbf{A_c}$: matrix $\mathbf{A}$ where *columns* normalized to sum to $\mathbf{1}$
3: $\mathbf{A_r}$: matrix $\mathbf{A}$ where *rows* normalized to sum to $\mathbf{1}$
4: $\tilde{\mathbf{A}}$: non-zero rows and columns of $\mathbf{A_c^T A_r}$
5: $\tilde{\mathbf{H}}$: non-zero rows and columns of $\mathbf{A_r A_c^T}$
6: a^0 : set the initial value of the authority vector
7: h^0 : set the initial value of the hub vector
8: **while** not converged **do**
9: $\quad \mathbb{I}: \ a^k \Leftarrow \tilde{\mathbf{A}}\ a^{k-1}$
10: $\quad \mathbb{O}: \ h^k \Leftarrow \tilde{\mathbf{H}}\ h^{k-1}$
11: $\quad a^{k'} \Leftarrow a^k$ Normalize
12: $\quad h^{k'} \Leftarrow h^k$ Normalize
13: $\quad k \Leftarrow k+1$
14: $\quad$ {Compute the convergence}
15: **end while**

The Algorithm 3 reflects all the findings that we had in the previous sections. We have experimented with the Algorithm 3 in Chapter 5. SALSA is considered to be less vulnerable to TKC effect (described in Section 3.4.9). Najork [Naj07a] had conducted experiments on large Webgraph to evaluate HITS and SALSA. He found that SALSA "substantially outperformed" HITS. The results that we have found in our study are a bit different, in our own specific settings. See Appendices A and B for comparison.

Part II

Evaluations, Analyses and Experiments

Chapter 4

Evaluations and Analyses of LAR Models

4.1 Link Analysis Ranking Process

The presence of (hyper-) link information clearly augmented a great deal to the characterization of the *informative content* present in the documents. LAR approaches are intended to resolve some of the intrinsic weaknesses of the content based IR models. Through the analysis of network or web structure of the documents (due to citation structure) LAR approaches bring in a whole new horizon to IR space.

The essence of LAR therefore is that the "overall information" of a hyperlink database of documents is not composed of only static "textual information", but also another, the "hyper" information.

The LAR models rely on the documents database with index structure having entries equipped with *hyperlink information*. The first step therefore towards analyses of the LAR algorithms is to have a hyperlinked collection of documents. From the hyperlinked information stored in index structure we create a hyperlink or *adjacency graph*, similar to the one described in last chapter, figure 3.2. In this graph each document is represented as a *node* and each hyperlink is represented as a *directed edge* between the nodes (documents) similar to the webgraph description in Section 3.3.

Usually in most of the LAR models, the edges between the documents are un-weighted, i.e., the costs of transition between the pages are assumed to be *uniform*. But there are models that work on weighted graphs, the study conducted by Bharat and Henzinger [Bha98] concerns with weighted graph by using measures from content analysis of the documents, together with LAR. In another study by Cohn and Hofmann [Coh01], a joint probabilistic model has been introduced for modelling the contents and hyperlinks. Hence probabilistic theories are used to make predictions about the existence and strengths of hyperlinks and citations.

The graph can be generated for the entire collection of documents independent of any query (*query independent*) or it can be generated from a subset of the documents collection based on a given query (*query dependent*). *PageRank* as discussed in Section 3.3, works on whole webgraph irrespective to any query. While *HITS* and *SALSA* in Sections 3.4 and 3.5 work on a smaller or focused subgraph based on a given query. This idea of query (in)dependence correspond to the initial perspectives of IR models (see descriptively in Section 2.2 and graphically in figure 2.1).

The resultant hyperlinked graph (query dependent or independent) will be given as an input to the LAR algorithms. This graph will be encoded in an adjacency matrix $\mathbf{A}$, where $\mathbf{A}[i, j] = 1$ if there is a link from node i to node j and 0 otherwise (see also Section 3.3.3). The LAR algorithm iteratively operate on the hyperlinked graph (the adjacency matrix $\mathbf{A}$) and returns the rank vector (n-dimensional) $\vec{x}$ with weights computed for each

node in the graph, where x_i is the weight of i^{th} node in the graph. The weights are actually the *probabilities of relevance* of each document to the user query. They are therefore used to rank the retrieved documents (sort by weights). LAR algorithms are intended to *discover* ranking of documents which *maximize* the relevance of user query to the ranked and retrieved documents. LAR algorithms are thus meant to discover *authoritative* documents through analyzing the hyperlink graph [Tsa04a].

Many interesting ideas have been brought in LAR space over the years. Starting from *Bibliometrics* [Lar96], then moving towards making more active use of "hyper" information together with "textual" information [Mar97], and then exclusively using only hyperlink information with objectives to maximize relevancy between user query and search outcomes. The application of core concepts from linear algebra, such as; the Markov chain theory (as described in Chapter 3), LAR as a system of linear equations, the use of Singular Value Decomposition, Dynamic programming, Extrapolation techniques, Optimization algorithms (such as Gradient Ascent), and many more intriguing ideas have been investigated and incorporated in LAR.

In the Chapter 3 we have reviewed the existing research in LAR. There were broadly two families of LAR algorithms, *query-independent* algorithms and *query-dependent* algorithms. In the former case, the algorithm ranks the whole document corpus independent of any query. While the latter ranks a subset of the document corpus depending on some specific query. The two pioneer algorithms PageRank and HITS are query-independent and query-dependent respectively. They were followed by substantial amount of research [Bri06; Hav03a; Jeh03; Kam03b; Dri99], to name just a few. IR practitioners now have a wider range of LAR algorithms to choose from.

In this chapter our focus is to purposely explore the intrinsic and implicit properties of query-dependent LAR algorithms. We will also study some of interesting properties and characteristics of query-independent algorithms and try to incorporate them in the query-dependent ones. We will go through different improvements (both in query-dependent and independent) and try to relate them to inherent properties of query-dependent algorithms particularly HITS and SALSA. We will also study the behaviours of different algorithms. The ideas discussed in this chapter and Chapter 3 will be supplemented by the empirical analyses, in Chapter 5.

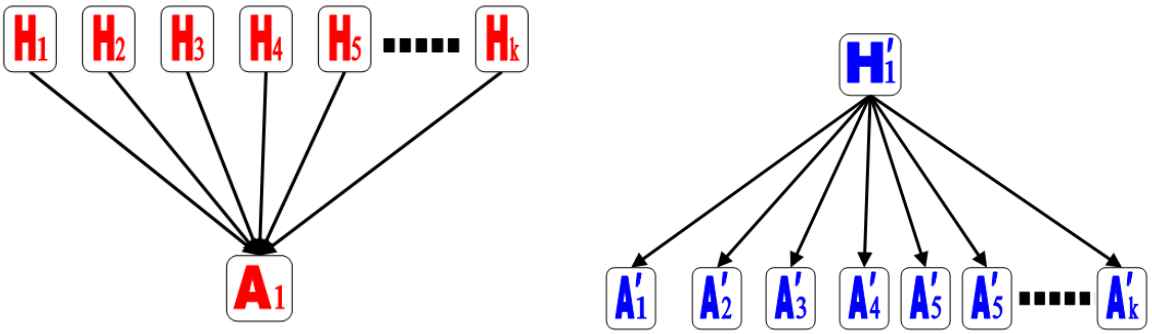
4.2 Implicit Properties of HITS algorithm and Problems

HITS algorithm can be analyzed from its inherent idea, i.e., the *mutually reinforcing relationship* between hubs and authorities. "A good authority is the one that is pointed by many good hubs and a good hub is the one that points to many good authorities" [Kle99]. Thus, hubs and authorities mutually strengthen each other. The quality of hub depends on the quality of the pages *pointed to* by this hub (the authorities) and the quality of authority depends on the quality of pages *points to* this authority (the hubs). Hub and Authority weight are calculated in a cyclic fashion; hub weights are calculated from authorities and authority weights are calculated from hubs.

The association described above between hub and authority weights is captured through the *addition operation* [Kle99]. Hub weight of a page is calculated as the sum of the authority weights of the pages that pointed to by this hub, similarly authority weight of a page is calculated as the sum of the hub weights of the pages that point to this authority node. As identified by [Tsa04a], the definition of mutually reinforcing relationship has *two implicit properties*. First that it is *symmetric*, because both hub and authority weights are calculated in same way. This means that if the orientation of graph is reversed then authority and hub weights of original graph will be hub and authority weights of reversed graph respectively. Secondly that it is *egalitarian*, while computing the authority weight of a page, the hub weights of the pages that point to this page are all treated

equally. Similarly, when calculating the hub weight of a page, the authority weights of the pages that it points to are considered equally. We will further elaborate these two properties with a help of example graphs.

4.2.1 Example



(a) Common Authority (b) Common Hub

Figure 4.1: *Problems in HITS.*

In figure 4.1, we see two components, the one in figure 4.1(a) with a central authority and the other one in figure 4.1(b) with a central hub. In the red component (figure 4.1(a)) many hubs points to one authority while in the blue component (figure 4.1(b)) one hub points to many authorities. If the blue hub points to more authorities than the number of red hubs, HITS will allocate all authority weight to the blue authorities and giving zero weight to the red authority. It is because the blue hub is considered as best hub and therefore the blue authorities receive more weight than the red authority.

Intuitively the red authority seems to be better than blue authorities and should have been ranked higher. But HITS produced a non-intuitive result and therefore ranked the blue authorities higher than the red one. This means that by *linking to* a lot of pages it will end up in a non-intuitive result. And therefore HITS is susceptible to *link spamming*.

Symmetry relationship between hubs and authorities treats both hub and authority the same. A node with high *in-degree* is likely to be a good authority, but a node with high *out-degree* should not necessarily be a good hub. Because with such assumption there are high possibilities of spamming. In such a situation it would be very easy to inflict ranking by artificially uplifting the hub weights by adding links to a lot of pages. In the Section 4.2.2 we will present some approaches to cope up with this problem, and some more problems identified there.

There are similar kind of problems identified by [Bha98]. Three problems are identified in connectivity analyses of HITS.

Firstly the *mutually reinforcing relationships* between hosts; certain arrangements of pages conspire to dominate and inflict the computation of ranking. For example, a set of pages on one host point to a single page on a second host, same as the red component in the figure 4.1(a). This drives up the hub scores of the pages on the first host and the authority score of the page on the second host [Bha98]. The reverse case is similar to the blue component in the figure 4.1(b). Secondly the *automatically generated links*; to insert links automatically or spam links in order to conspire the ranking. These two problems are the same as described in the figure 4.1.

Thirdly the *non-relevant documents*; it is highly probable that the focused subgraph (the base-set) contain

non-relevant nodes to the query topic. If the non-relevant nodes are well connected then the *topic drift* problem can arise, which means that the most highly ranked authorities and hubs will not be relevant to query topic. The consequences of the tightly connected non-relevant documents on ranking are exaggerated by the TKC effect of HITS, see Section 3.4.9.

4.2.2 Approaches to address the problems

The problems identified in the last section have significant side-effects on HITS algorithm. The example showed that blue hub (figure 4.1(b)) can inflict the hub and authority weights by linking to a lot of pages. The blue hub will have more score if it points to a lot of authorities irrespective of its authoritativeness. And similarly the blue authorities will get more score from a hub pointing to a lot of authorities. This means that there are definite loopholes for the spammers, to illicitly use link spamming to *conspire* the ranking.

Hub Average

As proposed by [Tsa04a] a modification to HITS algorithm could possibly solve the problems identified in figure 4.1. There is one natural modification needed in $\mathbb{O}$ operation, i.e., the hub weight of a node i is set to be the average authority weights of the authorities pointed by i. Hence,

$$a_i = \sum_{j \in B(i)} h_j \quad \text{and} \quad h_i = \frac{1}{|F(i)|} \sum_{j \in F(i)} a_j \tag{4.1}$$

where $B(i)$ is the set of *back-links* of node i, while $F(i)$ is the set of *forward-links* for node i. $|F(i)|$ is cardinality or the number of the forward-links.

This modification was originally presented by [Bor01]. The resultant algorithm is called as *HubAvg* (Hub Average). The intuition is that a good hub should point only (or at least mainly) to good authorities [Tsa04a]. In the example identified in the last section HubAvg assigns the same weights to both the red and blue hubs, and it identifies the red authority as the better authority than the blue ones.

HubAvg can be seen as 'hybrid' of HITS and SALSA algorithm [Bor01] (see Section 3.5). The $\mathbb{O}$ operation is the same as HITS, i.e., *broadcasting* the authority weight to hubs, and $\mathbb{I}$ operation is like SALSA, i.e., *dividing* the hub score to the authorities. If the same modification in $\mathbb{O}$ operation can be applied to $\mathbb{I}$ operation too, i.e., both hub and authority nodes divides their weights (instead of broadcasting) to the subsequent nodes, then HITS algorithm will reduce to SALSA algorithm, as described in Section 3.5.4.

Expressing these changes in matrix notation, we have:

$$\mathbf{A_r} = \mathbf{FA} \tag{4.2}$$

where $\mathbf{A}$ is the adjacency matrix and $\mathbf{F}$ is a diagonal matrix with $\mathbf{F(i,i)} = 1/|F(i)|$. While $F(i)$ is the set of *forward-links* as defined before. Thus we have:

$$\mathbf{M_{HA}} = \mathbf{A^T A_r} \tag{4.3}$$

$$\mathbf{M_{HA}} = (\mathbf{F^{\frac{1}{2}} A})^T (\mathbf{F^{\frac{1}{2}} A}) \tag{4.4}$$

The modified authority weight will be the principal right eigenvector of the matrix $\mathbf{M_{HA}}$. The power iterations for authority vector a will be:

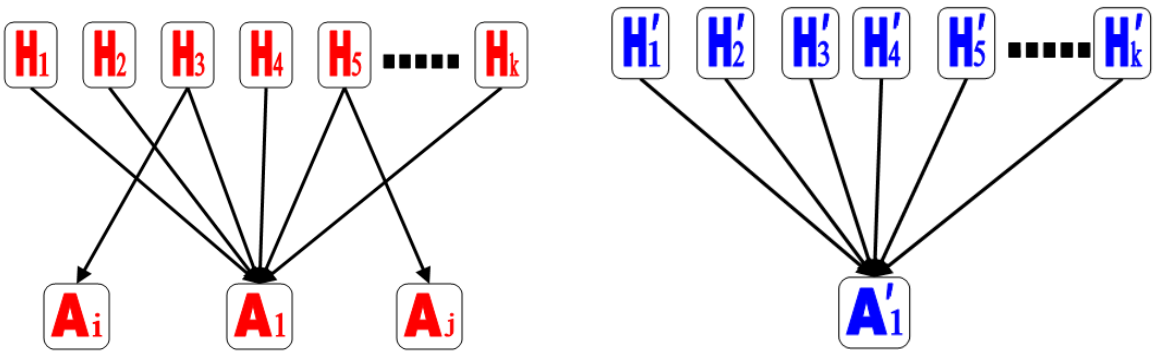
$$a^k = \mathbf{M_{HA}} a^{k-1} \tag{4.5}$$

Similarly, the hub weight vector will be the principal right eigenvector of the matrix $\mathbf{M_{AH}}$, hence:

$$\mathbf{M_{AH}} = (\mathbf{F}^{\frac{1}{2}}\mathbf{A})(\mathbf{F}^{\frac{1}{2}}\mathbf{A})^T \tag{4.6}$$

$$h^k = \mathbf{M_{AH}}\ h^{k-1} \tag{4.7}$$

Another Example



(a) Hubs pointing to weak authorities

(b) Common Authority

Figure 4.2: *Another problem in HITS.*

The solution provided by HubAvg still contains a loophole. In figure 4.2(a) there are two extra red authorities (A_i and A_j) that are pointed to by the two hubs (H_3 and H_5) respectively. HubAvg will assign more weight to the blue authority than the red one, although they are identical. The red hubs are *penalized* for pointing to the extra two weak authorities. And because they have divided their weights among the authorities that they point to, so, the share of the hubs that point to the weak authorities is less to the red authority in centre, A_1. Thus the red authority A_1 will be weighted less than the blue authority $A_1^{'}$. This is also an undesirable situation, a node which is simultaneously a strong hub on one topic while a weak hub on another topic, should not be penalized for being a weak hub on another topic.

Authority threshold

The effect of such a situation can be lessened simply by applying a *threshold operator*, that is, to retain only the highest authority weights (depending on threshold). *Authority-threshold* algorithm sets the hub weight of node i to be the sum of $k-$largest authority weights of the authorities pointed to by i. This means a good hub points to at least k good authorities.

$$a_i = \sum_{j \in B(i)} h_j \quad \text{and} \quad h_i = \sum_{j \in F_k(i)} a_j \tag{4.8}$$

where $F_k(i)$ is a subset of $F(i)$ containing k nodes with highest authority weights. This means that for any node $p \in F(i)$, such that $p \notin F_k(i)$, than $a_p \leq a_q \ \ \forall\ q \in F_k(i)$.

Max operator

The extreme values of k in authority threshold algorithm provide interesting insight into the algorithm. If k is greater than or equal to the maximum *out-degree* of the graph $\mathbf{G}$ then the authority threshold algorithm

reduces to the original HITS algorithm. And for $k = 1$ we get a *max* operator (only the topmost authority weights is set to hub score). *Max algorithm* is to say that a hub node is as good as the *best* authority that it points to [Tsa04b]. Applying Max algorithm to the example in figure 4.2(a), the red hubs pointing to the weak authorities would only keep the authority score of the authority node in centre A_1.

Norm family

We want to *scale* the weights, so that lower authority weights contribute less to the hub weight, instead of applying threshold or max operator. Through their weights we can determine the scaling factors. This idea is implemented in the *Norm family* of algorithms. The hub weight of node i is set to the $p - norm$ of the vector of the authority weights of the nodes pointed to by node i [Tsa04b; Tsa04a]. Therefore we have:

$$a_i = \sum h_j \qquad \text{and} \qquad h_i = (\sum a_j^p)^{1/p} \tag{4.9}$$

The extreme values of p in the Norm algorithm also provides interesting insights. For $p \in [1, \infty]$ as we increase p, the value of the $p - norm$ in hub vector is more and more dominated by the highest weights. For example for $p = 1$ the norm algorithm reduces to HITS, and for $p = \infty$ we will have max operator [Tsa04b]. And for $p = 2$ the hub vector will be the *Euclidean norm* of authority weights.

The modifications that are applied to only $\mathbb{O}$ operation of HITS, can also be applied to its $\mathbb{I}$ operation, symmetrically. For example, applying Norm operator symmetrically on both hub and authority vectors (instead of just hub vector) will have the same effects on the authority vector as it had on hub vector (in the discussion above). Max operator applied only to hub vector enables the hubs to only receive the top most authority scores of the nodes that they point to. And symmetrically applying max operator on the authority vector enables authority nodes in the graph to receive only the topmost hub scores of the nodes that point to them. Applying max operator on authority vector in figure 4.2(a); the red authority A_1 in centre will only receive the topmost hub score instead of accumulating all the hub scores.

It requires further analyses to observe the differences between the results of symmetrical application of operators (discussed in last paragraph) and improvements described earlier in equations (4.5), (4.8) and (4.9). Comparing them is out of scope of this study, but we will use the ideas described in this section as a ground to explore improvements in convergences and personalization of the results. In Chapter 5, we would extensively experiment with the algorithms defined in this section together with improvements in the coming sections.

4.3 Extrapolation Techniques to accelerate the Convergence

In this section we will probe into a new and genuine formulation of query-dependent LAR using a technique called *Extrapolation*. The Extrapolation techniques can be used to accelerate the convergence of Power Method (see Section 3.3.6). The technique is novel as it offers a new approach of taking into consideration important properties of the iterative method for effectively accelerating the computation of the query-dependent family of algorithms (e.g., HITS, its improvements and SALSA).

In mathematics, **Extrapolation** is the process of constructing new data points outside a discrete set of *known* data points. It is similar to the process of *Interpolation*, which constructs new points between known points, but its results are often less meaningful, and are subject to greater uncertainty. Interpolation is a specific case of curve fitting, in which the function must go exactly through the data points. In case of convergence, Extrapolation techniques can be employed to accelerate the convergence by using the known data points (values

from successive iterates) to construct new data points (principal eigenvector(s)). Techniques for accelerating the *convergent series* are often applied in *numerical analysis*, where they are used to improve the speed of numerical integration, and other well-known series.

Fast convergence and efficient computational speed in *query-dependent* algorithms are quite crucial, because of the fact that they operate on *query time*. They should therefore be proficient in terms of time they consume to converge and computations per iteration. The response time in query-dependent algorithms have a direct impact on the overall interaction between user and IR system. For example for an adjacency matrix of size 7000×7000, it is fairly expensive to compute the operation $\vec{x}^k = \mathbf{A}\vec{x}^{k-1}$ several times as $k \to \infty$.

The standard power method is used to compute the principal eigenvector and eigenvalue by computing the successive iterates until convergence. In the power method implementation of the query-dependent algorithms (see Section 3.4) the computations per iteration are usually expensive, depending on the size of matrix. There is a dire need for reducing the number of iterations as there is also an essential need to improve the computations per iteration too.

Many algorithms have been proposed over the years to improve the computational abilities of the power method (for example, Power method with Rayleigh quotient, shifted power method, inverse power method, etc) [Lay94; Gol96]. The fast eigenvector solvers and other numerical improvements involving vigorous use of the matrices are unsuitable for LAR problem, because of the size and sparsity of the matrices.

In this section we will primarily focus on investigating efficiencies in terms of the rate of convergence. In the extrapolation techniques, we largely take advantage of the fact that the principal eigenvalue of the *Markov matrix* is 1 in theory. This information can be used to compute the estimates of the nonprincipal eigenvectors during the iterations. Through the *estimates* computed during the successive iterates of Power method, we expect to extrapolate the value of the principal eigenvector.

Extrapolations techniques were previously used by Kamvar et al., [Kam03b], specifically tailored to the PageRank problem, but in our study we have employed it in the query-dependent algorithms such as HITS, its improvements and SALSA.

4.3.1 Aitken's Δ^2 Extrapolation

A technique called *Aitken's* Δ^2 (three-points) extrapolation can be used to speed up the convergence of any sequence that is *linearly convergent*[1] [Cio88]. Aitken Extrapolation method makes use of another popular method called *fixed point iteration*. If we have an equation of the form $x = f(x)$ we can attain the expected solution simply by iterating through the sequence, $x(k+1) = f(x(k))$. Or more specifically for a given function f defined on *real numbers* and a given initial point x_0 in the domain of f, the fixed point iteration is:

$$x_{k+1} = f(x_k), k = 0, 1, 2, \ldots \tag{4.10}$$

The series $x_0, x_1, \ldots$ are expected to converge to x. If the function f is continuous, then x is a *fixed point* of f, i.e., $x = f(x)$.

The equation (4.10) is the standard fixed point iteration. Now consider the standard power iteration method, we will get a correspondence with fixed point iteration, i.e.,:

$$\vec{x}^{(k)} = \mathbf{A}\vec{x}^{(k-1)} \tag{4.11}$$

[1] A sequence $\{x_i\}$ is said to converge linearly to x^* if there is constant $1 > c > 0$ such that $||x_{i+1} - x^*|| \leq c||x_i - x^*||$ or alternatively $||x_{i+1} - x_i|| \leq c||x_i - x_{i-1}||$

Let us consider f in equation (4.10) as an iterative numerical process, then the intermediate iterates of the linear convergent series, $x_i,\ x_{i+1}$ and x_{i+2} can be used to extrapolate the fixed point x. This three-point extrapolation scheme is well known as Aitken Δ^2 extrapolation.

Aitken Δ^2 extrapolation is oldest and most popular extrapolation technique. It forms the basis for other extrapolation techniques. It has also been used to speedup the convergence of the Power method for faster computation of PageRank [Kam03b].

In power method, Aitken acceleration computes the principal eigenvector of the Markov matrix in *one step*, under the assumption that the power iteration estimate $\vec{x}^{(k-2)}$ can be expressed as the *linear combination* of the first two eigenvectors, $\vec{u}_1$ and $\vec{u}_2$.

$$\vec{x}^{(k-2)} = \vec{u}_1 + \alpha_2\ \vec{u}_2 \tag{4.12}$$

where $\vec{u}_1$ is the principal eigenvector and $\vec{u}_2$ is the second eigenvector of Markov matrix in power method.

The equation (4.12) shows that from the nonprincipal eigenvectors (the values of $\vec{x}^{(k-2)}$ from successive iterates), we can extrapolate the value of the principal eigenvector $\vec{u}_1$. The previous values calculated in the successive iterates could be used to extrapolate the new value (the new data point outside the known data points), the principal eigenvector. This way we could accelerate the rate convergence of the already convergent series produced by Power Method.

The Aitken extrapolation step when applied periodically, enables us to subtract off the estimates of the nonprincipal eigenvectors from the current iterates $\vec{x}^{(k)}$. For the derivation of Aitken acceleration and the empirical proof that it can extrapolate the principal eigenvector for power method see [Kam03b; Cio88]. Aitken extrapolation technique is crucial primarily because the subsequent extrapolation techniques build upon the ideas advocated in this technique. It serves to provide a general *premise* for extrapolation. It is therefore essential to have a sound appreciation of this technique to comprehend the newer more sophisticated techniques of extrapolation used for accelerating convergence.

In a nutshell we want to use the *priori knowledge* (which we acquire from the prior iterates) and use that knowledge as a *basis* to extrapolate the new and better value (the principal eigenvector). We use the assumption that the new iterate(s) can be expressed as a linear combination of the *last few* iterates. With some changes to this basic assumption various extrapolation techniques can be formulated (for example, *Quadratic Extrapolation* assumes that last three iterates $\vec{x}^{(k-3)},\ \vec{x}^{(k-2)}$ and $\vec{x}^{(k-1)}$ together with current iterate $\vec{x}^{(k)}$ can be used to express the new and improved iterate value, see next section). In case of Aitken Extrapolation we are using three-points $\vec{x}^{(k-2)},\ \vec{x}^{(k-1)}$ and $\vec{x}^{(k)}$ to extrapolate the next point $\vec{u}_1$.

The extrapolation methods are different from standard fast eigensolvers, which mostly relies on the matrix factorization and/or matrix inversion. The extrapolation methods that we study here rely upon the fact that the principal (first) eigenvalue of the Markov matrix is, $\lambda_1 = 1$, in order to find an approximation to the principal eigenvector.

What is crucial here is to identify *theoretically* and *empirically* that the extrapolation methods accelerate the convergence, and the computed value is actually the principal eigenvector. In the section below we will provide theoretical proof for the importance of *Quadratic extrapolation* and its capability to extrapolate the principal eigenvector of the Markov matrix. In Chapter 5 we will provide experimental evidences of the capabilities of Extrapolation to improve the convergent sequence in query-dependent LAR algorithms.

4.3.2 Quadratic Extrapolation

In this section we will closely examine *Quadratic Extrapolation technique* [Kam03b]. Like Aitken Δ^2 Extrapolation, Quadratic extrapolation also uses the idea of taking the *linear combination* of last few iterates. Unlike Aitken, Quadratic extrapolation uses the *first three* (instead of first two) eigenvectors of Markov matrix to express the iterate ($\vec{x}^{(k-3)}$). Therefore, it assumes that the iterate $\vec{x}^{(k-3)}$ can be expressed as the *linear combination* of first three eigenvectors ($\vec{u}_1$, $\vec{u}_2$ and $\vec{u}_3$) of Markov matrix.

These assumptions in Quadratic extrapolation enable us to approximate the principal eigenvector in *closed form*[2] using iterates $\vec{x}^{(k-3)}$, $\vec{x}^{(k-2)}$, $\vec{x}^{(k-1)}$ and $\vec{x}^{(k)}$.

Formulation

From the theory of Power method, we know that the seed vector $\vec{x}^{(0)}$ can be expressed as the linear combination of *all* the eigenvectors of the Markov matrix (see Section 3.3.6). Thus,

$$\vec{x}^{(0)} = \vec{u}_1 + \alpha_2\vec{u}_1 + \ldots + \alpha_m\vec{u}_m$$

While in the specific settings of Quadratic Extrapolation it is assumed that the Markov matrix $\mathbf{A}$ in equation (4.11) has only 3 eigenvectors. Based on this assumption the iterate $\vec{x}^{(k-3)}$ can be expressed as linear combination of these 3 eigenvectors.

The quadratic extrapolation can be formulated from this premise that the matrix $\mathbf{A}$ has only 3 eigenvectors and therefore we can approximate the iterate $\vec{x}^{(k-3)}$ as:

$$\vec{x}^{(k-3)} = \vec{u_1} + \alpha_2\vec{u_2} + \alpha_3\vec{u_3} \tag{4.13}$$

The assumption is not in contrast to reality rather it is used to form the much stronger relation later in the derivation. Of course, the matrix $\mathbf{A}$ can have more than 3 eigenvectors. Later in the section we will form a relation based on this assertion. In the experimental analyses in Chapter 5 we have also provided empirical results verifying the validity of this assumption. It is shown in the experiments that Quadratic extrapolation derived from this assumption provides much better rate of convergence than the original algorithms (see Appendix A).

Now we are in a position to derive the required model using equation (4.13). From the assumption (equation (4.13)), the characteristic polynomial $p_A(\lambda)$ of the Markov matrix $\mathbf{A}$ can now be written as:

$$p_A(\lambda) = \gamma_0 + \gamma_1\lambda + \gamma_2\lambda^2 + \gamma_3\lambda^3 \tag{4.14}$$

Since the Markov matrix $\mathbf{A}$ is stochastic, we know from the theory of Markov chain that the first eigenvalue of $\mathbf{A}$ is $\lambda_1 = 1$ [Gol96; Lay94]. Thus:

$$p_A(\lambda = 1) = 0 \Rightarrow \gamma_0 + \gamma_1 + \gamma_2 + \gamma_3 = 0 \tag{4.15}$$

According to *Cayley-Hamilton theorem* [Gol96] any matrix $\mathbf{A}$ satisfies it's own characteristic polynomial, i.e., $p_A(\mathbf{A}) = 0$. Therefore multiplying $\vec{x}^{(k-3)}$ with the characteristic polynomial, we have:

$$p_A(\mathbf{A})\vec{x}^{(k-3)} = [\gamma_0\mathbf{I} + \gamma_1\mathbf{A} + \gamma_2\mathbf{A}^2 + \gamma_3\mathbf{A}^3]\vec{x}^{(k-3)} = 0 \tag{4.16}$$

[2] An *equation* or *system of equations* is said to have a *closed-form solution* if, and only if, at least one solution can be expressed analytically in terms of a bounded number of certain "well-known" functions [wik].

This can be simplified as:

$$\gamma_0\vec{x}^{(k-3)} + \gamma_1\vec{x}^{(k-2)} + \gamma_2\vec{x}^{(k-1)} + \gamma_3\vec{x}^{(k)} = 0 \tag{4.17}$$

Since we knew from the power iterations:

$$\vec{x}^{(k-2)} = \mathbf{A}\vec{x}^{(k-3)} \dots \quad \vec{x}^{(k)} = \mathbf{A}\vec{x}^{(k-1)}$$

From the above equations and equation (4.17), after simple steps we have:

$$\vec{x}^{(k-3)}(-\gamma_1 - \gamma_2 - \gamma_3) + \gamma_1\vec{x}^{(k-2)} + \gamma_2\vec{x}^{(k-1)} + \gamma_3\vec{x}^{(k)} = 0 \tag{4.18}$$

This can be further simplified as:

$$(\vec{x}^{(k-2)} - \vec{x}^{(k-3)})\gamma_1 + (\vec{x}^{(k-1)} - \vec{x}^{(k-3)})\gamma_2 + (\vec{x}^{(k)} - \vec{x}^{(k-3)})\gamma_3 = 0 \tag{4.19}$$

Define the following:

$$\vec{y}^{(k-2)} = \vec{x}^{(k-2)} - \vec{x}^{(k-3)} \tag{4.20}$$

$$\vec{y}^{(k-1)} = \vec{x}^{(k-1)} - \vec{x}^{(k-3)} \tag{4.21}$$

$$\vec{y}^{(k)} = \vec{x}^{(k)} - \vec{x}^{(k-3)} \tag{4.22}$$

Inserting equation (4.22) in equation (4.19) gives:

$$\vec{y}^{(k-2)}\gamma_1 + \vec{y}^{(k-1)}\gamma_2 + \vec{y}^{(k)}\gamma_3 = 0 \tag{4.23}$$

and

$$\left(\vec{y}^{(k-2)} \;,\; \vec{y}^{(k-1)} \;,\; \vec{y}^{(k)}\right)\vec{\gamma} = 0 \tag{4.24}$$

For the solution of the above system we don't want to have the trivial solution $\gamma = 0$, thus we constrain the leading term of the characteristic polynomial γ_3 as:

$$\gamma_3 = 1 \tag{4.25}$$

After substituting the value of γ_3, equation (4.24) can be written as:

$$\left(\vec{y}^{(k-2)} \quad \vec{y}^{(k-1)}\right)\begin{pmatrix}\gamma_1\\ \gamma_2\end{pmatrix} = -\vec{y}^{(k)} \tag{4.26}$$

Hence we have an *overdetermined* system of linear equations:

$$\begin{pmatrix}\gamma_1\\ \gamma_2\end{pmatrix} = -\mathbf{Y}^{\dagger}\vec{y}^{(k)} \tag{4.27}$$

Here $\mathbf{Y}^{\dagger}$ is the *pseudoinverse* of the matrix shown in the left side of the equation (4.26), $(\vec{y}^{(k-2)},\ \vec{y}^{(k-1)})$. From the equation (4.14) and above equations we can therefore find the coefficient of the characteristic polynomial $q_A(\lambda)$.

We may divide the characteristic polynomial with $(\lambda-1)$ to get $q_A(\lambda) = p_A(\lambda)/(\lambda-1)$. Hence we have now:

$$q_A(\lambda) = \frac{(\gamma_0 + \gamma_1\lambda + \gamma_2\lambda^2 + \gamma_3\lambda^3)}{(\lambda - 1)} = \beta_0 + \beta_1\lambda + \beta_2\lambda^2 \tag{4.28}$$

By polynomial division and after some simple algebraic operations we get the values for beta:

$$\beta_0 = \gamma_1 + \gamma_2 + \gamma_3 \tag{4.29}$$
$$\beta_1 = \gamma_2 + \gamma_3 \tag{4.30}$$
$$\beta_2 = \gamma_3 \tag{4.31}$$

By Cayley-Hamilton theorem, for any vector $\vec{z}$ in $\mathbb{R}^n$ we also have:

$$q_A(\mathbf{A})\vec{z} = \vec{u}_1 \tag{4.32}$$

where $\vec{u}_1$ is the principal eigenvector of matrix $\mathbf{A}$ corresponding to eigenvalue $\lambda_1 = 1$. Thus by letting $\vec{z} = \vec{x}^{(k-2)}$:

$$\vec{u}_1 = q_A(\mathbf{A})\vec{x}^{(k-2)} \tag{4.33}$$
$$= [\beta_0 + \beta_1\mathbf{A} + \beta_2\mathbf{A}^2]\vec{x}^{(k-2)} \tag{4.34}$$
$$= \beta_0\vec{x}^{(k-2)} + \beta_1\mathbf{A}\vec{x}^{(k-2)} + \beta_2\mathbf{A}^2\vec{x}^{(k-2)} \tag{4.35}$$

Using the power iterations in above equations:

$$\vec{x}^{(k-2)} = \mathbf{A}\vec{x}^{(k-3)} \ldots \quad \vec{x}^{(k)} = \mathbf{A}\vec{x}^{(k-1)}$$

Thus we get the closed form solution for $\vec{u}_1$, as:

$$\vec{u}_1 = \beta_0\vec{x}^{(k-2)} + \beta_1\vec{x}^{(k-1)} + \beta_2\vec{x}^{(k)} \tag{4.36}$$

The equation (4.36) together with equations (4.29) - (4.31) and equation (4.27) can be used to implement the Quadratic extrapolation. Hence together they will help to provide an approximation to the principal eigenvector of the Markov matrix $\mathbf{A}$. The above derivation steps are inspired from the work in [Kam03b; Cio88].

Discussion

In the above formulation in equation (4.27) we have to solve an *overdetermined system* of linear equations:

$$\begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} = -\mathbf{Y}^\dagger\vec{y}^{(k)}$$

The above overdetermined system can be solved through any *least-square method*, for example, through the QR factorization of *Gram-Schmidt algorithm* [Gol96; Lay94]. The Quadratic extrapolation technique can be further optimized by applying a better solver to the overdetermined system above.

The Quadratic Extrapolation improves convergence much better than the original rate of convergence of the algorithm, based on the empirical results (see Chapter 5 and Appendix A). In the slow convergent series, the Quadratic Extrapolation is proved to be an effective technique. For example, if the second eigenvalue of the Markov matrix is close to 1, i.e., $\lambda_2 \to 1$, theoretically and empirically the convergence of power method tends to slow down. In such a situation the slow converging sequence of Power method can be accelerated radically by Quadratic Extrapolation (see Section 5.2.1).

The important thing in Quadratic Extrapolation is that it should be applied periodically. Once the parameters for Extrapolation (such as $\vec{x}^{(k-3)}$, $\vec{x}^{(k-2)}$, $\vec{x}^{(k-1)}$ and $\vec{x}^{(k)}$) are ready we could either apply Quadratic extrapolation step immediately or apply it at any other instance with appropriate values. It doesn't necessarily

need to be applied too often to achieve maximum benefit. Experiments in Chapter 5 reveal interesting insights about the potential of Quadratic Extrapolation in various settings. By manipulating the periodic application of Quadratic extrapolation we can administer the convergence behaviour of an algorithm.

Theoretically, Quadratic extrapolation technique is used to *subtract off* the errors in the current iterate along the direction of the second and third eigenvectors, as mathematically represented by equation (4.13). By doing that it enhances the convergence for the future application of the power method. The approximate principal eigenvector as a result of Extrapolation step serves as a good approximation for the further iterates, which help to converge much faster. For more detailed discussion on the experimental manifestation of the extrapolation techniques see Section 5.2.1.

In the next section we will explore another interesting technique for extrapolation, where some important properties of the Markov matrix are exploited to make a more generic and cleaner formulation of Extrapolation.

4.3.3 Power $(\mathbf{A}^d)$ Extrapolation

Based on the ideas initially put forward in Aitken Extrapolation and Quadratic Extrapolation discussed in previous sections Haveliwala et al., [Havb] construct another interesting formulation of extrapolation. Similar to both Aitken and Quadratic extrapolation, here also by subtracting off the errors along several nonprincipal eigenvectors from the current iterates, it is intended to accelerate the rate of convergence of Power method. Not just relying on the values of successive iterates but other important properties of Markov matrix, i.e., the *nonprincipal eigenvalues* could be exploited to accelerate the convergence.

In linear algebra, finding the nonprincipal eigenvalues of a Markov matrix is a problem in itself. Calculating nonprincipal eigenvalues of Markov matrix may increase the computational overheads, instead of providing any improvements. Thus apparently the idea of using the nonprincipal eigenvalues for acceleration may not seem conducive in general. But in a study by Haveliwala and Kamvar [Hava], they discovered interesting insights about the nonprincipal eigenvalues of the Markov matrix in PageRank algorithm. They have proved that the modulus of second eigenvalue of the Markov matrix (or the *Google matrix*) is given by the damping factor 'c' in PageRank algorithm (see Section 3.3, where $\alpha = c$, in the original formulation). Thus if the row stochastic matrix $\mathbf{S}$ in equation (3.6) has at least two irreducible closed subsets, then the second eigenvalue of $\mathbf{S}$ is given by:

$$\lambda_2 = c$$

Note that the webgraph can have many eigenvalues with modulus of 'c' (i.e., one of c, $-c$, ci, and $-ci$). These eigenvalues of the webgraph has been exploited in the power extrapolation to approximate the principal eigenvector of the hyperlink matrix $\mathbf{A}$ corresponding to the webgraph. In the next section we will briefly formulate the use of second eigenvalue for accelerating the convergent series of power method, and in Chapter 5 we will also present experimental findings for this approach.

Formulation

In Power Extrapolation the iterate $\vec{x}^{(k-2)}$ can be represented as linear combination of three eigenvectors ($\vec{u}_1, \vec{u}_2$ and $\vec{u}_3$) of Markov matrix. Making use of the same assumption as it was in *Quadratic Extrapolation*. Thus:

$$\vec{x}^{(k-2)} = \vec{u_1} + \alpha_2 \vec{u_2} + \alpha_3 \vec{u_3} \tag{4.37}$$

The nonprincipal eigenvalues corresponding to nonprincipal eigenvectors $\vec{u_2}$ and $\vec{u_3}$ are 'c' and '$-c$' respectively, according to the results in [Hava]. From Power Iterations we have:

$$\vec{x}^{(k)} = \mathbf{A}^2\vec{x}^{(k-2)} \quad (\text{since} \quad \vec{x}^{(k)} = \mathbf{A}\ (\mathbf{A}\vec{x}^{(k-2)}) \quad \text{as} \quad \vec{x}^{(k-1)} = \mathbf{A}\vec{x}^{(k-2)})$$

Substituting the above relations in equation (4.37) we get:

$$\vec{x}^{(k)} = \mathbf{A}^2(\vec{u_1} + \alpha_2\vec{u_2} + \alpha_3\vec{u_3}) \tag{4.38}$$

$$\vec{x}^{(k)} = \vec{u_1} + \alpha_2\lambda_2^2\vec{u_2} + \alpha_3\lambda_3^2\vec{u_3} \tag{4.39}$$

We can replace $\lambda_2 = c$ and $\lambda_3 = -c$ in equation (4.39), thus we have:

$$\vec{x}^{(k)} = \vec{u_1} + c^2(\alpha_2\vec{u_2} + \alpha_3\vec{u_3}) \tag{4.40}$$

$$\vec{x}^{(k)} = \vec{u_1} + c^2(\vec{x}^{(k-2)} - \vec{u_1}) \tag{4.41}$$

From equation (4.41) the closed form of the approximated principal eigenvector $\vec{u_1}$ will be:

$$\vec{u_1} = \frac{\vec{x}^{(k)} - c^2\vec{x}^{(k-2)}}{1 - c^2} \tag{4.42}$$

The above derivation leads to $\mathbf{A}^2$ Extrapolation which subtract off error along the eigenspaces corresponding to eigenvalues c and $-c$. There is a general derivation as well to the case where the eigenvalues of modulus of c given by $c \times d_i$, where d_i are the d^{th} root of unity, are used to form a generalized closed form based on variable d [Havb]. For example, for $d = 4$ the nonprincipal eigenvalues of modulus of c are given by c, ci, and $-ci$, which means 4^{th} roots of unity.

The generalized case will have the closed form of the principal eigenvector $\vec{u_1}$ as:

$$\vec{u_1} = \frac{\vec{x}^{(k)} - c^d\vec{x}^{(k-d)}}{1 - c^d} \tag{4.43}$$

For details about the derivation of equation (4.43) see [Havb].

The implementation of $\mathbf{A}^d$ Extrapolation is much more simpler than the Quadratic Extrapolation, just to implement equation (4.43). Theoretically the overhead due to $\mathbf{A}^d$ extrapolation is negligible since it is applied only *once*. The convergence is also found to be similar to the Quadratic Extrapolation, but the wallclock-speedup is higher in $\mathbf{A}^d$ extrapolation [Havb]. In contrast to the findings by Haveliwala et al., our results show that the convergence behaviour due to power extrapolation in query-dependent algorithms are not comparable to that of Quadratic extrapolation in the empirical settings, see Chapter 5

Discussion

In the case of PageRank the *Eigengap* $(1 - |\lambda_2|)$ for the Markov matrix $\mathbf{A}$ is given exactly by the teleportation probability $1 - c$, in accordance to the findings by Haveliwala et al., discussed above. Theoretically, if the second eigenvalue λ_2 is close to 1, then the convergence of the power method will be slow. Because of the fact that convergence of the power method depends on $|\lambda_2|/|\lambda_1|$ factor, and k must be fairly large before $(|\lambda_2|/|\lambda_1|)^k$ converge to $\mathbf{0}$ [Lay94; Gol96]. In PageRank reducing the factor $|\lambda_2|/|\lambda_1|$ correspond to the reduction of the damping factor c, because we can only change the numerator which is $|\lambda_2|$. But reduction of damping factor means increase in teleportation $(1 - c)$ and hence increasing the chance for the spammers to inflict the rankings. However a high teleportation probability $((|\lambda_2|/|\lambda_1|) \rightarrow 1)$ constitutes slow convergence of power method.

Under such circumstances it is highly rewarding to accelerate the slow convergence of power method. The methods of Extrapolation discussed above supposedly provide a faster convergence even with high teleportation probabilities.

The $\mathbf{A}^d$ Extrapolation is based on second eigenvalue of the Markov matrix. In HITS and other query-dependent algorithms such as SALSA, the power extrapolation cannot be directly applicable. The second eigenvalue of the Markov matrix should be calculated independently. It is not a trivial problem to compute the nonprincipal eigenvalues of Markov matrix in general case. The teleportation scheme cannot be directly employed in the query-dependent algorithms, unless we incorporate the damping factor the same way as we did in PageRank (see Section 3.3.5).

There is a phenomenon called *Deflation* in linear algebra which can be used to compute the nonprincipal eigenvalues of the Markov matrix. Deflation is a technique of reducing the dimension of Markov matrix corresponding to its dominant eigenvalue. Once we compute the dominant eigenvalue, the Markov matrix can be reduced to one lower dimension (by subtracting the column corresponding to the dominant eigenvalue). We can now compute the dominant eigenvalue of the reduced dimension matrix, which will be the second eigenvalue of the original Markov matrix. Repeat this process until all the eigenvalues are computed. Finding the non-principal eigenvalues of Markov matrix is therefore not an easy task. The main techniques for dimensionality reduction are principal component analysis (PCA) and the famous singular value decomposition [Lay94] .

In the Section 4.4 we will discuss one way to incorporate the damping factor into the query-dependent HITS and possibly SALSA too. After that modification it will be then possible to apply power extrapolation under the same assumption as we had applied for PageRank algorithm.

4.3.4 Insights into Extrapolation

Extrapolation is one of the effective techniques in numerical analysis, but its use for acceleration of convergence in power method is novel. There are a lot of things to explore in the topic of Extrapolation. The studies so far just provide a definition level insight into the topic. There could be a lot of different and useful insights into much dynamic aspects of Extrapolation.

A new premise

The Extrapolation methods described in the previous sections are built upon the *premise* constructed about the initial function (equations (4.12), (4.13) and (4.37)). The common thread among all the extrapolations techniques is the initial assumption. In case of Quadratic Extrapolation we assumed that the matrix $\mathbf{A}$ has 3 eigenvectors, and expressed the current iterate as the linear combination of these 3 eigenvectors. A quadratic function illustrates a much closer representation to the reality in this case, and hence provides much better convergence.

One of the possible prospects in Extrapolation could be to start with another new *premise*. A much deeper understanding of Markov matrix and the insight into properties of LAR algorithms in question could be valuable to form a new premise(s). It could be possible to also consider the *personalization* (next section) factor in the construction of the new premise in Extrapolation. A possible future work from this study could be to explore extrapolation independently with the focus on formation of a new premise.

Extrapolation parameters

From experiments in the Chapter 5 we found that from the behaviour of the *convergence graph* we could manipulate the extrapolations parameters to control the convergence. In this regard the number of times we apply extrapolation step is quite crucial. If extrapolation is applied more often than required it might increase the overhead instead of improving. We will discuss in detail about the exploitations of the parameters in extrapolation during experimental analyses in the next chapter.

Hybrid Extrapolation technique

We might as well think of extrapolation in a 'hybrid' environment. That is, considering the properties of different extrapolation techniques and depending on the behaviour of the graph we could apply different extrapolation techniques at different instances during iterations of the power method. We will experiment with the hybrid of Quadratic Extrapolation and Power Extrapolation to experimentally examine this approach. But it also requires an independent study to come up with a possible *framework* which could be used to exploit different extrapolation techniques in a hybrid environment in order to achieve a much controlled convergence in power method. And to observe more closely the dependences and dynamics of different extrapolation techniques applied simultaneously. It could turn out to be a novel approach towards active use of extrapolation. And again, the factor of personalization can also be employed in the hybrid extrapolation scheme.

The findings discussed in this section just depict limited implications of extrapolation. There are quite a lot of other possibilities too for further innovation in the field of Extrapolation.

4.4 Personalization

There is no exact mathematically precise measure of "best", indeed it lies in the eyes of the *beholder*.

Personalization is one of the very significant and key topics in the relevancy ranking. A user is conventionally expected to solicit queries like: "what sources are popular or authoritative on some topic (query)?". In contrast, the Web user, with sophisticated or individualistic needs might query: "what sources are authoritative, given that some sources are *preferred*, which might not be authoritative in general". A user therefore may have opinions or inclinations on certain topics and may consider them better than the authorities in general on the topics in question. Under such a situation a user would like to see authoritative sources that coincide with their own preconceived conception of authoritativeness for certain topics. Or in general a mechanism is expected through which users can enforce their opinions on the search outcomes. Personalization is therefore the mechanism to *bridge* the user's opinions and conceptions (preferences in general) with the system's measure of authoritativeness.

The user of IR system now doesn't want to see the search outcomes the way system produces it, rather they need *personalized* results, specific to their own behaviours and moods. The user expects that the IR systems somehow *implicitly* monitor their behaviours, organize and update their behavioural information (the implicit feedbacks), and when they query, the search outcomes should be in accordance to their behaviours. For example, for the search query "movies", the page 'www.imdb.com' (Internet Movie Data Base) might be considered as top authority by most of relevancy ranking algorithms, but a user interested in theatres might want 'www.amctheatres.com' (AMC Theatres) as the most relevant hit to his/her query "movies".

Users on the Web usually have a set of preferred pages (e.g., bookmarks, usage history or click-through) and

they require that the search engines capture these information and based on that prepare the search outcomes, instead of producing generally relevant pages.

The *one-size fit all* approach by the conventional search-engines requires a placeholder or accommodation for personalization. Hence a search by user A should have different outcomes than a search by user B. The users want to *inject* their own opinions in the search process, about which type of sources are considered important by them, and what should the search outcomes look like. The IR model must therefore cater the user needs more advantageously by either implicitly decipher the trends in their searching behaviours or explicitly utilize any other relevant information, such explicit user relevance feedback.

The personalization is not only centred towards users' needs and preferences, but also through personalization the IR system can be in a better control of the search outcomes. One of the ways of making money by the search engines today is to boost certain sources based on some query topics, e.g., for advertising purpose. Advertising accounted for over 97% of *Google* revenue in 2003 [Lan06]. But also personalization is used to counter the bigger problem of *spamming* to a certain extent. Through personalization the possible spamming sources can be prevented to inflict the rankings, by assigning them negative weights with the facility of personalization schemes. Thus personalization can better serve the cause of *anti-spamming* and could be extensively and strategically employed to create firm barriers for spammers.

There has been numerous recent developments done in personalizing PageRank algorithm using several approaches, such as [Jeh03; Hav03b; Hav03a; Kam03a; Ric02]. The initial idea of personalizing PageRank was first proposed by Page et al., [Pag98], but was never fully explored. There has been some marginal amount of work done to personalize HITS too, such as, the *Gradient Ascent* approach by Chang et al. [Cha00].

The user search or usage history (e.g., usage logs) can be mapped to the set of representative content categories drawn from some sources, e.g., *Open Directory Project*[3]. The topic-sensitive approach by Haveliwala [Hav03a], take the linear combination of 16 topic-sensitive PageRank vectors (16 topics taken from Open Directory Project) to calculate the overall *Personalized PageRank vector*.

"To encompass different notions of importance for different users and queries, the basic PageRank algorithm can be modified to create 'personalized views' of the web, redefining importance according to user preference" [Jeh03].

These studies are just a small subset of the ongoing research in personalization. In this section we will go through a few of the important approaches towards personalization. We will then present a general approach towards personalization which can be tailored to bring the query-dependent models like HITS and SALSA within the framework of personalization. At the end we will also present some of the approaches of personalization in HITS algorithm, and some other approaches which offer the possibilities for personalization in HITS and its descendants.

Personalization in PageRank

In the Section 3.3.5 we derived equation (3.7) after the necessary adjustments, the factor $\frac{1}{n}\, ee^T$ in that equation can be called as the *teleportation matrix*, $\mathbf{E}$. Thus we could rewrite equation (3.7) as:

$$\mathbf{G} = \alpha\mathbf{S} + (1-\alpha)\mathbf{E} \tag{4.44}$$

The adjustments to the basic PageRank model hence created a possibility for personalization with the help of the teleportation matrix $\mathbf{E}$. Initially $\mathbf{E} = \frac{1}{n}\, ee^T$, but in the modified (personalized) version, $\mathbf{E}$ will be an

[3]Web directory for over 2.5 million URLs, http://www.dmoz.org/

$n \times n$ rank-one row stochastic matrix:

$$\mathbf{E} = e\vec{v}^T$$

where $\vec{v}$ is an $n-$vector whose elements are all non-zeros containing probability distributions, hence it is called the *personalization vector*. In terms of personalization vector $\vec{v}$, the PageRank model will be the same as in the Section 3.3.5, except the last step of derivation, which could be replace by:

$$\mathbf{G} = \alpha\mathbf{A} + (\alpha\mathbf{a} + (1-\alpha)e)\vec{v}^T \tag{4.45}$$

Equation (4.45) signifies that; at each time step, a random surfer who hops from webpage to webpage, with probability α, he randomly chooses the next page from the set of out-links of the current page. Occasionally with a probability $(1-\alpha)$ the surfer gets bored (or tend to avoid dead-ends or endless cycles), instead of following out-links, he jumps to a random webpage, i.e., he *teleports*. The destination of the random jump will be chosen according to the probability distribution given in $\vec{v}$. The preferences of the user therefore can be stored in the probability distribution $\vec{v}$. Thus $\vec{v}$ will automatically give preference to those pages which are of interest to the user.

In terms of the basic PageRank model the probability distribution for the random jump is set to be *uniform*, i.e., when the random surfer gets bored, he jumps to a random page based on uniform distribution (e.g., $\frac{1}{n}\ \vec{\mathbf{1}}$). If that probability distribution is set to be *non-uniform*, the resultant PageRank vector can be biased to prefer certain kinds of pages. That is why probability vector $\vec{v}$ is called as *personalization* vector.

Through *biasness* (in terms of probability distribution) in the probability vector $\vec{v}$, we could personalize the outcomes of the search. This is how the *user-independent* PageRank can be specialized to the *user-dependent* PageRank, with the personalization based on users' own choice. The only limiting factor here will be the computational cost of calculating the personalized PageRank for every user, which will be fairly prohibitive. We will talk about the computational concerns of personalization in the later sections.

4.4.1 Intelligent Surfer

The random surfer in personalized model can also be called as an *intelligent surfer*, one that is guided by a controlled probabilistic model of the relevance of a page to a query. This means that when a random surfer wants to select a destination at the time of jump, he will select them *intelligently*, instead of randomly.

By enriching the user query with their contextual information such as their behavioural trends (preferences) we could enforce the intelligent surfer to focus on the user specific outcomes. An intelligent surfer is expected not to loose the query context (memory of query topic) as quickly as the undirected or unintelligent random surfer. The jumps of the intelligent surfer must be guided by the contextual information in the personalization vector.

In the study by Richardson and Domingos [Ric02], a *query-dependent*, *content-sensitive* version of PageRank has been introduced. The random surfer (or *directed* surfer) in that case, when choosing among multiple *out-links* from a page, it tends to follow those which lead to pages whose content has been deemed relevant to the query[4]. Similarly, at the time of random jump the directed surfer tends to choose the page which is content-wise more relevant to the query at hand. Thus a query *enriched* with user specific details would automatically personalize the outcomes, by precisely using the most representative content depicting the user's preferences.

Although the idea initially presented in directed surfer model doesn't intended to cope with personalization. The concept can be used to personalize the outcomes based on content-sensitive approach. If content-wise the

[4]For example by using the traditional IR measures like $TF-IDF$ see Section 2.3.2 as the content dependent relevance function.

query contains the user specific details, then the directed surfer would automatically take care of personalizing the search outcomes. This could be a possible future work to make the *query subsystem* in an IR model accountable for the task of enriching the user query with user related information. The work done in directed surfer model by Richardson and Domingos would then be used for constructing the personalized results. In the modified directed surfer model (expected), the task of personalization is shifted from the relevancy subsystem to the query subsystem.

4.4.2 General Formulation of Personalization

To formulate a general representation for personalization using the random surfer model of PageRank, we need to consider equation (4.44) in the previous section. We will use this formulation as a basis to construct a general model for personalization, which could be subsequently used in the query-dependent algorithms such as HITS and SALSA.

Taking transpose of equation (4.44):

$$\mathbf{G}^T = \alpha \mathbf{A}^T + (1-\alpha)\vec{v}e^T \tag{4.46}$$

From the power iterations we have:

$$\vec{x} = \mathbf{G}^T\vec{x} \tag{4.47}$$

Substituting equation (4.46) in equation (4.47) we get:

$$\begin{aligned}
\vec{x} &= (\alpha \mathbf{A}^T + (1-\alpha)\vec{v}e^T)\vec{x} && (4.48)\\
\vec{x} &= \alpha \mathbf{A}^T\vec{x} + (1-\alpha)\vec{v}e^T\vec{x} && (4.49)\\
\vec{x} &= \alpha \mathbf{A}^T\vec{x} + (1-\alpha)\vec{v} \quad (\text{since} \quad e^T\vec{x} = 1,\ as\ \vec{x} \text{ is probability vector}) && (4.50)\\
\vec{x} - \alpha \mathbf{A}^T\vec{x} &= (1-\alpha)\vec{v} && (4.51)\\
(\mathbf{I} - \alpha \mathbf{A}^T)\vec{x} &= (1-\alpha)\vec{v} && (4.52)\\
\vec{x} &= (1-\alpha)(\mathbf{I} - \alpha \mathbf{A}^T)^{-1}\vec{v} && (4.53)
\end{aligned}$$

Since the matrix $(\mathbf{I} - \alpha \mathbf{A}^T)$ is strictly diagonally dominant therefore the inverse exists [Lay94; Gol96]. Because:

$$(\mathbf{I} - \alpha \mathbf{A}^T)^T = (\mathbf{I} - \alpha \mathbf{A}^T)$$

Let

$$Q = (1-\alpha)(\mathbf{I} - \alpha \mathbf{A}^T)^{-1}$$

Substituting $\mathbf{Q}$ above in equation (4.53), we have:

$$\vec{x} = \mathbf{Q}\vec{v} \tag{4.54}$$

The vector $\vec{x}$ is the *Personalized Rank Vector* corresponding to the personalization vector $\vec{v}$. In linear algebra, columns of matrix $\mathbf{Q}$ in equation (4.54) provide the complete *basis*[5] for the personalized rank vectors. Because any personalized rank vector can be expressed as a *convex combination* (hence, linear combination) of the columns of $\mathbf{Q}$ from equation (4.53).

[5] A *basis* is a set of vectors that, in a *linear combination*, can represent every vector in a given vector space, and such that no element of the set can be represented as a linear combination of the others. In other words, a basis is a linearly independent spanning set [Lay94].

In case of PageRank, the calculation of the matrix $\mathbf{Q}$ is computationally prohibitive. Taking the inverse of huge webgraph's adjacency matrix $\mathbf{A}$ would be a computational nightmare.

However low-rank approximation of $\mathbf{Q}$ can be still be used to achieve personalization partly. In that case we won't express all the personalized rank vectors, but rather focus on those corresponding to the convex combinations of the rank vectors in the reduced basis set.

Equation (4.54) serve to be the general model for personalization which could be moulded in other algorithms as well, not just PageRank. We will study different approaches that exploit the general model of personalization in the next section. We will also present other motivating models that attempt personalization in other prospects. That is, instead of just employing a personalization vector, the adjacency matrix can be equipped with personalization information.

4.4.3 Approaches towards Personalization

Topic-sensitive approach for Personalization

Introduced by Haveliwala [Hav03a], topic-sensitive approach computes a set of the PageRank vectors, based on a set of representative topics. This captures more accurately the notion of importance with respect to a particular topic. Thus instead of using single global PageRank vector, he takes linear combination of the *pre-calculated topic-sensitive* PageRank vectors at query time, weighted using the similarities of the query to the topics.

Topic-sensitive PageRank scheme was devised to compute an $n \times k$ approximation to $\mathbf{Q}$ in equation (4.54) using 16 topics. This scheme of personalization therefore uses a very coarse basis set to personalize the rankings. But its not truly personalizing for each and every individual rather the personalization is based on the topic of the query and query context.

The columns of the reduced-rank matrix $\mathbf{Q}$ are generated offline independent of any query (the topic-sensitive PageRank vectors). At the query time the convex combination of the columns of reduced-rank matrix $\mathbf{Q}$ is taken, using the context of the query to compute the appropriate topic weights.

In terms of the equation (4.54) the vector $\vec{v_j}$ (j^{th} column of $\mathbf{Q}$) generated by the topic-sensitive scheme is a dense vector, generated using a classifier for topic T_j; hence $(v_j)_i$ represents the (normalized) degree of membership of page i to topic j. Thus in this scheme a random surfer can teleport to a topic T_j with some probability w_j, followed by teleport to a particular page i with probability $(v_j)_i$. That is how the problem of heavily linked pages getting highly ranked for the queries for which they have no authority, can be avoided. For more information see [Hav03a; Hav03b].

Dynamic programming and Personalization

In the study by Jeh and Widom [Jeh03], *dynamic programming algorithms* for computing the *partial vectors* were employed, in order to construct personalized view from the partial vectors. Partial vectors are encoded form of personalized views. In this case the algorithms compute an $n \times k$ low-rank matrix $\mathbf{Q}$ (i.e., each columns corresponding to the partial vectors) corresponding to the highly ranked pages. A suitable number of partial vectors are computed offline and at query time those vectors are "constructed quickly" to form the Personalized PageRank score (personalized rank vector, in equation (4.54)).

This approach could be considered analogous to the topic-sensitive approach; here partial vectors are calculated using the concepts of dynamic programming while in topic sensitive approach personalized vectors are calculated based on the selected topics. Going into minor details of this algorithm is out of scope of this study. For more information about this approach, see [Jeh03].

Graph structure and Personalization

There is another study by Sepandar et al. [Kam03a], they have identified the web as having *nested block structure*. That is, the vast majority of the hyperlinks on the Web, link pages on a host to other pages on the same host. There are chunks of pages which point mostly to each other in the same host instead of pointing to other hosts. Most of the links on the web are therefore *inter-host* links rather than *intra-host* links. The overall web's network structure has clusters or chunks of pages connected tightly together within each chunk. They have formed a block structure where each block is *loosely* connected with others and *tightly* connected within itself.

The graph structure of web has been analyzed by Broder et al., [Bro00] where they also think that the connectivity in web is strongly limited by a high-level global structure (see figure 3.5). The graph structure of web has been characterized by the "bow-tie" structure. It is suggested that exploiting the "bow tie" structure of the web would be fairly useful in computing PageRank. The webgraph is also proved to follow the *power law distribution*, i.e., "the number of web pages with in-degree k is proportional to $k^{-\beta}$". For more information on web's graph structure, see [Lu04; Bro00; Ara02; Cha02].

The study of graph structure of web and its implication on IR is also very exciting and prospective. Discovering, classifying and organizing structures in the webgraph are quite evolutionary and intrinsically innovative. The study of webgraph is peculiar and essential because through such studies we try to measure the distribution of broad topics and topical clusters on the Web. The knowledge of different techniques such as aggregation/disaggregation, composition/decomposition and classification of webgraph would significantly stimulate the new advances. These techniques and other strategies have clearly predicated on our understanding of the social processes which shape the webgraph. Here we are more concerned with the indirect but essentially valuable use of the graph structure of web, i.e., the implication of graph structure of the web on *personalization*.

The *BlockRank* algorithm introduced by Sepandar et al. [Kam03a], exploits the block structure of the web by calculating the local PageRank vector of each block. These local PageRank vectors also corresponds to the columns of the low-rank matrix $\mathbf{Q}$. The local PageRank vectors are weighted according to their blocks. These weighted local PageRank vectors (columns of $\mathbf{Q}$) are aggregated and the resultant vector is used as an starting vector to the standard PageRank algorithm.

The analogy of the random surfer model to the *BlockRank* algorithm will be; once the random surfer gets bored he will teleport to a particular block or host B_j with probability w_j, instead of arbitrary set of pages. And following that he will further teleport to page i in the block B_j with probability $(v_j)(i)$. For example, a surfer interested in sports may teleport from one host page, e.g., 'www.espn.com' to another host, e.g., 'www.cnn.com' instead of teleporting to the same host. The primary purpose of the BlockRank algorithm is to exploit the block structure of the webgraph in order to achieve speedup in PageRank. Locality of references in each block reduces disk i/o and allows parallel computation of the local PageRank vectors and sometimes reuse of the old values too.

The implication of the graph structure of web for personalization is an innovative approach. The BlockRank method provides a broad perspective of the use of graph structure which can be indirectly used for personalization. But a more active and thorough use of graph structure for personalization can give more insight into the problem of personalization. For example, if the adjacency matrix of the webgraph can possibly contain user specific personalization data, that kind of data could be used to generalize and personalize the user preferences. Hence personalization based on graph structure either fetches the preferred pages or the pages related to the preferences of users.

A further in depth study of purely graph structure of web together with personalization can reveal interesting features and characteristics of graph structure in exposition with personalization and additional magnificence in personalization.

4.4.4 Personalized and Stabilized HITS

The basic HITS algorithm doesn't support user-specific personalization; therefore it is worth considering personalization in HITS and in other query-dependent alternatives such as SALSA. HITS does provide authoritative results on query topics, but it does not support personalized outcomes according to users specific preferences. It does measure a general *notion of authority* but does not cater the user's *internal* notion of authority of which sources are important.

One of the ways for personalization in HITS would be to restructure it according to the general model for personalization given in Section 4.4.2. In order to do that we have to incorporate the random surfer model into HITS, i.e., random surfer(s) follows the hyperlinks and occasionally, depending on the personalization vector, jump to another destination (based on non-uniform distribution). In other words we have to have adjustments in HITS analogous to the ones in the case of PageRank, see Section 3.3.5.

We will explore some techniques through which we could possibly incorporate user preferences in the model and hoist the authoritativeness of the sources considered favourable by user. We will focus on implicit and explicit properties of HITS and also some of its limitations in order to exploit them and formulate a personalization scheme in the algorithm. The focus is therefore to bring the query-dependent HITS and SALSA within the framework of personalization.

Gradient Ascent

There could be two possible ways to cater personalization in general; one is to enrich the *user query* with his personalization information, and second is to alter the weighting of the *link matrix* to influence relevance of the retrieved documents, as discussed in previous sections. In this subsection and in the next we will focus on the second approach, i.e., to manipulate the entries in the link matrix and use them for personalization.

Here we would like to have a mean of augmenting the information in the link matrix $\mathbf{A}$ with the users' preferences to personalize the retrieved pages. This way we could boost the weights of sources which are considered important by the user. By doing that it is expected that the system generalizes the users' preferences by not just boosting the weights of sources in question, but also boost the scores of the related sources as appropriate. It is true in HITS that if we alter the link matrix (by incorporating user preference), the mutual reinforcement relation will automatically distribute the authorities among the neighbourhoods of the boosted pages.

In the study [Cha00], Chang, Cohn and McCallum believe that instead of relying on automatically distribution of authority by HITS, we can perform *gradient ascent*[6] on the elements of the link matrix. The automatic distribution of authority by HITS suffers from the intrinsic prejudice exist in the hyperlink structures. The TKC effect of HITS will possibly forbid a *fair* distribution of authorities. Instead of relying on distribution based on tightly linked pages, the geometry of hyperlink space (eigenvector space) is considered for taking care of spread of authorities.

[6] *Gradient ascent* is an optimization algorithm. To find the *local maximum* of a function using gradient ascent, one takes steps proportional to the gradient of the function at the current point.

This way we could possibly alter the link matrix in order to more closely *align* its principal eigenvector with the preferences of the user (recall the theory of linear trend in HITS, see figure 3.7). To perform gradient ascent step to the link matrix, we simply add a fraction γ of the gradient ($\forall_{k,i} \;\; \Delta\mathbf{A}_{k,i} = \mathbf{A}_{k,j}a_i$ to boost document j) to each element of the link matrix, hence;

$$\mathbf{A}^{+}_{k,i} = \mathbf{A}_{k,i} + \gamma \cdot \frac{\Delta\mathbf{A}_{k,i}}{\sum_i \Delta\mathbf{A}_{k,i}} \tag{4.55}$$

The idea of spread of authorities in gradient ascent approach is the same as the *spread of activation* model defined earlier in 1984 by Anderson and Pirolli [AND84]. Which is also later used in individual studies [Pir96; Car97], in order to create a better visualization of information on web. For detailed information on gradient ascent approach see [Cha00; Tan02].

In the next section we will use the same approach as in this section – we will alter the link matrix, and then apply the personalization scheme.

Exponentiated Input

As identified in [Mil01; Far06; Ng01b], the HITS algorithm doesn't always behave as expected. It is possible that HITS can return non-unique weight vector and can inappropriately assign zero weights to certain sources in the network (see experimental results in Section 5.2.2).

Technically, the *dominant eigenvalue* of the authority matrix $\mathbf{A}^T\mathbf{A}$ can be repeated, which causes the authority vector to be non-unique. It is not always the case to get a unique dominant eigenvalue for the matrix $\mathbf{A}^T\mathbf{A}$. In terms of linear algebra this means that the authority matrix $\mathbf{A}^T\mathbf{A}$ is *reducible* [Lay94]. That is, there are set of states that it's possible to enter, but once entered it's impossible to exit.

When the principal eigenvalue is repeated, the ranking vector or principal eigenvector can be any non-negative vector in multi-dimensional dominant eigenspace. Repeated principal eigenvalue will have the corresponding multiple principal eigenvectors, which will span the dominant eigenspace. For example, for a two-level reversed *binary tree* whose edges point upward towards the root, the eigenvalue of the authority matrix $\mathbf{A}^T\mathbf{A}$ can be repeated, see the examples in Section 5.2.2. There are some characteristic graphs that give rise to the repeated eigenvalues problem. We have experimented with some of them in Chapter 5, also see [Far06; Mil01].

The second problem is that HITS algorithm yields *zero authority weight* for apparently important nodes in certain types of graph see Section 5.2.2. If the dominant eigenvalue of matrix $\mathbf{A}^T\mathbf{A}$ is repeated, or the principal eigenvector has zero entries for nodes with nonzero out-degree, then the matrix $\mathbf{A}^T\mathbf{A}$ must be reducible (proof available in [Far06]). It has been formally proved in [Far06; Mil01] and experimentally ascertained in Section 5.2.2, that HITS algorithm is badly behaved on certain kind of graphs, and therefore it will have *non-unique* and *zero authority* weights for important nodes.

Farahat et al., [Far06] developed *Exponentiated Input* method to address the limitations of HITS described above. HITS directly depends on adjacency matrix $\mathbf{A}$, which means in a given iteration step only paths of length $\mathbf{1}$ is considered. That is, to determine the authority score of a page p_i, at each iteration HITS looks only at the hub scores of the adjacent pages (pages that directly point to page p_i) and similarly for hub score. In the Exponentiated HITS we consider paths of length greater than $\mathbf{1}$, by exploiting the exponentiation of matrix $\mathbf{A}$ in a controlled fashion. The matrix $\mathbf{A}$ is thus replaced with the "Taylor series" matrix:

$$\mathbf{A} + \mathbf{A}^2/2! + \mathbf{A}^3/3! + \ldots + \mathbf{A}^m/m! + \ldots = \; e^{\mathbf{A}} - \mathbf{I} \tag{4.56}$$

The path of length **1** is considered more important than paths of length greater than **1** by using the scaling factor $1/m!$. Now the $\mathbb{I}$ and $\mathbb{O}$ operations of HITS can be re-written as:

$$a_k = (e^{\mathbf{A}} - \mathbf{I})\ h_{k-1} \tag{4.57}$$

$$h_k = (e^{\mathbf{A}} - \mathbf{I})\ a_{k-1} \tag{4.58}$$

Given a *weakly connected*[7] input graph, the Exponentiated HITS prevents both the problems identified above; see the proof on [Far06]. With the above adjustment now Exponentiated HITS can take into account paths of length greater than **1** in a given iteration.

Suppose,

$$\mathbf{A}_{EI} = (e^{\mathbf{A}} - \mathbf{I})$$

After having a proof of uniqueness and non-nil-weighting now with matrix $\mathbf{A}_{EI}$ we are in a position to directly address personalization. The Exponentiated matrix $\mathbf{A}_{EI}$ can be used as a step towards the personalized HITS algorithm. By incorporating user's personalization data into the base matrix $\mathbf{A}$, the subsequent Exponentiated matrix $\mathbf{A}_{EI}$ will provide a better representation of the users' personalization. With every power of $\mathbf{A}$ the personalization data will get boosted and propagated. It requires a further in depth analyses to observe personalization together with Exponentiated Input. A possible future work will be to observe the calibrated effect of personalization due to different Exponentiated inputs. In the next chapter we will experiment with the Exponentiated Input in HITS together with personalization scheme similar to the PageRank model, also described in the next subsection.

We can use also the link matrix $\mathbf{A}'$ to hold the weights of connections between the pages based on *usage data* from webserver log of traffic. The link matrix can be initialized to **0**, and then increment the link from node i to j every time a user travels from page i to j. The effect of the matrix $\mathbf{A}'$ would be similar to the gradient ascent method described in previous section. The link matrix now contains the usage information, and the entries in the matrix are larger when certain pages are visited more often. Thus in each iteration the most frequently followed links play larger role in determining new authority weights. The matrix $\mathbf{A}'$ can be called as the *usage weighted input matrix*.

The usage logs provide a wealth of information that the relevancy ranking process can harness to improve search quality. How to traverse them, how to represent them in the link matrix and their expositions in personalization are quite motivating inquisitions in the field of usage log exploitation. Interpreting and utilizing these information is certainly not easy task. There has been some work done on exploiting usage or interaction information obtained from search engine logs, in order to understand how these *implicit feedbacks* (the logs) could be used to tune the relevancy ranking [Agi06a; Agi06b].

In Chapter 5 we will experiment with the personalized Exponentiated HITS and compare the results with other alternative algorithms, e.g., comparing the results with basic HITS and its improvements.

Randomized HITS

If link analysis is to provide a robust notion of authoritativeness in such a setting where information is dynamic in nature, it is natural to ask that is it also robust in the sense of being *stable* to perturbation of the link structure [Ng01b]. We can view the perturbations to network as continuous, abrupt and drastic changes in the adjacency matrix (changes in hyperlink structure). *Perturbation analyses* in linear algebra has very widespread

[7] A directed graph **G** is weakly connected if any node can be reached from any other node by traversing edges either in their indicated direction or in the opposite direction.

applications and implications. Through such analyses we would like to measure the consistency of the algorithms by exposing it to diverse and dynamic sets of input data.

The robustness of HITS algorithm like other link analysis algorithms requires an analysis of the stability of the eigenvector calculations. Issues like relationship between multiple eigenvectors and invariant subspaces, and the effect of random jumps must be considered [Ng01b; Ng01a]. In perturbation analysis, small changes to a matrix tend to result in large changes to eigenvectors only when eigenvalues are sufficiently close to each other [Mot95]. That is, when the *eigengap*[8] of the matrix $\mathbf{A^T A}$ is large HITS is expected to be insensitive to small changes in the link structure of the matrix $\mathbf{A}$. But if the eigengap is small then slight changes in the link structure of matrix $\mathbf{A}$ will lead to large changes in ranking produced by HITS algorithm. Stability is therefore a desirable property of LAR algorithms, given the fact that the link structure of web changes quite frequently.

In the study by Ng et al., [Ng01a], they have experimentally found that HITS under certain conditions is quite instable, and also provided conditions under which HITS can be stable. They also found that PageRank algorithm is relatively *immune* to stability concerns. Ng et al., [Ng01b] later developed a modification to HITS algorithm, inspired by PageRank's immunity to stability conditions. They considered a random surfer model similar to PageRank case, but in HITS the random surfer follows the hyperlinks both in the *forward* and *backward* directions, and occasionally "resets" and jumps to a page chosen uniformly at random. The method is named as "Randomized HITS", because of the random surfer model. Mathematically, after the modification the $\mathbb{I}$ and $\mathbb{O}$ operations in HITS can be rewritten as:

$$a^{(k+1)} = \alpha\vec{\mathbf{1}} + (1-\alpha)\mathbf{A}_{row}^T\ h^{(k)} \tag{4.59}$$

$$h^{(k+1)} = \alpha\vec{\mathbf{1}} + (1-\alpha)\mathbf{A}_{col}\ a^{(k+1)} \tag{4.60}$$

With the help of the above modifications, analogous to the PageRank, now the Randomized HITS is more stable to perturbations than the original basic HITS.

The Markov chain constructed as a result of the above two operations can be interpreted as: with probability $(1-\alpha)$ the random surfer follows the hyperlink from the current page in forward direction when $\mathbb{I}$ operation is executed, and follows the hyperlink from current page in backward direction when $\mathbb{O}$ operation is executed. The random surfer *teleports* or jumps to a uniformly chosen destination with probability α.

The adjustment proposed in equations (4.59) and (4.60) bring in the damping factor in essentially the same way as that of PageRank. A consequence is that, we now have a room for personalization in HITS analogous to the PageRank model. In equations (4.59) and (4.60) if we manage to control the teleportation of the random surfer we could incorporate personalization. Notice that if we could replace the vector $\vec{\mathbf{1}}$ with non-uniform probability vector $\vec{v}$, we will have the personalization in Randomized HITS, thus;

$$a^{(k+1)} = \alpha\vec{v} + (1-\alpha)\mathbf{A}_{row}^T h^{(k)} \tag{4.61}$$

$$h^{(k+1)} = \alpha\vec{v} + (1-\alpha)\mathbf{A}_{col} a^{(k+1)} \tag{4.62}$$

With the above modification to the Randomized HITS, now the random surfer who teleports with probability α, will be enforced to jump to *non-uniformly* chosen destination based on the personalization vector $\vec{v}$.

Motivated further by the above discussion a modification similar to the PageRank primitivity trick can also be applied to HITS as well. Making the HITS quite conducive for the generic formulation of personalization given in Section 4.4.2. Thus mathematically the modified matrix can be written as:

$$\alpha\mathbf{A}^T\mathbf{A} + (1-\alpha)\mathbf{E} \tag{4.63}$$

[8]The difference between largest and second largest eigenvalues of matrix $\mathbf{A}$, i.e., $|\lambda_1 - \lambda_2|$, and for the matrix $\mathbf{A^T A}$ the eigengap will be $|1 - \lambda_2|$.

With the above modification, the authority matrix becomes irreducible and therefore by *Perron-Frobenius theorem*, it possesses a unique and positive dominant eigenvector [Lan06; Kit98]. With the above modifications in equations (4.61), (4.62) and (4.63) we are now convinced that HITS can be both stabilized and personalized. We will experimentally ascertain the viability of these modifications in the next chapter.

Personalization is relatively new phenomenon in IR community, but the implications are quite futuristic. Therefore, there is a lot of scope available for further research in the field of personalization. The area is very fascinating and rewarding, and you have very widespread opportunities for manoeuvring. Personalization is employed also to improve performances apart from just personalizing the results, e.g., the BlockRank algorithm described in previous sections. Personalization can also be employed to control the search outcomes. It spans from the query subsystem to the relevancy subsystem in IR. In the query disambiguation process, personalization can be employed to obtain user specific information. It is therefore not far from truth that the prevalent and exclusive use of personalization in retrieval is the *current* and will be the *future* outlook of IR.

In the next chapter we will investigate further into empirical implications and results of the methods discussed in this chapter.

Chapter 5

Experimental Evaluations

Experimental evaluation is to assess the quality of the algorithms and, more importantly to understand and observe how theoretically predicted properties manifest themselves in the practical setups. We will explore the practical exhibitions of ideas discussed in LAR algorithms earlier in this study.

In this chapter the focus is on the empirical evaluations of the algorithms discussed in Chapters 3 and 4. And also provide the practical manifestations of the key ideas discussed. We will specifically observe the effectiveness of the query-dependent LAR algorithms, such as; HITS, SALSA and their improvements. We will study their convergence behaviours and the significance of the modifications and improvements applied on them. Specifically we will observe the peculiarities of *Extrapolation techniques* in improving the rate of convergence in query-dependent algorithms.

The algorithms will be exposed against different sets of data. We have described the datasets that we will use for our experiments in next section, the experimental setup. Because of unavailability of sample and assessed datasets, we primarily rely on the dataset used in [Bor05]. We will thus compare the effectiveness of our findings with the findings described in the work by Borodin et al. We will also compare our findings with findings in both [Naj07b; Naj07a].

We will start with the experimental setup, and present the set of queries that will be used to test the algorithms. Those queries have already appeared in the previous researches [Cha98; Kle99; Bor05].

The overall focus of the chapter is to see the relationship between the theoretical properties of the query-dependent algorithms (defined earlier in Chapters 3 and 4) and their empirical outcomes.

5.1 Experimental Setup

5.1.1 The Graph and the Dataset

The algorithms that we will test, operate on a collection of pages that is created following the guidelines of Kleinberg (see Section 3.4). As is described in the work done in [Bor05] the search engine Google is queried for each of the queries shown in Table 5.1 (when a query consists of more than one word, we put the '+' symbol in front, so as to ensure that all pages contain the query terms). The first 200 pages returned by Google form the *Root-set* as prescribed by Kleinberg. For each page in the Root-set, all the *out-links* are stored of that page, and the first 50 *in-links*, in the order they are returned by Google, as described in Section 3.4.3. One way of obtaining in-links is to use Google queries of form $link : url$ (e.g., $link : www.fastsearch.com$), which returns

a list of documents that point to the *url*. The Root-set is then expanded into the Base-set by including the in-links, and out-links of the pages in the Root-set (see the graphical representation of this process in figure 3.4). Given the Base-set, the underlying graph is constructed, induced by this set of pages: each page is represented as a node, and a (directed) edge is placed for each link between the pages (nodes). Edges that connect two nodes within the same domain are removed since they usually serve navigation purposes, and isolated or dangling nodes are also deleted. The final graph will be given as input to the LAR algorithms.

The running time of creating the base-set graph is completely dominated by the time it takes to fetch the documents. It takes quite some time to download the documents, given that every neighbourhood graph contains nodes on the order of at least 2000+ documents.

The dataset that we will use in our experimentations has been taken from [Bor01]. On the prescription of Kleinberg as described in last two paragraphs, the dataset are stored in an *inverted file* format. Every page is first assigned a '*docid*' (document id), and from the pages' *docids* an inverted file will be generated corresponding to each query. An inverted file (for query "amusement parks") might look like:

$$
\begin{array}{lcl}
\mathbf{1} & : & \mathbf{182}, \mathbf{183}, \mathbf{12}, \mathbf{-1} \\
\vdots & & \\
5 & : & 325, 326, 327, 328, 329, -1 \\
\vdots & & \\
51 & : & 1296, 1297, 694, 707, 715, 789, 502, -1 \\
\vdots & & \\
129 & : & 2992, 3304, 2994, 3306, 3307, 3308, -1 \\
130 & : & 1122, 3314, 3315, -1 \\
\vdots & & \\
3403 & : & 3405, 3406, -1
\end{array}
$$

This means that *docid* **1** the first row (boldface), contains link to *docids* 182, 183 and 12, and -1 indicate end of list (out-links). Using the inverted file as an input to a *script* (such as, a bash-script, or a python code or a matlab code), we could convert the inverted file to an adjacency matrix $\mathbf{A}$ form. Where the entries of $\mathbf{A}$ corresponding to *docid*-**1** will be; $\mathbf{A}(1, 182) = 1$, $\mathbf{A}(1, 183) = 1$ & $\mathbf{A}(1, 12) = 1$ and rest of the entries in 1^{st} row of matrix $\mathbf{A}$ will be; $\mathbf{A}(1, j) = 0$, where $j \notin (182, 183, 12)$. The resultant adjacency matrix $\mathbf{A}$ corresponding to each query given in Table 5.1, can be given as an input to the LAR algorithms.

5.1.2 The Queries

There are some standard set of queries appeared in the literature and used in previous works [Bor05; Kle99; Lem00]. The choices of queries are driven by the fact that they become a representative of the whole Web. Therefore it is expected that through their representativeness they unveil the implicit properties of the algorithms, e.g., HITS support of tightly knit communities (see section 3.4). Every query represents a topic on web. Webgraph can be considered as a set of clusters of *strongly connected nodes*, each cluster theoretically represents some topic(s). In principle we want to align query topic(s) with the topic(s) on the webgraph. By testing the algorithms with different representative queries (topics), we tend to observe the behaviour of the algorithms on these topic(s), using single or multi-topic queries at a time.

There are queries where the most relevant results are not textually expressed in the most relevant documents, e.g., the phrase "search engine" doesn't appear in the most of the search engines main pages. Thus we would have those types of queries also which are usually not expressed within the relevant documents, such as, "search engines", "automobile industries", etc.

Query	Nodes	Hubs	Authorities	Links	Avg out
abortion	3340	2299	1666	22287	9.69
affirmative action	2523	1954	4657	866	2.38
alcohol	4594	3918	1183	16671	4.25
amusement parks	3410	1893	1925	10580	5.58
architecture	7399	5302	3035	36121	6.81
armstrong	3225	2684	889	8159	9.17
automobile industries	1196	785	561	3057	3.89
basketball	6049	5033	1989	24409	4.84
blues	5354	4241	1891	24389	5.75
cheese	3266	2700	1164	11660	4.31
classical guitar	3150	2318	1350	12044	5.19
complexity	3564	2306	1951	13481	5.84
computational complexity	1075	674	591	2181	3.23
computational geometry	2292	1500	1294	8189	5.45
death penalty	4298	2659	2401	21956	8.25
genetic	5298	4293	1732	19261	4.48
geometry	4326	3164	1815	13363	4.22
globalization	4334	2809	2135	17424	8.16
gun control	2955	2011	1455	11738	5.83
iraq war	3782	2604	1860	15373	5.90
jaguar	2820	2268	936	8392	3.70
jordan	4009	3355	1061	10937	3.25
moon landing	2188	1316	1179	5597	4.25
movies	7967	6624	2573	28814	4.34
national parks	4757	3968	1260	14156	3.56
net censorship	2598	1618	1474	7888	4.87
randomized algorithms	742	502	341	1205	2.40
recipes	5243	4375	1508	18152	4.14
roswell	2790	1973	1303	8487	4.30
search engines	11659	7577	6209	292236	38.56
shakespeare	4383	3660	1247	13575	3.70
table tennis	1948	1489	803	5465	3.67
vintage cars	3460	2044	1920	12796	6.26
weather	8011	6464	2852	34672	5.36

Table 5.1: *Query Statistics*

There are also queries for which we have conflicting communities on the webgraph. For example the query "iraq war" and "abortion" can have sets of conflicting clusters of webgraph. It is interesting to observe how different algorithms treat these queries.

The queries in the Table 5.1 with the statistical information will be used for the experiments; the queries are exactly the same as is used in [Bor05].

5.1.3 Query Statistics

The query statistics provided in the [Tsa04a] are used in our study (see Table 5.1). The table provided here shows the analytical study of the datasets corresponding to each query. The assessment information of the dataset gives a broad picture of the neighbourhood graph (the base-set). The information therefore is useful for conceptualizing and understanding the underlying structure of graph and broad picture of manifestation of

algorithms. Sometimes from the assessed dataset we could predict the expected behaviours, performances or outcomes of the algorithms. For example, for the query "search engines" there are $11,659$ nodes and $292,236$ links, which means the underlying graph for this query is quite big $(11,659 \times 11,659)$ and dense. There will be memory contention issues for such a big graph. It would be interesting to observe how different algorithms will react to such a big graph.

In case of PageRank, Haveliwala [Hav99], presents a memory efficient approach that lowers the main memory requirements for the huge webgraph (using *Block-Based strategy*). In HITS usually the memory concern is not that terrible, because the graphs are usually of order $1,000 - 5,000$ nodes. Nevertheless, efficient usage of memory is a favourable property for HITS too, given the fact that it is computed at query time, and the retrieved pages could be sizable.

Profoundly analyzed datasets provide valuable input for assessment or comparison of the LAR algorithms. Therefore it's very crucial to have a sample and representative datasets with statistical information available, which could be used for experimentation. We could have done our own evaluation based on the pre-labelled corpus, such as *TREC collection*[1], but due to limited time constraint, we primarily rely on the dataset provided on [Tsa].

We refer and confide on the query statistics given in the Table 5.1 during experimentations.

5.1.4 Measures

To assess the quality and accuracy of the results of an algorithm we will compare the 'precision' over $top - 15$ results with the results in [Bor05]. Usually for the relevancy ranking algorithms the measures used to assess the algorithms are 'precision' and 'recall'. But we expect the result of the text based search to have high recall, and therefore we only consider the precision over $top - 15$ results, and also considering the behaviour web user, we can only consider the accuracy of the $top - 15$ results or the first page of results.

Given a query the dataset that we are using has been classified as *non-relevant*, *relevant* or *highly relevant* to the topic of the query (see [Tsa]). The classification is based on the *relevance ratio* of the query topic with page in question. Almost similar notion of relevance is employed in the *TREC conference* for topic distillation queries. In TREC the data relevance and high relevance is usually predefined by a set of experts.

5.1.5 Convergence

Convergence is the measure of how the ordering of the pages changes as the number of iterations increases. In general an algorithm should declare convergence once the value of $Residual_i = Rank_{k+1} - Rank_k$ *stabilizes*. This means that the new iterations cannot change the ranking order, the ranking order has now stabilized.

In most of the literature L_1 norm or residual of the authority weights of two successive iterates is used to detect convergence in query-dependent LAR algorithms. For measuring the convergence in all of our algorithms we will therefore rely on the following measure:

$$\delta_k = ||\mathbf{A}\vec{x}^{(k)} - \vec{x}^{(k)}||_p \quad p = 1 \tag{5.1}$$

In linear algebra, equation (5.1) is generally used as an indicator of convergence for most of the iterative algorithms. In almost all of the experiments, L_1 norm in equation (5.1) is compared against $\epsilon \in (10^{-16} - 10^{-5})$. We will frequently refer to the variable ϵ, during the proceedings of this chapter, therefore it is wise to keep it in mind.

[1] **T**ext **RE**trieval **C**onference(TREC) is the primary benchmark for information retrieval.

There are other possible ways also to measure convergence, for instance in [Kam03b] *Kendall's-tau rank correlation (KDist)* measure is used, to see if the residual, L_1 norm is a good measure of convergence. Haveliwala [Hav99] suggests to use *induced orderings*, rather than residuals, by looking at the ordering of the pages *induced* by the rank vector to measure convergence. With induced ordering, PageRank vector converges in as few as 10 iterations in comparison with convergence with L_1 residual which takes about 50+ iterations.

We choose L_1 norm as a measure of convergence for the sake of simplicity. But it's a good idea to try different measures of convergence as described in the previous paragraph, to see their effectiveness on the number of iterations and the accuracy of rankings.

5.1.6 User Study

In the user study, a group of users have been shown the results of different algorithms (see [Tsa]). The user study has been performed on the dataset that we are using. The results were permuted, so as that they appeared in random order, no information about the algorithm(s) was revealed to the users. The users ranked every document as, "Highly relevant", "Relevant", "Non-Relevant" or "Don't know". User feedback on the results produced by the algorithm is quite important and therefore provides information about the practical usefulness of different approaches. User feedback can be used to assess the quality of algorithms. The number of users feedbacks per query is shown in Table 5.2.

In [Bor05] the users' information has been collected online, for each of the algorithms they tested. Average is taken over all users and all queries on the feedbacks received. Considering the lack of information of users on some of the topics, their feedback introduced some noise too. But due to limited number of users per query the quantity of errors were measurable. On the user rating for a specific document, the document is rated as "Relevant" if the "Relevant" and "Highly Relevant" votes are more than "Non-Relevant" votes, but in case of ties "Non-Relevant" votes are favoured (pessimistic approach). For more information about the user study and the dataset see [Bor05; Tsa].

In the next sections the identified algorithms will be tested and evaluated in the light of experimental setups described in this section.

5.2 Algorithms and Results of Experiments

We have implemented most of the algorithms discussed in Chapter 4, to observe them in different experimental settings. In this section we will test them according to the experimental setups described in previous sections. The algorithms will be tested against the datasets corresponding to each of the 34 queries defined in Table 5.1. That is, the adjacency matrix $\mathbf{A}$ corresponding to each query will be given as an input to the algorithms.

Broadly the following algorithms will be implemented:

- HITS (Basic, Extrapolated and Personalized)
- AT (Authority threshold - threshold set as average out-degree)
- Norm (Norm Family - with Euclidean Norm)
- HubAvg (Hub Average)
- Max (Max Operator)
- SALSA

query	**users**
abortion	22
affirmative action	7
alcohol	8
amusement parks	8
architecture	7
armstrong	8
automobile industries	7
basketball	12
blues	8
cheese	5
classical guitar	8
complexity	4
computational complexity	4
computational geometry	3
death penalty	9
genetic	7
geometry	7
globalization	5
gun control	7
iraq war	8
jaguar	5
jordan	4
moon landing	8
movies	10
national parks	6
net censorship	4
randomized algorithms	5
recipes	10
roswell	4
search engines	5
shakespeare	6
table tennis	6
vintage cars	5
weather	9
average	**7**

Table 5.2: *Users per query*

- Extrapolation – we will expose the above mentioned algorithms to the following two extrapolation techniques:
  - A^d Power Extrapolation
  - Quadratic Extrapolation
- Hybrid Extrapolation – we will briefly observe the combined effects of the Power and Quadratic extrapolation in a hybrid environment.
- Personalization – we will boost certain pages from the datasets and observe the outcomes of some of the algorithms above within the framework of personalization. And also the following algorithms will be tested with and without personalization:
  - Exponentiated HITS
  - Randomized HITS

For each of the above algorithms we will extensively experiment and compare *Extrapolated, Exponentiated* and *Randomized* versions with the original algorithms. In case of Extrapolation we would like to see the effects of acceleration step applied in various settings on the convergence behaviour of the algorithms. We will also look into the hybrid extrapolation case also, where we would employ Quadratic and Power extrapolation interchangeable in different scenarios.

We will experimentally assess the concepts in *Personalization*. We will also experiment personalization together with Exponentiated and Randomized versions of HITS, in order to see the cumulative effect of exponentiation or randomization accompanied by personalization.

The $top - 15$ results corresponding to each of queries and algorithms are available on Appendix B. Some of the significant ones are also discussed during the proceedings of the chapter.

We will take the same approach as in Chapter 4; starting with Extrapolation, by carefully manipulating the parameters on each of the algorithms to detect any observable manifestations, and similarly pursue the same motivation with Personalization. We will try to sufficiently explain the context in which the results presented in this study can be regenerated for further studies or evidences.

5.2.1 Extrapolation

In this section we will assess the algorithms experimentally by comparing them with their *Extrapolated* version. As discussed in Section 4.3 the purpose of employing extrapolation is to understand convergence property more closely. Extrapolation therefore, is to exploit the convergence property and principally accelerate convergence by subtracting off estimates of the nonprincipal eigenvectors.

In this section we will exclusively and profoundly look into convergence property of the algorithms. The application of a single extrapolation step is considered to be equivalent to 0.5 times or less the cost of an iteration of power method (32% of cost of an iteration of power method [Havb]). Therefore the effects of extrapolation step is not that severe. We will interpret the improvements in number of iterations almost equivalently to the improvements in time, e.g., a 2 times improvement in number of iterations will be treated as $1.8 - 2.0$ times speedup in the wall-clock time.

Specifically in this section we will make use of *Quadratic Extrapolation* (see Section 4.3.2). We will primarily show that the speedup due to Quadratic Extrapolation is on the scale of $30 - 900\%$, when we apply it carefully.

In a few of the experiments we even got a higher speedup than this scale (see Table 5.3), we will present them too during the proceedings.

The Extrapolated versions of each of the aforementioned algorithms are implemented according to the formulations defined in Section 4.3. We will take one algorithm at a time and analyze and compare its convergences in terms of how many times extrapolation step is applied, and how the algorithms respond to it. We will present the convergence trend in a graphical form. In the convergence graph the $y - axis$ represents the log of the L_1 norm (see equation (5.1)) of the two successive iterates, and $x - axis$ represents the number of iterations.

HITS

Recall HITS algorithm (see section 3.4), in this section we will compare experimentally HITS and *Extrapolated HITS*. The adjacency matrix $\mathbf{A}$ (the base-set) corresponding to each query is used as an input to each of the algorithms. The adjacency matrix is constructed according to the prescription described in Section 5.1.1.

HITS will compute authority and hub vectors by iteratively reaching to a stable state by determining the dominant eigenvectors of the matrices $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$. We will compare the convergence trend of the basic HITS with the Extrapolated HITS in numerous expositions.

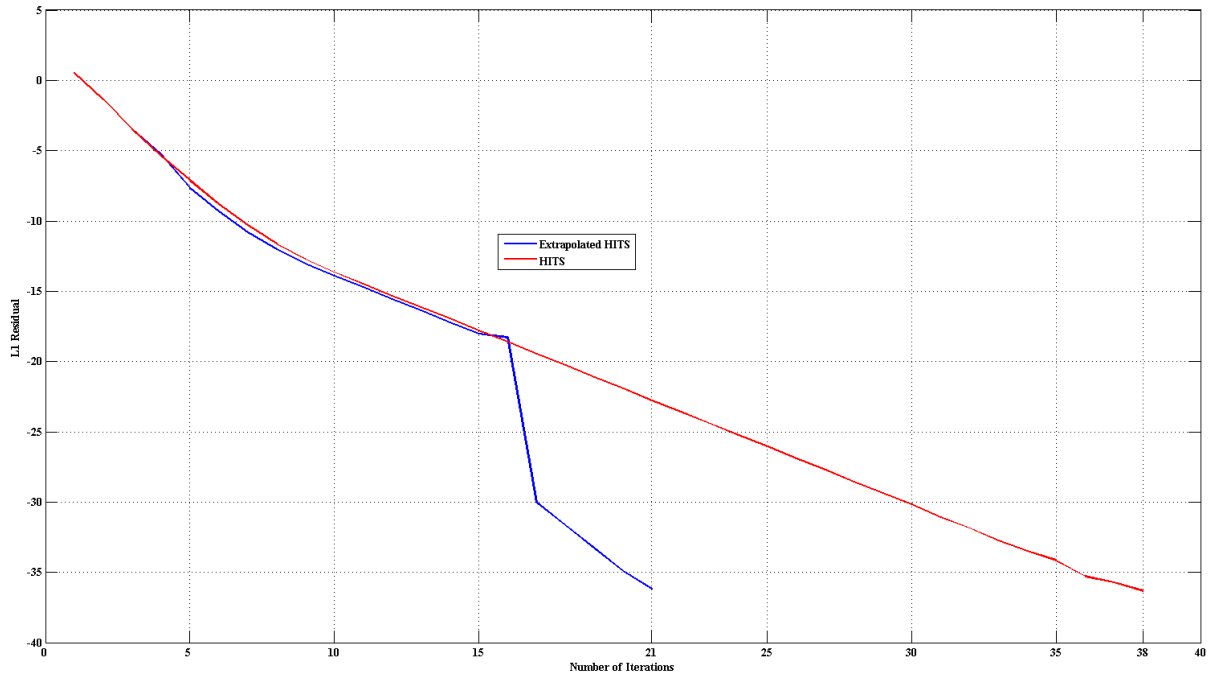


Figure 5.1: *Convergence graph for query "alcohol"*

Starting off from the query "alcohol", the adjacency matrix $\mathbf{A}$ corresponding to the base-set (for query "alcohol") is given as an input to HITS and Extrapolated HITS. The resultant curves portraying the rate of convergence are shown in the figure 5.1.

Clearly the convergence is significantly better in Extrapolated version of HITS. It reaches to the residual ϵ much quicker and with less number of steps than original HITS. We have tailored the parameters in Extrapolation to see any further gains or irregular behaviours. Applying extrapolation *twice* provides a better convergence

than applying it 5 times in this example. When extrapolation is applied twice we see convergence in 21^{st} iterations while applying extrapolation 5 times the convergence occur after 22^{nd} iterations.

The quadratic extrapolation technique should be applied *periodically* to subtract off the errors in the current iterate (along the direction of the second and third eigenvectors). It improves convergence only when it is applied *carefully*. It is interesting to observe empirically that extrapolation applied too frequently doesn't really achieve any further benefits. Because by doing that we are not allowing the power method to use the new computed iterate to annihilate error components of the iterate in directions along the eigenvectors with *small* eigenvalues. Quadratic Extrapolation step leaves error components primarily along the smaller eigenvectors, which the power method is better equipped to eliminate. Thus we need to allow power method to eliminate errors instead of applying extrapolation step frequently and accumulating the errors after every application. That is why it is very much important to apply the right number of extrapolation steps to gain the required improvements.

In the convergence graph (figure 5.1) only extrapolation is applied twice, i.e., on 3^{rd} and 7^{th} iterations. Increase in the number of extrapolation step therefore doesn't really offer any significant change in this example. So, it's a matter of applying the sufficient number of extrapolations in different time periods in order to enhance the convergence of the future application of the power method.

Note that the ranking produced by both the Extrapolated HITS and original HITS are exactly the same. Extrapolation only improves convergence and the ranking order remains the same as in the original algorithm. It does not affect the ranking order at all.

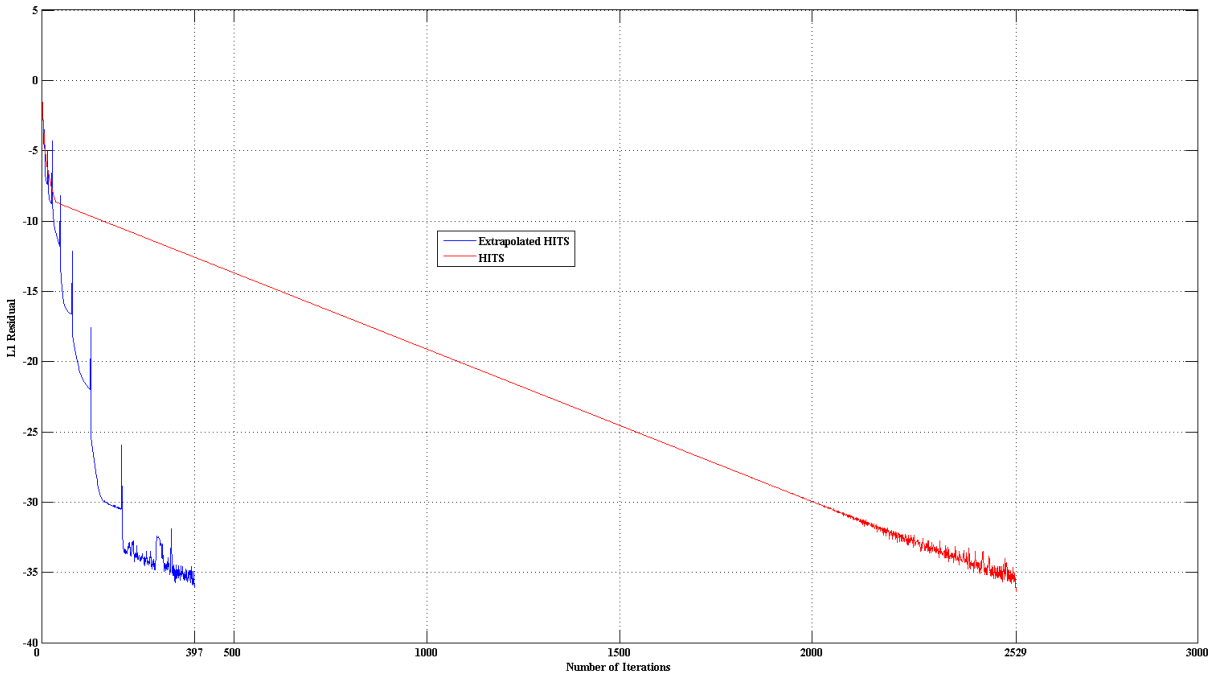


Figure 5.2: *Convergence for query "affirmative action"*

In another example we consider a multi-topic query "affirmative action". The normal HITS will take $2,529$ iterations to reach ϵ residual. While the extrapolated HITS converge significantly faster in just 397 iterations – extrapolation is applied 9 times only. From the figure 5.2 we can visibly notice a considerable speedup in

Extrapolated HITS.

The *spikes* seen in the graph (see the blue curve in figure 5.2) are due to the acceleration step, but speedup occurs nevertheless. The spikes apparently seem as poor approximation due to the extrapolation, but immediately after the spike we see a very drastic drift in the convergence. We have seen this behaviour in most of the cases, i.e., there is a spike at acceleration step and immediately after the spike we see a very rapid drop in convergence (see all the convergence curves in Appendix A).

From the experiments with HITS it's now quite evident that for extrapolation to be effective we have to apply it on the *right time* and the *sufficient number* of times during iterations. Sometimes applying late, sometimes applying early in iterations, sometimes applying more frequently and sometimes applying less frequently will give us incredible improvements in convergences, in an order of $100 - 900\%$ (see also Appendix A and Table 5.3. It's strongly encouraged that the reader goes through the results in appendices while reading the description here).

It requires further in depth analyses to *optimize* and thus *automate* the application of the extrapolation. That is, to read the properties of the convergence graph and also the behaviours of algorithms and figure out *when* and *how many times* extrapolation should be applied to get required improvements in convergence and the wall-clock time. If the convergence curve is already steep enough, applying extrapolation at that part of graph might not be helpful. But if the curve is mildly converging it might be very helpful to apply extrapolation step on that part of the convergence curve.

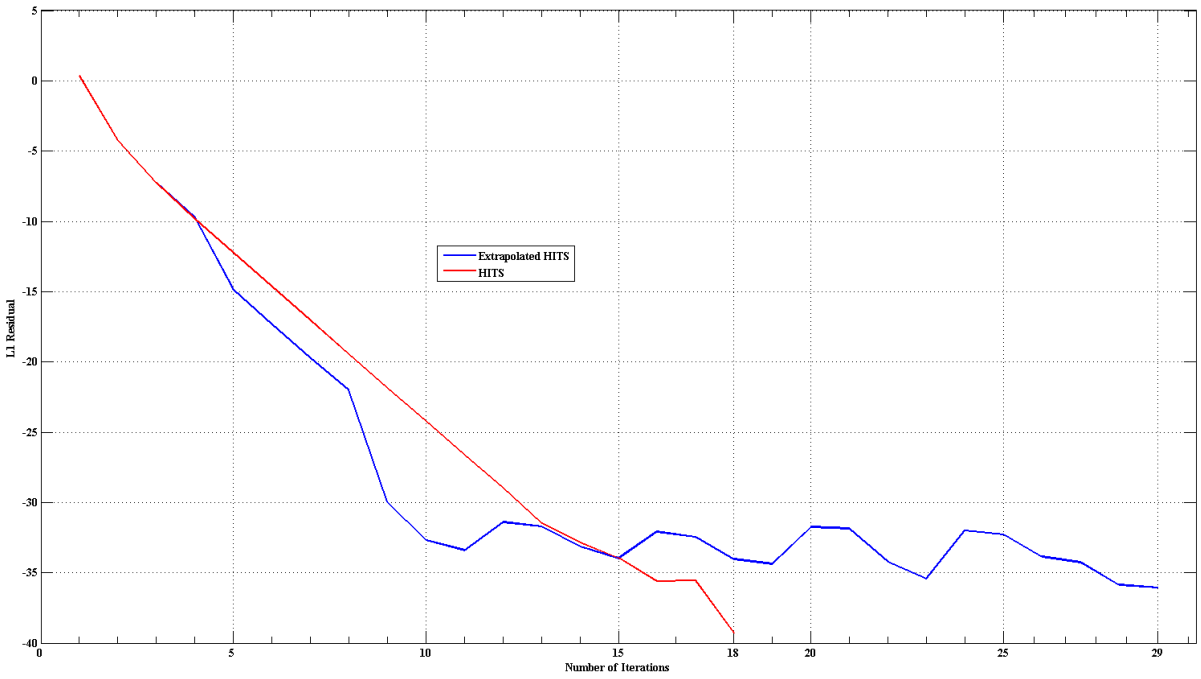


Figure 5.3: *Convergence for query "death penalty"*

Lack of good understanding of the extrapolation technique and convergence behaviour of algorithm might lead to a poor performance also due to extrapolation (see figure 5.3).

Extrapolation with HITS provides a substantial performance gain on the slow convergent sequences and

necessary insight about the capabilities of the extrapolation. We also observed that as the number of iterations increases we see more *noticeable* and *drastic* improvement in the Extrapolated HITS (see Appendix A).

HubAvg (Hub Average)

Recall the improvement of HITS, HubAvg (see Section 4.2.2), in this section we will compare HubAvg and *Extrapolated* HubAvg. Here also we intend to see any further outstanding variations in the convergence behaviour of algorithm due to extrapolation.

HubAvg generally performs quite good. The $top-15$ results capture most of the authoritative sources (see Appendix B). As far the implementation is concerned it uses most of its time in calculating the average matrix, $\mathbf{F}$, see equation (4.1). This means that if the adjacency matrix $\mathbf{A}$ is large, the cost of computing the average matrix $\mathbf{F}$ might cause some delays. For example, for queries "search engines", "weather", "architecture", "basketball" and "genetic", it took quite some time to just compute $\mathbf{F}$. Because the size of adjacency matrices are in range $(6,000 \times 6,000) - (12,000 \times 12,000)$, hence it requires some computational time to operate on them and a huge memory to contain them. Hence there is a need to optimize the computation of average matrix $\mathbf{F}$ in HubAvg algorithm.

Extrapolation is reasonably effective in HubAvg too, in most of the cases the convergence speedup are even better than its predecessor HITS. The speedup due to extrapolation is observed to be of order $150-930\%$, see for example convergence curves in figure 5.4.

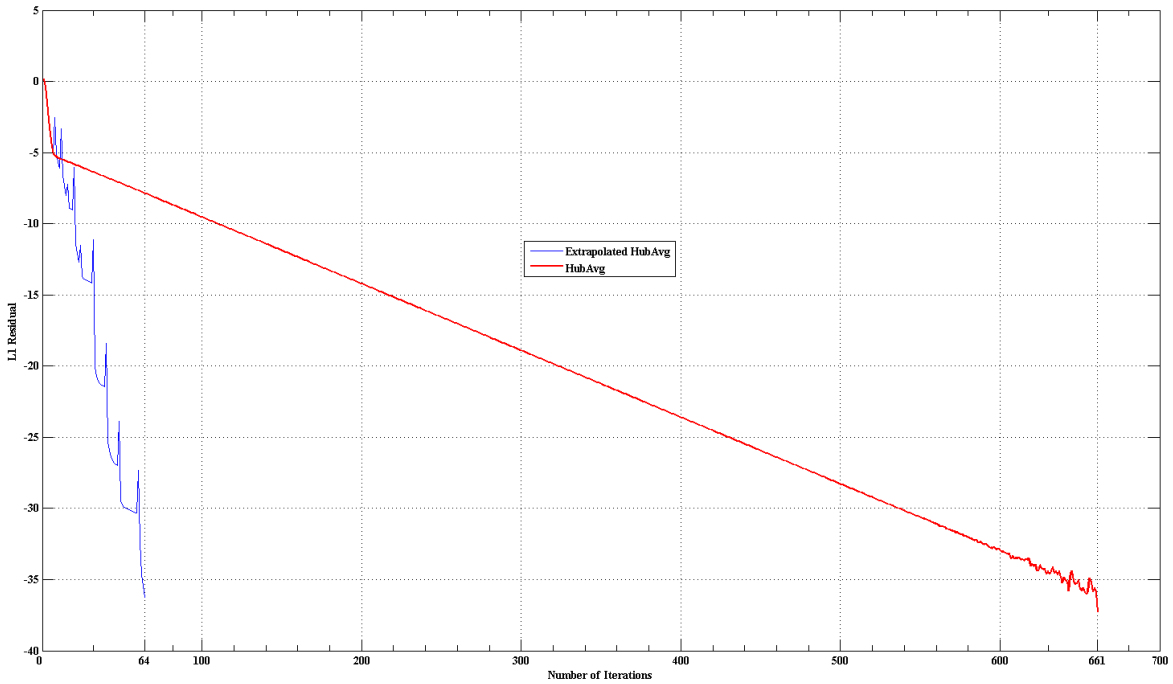


Figure 5.4: *Convergence graph for query "computational complexity"*

In almost all the experiments with HubAvg, we see a sharp spike when the extrapolation step is applied (see Appendix A). But immediately after the spike we also observe sharp steepness in the curve, as in Extrapolated HITS.

Considering the case of the multi-topic query "computational complexity", there is an astonishing speedup in convergence (see figure 5.4). In this case extrapolation is applied only 9 times, and each time it's applied we see a drastic decline in the convergence curve. See the rate of change of residual in extrapolated HubAvg and normal HubAvg, visibly the rate at which the extrapolated HubAvg drive towards the solution is almost vertical in comparison with the normal algorithm.

We regulate the parameters by applying extrapolation either too often or seldom, early in iterations or late in iterations, in order to gain a better appreciation. We also intend to see the order of improvements in convergence due to extrapolation. We have established that applying extrapolation late when an uneven trend is found in convergence, and applying it frequently if each application of the extrapolation gives additional improvements. Thus applying extrapolation less frequent only if every application does not bring any radical improvements (see this behaviour in convergence curves in Appendix A).

The performance boost both in wall-clock time and convergence due to extrapolation is promising. The cost of extrapolation step is in order of 0.5 times a single iteration of the Power Method or even less, given the fact that we apply extrapolation only periodically (in the case of HubAvg, on an average we applied extrapolation 7 times, see Table 5.3), we can highlight that the extrapolation has quite minimal overhead but has quite impressive outcomes.

Authority Threshold - AT (k)

To reduce the effects of the weak authorities on the computation of the hub weight, while at the same time retain the positive effects of the strong authorities, *Authority Threshold* is applied to retain only the top *few* authorities. Therefore Authority threshold, *AT (k)*, is to say that a node is good hub if it points to at least k good authorities (see Section 4.2.2).

In this subsection we will consider *AT - Avg* (taking the average out-degree from Table 5.1 as k), and compare it with its Extrapolated counterpart. Each iteration is computationally intense because computing the hub scores ($\mathbb{O}$ operation) entails accumulating k good authorities. A layman implementation of threshold-operator therefore doesn't offer any computational lift. After SALSA, one of the algorithms which took a lot of time in experiments is AT-Avg. Unlike SALSA (which delays largely due to slow convergence), the delay here is mainly due to inefficiency of the $\mathbb{O}$ operation.

It has been noticed (in AT-Avg) that the *size of matrix* $\mathbf{A}$ does not really influence the number of iterations required to converge. For example for the query "search engines", with a big matrix of size $(11,659 \times 11,659)$ (see Table 5.1), the original AT-Avg took 40 iterations while the extrapolated version took 17 iterations (with 4 application of extrapolation step) to converge, see figure 5.5. Of course a big matrix can slow down the $\mathbb{O}$ operation in AT-Avg, but it has negligible effects on the number of iterations.

Again applying extrapolation frequently or seldom and applying it earlier or late depends on the convergence trend of the algorithm. But the extrapolated AT-Avg's performance is quite promising too, i.e., a performance gain of order $100 - 900\%$, see Table 5.3.

As the cost of a single iteration of AT-Avg is significant, therefore it's very much important to have less number of iterations and hence a sharp reduction in the residuals is highly essential. But applying any other efficient method for the threshold-operator together with Extrapolation would have more promising results in AT-Avg in terms of wall-clock time.

See figure 5.6 for the query "abortion", the extrapolated AT-Avg converges after just 25^{th} iterations when extrapolation is applied only 6 times, while the normal AT-Avg algorithm converges after 171^{st} iterations.

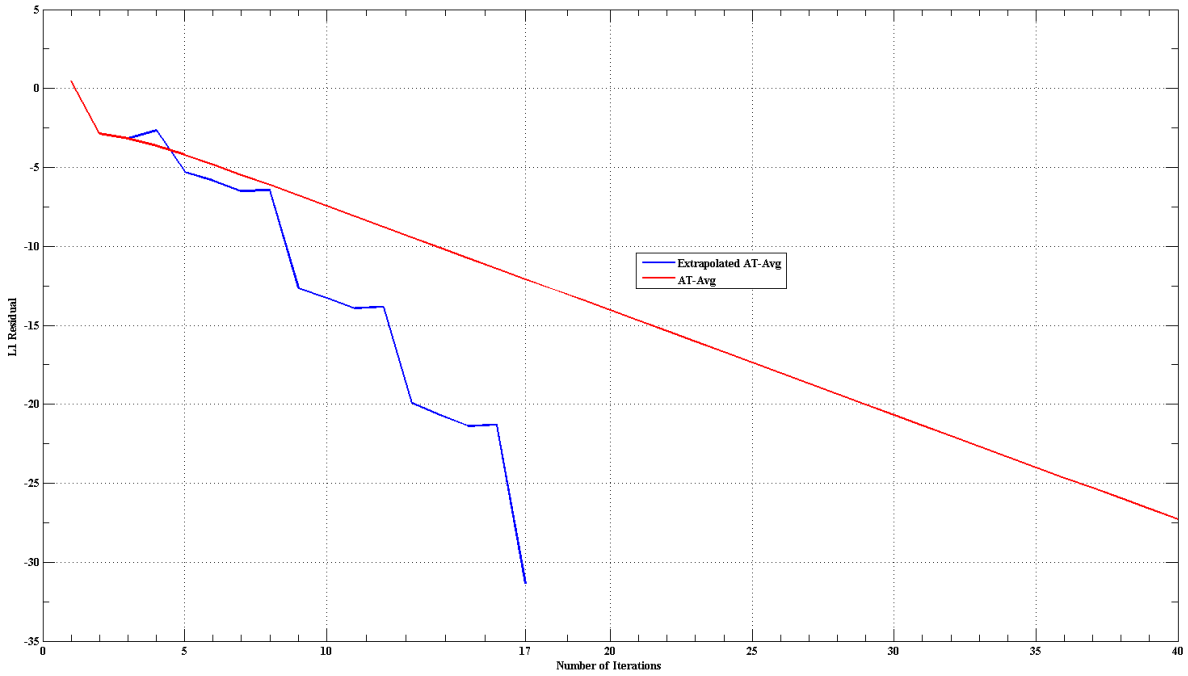


Figure 5.5: *Convergence graph for query "search engines"*

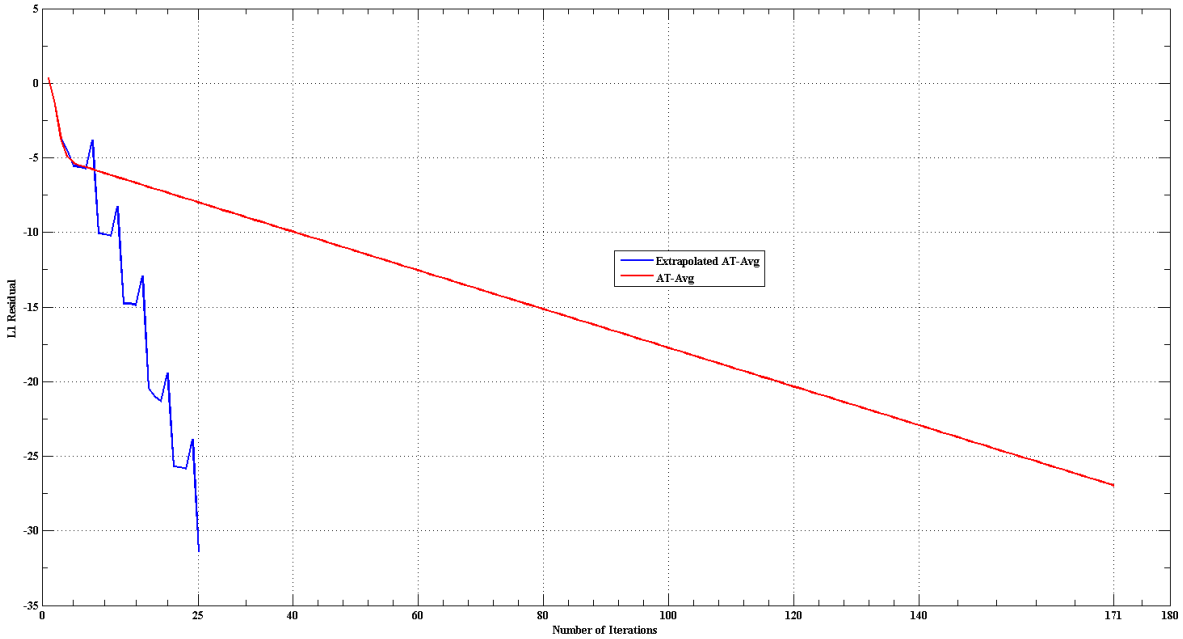


Figure 5.6: *Convergence graph for query "abortion"*

The irregular and steep convergence curve with spikes at the extrapolation point ascertains the theoretical essence of extrapolation, i.e., it subtract off approximation to the second and third eigenvectors, see Section 4.3.2, and therefore provide a much quicker convergence than normal.

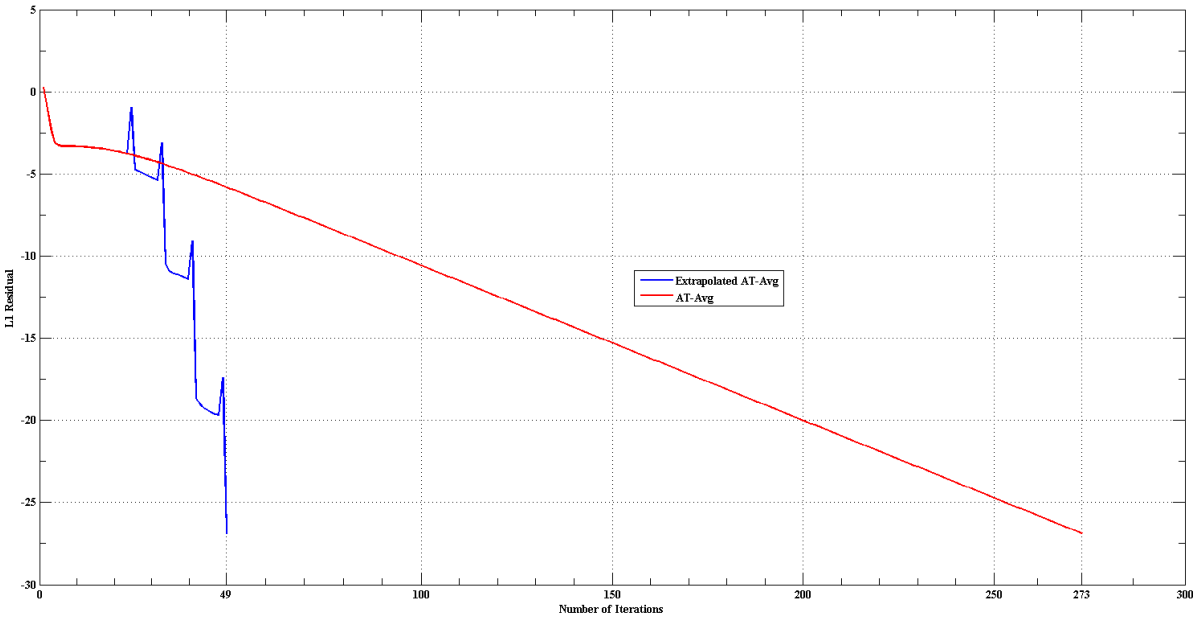


Figure 5.7: *Convergence graph for query "vintage cars"*

Consider now the multi-topic query "vintage cars", it takes only 49 iterations and extrapolation applied 4 times verses 273 iteration of the normal AT-Avg. The extrapolation steps are employed after 5^{th} iteration. As seen from the figure 5.7 and as expected, after every application of extrapolation there is a sharp reduction in L_1 Residual.

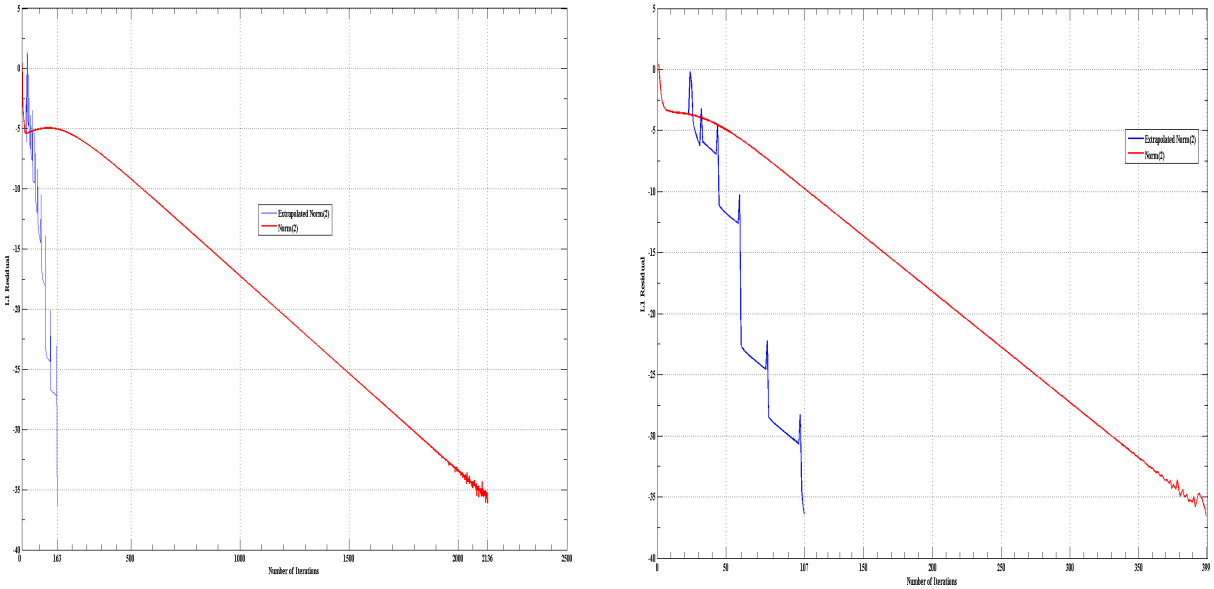
The observations about convergence due to extrapolation in the previous algorithms are also valid in the case of Authority threshold.

Norm(2)

Recall *Norm* family of algorithms from last chapter (see Section 4.2.2) – to scale the weights in a way so that lower authority weights contribute less to the hub weight. Hub weights of the nodes are taken to be the *p-norm* of the vectors of the authority weights. In the case of *Norm (2)*, it's simply the *vector norm* or *Euclidean norm* (see equation (4.8)).

The efforts per iteration in *Norm (2)* is not that much in comparison with AT-Avg and HubAvg, but the improvement due to extrapolation is reasonably comparable. An improvement of order 12; (163 iterations by extrapolated version and 2136 by original *Norm (2)*), provides a very good reason to prefer extrapolation over the original Norm.

It should also be noticed (from Appendix B) that $top-15$ results of the *Norm (2)* is the subset of the union of $top-15$ results of HITS and HubAvg. The TKC effect in HITS appears to concur in both HubAvg and Norm (2) algorithms. Looking deeper into TKC effect is out of scope of this study, for detailed analyses on TKC effect see [Tsa04a; Bor05].



(a) Convergence for query "recipes"

(b) Convergence for query "jordan"

Figure 5.8: *The convergence graphs - $Norm(2)$ algorithm.*

Max

Max algorithm, see Section 4.2.2, a special case of both AT algorithm for threshold value $k = 1$ and of *Norm (p)* algorithm for value $p = \infty$, i.e., $\infty - norm$. The hub score in $\mathbb{O}$ operation is therefore computed by applying max operator to authority weights.

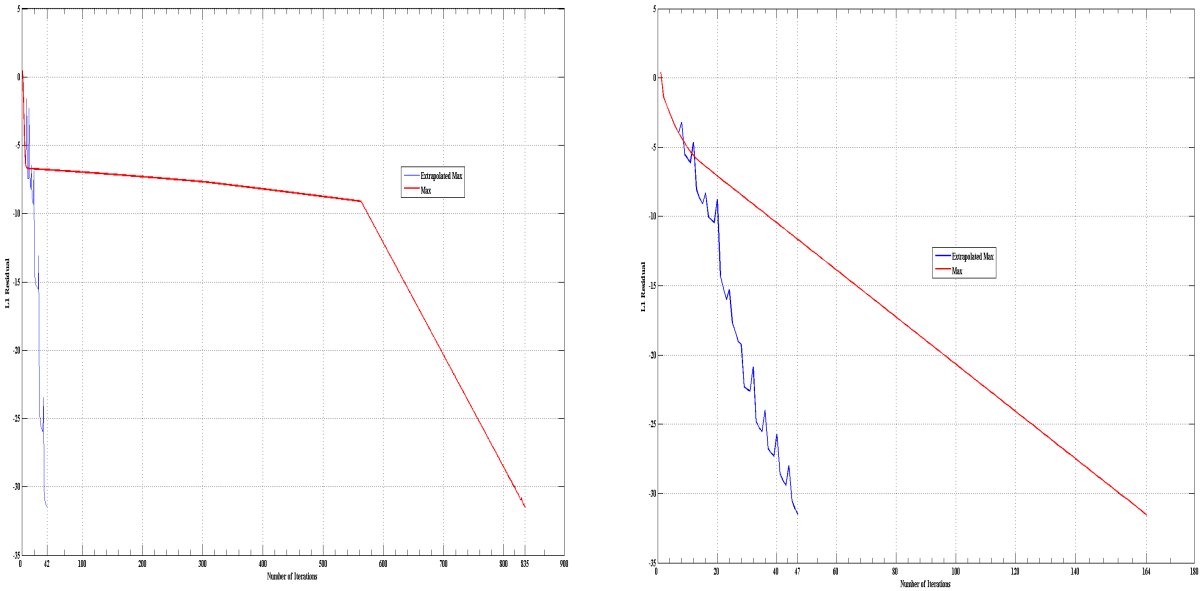
Like AT algorithm, Max algorithm is also computationally expensive per iteration because of the application of max operator. Max algorithm is deemed as second best algorithm in their study by Borodin et al., [Bor05]. To an extent Max algorithm gets rid of the TKC effect present in it predecessor HITS (see $top - 15$ results in Appendix B). It is also deemed to be good enough in finding relevant webpages to a *webpage query*. That is, Max algorithm is capable of discovering webpages, related to a query webpage [Tsa04b].

The improvement in convergence by extrapolated Max algorithm is quite analogous to that of Extrapolated AT-Avg and Norm algorithms.

For the query "basketball", the extrapolated Max is 19 times (the most astonishing result) faster than the normal Max (see figure 5.9). Extrapolation applied just 6 times (after 5^{th} iteration), forced the slowly converging sequence to rapidly converge in just 42 iterations. Without extrapolation Max algorithm took 835 iterations to converge.

SALSA

As introduced by Lempel and Moran (see Section 3.5), we have also exposed SALSA to the experimental evaluation, in order to observe its convergence behaviour in exposition to Extrapolation. As with the previous algorithms, we tweak the parameters in the algorithm to sort out any observable conduct in the convergence curves.



(a) Convergence for query "basketball"

(b) Convergence for query "architecture"

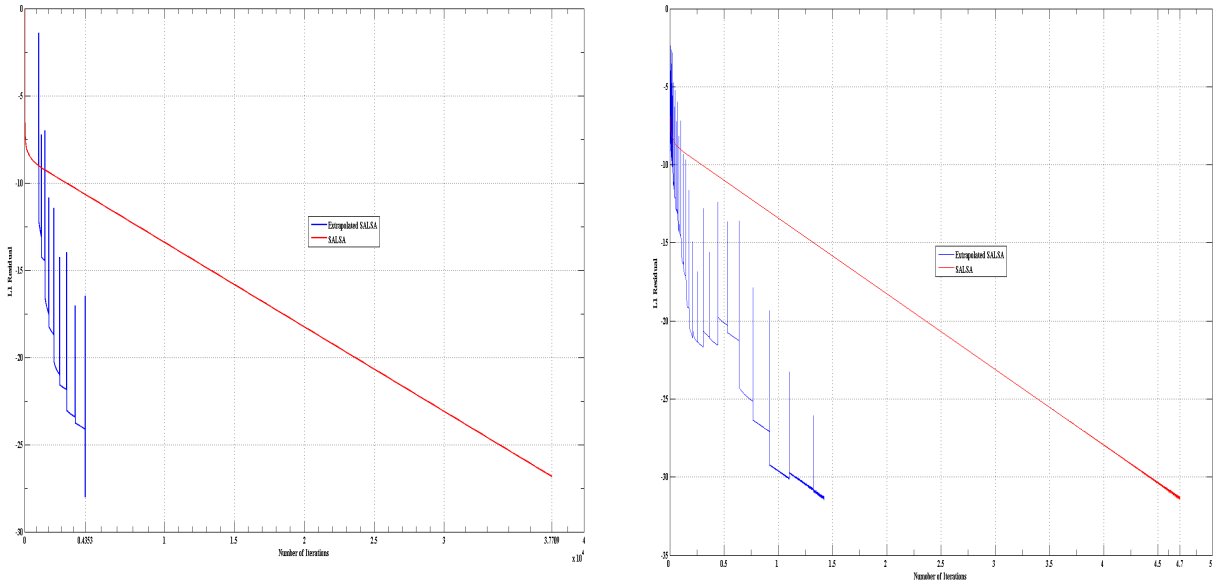
Figure 5.9: *The convergence graphs - Max algorithm.*

From the experiments, SALSA takes a lot of iterations to converge, in average more than 1000 iterations to reach to residual ϵ. In contrast to the findings by Najork [Naj07a], SALSA does not really outperforms HITS (or its descendants) in convergence behaviours. But in terms of relevancy, the $top - 15$ results produced by SALSA seem to resemble the antecedent of all LAR algorithms, the *InDegree* algorithm (see Section 3.2 and for results see Appendix B, and compare it with [Tsa]).

The worst case is when query "complexity" took $37,709$ iterations to converge, while the extrapolated version took about $13,000$ iterations (where extrapolation is applied 39 times) to reach to a L_1 residual 10^{-12}. By changing the starting point and frequency of the extrapolation we get convergence after $4,353$ iterations; extrapolation is applied 9 times starting after 1000^{th} iterations. Thus it took half the number of iterations of the case when extrapolation is applied 39 times (see figure 5.10).

Careful application of extrapolation therefore gives a speedup of a factor of 2 over applying it haphazardly. Therefore it's very much important to see when it is required and appropriate to apply extrapolation and how often it should be applied to gain a pragmatic and significant improvement.

Any other measure of convergence might be helpful; L_1 residual doesn't really scale very well with SALSA. A similar measure as described by Haveliwala (see Section 5.1.5) applied to SALSA might provide any interesting insight about its convergence sensitivity. It might be that we donot need to iterate so much to converge to an almost good solution. Thus detecting convergence from *induced ordering* might be helpful in case of SALSA. If the ranking order stays the same irrespective of the weights (values), then applying any further iteration might not bring any significant changes except the slight changes in values (weights).



(a) Convergence for query "complexity" - 9 times Extrapolation applied

(b) Convergence for query "complexity"- 39 times Extrapolation applied

Figure 5.10: *The convergence graphs - SALSA algorithm.*

Summary and Results of Extrapolation

Extrapolation techniques are therefore very effective, the extrapolated algorithms in our experiments yielded a net speedup of over 3 (see Table 5.3), the speedup could be even more significant in practice; for example we even got a speedup of 19 on our dataset depending on careful application of *extrapolation step* (see figure 5.9(a) for query "basketball").

See table 5.3 for a comprehensive overview of all the results for each of 34 queries defined in Table 5.1 that we used, and for all of the algorithms that we discussed in this section. In the table:

-*itr*, is the number of iterations
-*ext*, is the number of times extrapolation applied
-E, is the Extrapolated version of the algorithms, and
-N refers to normal version

		HITS		**HubAvg**		**AT-avg**		**Norm**		**Max**		**SALSA**	
Query	#	**E**	**N**	**E**	**N**	**E**	**N**	**E**	**N**	**E**	**N**	**E**	**N**
abortion	itr	12	17	43	106	25	171	25	39	33	113	323	1436
	ext	3		8		6		3		7		14	
affirmative action	itr	397	2529	71	199	79	327	125	356	75	182	375	2195
	ext	3		10		7		17		9		19	
alcohol	itr	21	38	41	135	17	35	23	42	21	33	281	1415
	ext	2		6		4		5		4		18	
amusement parks	itr	26	57	34	61	37	86	95	193	37	81	648	4106
	ext	5		8		6		10		5		21	
architecture	itr	45	98	155	817	109	792	48	131	47	164	1021	7428
	ext	4		17		11		9		10		24	
armstrong	itr	30	68	26	44	20	33	32	51	61	502	6400	62545
	ext	5		4		4		5		8		35	
automobile industries	itr	50	175	64	217	25	51	35	60	29	56	1171	7951
	ext	9		10		5		5		5		25	

		HITS		HubAvg		AT-avg		Norm(2)		Max		SALSA	
Query	#	**E**	**N**	**E**	**N**	**E**	**N**	**E**	**N**	**E**	**N**	**E**	**N**
basketball	itr	20	27	49	119	29	76	35	66	42	835	590	3885
	ext	3		3		5		6		6		22	
blues	itr	29	52	48	163	25	64	34	88	37	93	124	672
	ext	6		11		5		5		6		13	
cheese	itr	17	24	44	87	46	156	38	100	33	60	497	3732
	ext	3		7		4		5		6		21	
classical guitar	itr	43	160	38	88	34	71	49	128	41	117	313	1738
	ext	10		8		3		6		8		18	
complexity	itr	21	32	101	321	26	66	24	43	46	132	4353	37709
	ext	3		11		5		5		9		9	
comput. complexity	itr	53	144	64	661	18	29	25	45	21	38	90	415
	ext	5		9		3		3		4		12	
comput. geometry	itr	28	58	44	98	33	161	43	178	25	43	73	320
	ext	6		6		7		5		4		11	
death penalty	itr	12	18	36	70	14	21	25	49	21	36	329	1902
	ext	1		5		3		5		4		18	
genetic	itr	26	40	43	91	17	30	25	42	21	38	169	908
	ext	3		7		4		5		4		15	
geometry	itr	25	45	45	108	23	44	35	18	37	111	161	992
	ext	5		6		5		7		8		15	
globalization	itr	17	24	52	139	17	40	19	28	25	47	406	2529
	ext	3		5		4		3		4		20	
gun control	itr	43	149	100	448	25	51	51	133	21	32	281	1622
	ext	4		11		6		7		4		18	
iraq war	itr	146	736	39	70	22	35	34	62	69	197	953	5032
	ext	14		5		4		4		7		24	
jaguar	itr	22	31	131	879	49	166	25	43	55	218	2561	13420
	ext	4		6		9		5		8		30	
jordan	itr	30	78	62	131	61	214	107	399	30	62	513	3053
	ext	3		6		9		6		6		21	
moon landing	itr	34	79	59	178	25	83	31	54	46	126	853	7558
	ext	2		4		6		4		5		24	
movies	itr	131	568	41	72	21	39	28	44	26	40	698	7484
	ext	10		4		5		5		6		22	
national parks	itr	17	21	33	95	81	543	25	62	19	31	2133	24404
	ext	2		4		3		4		3		29	
net censorship	itr	35	76	284	1048	18	31	31	63	22	46	269	1654
	ext	3		18		4		4		5		17	
randomized algorithms	itr	63	193	80	239	61	235	164	692	43	139	281	1554
	ext	13		9		10		9		10		18	
recipes	itr	90	397	53	113	28	83	163	2136	26	61	3689	16299
	ext	12		7		6		11		6		32	
roswell	itr	100	341	175	538	35	78	52	101	41	115	424	3261
	ext	13		11		5		7		10		20	
search engines	itr	9	13	23	41	17	40	14	39	15	21	1585	9904
	ext	2		3		4		3		3		27	
shakespeare	itr	29	72	64	270	23	49	25	51	26	49	194	1243
	ext	5		6		3		5		5		16	
table tennis	itr	26	45	42	114	16	24	22	34	18	27	193	1305
	ext	6		9		4		4		4		16	
vintage cars	itr	35	60	91	587	49	273	44	96	33	77	429	3388
	ext	3		7		4		5		7		20	
weather	itr	26	53	32	54	17	33	23	36	20	29	337	3611
	ext	6		4		3		4		4		19	
Average	itr	**50.2**	**191.7**	**67.9**	**247.1**	**33.6**	**124.4**	**46.3**	**167.7**	**34.2**	**116.2**	**962.3**	**7255.0**
	ext	**5.3**		**7.5**		**5.2**		**5.8**		**6.0**		**20.1**	
Median	itr	**29.0**	**59.0**	**48.5**	**125.0**	**25.0**	**65.0**	**33.0**	**61.0**	**31.5**	**61.5**	**415.0**	**3157.0**
	ext	**4.0**		**7.0**		**5.0**		**5.0**		**6.0**		**19.5**	
StdDev	itr	**69.1**	**443.3**	**52.0**	**263.8**	**21.9**	**160.9**	**38.4**	**372.4**	**14.8**	**155.3**	**1382.0**	**12390.5**
	ext	**3.6**		**3.5**		**2.1**		**2.8**		**2.1**		**5.9**	

Table 5.3: *Results of the experiments with* ***Extrapolation***

The extrapolation columns in the Table 5.3 indicate the best performance in terms of number of iterations that we got as a result of tweaking the parameters. Overall we have applied extrapolation steps **8** times on an average to get rapid convergences. So, the overhead of net application of extrapolation is very less, almost $3-4$ iterations of power method. The last two columns in the Table 5.3 show the slow converging behaviours of SALSA (on average **7255** iterations). But the other columns on average converge after just **46** as a result of extrapolation in comparison to the average **170** iterations of the original algorithms. A net average speedup of

order **5.78** in all the algorithms presents a very good reason for the usefulness of extrapolation techniques.

Note that Extrapolation technique can also be applied in conjunction with other acceleration techniques, such as *BlockRank* [Kam03a], or other iterative algorithms e.g., *Gauss-Seidel*, *Successive Over Relaxation*, *Conjugate Gradient* or any other methods [Gol96; Lan06]. When used in conjunction with any other methods, we might expect more insights about the effectiveness of Extrapolation, both in terms of time and convergence.

We have had a limited evaluation of the hybrid implementation of extrapolation technique (defined in Section 4.3.4). When we discuss the findings of Power $\mathbf{A}^d$ Extrapolation we will also put forward the findings from hybrid approach of extrapolation. We have tested the application of power extrapolation interchangeable with quadratic extrapolation to discover any further speedup or insights.

5.2.2 Personalization

As explained in the last chapter (see Section 4.4) there has been some work done to personalize the search results according to the user's internal model of "authority", in this section we will provide empirical implications of some of those models. We will mostly experiment with the generic model for personalization described in Section 4.4.2. A slight effort towards modifying the weights of entries in the adjacency matrix, such as with *Exponentiated HITS*, in order to influence the authoritative sources.

In general, through personalization, we intend to boost certain interesting documents according to the user's (or user group's) preferences, usually represented by the personalization vector. Hence we will modify the original query-dependent LAR algorithms identified earlier, in order to incorporate them within personalization model. We will primarily expose HITS within the generic framework of personalization (see Section 4.4.2), and then present other possible ways for personalization, e.g., personalizing the *Randomized* and *Exponentiated* HITS.

The results produced by HITS and SALSA are maybe generally *authoritative* considering the network structure of pages, but may not reflect the user's *preferences*. There is a valid rationale to modify basic HITS to take into account the user's interests and therefore personalization is an effort towards changing the general notion of authority to user's own and internal notion of authority and importance.

Personalized HITS

See equation (4.63) in the last chapter, according to the formulation $x = \mathbf{Q}\vec{v}$ in equation (4.54), we could also incorporate the same modification to the HITS equation (4.63). Thus the $\mathbb{I}$ operation in HITS with personalization, will look like:

$$\vec{a}^{(k)} = \alpha \mathbf{A}^T \mathbf{A}\ \vec{a}^{(k-1)} + (1-\alpha)\vec{v} \tag{5.2}$$

where $\vec{v}$ is the personalized probability vector. According to the random surfer model, when random surfer gets bored, he jumps to randomly chosen destination according to the probability distribution given in $\vec{v}$. If $\vec{v}$ contains uniform probability distribution (e.g., $\vec{v} = \frac{1}{n}e$), then the above equation will be the same as that of the original HITS. The convex combination of authority vector $\vec{a}$ and uniformly distributed personalized vector $\vec{v} = \frac{1}{n}e$, experimentally gives the same results as the original HITS algorithm.

Consider the $top-15$ result in Table 5.4 for query "search engines", if we want to boost the page 'About Web Search' on 7^{th} position (in boldface), we must change the *uniformly* distributed personalized vector $\vec{v}$ in equation (5.2) to a *non-uniform* probability vector. By giving a higher probability to the index 143 (representing the page index **P-143**) in the personalized vector $\vec{v}$, we are implicitly provoking the random surfer to choose

Rank	Page Index	Title	URL
1.	P-135	AltaVista	www.altavista.com
2.	P-1401	Ego Surf - EgoSurf - egosurf.com	www.egosurf.com
3.	P-5225	Yahoo! Danmark	www.yahoo.dk
4.	P-3390	AltaVista Text-Only Search	ragingsearch.altavista.com
5.	P-4539	Euroseek	euroseek.net
6.	P-1721	Your Search Engine Internet directory, information, search engine	www.searchpalm.com
7.	**P-143**	**About Web Search - Guide to Search Engine Optimization & Online S**	**websearch.about.com**
8.	P-4979	Abacho - THE POWERFUL SEARCHENGINE!!	www.abacho.co.uk
9.	P-1844	careerhighway.com	www.careerhighway.com
10.	P-5766	Ananzi South Africa - Search Engine	www.ananzi.co.za
11.	P-1303	Empty title field	www.portalhub.com
12.	P-5283	DINO-Online Suchmaschine Webkatalog Linkliste Internet Verzeichni	www.dino-online.de
13.	P-5296	Lycos.de	www.lycos.de
14.	P-5264	Excite France	www.excite.fr
15.	P-5285	Fireball - Die Suchmaschine	www.fireball.de

Table 5.4: *Top 15 results for query "search engines"*

page **P-143** when he has to jump to a destination. The probability of page index **P-143** should be raised relative to the probabilities assigned to other indices in vector $\vec{v}$.

Rank	Page Index	Title	URL
1.	P-135	AltaVista	www.altavista.com
2.	P-1401	Ego Surf - EgoSurf - egosurf.com	www.egosurf.com
3.	P-5225	Yahoo! Danmark	www.yahoo.dk
4.	**P-143**	**About Web Search - Guide to Search Engine Optimization & Online S**	**websearch.about.com**
5.	P-3390	AltaVista Text-Only Search	ragingsearch.altavista.com
6.	P-4539	Euroseek	euroseek.net
7.	P-1721	Your Search Engine Internet directory, information, search engine	www.searchpalm.com
8.	P-4979	Abacho - THE POWERFUL SEARCHENGINE!!	www.abacho.co.uk
9.	P-1844	careerhighway.com	www.careerhighway.com
10.	P-5766	Ananzi South Africa - Search Engine	www.ananzi.co.za
11.	P-1303	Empty title field	www.portalhub.com
12.	P-5283	DINO-Online Suchmaschine Webkatalog Linkliste Internet Verzeichni	www.dino-online.de
13.	P-5296	Lycos.de	www.lycos.de
14.	P-5264	Excite France	www.excite.fr
15.	P-5285	Fireball - Die Suchmaschine	www.fireball.de

Table 5.5: *Top 15 personalized results for query "search engines"* ($\alpha = 0.6$)

The $top - 15$ results in Table 5.5 shows that now the page $P - 143$ is moved from previous 7^{th} position to the 4^{rth} position higher in the ranking. The damping factor is set to $\alpha = 0.6$ in this case. But in case if users like to see the interested page much higher, then we have to further reduce the damping factor. For example, if we have $\alpha = 0.15$, the resultant ranking order will be now given in Table 5.6. The other way to elevate page's ranking is to raise it's corresponding probability weight in the personalization vector $\vec{v}$.

If the page(s) to boost are lower down in the ranking, e.g., 'WoYaa search engine'[2] appears on 38^{th} position; boosting this page would require the random surfer to jump more often (i.e., $\alpha \leq 0.2$) and also the weight in the personalization vector should be relatively higher.

Hence it is highly rewarding to have an *efficient* and *automated* way, to orchestrate the values of the vector $\vec{v}$, the adjacency matrix $\mathbf{A}$, and the damping factor α. The automated solution can provide further insight about the effectiveness of these parameters on the ranking of the search outcomes. In our study we have manually

[2]African search engine and Website's Directory

Rank	Page Index	Title	URL
1.	P-135	AltaVista	www.altavista.com
2.	**P-143**	**About Web Search - Guide to Search Engine Optimization & Online S**	**websearch.about.com**
3.	P-1401	Ego Surf - EgoSurf - egosurf.com	www.egosurf.com
4.	P-5225	Yahoo! Danmark	www.yahoo.dk
5.	P-3390	AltaVista Text-Only Search	ragingsearch.altavista.com
6.	P-4539	Euroseek	euroseek.net
7.	P-1721	Your Search Engine Internet directory, information, search engine	www.searchpalm.com
8.	P-4979	Abacho - THE POWERFUL SEARCHENGINE!!	www.abacho.co.uk
9.	P-1844	careerhighway.com	www.careerhighway.com
10.	P-5766	Ananzi South Africa - Search Engine	www.ananzi.co.za
11.	P-1303	Empty title field	www.portalhub.com
12.	P-5283	DINO-Online Suchmaschine Webkatalog Linkliste Internet Verzeichni	www.dino-online.de
13.	P-5296	Lycos.de	www.lycos.de
14.	P-5264	Excite France	www.excite.fr
15.	P-5285	Fireball - Die Suchmaschine	www.fireball.de

Table 5.6: *Top 15 personalized results for query "search engines"* ($\alpha = 0.15$)

adjusted the parameters in order to observe their impact on ranking.

The other factor that is decisive in personalization, is the configuration of the personalized vector $\vec{v}$ for a particular user. The question is, how to accumulate and update the users specific personalized vectors $\vec{v}$, to capture or represent their interests. As discussed in the last chapter, the matrix $\mathbf{Q}$, in the expression $x = \mathbf{Q}\vec{v}$, provides the complete *basis* for personalized vector $\vec{v}$. But here we are only concerned with the one vector that represents a particular user's preferences. There could be several ways to acquire the users' preference information, for example, through *relevance feedback* cycles. Depending on the requirement and specification of the overall system, either each *user* or *user group* could be assigned individual personalization vector(s) representing their interests.

The personalization vector specific to a user or user group can be updated *explicitly* or *implicitly* as described in Section 4.4. Either they should be explicitly asked for their preferences, or their history (e.g., usage or interaction information from logs, such as click-through, time on page, etc) or bookmark information could be monitored and therefore could be used *intelligently* by the system to form their personalization vector. Acquiring meaningful implicit feedback from the tracing of the "noisy" users' interaction information in search logs is yet another interesting area in personalization. Implicit relevance measures and feedbacks have been recently studied by several research groups and its quite an active and contemporary area in ranking and personalization [Agi06a; Agi06b].

User-profiles (capturing the implicit or explicit feedback) can be used to maintain the representation of the users' interests. Considering the behaviour of web user, their interests and preferences keeps on changing repeatedly and unexpectedly. The IR system must keep track of the updates in the user preferences and hence make appropriate and necessary modifications in their assigned personalization vectors. In the study by Baudisch [Bau01], an information filtering architecture has been designed which is capable of dealing with frequently changing user interests. The architecture called as *QuerySet filtering architecture (QSA)* uses *user-profiles* to handle both gradual and abrupt changes in user preferences. Different interfaces are discussed through which user-aggregated relevance feedback are gathered by allowing them to manually handle major profile inaccuracies. For more information see [Bau01].

The next and main concern is the *dissemination* of user preferences and its generalization (as discussed in Section 4.4). Users do not really want to only boost the ranks of the specific pages that they consider

authoritative (preferences), but they expect a *generalization* of their preferences. The set of authoritative pages that is maintained in the personalization vector (or user-profiles) should be used rather more actively than just to boost their rankings. This means that the system must not only raise the relevancy of the pages preferred by the user, but also boost the related pages to the preferred set. The authority given to the preferred pages must be extended to the related pages as well, which might be more authoritative. The idea of dissemination of user preferences is same as that of the *spread of activation* presented by Anderson and Pirolli [AND84].

As described in Section 4.4.4, through *gradient ascent* we could achieve the dissemination of authorities. The effect of gradient ascent would be quite similar to the approaches toward interpretation of user feedbacks from their interaction logs.

The empirical analyses for these approaches would be out of scope of this work. For further information on the approaches described here see [Mil01; Cha00; Pir96; Agi06b].

Power Extrapolation (A^d)

As described in last chapter, there is another extrapolation method proposed by Haveliwala et al., (see Section 4.3.3). This method of extrapolation takes into account the *second eigenvalue* of the Markov matrix and uses it to subtract off the error along several non-principal eigenvectors.

There are numerous ways for finding the second eigenvalue of the Markov matrix ($\mathbf{A}^T\mathbf{A}$ in case of HITS), as discussed in last chapter, e.g., by *deflation* (which is computational prohibitive in our case). As described in Section 4.3.3, to get an approximation of the second eigenvalue of the row-stochastic Markov matrix $\mathbf{A}$, in the equation (4.45) the damping factor α serve to be the second eigenvalue of the $\mathbf{A}$, i.e.,:

$$\lambda_2 = \alpha \tag{5.3}$$

The above expression is formally proved by Haveliwala and Kamvar in [Hava].

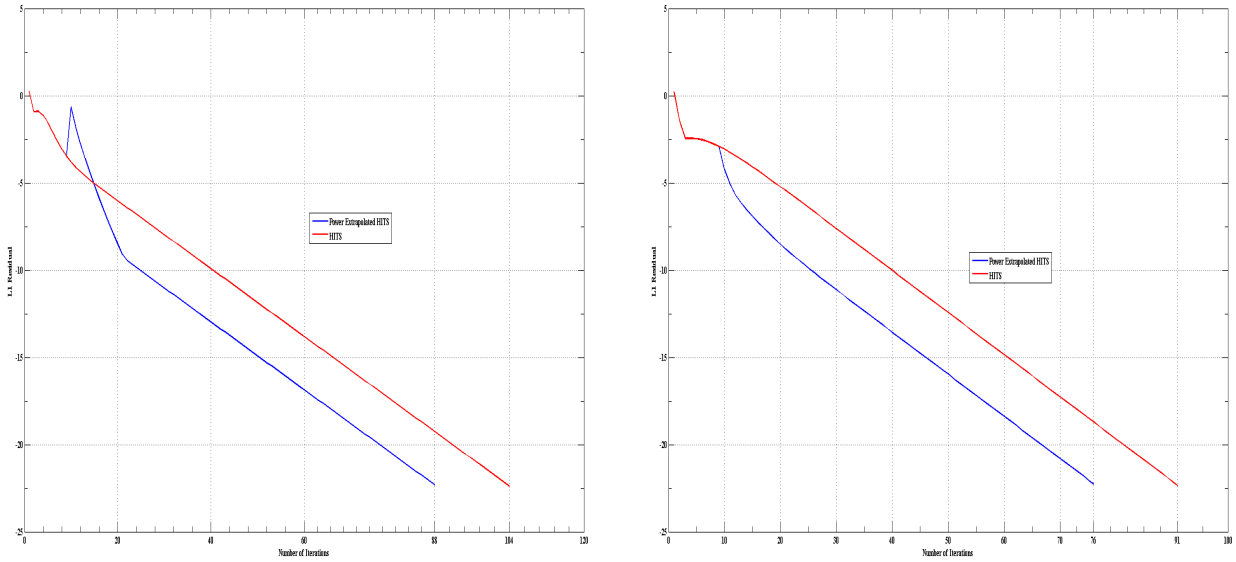
Therefore by using the damping factor as the second eigenvalue we could easily implement Power Extrapolation method. The purpose of evaluating power extrapolation under personalization is also because we will use the personalization scheme as a way to do the *power extrapolation*.

The convergence due to *Power Extrapolation* as observed by Haveliwala et al. [Havb], should be similar to *Quadratic extrapolation* evaluated in the last section. But with the query-dependent LAR algorithms, Power Extrapolation doesn't manifest as desirable speedup as in Quadratic extrapolation case. We are getting on average improvements of order $10 - 25\%$. Hence the application of Power Extrapolation doesn't seem to be equivalent or comparable to that of Quadratic extrapolation in query-dependent algorithms.

Consider for example, the query "automobile industries" (see figure 5.11), the power extrapolated HITS converges in 88 iterations while normal HITS takes 104 iterations, the damping factor $\alpha = 0.9$ and $d = 6$ (recall $\mathbf{A}^d$). Notice that extrapolation is only applied once. Also notice that the convergence curve of the power extrapolated HITS resembles the curve of the query independent PageRank shown in [Havb].

Note that in most of our experiments we kept the damping factor reasonably *high* in order to achieve better convergence. Thus, by doing that we are indirectly preventing the spammers to artificially inflict the rankings (recall the effect of the damping factor on spamming, Section 3.3.7).

The application of power extrapolation to query-dependent algorithms HITS and SALSA therefore does not offer any practically prominent improvements. But the application of power extrapolation is useful in a sense that it exploits the damping factor to accelerate the convergence. That is, it relies on already existing information and it's quite simple to implement. The cost of one time application of power extrapolation is



(a) Convergence for query "automobile industries"

(b) Convergence for query "computational complexity"

Figure 5.11: *Convergence graphs - Power Extrapolated HITS algorithm.*

almost negligible. And finally together with extrapolation, we now have a room for personalization as well with the help of power extrapolation.

Hybrid Extrapolation

The *Hybrid* approach discussed in Section 4.3.3 could be a novel approach towards extrapolation. We have applied Quadratic Extrapolation interchangeably with Power Extrapolation to see any abrupt trends in the convergences. But the convergences produced are almost comparable to the results in Quadratic Extrapolation. The limited evaluation of the hybrid doesn't offer any significant insights into the usefulness of the hybrid approach, but the idea of exploiting multiple extrapolation techniques is well worth. The primary purpose of using the hybrid approach is to observe the inter-dependencies of the extrapolation techniques and the effect those dependencies on the convergence graph.

Consider for example, the multi-topic query "affirmative action", we have applied Power Extrapolation step after 6^{th} iteration, and after that Quadratic extrapolation took the charge. The convergence graph is shown in figure 5.12. There is a slight improvement in comparison to the figure 5.2, which was when we only applied Quadratic Extrapolation. In only Quadratic extrapolation approach we got convergence in 397^{th} iteration while in the hybrid approach we are getting convergence in just 314^{th} iteration. We have tested the hybrid approach with different settings, e.g., we have applied power extrapolation step at different *time steps* together with quadratic extrapolation. Sometimes applied at the start of the convergence curve, sometimes in the middle and sometimes at the end of the curve.

There is certainly more to the hybrid approach of extrapolation. The figure 5.12 provides a very cursory benefit of the hybrid approach of extrapolation, it could be influential in other cases. Instead of just using power and quadratic extrapolations, if any other new or existing extrapolation techniques could be used, it

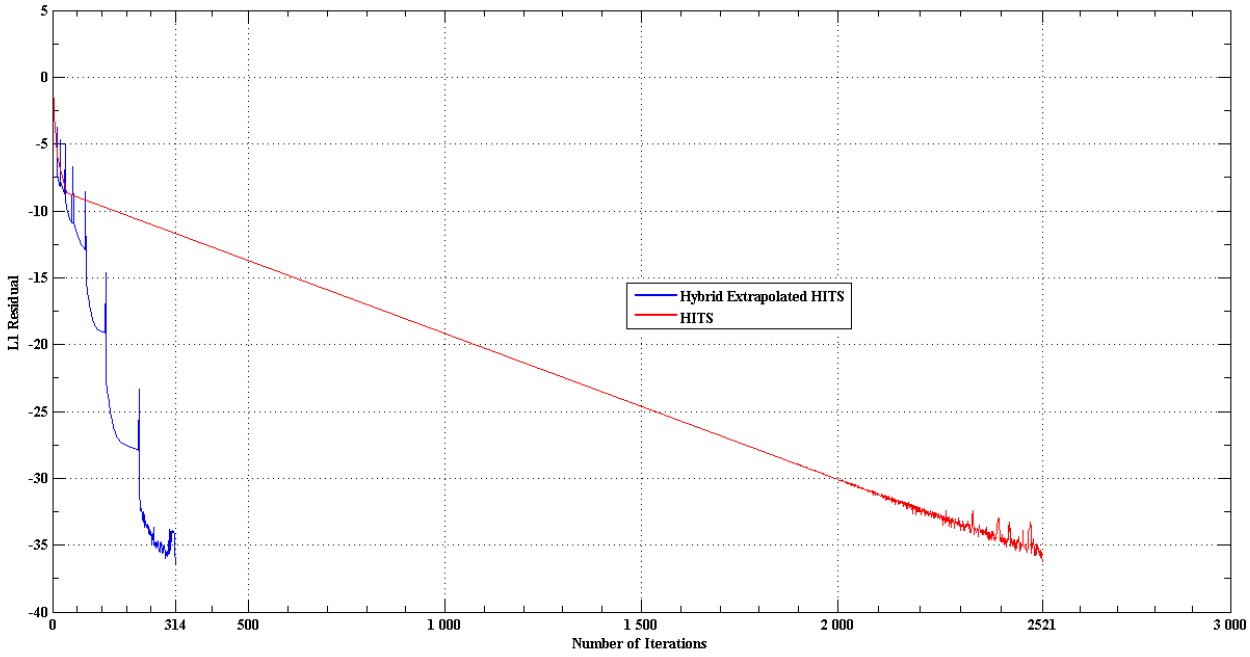


Figure 5.12: *Convergence for query "affirmative action" - Hybrid Extrapolation*

might offer some more desirable improvements. It might as well be used for personalization more actively, e.g., now we could only boost certain pages through the help of power extrapolation. If the extrapolation premise or assumption (as defined in Section 4.3.4) is based on the personalization scheme then it could be possible to have twofold benefits of extrapolation, both in terms of convergence speedup and personalized outcomes.

Exponentiated HITS

Recall *Exponentiated HITS* (see Section 4.4.4), to realize the motivation we start with an example (see figure 5.13); running the original HITS algorithm first on the figure 5.13(a), yields the following authority and hub vectors:

```
a = (0      0.1667  0.1667  0.0  0.0  0.0    0.0    0.1667  0.1667  0.1667  0.1667)
h = (0.3333 0.0     0.0     0.0  0.0  0.3333 0.3333 0       0       0       0)
```

We see a lot of *zeros* and the weights are haphazardly *repeated* too, and hence the ranking appears to be impractical. Now running Exponentiated HITS on the same graph we get the following *unique* authority and hub vectors:

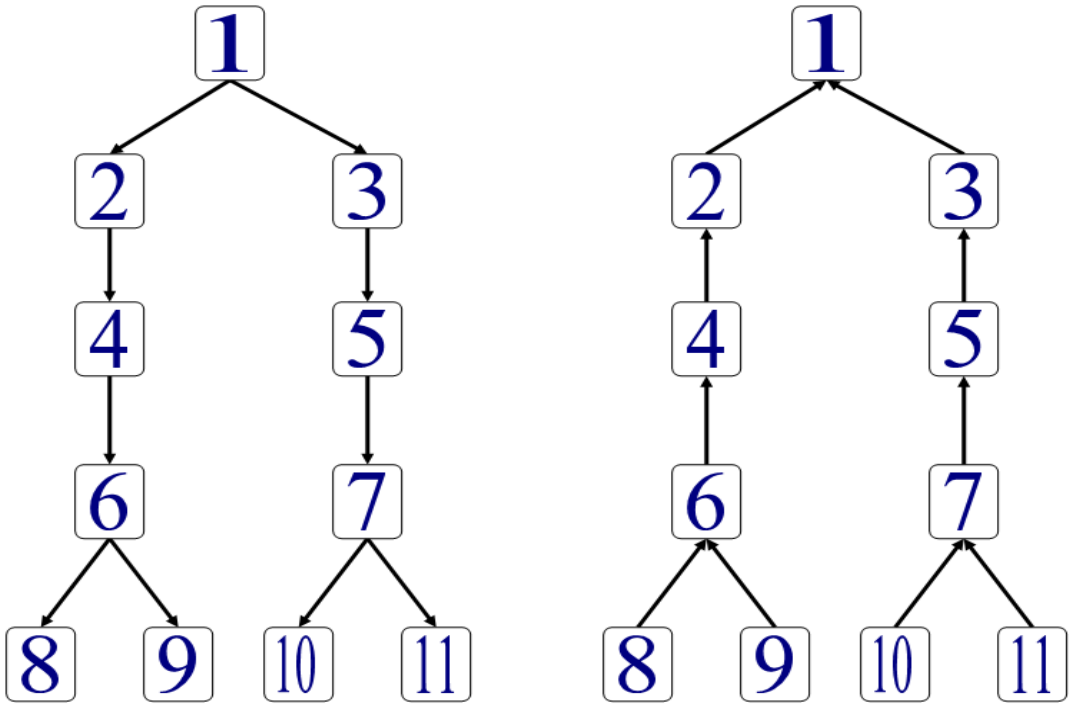
```
a = (0      0.1029 0.1029 0.1048 0.1048 0.099   0.099  0.0962 0.0962 0.0962  0.0962)
h = (0.2382 0.1237 0.1237 0.1298 0.1298 0.1274  0.1274 0      0      0       0)
```

The Exponentiated HITS gives the following weights for the figure 5.13(b):

```
a = (0.2382 0.1237 0.1237 0.1298 0.1298 0.1274  0.1274 0      0      0       0)
h = (0      0.1029 0.1029 0.1048 0.1048 0.099   0.099  0.0962 0.0962 0.0962  0.0962)
```

And the normal HITS:

```
a = (0.3333 0.0     0.0     0.0  0.0  0.3333 0.3333 0       0       0       0)
h = (0      0.1667  0.1667  0.0  0.0  0.0    0.0    0.1667  0.1667  0.1667  0.1667)
```



(a) Problem a (b) Problem b

Figure 5.13: *Two examples.*

Page 1, 6 and 7 have a tie in case of normal HITS on figure 5.13(b), while the Exponentiated HITS clearly provide a more practical ranking. Pages 2, 3, 4, and 5 have zero authority just like pages 8, 9, 10 and 11, on the same graph, which isn't favourable.

From the above two examples in figure 5.13 we see that HITS actually doesn't perform well on some types of graph, the rank produced are *inconsistent* (containing a lot of zero and repeated values). Exponentiated HITS in such types of situations provide a much practical relevancy scores by supplementing the inconsistencies in the basic HITS. Hence Exponentiated HITS provides authoritativeness more precisely and distinctively, consistent with the graph structure. Exponentiated HITS addresses both the repeated weights and zero weights problems existed in the original HITS.

After this motivation we are now ready to try out the personalized Exponentiated HITS. Like its predecessors Exponentiated HITS should be formulated within the general formulation of personalization.

On our given dataset corresponding to the queries in Table 5.1 Exponentiated HITS performs reasonably good. The $top-15$ results depict an intersection of the authoritative sources in all the algorithms that we have discussed here. But the operation $(e^A - I)$ is computationally stiff, for a matrix $\mathbf{A}$ of size $(5,000 \times 5,000)$, it takes about $2-5$ minutes just to compute the exponentiation. But after exponentiation, the convergences are usually quite fast, on an average we had convergence in order $15-30$ iterations.

The $top-15$ results for the query "computational complexity" are shown in Table 5.7. The ranking order is different from HITS's, but we see a lot of important authoritative pages in this list in comparison to the other algorithms (see Appendix B for comparison).

Rank	Page Index	Title	URL
1.	P-1	ECCC - The Electronic Colloquium on Computational Complexity	eccc.uni-trier.de/eccc
2.	P-9	Computational Complexity Conference	computationalcomplexity.org
3.	P-384	Center for Discrete Mathematics and Theoretical Computer Science (DIMACS)	dimacs.rutgers.edu
4.	P-370	IEEE Conference on Computational Complexity	www.cs.utep.edu/longpre/complexity.html
5.	P-242	The Complexity Zoo	www.cs.berkeley.edu/~aaronson/zoo.html
6.	P-256	ACM: Association for Computing Machinery, the world's first educational and scientific computing society	www.acm.org
7.	P-247	CCC'03 Research Abstracts Download	www.cse.sc.edu/~fenner/CCC03/abstract03.html
8.	P-5	My Computational Complexity Web Log	www.fortnow.com/lance/complog
9.	P-372	www.ncstrl.org	www.ncstrl.org
10.	P-328	European Association for Theoretical Computer Science	www.eatcs.org
11.	P-254	SIGACT News	sigact.acm.org/sigactnews
12.	P-241	Lance Fortnow	www.neci.nj.nec.com/homepages/fortnow
13.	P-364	Complexity 2003 - Kolmogorov day	www.lri.fr/~laplante/kolmogorov.htm
14.	P-661	Ian Parberry's Home Page	hercule.csci.unt.edu/~ian
15.	P-666	Empty title field	www.math.cas.cz/~pudlak

Table 5.7: *Top 15 results for query "computational complexity", Exponentiated HITS*

The Personalized Exponentiated HITS also offers twofold benefits – first that pages returned are more relevant and authoritative in general and secondly the possibility for personalization, like HITS.

For example, in the Table 5.7, in order to boost pages $P-241$, $P-661$ and $P-666$, to appear in $top-10$ or higher in the ranking, we will apply Personalized implementation of Exponentiated HITS. Notice that as the interested documents are lower down in the ranking therefore random surfer has to teleport frequently, thus $\alpha = 0.05$. The results of personalization in Exponentiated HITS is shown in Table 5.8.

Randomized HITS

Random surfer model as implemented in the previous sections and also in PageRank model. Here in the case of *Randomized HITS* (see Section 4.4.4) the primary purpose of random surfer model is to make HITS more *stable* to perturbations. Unlike the random surfer model in PageRank, the random surfer in Randomized HITS follows hyperlinks both in forward direction and in backward direction. The random surfer with guided teleportation will help to *bias* the rankings produced by Randomized HITS. In the original Randomized HITS model the jumps of the random surfer are chosen uniformly. But with the modifications in equations (4.61) and (4.62) the jumps of the random surfer can now be controlled with the *personalization vector* $\vec{v}$, in order to control the search outcomes. Hence this way we can direct the random surfer's jumps based on the non-uniform probability distributions.

We have first implemented equations (4.59) and (4.60) and then tested it against the set of queries given in the Table 5.1. From the $top-15$ results on almost all queries, it is found that the Randomized HITS outcomes resemble the $top-15$ results of the primitive algorithm, the *InDegree* (described in Section 3.2). It is also observed that here the rate of convergence is quite rapid, on an average we get convergence in $10-20$ iterations.

The query "affirmative action" gives the $top-15$ results from running the Randomized HITS (see table 5.9). Now consider Table 5.10 the $top-15$ results of *InDegree* Algorithm. Comparing them we visibly observe that the $top-15$ results are ranked the same by the two algorithms. Randomized HITS hence resembles InDegree

Rank	Page Index	Title	URL
1.	P-1	ECCC - The Electronic Colloquium on Computational Complexity	eccc.uni-trier.de/eccc
2.	P-9	Computational Complexity Conference	computationalcomplexity.org
3.	P-384	Center for Discrete Mathematics and Theoretical Computer Science (DIMACS)	dimacs.rutgers.edu
4.	P-370	IEEE Conference on Computational Complexity	www.cs.utep.edu/longpre/complexity.html
5.	P-242	The Complexity Zoo	www.cs.berkeley.edu/~aaronson/zoo.html
6.	**P-241**	**Lance Fortnow**	**www.neci.nj.nec.com/homepages/fortnow**
7.	P-256	ACM: Association for Computing Machinery, the world's first educational and scientific computing society	www.acm.org
8.	P-247	CCC'03 Research Abstracts Download	www.cse.sc.edu/~fenner/CCC03/abstract03.html
9.	P-5	My Computational Complexity Web Log	www.fortnow.com/lance/complog
10.	**P-661**	**Ian Parberry's Home Page**	**hercule.csci.unt.edu/~ian**
11.	P-372	www.ncstrl.org	www.ncstrl.org
12.	**P-666**	**Empty title field**	**www.math.cas.cz/~pudlak**
13.	P-328	European Association for Theoretical Computer Science	www.eatcs.org
14.	P-254	SIGACT News	sigact.acm.org/sigactnews
15.	P-364	Complexity 2003 - Kolmogorov day	www.lri.fr/~laplante/kolmogorov.htm

Table 5.8: *Top 15 results for query "computational complexity", Personalized Exponentiated HITS* ($\alpha = 0.05$)

Rank	Page Index	Title	URL
1.	P-2026	Copyright Information	www.psu.edu/copyright.html
2.	P-1	Affirmative Action and Diversity Page	aad.english.ucsb.edu
3.	P-739	Adobe Acrobat Reader - Download	www.adobe.com/. . . /readstep.html
4.	P-280	U.S. Equal Employment Opportunity Commission Home Page	www.eeoc.gov
5.	P-2	American Association for Affirmative Action	www.affirmativeaction.org
6.	P-2381	Site Meter - Counter and Statistics Tracker	sm6.sitemeter.com /stats.asp?site=sm6wobbly123
7.	P-2382	Free web counter - Site access tracker - CQ Counter	cqcounter.com /?_id=nsnewman&_lo=us
8.	P-2383	Free web counter - Site access tracker - CQ Counter	cqcounter.com
9.	P-316	National Organization for Women	www.now.org
10.	P-3	Affirmative Action Register	www.aar-eeo.com
11.	P-72	TEXT	www.eoaa.vt.edu
12.	P-7	CAA	www.caasf.org
13.	P-1741	WIU - Division of Student Services	student.services.wiu.edu
14.	P-41	PSU Affirmative Action	www.psu.edu/dept/aaoffice
15.	P-904	The United States Department of Labor Home Page, Secretary of Lab	www.dol.gov

Table 5.9: *Top 15 results for query "affirmative action", Randomized HITS*

Rank	Page Index	Title	URL
1.	P-2026	Copyright Information	www.psu.edu/copyright.html
2.	P-1	Affirmative Action and Diversity Page	aad.english.ucsb.edu
3.	P-739	Adobe Acrobat Reader - Download	www.adobe.com/. . . /readstep.html
4.	P-280	U.S. Equal Employment Opportunity Commission Home Page	www.eeoc.gov
5.	P-2	American Association for Affirmative Action	www.affirmativeaction.org
6.	P-2381	Site Meter - Counter and Statistics Tracker	sm6.sitemeter.com /stats.asp?site=sm6wobbly123
7.	P-2382	Free web counter - Site access tracker - CQ Counter	cqcounter.com /?_id=nsnewman&_lo=us
8.	P-2383	Free web counter - Site access tracker - CQ Counter	cqcounter.com
9.	P-316	National Organization for Women	www.now.org
10.	P-3	Affirmative Action Register	www.aar-eeo.com
11.	P-72	TEXT	www.eoaa.vt.edu
12.	P-7	CAA	www.caasf.org
13.	P-1741	WIU - Division of Student Services	student.services.wiu.edu
14.	P-41	PSU Affirmative Action	www.psu.edu/dept/aaoffice
15.	P-904	The United States Department of Labor Home Page, Secretary of Lab	www.dol.gov

Table 5.10: *Top 15 results for query "affirmative action", InDegree*

Rank	Page Index	Title	URL
1.	**P-2**	**American Association for Affirmative Action**	**www.affirmativeaction.org**
2.	**P-3**	**Affirmative Action Register**	**www.aar-eeo.com**
3.	P-2026	Copyright Information	www.psu.edu/copyright.html
4.	P-1	Affirmative Action and Diversity Page	aad.english.ucsb.edu
5.	P-739	Adobe Acrobat Reader - Download	www.adobe.com/. . . /readstep.html
6.	P-280	U.S. Equal Employment Opportunity Commission Home Page	www.eeoc.gov
7.	P-2381	Site Meter - Counter and Statistics Tracker	sm6.sitemeter.com /stats.asp?site=sm6wobbly123
8.	P-2382	Free web counter - Site access tracker - CQ Counter	cqcounter.com /?_id=nsnewman&_lo=us
9.	P-2383	Free web counter - Site access tracker - CQ Counter	cqcounter.com
10.	P-316	National Organization for Women	www.now.org
11.	P-72	TEXT	www.eoaa.vt.edu
12.	P-7	CAA	www.caasf.org
13.	P-1741	WIU - Division of Student Services	student.services.wiu.edu
14.	P-41	PSU Affirmative Action	www.psu.edu/dept/aaoffice
15.	P-904	The United States Department of Labor Home Page, Secretary of Lab	www.dol.gov

Table 5.11: *Top 15 results for query "affirmative action", Personalized Randomized HITS* ($\alpha = 0.85$)

algorithm. In almost all the 34 queries, randomized HITS produce a result quite similar to that of InDegree algorithm.

Now let's observe the effects of Personalization on Randomized HITS, by implementing the equations (4.61) and (4.62). Hence this way we can include personalization vector $\vec{v}$, which could be exploited to interleave the outcomes.

For example, consider pages $P-2$ and $P-3$ in Table 5.9, for query "affirmative action". To boost these two pages, and bring them in the top two positions we must appropriately alter the values in personalization vector $\vec{v}$. Notice that α here refers to teleportation probability. In this example setting $\alpha = 0.85$, which means 85% of time the random surfer teleports or jumps to new destination based on vector $\vec{v}$, and 15% of time follows the hyperlinks in forward and backward directions based on the $\mathbb{I}$ and $\mathbb{O}$ operations respectively. With these settings we have the personalized outcomes of randomized HITS shown in Table 5.11. Clearly the pages $P-2$ and $P-3$ now appear in $top-2$ positions.

We haven't observed Randomized HITS in relation to the perturbations. We therefore rely on the findings from [Ng01b] in case of stability, but personalization in Randomized HITS performs as theoretically predicted in Section 4.4.4.

Summary and Reflection on Personalization

The results of the experiments due to personalization portray very broad and general implications of personalization in query-dependent LAR algorithms. It requires a further in depth study, independently conducted with focus on personalization. The active integration of personalization information in the link matrices can be quite a fascinating extension. For example, by using the entries in the link matrix as weights representing the personalization information, possibly utilize the user interaction or usage information from logs in the link matrix to personalize the results based on user behaviour. In our study we could only manage to have an overview of these concepts, and a quite cursory experimental analysis.

There is certainly more to personalization than just boosting a couple of pages. As discussed in Section 4.4, we could employ personalization for other purposes such as anti-spamming and having a control on the set of relevant outcomes. The experiments that we conducted in personalization are aimed to see whether we can control the outcomes to an extent. And we found that it sufficiently possible to use personalization in the query-dependent algorithms without any major barriers. While exploring personalization we have also talked about various capabilities and similarities of different algorithms.

Of course there could more to be done with the ideas that are explored empirically and theoretically in this chapter. In the next chapter we will present the possible extensions and future work out of this study. We will also discuss the broad implications of this study.

Part III

Conclusions, Recommendations & Future Study

Chapter 6

Conclusions

IR is a very wide area; spanning all types and kinds of data and information retrieval. From its broad and wide range applications in databases and image processing, to more specific and active applications in multimedia retrieval and search engines, IR research has been around from decades to mesmerize these key areas. All sorts of dynamic and core concepts in mathematics, informatics and other relevant fields have been applied in IR research to make a difference.

Search based on pure *content analysis* of documents in the huge network of web was not enough. The reasons could be computational limitations on how to process such a huge amount of diverse and chaotic content, but also because of the innate limitations in content analysis; its incapability to provide good match to the user interests. Therefore there was a dire need to take the next step towards better understanding the match between the users' interests and retrieval processes. Somehow extract or mimic the user behavioural information and use that as a valuable source of information for tuning and improving retrieval. This was the stimulus which invoked the efforts towards the active use of the *latent information* like 'citation structure'.

Link Analysis Ranking was the next step from just content-analyses. It involves analyses and understanding of a very huge and jumbled network(s) of documents. From such a huge and massive network extracting useful information is quite a challenging and difficult task. The challenge is not just because of size of network, but also because of its diversity and unpredictability. The huge network(s) of documents hence forms the core of link analysis ranking.

Not only the challenge is due to huge network but also the interests of the users of IR have become very unpredictable. The actual challenge therefore is not just to understand or classify the huge network of documents but also to understand users' changing needs and provide them a good match from highly dynamic sources of information.

In this study we had extensively explored IR problem from a general perspective and particularly went into certain depth on the topic of Link Analysis Ranking. We have discussed different perspectives in several contexts, which subsequently formed a strong basis for this research.

The goals from the general exploration of IR models were to have a strong standing and background of the problem in IR. We have explored wide varieties of novel approaches towards IR, for example, the use of Singular Value Decomposition (SVD) in latent semantic indexing (LSI) and HITS, the idea neural networks and inference networks. SVD, LSI and inference networks therefore formed a realistic motivation towards the focus of our study the Link Analysis Ranking.

In general, the task of digesting, conceptualizing and reasoning from a wide-range plethora of ideas and

concepts within IR is a difficult task in itself. After acquiring background knowledge in IR, figuring out any peculiarity, in the form of improvements of old ideas or discovery of new ideas is yet another challenge. Overall the topic of IR is a very difficult topic to deal with, but at the same very interesting, dynamic and rewarding topic as well.

6.1 Objectives of the study

The objectives of the study primarily were to explore and evaluate different models in IR particularly with a focus on relevancy ranking strategies. The following objectives were in mind when we started the study:

- To study document ranking strategies. To have an overview and understanding of the different concepts and ideas available in IR and their mathematical and logical fundaments.
- To come up with some radical improvements and evaluation of the existing models or propose some new strategy or model.
- To experimentally evaluate those models, verify the theoretical implications with empirical manifestations.

In this section we will take each of the objectives and relate it to the outcomes of this study, to see if the study outcomes are in line with the objectives.

6.1.1 Document ranking strategies

From the start of the study we kept our scope quite wide and flexible. We have started with the general formulation of the IR process in order to gather bits and pieces to be able to perceive the whole picture. By doing that it was intended to explore the whole spectrum of relevancy in IR; from classical content-analysis based models to vector based models, and probabilistic models to inference and neural network based models and ultimately to the hyperlinked and citation based models. The objective was to relate or classify them in a way to generalize the concepts in order to reuse or innovate them in some other situations.

We came across a lot of original and provocative models which used central concepts from mathematics and informatics. A lot of different mathematical notions are applied in a variety of situations. There is an extensive application of Boolean algebra, vector algebra, linear algebra, probability theory, artificial intelligence and other interesting areas of knowledge.

In link analysis ranking there is an extensive use of the core and theoretic notions of linear algebra. The use of Markov chain theory, Singular Value Decomposition, Optimization theory and sophisticated techniques from graph theory are few of the most used and highly talked about concepts in LAR. We have explored different ideas to certain comprehendible levels, which enabled to propose some fruitful contributions in our limited study.

6.1.2 Improvements and Contributions

After having a formal background of IR in general and link analysis ranking in particular, we then had to perform closer examinations of the models. During the analysis and evaluation we have discussed strengths, weaknesses and possible prospects of the different approaches. This way we were able to explore the implicit properties of numerous algorithms and exploit those properties to enhance their strengths and trim down their weaknesses.

Within the link analysis ranking we have primarily focused on query-dependent approaches for relevancy ranking. Query-dependent approaches usually have a direct consequence on the interaction time between user and system. Therefore it is highly appreciated to optimize or accelerate certain processes in the system in order to prevent any kind of delays or overheads at query-time. Our major contribution therefore is the improvements primarily in the convergence behaviour of the query-dependent algorithms. Speedup in convergence by reducing the number of iterations directly improves the interaction time.

Extrapolations are simple and unique techniques that require little additional infrastructure that needs to be incorporated in the existing LAR algorithms. We have distinctively applied extrapolation techniques to query-dependent LAR algorithms such as HITS and SALSA. We have manipulated the parameters extensively and extract a very novel performance gain due to Extrapolation (see Appendix A). By periodic application of extrapolation we have exceedingly enhanced the rate of convergence of the query-dependent LAR algorithms.

In the study by Kamvar et al., [Kam03b] they found an improvement of order 3 at-most due to extrapolation, in PageRank algorithm. By applying extrapolation much more carefully in the query-dependent algorithms, we have had improvements of order in range $(3 - 19)$, see Appendix A.

We have also incorporated *personalization* in the query-dependent algorithms. There is wide range of studies on personalization in IR research communities. We went through a handful of approaches towards personalization. We have formed a generic formulation for personalization, based on the random surfer model of PageRank. We have casted the query-dependent LAR algorithms in the generic framework of personalization. There were few of the query-dependent algorithms which had an intrinsic support for personalization, only it needed to be used in that context (the idea of random jumps).

6.1.3 Experiments and Evaluations

We have experimented with the algorithms that we proposed, as well as the ones that we studied. We have performed an extensive experimental study (see Appendices A and B). The purpose of the widespread experimental evaluation was primarily to observe the effects of the changes and improvements on the overall algorithm, and consequently to examine the manifestation of theoretical predicted properties.

For experiments we used a standard set of queries, based on their representativeness and on how they invoke some of the intrinsic properties and problems in LAR algorithms (e.g., TKC effect in HITS). The dataset corresponding the standard query set were previously used and analyzed by Borodin et al. [Bor05]. For each of the queries we had a neighbourhood graph (the base-set as prescribed by Kleinberg, see Section 3.4), which can be formed from an index structure with hyperlink information. The adjacency matrix corresponding to the base-set can be given as an input to the algorithms.

Setting up the experimental environment is the first major step, and hence not an easy task. The difficulty primarily is due to the availability of a representative set of data, and secondly the formations of an appropriate adjacency matrix that does not contain *non-informative* links and only contains*informative* links. And thirdly a computational capability is also required to test the algorithms in a realistic environment.

We have extensively experimented with the extrapolated and normal algorithms, and compared them with each other to see their responses to the improvements due to extrapolation. We therefore observed their convergence behaviour quite closely (see Appendix A).

The speedup in convergence due to extrapolation came first as a surprise. There were some interesting observations that came out as a result of the experiments. We observed that a careful application of extrapolation can improve convergence inevitably. Therefore, it matters when, where and how many times during iterations

you apply extrapolation step to gain the required acceleration. The importance of extrapolation in accelerating the convergence is hence remarkable. We have also tested hybrid extrapolation technique, where we observed the effect of extrapolation based on different techniques applied togther.

We have experimented with personalization as well. Tested different algorithms and experimented with different personalization vectors. The query-dependent algorithms within the generic framework of personalization have been incorporated and observed. We have boosted a set of documents with help of personalization and observed their effects on the $top - 15$ results. We also felt the need for more sophisticated approaches for personalization, such as, the spread of authority from the boosted document(s) to the related documents, and an active incorporation of usage information in relevancy ranking model.

As a result of experiments we also found some resemblance between few of the algorithms as well. For example we discovered that the *Randomized HITS* resembles the simpler and primitive algorithm, the *InDegree* algorithm.

6.2 Future Work

Despite the already huge amount of available research on IR and relevancy ranking, there are still a lot of interesting areas open for more in depth exploration. We have also identified quite a few of the possible future work during the evaluations and analyses. There is still a lot of work to be done within Link Analysis Ranking. There were numerous things that we touched upon during this study which also require a further in depth research.

Analyses and observations of *Graph Structure of Web* is one of the most primitive and motivational areas in Link Analysis ranking. There is already extensive work done in the area of pure graph analysis for LAR research. But there is still a greater need to do more close analyses of the complicated network of documents, the webgraph. Because the webgraph is changing at an incredible rate, the mere addition or removal of information on web at any moment should be enough to motivate more focus on this area.

The existing researches formulated an overall structure of the web, e.g., the "bow tie" formulation, and the discovery of power law distribution. But a study totally based on the clustering of the graph, in terms of both finding out the *good clusters* and *bad clusters*, or a characterization of the good and bad clusters in webgraph can be very exemplary future work.

Extrapolation techniques can be further extended, by exploring other interesting ideas which could be used to either further improve the existing work or become a new entry. As a result of our study, it is possible to read much more closely the convergence graph from another perspective. For example, use any other convergence measure instead of just L_1 norm. The question is from the convergence behaviours, is it possible to automate *when*, *where* and *how* many times extrapolation should be applied to gain certain acceleration. Also the more active use of *induced ordering* [Hav99] to measure convergence together with extrapolation could be a possible future work. Hybrid approach of extrapolation could also be further observed to formulate a better framework for extrapolation, which could possibly be used for personalization too, apart from just accelerating the convergences.

Users' usage data in the webserver logs is currently used as a heuristic for personalization and other purposes, for example, for upgrading the dictionaries with new keywords. But its rigorous use as weights in the link matrix could be appealing towards more active use of personalization.

IR can be thought of as a flow problem (law of conservation). We can make use of the pattern of traffic (e.g., from webserver logs), and the paths that users implicitly traverse in the graph, which means we need to

exploit information gained from traffic pattern on the web traffic. A possible future work could be to have a traffic model formulation of IR, which has a direct implication on personalization.

Personalization therefore is quiet central and contemporary topic in the research communities in IR. There is a need that one read more in depth about the spreading activation of the authoritative sources which is important in users' perspective. Not just boost the documents in question but also spread or propagate the authoritativeness/importance among the related documents. The notion of authority must therefore be shifted from system to user.

Use of *structural information* of document can be beneficial in some places but can also be the requirement in another place. Structured information can be thought in realistic sense (i.e., chapters, sections, subsections), but it can also be thought in other settings. A nested structure can be *artificially* created for the documents (structured or un-structured) in the collection, in order to personalize or improve the retrieval process. A possible future work therefore could be to consider documents in nested structured approach.

6.3 Problems and impediments

In order to get a good theoretical background in IR you have to spend quite some time. There is a huge amount of work done in IR which spread across many different and diverse disciplines. The varieties of areas that we have discussed here in this study are just a very few of them. There is so much *width* and *depth* in IR problem space, at times it becomes difficult to just go through already existing work in any particular area. You would never reach to the bottom of any area in IR to have a comprehensive overview. It's better to get contended at certain depth during the study.

For experimental setup and evaluations of the algorithms it's quite hard to find out a sufficiently representative set of data. The data used should be assessed, analyzed and a good representative of the actual environment. This means you would confine your analyses on the representativeness and statistical information of the dataset. And if the dataset is not properly chosen and analyzed, the findings of the study might contradict in practice. And it's usually true that algorithms that perform quite well on the experimental environment might not perform as expected in the realistic setup. Therefore it is increasingly important to choose a dataset quite carefully.

There are some road blockers and impediments on the way, and to overcome those blockers you really need to put some firm efforts. Therefore it's quite necessary to be aware of the complexity of the problem before engaging in a formal study.

6.4 Recommendations and Broader Implications

A good background in the knowledge areas like *Linear Algebra*, *Numerical analysis*, *Combinatorial Optimizations*, *Compiler construction* (especially the lexical analyzer, for understanding the text operations), *finite automaton*, and also knowledge of *Concurrent systems* can be valuable before engaging in formal research in IR. A good background in *Artificial intelligence*, especially in neural networks and expert systems could be a great support for further studies.

This study is just a very broad and exterior overview of the IR models. It is very worthy to come within the spectrum of IR; primarily because the area is quite wide, your interests won't remain unfulfilled in such wide range gamut of possibilities. Introducing new ideas are usually considered as valuable as improving the old ones in IR. There is always a room for new ideas to be used in this spectrum.

The task of studying IR is inevitably difficult but the difficulty also entails within it a wholesome of interests, rewards and desire for exploration. You never feel that you have nothing to do at any instance. Every new article and every new book that you read, you get more motivations and more understandings, and hence a stronger background. Writing this last part of thesis I still have to read some more articles. In the end, it has never been unfulfilled neither un-inspirational to study the Information Retrieval (IR).

Bibliography

[Agi06a] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 19–26, 2006.

[Agi06b] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 3–10, 2006.

[AND84] J. ANDERSON and P. PIROLLI. Spread of activation. *Journal of experimental psychology Learning, memory, and cognition*, vol. 10(4):pp. 791–798, 1984.

[Ara01] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the Web. *ACM Transactions on Internet Technology*, vol. 1(1):pp. 2–43, 2001.

[Ara02] A. Arasu, J. Novak, A. Tomkins, and J. Tomlin. PageRank computation and the structure of the web: Experiments and algorithms. *Proceedings of the Eleventh International World Wide Web Conference, Poster Track*, 2002.

[Bar94] R. Barrett. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*. Society for Industrial Mathematics, 1994.

[Bau01] P. Baudisch. *Dynamic Information Filtering*. GMD-Forschungszentrum Informationstechnik, 2001.

[Bha98] K. Bharat and M. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 104–111, 1998.

[Bia05] M. Bianchini, M. Gori, and F. Scarselli. Inside PageRank. *ACM Transactions on Internet Technology*, vol. 5(1):pp. 92–128, 2005.

[Bor01] A. Borodin, G. Roberts, J. Rosenthal, and P. Tsaparas. Finding authorities and hubs from link structures on the World Wide Web. *Proceedings of the 10th international conference on World Wide Web*, pp. 415–429, 2001.

[Bor05] A. Borodin, G. Roberts, J. Rosenthal, and P. Tsaparas. Link analysis ranking: algorithms, theory, and experiments. *ACM Transactions on Internet Technology*, vol. 5(1):pp. 231–297, 2005.

[Bré99] P. Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, 1999.

[Bri98] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *WWW7 / Computer Networks*, vol. 30(1-7):pp. 107–117, 1998.

[Bri06] M. Brinkmeier. PageRank revisited. *ACM Transactions on Internet Technology (TOIT)*, vol. 6(3):pp. 282–301, 2006.

[Bro97] A. Broder, S. Glassman, M. Manasse, and G. Zweig. Syntactic clustering of the Web. *Computer Networks and ISDN Systems*, vol. 29(8-13):pp. 1157–1166, 1997.

[Bro00] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the Web. *Computer Networks*, vol. 33(1-6):pp. 309–320, 2000.

[Bry06] K. Bryan and T. Leise. The $25,000,000,000 Eigenvector: The Linear Algebra behind Google. *SIAM Review*, vol. 48(3):pp. 569–81, 2006.

[BY99] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*. Addison-Wesley Harlow, England, 1999.

[Car97] S. Carrière and R. Kazman. WebQuery: searching and visualizing the Web through connectivity. *Computer Networks and ISDN Systems*, vol. 29(8-13):pp. 1257–1267, 1997.

[Cha98] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. Automatic resource compilation by analyzing hyperlink structure and associated text. *Computer Networks and ISDN Systems*, vol. 30(1-7):pp. 65–74, 1998.

[Cha99] S. Chakrabarti, B. Dom, S. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Hypersearching the Web. 1999.

[Cha00] H. Chang, D. Cohn, and A. McCallum. Creating customized authority lists. *Proceedings of the Seventeenth International Conference of Machine Learning*, 2000.

[Cha02] S. Chakrabarti, M. Joshi, K. Punera, and D. Pennock. The structure of broad topics on the web. *Proceedings of the eleventh international conference on World Wide Web*, pp. 251–262, 2002.

[Cio88] J. Cioslowski. Why does the Aitken extrapolation often help to attain convergence in self-consistent field calculations? *The Journal of Chemical Physics*, vol. 89:p. 2126, 1988.

[Coh01] D. Cohn and T. Hofmann. The missing link-a probabilistic model of document content and hypertext connectivity. *Advances in Neural Information Processing Systems*, vol. 13:pp. 430–436, 2001.

[com] Models and algorithms for complex networks. `http://www.cs.helsinki.fi/u/tsaparas/MACN2006/index.html`.

[Dod01] M. Dodge and R. Kitchin. *Atlas of cyberspace*. Addison-Wesley New York, 2001.

[Dri99] P. Drineas, A. Frieze, R. Kannan, S. Vempala, and V. Vinay. Clustering in large graphs and matrices. *Proceedings of the tenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 291–299, 1999.

[Far06] A. Farahat, T. LoFaro, J. Miller, G. Rae, and L. Ward. Authority Rankings from HITS, PageRank, and SALSA: Existence, Uniqueness, and Effect of Initialization. *SIAM Journal on Scientific Computing*, vol. 27(4):pp. 1181–1201, 2006.

[fas] The fastforward blog. `http://www.fastforwardblog.com/2006/12/15/interview-john-markus-lervik-ceo-of-fast/`.

[Fur88] G. Furnas, S. Deerwester, S. Dumais, T. Landauer, R. Harshman, L. Streeter, and K. Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 465–480, 1988.

[Gol96] G. Golub and C. Van Loan. *Matrix Computations*. Johns Hopkins University Press, 1996.

[goo] Google, the www search engine. `http://www.google.com/technology/index.html`.

[Hava] T. Haveliwala and S. Kamvar. The second eigenvalue of the Google matrix. *A Stanford University Technical Report http://dbpubs stanford edu*, vol. 8090:pp. 2003–20.

[Havb] T. Haveliwala, S. Kamvar, D. Klein, C. Manning, and G. Golub. Computing PageRank using Power Extrapolation. Tech. rep., Tech. Rep. 2003-45, Stanford University, http://dbpubs. stanford. edu/pub/2003-45, July 2003.

[Hav99] T. Haveliwala et al. Efficient computation of PageRank. *Stanford University, http://dbpubs stanford edu*, vol. 8090:pp. 1998–31, 1999.

[Hav03a] T. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, vol. 15(4):pp. 784–796, 2003.

[Hav03b] T. Haveliwala, S. Kamvar, and G. Jeh. An analytical comparison of approaches to personalizing PageRank. *Preprint, June*, 2003.

[He05] B. He and I. Ounis. A study of the dirichlet priors for term frequency normalisation. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 465–471, 2005.

[Hir00] J. Hirai, S. Raghavan, H. Garcia-Molina, and A. Paepcke. WebBase: a repository of Web pages. *Computer Networks*, vol. 33(1-6):pp. 277–293, 2000.

[Jeh03] G. Jeh and J. Widom. Scaling Personalized Web Search. *Proceedings of the Twelfth International World Wide Web Conference*, 2003.

[Jon] K. Jones, S. Walker, and S. Robertson. A probabilistic model of information retrieval: development and status. Tech. rep., A Technical Report of the Computer Laboratory, University of Cambridge, UK, 1998.

[Kai98] H. Kaindl, S. Kramer, and L. Afonso. Combining structure search and content search for the World-Wide Web. *Proceedings of the ninth ACM conference on Hypertext and hypermedia: links, objects, time and space—structure in hypermedia systems: links, objects, time and space—structure in hypermedia systems*, pp. 217–224, 1998.

[Kam03a] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub. Exploiting the block structure of the web for computing pagerank. *Preprint, March*, 2003.

[Kam03b] S. Kamvar, T. Haveliwala, C. Manning, and G. Golub. Extrapolation methods for accelerating PageRank computations. *Proceedings of the 12th international conference on World Wide Web*, pp. 261–270, 2003.

[Kam04] S. Kamvar, T. Haveliwala, and G. Golub. Adaptive methods for the computation of pagerank. *Linear Algebra Appl*, vol. 386:pp. 51–65, 2004.

[Kit98] B. Kitchens. *Symbolic Dynamics: One-sided, Two-sided, and Countable State Markov Shifts*. Springer, 1998.

[Kle99] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, vol. 46(5):pp. 604–632, 1999.

[Lan06] A. Langville and C. Meyer. *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, 2006.

[Lar96] R. Larson. Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace. *Ann Meeting of the American Soc Info Sci*, pp. 71–78, 1996.

[Lay94] D. Lay. *Linear algebra and its applications*. Addison-Wesley Reading, Mass, 1994.

[Lem00] R. Lempel and S. Moran. The stochastic approach for link-structure analysis (SALSA) and the TKC effect. *Computer Networks*, vol. 33(1-6):pp. 387–401, 2000.

[Lu04] Y. Lu, B. Zhang, W. Xi, Z. Chen, Y. Liu, M. Lyu, and W. Ma. The PowerRank web link analysis algorithm. *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters*, pp. 254–255, 2004.

[Mar97] M. Marchiori. The quest for correct information on the Web: hyper search engines. *Computer Networks and ISDN Systems*, vol. 29(8-13):pp. 1225–1235, 1997.

[Met04] D. Metzler and W. Croft. Combining the language model and inference network approaches to retrieval. *Information Processing and Management*, vol. 40(5):pp. 735–750, 2004.

[Mil01] J. Miller, G. Rae, F. Schaefer, L. Ward, T. LoFaro, and A. Farahat. Modifications of Kleinberg's HITS algorithm using matrix exponentiation and web log records. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 444–445, 2001.

[Mot95] R. Motwani and P. Raghavan. *Randomized Algorithms*. Cambridge University Press, 1995.

[Naj07a] M. Najork. Comparing the effectiveness of hits and salsa. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 157–164, 2007.

[Naj07b] M. Najork, H. Zaragoza, and M. Taylor. Hits on the web: how does it compare? *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 471–478, 2007.

[Ng01a] A. Ng, A. Zheng, and M. Jordan. Link analysis, eigenvectors and stability. *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 903–910, 2001.

[Ng01b] A. Ng, A. Zheng, and M. Jordan. Stable algorithms for link analysis. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 258–266, 2001.

[Pag98] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web, 1998.

[Pir96] P. Pirolli, J. Pitkow, and R. Rao. Silk from a sow's ear: extracting usable structures from the Web. *Proceedings of the SIGCHI conference on Human factors in computing systems: common ground*, pp. 118–125, 1996.

[Raf00] D. Rafiei and A. Mendelzon. What is this page known for? Computing Web page reputations. *Computer Networks*, vol. 33(1-6):pp. 823–835, 2000.

[Ric02] M. Richardson and P. Domingos. The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank. *Advances in Neural Information Processing Systems*, vol. 14:pp. 1441–1448, 2002.

[Rob04] S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. *Proceedings of the Thirteenth ACM conference on Information and knowledge management*, pp. 42–49, 2004.

[Sal83] G. Salton, E. Fox, and H. Wu. Extended boolean information retrieval. *Communications of the ACM*, vol. 26(11):pp. 1022–1036, 1983.

[Sin01] A. Singhal. Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin*, vol. 24, 2001.

[Tan02] F. Tanudjaja and L. Mui. Persona: a contextualized and personalized web search. *System Sciences, 2002 HICSS Proceedings of the 35th Annual Hawaii International Conference on*, pp. 1232–1240, 2002.

[Tsa] P. Tsaparas. Link analysis ranking - experiments. `http://www.cs.toronto.edu/~tsap/experiments/thesis/`.

[Tsa04a] P. Tsaparas. *Link Analysis Ranking*. Ph.D. thesis, University of Toronto, 2004.

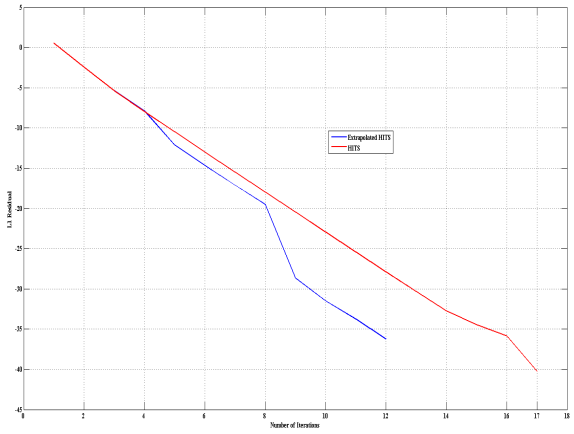
[Tsa04b] P. Tsaparas. Using non-linear dynamical systems for web searching and ranking. *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 59–70, 2004.

[Tur91] H. Turtle and W. B. Croft. Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, vol. 9, 1991.

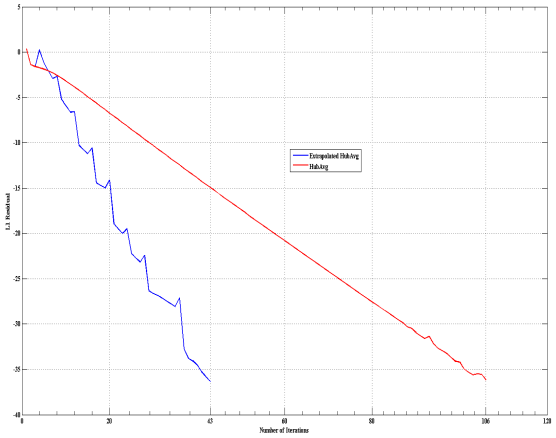
[wik] Wikipedia, the free encyclopedia. `http://www.wikipedia.org`.

Appendix A

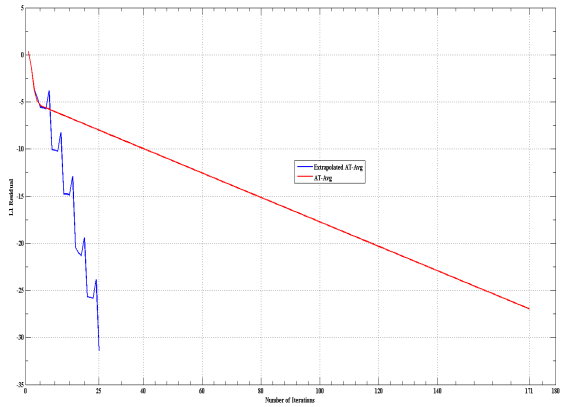
Experiments - Extrapolation



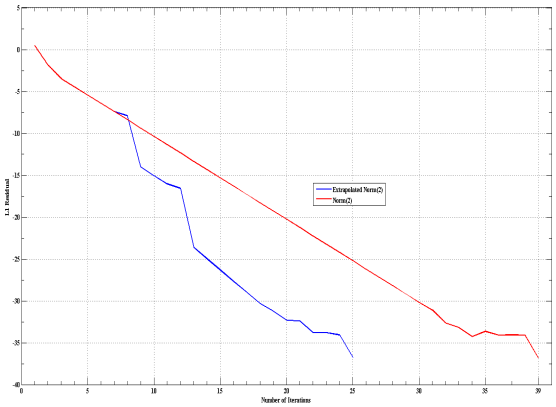
(a) HITS



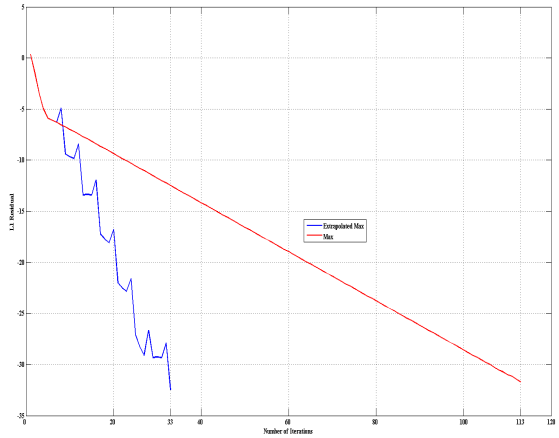
(b) HubAvg



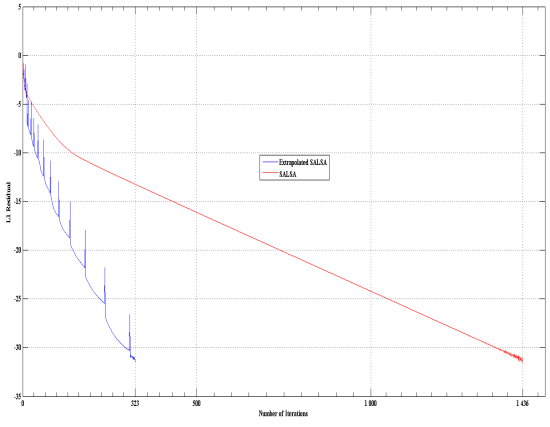
(c) AT-Avg



(d) Norm (2)

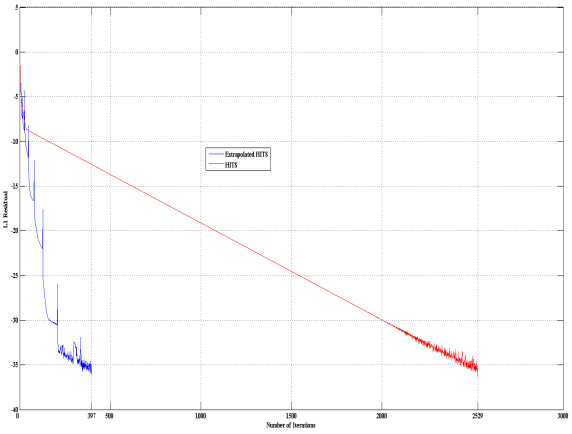


(e) Max

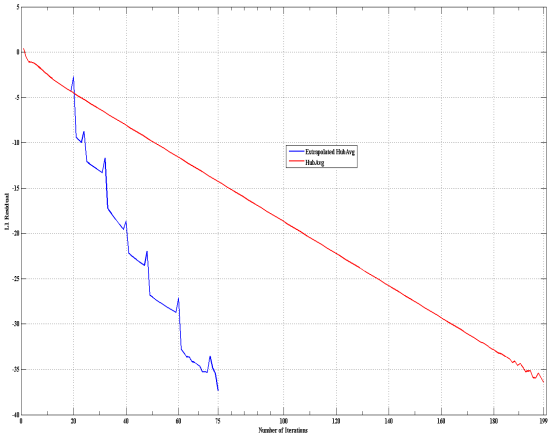


(f) SALSA

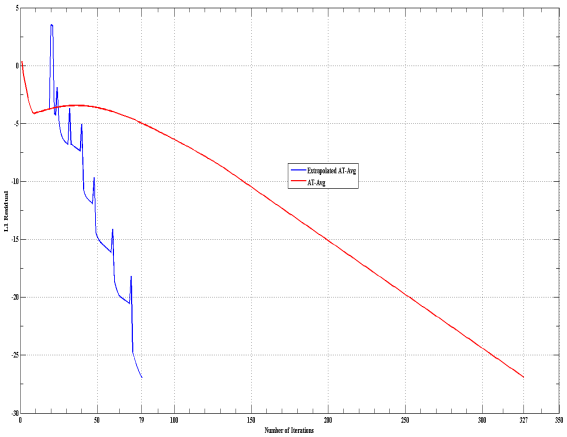
Figure A.1: *Convergence graphs for query "abortion"*



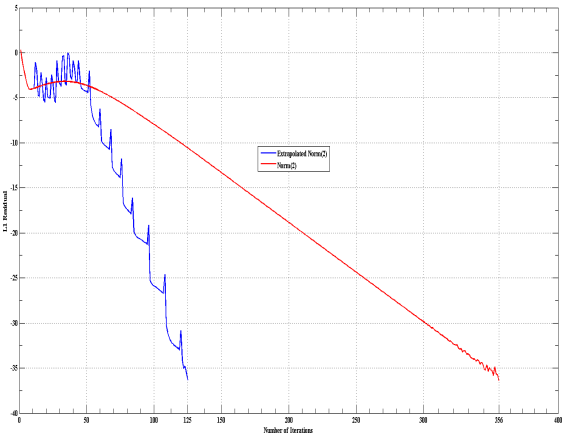
(a) HITS



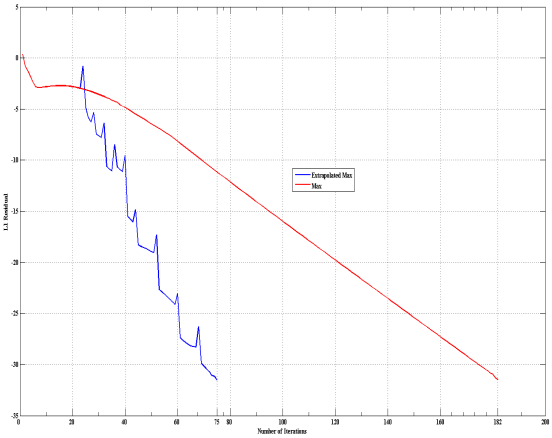
(b) HubAvg



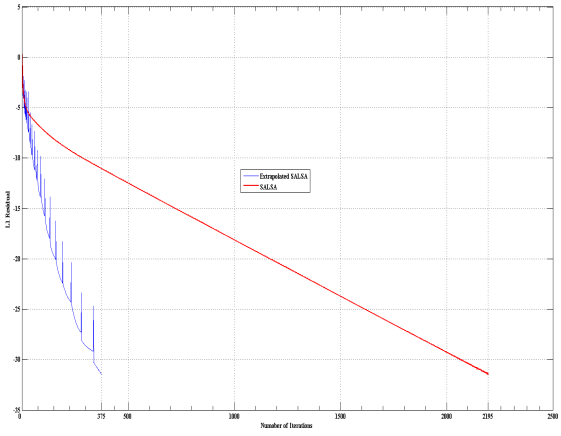
(c) AT-Avg



(d) Norm (2)

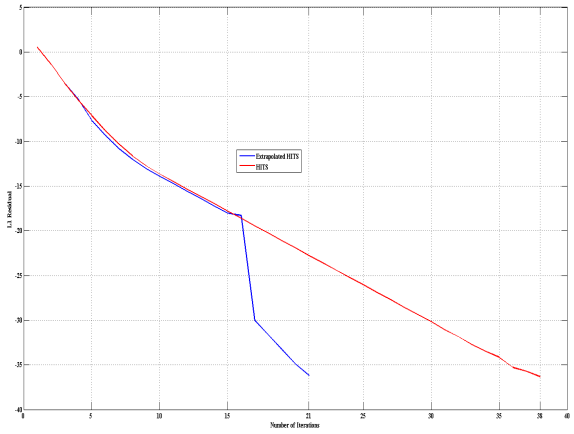


(e) Max

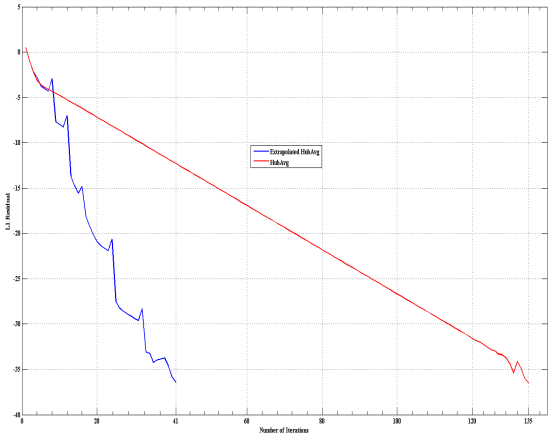


(f) SALSA

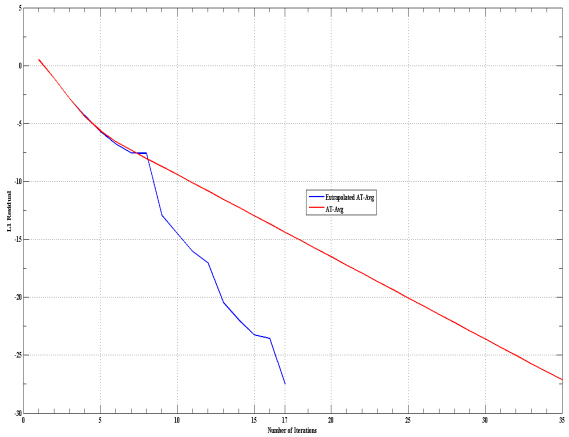
Figure A.2: *Convergence graphs for query "affirmative action"*



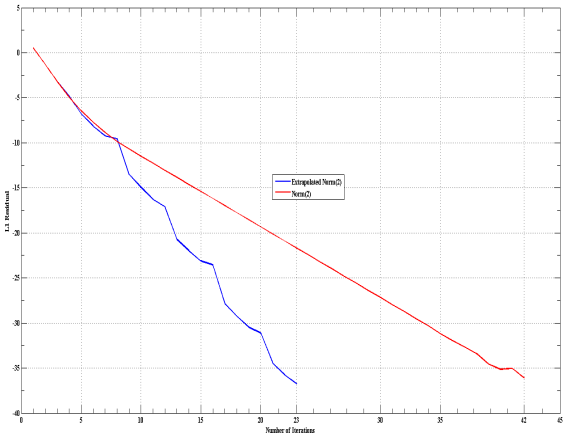
(a) HITS



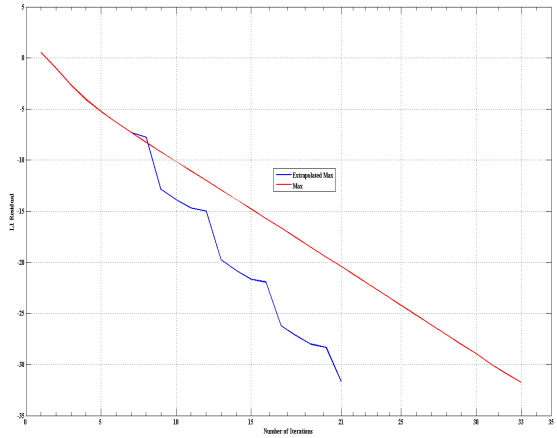
(b) HubAvg



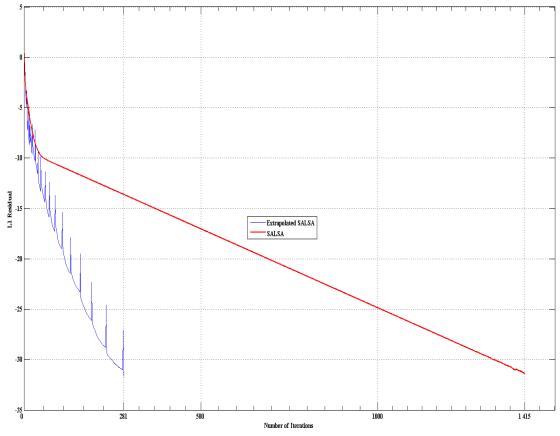
(c) AT-Avg



(d) Norm (2)

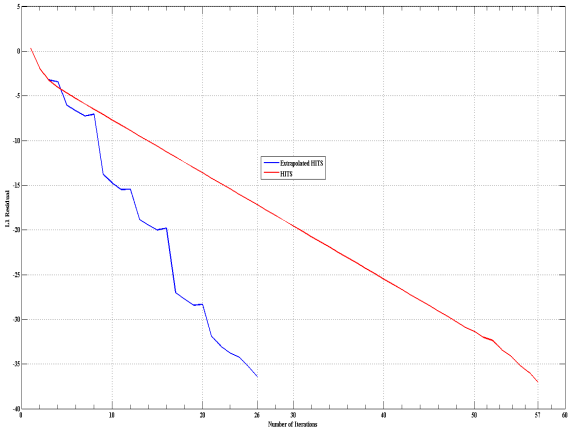


(e) Max

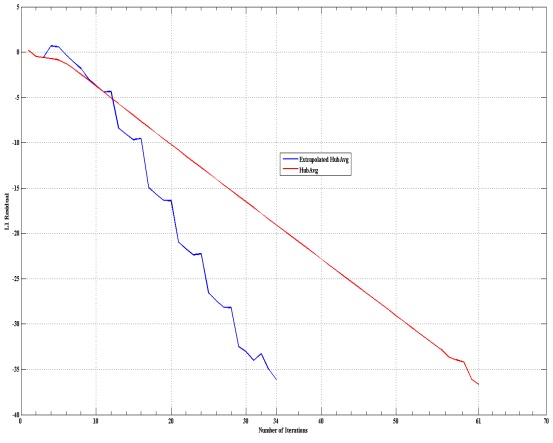


(f) SALSA

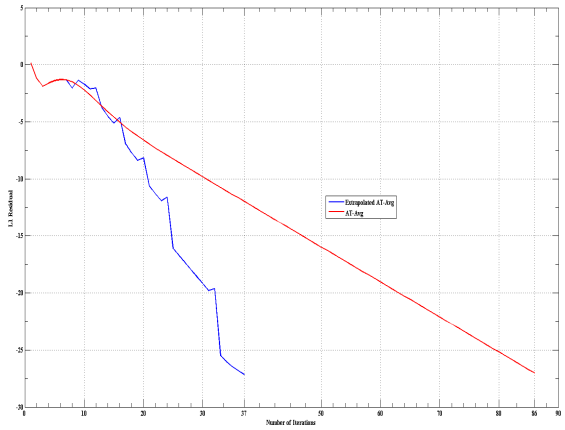
Figure A.3: *Convergence graphs for query "alcohol"*



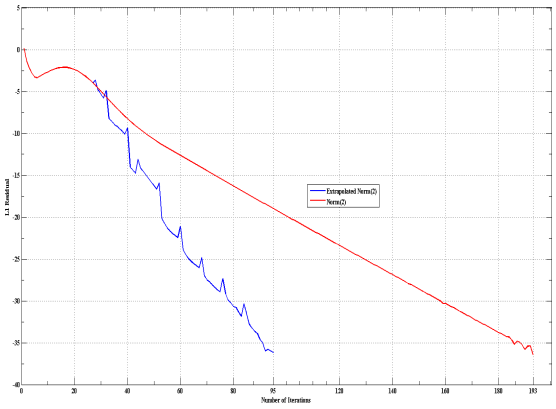
(a) HITS



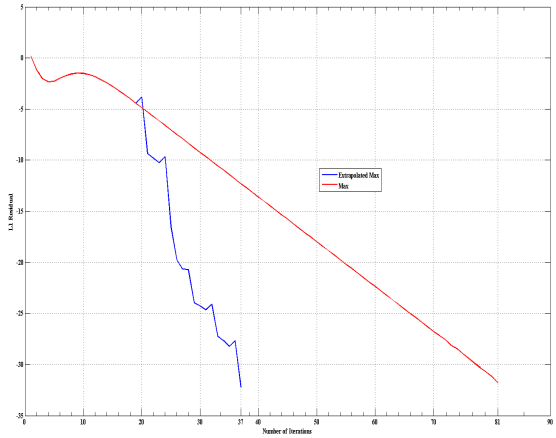
(b) HubAvg



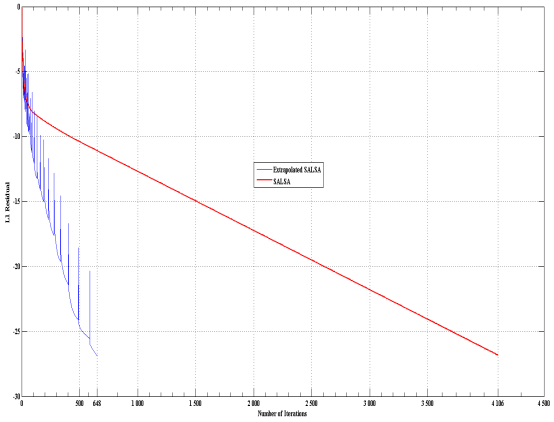
(c) AT-Avg



(d) Norm (2)

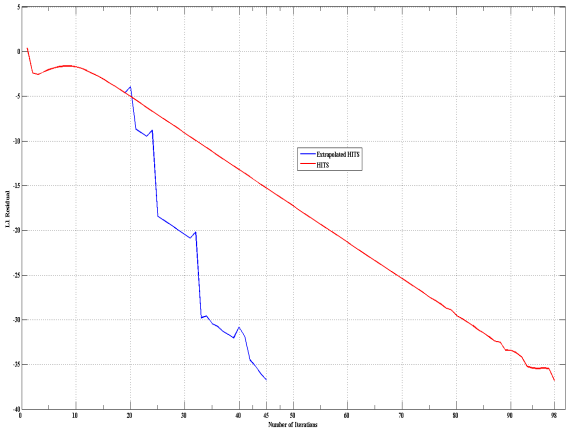


(e) Max

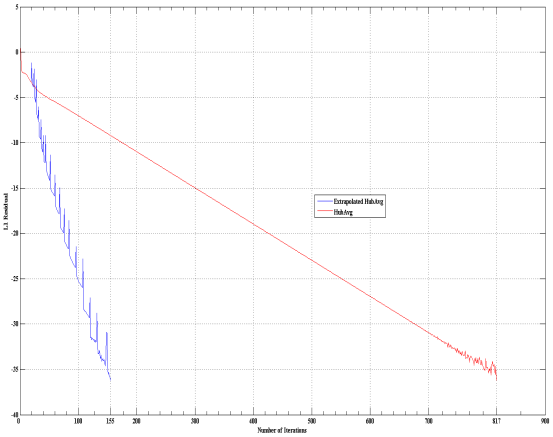


(f) SALSA

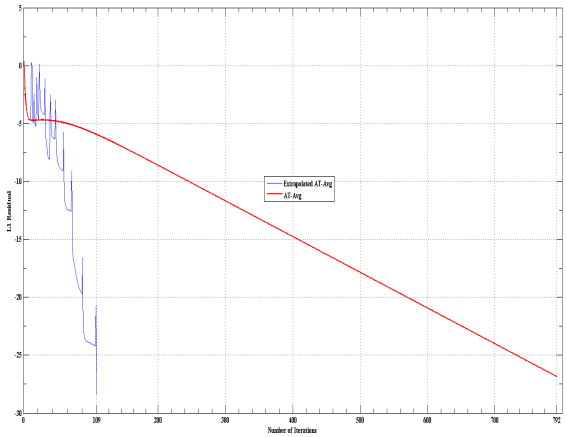
Figure A.4: *Convergence graphs for query "amusement parks"*



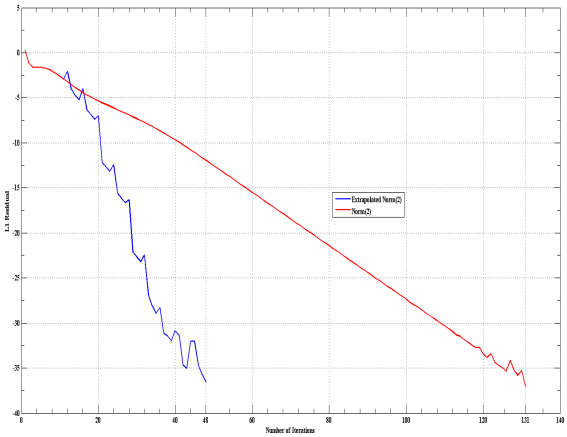
(a) HITS



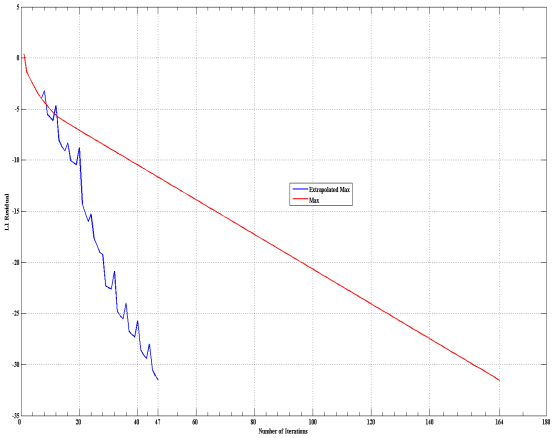
(b) HubAvg



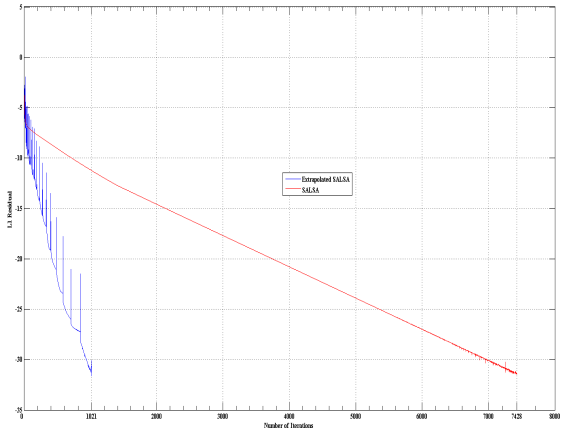
(c) AT-Avg



(d) Norm (2)

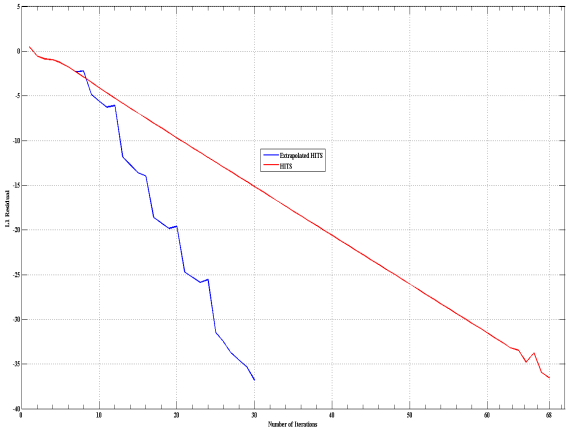


(e) Max

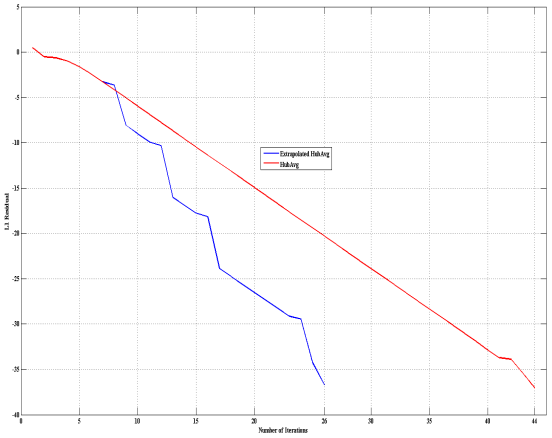


(f) SALSA

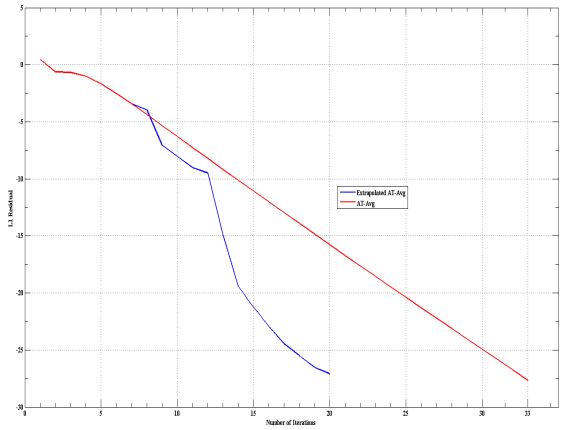
Figure A.5: *Convergence graphs for query "architecture"*



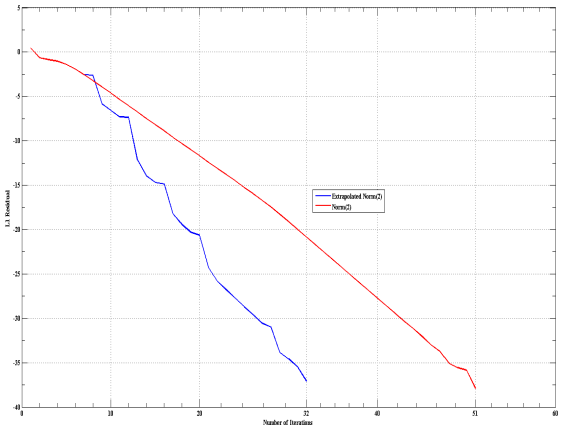
(a) HITS



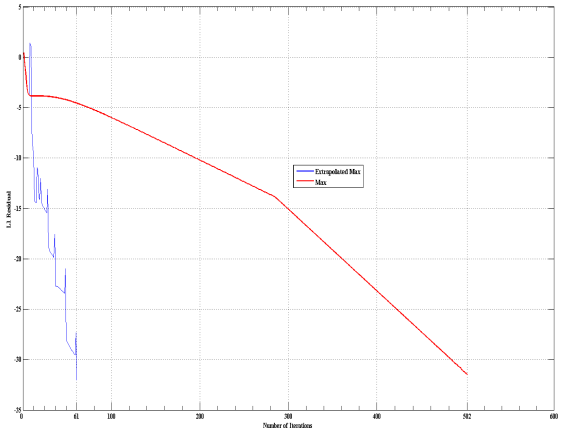
(b) HubAvg



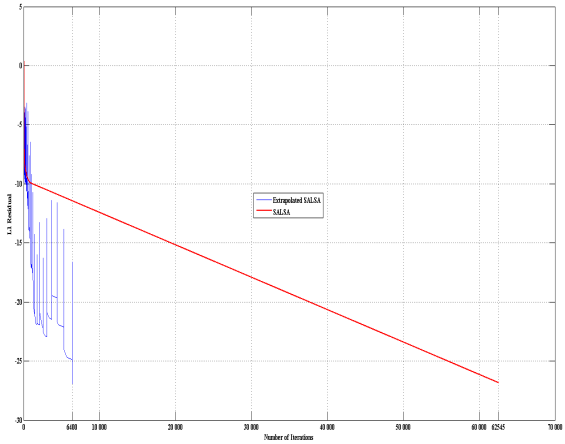
(c) AT-Avg



(d) Norm (2)

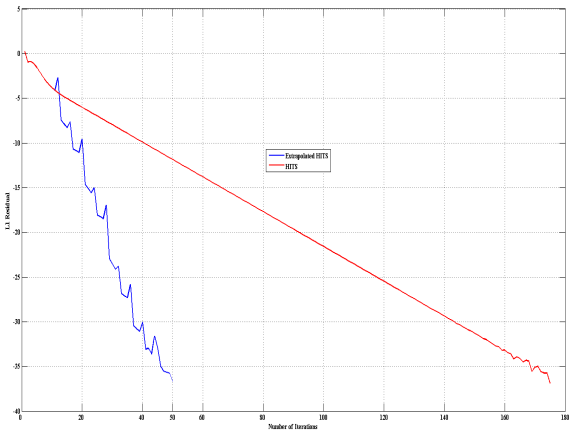


(e) Max

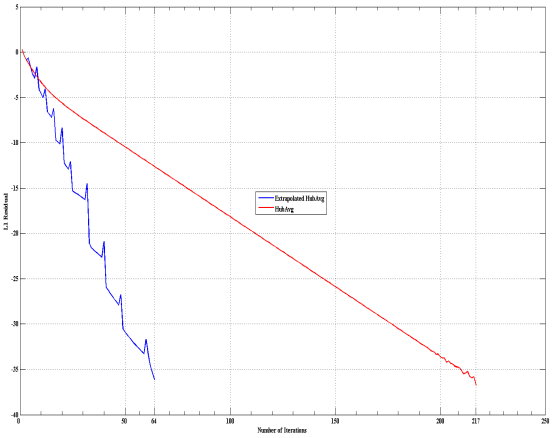


(f) SALSA

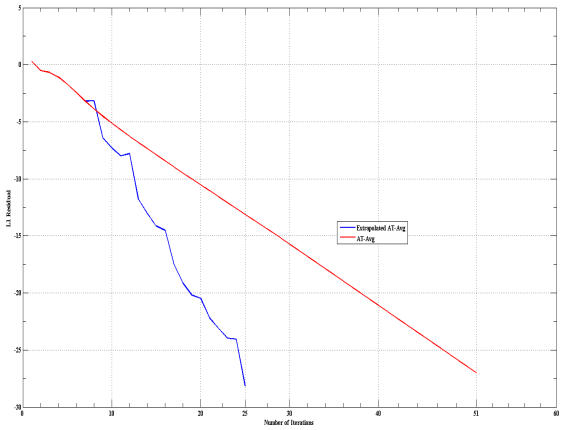
Figure A.6: *Convergence graphs for query "armstrong"*



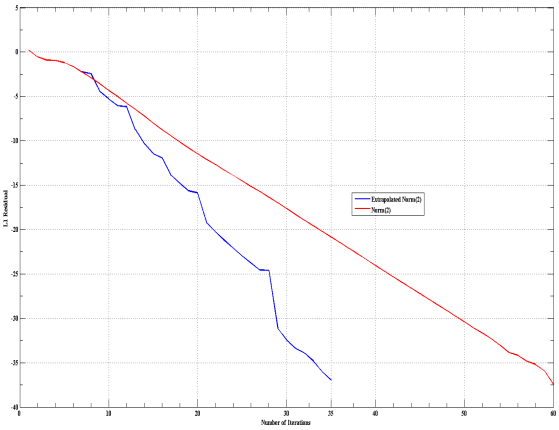
(a) HITS



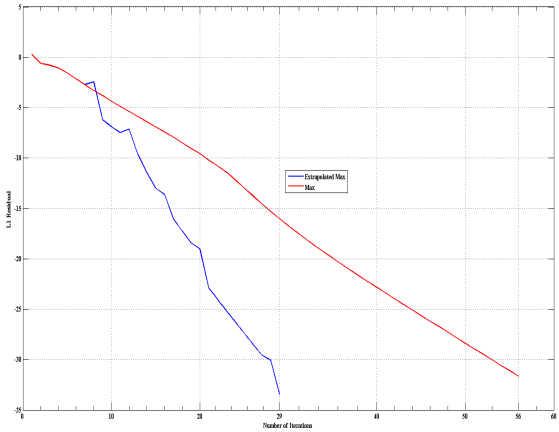
(b) HubAvg



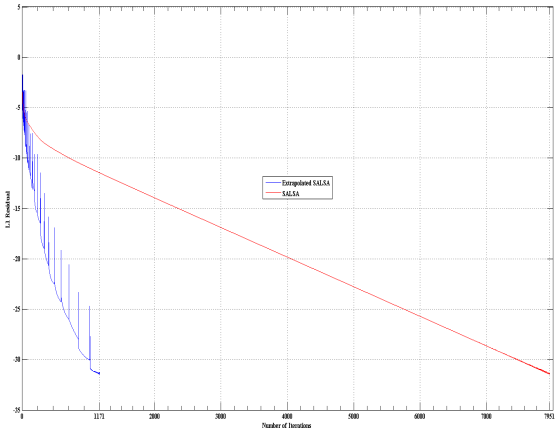
(c) AT-Avg



(d) Norm (2)

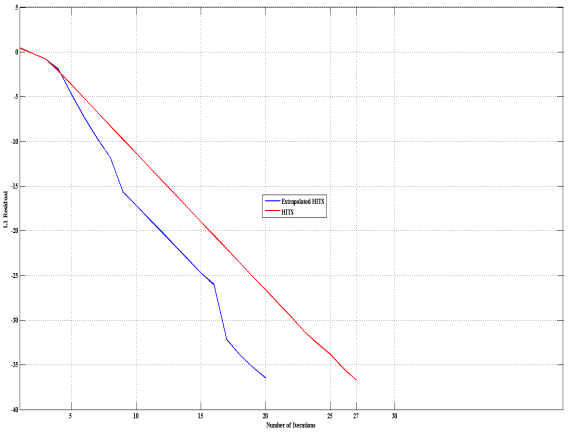


(e) Max

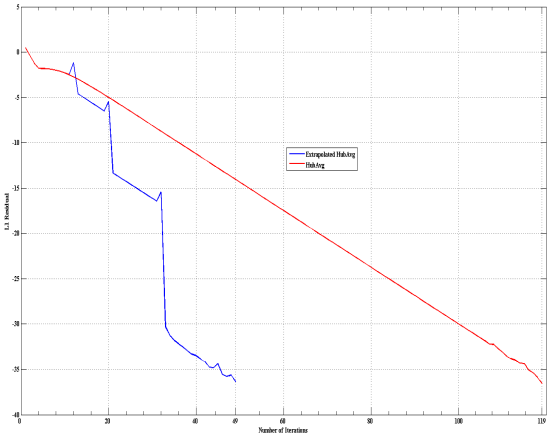


(f) SALSA

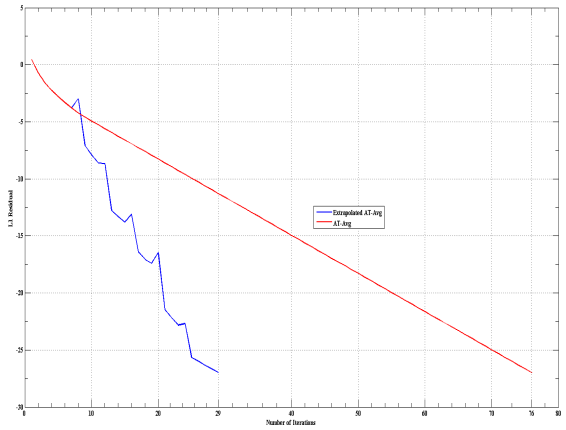
Figure A.7: *Convergence graphs for query "automobile industries"*



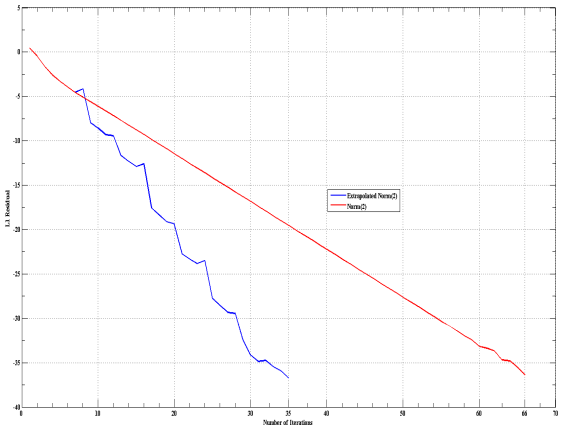
(a) HITS



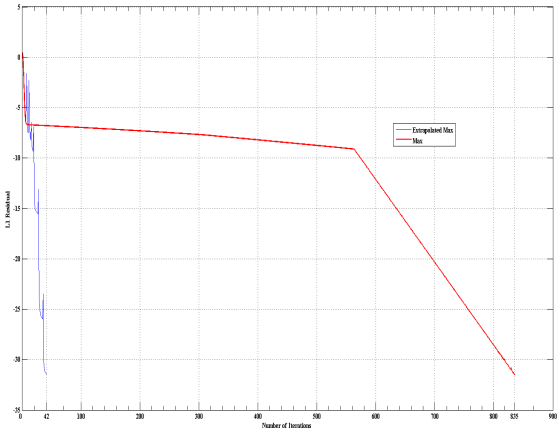
(b) HubAvg



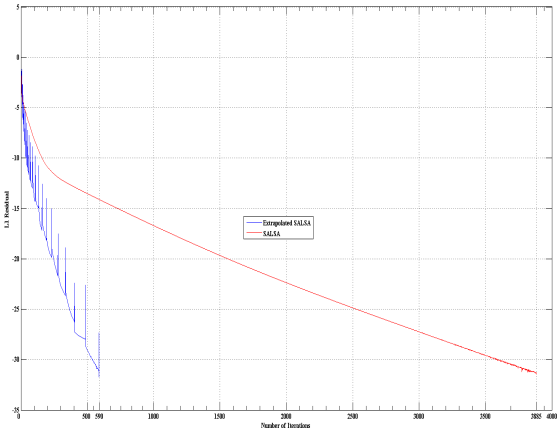
(c) AT-Avg



(d) Norm (2)

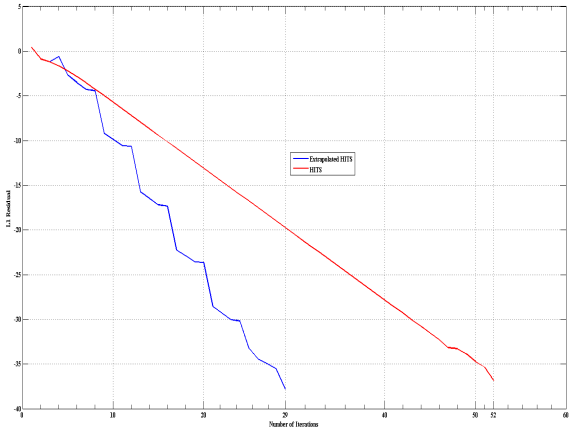


(e) Max

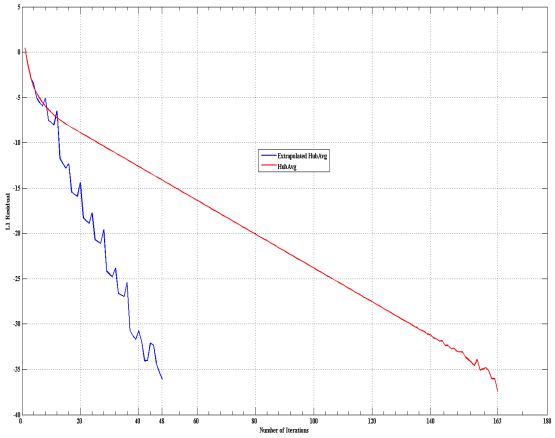


(f) SALSA

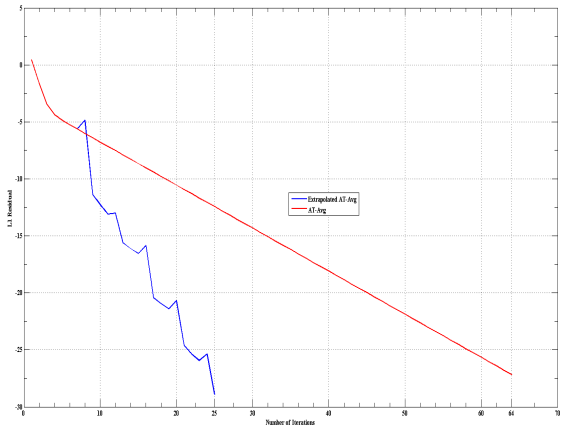
Figure A.8: *Convergence graphs for query "basketball"*



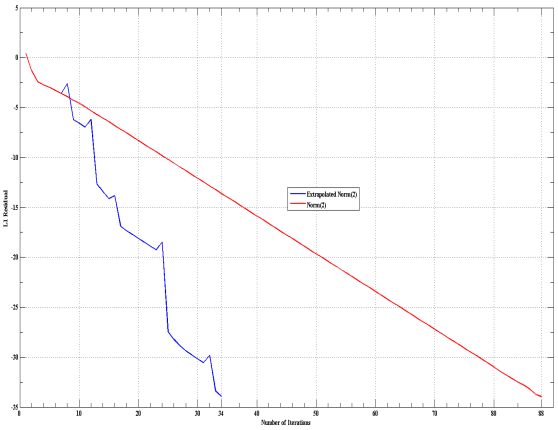
(a) HITS



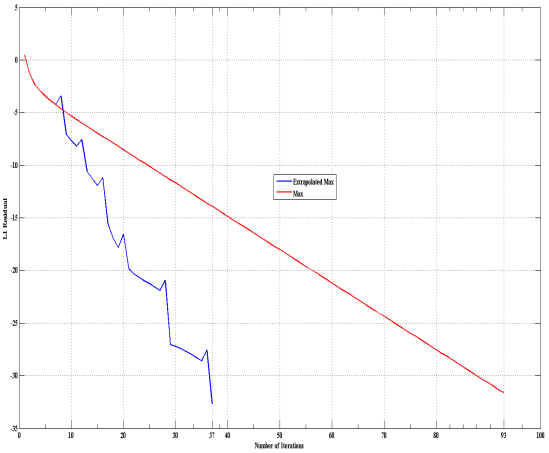
(b) HubAvg



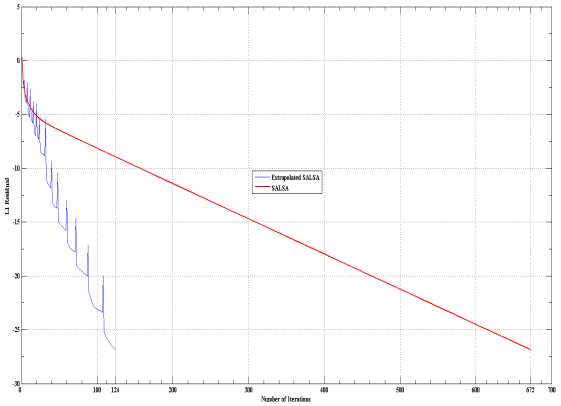
(c) AT-Avg



(d) Norm (2)

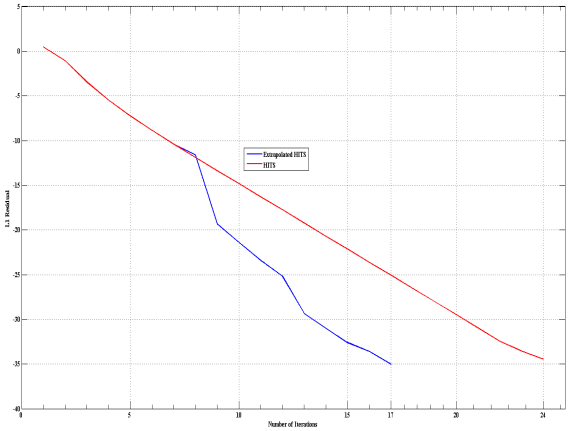


(e) Max

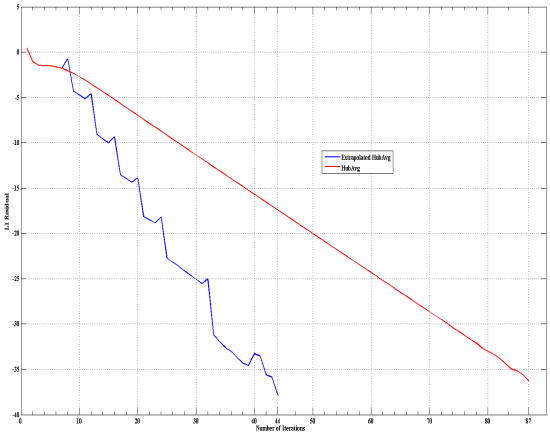


(f) SALSA

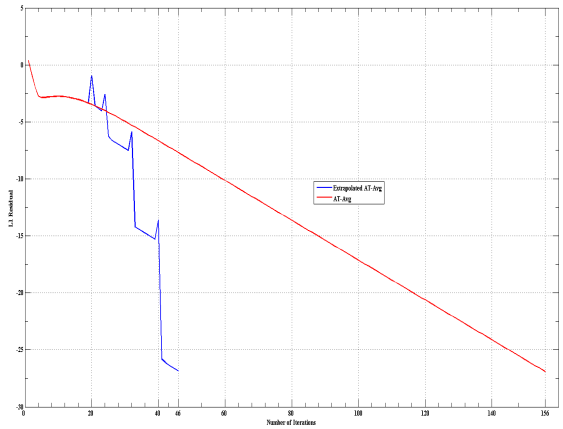
Figure A.9: *Convergence graphs for query "blues"*



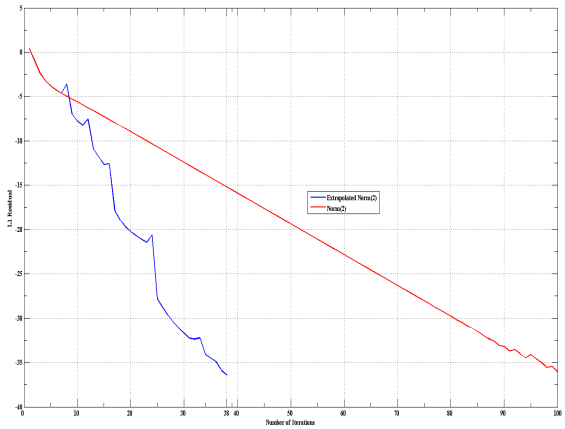
(a) HITS



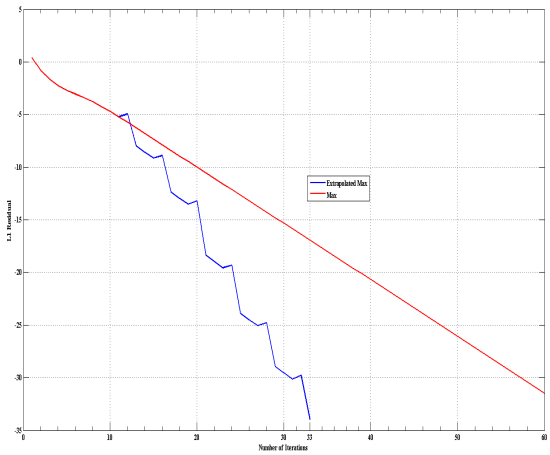
(b) HubAvg



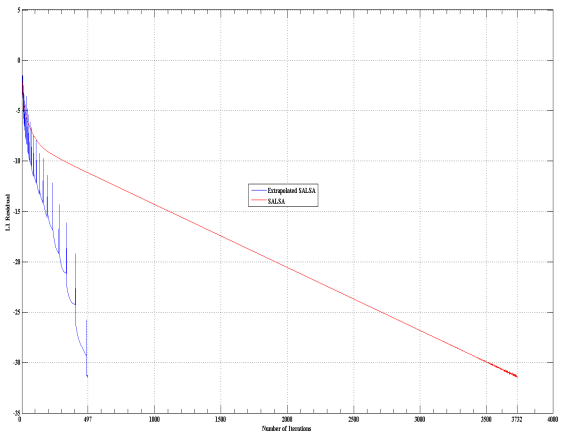
(c) AT-Avg



(d) Norm (2)

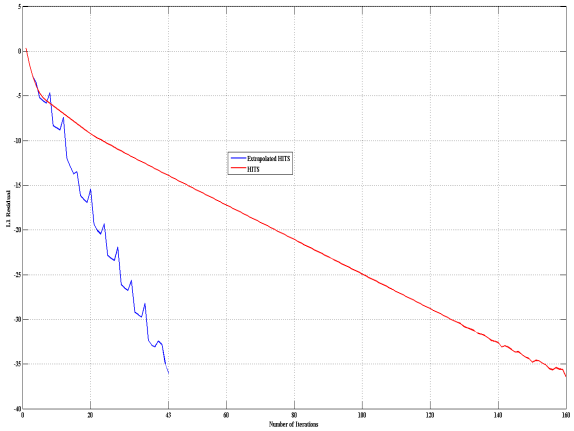


(e) Max

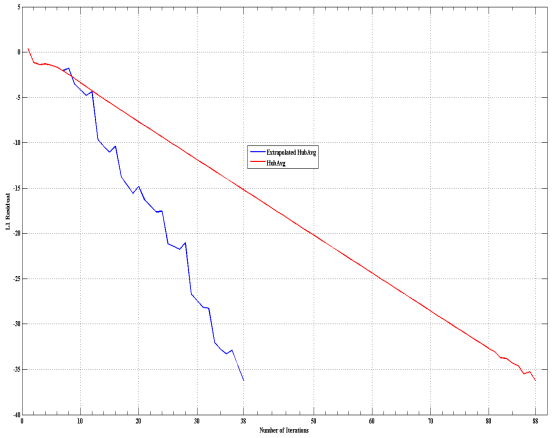


(f) SALSA

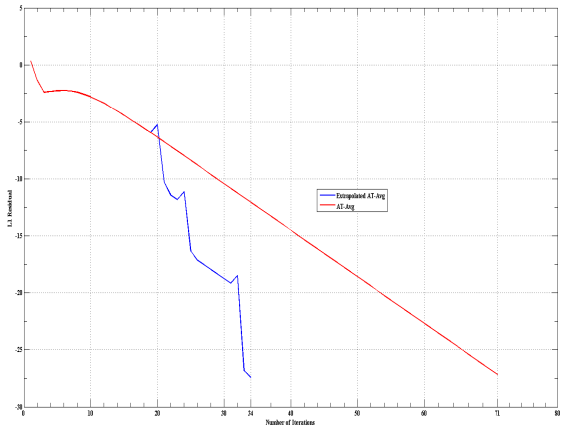
Figure A.10: *Convergence graphs for query "cheese"*



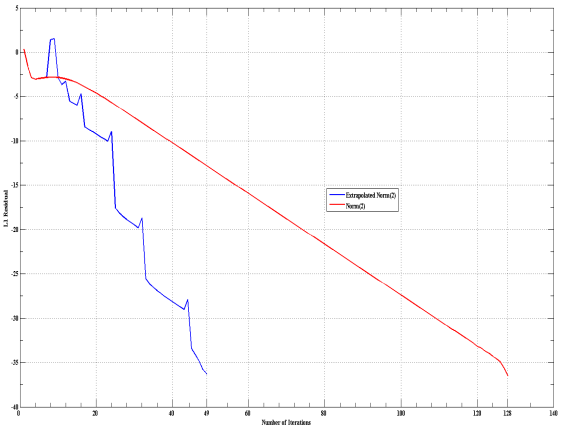
(a) HITS



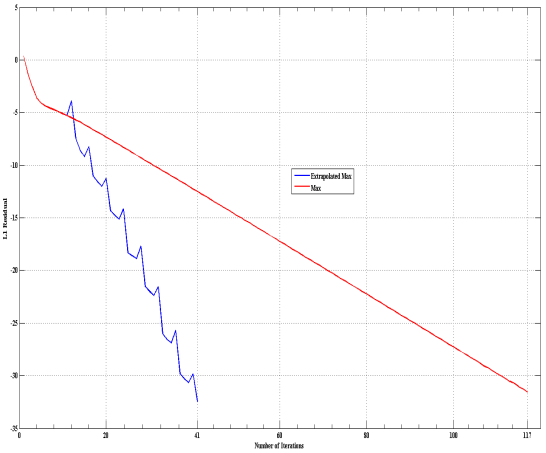
(b) HubAvg



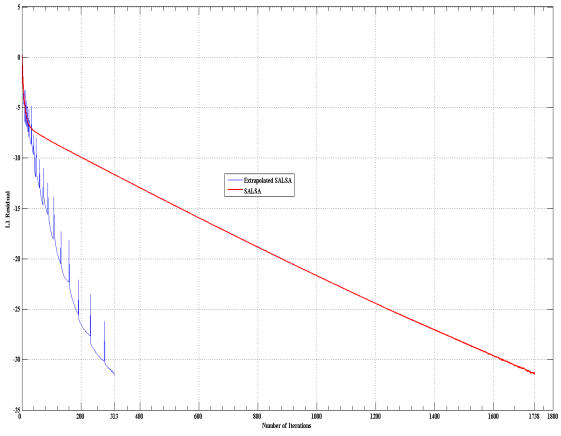
(c) AT-Avg



(d) Norm (2)

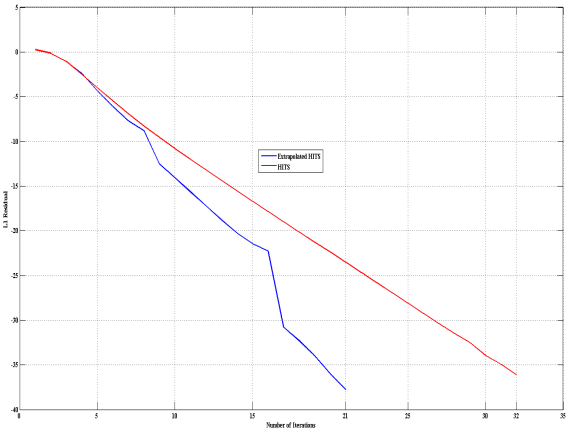


(e) Max

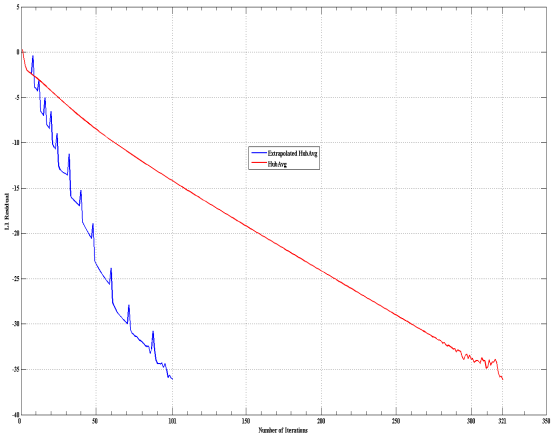


(f) SALSA

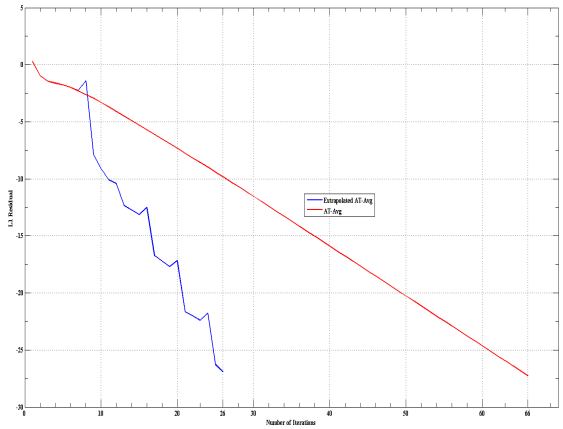
Figure A.11: *Convergence graphs for query "classical guitar"*



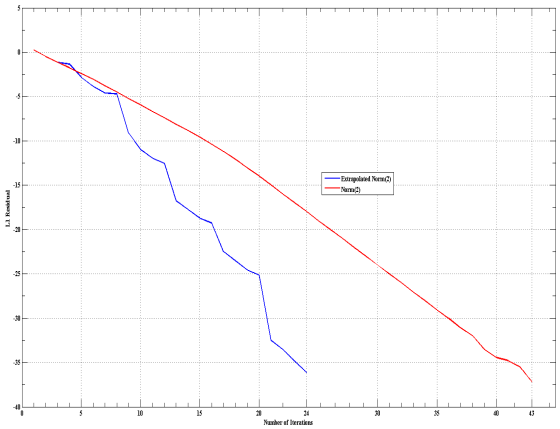
(a) HITS



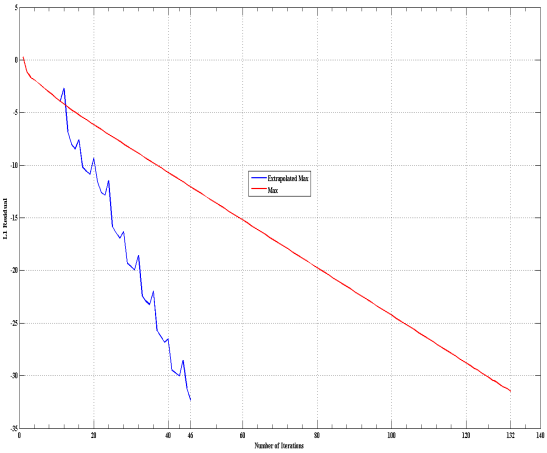
(b) HubAvg



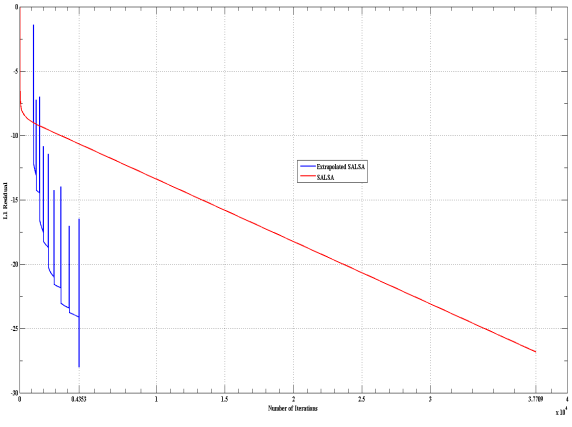
(c) AT-Avg



(d) Norm (2)

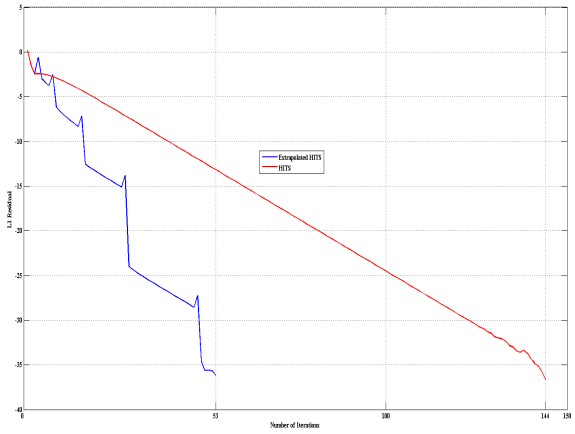


(e) Max

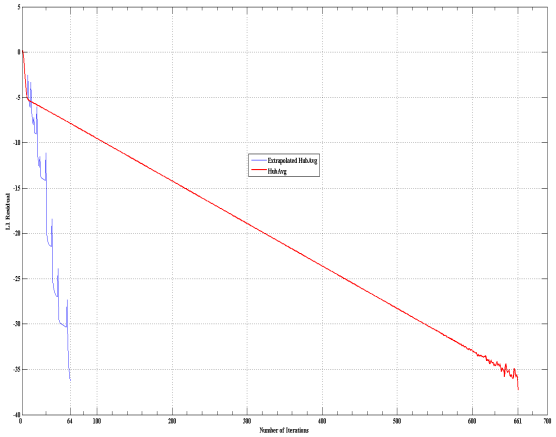


(f) SALSA

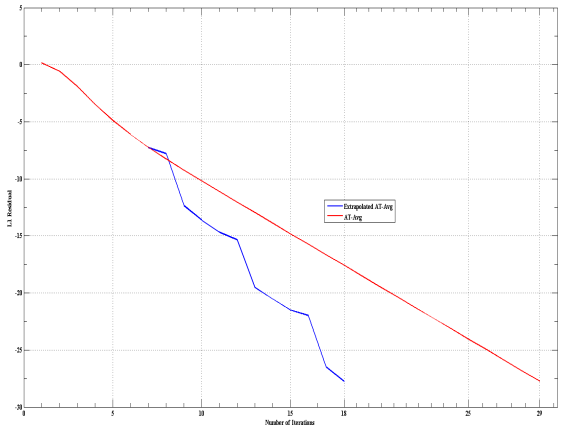
Figure A.12: *Convergence graphs for query "complexity"*



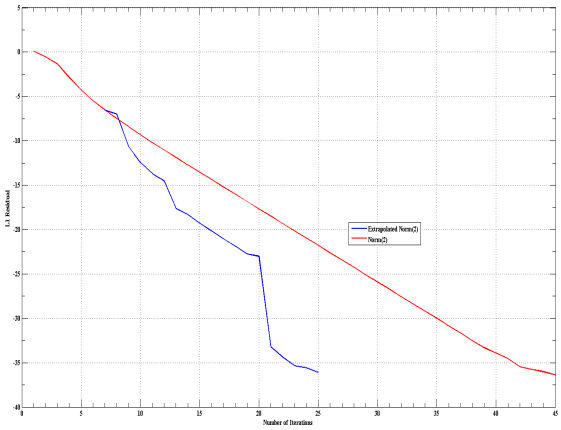
(a) HITS



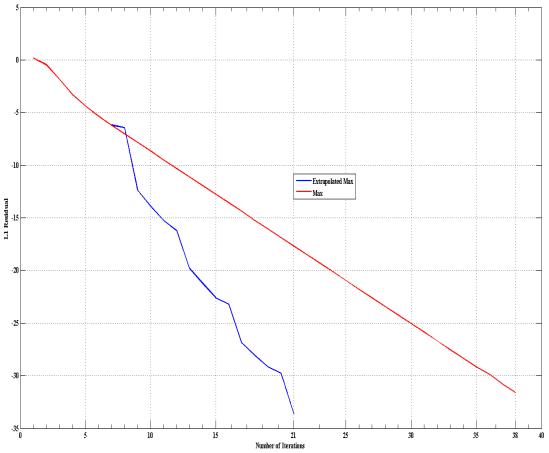
(b) HubAvg



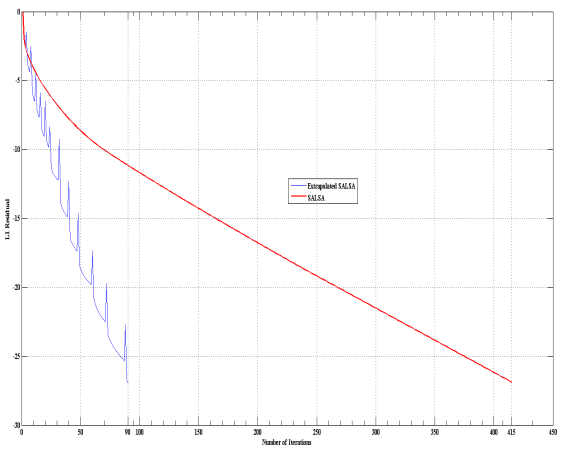
(c) AT-Avg



(d) Norm (2)

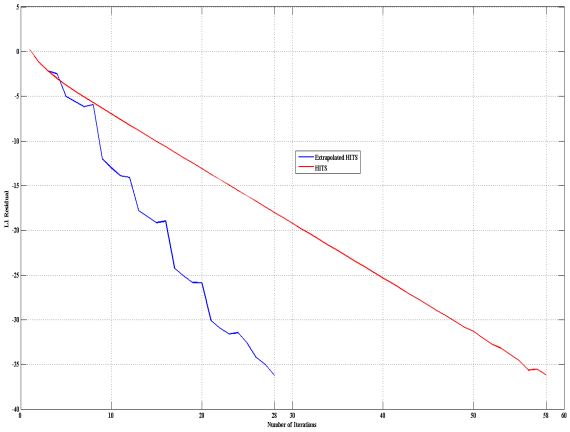


(e) Max

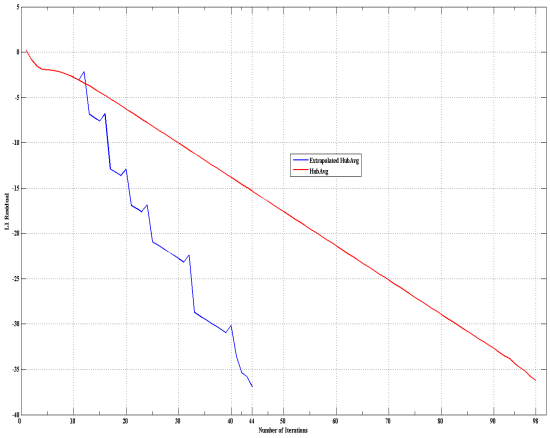


(f) SALSA

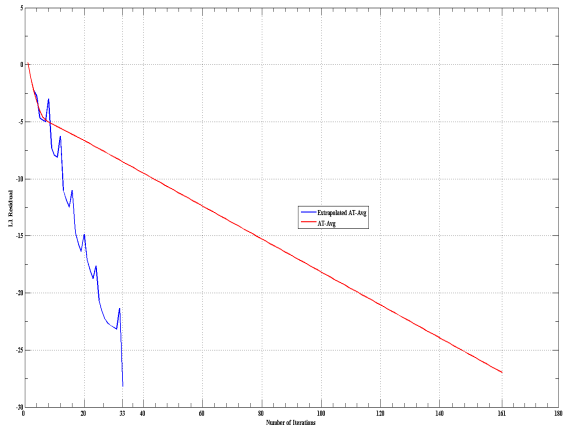
Figure A.13: *Convergence graphs for query "computational complexity"*



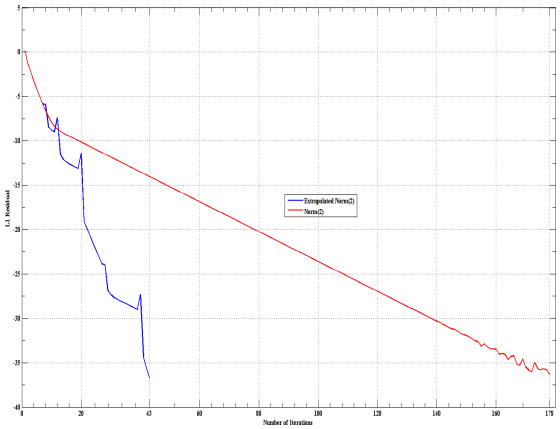
(a) HITS



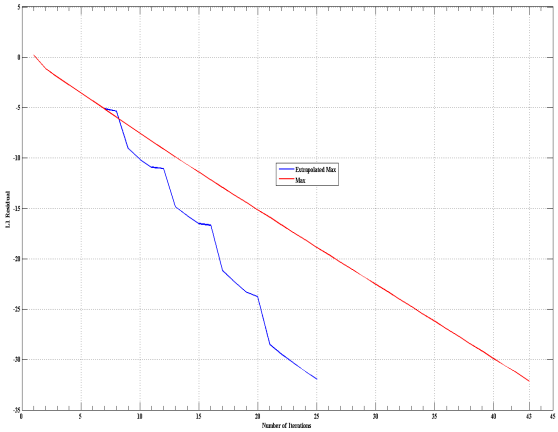
(b) HubAvg



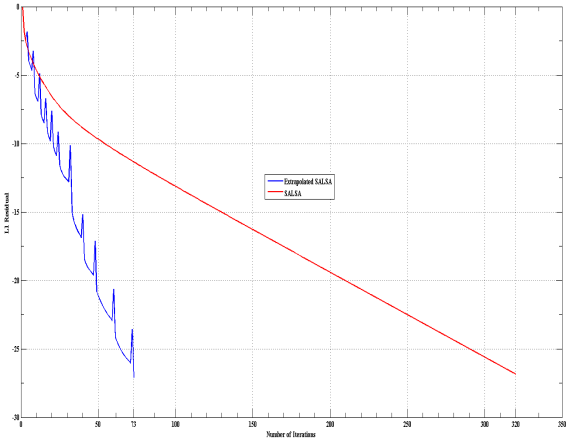
(c) AT-Avg



(d) Norm (2)

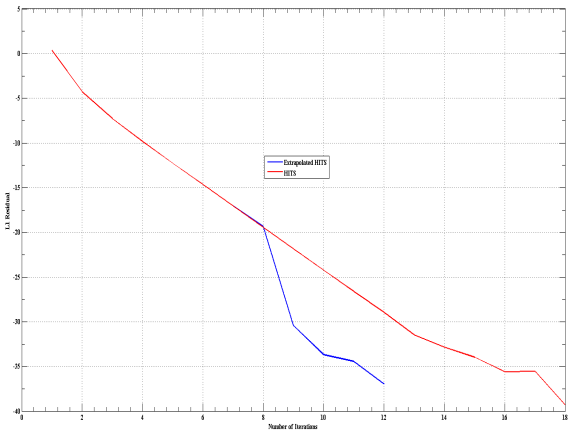


(e) Max

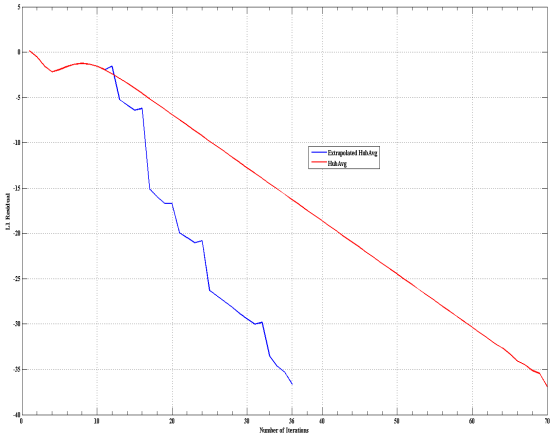


(f) SALSA

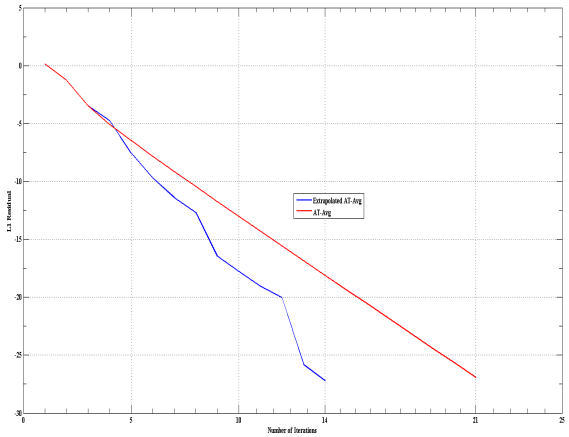
Figure A.14: *Convergence graphs for query "computational geometry"*



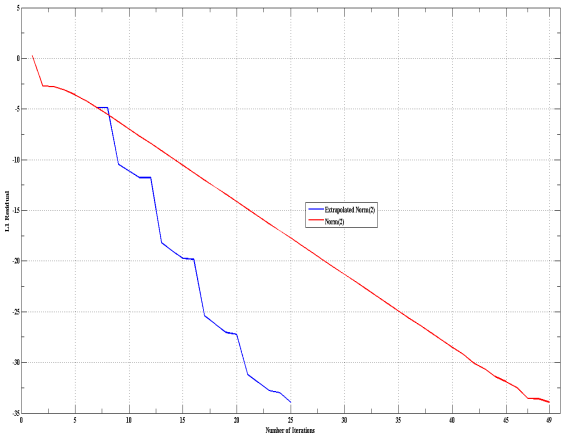
(a) HITS



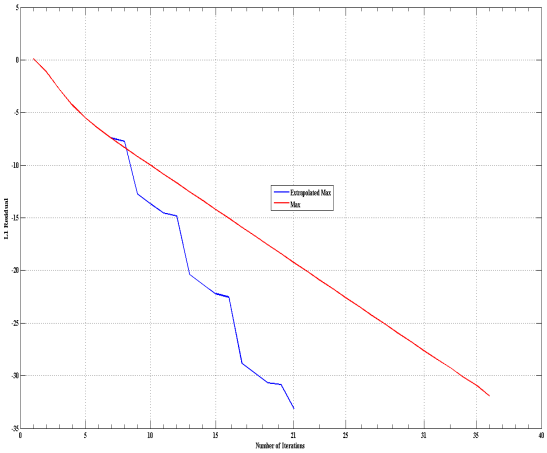
(b) HubAvg



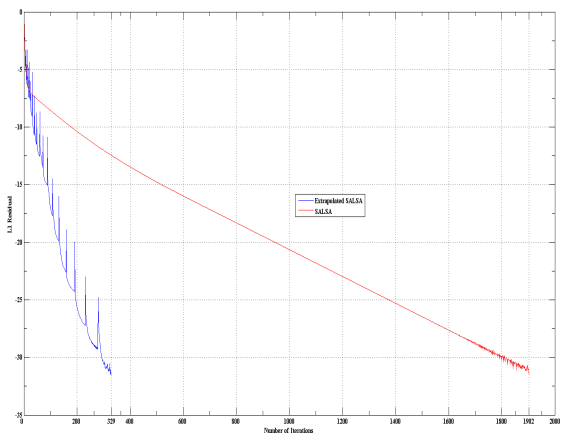
(c) AT-Avg



(d) Norm (2)

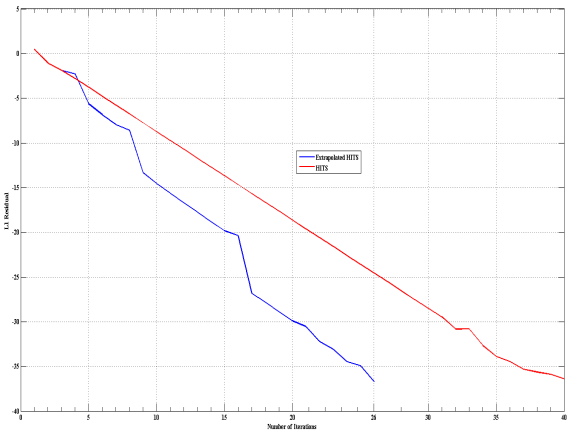


(e) Max

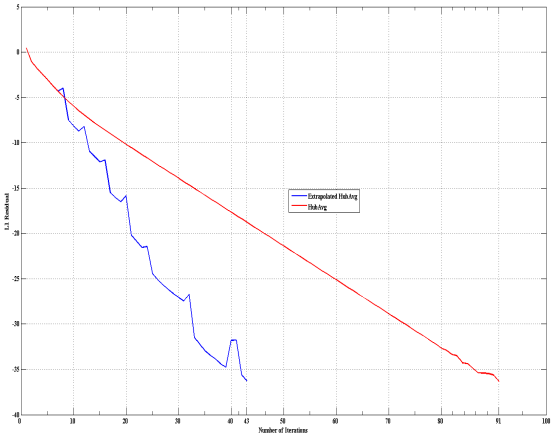


(f) SALSA

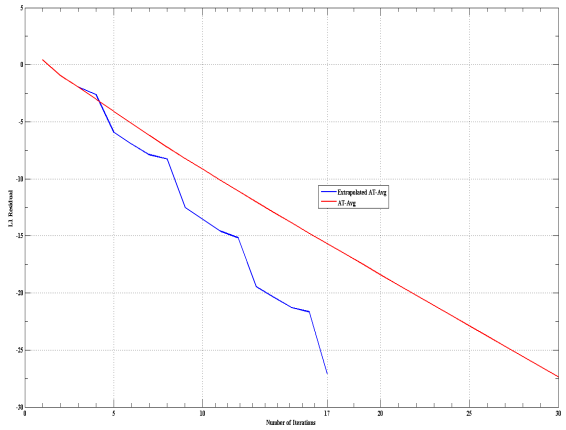
Figure A.15: *Convergence graphs for query "death penalty"*



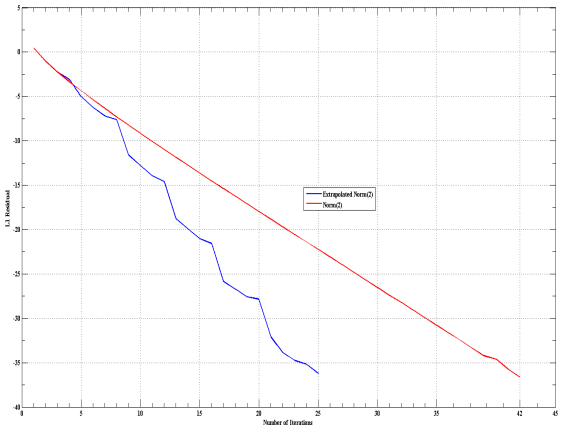
(a) HITS



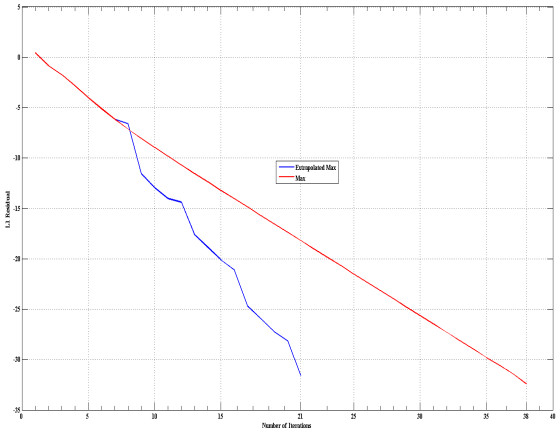
(b) HubAvg



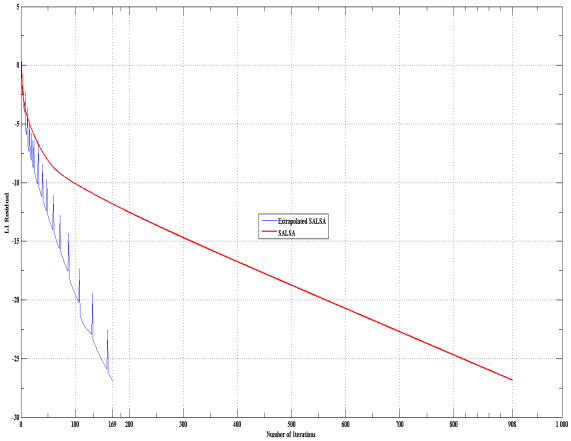
(c) AT-Avg



(d) Norm (2)

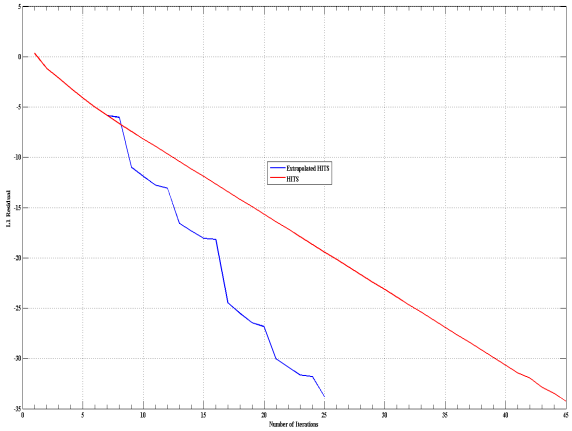


(e) Max

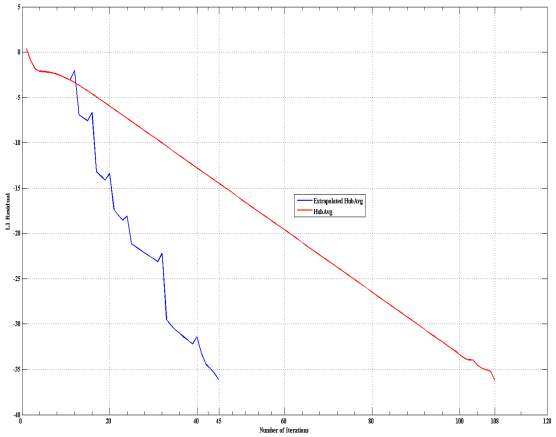


(f) SALSA

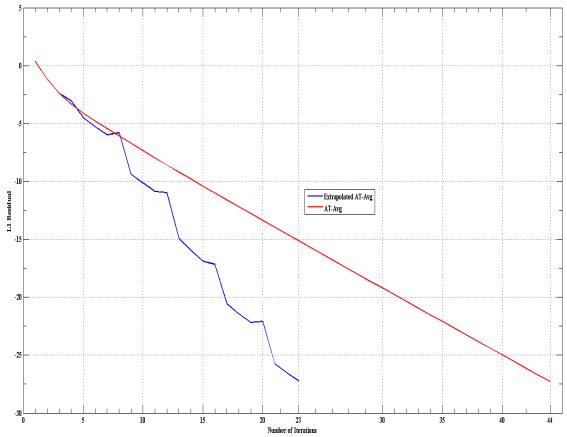
Figure A.16: *Convergence graphs for query "genetic"*



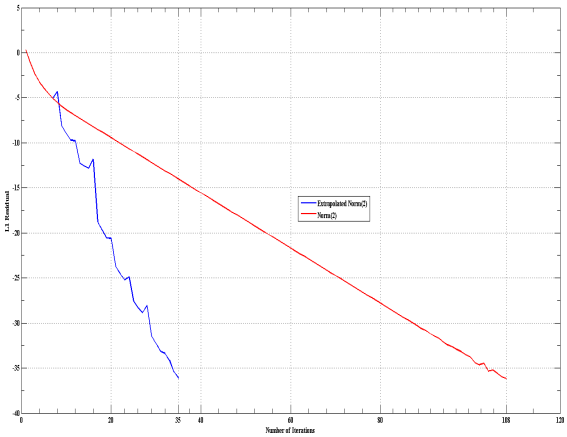
(a) HITS



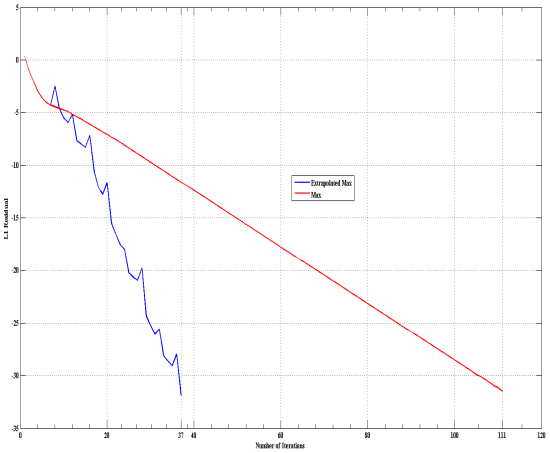
(b) HubAvg



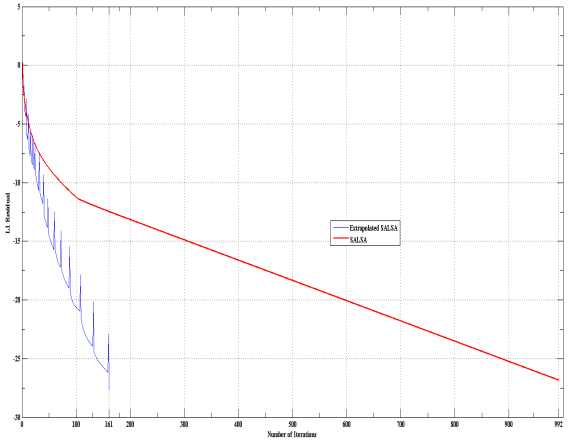
(c) AT-Avg



(d) Norm (2)

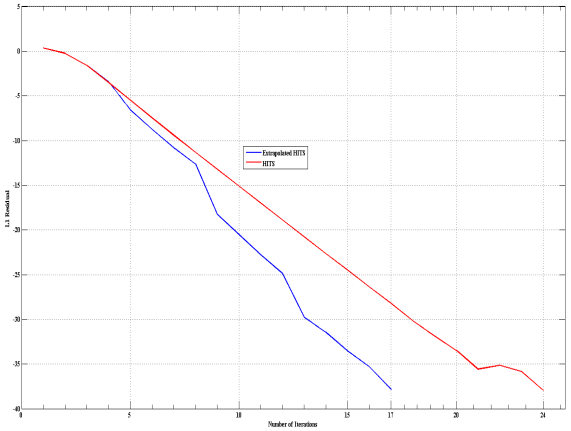


(e) Max

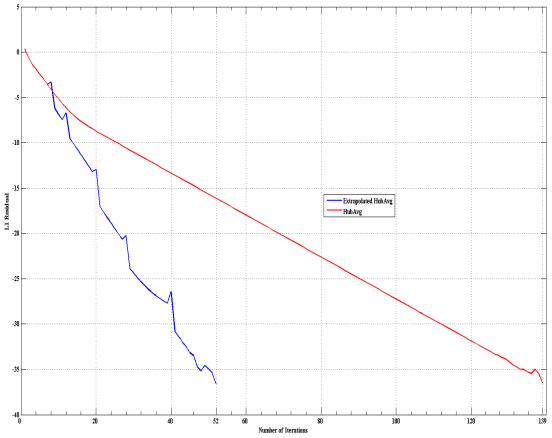


(f) SALSA

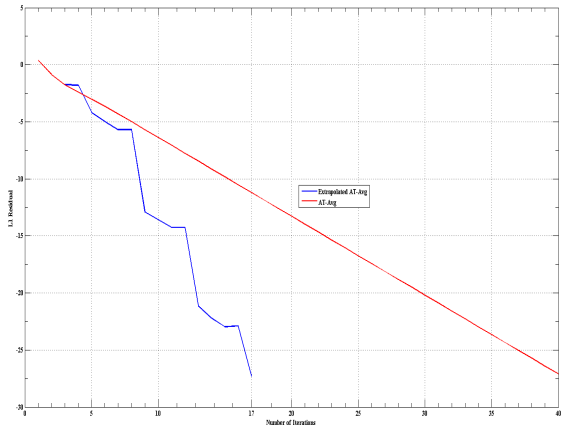
Figure A.17: *Convergence graphs for query "geometry"*



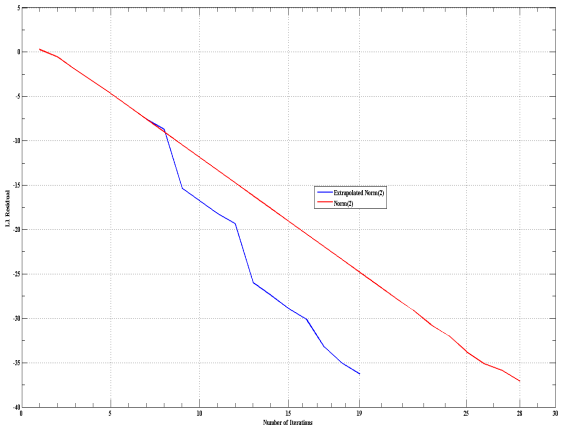
(a) HITS



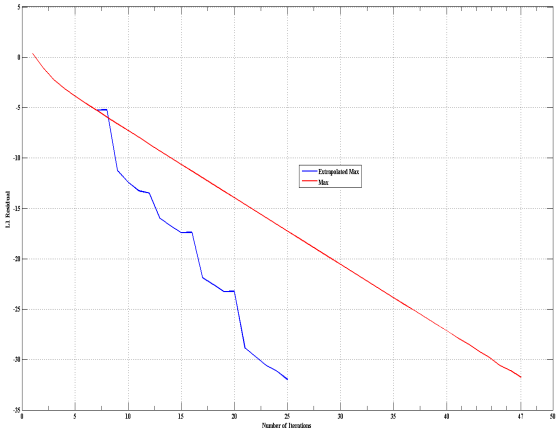
(b) HubAvg



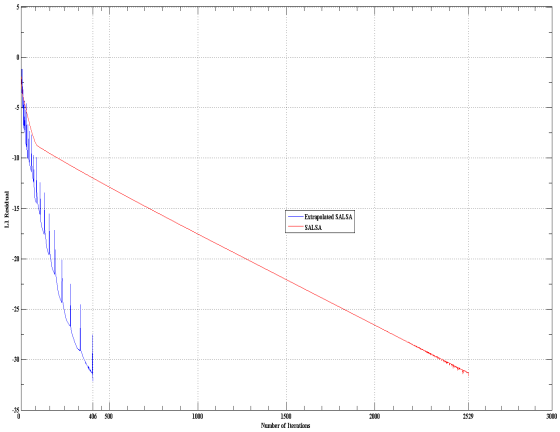
(c) AT-Avg



(d) Norm (2)

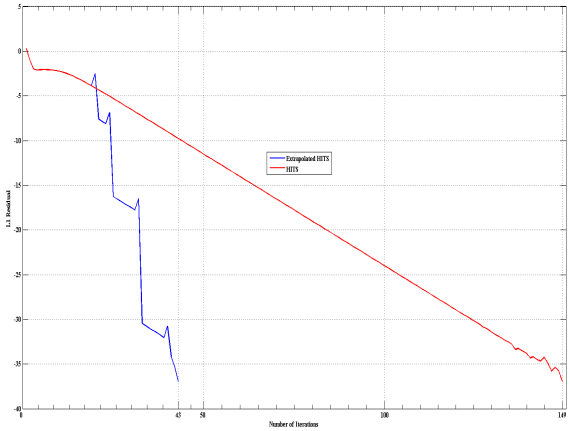


(e) Max

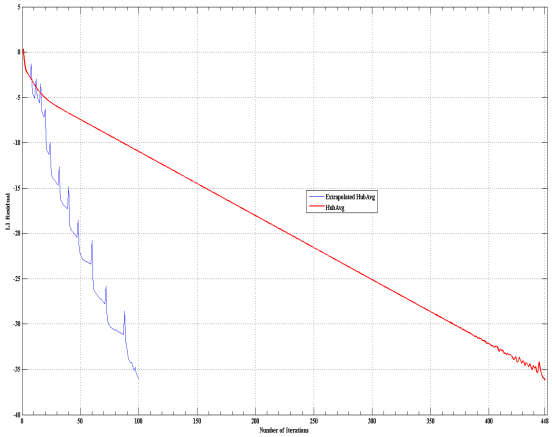


(f) SALSA

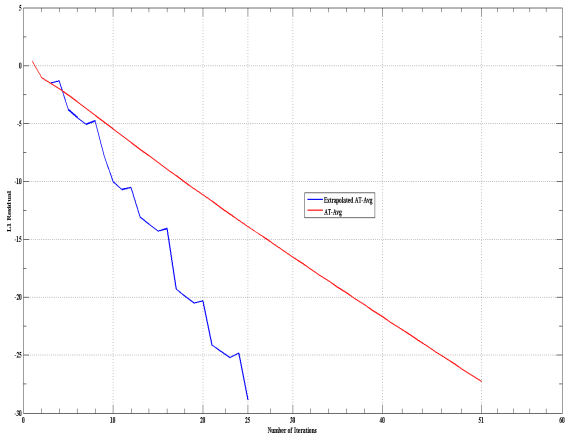
Figure A.18: *Convergence graphs for query "globalization"*



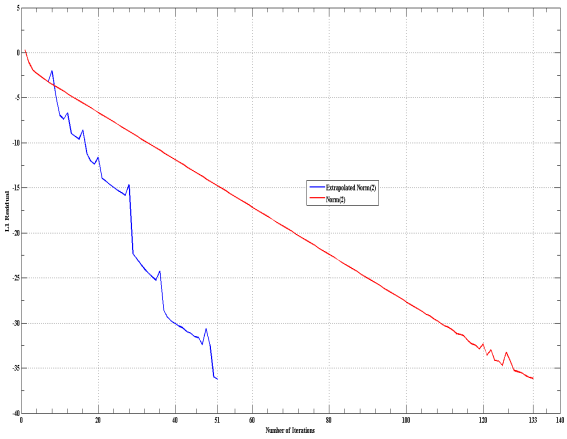
(a) HITS



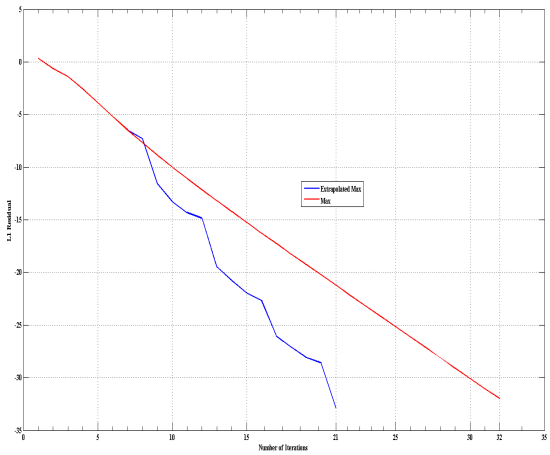
(b) HubAvg



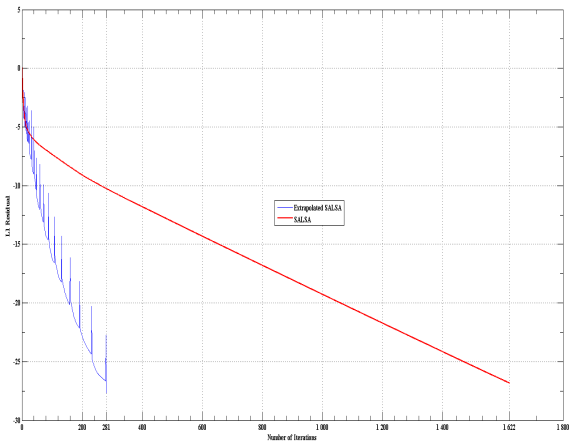
(c) AT-Avg



(d) Norm (2)

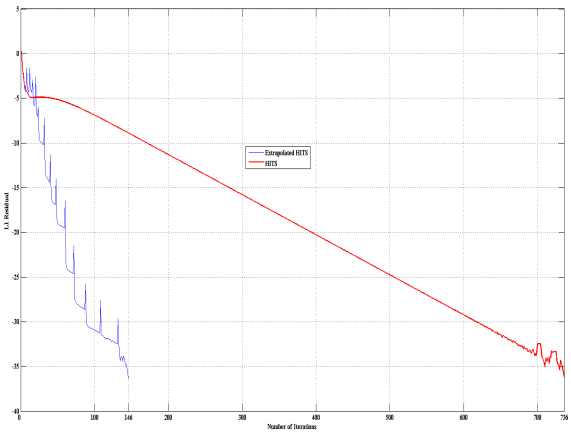


(e) Max

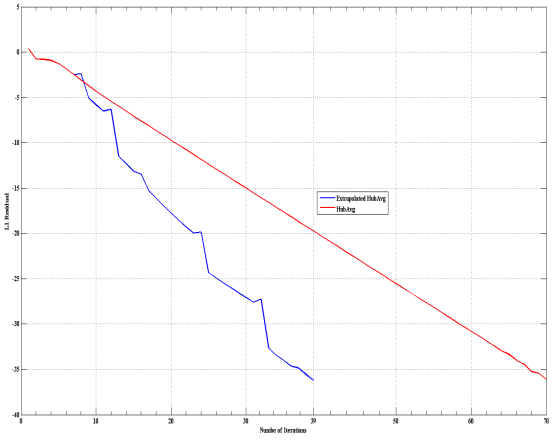


(f) SALSA

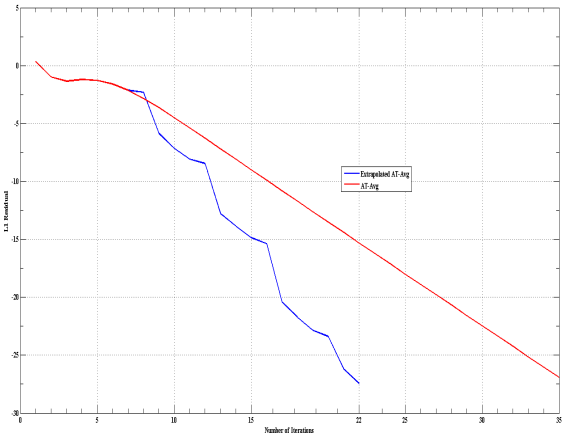
Figure A.19: *Convergence graphs for query "gun control"*



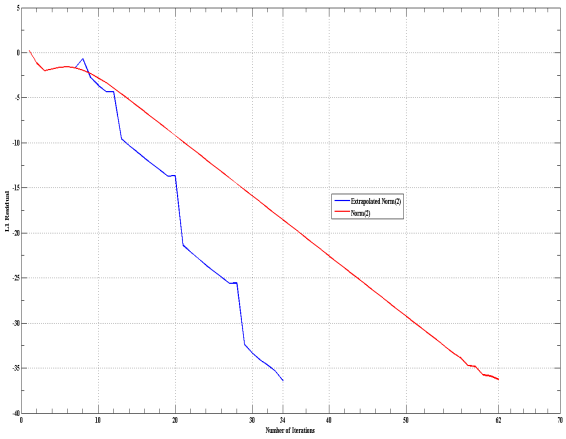
(a) HITS



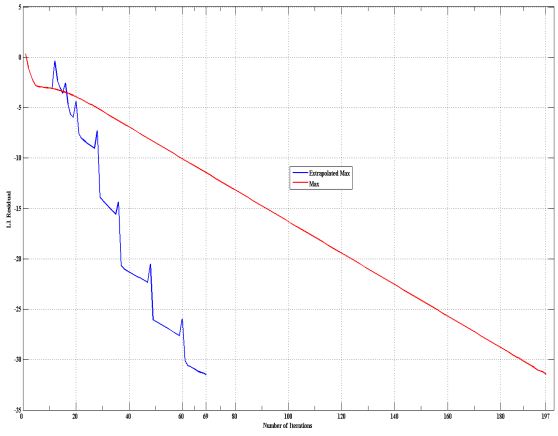
(b) HubAvg



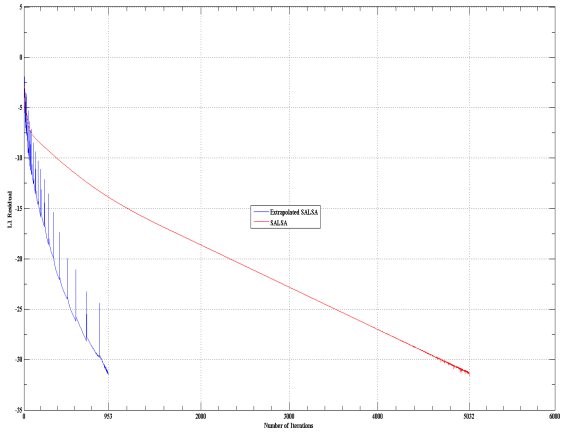
(c) AT-Avg



(d) Norm (2)

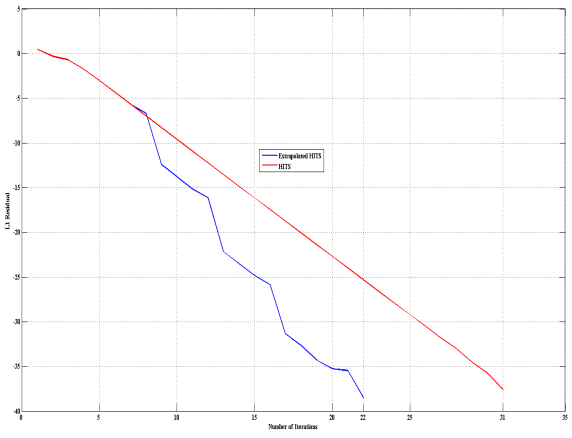


(e) Max

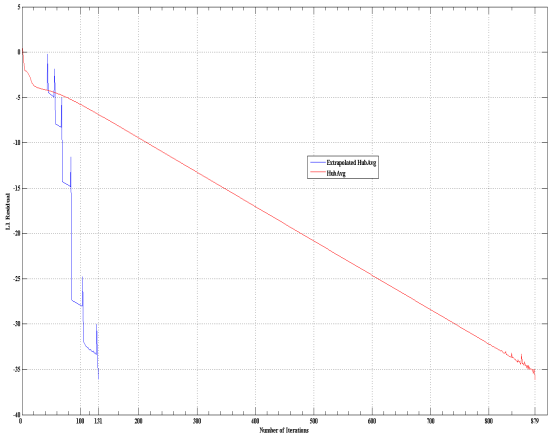


(f) SALSA

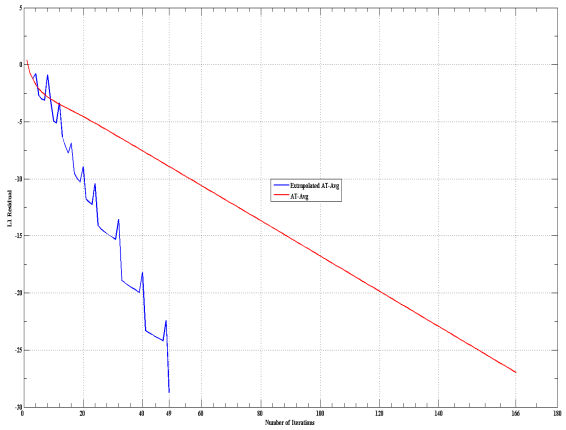
Figure A.20: *Convergence graphs for query "iraq war"*



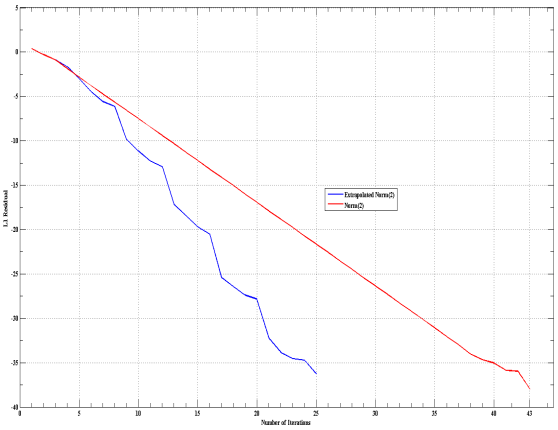
(a) HITS



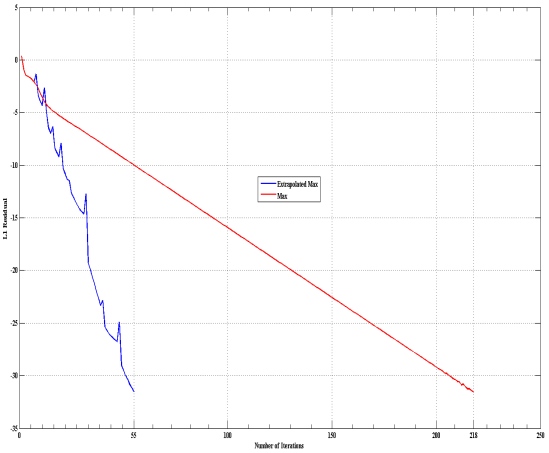
(b) HubAvg



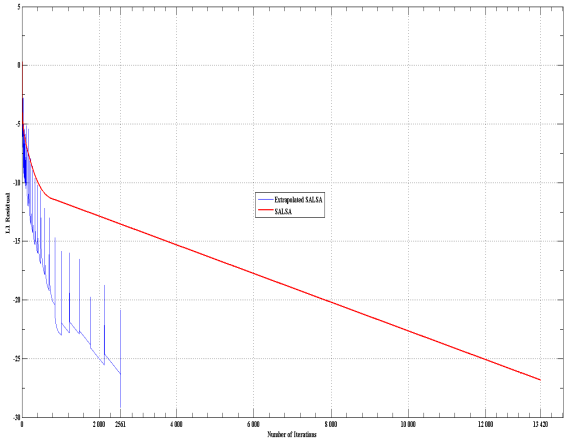
(c) AT-Avg



(d) Norm (2)

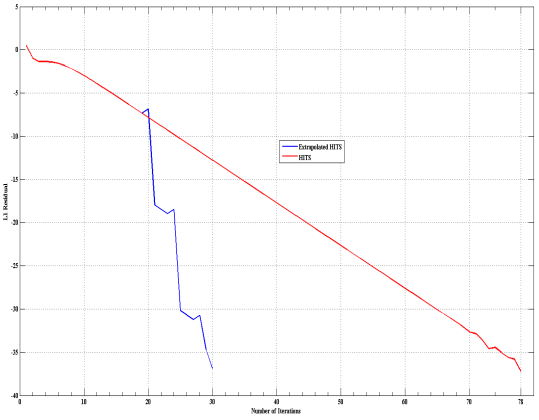


(e) Max

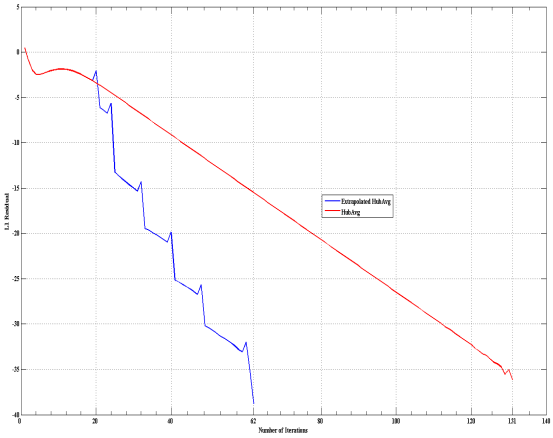


(f) SALSA

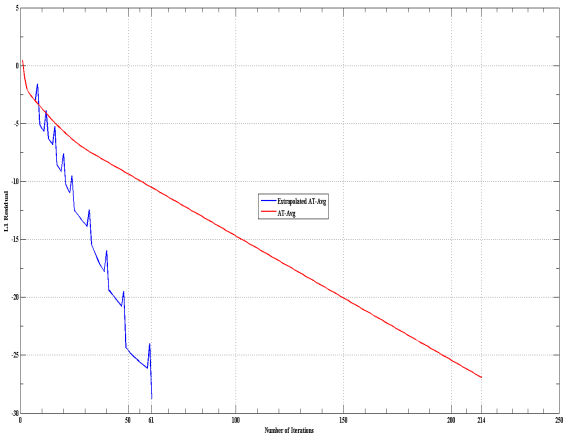
Figure A.21: *Convergence graphs for query "jaguar"*



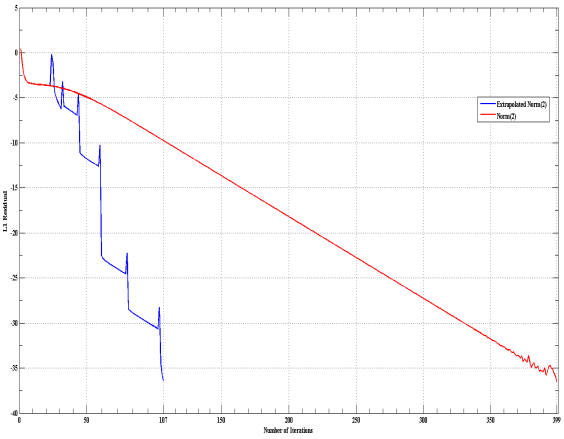
(a) HITS



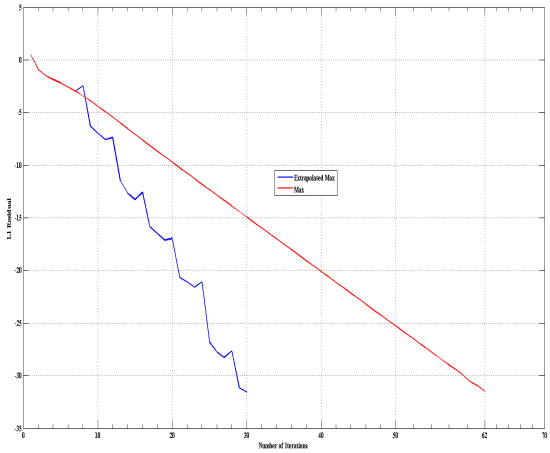
(b) HubAvg



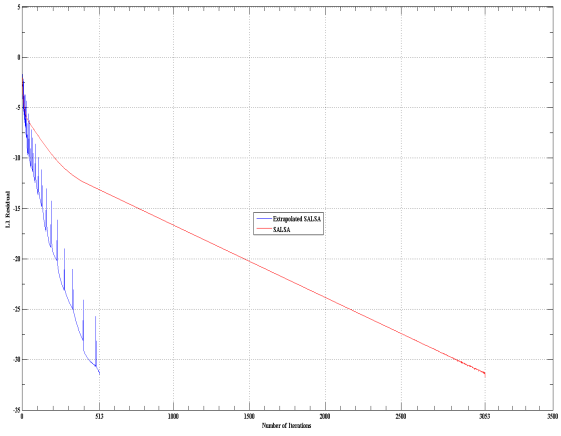
(c) AT-Avg



(d) Norm (2)

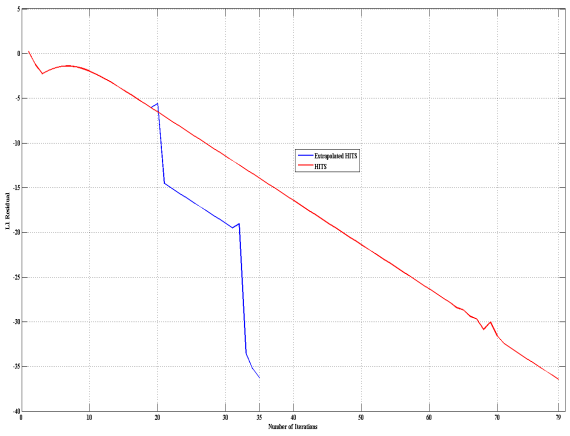


(e) Max

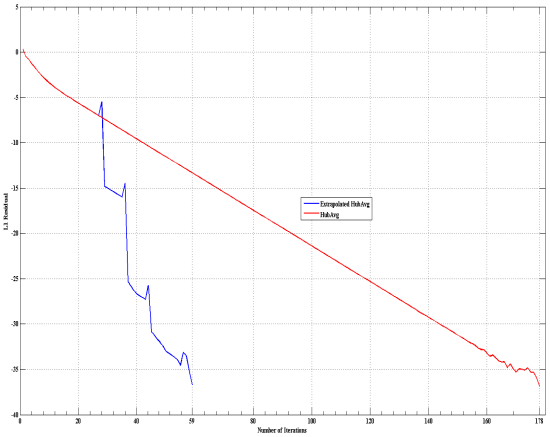


(f) SALSA

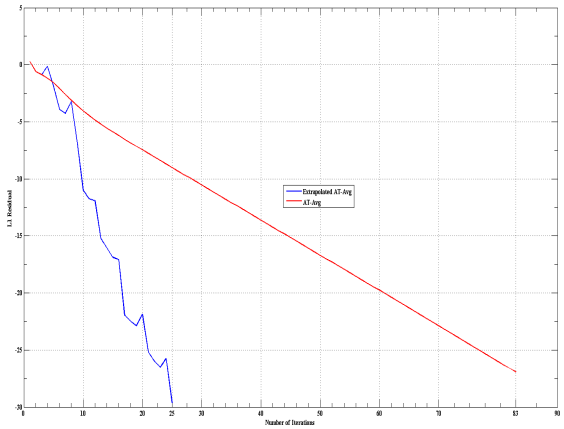
Figure A.22: *Convergence graphs for query "jordan"*



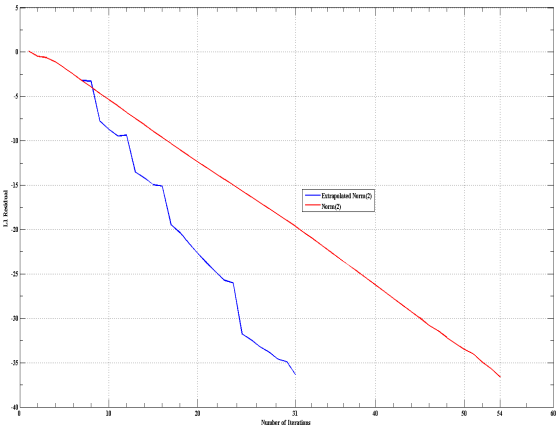
(a) HITS



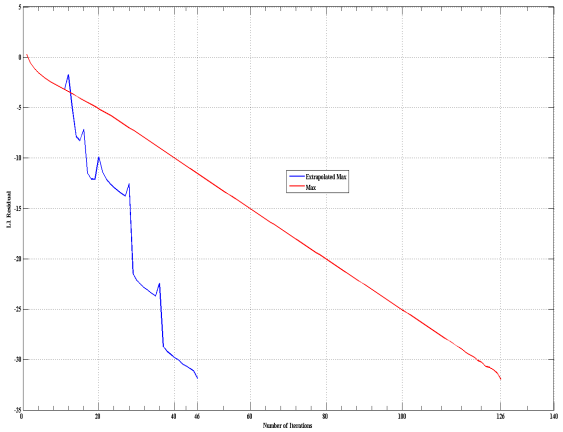
(b) HubAvg



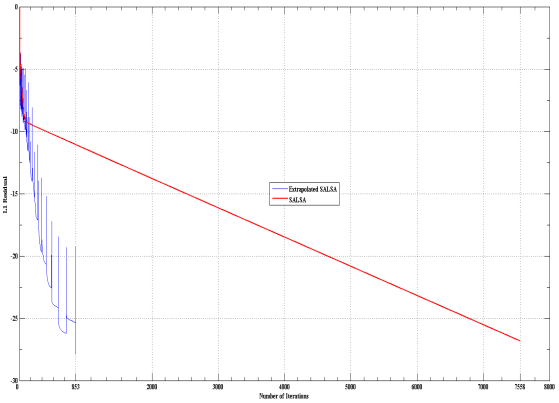
(c) AT-Avg



(d) Norm (2)

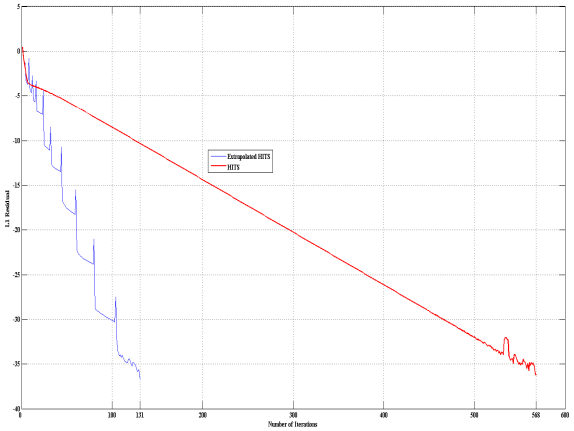


(e) Max

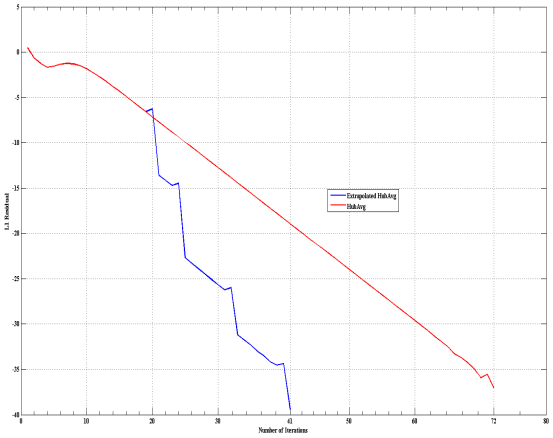


(f) SALSA

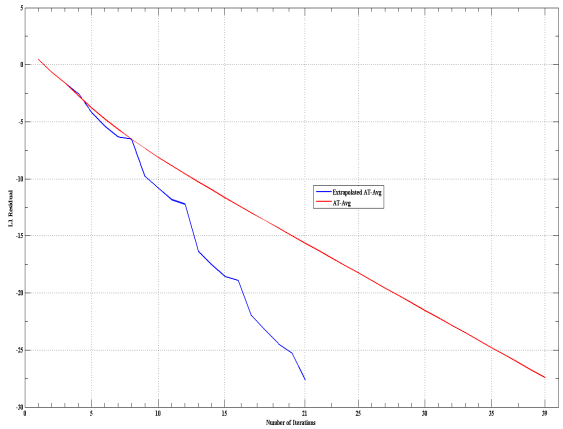
Figure A.23: *Convergence graphs for query "moon landing"*



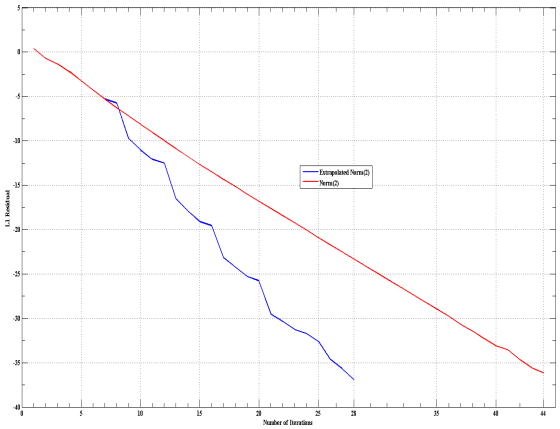
(a) HITS



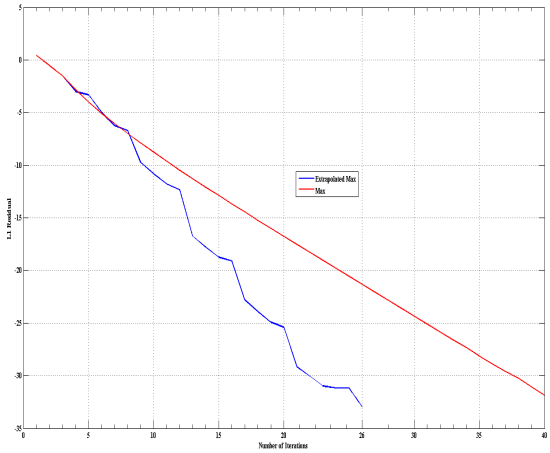
(b) HubAvg



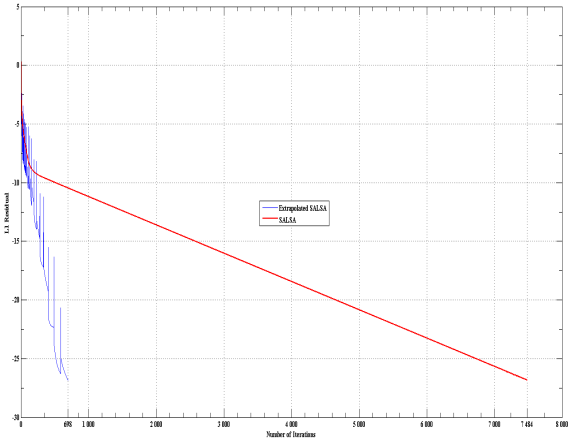
(c) AT-Avg



(d) Norm (2)

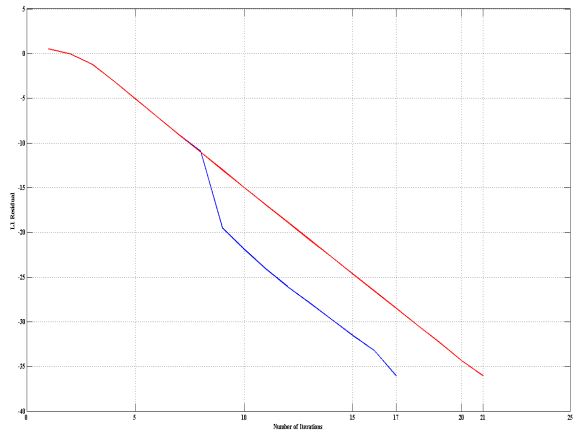


(e) Max

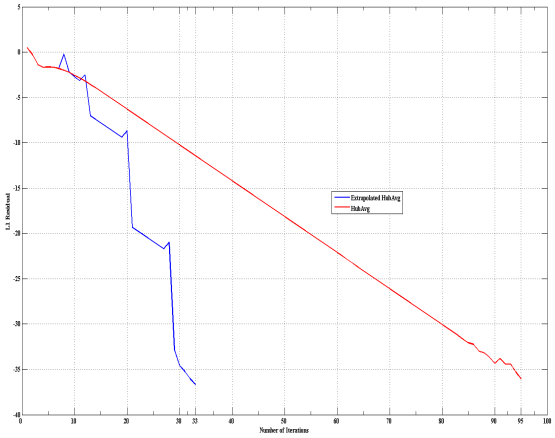


(f) SALSA

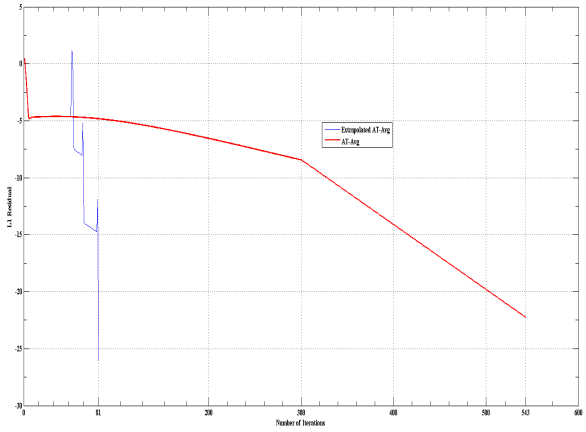
Figure A.24: *Convergence graphs for query "movies"*



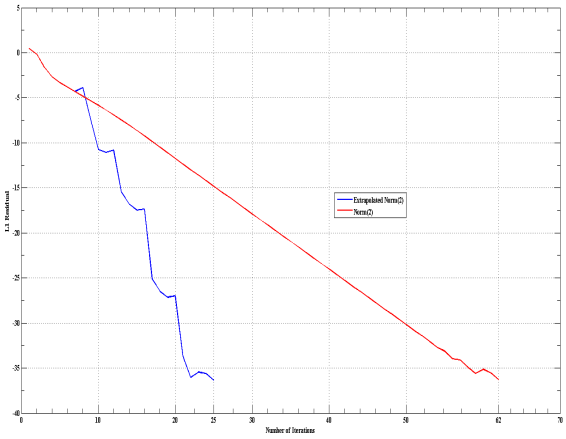
(a) HITS



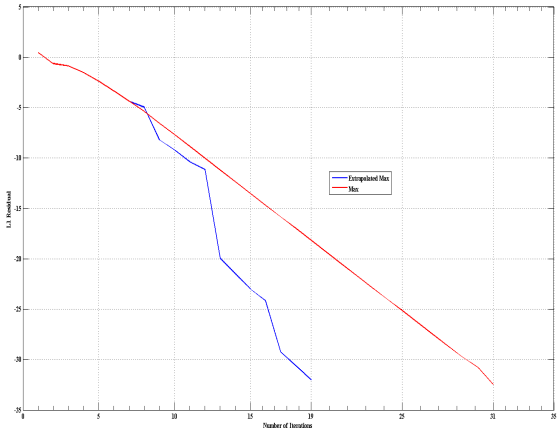
(b) HubAvg



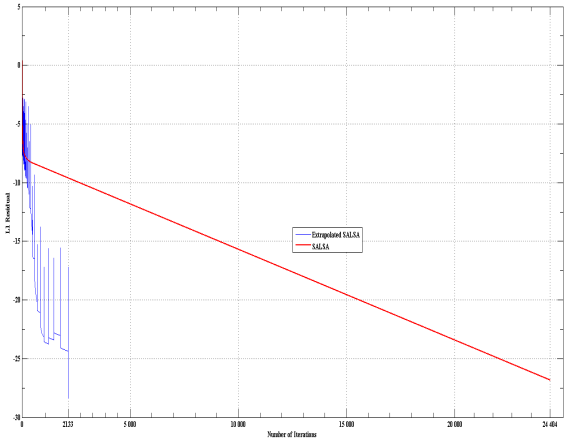
(c) AT-Avg



(d) Norm (2)

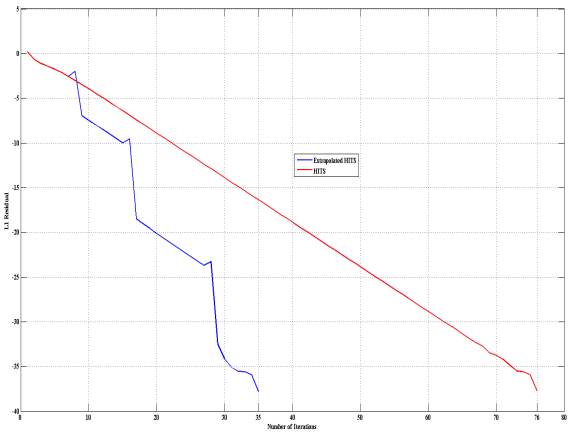


(e) Max

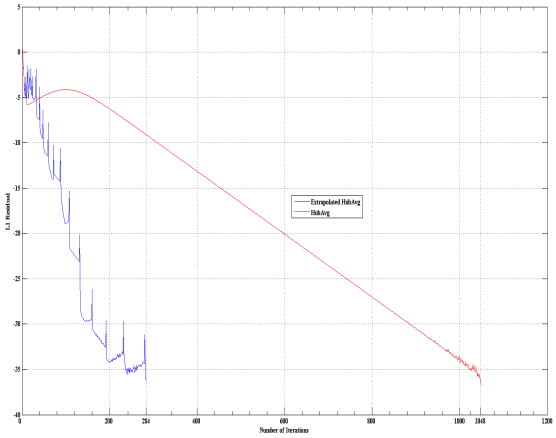


(f) SALSA

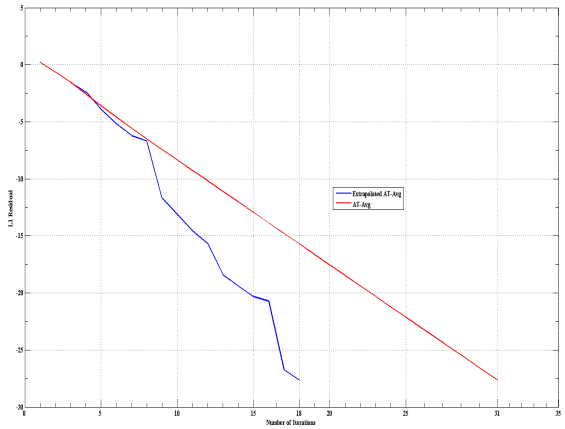
Figure A.25: *Convergence graphs for query "national parks"*



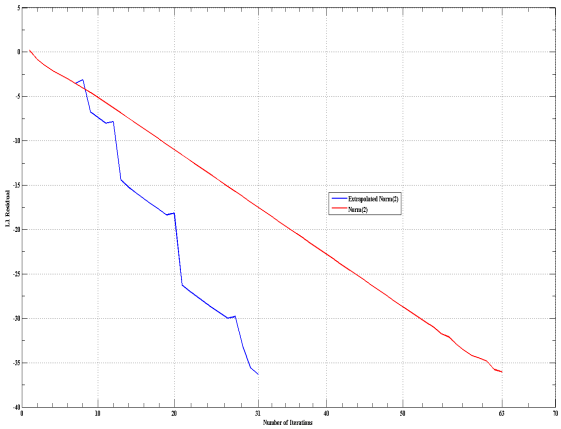
(a) HITS



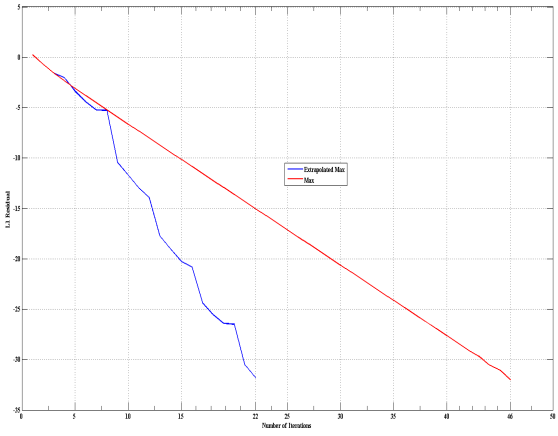
(b) HubAvg



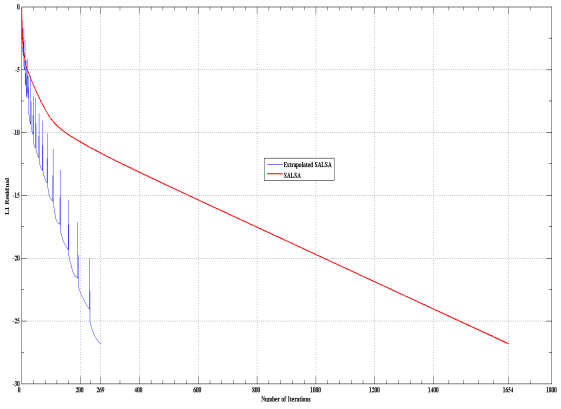
(c) AT-Avg



(d) Norm (2)

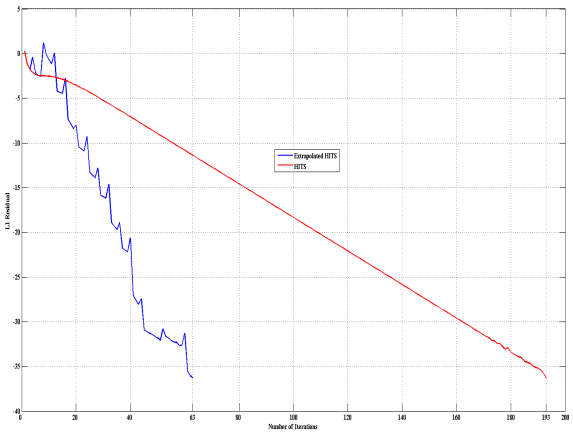


(e) Max

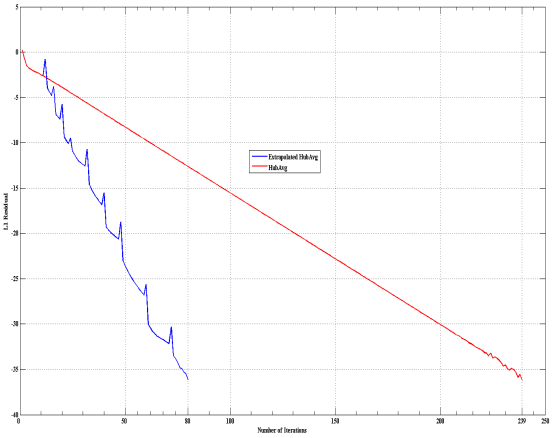


(f) SALSA

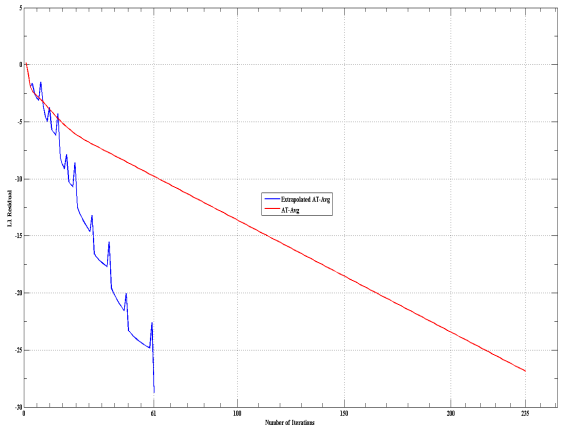
Figure A.26: *Convergence graphs for query "net censorship"*



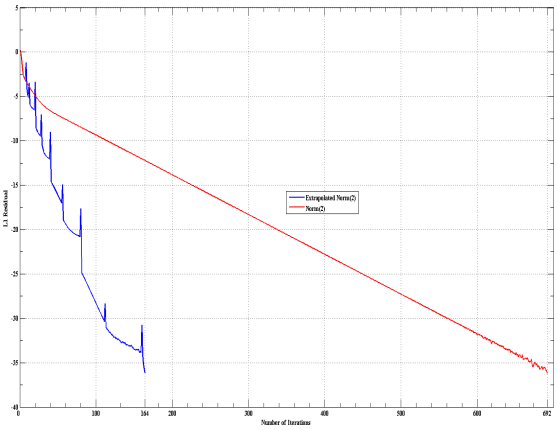
(a) HITS



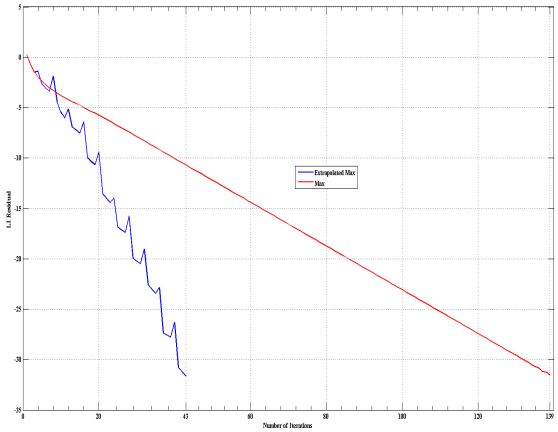
(b) HubAvg



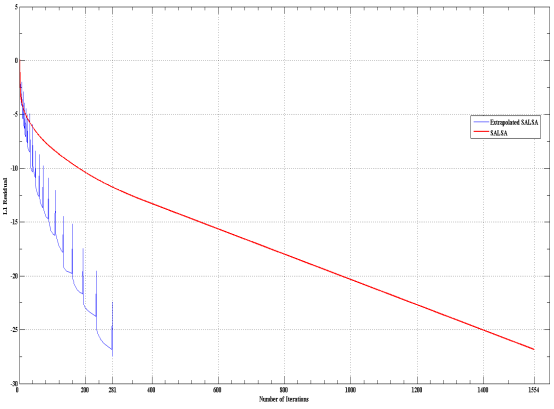
(c) AT-Avg



(d) Norm (2)

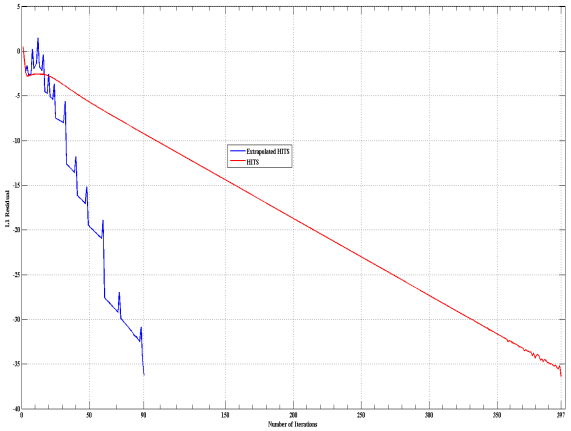


(e) Max

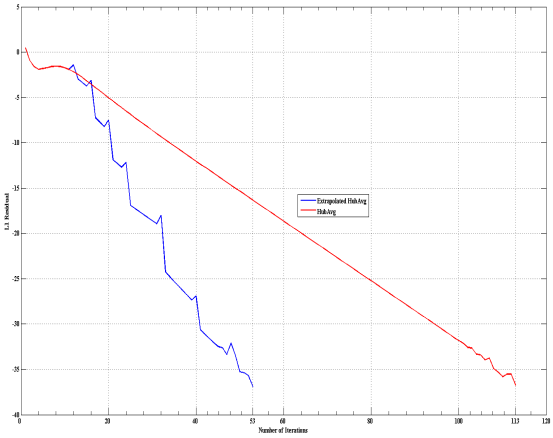


(f) SALSA

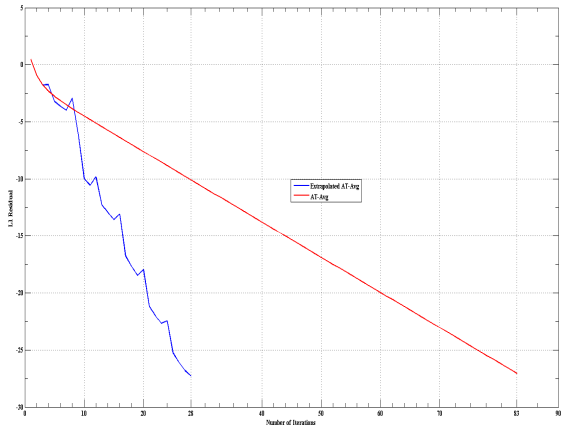
Figure A.27: *Convergence graphs for query "randomized algorithms"*



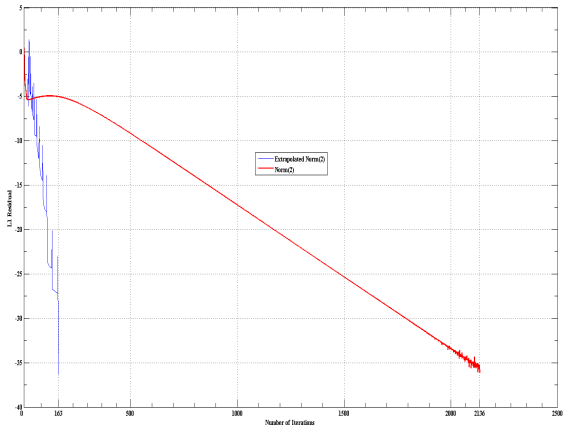
(a) HITS



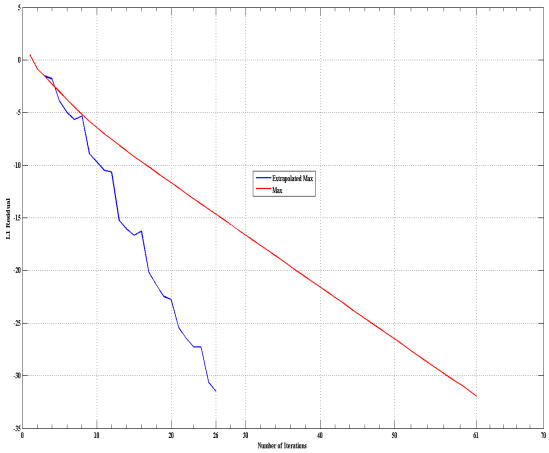
(b) HubAvg



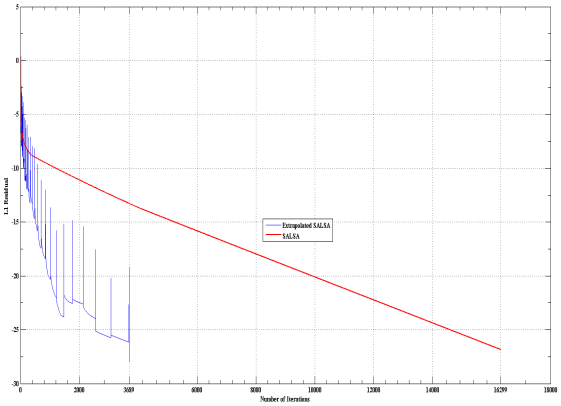
(c) AT-Avg



(d) Norm (2)

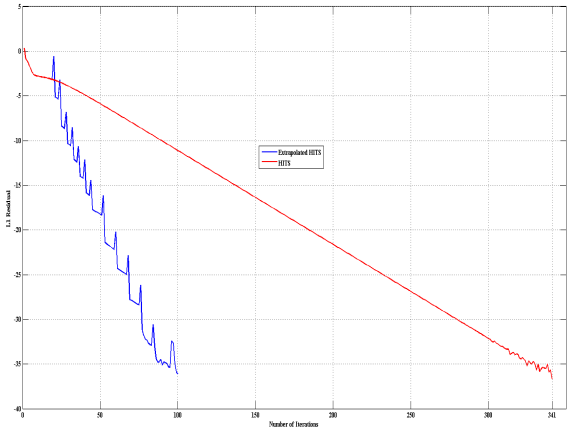


(e) Max

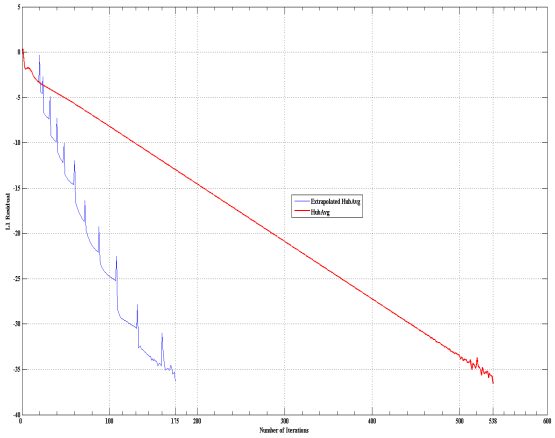


(f) SALSA

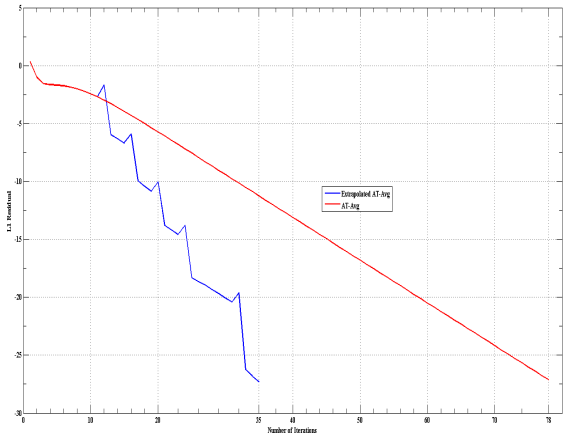
Figure A.28: *Convergence graphs for query "recipes"*



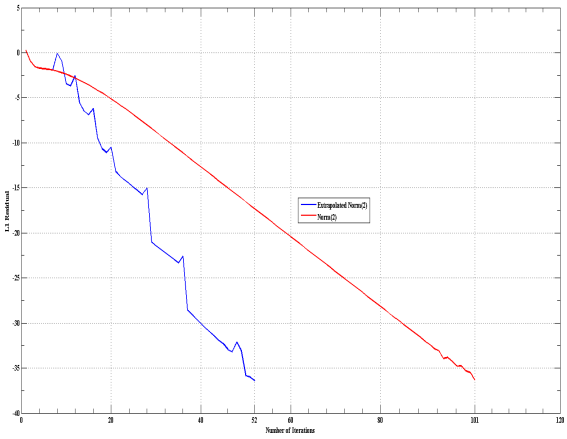
(a) HITS



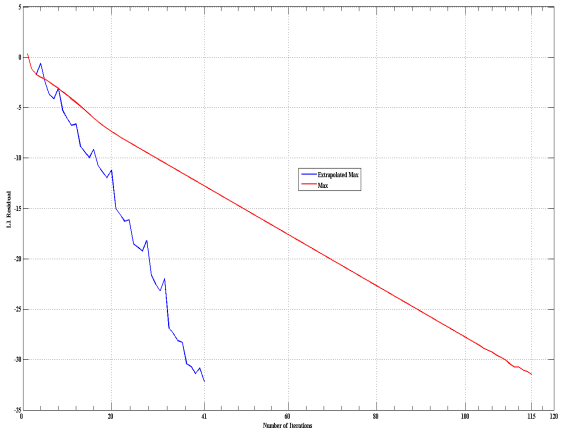
(b) HubAvg



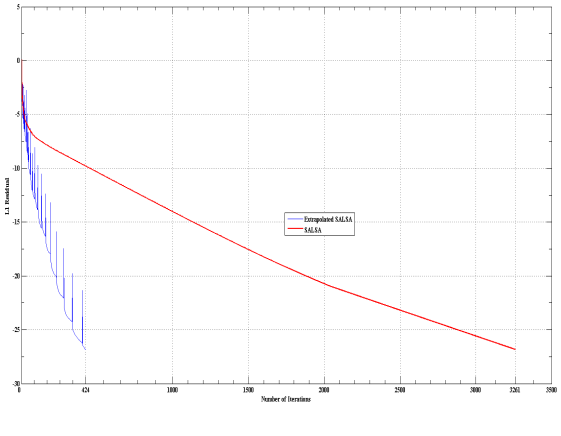
(c) AT-Avg



(d) Norm (2)

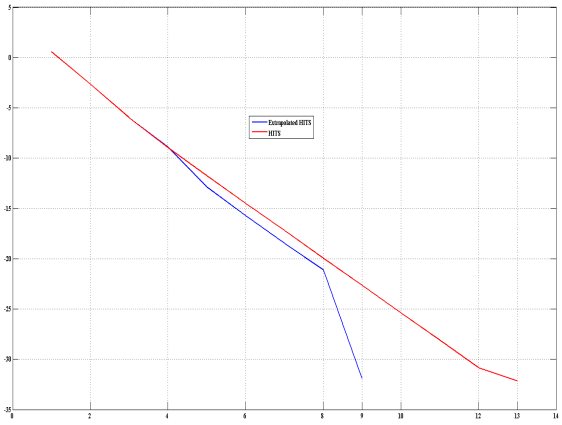


(e) Max

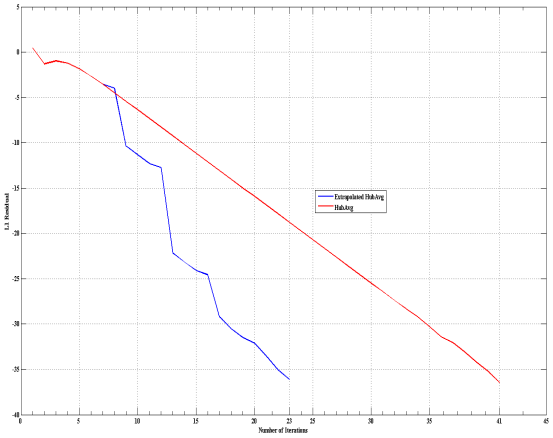


(f) SALSA

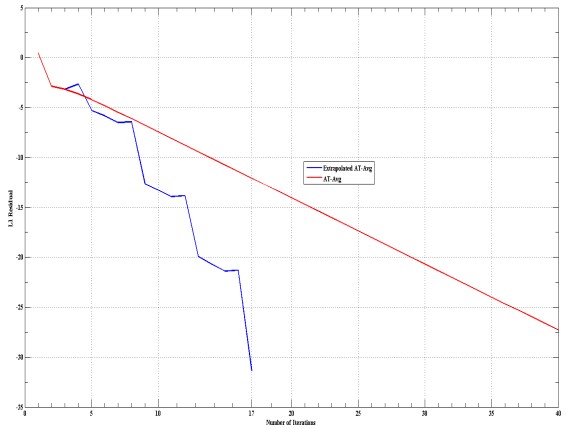
Figure A.29: *Convergence graphs for query "roswell"*



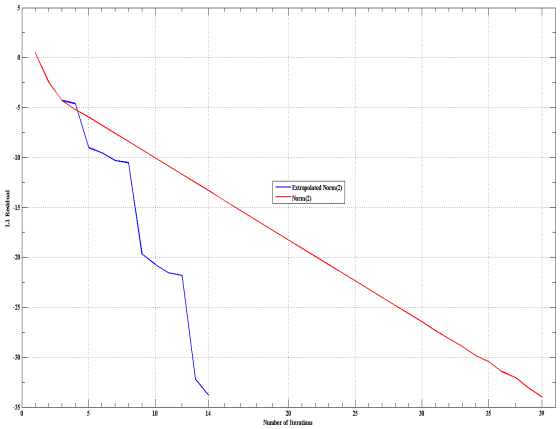
(a) HITS



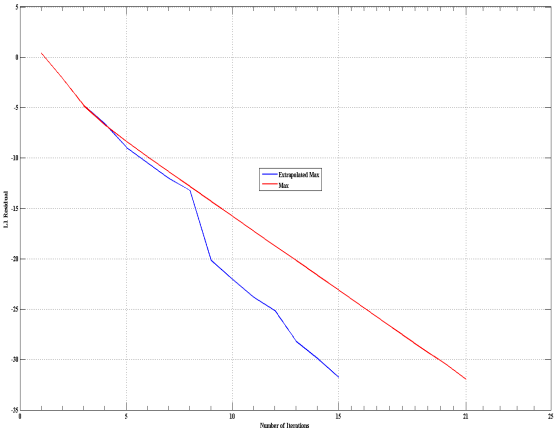
(b) HubAvg



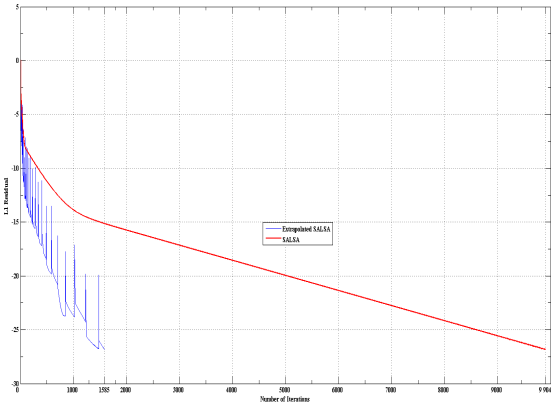
(c) AT-Avg



(d) Norm (2)

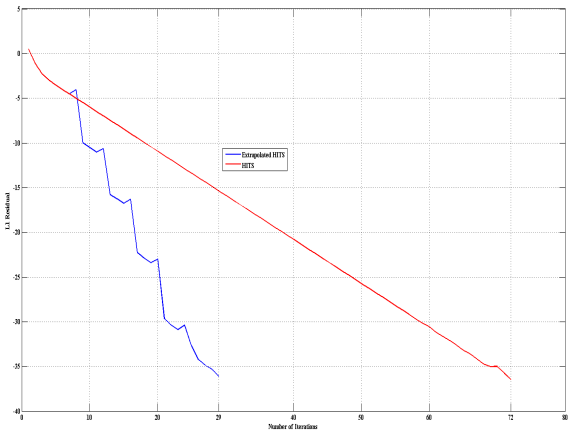


(e) Max

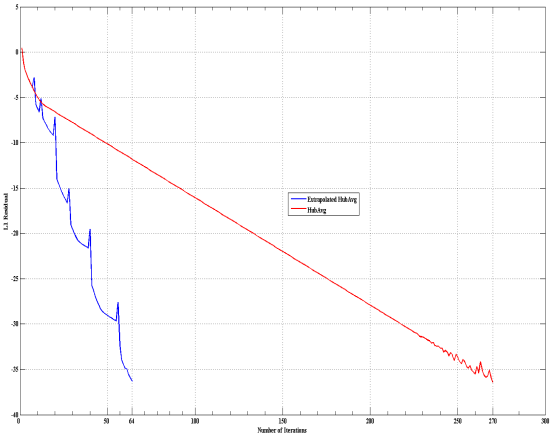


(f) SALSA

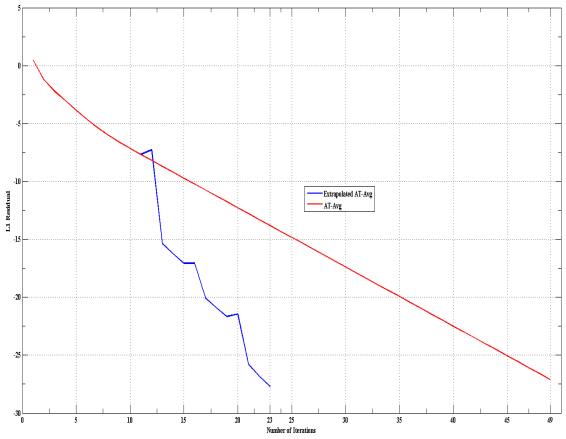
Figure A.30: *Convergence graphs for query "search engines"*



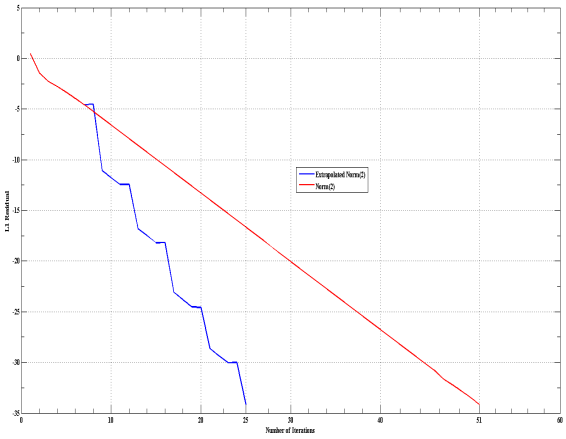
(a) HITS



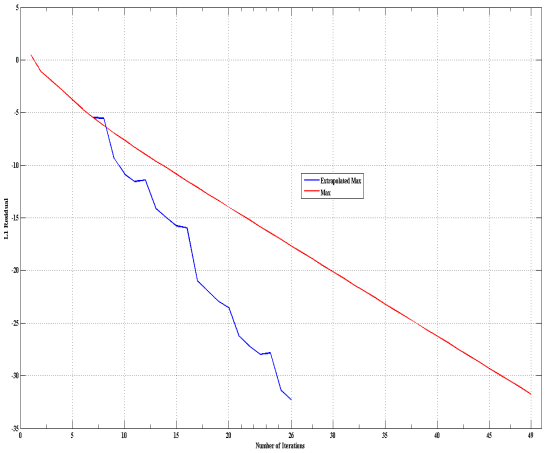
(b) HubAvg



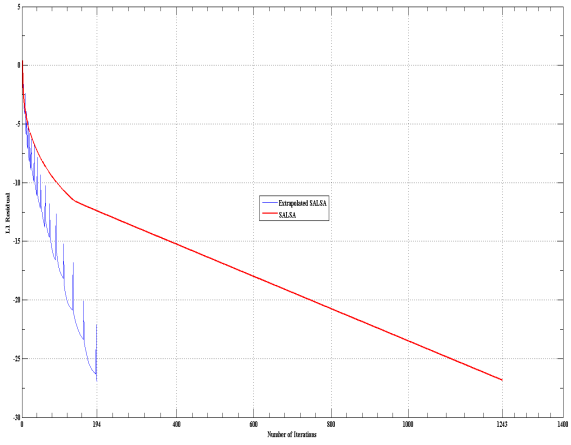
(c) AT-Avg



(d) Norm (2)

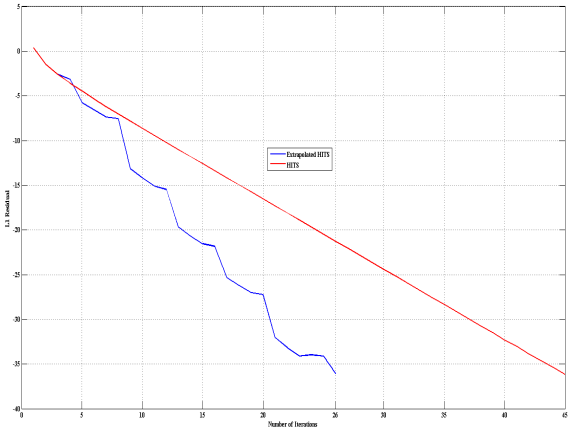


(e) Max

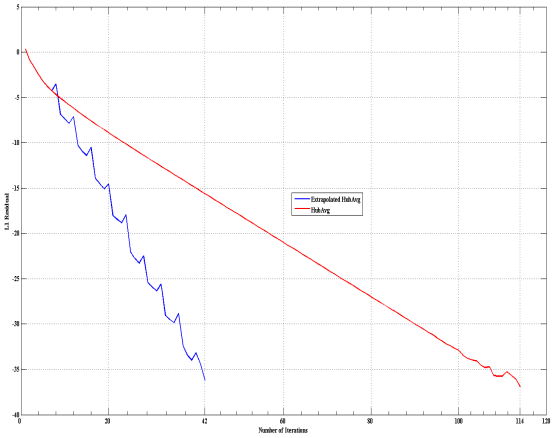


(f) SALSA

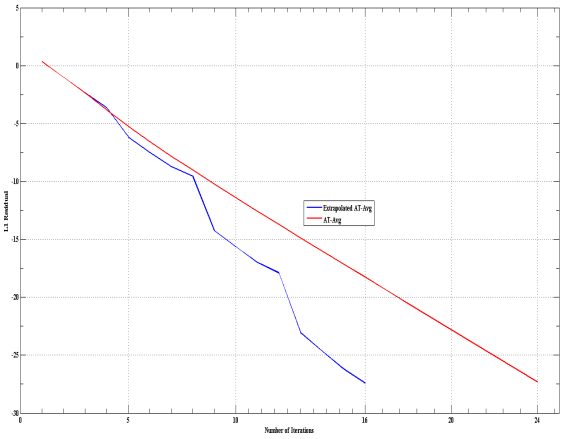
Figure A.31: *Convergence graphs for query "shakespeare"*



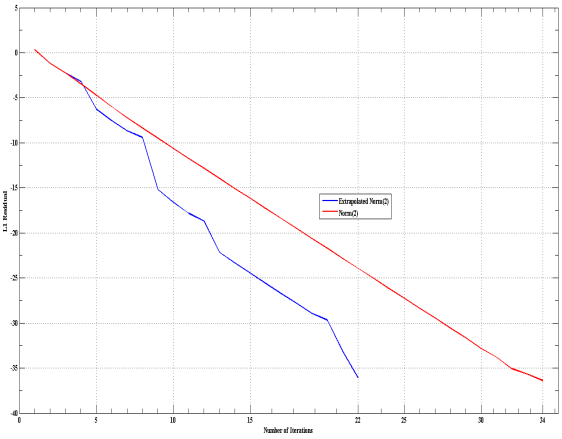
(a) HITS



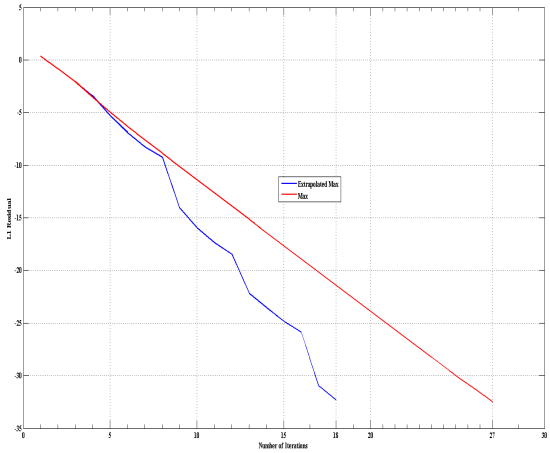
(b) HubAvg



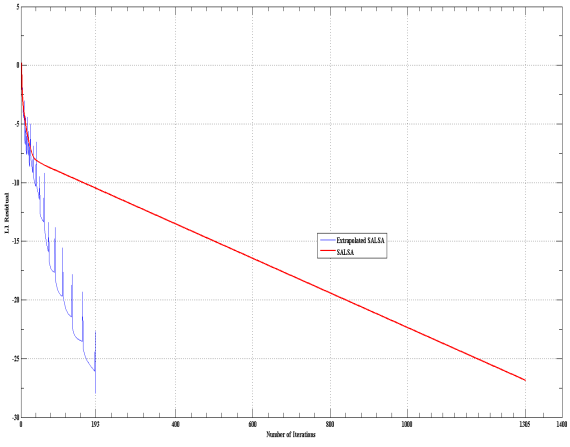
(c) AT-Avg



(d) Norm (2)

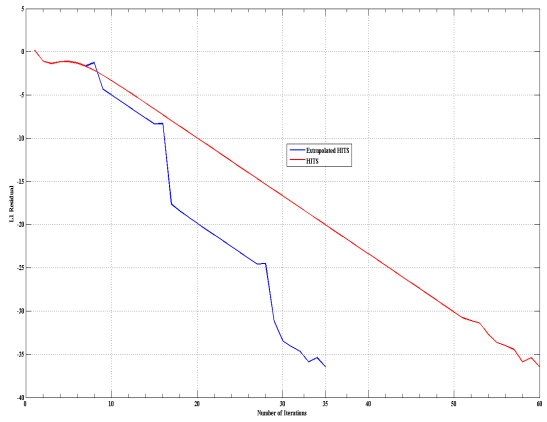


(e) Max

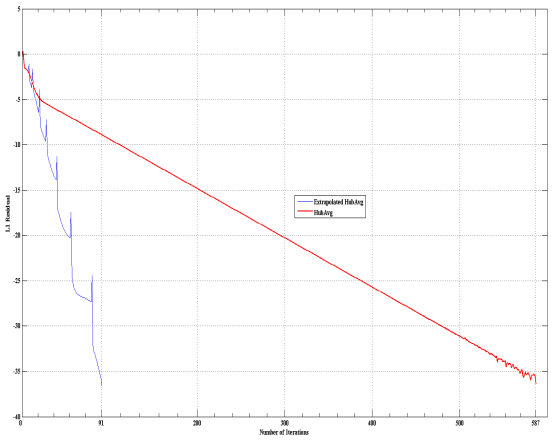


(f) SALSA

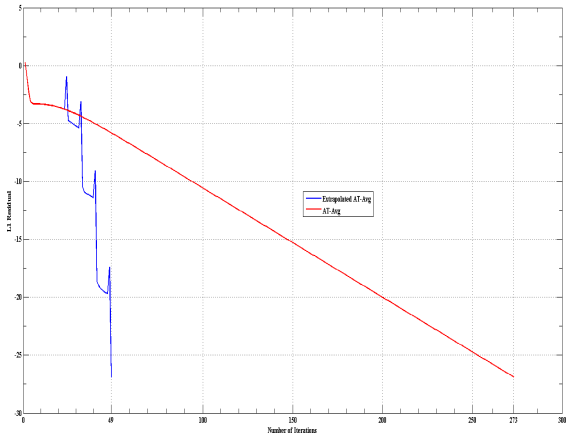
Figure A.32: *Convergence graphs for query "table tennis"*



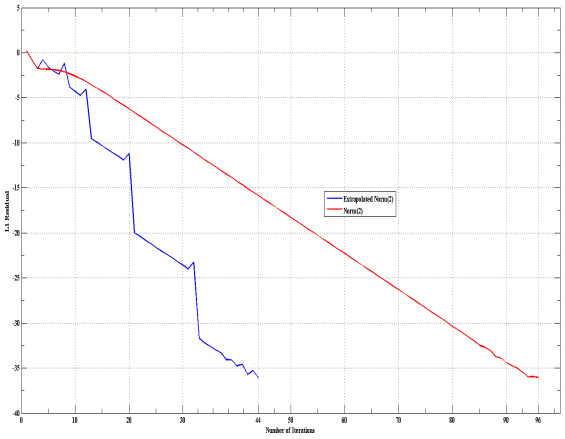
(a) HITS



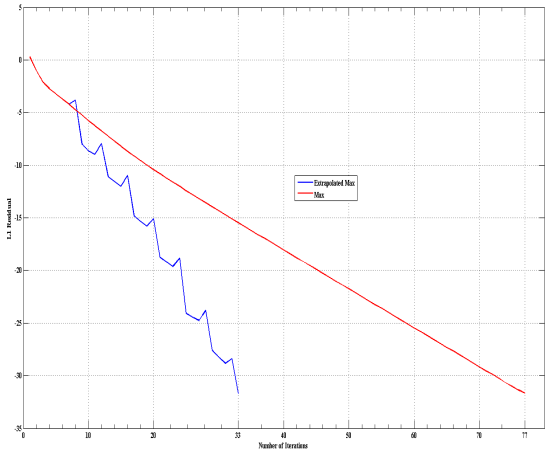
(b) HubAvg



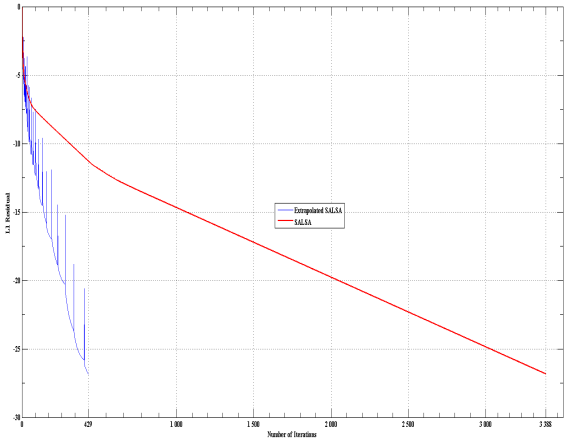
(c) AT-Avg



(d) Norm (2)

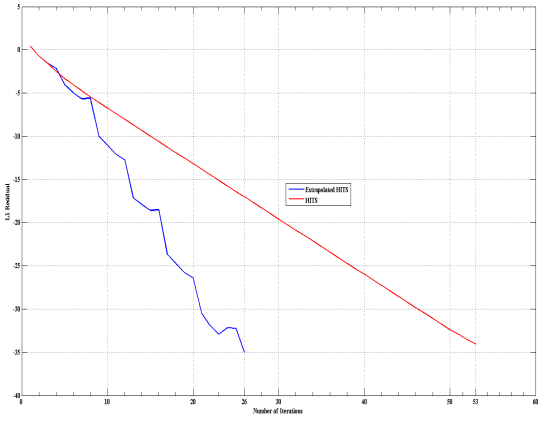


(e) Max

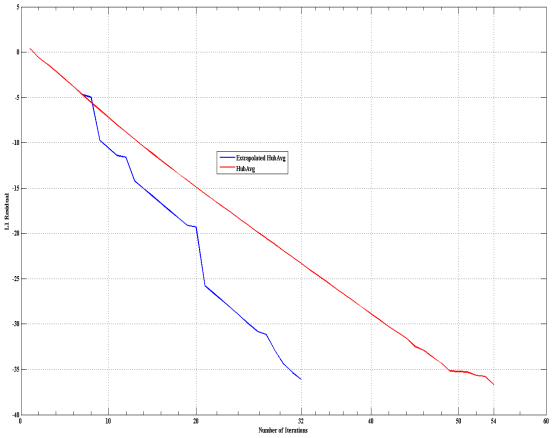


(f) SALSA

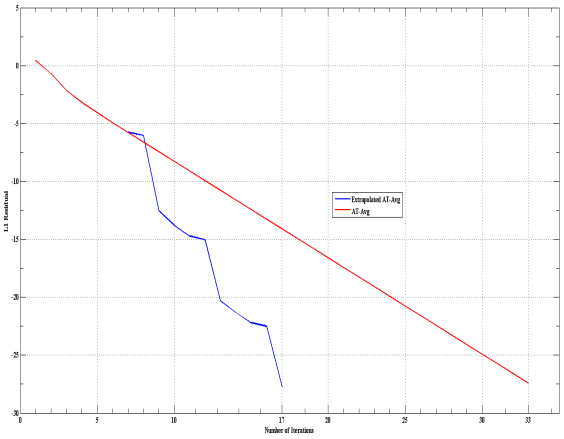
Figure A.33: *Convergence graphs for query "vintage cars"*



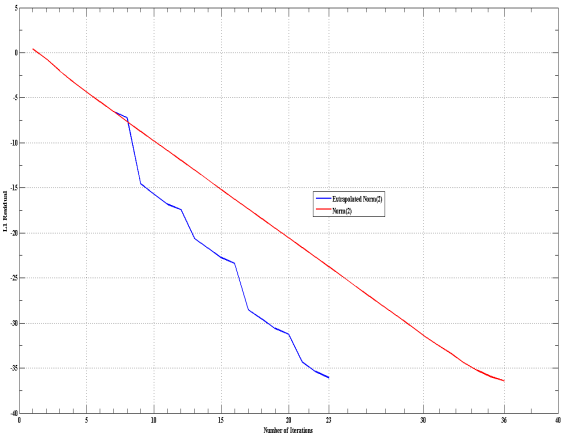
(a) HITS



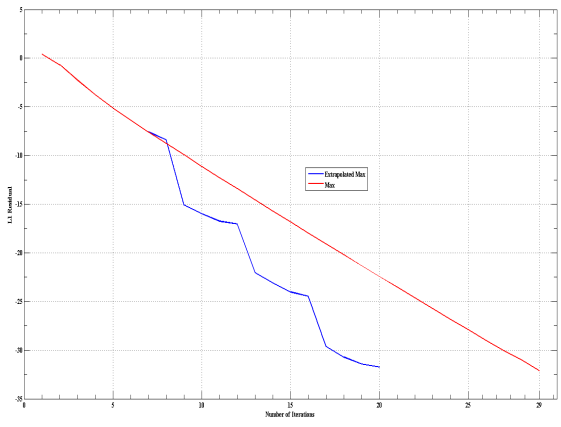
(b) HubAvg



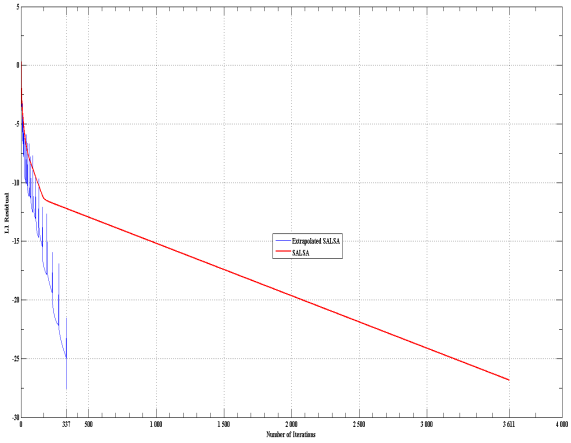
(c) AT-Avg



(d) Norm (2)



(e) Max



(f) SALSA

Figure A.34: *Convergence graphs for query "weather"*

Appendix B

Experiments - *Top-15* Results

HITS	**HubAvg**
1. [pid: 1537] Priests for Life Index URL: http://www.priestsforlife.org	1. [pid: 4] prochoiceamerica.org : NARAL Pro-Choice America URL: http://www.naral.org
2. [pid: 1023] National Right to Life URL: http://www.nrlc.org	2. [pid: 677] Planned Parenthood Federation of America URL: http://www.plannedparenthood.org
3. [pid: 7] After Abortion: Information on the aftereffects of abortion and p URL: http://www.afterabortion.org	3. [pid: 1] NAF - The Voice of Abortion Providers URL: http://www.prochoice.org
4. [pid: 1545] ProLifeInfo.org URL: http://www.prolifeinfo.org	4. [pid: 2] Abortion Clinics OnLine URL: http://www.gynpages.com
5. [pid: 1417] Pregnancy Centers Online URL: http://www.pregnancycenters.org	5. [pid: 232] FEMINIST MAJORITY FOUNDATION ONLINE HOMEPAGE URL: http://www.feminist.org
6. [pid: 1531] Human Life International URL: http://www.hli.org	6. [pid: 259] The Alan Guttmacher Institute: Home Page URL: http://www.guttmacher.org
7. [pid: 17] Abortion - Breast Cancer Link – Dr. Joel Brind Ph. D. URL: http://www.abortioncancer.com	7. [pid: 227] center for reproductive rights URL: http://www.crlp.org
8. [pid: 3] Abortion facts and information, statistics, hotlines and helpline URL: http://www.abortionfacts.com	8. [pid: 1798] The Religious Coalition for Reproductive Choice URL: http://www.rcrc.org
9. [pid: 1259] Campaign Life Coalition British Columbia URL: http://www.clcbc.org	9. [pid: 601] National Organization for Women URL: http://www.now.org
10. [pid: 300] Empty title field URL: http://www.heritagehouse76.com	10. [pid: 1780] Medical Students for Choice: Home URL: http://www.ms4c.org
11. [pid: 21] Coalition on Abortion/Breast Cancer URL: http://www.abortionbreastcancer.com	11. [pid: 22] NCAP URL: http://www.ncap.com
12. [pid: 1586] Americans United for Life URL: http://www.americansunitedforlife.org	12. [pid: 1774] CFFC Splash Screen URL: http://www.cath4choice.org
13. [pid: 38] Lifecall... Resources for Pregnant Woment and Their Babies URL: http://www.lifecall.org	13. [pid: 14] National Network of Abortion Funds URL: http://www.nnaf.org
14. [pid: 1590] The Justice Foundation's Home Page URL: http://justice.oct.net	14. [pid: 1023] National Right to Life URL: http://www.nrlc.org
15. [pid: 1418] just the facts- life before birth, your first nine months URL: http://www.justthefacts.org	15. [pid: 8] California Abortion & Reproductive Rights Action League (CARAL) URL: http://www.caral.org
AT-Avg	**Norm (2)**
1. [pid:1023] National Right to Life URL: http://www.nrlc.org	1. [pid:1023] National Right to Life URL: http://www.nrlc.org
2. [pid:1537] Priests for Life Index URL: http://www.priestsforlife.org	2. [pid:1537] Priests for Life Index URL: http://www.priestsforlife.org
3. [pid:1545] ProLifeInfo.org URL: http://www.prolifeinfo.org	3. [pid:1417] Pregnancy Centers Online URL: http://www.pregnancycenters.org
4. [pid:1417] Pregnancy Centers Online URL: http://www.pregnancycenters.org	4. [pid:1545] ProLifeInfo.org URL: http://www.prolifeinfo.org
5. [pid:7] After Abortion: Information on the aftereffects of abortion and p URL: http://www.afterabortion.org	5. [pid:7] After Abortion: Information on the aftereffects of abortion and p URL: http://www.afterabortion.org
6. [pid:4] prochoiceamerica.org : NARAL Pro-Choice America URL: http://www.naral.org	6. [pid:1531] Human Life International URL: http://www.hli.org
7. [pid:1531] Human Life International URL: http://www.hli.org	7. [pid:17] Abortion - Breast Cancer Link – Dr. Joel Brind Ph. D. URL: http://www.abortioncancer.com
8. [pid:677] Planned Parenthood Federation of America URL: http://www.plannedparenthood.org	8. [pid:3] Abortion facts and information, statistics, hotlines and helpline URL: http://www.abortionfacts.com
9. [pid:3] Abortion facts and information, statistics, hotlines and helpline URL: http://www.abortionfacts.com	9. [pid:1259] Campaign Life Coalition British Columbia URL: http://www.clcbc.org
10. [pid:17] Abortion - Breast Cancer Link – Dr. Joel Brind Ph. D. URL: http://www.abortioncancer.com	10. [pid:21] Coalition on Abortion/Breast Cancer URL: http://www.abortionbreastcancer.com
11. [pid:1] NAF - The Voice of Abortion Providers URL: http://www.prochoice.org	11. [pid:300] Empty title field URL: http://www.heritagehouse76.com
12. [pid:1259] Campaign Life Coalition British Columbia URL: http://www.clcbc.org	12. [pid:38] Lifecall... Resources for Pregnant Woment and Their Babies URL: http://www.lifecall.org
13. [pid:300] Empty title field URL: http://www.heritagehouse76.com	13. [pid:1586] Americans United for Life URL: http://www.americansunitedforlife.org
14. [pid:2879] Pro-life news and information from American Life League URL: http://www.all.org	14. [pid:1590] The Justice Foundation's Home Page URL: http://justice.oct.net
15. [pid:22] Coalition on Abortion/Breast Cancer URL: http://www.abortionbreastcancer.com	15. [pid:137] John Kindley Home Page URL: http://www.johnkindley.com
Max	**SALSA**
1. [pid:4] prochoiceamerica.org : NARAL Pro-Choice America URL: http://www.naral.org	1. [pid:4] prochoiceamerica.org : NARAL Pro-Choice America URL: http://www.naral.org
2. [pid:677] Planned Parenthood Federation of America URL: http://www.plannedparenthood.org	2. [pid:1023] National Right to Life URL: http://www.nrlc.org
3. [pid:1023] National Right to Life URL: http://www.nrlc.org	3. [pid:677] Planned Parenthood Federation of America URL: http://www.plannedparenthood.org
4. [pid:1] NAF - The Voice of Abortion Providers URL: http://www.prochoice.org	4. [pid:1] NAF - The Voice of Abortion Providers URL: http://www.prochoice.org
5. [pid:1537] Priests for Life Index URL: http://www.priestsforlife.org	5. [pid:1537] Priests for Life Index URL: http://www.priestsforlife.org
6. [pid:1417] Pregnancy Centers Online URL: http://www.pregnancycenters.org	6. [pid:1417] Pregnancy Centers Online URL: http://www.pregnancycenters.org
7. [pid:1545] ProLifeInfo.org URL: http://www.prolifeinfo.org	7. [pid:1545] ProLifeInfo.org URL: http://www.prolifeinfo.org
8. [pid:2] Abortion Clinics OnLine URL: http://www.gynpages.com	8. [pid:7] After Abortion: Information on the aftereffects of abortion and p URL: http://www.afterabortion.org
9. [pid:7] After Abortion: Information on the aftereffects of abortion and p URL: http://www.afterabortion.org	9. [pid:2] Abortion Clinics OnLine URL: http://www.gynpages.com
10. [pid:232] FEMINIST MAJORITY FOUNDATION ONLINE HOMEPAGE URL: http://www.feminist.org	10. [pid:17] Abortion - Breast Cancer Link – Dr. Joel Brind Ph. D. URL: http://www.abortioncancer.com
11. [pid:259] The Alan Guttmacher Institute: Home Page URL: http://www.guttmacher.org	11. [pid:232] FEMINIST MAJORITY FOUNDATION ONLINE HOMEPAGE URL: http://www.feminist.org
12. [pid:227] center for reproductive rights URL: http://www.crlp.org	12. [pid:2879] Pro-life news and information from American Life League URL: http://www.all.org
13. [pid:1531] Human Life International URL: http://www.hli.org	13. [pid:3] Abortion facts and information, statistics, hotlines and helpline URL: http://www.abortionfacts.com
14. [pid:17] Abortion - Breast Cancer Link – Dr. Joel Brind Ph. D. URL: http://www.abortioncancer.com	14. [pid:1531] Human Life International URL: http://www.hli.org
15. [pid:1798] The Religious Coalition for Reproductive Choice URL: http://www.rcrc.org	15. [pid:227] center for reproductive rights URL: http://www.crlp.org

Table B.1: *Top 15 results for query "abortion"*

HITS	**HubAvg**
1. [pid:1] Affirmative Action and Diversity Page URL: http://aad.english.ucsb.edu	1. [pid:2026] Copyright Information URL: http://www.psu.edu/copyright.html
2. [pid:2] American Association for Affirmative Action URL: http://www.affirmativeaction.org	2. [pid:41] PSU Affirmative Action URL: http://www.psu.edu/dept/aaoffice
3. [pid:280] U.S. Equal Employment Opportunity Commission Home Page URL: http://www.eeoc.gov	3. [pid:1018] Welcome to Penn State's Home on the Web URL: http://www.psu.edu
4. [pid:316] National Organization for Women URL: http://www.now.org	4. [pid:1168] PSU Office for Disability Services URL: http://www.lions.psu.edu/ods
5. [pid:904] The United States Department of Labor Home Page, Secretary of Lab URL: http://www.dol.gov	5. [pid:966] University of Illinois URL: http://www.uiuc.edu
6. [pid:519] DiversityWeb - A Resource Hub for Higher Education URL: http://www.diversityweb.org	6. [pid:1457] Purdue University-West Lafayette, Indiana URL: http://www.purdue.edu
7. [pid:2278] Diversity Database, University of Maryland URL: http://www.inform.umd.edu/EdRes/Topic/Diversity	7. [pid:1169] University of Michigan URL: http://www.umich.edu
8. [pid:2381] Site Meter - Counter and Statistics Tracker URL: http://sm6.sitemeter.com/stats.asp?site=sm6wobbly123	8. [pid:1982] UC Berkeley home page URL: http://www.berkeley.edu
9. [pid:2382] Free web counter - Site access tracker - CQ Counter URL: http://cqcounter.com/?_id=nsnewman&_lo=us	9. [pid:820] The University of Arizona URL: http://www.arizona.edu
10. [pid:2383] Free web counter - Site access tracker - CQ Counter URL: http://cqcounter.com	10. [pid:1009] The University of Iowa Homepage URL: http://www.uiowa.edu
11. [pid:615] CIR Home URL: http://www.cir-usa.org	11. [pid:1427] Penn: University of Pennsylvania URL: http://www.upenn.edu
12. [pid:835] Empty title field URL: http://www.auaa.org	12. [pid:1250] Welcome to the University of Vermont URL: http://www.uvm.edu
13. [pid:616] civilrights.org – The Progressive Coalition for Equal Opportunit URL: http://www.civilrights.org	13. [pid:570] University of Colorado at Boulder URL: http://www.colorado.edu
14. [pid:887] Welcome to SHRM Online URL: http://www.shrm.org	14. [pid:1255] University of Oregon Home Page URL: http://www.uoregon.edu
15. [pid:796] Welcome to aadap.org. Here you'll find news updates on affirmati URL: http://www.aadap.org	15. [pid:1698] Stony Brook University URL: http://www.sunysb.edu

AT-Avg	**Norm (2)**
1. [pid:2026] Copyright Information URL: http://www.psu.edu/copyright.html	1. [pid:2026] Copyright Information URL: http://www.psu.edu/copyright.html
2. [pid:41] PSU Affirmative Action URL: http://www.psu.edu/dept/aaoffice	2. [pid:41] PSU Affirmative Action URL: http://www.psu.edu/dept/aaoffice
3. [pid:1018] Welcome to Penn State's Home on the Web URL: http://www.psu.edu	3. [pid:1018] Welcome to Penn State's Home on the Web URL: http://www.psu.edu
4. [pid:966] University of Illinois URL: http://www.uiuc.edu	4. [pid:966] University of Illinois URL: http://www.uiuc.edu
5. [pid:1457] Purdue University-West Lafayette, Indiana URL: http://www.purdue.edu	5. [pid:1457] Purdue University-West Lafayette, Indiana URL: http://www.purdue.edu
6. [pid:1982] UC Berkeley home page URL: http://www.berkeley.edu	6. [pid:1982] UC Berkeley home page URL: http://www.berkeley.edu
7. [pid:1169] University of Michigan URL: http://www.umich.edu	7. [pid:1169] University of Michigan URL: http://www.umich.edu
8. [pid:820] The University of Arizona URL: http://www.arizona.edu	8. [pid:820] The University of Arizona URL: http://www.arizona.edu
9. [pid:1009] The University of Iowa Homepage URL: http://www.uiowa.edu	9. [pid:1009] The University of Iowa Homepage URL: http://www.uiowa.edu
10. [pid:1427] Penn: University of Pennsylvania URL: http://www.upenn.edu	10. [pid:1427] Penn: University of Pennsylvania URL: http://www.upenn.edu
11. [pid:1250] Welcome to the University of Vermont URL: http://www.uvm.edu	11. [pid:1250] Welcome to the University of Vermont URL: http://www.uvm.edu
12. [pid:570] University of Colorado at Boulder URL: http://www.colorado.edu	12. [pid:570] University of Colorado at Boulder URL: http://www.colorado.edu
13. [pid:1255] University of Oregon Home Page URL: http://www.uoregon.edu	13. [pid:1255] University of Oregon Home Page URL: http://www.uoregon.edu
14. [pid:1698] Stony Brook University URL: http://www.sunysb.edu	14. [pid:1698] Stony Brook University URL: http://www.sunysb.edu
15. [pid:1918] Welcome to Ohio University URL: http://www.ohiou.edu	15. [pid:1918] Welcome to Ohio University URL: http://www.ohiou.edu

Max	**SALSA**
1. [pid:2026] Copyright Information URL: http://www.psu.edu/copyright.html	1. [pid:2026] Copyright Information URL: http://www.psu.edu/copyright.html
2. [pid:41] PSU Affirmative Action URL: http://www.psu.edu/dept/aaoffice	2. [pid:1] Affirmative Action and Diversity Page URL: http://aad.english.ucsb.edu
3. [pid:1018] Welcome to Penn State's Home on the Web URL: http://www.psu.edu	3. [pid:739] Adobe Acrobat Reader - Download URL: http://www.adobe.com/products/acrobat/readstep.html
4. [pid:966] University of Illinois URL: http://www.uiuc.edu	4. [pid:280] U.S. Equal Employment Opportunity Commission Home Page URL: http://www.eeoc.gov
5. [pid:1457] Purdue University-West Lafayette, Indiana URL: http://www.purdue.edu	5. [pid:2] American Association for Affirmative Action URL: http://www.affirmativeaction.org
6. [pid:1982] UC Berkeley home page URL: http://www.berkeley.edu	6. [pid:2381] Site Meter - Counter and Statistics Tracker URL: http://sm6.sitemeter.com/stats.asp?site=sm6wobbly123
7. [pid:1169] University of Michigan URL: http://www.umich.edu	7. [pid:2382] Free web counter - Site access tracker - CQ Counter URL: http://cqcounter.com/?_id=nsnewman&_ lo=us
8. [pid:820] The University of Arizona URL: http://www.arizona.edu	8. [pid:2383] Free web counter - Site access tracker - CQ Counter URL: http://cqcounter.com
9. [pid:1009] The University of Iowa Homepage URL: http://www.uiowa.edu	9. [pid:316] National Organization for Women URL: http://www.now.org
10. [pid:1427] Penn: University of Pennsylvania URL: http://www.upenn.edu	10. [pid:3] Affirmative Action Register URL: http://www.aar-eeo.com
11. [pid:1250] Welcome to the University of Vermont URL: http://www.uvm.edu	11. [pid:72] TEXT URL: http://www.eoaa.vt.edu
12. [pid:570] University of Colorado at Boulder URL: http://www.colorado.edu	12. [pid:7] CAA URL: http://www.caasf.org
13. [pid:1255] University of Oregon Home Page URL: http://www.uoregon.edu	13. [pid:1741] WIU - Division of Student Services URL: http://student.services.wiu.edu
14. [pid:1698] Stony Brook University URL: http://www.sunysb.edu	14. [pid:41] PSU Affirmative Action URL: http://www.psu.edu/dept/aaoffice
15. [pid:1918] Welcome to Ohio University URL: http://www.ohiou.edu	15. [pid:904] The United States Department of Labor Home Page, Secretary of Lab URL: http://www.dol.gov

Table B.2: *Top 15 results for query "affirmative action"*

HITS	**HubAvg**
1. [pid:1] NCADI: SAMHSA's The National Clearinghouse for Alcohol URL: http://www.health.org	1. [pid:285] The Substance Abuse and Mental Health Services Admin URL: http://www.samhsa.gov
2. [pid:2] National Institute on Alcohol Abuse and Alcoholism URL: http://www.niaaa.nih.gov	2. [pid:1] NCADI: SAMHSA's The National Clearinghouse for Alcohol URL: http://www.health.org
3. [pid:285] The Substance Abuse and Mental Health Services Admin URL: http://www.samhsa.gov	3. [pid:2] National Institute on Alcohol Abuse and Alcoholism URL: http://www.niaaa.nih.gov
4. [pid:1547] National Institute on Drug Abuse URL: http://www.nida.nih.gov	4. [pid:14] ETOH Home Page URL: http://etoh.niaaa.nih.gov
5. [pid:1540] Alcoholics Anonymous URL: http://www.alcoholics-anonymous.org	5. [pid:287] SAMHSA Web: Center for Substance Abuse Prevention (CSAP) URL: http://www.samhsa.gov/centers/csap/csap.html
6. [pid:2381] Join Together Online - Take Action Against Substance Abuse URL: http://www.jointogether.org	6. [pid:276] Facility Locator URL: http://findtreatment.samhsa.gov/facilitylocatordoc.htm
7. [pid:1209] National Council on Alcoholism and Drug Dependence - NCADD - figh URL: http://www.ncadd.org	7. [pid:288] SAMHSA Web: Center for Substance Abuse Treatment (CSAT) URL: http://www.samhsa.gov/centers/csat2002/csat_frame.html
8. [pid:1613] Welcome to the Office of National Drug Control Policy - ONDCP URL: http://www.whitehousedrugpolicy.gov	8. [pid:286] SAMHSA Web: Center for Mental Health Services (CMHS) URL: http://www.samhsa.gov/centers/cmhs/cmhs.html
9. [pid:58] Center on Alcohol Marketing and Youth URL: http://camy.org	9. [pid:1547] National Institute on Drug Abuse URL: http://www.nida.nih.gov
10. [pid:1538] Al-Anon/Alateen URL: http://www.al-anon.org	10. [pid:1540] Alcoholics Anonymous URL: http://www.alcoholics-anonymous.org
11. [pid:6] Higher Education Center for Alcohol and Other Drug Prevention URL: http://www.edc.org/hec	11. [pid:336] United States Department of Health and Human Services URL: http://www.os.dhhs.gov
12. [pid:2380] Partnership for a Drug-Free America® URL: http://www.drugfreeamerica.org	12. [pid:1744] Centers for Disease Control and Prevention URL: http://www.cdc.gov
13. [pid:1744] Centers for Disease Control and Prevention URL: http://www.cdc.gov	13. [pid:2381] Join Together Online - Take Action Against Substance Abuse and Gu URL: http://www.jointogether.org
14. [pid:1202] MADD Online: Home URL: http://www.madd.org	14. [pid:1613] Welcome to the Office of National Drug Control Policy - ONDCP URL: http://www.whitehousedrugpolicy.gov
15. [pid:1193] ASAM - American Society of Addiction Medicine Home Page URL: http://www.asam.org	15. [pid:160] Web of Addictions URL: http://www.well.com/user/woa

AT-Avg	**Norm (2)**
1. [pid:1] NCADI: SAMHSA's The National Clearinghouse for Alcohol URL: http://www.health.org	1. [pid:1] NCADI: SAMHSA's The National Clearinghouse for Alcohol URL: http://www.health.org
2. [pid:2] National Institute on Alcohol Abuse and Alcoholism URL: http://www.niaaa.nih.gov	2. [pid:2] National Institute on Alcohol Abuse and Alcoholism URL: http://www.niaaa.nih.gov
3. [pid:285] The Substance Abuse and Mental Health Services Administration SAM URL: http://www.samhsa.gov	3. [pid:285] The Substance Abuse and Mental Health Services Administration SAM URL: http://www.samhsa.gov
4. [pid:1547] National Institute on Drug Abuse URL: http://www.nida.nih.gov	4. [pid:1547] National Institute on Drug Abuse URL: http://www.nida.nih.gov
5. [pid:1540] Alcoholics Anonymous URL: http://www.alcoholics-anonymous.org	5. [pid:1540] Alcoholics Anonymous URL: http://www.alcoholics-anonymous.org
6. [pid:2381] Join Together Online - Take Action Against Substance Abuse URL: http://www.jointogether.org	6. [pid:2381] Join Together Online - Take Action Against Substance Abuse URL: http://www.jointogether.org
7. [pid:1209] National Council on Alcoholism and Drug Dependence - NCADD - figh URL: http://www.ncadd.org	7. [pid:1209] National Council on Alcoholism and Drug Dependence - NCADD - figh URL: http://www.ncadd.org
8. [pid:1613] Welcome to the Office of National Drug Control Policy - ONDCP URL: http://www.whitehousedrugpolicy.gov	8. [pid:1613] Welcome to the Office of National Drug Control Policy - ONDCP URL: http://www.whitehousedrugpolicy.gov
9. [pid:1744] Centers for Disease Control and Prevention URL: http://www.cdc.gov	9. [pid:58] Center on Alcohol Marketing and Youth URL: http://camy.org
10. [pid:58] Center on Alcohol Marketing and Youth URL: http://camy.org	10. [pid:1744] Centers for Disease Control and Prevention URL: http://www.cdc.gov
11. [pid:6] Higher Education Center for Alcohol and Other Drug Prevention URL: http://www.edc.org/hec	11. [pid:6] Higher Education Center for Alcohol and Other Drug Prevention URL: http://www.edc.org/hec
12. [pid:1538] Al-Anon/Alateen URL: http://www.al-anon.org	12. [pid:1538] Al-Anon/Alateen URL: http://www.al-anon.org
13. [pid:160] Web of Addictions URL: http://www.well.com/user/woa	13. [pid:2380] Partnership for a Drug-Free America® URL: http://www.drugfreeamerica.org
14. [pid:2380] Partnership for a Drug-Free America® URL: http://www.drugfreeamerica.org	14. [pid:160] Web of Addictions URL: http://www.well.com/user/woa
15. [pid:336] United States Department of Health and Human Services URL: http://www.os.dhhs.gov	15. [pid:336] United States Department of Health and Human Services URL: http://www.os.dhhs.gov

Max	**SALSA**
1. [pid:1] NCADI: SAMHSA's The National Clearinghouse for Alcohol URL: http://www.health.org	1. [pid:1] NCADI: SAMHSA's The National Clearinghouse for Alcohol URL: http://www.health.org
2. [pid:2] National Institute on Alcohol Abuse and Alcoholism URL: http://www.niaaa.nih.gov	2. [pid:2] National Institute on Alcohol Abuse and Alcoholism URL: http://www.niaaa.nih.gov
3. [pid:285] The Substance Abuse and Mental Health Services Admin URL: http://www.samhsa.gov	3. [pid:285] The Substance Abuse and Mental Health Services Admin URL: http://www.samhsa.gov
4. [pid:1547] National Institute on Drug Abuse URL: http://www.nida.nih.gov	4. [pid:1547] National Institute on Drug Abuse URL: http://www.nida.nih.gov
5. [pid:1540] Alcoholics Anonymous URL: http://www.alcoholics-anonymous.org	5. [pid:1540] Alcoholics Anonymous URL: http://www.alcoholics-anonymous.org
6. [pid:2381] Join Together Online - Take Action Against Substance Abuse and Gu URL: http://www.jointogether.org	6. [pid:34] Welcome to APOLNET URL: http://www.apolnet.org
7. [pid:1209] National Council on Alcoholism and Drug Dependence - NCADD - figh URL: http://www.ncadd.org	7. [pid:1744] Centers for Disease Control and Prevention URL: http://www.cdc.gov
8. [pid:1744] Centers for Disease Control and Prevention URL: http://www.cdc.gov	8. [pid:14] ETOH Home Page URL: http://etoh.niaaa.nih.gov
9. [pid:1613] Welcome to the Office of National Drug Control Policy - ONDCP URL: http://www.whitehousedrugpolicy.gov	9. [pid:58] Center on Alcohol Marketing and Youth URL: http://camy.org
10. [pid:6] Higher Education Center for Alcohol and Other Drug Prevention URL: http://www.edc.org/hec	10. [pid:336] United States Department of Health and Human Services URL: http://www.os.dhhs.gov
11. [pid:14] ETOH Home Page URL: http://etoh.niaaa.nih.gov	11. [pid:2381] Join Together Online - Take Action Against Substance Abuse URL: http://www.jointogether.org
12. [pid:58] Center on Alcohol Marketing and Youth URL: http://camy.org	12. [pid:6] Higher Education Center for Alcohol and Other Drug Prevention URL: http://www.edc.org/hec
13. [pid:160] Web of Addictions URL: http://www.well.com/user/woa	13. [pid:287] SAMHSA Web: Center for Substance Abuse Prevention (CSAP) URL: http://www.samhsa.gov/centers/csap/csap.html
14. [pid:1538] Al-Anon/Alateen URL: http://www.al-anon.org	14. [pid:2866] College Drinking: Changing the Culture URL: http://www.collegedrinkingprevention.gov
15. [pid:336] United States Department of Health and Human Services URL: http://www.os.dhhs.gov	15. [pid:1209] National Council on Alcoholism and Drug Dependence - NCADD - figh URL: http://www.ncadd.org

Table B.3: *Top 15 results for query "alcohol"*

HITS	**HubAvg**
1. [pid:2025] Welcome to RENOLDI rides & parts Inc. URL: http://www.renoldi.com	1. [pid:1438] HONcode: Principles URL: http://www.hon.ch/HONcode/Conduct.html?HONConduct151253
2. [pid:1] IAAPA URL: http://www.iaapa.org	2. [pid:1679] AttorneyPages Helps You Find the Best Attorney, Lawyer or Law Fir URL: http://attorneypages.com
3. [pid:708] Knott's URL: http://www.knotts.com	3. [pid:1680] Do It Yourself Home Improvement, Repair, Remodeling and Hardware URL: http://doityourself.com
4. [pid:2484] Traditional Amusement parks of the past and present..... URL: http://www.tradition.cjb.net	4. [pid:1678] Free Legal Advice in 100+ Law Topics - Law Attorney URL: http://freeadvice.com
5. [pid:1987] HUSS URL: http://www.hussrides.com	5. [pid:1676] ExpertPages.com - Books, Tapes and Seminars for Experts URL: http://expert-pages.com/books.htm
6. [pid:467] Empty title field URL: http://www.aimsintl.org	6. [pid:1668] Accidents Happen - Why are Lawyers involved? URL: http://law.freeadvice.com/resources/contact_us.htm
7. [pid:2436] Screamscape URL: http://www.screamscape.com	7. [pid:885] Empty title field URL: http://imgserv.adbutler.com/go2/;ID=129392;size=120x60;setID=4545
8. [pid:2026] REVERCHON : HOME PAGE URL: http://www.reverchon.com	8. [pid:955] Empty title field URL: http://imgserv.adbutler.com/go2/;ID=129392;size=468x60;setID=4970
9. [pid:2053] Empty title field URL: http://www.zierer.com	9. [pid:956] Site Meter - Counter and Statistics Tracker URL: http://s10.sitemeter.com/stats.asp?site=s10preschoolcoloringbook
10. [pid:2116] DE URL: http://www.drewexpo.com	10. [pid:887] Empty title field URL: http://imgserv.adbutler.com/go2/;ID=129392;size=125x125;setID=497
11. [pid:835] Joyrides - Amusement Park and Roller Coaster Photos URL: http://www.joyrides.com	11. [pid:1600] Expert Witness Directory — Forensic, Technical, Investigative URL: http://expertpages.com
12. [pid:802] Schlitterbahn Waterparks - The 2 Hottest Coolest Times in Texas! URL: http://www.schlitterbahn.com	12. [pid:1094] Get Your Discount Card Today URL: http://www.usaphonetime.com
13. [pid:3241] Great Adventure Source URL: http://greatadventure.8m.com	13. [pid:1669] The Expert Pages - About Advice & Counsel Corp. URL: http://expertpages.com/about.htm
14. [pid:2312] Amusement Business.com URL: http://www.amusementbusiness.com	14. [pid:1670] Terms & Conditions at ExpertPages.com URL: http://expertpages.com/conditions.htm
15. [pid:1959] ChanceMorgan/CRM Portal Page URL: http://www.chancemorgan.com	15. [pid:1675] Expert Pages Privacy Policy URL: http://expertpages.com/privacy.htm
AT-Avg	**Norm (2)**
1. [pid:1679] AttorneyPages Helps You Find the Best Attorney, Lawyer or Law Fir URL: http://attorneypages.com	1. [pid:1679] AttorneyPages Helps You Find the Best Attorney, Lawyer or Law Fir URL: http://attorneypages.com
2. [pid:1680] Do It Yourself Home Improvement, Repair, Remodeling and Hardware URL: http://doityourself.com	2. [pid:1680] Do It Yourself Home Improvement, Repair, Remodeling and Hardware URL: http://doityourself.com
3. [pid:1678] Free Legal Advice in 100+ Law Topics - Law Attorney URL: http://freeadvice.com	3. [pid:1678] Free Legal Advice in 100+ Law Topics - Law Attorney URL: http://freeadvice.com
4. [pid:1676] ExpertPages.com - Books, Tapes and Seminars for Experts URL: http://expert-pages.com/books.htm	4. [pid:1676] ExpertPages.com - Books, Tapes and Seminars for Experts URL: http://expert-pages.com/books.htm
5. [pid:1668] Accidents Happen - Why are Lawyers involved? URL: http://law.freeadvice.com/resources/contact_us.htm	5. [pid:1668] Accidents Happen - Why are Lawyers involved? URL: http://law.freeadvice.com/resources/contact_us.htm
6. [pid:1600] Expert Witness Directory — Forensic, Technical, Investigative URL: http://expertpages.com	6. [pid:1600] Expert Witness Directory — Forensic, Technical, Investigative URL: http://expertpages.com
7. [pid:1094] Get Your Discount Card Today URL: http://www.usaphonetime.com	7. [pid:1094] Get Your Discount Card Today URL: http://www.usaphonetime.com
8. [pid:1669] The Expert Pages - About Advice & Counsel Corp. URL: http://expertpages.com/about.htm	8. [pid:374] MapQuest: Home URL: http://www.mapquest.com
9. [pid:1670] Terms & Conditions at ExpertPages.com URL: http://expertpages.com/conditions.htm	9. [pid:1669] The Expert Pages - About Advice & Counsel Corp. URL: http://expertpages.com/about.htm
10. [pid:1675] Expert Pages Privacy Policy URL: http://expertpages.com/privacy.htm	10. [pid:1670] Terms & Conditions at ExpertPages.com URL: http://expertpages.com/conditions.htm
11. [pid:374] MapQuest: Home URL: http://www.mapquest.com	11. [pid:1675] Expert Pages Privacy Policy URL: http://expertpages.com/privacy.htm
12. [pid:386] Adventure travel & outdoor recreation from Outside Magazine: trav URL: http://www.outsidemag.com	12. [pid:386] Adventure travel & outdoor recreation from Outside Magazine: trav URL: http://www.outsidemag.com
13. [pid:695] Disneyland Resort - The official Web site for the Disneyland Reso URL: http://www.disneyland.com	13. [pid:695] Disneyland Resort - The official Web site for the Disneyland Reso URL: http://www.disneyland.com
14. [pid:3] Theme Parks- Info and reviews about theme parks and amusement par URL: http://themeparks.about.com	14. [pid:3] Theme Parks- Info and reviews about theme parks and amusement par URL: http://themeparks.about.com
15. [pid:861] National Park Service - Experience Your America URL: http://www.nps.gov	15. [pid:815] Disney Online - The Official Home Page of The Walt Disney Company URL: http://www.disney.com
Max	**SALSA**
1. [pid:1438] HONcode: Principles URL: http://www.hon.ch/HONcode/Conduct.html?HONConduct151253	1. [pid:4] Empty title field URL: http://www.sixflags.com
2. [pid:4] Empty title field URL: http://www.sixflags.com	2. [pid:59] Busch Gardens Adventure Parks URL: http://www.buschgardens.com
3. [pid:59] Busch Gardens Adventure Parks URL: http://www.buschgardens.com	3. [pid:1] IAAPA URL: http://www.iaapa.org
4. [pid:27] Cedar Point Amusement Park, The Roller Coaster Capital of the Wor URL: http://www.cedarpoint.com	4. [pid:1678] Free Legal Advice in 100+ Law Topics - Law Attorney URL: http://freeadvice.com
5. [pid:1] IAAPA URL: http://www.iaapa.org	5. [pid:1679] AttorneyPages Helps You Find the Best Attorney, Lawyer or Law Fir URL: http://attorneypages.com
6. [pid:708] Knott's URL: http://www.knotts.com	6. [pid:1680] Do It Yourself Home Improvement, Repair, Remodeling and Hardware URL: http://doityourself.com
7. [pid:379] Universal Studios URL: http://www.usf.com	7. [pid:1676] ExpertPages.com - Books, Tapes and Seminars for Experts URL: http://expert-pages.com/books.htm
8. [pid:2025] Welcome to RENOLDI rides & parts Inc. URL: http://www.renoldi.com	8. [pid:1668] Accidents Happen - Why are Lawyers involved? URL: http://law.freeadvice.com/resources/contact_us.htm
9. [pid:808] Kennywood : America's Finest Traditional Amusement Park in Pittsb URL: http://www.kennywood.com	9. [pid:5] Exhibits Collection – Amusement Park Physics URL: http://www.learner.org/exhibits/parkphysics
10. [pid:5] Exhibits Collection – Amusement Park Physics URL: http://www.learner.org/exhibits/parkphysics	10. [pid:27] Cedar Point Amusement Park, The Roller Coaster Capital of the Wor URL: http://www.cedarpoint.com
11. [pid:695] Disneyland Resort - The official Web site for the Disneyland Reso URL: http://www.disneyland.com	11. [pid:415] Microsoft bCentral URL: http://www.linkexchange.com
12. [pid:793] Paramount Parks :: Kings Island URL: http://www.pki.com	12. [pid:632] FreeFind Site Search URL: http://search.freefind.com/find.html?id=7081443
13. [pid:503] Beachboardwalk.com: California's Seaside Amusement Park URL: http://www.beachboardwalk.com	13. [pid:374] MapQuest: Home URL: http://www.mapquest.com
14. [pid:1817] LEGO.com Plug-in Download URL: http://www.legolandca.com	14. [pid:695] Disneyland Resort - The official Web site for the Disneyland Reso URL: http://www.disneyland.com
15. [pid:11] FunGuide - the Internet Directory of Fun Places URL: http://www.funguide.com	15. [pid:2025] Welcome to RENOLDI rides & parts Inc. URL: http://www.renoldi.com

Table B.4: *Top 15 results for query "amusement parks"*

HITS	**HubAvg**
1. [pid:3522] All Conferences . Com URL: http://www.allconferences.com	1. [pid:3522] All Conferences . Com URL: http://www.allconferences.com
2. [pid:3514] Castles of the World Tours URL: http://www.castlesoftheworld.com	2. [pid:3514] Castles of the World Tours URL: http://www.castlesoftheworld.com
3. [pid:3559] Affordable Globus Tours 11% Discount on Globus and Cosmos URL: http://www.affordableglobustours.com	3. [pid:3520] Crosses.org URL: http://www.crosses.org
4. [pid:3561] Trafalgar Tours: 12% Discount on Trafalgar Tour Prices - Trafalga URL: http://www.affordableTrafalgartours.com	4. [pid:3515] Castles Hotels URL: http://www.castles-hotels.com
5. [pid:3562] Contiki Tours: 5% discount on Contiki Tours - Contiki URL: http://www.affordableContikitours.com	5. [pid:3516] Castles For Sale URL: http://www.castles-for-sale.com
6. [pid:3563] Ireland Tours: 5% Off CIE Tours. URL: http://www.affordableIrelandtours.com	6. [pid:3517] Past Tours from Castles of the World URL: http://www.castlesoftheworld.com/PastTours
7. [pid:3570] Collette Vacations: 5% discount on Collette Vacations URL: http://www.affordablecollettetours.com	7. [pid:3559] Affordable Globus Tours 11% Discount on Globus and Cosmos URL: http://www.affordableglobustours.com
8. [pid:3557] Affordable Resorts - Discounts on: Apple Vacations, Club Med, San URL: http://www.affordableresorts.com	8. [pid:3561] Trafalgar Tours: 12% Discount on Trafalgar Tour Prices - Trafalga URL: http://www.affordableTrafalgartours.com
9. [pid:3574] Club Med Resorts- Discounted Club Med Vacations URL: http://www.affordableclubmedresorts.com	9. [pid:3562] Contiki Tours: 5% discount on Contiki Tours - Contiki URL: http://www.affordableContikitours.com
10. [pid:3579] Disney Vacation - Disney Vacations - Disney Resort URL: http://www.affordabledisneyresorts.com	10. [pid:3563] Ireland Tours: 5% Off CIE Tours. URL: http://www.affordableIrelandtours.com
11. [pid:3581] Breezes Superclub - Breezes Superclubs URL: http://www.affordablesuperclubsresorts.com	11. [pid:3570] Collette Vacations: 5% discount on Collette Vacations URL: http://www.affordablecollettetours.com
12. [pid:3564] General Tours: 10% discount on General Tours - General Tour. URL: http://www.affordablegeneraltours.com	12. [pid:3525] Affordable Tours - Discounts on: Globus Tours, Trafalgar Tours, C URL: http://www.AffordableTours.com
13. [pid:3525] Affordable Tours - Discounts on: Globus Tours, Trafalgar Tours, C URL: http://www.AffordableTours.com	13. [pid:3564] General Tours: 10% discount on General Tours - General Tour. URL: http://www.affordablegeneraltours.com
14. [pid:3558] Affordable Cruises Web: Celebrity Cruises, Crystal Cruises, Holla URL: http://www.affordablecruisesweb.com	14. [pid:3557] Affordable Resorts - Discounts on: Apple Vacations, Club Med, San URL: http://www.affordableresorts.com
15. [pid:3578] Air Jamaica Vacations - 5% Discount off any Air Jamaica URL: http://www.affordablejamaicavacations.com	15. [pid:3574] Club Med Resorts- Discounted Club Med Vacations URL: http://www.affordableclubmedresorts.com

AT-Avg	**Norm (2)**
1. [pid:3522] All Conferences . Com URL: http://www.allconferences.com	1. [pid:3522] All Conferences . Com URL: http://www.allconferences.com
2. [pid:3514] Castles of the World Tours URL: http://www.castlesoftheworld.com	2. [pid:3514] Castles of the World Tours URL: http://www.castlesoftheworld.com
3. [pid:3520] Crosses.org URL: http://www.crosses.org	3. [pid:3559] Affordable Globus Tours 11% Discount on Globus and Cosmos URL: http://www.affordableglobustours.com
4. [pid:3559] Affordable Globus Tours 11% Discount on Globus and Cosmos URL: http://www.affordableglobustours.com	4. [pid:3561] Trafalgar Tours: 12% Discount on Trafalgar Tour Prices - Trafalga URL: http://www.affordableTrafalgartours.com
5. [pid:3570] Collette Vacations: 5% discount on Collette Vacations URL: http://www.affordablecollettetours.com	5. [pid:3562] Contiki Tours: 5% discount on Contiki Tours URL: http://www.affordableContikitours.com
6. [pid:3561] Trafalgar Tours: 12% Discount on Trafalgar Tour Prices - Trafalga URL: http://www.affordableTrafalgartours.com	6. [pid:3563] Ireland Tours: 5% Off CIE Tours. URL: http://www.affordableIrelandtours.com
7. [pid:3562] Contiki Tours: 5% discount on Contiki Tours URL: http://www.affordableContikitours.com	7. [pid:3570] Collette Vacations: 5% discount on Collette Vacations URL: http://www.affordablecollettetours.com
8. [pid:3563] Ireland Tours: 5% Off CIE Tours. URL: http://www.affordableIrelandtours.com	8. [pid:3557] Affordable Resorts - Discounts on: Apple Vacations, Club Med, San URL: http://www.affordableresorts.com
9. [pid:3525] Affordable Tours - Discounts on: Globus Tours, Trafalgar Tours, C URL: http://www.AffordableTours.com	9. [pid:3574] Club Med Resorts- Discounted Club Med Vacations - Club Med Vacati URL: http://www.affordableclubmedresorts.com
10. [pid:3557] Affordable Resorts - Discounts on: Apple Vacations, Club Med, San URL: http://www.affordableresorts.com	10. [pid:3579] Disney Vacation - Disney Vacations - Disney Resort - Disney Resor URL: http://www.affordabledisneyresorts.com
11. [pid:3564] General Tours: 10% discount on General Tours - General Tour. URL: http://www.affordablegeneraltours.com	11. [pid:3581] Breezes Superclub - Breezes Superclubs URL: http://www.affordablesuperclubsresorts.com
12. [pid:3574] Club Med Resorts- Discounted Club Med Vacations - Club Med Vacati URL: http://www.affordableclubmedresorts.com	12. [pid:3564] General Tours: 10% discount on General Tours - General Tour. URL: http://www.affordablegeneraltours.com
13. [pid:3579] Disney Vacation - Disney Vacations - Disney Resort URL: http://www.affordabledisneyresorts.com	13. [pid:3525] Affordable Tours - Discounts on: Globus Tours, Trafalgar ... URL: http://www.AffordableTours.com
14. [pid:3581] Breezes Superclub - Breezes Superclubs URL: http://www.affordablesuperclubsresorts.com	14. [pid:3558] Affordable Cruises Web:..., Crystal Cruises, Holla URL: http://www.affordablecruisesweb.com
15. [pid:3558] Affordable Cruises Web: Celebrity Cruises, Crystal Cruises, Holla URL: http://www.affordablecruisesweb.com	15. [pid:3578] Air Jamaica Vacations - 5% Discount off any Air Jamaica Vacation URL: http://www.affordablejamaicavacations.com

Max	**SALSA**
1. [pid:5251] - - totemweb.com URL: http://www.totemweb.com	1. [pid:5251] - - totemweb.com URL: http://www.totemweb.com
2. [pid:50] Architecture Design Images History 3D Models and more - Artifice URL: http://www.greatbuildings.com	2. [pid:50] Architecture Design Images History 3D Models and more - Artifice URL: http://www.greatbuildings.com
3. [pid:177] Empty title field URL: http://www.aia.org	3. [pid:5523] Google URL: http://www.google.com
4. [pid:5523] Google URL: http://www.google.com	4. [pid:177] Empty title field URL: http://www.aia.org
5. [pid:1001] e-Architect URL: http://www.e-architect.com	5. [pid:11] ADAM, the Art, Design, Architecture & ... URL: http://adam.ac.uk
6. [pid:11] ADAM, the Art, Design, Architecture & ... URL: http://adam.ac.uk	6. [pid:1] Architecture.com URL: http://www.architecture.com
7. [pid:6736] ReSources Home Page URL: http://www.resources.com	7. [pid:1001] e-Architect URL: http://www.e-architect.com
8. [pid:1] Architecture.com URL: http://www.architecture.com	8. [pid:118] Architecture Centre Bristol URL: http://www.arch-centre.demon.co.uk
9. [pid:1221] What You Need to Know About^{tm} URL: http://www.about.com	9. [pid:6736] ReSources Home Page URL: http://www.resources.com
10. [pid:118] Architecture Centre Bristol URL: http://www.arch-centre.demon.co.uk	10. [pid:113] The Institute of Classical Architecture URL: http://www.classicist.org
11. [pid:113] The Institute of Classical Architecture URL: http://www.classicist.org	11. [pid:4671] Fine Art - World Wide Arts Resources - ... URL: http://wwar.com
12. [pid:3162] Metropolis Magazine URL: http://www.metropolismag.com	12. [pid:8] CCA URL: http://cca.qc.ca
13. [pid:4671] Fine Art - World Wide Arts Resources - Contemporary... URL: http://wwar.com	13. [pid:23] Archeire - Irish Architecture Online - ... URL: http://www.archeire.com
14. [pid:16] ArtServe at the Australian National University URL: http://rubens.anu.edu.au	14. [pid:3522] All Conferences . Com URL: http://www.allconferences.com
15. [pid:7] Architecture Web Resources URL: http://library.nevada.edu/arch/rsrce/webrsrce/contents.html	15. [pid:1221] What You Need to Know About^{tm} URL: http://www.about.com

Table B.5: *Top 15 results for query "architecture"*

HITS	**HubAvg**
1. [pid:1450] Town Hall Book Service URL: http://www.thbookservice.com	1. [pid:1450] Town Hall Book Service URL: http://www.thbookservice.com
2. [pid:1451] World Vision URL: http://etools.ncol.com/a/jgroup/bg_worldvision_townhall_52.html	2. [pid:1451] World Vision URL: http://etools.ncol.com/a/jgroup/bg_worldvision_townhall_52.html
3. [pid:1452] Town Hall - Letter from David Limbaugh URL: http://cf.heritage.org/rd.cfm?id=36	3. [pid:1452] Town Hall - Letter from David Limbaugh URL: http://cf.heritage.org/rd.cfm?id=36
4. [pid:1453] World Vision URL: http://etools.ncol.com/a/jgroup/bg_worldvision_townhall468x60_65.	4. [pid:1453] World Vision URL: http://etools.ncol.com/a/jgroup/bg_worldvision_townhall468x60_65.
5. [pid:1454] Patterson, Colonel Robert: Dereliction of Duty URL: http://www.thbookservice.com/BookPage.asp?prod_cd=C6153	5. [pid:1454] Patterson, Colonel Robert: Dereliction of Duty URL: http://www.thbookservice.com/BookPage.asp?prod_cd=C6153
6. [pid:1455] Welcome to the Alexa Toolbar Download URL: http://cf.heritage.org/rd.cfm?id=286	6. [pid:1455] Welcome to the Alexa Toolbar Download URL: http://cf.heritage.org/rd.cfm?id=286
7. [pid:2426] Amazon.com: Books: Letters to a Young Victim: Hope and Healing in URL: http://www.amazon.com/. . . /townhall	7. [pid:2426] Amazon.com: Books: Letters to a Young Victim: Hope and Healing in URL: http://www.amazon.com/. . . /townhall
8. [pid:1615] Amazon.com: Books: Letters to a Young Victim: Hope and Healing in URL: http://www.amazon.com/exec/obidos/ASIN/0684824663/townhallcom	8. [pid:1615] Amazon.com: Books: Letters to a Young Victim: Hope and Healing in URL: http://www.amazon.com/exec/obidos/ASIN/0684824663/townhallcom
9. [pid:1616] Amazon.com: Books: Beyond Blame: How We Can Succeed by Breaking t URL: http://www.amazon.com/exec/obidos/ASIN/0029353653/townhallcom	9. [pid:1616] Amazon.com: Books: Beyond Blame: How We Can Succeed by Breaking t URL: http://www.amazon.com/exec/obidos/ASIN/0029353653/townhallcom
10. [pid:1449] Empty title field URL: http://www.townhall.com	10. [pid:1449] Empty title field URL: http://www.townhall.com
11. [pid:56] Townhall.com: Conservative Columnists: Armstrong Williams URL: http://www.townhall.com/columnists/armstrongwilliams/archive.shtm	11. [pid:56] Townhall.com: Conservative Columnists: Armstrong Williams URL: http://www.townhall.com/columnists/armstrongwilliams/archive.shtm
12. [pid:1276] SpaceWeather.com – News and information URL: http://leonids.com	12. [pid:1276] SpaceWeather.com – News and information URL: http://leonids.com
13. [pid:2595] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/00	13. [pid:57] Neil Armstrong URL: http://starchild.gsfc.nasa.gov/. . . /_level2/arm
14. [pid:2596] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/01	14. [pid:1277] Empty title field URL: http://www.nasa.gov
15. [pid:2597] Empty title field URL: http://g.msn.com/0nwenus0/AK/02	15. [pid:1275] Flight Journal magazine - airplane history, photos, technology URL: http://www.flightjournal.com
AT-Avg	**Norm (2)**
1. [pid:1450] Town Hall Book Service URL: http://www.thbookservice.com	1. [pid:1450] Town Hall Book Service URL: http://www.thbookservice.com
2. [pid:1451] World Vision URL: http://etools.ncol.com/a/jgroup/bg_worldvision_townhall_52.html	2. [pid:1451] World Vision URL: http://etools.ncol.com/a/jgroup/bg_worldvision_townhall_52.html
3. [pid:1452] Town Hall - Letter from David Limbaugh URL: http://cf.heritage.org/rd.cfm?id=36	3. [pid:1452] Town Hall - Letter from David Limbaugh URL: http://cf.heritage.org/rd.cfm?id=36
4. [pid:1453] World Vision URL: http://etools.ncol.com/a/jgroup/bg_worldvision_townhall468x60_65.	4. [pid:1453] World Vision URL: http://etools.ncol.com/a/jgroup/bg_worldvision_townhall468x60_65.
5. [pid:1454] Patterson, Colonel Robert: Dereliction of Duty URL: http://www.thbookservice.com/BookPage.asp?prod_cd=C6153	5. [pid:1454] Patterson, Colonel Robert: Dereliction of Duty URL: http://www.thbookservice.com/BookPage.asp?prod_cd=C6153
6. [pid:1455] Welcome to the Alexa Toolbar Download URL: http://cf.heritage.org/rd.cfm?id=286	6. [pid:1455] Welcome to the Alexa Toolbar Download URL: http://cf.heritage.org/rd.cfm?id=286
7. [pid:2426] Amazon.com: Books: Letters to a Young Victim: Hope and Healing in URL: http://www.amazon.com/. . . /townhall	7. [pid:2426] Amazon.com: Books: Letters to a Young Victim: Hope and Healing in URL: http://www.amazon.com/. . . /townhall
8. [pid:1615] Amazon.com: Books: Letters to a Young Victim: Hope and Healing in URL: http://www.amazon.com/exec/obidos/ASIN/0684824663/townhallcom	8. [pid:1615] Amazon.com: Books: Letters to a Young Victim: Hope and Healing in URL: http://www.amazon.com/exec/obidos/ASIN/0684824663/townhallcom
9. [pid:1616] Amazon.com: Books: Beyond Blame: How We Can Succeed by Breaking t URL: http://www.amazon.com/exec/obidos/ASIN/0029353653/townhallcom	9. [pid:1616] Amazon.com: Books: Beyond Blame: How We Can Succeed by Breaking t URL: http://www.amazon.com/exec/obidos/ASIN/0029353653/townhallcom
10. [pid:1449] Empty title field URL: http://www.townhall.com	10. [pid:1449] Empty title field URL: http://www.townhall.com
11. [pid:56] Townhall.com: Conservative Columnists: Armstrong Williams URL: http://www.townhall.com/columnists/armstrongwilliams/archive.shtm	11. [pid:56] Townhall.com: Conservative Columnists: Armstrong Williams URL: http://www.townhall.com/columnists/armstrongwilliams/archive.shtm
12. [pid:1276] SpaceWeather.com – News and information URL: http://leonids.com	12. [pid:1276] SpaceWeather.com – News and information URL: http://leonids.com
13. [pid:167] Herbert W. Armstrong Library and Archives URL: http://www.herbertwarmstrong.org	13. [pid:1277] Empty title field URL: http://www.nasa.gov
14. [pid:1277] Empty title field URL: http://www.nasa.gov	14. [pid:167] Herbert W. Armstrong Library and Archives URL: http://www.herbertwarmstrong.org
15. [pid:1789] Geist: Canadian Ideas, Canadian Culture URL: http://geist.com	15. [pid:1789] Geist: Canadian Ideas, Canadian Culture URL: http://geist.com
Max	**SALSA**
1. [pid:1450] Town Hall Book Service URL: http://www.thbookservice.com	1. [pid:1450] Town Hall Book Service URL: http://www.thbookservice.com
2. [pid:1451] World Vision URL: http://etools.ncol.com/a/jgroup/bg_worldvision_townhall_52.html	2. [pid:1451] World Vision URL: http://etools.ncol.com/a/jgroup/bg_worldvision_townhall_52.html
3. [pid:1452] Town Hall - Letter from David Limbaugh URL: http://cf.heritage.org/rd.cfm?id=36	3. [pid:1452] Town Hall - Letter from David Limbaugh URL: http://cf.heritage.org/rd.cfm?id=36
4. [pid:1453] World Vision URL: http://etools.ncol.com/a/jgroup/bg_worldvision_townhall468x60_65.	4. [pid:1453] World Vision URL: http://etools.ncol.com/a/jgroup/bg_worldvision_townhall468x60_65.
5. [pid:1454] Patterson, Colonel Robert: Dereliction of Duty URL: http://www.thbookservice.com/BookPage.asp?prod_cd=C6153	5. [pid:1454] Patterson, Colonel Robert: Dereliction of Duty URL: http://www.thbookservice.com/BookPage.asp?prod_cd=C6153
6. [pid:1455] Welcome to the Alexa Toolbar Download URL: http://cf.heritage.org/rd.cfm?id=286	6. [pid:1455] Welcome to the Alexa Toolbar Download URL: http://cf.heritage.org/rd.cfm?id=286
7. [pid:2426] Amazon.com: Books: Letters to a Young Victim URL: http://www.amazon.com/. . . /townhall	7. [pid:1897] FIETSEN TEGEN KANKER URL: http://www.fietsentegenkanker.org
8. [pid:1615] Amazon.com: Books: Letters to a Young Victim URL: http://www.amazon.com/exec/obidos/ASIN/0684824663/townhallcom	8. [pid:167] Herbert W. Armstrong Library and Archives URL: http://www.herbertwarmstrong.org
9. [pid:1616] Amazon.com: Books: Beyond Blame URL: http://www.amazon.com/exec/obidos/ASIN/0029353653/townhallcom	9. [pid:122] Biblical Evidence for Catholicism URL: http://www.biblicalcatholic.com
10. [pid:1449] Empty title field URL: http://www.townhall.com	10. [pid:129] R.V. Armstrong & Associates IS0 9000 Trainig Lead Auditor Inte URL: http://www.rvarmstrong.com
11. [pid:56] Townhall.com: Conservative Columnists: Armstrong Williams URL: http://www.townhall.com/columnists/armstrongwilliams/archive.shtm	11. [pid:116] CleanReg by Armstrong's Systems House, Inc. URL: http://www.cleanreg.com
12. [pid:1276] SpaceWeather.com – News and information URL: http://leonids.com	12. [pid:1789] Geist: Canadian Ideas, Canadian Culture URL: http://geist.com
13. [pid:167] Herbert W. Armstrong Library and Archives URL: http://www.herbertwarmstrong.org	13. [pid:2589] Visual Escapes - artist directory and gallery. Imaginative art URL: http://surrealities.cjb.net
14. [pid:1897] FIETSEN TEGEN KANKER URL: http://www.fietsentegenkanker.org	14. [pid:104] The Unofficial Lance Armstrong Fan Club URL: http://www.lancearmstrongfanclub.com
15. [pid:129] R.V. Armstrong & Associates IS0 9000 Trainig Lead Auditor URL: http://www.rvarmstrong.com	15. [pid:5] LAF URL: http://www.laf.org

Table B.6: *Top 15 results for query "armstrong"*

HITS	**HubAvg**
1. [pid:427] E Business Solutions,Website Promotion Services URL: http://www.intermesh.net/advertis.html	1. [pid:425] Travel - India Travel,Tourism In India,Travel to India URL: http://www.indiantravelportal.com
2. [pid:491] Empty title field URL: http://auto.indiamart.com	2. [pid:998] Empty title field URL: http://www.indiantravelportal.com/tajmahal
3. [pid:418] Empty title field URL: http://www.indiamart.com	3. [pid:983] Empty title field URL: http://www.indiantravelportal.com/indian-cities
4. [pid:482] Empty title field URL: http://www.indiangiftsportal.com	4. [pid:158] Empty title field URL: http://www.bombaymotor.com
5. [pid:483] Jewelry Box, Jewelry Gift Box, Jewelry Box Shopping Online URL: http://www.indiangiftsportal.com/india-shopping/exclusives/jewell	5. [pid:480] Adventure Tour Travel,India Adventure Travel,India Adventure URL: http://www.indiantravelportal.com/adventure
6. [pid:495] Birthday Gifts,Birthday Gift Idea,Send Birthday Gifts,Unique URL: http://www.indiangiftsportal.com/india-shopping/occasions/birthda	6. [pid:501] Himalayas,Himalaya,India Himalayas,Himalaya Trekking URL: http://www.indiantravelportal.com/himalayas
7. [pid:496] Anniversary Gifts,Wedding Anniversary Gift,Anniversary Gift URL: http://www.indiangiftsportal.com/india-shopping/occasions/anniver	7. [pid:984] Empty title field URL: http://www.indiantravelportal.com/trekking
8. [pid:497] Wedding Gifts,Wedding Anniversary Gift,Wedding Gift Idea URL: http://www.indiangiftsportal.com/india-shopping/occasions/wedding	8. [pid:499] Empty title field URL: http://www.indiantravelportal.com/fairs
9. [pid:498] Mixed Bag, Exclusives, Indian Gifts Portal URL: http://www.indiangiftsportal.com/india-shopping/exclusives/mixed-	9. [pid:500] Empty title field URL: http://www.indiantravelportal.com/festivals
10. [pid:502] Business Solutions,Ecommerce Business Solutions URL: http://www.intermesh.net	10. [pid:481] Tripura,Tripura India,Tourism in Tripura,Tripura India URL: http://www.indiantravelportal.com/tripura
11. [pid:449] Empty title field URL: http://handicraft.indiamart.com	11. [pid:167] Empty title field URL: http://www.windsorauto.com
12. [pid:453] Empty title field URL: http://apparel.indiamart.com	12. [pid:986] Darjeeling,Darjeeling India,Darjeeling Tourism,Darjiling URL: http://www.indiantravelportal.com/west-bengal/darjeeling
13. [pid:490] Empty title field URL: http://health.indiamart.com	13. [pid:477] India Mysore Travel, Historical Tours of Mysore URL: http://mysore.indiantravelportal.com/historic-mysore.html
14. [pid:492] India Finance and Investment Guide, India URL: http://finance.indiamart.com	14. [pid:605] KineticIndia URL: http://www.kineticindia.com
15. [pid:493] Empty title field URL: http://news.indiamart.com	15. [pid:991] Rajasthan Tours,Rajasthan Tour,Rajasthan India Tours URL: http://www.rajasthan-travel-tours.com/rajasthan-tours.html

AT-Avg	**Norm (2)**
1. [pid:425] Travel - India Travel,Tourism In India,Travel to India URL: http://www.indiantravelportal.com	1. [pid:425] Travel - India Travel,Tourism In India,Travel to India URL: http://www.indiantravelportal.com
2. [pid:998] Empty title field URL: http://www.indiantravelportal.com/tajmahal	2. [pid:998] Empty title field URL: http://www.indiantravelportal.com/tajmahal
3. [pid:158] Empty title field URL: http://www.bombaymotor.com	3. [pid:158] Empty title field URL: http://www.bombaymotor.com
4. [pid:983] Empty title field URL: http://www.indiantravelportal.com/indian-cities	4. [pid:983] Empty title field URL: http://www.indiantravelportal.com/indian-cities
5. [pid:480] Adventure Tour Travel,India Adventure Travel,India Adventure URL: http://www.indiantravelportal.com/adventure	5. [pid:480] Adventure Tour Travel,India Adventure Travel,India Adventure URL: http://www.indiantravelportal.com/adventure
6. [pid:501] Himalayas,Himalaya,India Himalayas,Himalaya Trekking URL: http://www.indiantravelportal.com/himalayas	6. [pid:501] Himalayas,Himalaya,India Himalayas,Himalaya Trekking URL: http://www.indiantravelportal.com/himalayas
7. [pid:984] Empty title field URL: http://www.indiantravelportal.com/trekking	7. [pid:984] Empty title field URL: http://www.indiantravelportal.com/trekking
8. [pid:499] Empty title field URL: http://www.indiantravelportal.com/fairs	8. [pid:499] Empty title field URL: http://www.indiantravelportal.com/fairs
9. [pid:500] Empty title field URL: http://www.indiantravelportal.com/festivals	9. [pid:500] Empty title field URL: http://www.indiantravelportal.com/festivals
10. [pid:481] Tripura,Tripura India,Tourism in Tripura,Tripura India URL: http://www.indiantravelportal.com/tripura	10. [pid:481] Tripura,Tripura India,Tourism in Tripura,Tripura India URL: http://www.indiantravelportal.com/tripura
11. [pid:986] Darjeeling,Darjeeling India,Darjeeling Tourism URL: http://www.indiantravelportal.com/west-bengal/darjeeling	11. [pid:986] Darjeeling,Darjeeling India,Darjeeling Tourism URL: http://www.indiantravelportal.com/west-bengal/darjeeling
12. [pid:991] Rajasthan Tours,Rajasthan Tour,Rajasthan India Tours,Rajasthan URL: http://www.rajasthan-travel-tours.com/rajasthan-tours.html	12. [pid:991] Rajasthan Tours,Rajasthan Tour,Rajasthan India Tours,Rajasthan URL: http://www.rajasthan-travel-tours.com/rajasthan-tours.html
13. [pid:477] India Mysore Travel, Historical Tours of Mysore URL: http://mysore.indiantravelportal.com/historic-mysore.html	13. [pid:477] India Mysore Travel, Historical Tours of Mysore URL: http://mysore.indiantravelportal.com/historic-mysore.html
14. [pid:982] Empty title field URL: http://www.indiantravelportal.com/zonesofindia/south.html	14. [pid:982] Empty title field URL: http://www.indiantravelportal.com/zonesofindia/south.html
15. [pid:985] Mumbai (Bombay) Travel Agents & Tour Operators in India URL: http://travel-agents.indiantravelportal.com/mumbai.html	15. [pid:985] Mumbai (Bombay) Travel Agents & Tour Operators in India URL: http://travel-agents.indiantravelportal.com/mumbai.html

Max	**SALSA**
1. [pid:425] Travel - India Travel,Tourism In India,Travel to India URL: http://www.indiantravelportal.com	1. [pid:425] Travel - India Travel,Tourism In India,Travel to India URL: http://www.indiantravelportal.com
2. [pid:998] Empty title field URL: http://www.indiantravelportal.com/tajmahal	2. [pid:998] Empty title field URL: http://www.indiantravelportal.com/tajmahal
3. [pid:158] Empty title field URL: http://www.bombaymotor.com	3. [pid:742] Government of Canada Site — Site du gouvernement du Canada URL: http://canada.gc.ca
4. [pid:983] Empty title field URL: http://www.indiantravelportal.com/indian-cities	4. [pid:158] Empty title field URL: http://www.bombaymotor.com
5. [pid:480] Adventure Tour Travel,India Adventure Travel URL: http://www.indiantravelportal.com/adventure	5. [pid:983] Empty title field URL: http://www.indiantravelportal.com/indian-cities
6. [pid:501] Himalayas,Himalaya,India Himalayas,Himalaya Trekking URL: http://www.indiantravelportal.com/himalayas	6. [pid:515] Mesurer et analyser l'audience d'un site web URL: http://www.xiti.com/xiti.asp?s=27855
7. [pid:984] Empty title field URL: http://www.indiantravelportal.com/trekking	7. [pid:517] HitBoxCentral - HitBox Central - Home URL: http://rd1.hitbox.com/rd?acct=WQ500719D8AF10FR0
8. [pid:499] Empty title field URL: http://www.indiantravelportal.com/fairs	8. [pid:516] Weborama leader européen de la mesure d'audience ... URL: http://www.weborama.com
9. [pid:500] Empty title field URL: http://www.indiantravelportal.com/festivals	9. [pid:1] Automotive Industries — Home URL: http://www.ai-online.com
10. [pid:481] Tripura,Tripura India,Tourism in Tripura,Tripura India URL: http://www.indiantravelportal.com/tripura	10. [pid:10] Empty title field URL: http://www.aiacanada.com
11. [pid:477] India Mysore Travel, Historical Tours of Mysore URL: http://mysore.indiantravelportal.com/historic-mysore.html	11. [pid:1130] The Ontario Neurotrauma Foundation (ONF) URL: http://www.onf.org
12. [pid:986] Darjeeling,Darjeeling India,Darjeeling Tourism URL: http://www.indiantravelportal.com/west-bengal/darjeeling	12. [pid:999] Introduction au site Web officiel du gouvernement du Canada URL: http://www.canada.gc.ca/main_f.html
13. [pid:991] Rajasthan Tours,Rajasthan Tour,Rajasthan India Tours,Rajasthan URL: http://www.rajasthan-travel-tours.com/rajasthan-tours.html	13. [pid:701] Empty title field URL: http://www.placementindia.com
14. [pid:167] Empty title field URL: http://www.windsorauto.com	14. [pid:703] Exporters India - Indian Exporters, Importers URL: http://www.exportersindia.com
15. [pid:20] Empty title field URL: http://www.suneetul.maharashtradirectory.com	15. [pid:699] Web Hosting, Web Development, Promotion Site Web URL: http://www.weblinkindia.net

Table B.7: *Top 15 results for query "automobile industries"*

HITS	**HubAvg**
1. [pid:1023] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/14	1. [pid:2266] Student Advantage Discount Card–Save money during college URL: http://www.studentadvantage.com
2. [pid:1142] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/07	2. [pid:3354] The University of North Carolina at Chapel Hill URL: http://www.unc.edu
3. [pid:1143] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/08	3. [pid:3367] University of North Carolina - Tickets - Official Athletic Site URL: http://www.mediateamlink.com/oas/unc
4. [pid:1144] Empty title field URL: http://g.msn.com/0nwenus0/AK/09	4. [pid:3401] UNC Rams Club URL: http://www.ramsclub.org/home/5805.asp
5. [pid:1145] MSN Search – More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/10	5. [pid:5631] Florida State University Varsity Club URL: http://www.fsuvarsityclub.org
6. [pid:1146] Welcome to MSN Shopping URL: http://g.msn.com/0nwenus0/AK/11	6. [pid:5626] www.seminole-boosters.com URL: http://www.seminole-boosters.com
7. [pid:1147] MSN Money - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/12	7. [pid:5636] Tallahassee Map URL: http://www.fsu.edu/Welcome/tallymaps/tallymap.html
8. [pid:1148] MSN People and Chat - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/13	8. [pid:2726] Welcome to Duke University Stores URL: http://www.dukestore.com
9. [pid:1016] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/00	9. [pid:2724] Empty title field URL: http://netstile.evenue.net/evenue/se/duke
10. [pid:1017] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/01	10. [pid:4173] University of Notre Dame URL: http://www.nd.edu
11. [pid:1018] Empty title field URL: http://g.msn.com/0nwenus0/AK/02	11. [pid:2085] NCAA Online URL: http://www.ncaa.org
12. [pid:1019] MSN Search – More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/03	12. [pid:4175] mail.und.com - Your Fighting Irish E-mail! URL: http://mail.und.com/email/scripts/useragreement.pl
13. [pid:1020] Welcome to MSN Shopping URL: http://g.msn.com/0nwenus0/AK/04	13. [pid:179] University of Kentucky Basketball Museum URL: http://www.ukbballmuseum.org
14. [pid:1021] MSN Money - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/05	14. [pid:4923] Welcome to the University of Maryland Terrapin Club URL: http://www.terrapinclub.com
15. [pid:1022] MSN People and Chat - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/06	15. [pid:4924] index URL: http://www.inform.umd.edu/CampusInfo/Departments/Athletics/OAC
AT-Avg	**Norm (2)**
1. [pid:1] NBA.com URL: http://www.nba.com	1. [pid:1] NBA.com URL: http://www.nba.com
2. [pid:1023] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/14	2. [pid:1023] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/14
3. [pid:1142] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/07	3. [pid:1142] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/07
4. [pid:1143] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/08	4. [pid:1143] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/08
5. [pid:1144] Empty title field URL: http://g.msn.com/0nwenus0/AK/09	5. [pid:1144] Empty title field URL: http://g.msn.com/0nwenus0/AK/09
6. [pid:1145] MSN Search – More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/10	6. [pid:1145] MSN Search – More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/10
7. [pid:1146] Welcome to MSN Shopping URL: http://g.msn.com/0nwenus0/AK/11	7. [pid:1146] Welcome to MSN Shopping URL: http://g.msn.com/0nwenus0/AK/11
8. [pid:1147] MSN Money - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/12	8. [pid:1147] MSN Money - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/12
9. [pid:1148] MSN People and Chat - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/13	9. [pid:1148] MSN People and Chat - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/13
10. [pid:1016] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/00	10. [pid:1016] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/00
11. [pid:1017] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/01	11. [pid:1017] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/01
12. [pid:1018] Empty title field URL: http://g.msn.com/0nwenus0/AK/02	12. [pid:1018] Empty title field URL: http://g.msn.com/0nwenus0/AK/02
13. [pid:1019] MSN Search – More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/03	13. [pid:1019] MSN Search – More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/03
14. [pid:1020] Welcome to MSN Shopping URL: http://g.msn.com/0nwenus0/AK/04	14. [pid:1020] Welcome to MSN Shopping URL: http://g.msn.com/0nwenus0/AK/04
15. [pid:1021] MSN Money - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/05	15. [pid:1021] MSN Money - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/05
Max	**SALSA**
1. [pid:1] NBA.com URL: http://www.nba.com	1. [pid:1] NBA.com URL: http://www.nba.com
2. [pid:1023] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/14	2. [pid:2266] Student Advantage Discount Card–Save money during college URL: http://www.studentadvantage.com
3. [pid:1142] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/07	3. [pid:5] FIBA - International Basketball Federation URL: http://www.fiba.com
4. [pid:1143] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/08	4. [pid:1023] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/14
5. [pid:1144] Empty title field URL: http://g.msn.com/0nwenus0/AK/09	5. [pid:1142] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/07
6. [pid:1145] MSN Search – More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/10	6. [pid:1143] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/08
7. [pid:1146] Welcome to MSN Shopping URL: http://g.msn.com/0nwenus0/AK/11	7. [pid:1144] Empty title field URL: http://g.msn.com/0nwenus0/AK/09
8. [pid:1147] MSN Money - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/12	8. [pid:1145] MSN Search – More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/10
9. [pid:1148] MSN People and Chat - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/13	9. [pid:1146] Welcome to MSN Shopping URL: http://g.msn.com/0nwenus0/AK/11
10. [pid:1016] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/00	10. [pid:1147] MSN Money - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/12
11. [pid:1017] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/01	11. [pid:1148] MSN People and Chat - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/13
12. [pid:1018] Empty title field URL: http://g.msn.com/0nwenus0/AK/02	12. [pid:1016] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/00
13. [pid:1019] MSN Search – More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/03	13. [pid:1017] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/01
14. [pid:1020] Welcome to MSN Shopping URL: http://g.msn.com/0nwenus0/AK/04	14. [pid:1018] Empty title field URL: http://g.msn.com/0nwenus0/AK/02
15. [pid:1021] MSN Money - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/05	15. [pid:1019] MSN Search – More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/03

Table B.8: *Top 15 results for query "basketball"*

HITS	HubAvg
1. [pid:2] The Blues Foundation. Your Home for Blues, Blues, Blues! URL: http://www.blues.org	1. [pid:2] The Blues Foundation. Your Home for Blues, Blues, Blues! URL: http://www.blues.org
2. [pid:2755] WebRing: addsite_login URL: http://l.webring.com/wrman?ring=bluessociety&addsite	2. [pid:27] Harry's Blues Lyrics Online, Home Page, US URL: http://blueslinks.tripod.com
3. [pid:153] A Cyber Blues Society for musicians, fans, and webmasters of the URL: http://www.bluessociety.net	3. [pid:9] The Blue Highway URL: http://www.thebluehighway.com
4. [pid:4042] Google Search: URL: http://www.google.com/search	4. [pid:101] Delta Blues - DeltaBlues - deltablues.com URL: http://www.deltablues.com
5. [pid:4519] BluesSociety.net - for musicians, fans, and webmasters of the blu URL: http://sitebuilder.bluessociety.net	5. [pid:887] home URL: http://www.fargobluesfest.com
6. [pid:4561] Your Mailinglist Provider URL: http://www.yourmailinglistprovider.com/	6. [pid:3460] stantonanderson.com URL: http://www.stantonanderson.com
7. [pid:4464] BluesSociety.net - Free Email Service. URL: http://mail.bluessociety.net	7. [pid:2077] Dan's Police Page URL: http://danspolice.8m.com
8. [pid:1915] A Cyber Blues Society Blues Links Page 3 URL: http://www.bluessociety.net/links03.html	8. [pid:13] Blues On Stage, your complete blues guide. Twin Cities Blues - M URL: http://www.mnblues.com
9. [pid:3973] A Cyber Blues Society Blues Links Page 5 URL: http://www.bluessociety.net/links05.html	9. [pid:153] A Cyber Blues Society for musicians, fans, and webmasters of the URL: http://www.bluessociety.net
10. [pid:4502] Blues Biographies - Artist of the Blues - www.bluessociety.net URL: http://www.bluessociety.net/greats.html	10. [pid:4] BLUES WORLD URL: http://www.bluesworld.com
11. [pid:4564] Add your website to The Cyber Blues Society Reciprocal Links Page URL: http://www.bluessociety.net/links04.html	11. [pid:2378] Empty title field URL: http://www.letthegoodtimesroll.com
12. [pid:4565] BluesSociety.net - Community Message Center URL: http://community.bluessociety.net/commun_v3/scripts/directory.pl	12. [pid:6] BluesWEB URL: http://www.island.net/ blues
13. [pid:4566] The Cyber Blues Society Newsletter sign up page. URL: http://www.bluessociety.net/newsletter.html	13. [pid:185] Natchel' Blues Network, Hampton Roads, VA URL: http://www.natchelblues.org
14. [pid:4567] A Cyber Blues Society Blues Links Page URL: http://www.bluessociety.net/links01.html	14. [pid:117] ROOSTER BLUES URL: http://www.roosterblues.com
15. [pid:4568] BLUES LINKS - The Cyber Blues Society Links page 2... URL: http://www.bluessociety.net/links02.html	15. [pid:2183] MO BLUES.ORG URL: http://www.moblues.org

AT-Avg	Norm (2)
1. [pid:2] The Blues Foundation. Your Home for Blues, Blues, Blues! URL: http://www.blues.org	1. [pid:2] The Blues Foundation. Your Home for Blues, Blues, Blues! URL: http://www.blues.org
2. [pid:27] Harry's Blues Lyrics Online, Home Page, US URL: http://blueslinks.tripod.com	2. [pid:153] A Cyber Blues Society for musicians, fans, and webmasters of the URL: http://www.bluessociety.net
3. [pid:9] The Blue Highway URL: http://www.thebluehighway.com	3. [pid:2755] WebRing: addsite_login URL: http://l.webring.com/wrman?ring=bluessociety&addsite
4. [pid:887] home URL: http://www.fargobluesfest.com	4. [pid:4519] BluesSociety.net - for musicians, fans, and webmasters of the blu URL: http://sitebuilder.bluessociety.net
5. [pid:3460] stantonanderson.com URL: http://www.stantonanderson.com	5. [pid:4042] Google Search: URL: http://www.google.com/search
6. [pid:101] Delta Blues - DeltaBlues - deltablues.com URL: http://www.deltablues.com	6. [pid:4464] BluesSociety.net - Free Email Service. URL: http://mail.bluessociety.net
7. [pid:2077] Dan's Police Page URL: http://danspolice.8m.com	7. [pid:4561] Your Mailinglist Provider URL: http://www.yourmailinglistprovider.com/
8. [pid:153] A Cyber Blues Society for musicians, fans, and webmasters of the URL: http://www.bluessociety.net	8. [pid:1915] A Cyber Blues Society Blues Links Page 3 URL: http://www.bluessociety.net/links03.html
9. [pid:13] Blues On Stage, your complete blues guide. Twin Cities Blues - M URL: http://www.mnblues.com	9. [pid:3973] A Cyber Blues Society Blues Links Page 5 URL: http://www.bluessociety.net/links05.html
10. [pid:185] Natchel' Blues Network, Hampton Roads, VA URL: http://www.natchelblues.org	10. [pid:4502] Blues Biographies - Artist of the Blues - www.bluessociety.net URL: http://www.bluessociety.net/greats.html
11. [pid:4519] BluesSociety.net - for musicians, fans, and webmasters of the blu URL: http://sitebuilder.bluessociety.net	11. [pid:4564] Add your website to The Cyber Blues Society Reciprocal Links Page URL: http://www.bluessociety.net/links04.html
12. [pid:2378] Empty title field URL: http://www.letthegoodtimesroll.com	12. [pid:4565] BluesSociety.net - Community Message Center URL: http://community.bluessociety.net/commun_v3/scripts/directory.pl
13. [pid:2755] WebRing: addsite_login URL: http://l.webring.com/wrman?ring=bluessociety&addsite	13. [pid:4566] The Cyber Blues Society Newsletter sign up page. URL: http://www.bluessociety.net/newsletter.html
14. [pid:2183] MO BLUES.ORG URL: http://www.moblues.org	14. [pid:4567] A Cyber Blues Society Blues Links Page URL: http://www.bluessociety.net/links01.html
15. [pid:4] BLUES WORLD URL: http://www.bluesworld.com	15. [pid:4568] BLUES LINKS - The Cyber Blues Society Links page 2... URL: http://www.bluessociety.net/links02.html

Max	SALSA
1. [pid:2] The Blues Foundation. Your Home for Blues, Blues, Blues! URL: http://www.blues.org	1. [pid:2] The Blues Foundation. Your Home for Blues, Blues, Blues! URL: http://www.blues.org
2. [pid:27] Harry's Blues Lyrics Online, Home Page, US URL: http://blueslinks.tripod.com	2. [pid:27] Harry's Blues Lyrics Online, Home Page, US URL: http://blueslinks.tripod.com
3. [pid:9] The Blue Highway URL: http://www.thebluehighway.com	3. [pid:101] Delta Blues - DeltaBlues - deltablues.com URL: http://www.deltablues.com
4. [pid:153] A Cyber Blues Society for musicians, fans, and webmasters of the URL: http://www.bluessociety.net	4. [pid:9] The Blue Highway URL: http://www.thebluehighway.com
5. [pid:2755] WebRing: addsite_login URL: http://l.webring.com/wrman?ring=bluessociety&addsite	5. [pid:887] home URL: http://www.fargobluesfest.com
6. [pid:4519] BluesSociety.net - for musicians, fans, and webmasters of the blu URL: http://sitebuilder.bluessociety.net	6. [pid:3460] stantonanderson.com URL: http://www.stantonanderson.com
7. [pid:4042] Google Search: URL: http://www.google.com/search	7. [pid:2077] Dan's Police Page URL: http://danspolice.8m.com
8. [pid:101] Delta Blues - DeltaBlues - deltablues.com URL: http://www.deltablues.com	8. [pid:13] Blues On Stage, your complete blues guide. Twin Cities Blues - M URL: http://www.mnblues.com
9. [pid:13] Blues On Stage, your complete blues guide. Twin Cities Blues URL: http://www.mnblues.com	9. [pid:153] A Cyber Blues Society for musicians, fans, and webmasters URL: http://www.bluessociety.net
10. [pid:4464] BluesSociety.net - Free Email Service. URL: http://mail.bluessociety.net	10. [pid:2378] Empty title field URL: http://www.letthegoodtimesroll.com
11. [pid:4561] Your Mailinglist Provider URL: http://www.yourmailinglistprovider.com/	11. [pid:185] Natchel' Blues Network, Hampton Roads, VA URL: http://www.natchelblues.org
12. [pid:1915] A Cyber Blues Society Blues Links Page 3 URL: http://www.bluessociety.net/links03.html	12. [pid:117] ROOSTER BLUES URL: http://www.roosterblues.com
13. [pid:3973] A Cyber Blues Society Blues Links Page 5 URL: http://www.bluessociety.net/links05.html	13. [pid:2183] MO BLUES.ORG URL: http://www.moblues.org
14. [pid:4502] Blues Biographies - Artist of the Blues - www.bluessociety.net URL: http://www.bluessociety.net/greats.html	14. [pid:1767] bluesfind.cjb.net URL: http://www.bluesfind.cjb.net
15. [pid:4564] Add your website to The Cyber Blues Society Reciprocal Links Page URL: http://www.bluessociety.net/links04.html	15. [pid:4] BLUES WORLD URL: http://www.bluesworld.com

Table B.9: *Top 15 results for query "blues"*

HITS	HubAvg
1. [pid:477] (caffdi) take & URL: http://caffeinediary.blogspot.com	1. [pid:1683] Coming Soon... URL: http://www.cheesegiftbasket.net
2. [pid:2402] wrongwaygoback : dynamic ribbon device : by neale talbot URL: http://www.wrongwaygoback.com	2. [pid:1785] Ethnic art : African dance and drum classes, tribal music URL: http://www.ethnicarts.org
3. [pid:2349] Boing Boing: A Directory of Wonderful Things URL: http://www.boingboing.net	3. [pid:1793] Snowboard Boots for sale at Snowboard-Boots.com URL: http://www.snowboard-boots.com
4. [pid:1502] movabletype.org URL: http://www.movabletype.org	4. [pid:1787] Half Moon Bay Bed and Breakfast Inn — Northern California URL: http://www.millroseinn.com
5. [pid:2372] Izzle! Izzle pfaff! URL: http://www.izzlepfaff.com	5. [pid:1735] Empty title field URL: http://www.cheese-express.com
6. [pid:720] Parasyte: Insanity of the Mind v.2 URL: http://parasyte.pitas.com	6. [pid:1789] Empty title field URL: http://www.santacruzwebdesign.com
7. [pid:2369] harrumph! still crazy. URL: http://www.harrumph.com	7. [pid:1791] Small business merchant accounts — ecommerce and small business — URL: http://www.ikorb.com
8. [pid:2048] guestofbeth.diaryland.com URL: http://guestofbeth.diaryland.com	8. [pid:1786] Cookie gifts : cookie delivery : cookie tins : cookie baskets : C URL: http://www.pacificcookie.com
9. [pid:2401] ::: wood s lot ::: "fictive things wink as they will" URL: http://www.ncf.ca/ ek867/wood_s_lot.html	9. [pid:1792] Bushrods BBQ Equipment — barbecue trailers — barbeque smokers — b URL: http://www.bushrods.com
10. [pid:2266] Caterina.net URL: http://caterina.net	10. [pid:1788] b.firm Skin Care Products - Try our Anti Cellulite & Skin Firming URL: http://www.tobfirm.com
11. [pid:2396] Travelers Diagram...an appreciation of culture and creation URL: http://www.travelersdiagram.com	11. [pid:1790] Food Safety — Food Security — Purity Assurance Technologies - HAC URL: http://www.purityassurance.com
12. [pid:2346] anil dash - New York Still Loves Funk URL: http://dashes.com/anil	12. [pid:1794] Dairy — Dairy Products — Cheese Products — Dairy-Express.com URL: http://www.dairy-express.com
13. [pid:2354] defective yeti URL: http://www.defectiveyeti.com	13. [pid:97] 800cheesecake.com URL: http://www.800cheesecake.com
14. [pid:2359] Emptybottle.org : A little song, a little dance, a little seltzer URL: http://www.emptybottle.org	14. [pid:66] Cheese of the month club — cheese gift baskets URL: http://www.cheeseexpress.com
15. [pid:2358] Eclogues URL: http://www.sargassosea.net	15. [pid:34] Say Cheese - SayCheese - saycheese.net URL: http://www.saycheese.net

AT-Avg	Norm (2)
1. [pid:1683] Coming Soon ... URL: http://www.cheesegiftbasket.net	1. [pid:477] (caffdi) take & URL: http://caffeinediary.blogspot.com
2. [pid:1785] Ethnic art : African dance and drum classes, tribal music, and af URL: http://www.ethnicarts.org	2. [pid:1502] movabletype.org URL: http://www.movabletype.org
3. [pid:1793] Snowboard Boots for sale at Snowboard-Boots.com URL: http://www.snowboard-boots.com	3. [pid:2402] wrongwaygoback : dynamic ribbon device : by neale talbot URL: http://www.wrongwaygoback.com
4. [pid:1787] Half Moon Bay Bed and Breakfast Inn — Northern California URL: http://www.millroseinn.com	4. [pid:2349] Boing Boing: A Directory of Wonderful Things URL: http://www.boingboing.net
5. [pid:1735] Empty title field URL: http://www.cheese-express.com	5. [pid:2372] Izzle! Izzle pfaff! URL: http://www.izzlepfaff.com
6. [pid:1789] Empty title field URL: http://www.santacruzwebdesign.com	6. [pid:2048] guestofbeth.diaryland.com URL: http://guestofbeth.diaryland.com
7. [pid:1791] Small business merchant accounts — ecommerce and small business — URL: http://www.ikorb.com	7. [pid:720] Parasyte: Insanity of the Mind v.2 URL: http://parasyte.pitas.com
8. [pid:1786] Cookie gifts : cookie delivery : cookie tins : cookie baskets : C URL: http://www.pacificcookie.com	8. [pid:2369] harrumph! still crazy. URL: http://www.harrumph.com
9. [pid:1792] Bushrods BBQ Equipment — barbecue trailers — barbeque smokers — b URL: http://www.bushrods.com	9. [pid:2346] anil dash - New York Still Loves Funk URL: http://dashes.com/anil
10. [pid:1788] b.firm Skin Care Products - Try our Anti Cellulite & Skin Firming URL: http://www.tobfirm.com	10. [pid:2396] Travelers Diagram...an appreciation of culture and creation URL: http://www.travelersdiagram.com
11. [pid:1790] Food Safety — Food Security — Purity Assurance Technologies - HAC URL: http://www.purityassurance.com	11. [pid:2266] Caterina.net URL: http://caterina.net
12. [pid:1794] Dairy — Dairy Products — Cheese Products URL: http://www.dairy-express.com	12. [pid:2401] ::: wood s lot ::: "fictive things wink as they will" URL: http://www.ncf.ca/ēk867/wood_s_lot.html
13. [pid:66] Cheese of the month club — cheese gift baskets URL: http://www.cheeseexpress.com	13. [pid:2354] defective yeti URL: http://www.defectiveyeti.com
14. [pid:97] 800cheesecake.com URL: http://www.800cheesecake.com	14. [pid:2359] Emptybottle.org : A little song, a little dance, a little seltzer URL: http://www.emptybottle.org
15. [pid:477] (caffdi) take & URL: http://caffeinediary.blogspot.com	15. [pid:2314] randomwalks.com URL: http://randomwalks.com

Max	SALSA
1. [pid:477] (caffdi) take & URL: http://caffeinediary.blogspot.com	1. [pid:477] (caffdi) take & URL: http://caffeinediary.blogspot.com
2. [pid:1502] movabletype.org URL: http://www.movabletype.org	2. [pid:97] 800cheesecake.com URL: http://www.800cheesecake.com
3. [pid:2402] wrongwaygoback : dynamic ribbon device : by neale talbot URL: http://www.wrongwaygoback.com	3. [pid:1683] Coming Soon... URL: http://www.cheesegiftbasket.net
4. [pid:2349] Boing Boing: A Directory of Wonderful Things URL: http://www.boingboing.net	4. [pid:66] Cheese of the month club — cheese gift baskets URL: http://www.cheeseexpress.com
5. [pid:2372] Izzle! Izzle pfaff! URL: http://www.izzlepfaff.com	5. [pid:1502] movabletype.org URL: http://www.movabletype.org
6. [pid:2048] guestofbeth.diaryland.com URL: http://guestofbeth.diaryland.com	6. [pid:1785] Ethnic art : African dance and drum classes, tribal music, and af URL: http://www.ethnicarts.org
7. [pid:720] Parasyte: Insanity of the Mind v.2 URL: http://parasyte.pitas.com	7. [pid:1787] Half Moon Bay Bed and Breakfast Inn — Northern California URL: http://www.millroseinn.com
8. [pid:2369] harrumph! still crazy. URL: http://www.harrumph.com	8. [pid:1793] Snowboard Boots for sale at Snowboard-Boots.com URL: http://www.snowboard-boots.com
9. [pid:2280] kottke.org :: home of fine hypertext products URL: http://kottke.org	9. [pid:1735] Empty title field URL: http://www.cheese-express.com
10. [pid:2346] anil dash - New York Still Loves Funk URL: http://dashes.com/anil	10. [pid:34] Say Cheese - SayCheese - saycheese.net URL: http://www.saycheese.net
11. [pid:2266] Caterina.net URL: http://caterina.net	11. [pid:8] Cheese Racing URL: http://www.cheeseracing.org
12. [pid:2396] Travelers Diagram...an appreciation of culture and creation URL: http://www.travelersdiagram.com	12. [pid:2402] wrongwaygoback : dynamic ribbon device : by neale talbot URL: http://www.wrongwaygoback.com
13. [pid:2401] ::: wood s lot ::: "fictive things wink as they will" URL: http://www.ncf.ca/ēk867/wood_s_lot.html	13. [pid:2048] guestofbeth.diaryland.com URL: http://guestofbeth.diaryland.com
14. [pid:109] cheesedip.com - occasionally cranky commentary URL: http://www.cheesedip.com	14. [pid:2349] Boing Boing: A Directory of Wonderful Things URL: http://www.boingboing.net
15. [pid:2354] defective yeti URL: http://www.defectiveyeti.com	15. [pid:1] CHEESE.COM - All about cheese!. URL: http://www.cheese.com

Table B.10: *Top 15 results for query "cheese"*

HITS	HubAvg
1. [pid:868] earlyromanticguitar.com URL: http://www.earlyromanticguitar.com	1. [pid:972] Hitsquad.com - Musicians Web Center - Music Software, MP3 Softwar URL: http://www.hitsquad.com
2. [pid:423] Empty title field URL: http://classicalguitar.freehosting.net	2. [pid:973] Hitsquad Privacy Policy URL: http://www.hitsquad.com/privacy.shtml
3. [pid:1969] Adirondack Spruce.com URL: http://adirondackspruce.com	3. [pid:974] Advertising on Hitsquad Music Industry Web Sites URL: http://www.hitsquad.com/advertising.shtml
4. [pid:87] The Classical Guitar Homepage of AK*Creations URL: http://www.ak-c.demon.nl	4. [pid:910] Empty title field URL: http://www.vicnet.net.au/ easyjamn
5. [pid:12] Guitar Alive - GuitarAlive - guitaralive.com URL: http://www.guitaralive.com	5. [pid:1644] AMG All Music Guide URL: http://www.allmusic.com
6. [pid:19] Empty title field URL: http://www.guitarfoundation.org	6. [pid:1643] Free Music Download, MP3 Music, Music Chat, Music Video, Music CD URL: http://ubl.com
7. [pid:9] GUITAR REVIEW URL: http://www.guitarreview.com	7. [pid:1645] 2000 Guitars Database URL: http://dargo.vicnet.net.au/guitar/list.cfm?category=Bands/Artists
8. [pid:1533] Avi Afriat - Classical guitar homepage URL: http://afriat.tripod.com	8. [pid:12] Guitar Alive - GuitarAlive - guitaralive.com URL: http://www.guitaralive.com
9. [pid:1] The Classical Guitar Home Page URL: http://www.guitarist.com/cg/cg.htm	9. [pid:1639] CDNOW URL: http://www.cdnow.com/from=sr-767167
10. [pid:131] Empty title field URL: http://www.duolenz.com	10. [pid:73] OLGA - The On-Line Guitar Archive URL: http://www.olga.net
11. [pid:2136] New Page 1 URL: http://www.inokuchiviolin.com	11. [pid:2379] GuitarsRule.com URL: http://www.guitarsrule.com
12. [pid:2655] Experimental Guitar URL: http://guitarsoc.homestead.com	12. [pid:3084] MP3.com: THE destination for digital music. URL: http://www.mp3.com
13. [pid:49] Welcome to the San Francisco Classical Guitar Society! URL: http://www.sfcgs.org	13. [pid:415] Bach for Guitar in Tablature URL: http://alan.melvin.com
14. [pid:2656] Free guitars tips tabs Tutors St Albans Hertfordshire Herts URL: http://www.houlston.freeserve.co.uk	14. [pid:2263] The Boston Classical Guitar Society URL: http://www.bostonguitar.org
15. [pid:2676] Bromley Guitar Society Index Page (Rev E) URL: http://www.bromleyguitarsociety.co.uk	15. [pid:2633] GearSearch.com - musical instruments - guitar, bass, drums, keybo URL: http://www.gearsearch.com

AT-Avg	Norm (2)
1. [pid:972] Hitsquad.com - Musicians Web Center - Music Software, MP3 Softwar URL: http://www.hitsquad.com	1. [pid:972] Hitsquad.com - Musicians Web Center - Music Software, MP3 Softwar URL: http://www.hitsquad.com
2. [pid:973] Hitsquad Privacy Policy URL: http://www.hitsquad.com/privacy.shtml	2. [pid:973] Hitsquad Privacy Policy URL: http://www.hitsquad.com/privacy.shtml
3. [pid:974] Advertising on Hitsquad Music Industry Web Sites URL: http://www.hitsquad.com/advertising.shtml	3. [pid:974] Advertising on Hitsquad Music Industry Web Sites URL: http://www.hitsquad.com/advertising.shtml
4. [pid:910] Empty title field URL: http://www.vicnet.net.au/ easyjamn	4. [pid:910] Empty title field URL: http://www.vicnet.net.au/ easyjamn
5. [pid:1644] AMG All Music Guide URL: http://www.allmusic.com	5. [pid:1644] AMG All Music Guide URL: http://www.allmusic.com
6. [pid:1643] Free Music Download, MP3 Music, Music Chat, Music Video, Music CD URL: http://ubl.com	6. [pid:1643] Free Music Download, MP3 Music, Music Chat, Music Video, Music CD URL: http://ubl.com
7. [pid:1645] 2000 Guitars Database URL: http://dargo.vicnet.net.au/guitar/list.cfm?category=Bands/Artists	7. [pid:12] Guitar Alive - GuitarAlive - guitaralive.com URL: http://www.guitaralive.com
8. [pid:1639] CDNOW URL: http://www.cdnow.com/from=sr-767167	8. [pid:1645] 2000 Guitars Database URL: http://dargo.vicnet.net.au/guitar/list.cfm?category=Bands/Artists
9. [pid:12] Guitar Alive - GuitarAlive - guitaralive.com URL: http://www.guitaralive.com	9. [pid:1639] CDNOW URL: http://www.cdnow.com/from=sr-767167
10. [pid:73] OLGA - The On-Line Guitar Archive URL: http://www.olga.net	10. [pid:868] earlyromanticguitar.com URL: http://www.earlyromanticguitar.com
11. [pid:423] Empty title field URL: http://classicalguitar.freehosting.net	11. [pid:423] Empty title field URL: http://classicalguitar.freehosting.net
12. [pid:868] earlyromanticguitar.com URL: http://www.earlyromanticguitar.com	12. [pid:73] OLGA - The On-Line Guitar Archive URL: http://www.olga.net
13. [pid:1] The Classical Guitar Home Page URL: http://www.guitarist.com/cg/cg.htm	13. [pid:1] The Classical Guitar Home Page URL: http://www.guitarist.com/cg/cg.htm
14. [pid:3084] MP3.com: THE destination for digital music. URL: http://www.mp3.com	14. [pid:87] The Classical Guitar Homepage of AK*Creations URL: http://www.ak-c.demon.nl
15. [pid:87] The Classical Guitar Homepage of AK*Creations URL: http://www.ak-c.demon.nl	15. [pid:1969] Adirondack Spruce.com URL: http://adirondackspruce.com

Max	SALSA
1. [pid:12] Guitar Alive - GuitarAlive - guitaralive.com URL: http://www.guitaralive.com	1. [pid:12] Guitar Alive - GuitarAlive - guitaralive.com URL: http://www.guitaralive.com
2. [pid:868] earlyromanticguitar.com URL: http://www.earlyromanticguitar.com	2. [pid:868] earlyromanticguitar.com URL: http://www.earlyromanticguitar.com
3. [pid:19] Empty title field URL: http://www.guitarfoundation.org	3. [pid:19] Empty title field URL: http://www.guitarfoundation.org
4. [pid:1969] Adirondack Spruce.com URL: http://adirondackspruce.com	4. [pid:972] Hitsquad.com - Musicians Web Center - Music Software, MP3 Softwar URL: http://www.hitsquad.com
5. [pid:423] Empty title field URL: http://classicalguitar.freehosting.net	5. [pid:973] Hitsquad Privacy Policy URL: http://www.hitsquad.com/privacy.shtml
6. [pid:9] GUITAR REVIEW URL: http://www.guitarreview.com	6. [pid:974] Advertising on Hitsquad Music Industry Web Sites URL: http://www.hitsquad.com/advertising.shtml
7. [pid:87] The Classical Guitar Homepage of AK*Creations URL: http://www.ak-c.demon.nl	7. [pid:1969] Adirondack Spruce.com URL: http://adirondackspruce.com
8. [pid:1] The Classical Guitar Home Page URL: http://www.guitarist.com/cg/cg.htm	8. [pid:423] Empty title field URL: http://classicalguitar.freehosting.net
9. [pid:972] Hitsquad.com - Musicians Web Center - Music Software, MP3 Softwar URL: http://www.hitsquad.com	9. [pid:910] Empty title field URL: http://www.vicnet.net.au/ easyjamn
10. [pid:973] Hitsquad Privacy Policy URL: http://www.hitsquad.com/privacy.shtml	10. [pid:87] The Classical Guitar Homepage of AK*Creations URL: http://www.ak-c.demon.nl
11. [pid:974] Advertising on Hitsquad Music Industry Web Sites URL: http://www.hitsquad.com/advertising.shtml	11. [pid:9] GUITAR REVIEW URL: http://www.guitarreview.com
12. [pid:1533] Avi Afriat - Classical guitar homepage URL: http://afriat.tripod.com	12. [pid:1] The Classical Guitar Home Page URL: http://www.guitarist.com/cg/cg.htm
13. [pid:910] Empty title field URL: http://www.vicnet.net.au/ easyjamn	13. [pid:73] OLGA - The On-Line Guitar Archive URL: http://www.olga.net
14. [pid:131] Empty title field URL: http://www.duolenz.com	14. [pid:1533] Avi Afriat - Classical guitar homepage URL: http://afriat.tripod.com
15. [pid:2317] Guitar Notes - guitar links, lessons, mp3s, tabs, shopping, and r URL: http://www.guitarnotes.com	15. [pid:131] Empty title field URL: http://www.duolenz.com

Table B.11: *Top 15 results for query "classical guitar"*

HITS	**HubAvg**
1. [pid:2435] Ziff Davis Media — Home URL: http://www.ziffdavis.com	1. [pid:2435] Ziff Davis Media — Home URL: http://www.ziffdavis.com
2. [pid:2527] Ziff Davis Media — Privacy Policy and Terms URL: http://www.ziffdavis.com/terms/index.asp?page=privacypolicy	2. [pid:2527] Ziff Davis Media — Privacy Policy and Terms URL: http://www.ziffdavis.com/terms/index.asp?page=privacypolicy
3. [pid:2528] Ziff Davis Media — Privacy Policy and Terms URL: http://www.ziffdavis.com/terms/index.asp?page=termsofservice	3. [pid:2528] Ziff Davis Media — Privacy Policy and Terms URL: http://www.ziffdavis.com/terms/index.asp?page=termsofservice
4. [pid:2474] :::::::::::::::::: Registration :::::::::::::::::: URL: http://webevents.broadcast.com/ziffdavis/062303/index.asp?loc=bot	4. [pid:2474] :::::::::::::::::: Registration :::::::::::::::::: URL: http://webevents.broadcast.com/ziffdavis/062303/index.asp?loc=bot
5. [pid:2475] :::::::::::::::::: Registration :::::::::::::::::: URL: http://webevents.broadcast.com/ziffdavis/062603/index.asp?loc=bot	5. [pid:2475] :::::::::::::::::: Registration :::::::::::::::::: URL: http://webevents.broadcast.com/ziffdavis/062603/index.asp?loc=bot
6. [pid:2476] :::::::::::::::::: Registration :::::::::::::::::: URL: http://webevents.broadcast.com/ziffdavis/063003/index.asp?loc=bot	6. [pid:2476] :::::::::::::::::: Registration :::::::::::::::::: URL: http://webevents.broadcast.com/ziffdavis/063003/index.asp?loc=bot
7. [pid:2477] eWeek Research Library: Wireless Security URL: http://eweek.bitpipe.com/data/rlist?t=1016747982_81244038&src=int	7. [pid:2529] Ziff Davis Media — About URL: http://www.ziffdavis.com/about/index.asp?page=contactus
8. [pid:2478] eWeek Research Library: Wireless LAN Security URL: http://eweek.bitpipe.com/...&type=RES&x	8. [pid:2477] eWeek Research Library: Wireless Security URL: http://eweek.bitpipe.com/...8&src=int
9. [pid:2479] eWeek Research Library: How to Ensure Your Wireless Deployment Is URL: http://eweek.bitpipe.com/...&type=RES&x=	9. [pid:2478] eWeek Research Library: Wireless LAN Security - Enterprise Rogue URL: http://eweek.bitpipe.com/...&type=RES&x
10. [pid:2480] eWeek Research Library: The CIO's Guide to Wireless URL: http://eweek.bitpipe.com/...&type=RES&x	10. [pid:2479] eWeek Research Library: How to Ensure Your Wireless Deployment Is URL: http://eweek.bitpipe.com/...&type=RES&x=
11. [pid:2481] eWeek Research Library: Who's Watching Your Wireless Network? URL: http://eweek.bitpipe.com/...&type=RES&x=	11. [pid:2480] eWeek Research Library: The CIO's Guide to Wireless URL: http://eweek.bitpipe.com/...&type=RES&x
12. [pid:2482] eWeek Research Library: Power over Ethernet for Wireless LANs URL: http://eweek.bitpipe.com/...&type=RES&x	12. [pid:2481] eWeek Research Library: Who's Watching Your Wireless Network? URL: http://eweek.bitpipe.com/...&type=RES&x=
13. [pid:2483] eWeek Research Library: IEEE 802.11g URL: http://eweek.bitpipe.com/...&type=RES&x	13. [pid:2482] eWeek Research Library: Power over Ethernet for Wireless LANs URL: http://eweek.bitpipe.com/...&type=RES&x
14. [pid:2484] PriceGrabber.com - The Smart Place to Start Your Shopping URL: http://eweek.pricegrabber.com/...&mode=e	14. [pid:2483] eWeek Research Library: IEEE 802.11g URL: http://eweek.bitpipe.com/...&type=RES&x
15. [pid:2485] PriceGrabber.com - The Smart Place to Start Your Shopping URL: http://eweek.pricegrabber.com/...mode=ew_rhmod061303	15. [pid:2484] PriceGrabber.com - The Smart Place to Start Your Shopping URL: http://eweek.pricegrabber.com/...&mode=e
AT-Avg	**Norm (2)**
1. [pid:2435] Ziff Davis Media — Home URL: http://www.ziffdavis.com	1. [pid:2435] Ziff Davis Media — Home URL: http://www.ziffdavis.com
2. [pid:2527] Ziff Davis Media — Privacy Policy and Terms URL: http://www.ziffdavis.com/terms/index.asp?page=privacypolicy	2. [pid:2527] Ziff Davis Media — Privacy Policy and Terms URL: http://www.ziffdavis.com/terms/index.asp?page=privacypolicy
3. [pid:2528] Ziff Davis Media — Privacy Policy and Terms URL: http://www.ziffdavis.com/terms/index.asp?page=termsofservice	3. [pid:2528] Ziff Davis Media — Privacy Policy and Terms URL: http://www.ziffdavis.com/terms/index.asp?page=termsofservice
4. [pid:2474] :::::::::::::::::: Registration :::::::::::::::::: URL: http://webevents.broadcast.com/ziffdavis/062303/index.asp?loc=bot	4. [pid:2474] :::::::::::::::::: Registration :::::::::::::::::: URL: http://webevents.broadcast.com/ziffdavis/062303/index.asp?loc=bot
5. [pid:2475] :::::::::::::::::: Registration :::::::::::::::::: URL: http://webevents.broadcast.com/ziffdavis/062603/index.asp?loc=bot	5. [pid:2475] :::::::::::::::::: Registration :::::::::::::::::: URL: http://webevents.broadcast.com/ziffdavis/062603/index.asp?loc=bot
6. [pid:2476] :::::::::::::::::: Registration :::::::::::::::::: URL: http://webevents.broadcast.com/ziffdavis/063003/index.asp?loc=bot	6. [pid:2476] :::::::::::::::::: Registration :::::::::::::::::: URL: http://webevents.broadcast.com/ziffdavis/063003/index.asp?loc=bot
7. [pid:2529] Ziff Davis Media — About URL: http://www.ziffdavis.com/about/index.asp?page=contactus	7. [pid:2529] Ziff Davis Media — About URL: http://www.ziffdavis.com/about/index.asp?page=contactus
8. [pid:2477] eWeek Research Library: Wireless Security URL: http://eweek.bitpipe.com/..._81244038&src=int	8. [pid:2477] eWeek Research Library: Wireless Security URL: http://eweek.bitpipe.com/..._81244038&src=int
9. [pid:2478] eWeek Research Library: Wireless LAN Security - Enterprise Rogue URL: http://eweek.bitpipe.com/...&type=RES&x	9. [pid:2478] eWeek Research Library: Wireless LAN Security - Enterprise Rogue URL: http://eweek.bitpipe.com/...&type=RES&x
10. [pid:2479] eWeek Research Library: How to Ensure Your Wireless Deployment Is URL: http://eweek.bitpipe.com/...&type=RES&x=	10. [pid:2479] eWeek Research Library: How to Ensure Your Wireless Deployment Is URL: http://eweek.bitpipe.com/...&type=RES&x=
11. [pid:2480] eWeek Research Library: The CIO's Guide to Wireless URL: http://eweek.bitpipe.com/...&type=RES&x	11. [pid:2480] eWeek Research Library: The CIO's Guide to Wireless URL: http://eweek.bitpipe.com/...&type=RES&x
12. [pid:2481] eWeek Research Library: Who's Watching Your Wireless Network? URL: http://eweek.bitpipe.com/...&type=RES&x=	12. [pid:2481] eWeek Research Library: Who's Watching Your Wireless Network? URL: http://eweek.bitpipe.com/...&type=RES&x=
13. [pid:2482] eWeek Research Library: Power over Ethernet for Wireless LANs URL: http://eweek.bitpipe.com/...&type=RES&x	13. [pid:2482] eWeek Research Library: Power over Ethernet for Wireless LANs URL: http://eweek.bitpipe.com/...&type=RES&x
14. [pid:2483] eWeek Research Library: IEEE 802.11g URL: http://eweek.bitpipe.com/...&type=RES&x	14. [pid:2483] eWeek Research Library: IEEE 802.11g URL: http://eweek.bitpipe.com/...&type=RES&x
15. [pid:2484] PriceGrabber.com - The Smart Place to Start Your Shopping URL: http://eweek.pricegrabber.com/...&mode=e	15. [pid:2484] PriceGrabber.com - The Smart Place to Start Your Shopping URL: http://eweek.pricegrabber.com/.../&mode=e
Max	**SALSA**
1. [pid:565] SFI Home Page URL: http://www.santafe.edu	1. [pid:565] SFI Home Page URL: http://www.santafe.edu
2. [pid:527] New England Complex Systems Institute URL: http://necsi.org	2. [pid:3] ECCC - The Electronic Colloquium on Computational Complexity URL: http://eccc.uni-trier.de/eccc
3. [pid:3] ECCC - The Electronic Colloquium on Computational Complexity URL: http://eccc.uni-trier.de/eccc	3. [pid:2435] Ziff Davis Media — Home URL: http://www.ziffdavis.com
4. [pid:6] Complexity Digest URL: http://www.comdig.org	4. [pid:178] Artificial Life VIII The 8th International Conference on the Simu URL: http://alife8.alife.org
5. [pid:178] Artificial Life VIII The 8th International Conference URL: http://alife8.alife.org	5. [pid:2527] Ziff Davis Media — Privacy Policy and Terms URL: http://www.ziffdavis.com/terms/index.asp?page=privacypolicy
6. [pid:10] The Complexity and Artificial Life Research Concept URL: http://www.calresco.org	6. [pid:2528] Ziff Davis Media — Privacy Policy and Terms URL: http://www.ziffdavis.com/terms/index.asp?page=termsofservice
7. [pid:1664] Google URL: http://www.google.com	7. [pid:1664] Google URL: http://www.google.com
8. [pid:11] Complexity, Self Adaptive Complex Systems, and Chaos Theory URL: http://www.brint.com/Systems.htm	8. [pid:6] Complexity Digest URL: http://www.comdig.org
9. [pid:1685] CCSR Homepage URL: http://www.ccsr.uiuc.edu	9. [pid:263] Empty title field URL: http://www.ams.org
10. [pid:39] Emergence - A Journal of Complexity Issues in Organizations and M URL: http://www.emergence.org	10. [pid:2474] :::::::::::::::::: Registration :::::::::::::::::: URL: http://webevents.broadcast.com/.../index.asp?loc=bot
11. [pid:903] Welcome to Principia Cybernetica Web URL: http://pespmc1.vub.ac.be	11. [pid:2475] :::::::::::::::::: Registration :::::::::::::::::: URL: http://webevents.broadcast.com/.../index.asp?loc=bot
12. [pid:1] Complexity International URL: http://www.csu.edu.au/ci	12. [pid:2476] :::::::::::::::::: Registration :::::::::::::::::: URL: http://webevents.broadcast.com/.../index.asp?loc=bot
13. [pid:1694] lanl.arXiv.org e-Print archive mirror URL: http://xyz.lanl.gov	13. [pid:2529] Ziff Davis Media — About URL: http://www.ziffdavis.com/about/index.asp?page=contactus
14. [pid:2095] Chaos at Maryland URL: http://www-chaos.umd.edu	14. [pid:2477] eWeek Research Library: Wireless Security URL: http://eweek.bitpipe.com/...&src=int
15. [pid:1156] The Collection of Computer Science Bibliographies URL: http://liinwww.ira.uka.de/bibliography	15. [pid:2478] eWeek Research Library: Wireless LAN Security URL: http://eweek.bitpipe.com/...&type=RES&x

Table B.12: *Top 15 results for query "complexity"*

HITS	**HubAvg**
1. [pid:1] ECCC - The Electronic Colloquium on Computational Complexity URL: http://eccc.uni-trier.de/eccc	1. [pid:1] ECCC - The Electronic Colloquium on Computational Complexity URL: http://eccc.uni-trier.de/eccc
2. [pid:256] ACM: Association for Computing Machinery URL: http://www.acm.org	2. [pid:5] My Computational Complexity Web Log URL: http://www.fortnow.com/lance/complog
3. [pid:557] The Electronic Journal of Combinatorics URL: http://www.combinatorics.org	3. [pid:2] Springer Link - Publication URL: http://link.springer-ny.com/link/service/journals/00037
4. [pid:384] Center for Discrete Mathematics and Theoretical Computer Science URL: http://dimacs.rutgers.edu	4. [pid:557] The Electronic Journal of Combinatorics URL: http://www.combinatorics.org
5. [pid:2] Springer Link - Publication URL: http://link.springer-ny.com/link/service/journals/00037	5. [pid:256] ACM: Association for Computing Machinery URL: http://www.acm.org
6. [pid:766] IEEE Computer Society URL: http://computer.org	6. [pid:384] Center for Discrete Mathematics and Theoretical Computer Science URL: http://dimacs.rutgers.edu
7. [pid:328] European Association for Theoretical Computer Science URL: http://www.eatcs.org	7. [pid:4] CC Published by Birkhäuser Verlag AG URL: http://www.birkhauser.ch/journals/3700/3700_tit.htm
8. [pid:23] Complexity People URL: http://eccc.uni-trier.de/eccc/info/people.html	8. [pid:3] IEEE Conference on Computational Complexity URL: http://cs.utep.edu/longpre/complexity.html
9. [pid:3] IEEE Conference on Computational Complexity URL: http://cs.utep.edu/longpre/complexity.html	9. [pid:729] Computer Science Papers NEC Research Institute CiteSeer URL: http://citeseer.nj.nec.com/cs
10. [pid:729] Computer Science Papers NEC Research Institute CiteSeer Publicati URL: http://citeseer.nj.nec.com/cs	10. [pid:23] Complexity People URL: http://eccc.uni-trier.de/eccc/info/people.html
11. [pid:9] Computational Complexity Conference URL: http://computationalcomplexity.org	11. [pid:766] IEEE Computer Society URL: http://computer.org
12. [pid:7] BEATCS Computational Complexity Column URL: http://external.nj.nec.com/homepages/fortnow/beatcs	12. [pid:245] NN Home URL: http://pro.blogger.com
13. [pid:242] The Complexity Zoo URL: http://www.cs.berkeley.edu/ aaronson/zoo.html	13. [pid:730] CiteSeer: The NEC Research Institute Scientific Literature URL: http://citeseer.org
14. [pid:730] CiteSeer: The NEC Research Institute Scientific Literature URL: http://citeseer.org	14. [pid:1032] A compendium of NP optimization problems URL: http://www.nada.kth.se/ viggo/problemlist/compendium.html
15. [pid:372] www.ncstrl.org URL: http://www.ncstrl.org	15. [pid:328] European Association for Theoretical Computer Science URL: http://www.eatcs.org
AT-Avg	**Norm (2)**
1. [pid:1] ECCC - The Electronic Colloquium on Computational Complexity URL: http://eccc.uni-trier.de/eccc	1. [pid:1] ECCC - The Electronic Colloquium on Computational Complexity URL: http://eccc.uni-trier.de/eccc
2. [pid:256] ACM: Association for Computing Machinery URL: http://www.acm.org	2. [pid:2] Springer Link - Publication URL: http://link.springer-ny.com/link/service/journals/00037
3. [pid:2] Springer Link - Publication URL: http://link.springer-ny.com/link/service/journals/00037	3. [pid:256] ACM: Association for Computing Machinery, the world's first educa URL: http://www.acm.org
4. [pid:557] The Electronic Journal of Combinatorics URL: http://www.combinatorics.org	4. [pid:557] The Electronic Journal of Combinatorics URL: http://www.combinatorics.org
5. [pid:384] Center for Discrete Mathematics and Theoretical Computer Science URL: http://dimacs.rutgers.edu	5. [pid:384] Center for Discrete Mathematics and Theoretical Computer Science URL: http://dimacs.rutgers.edu
6. [pid:766] IEEE Computer Society URL: http://computer.org	6. [pid:766] IEEE Computer Society URL: http://computer.org
7. [pid:23] Complexity People URL: http://eccc.uni-trier.de/eccc/info/people.html	7. [pid:23] Complexity People URL: http://eccc.uni-trier.de/eccc/info/people.html
8. [pid:328] European Association for Theoretical Computer Science URL: http://www.eatcs.org	8. [pid:328] European Association for Theoretical Computer Science URL: http://www.eatcs.org
9. [pid:729] Computer Science Papers NEC Research Institute CiteSeer Publicati URL: http://citeseer.nj.nec.com/cs	9. [pid:729] Computer Science Papers NEC Research Institute CiteSeer URL: http://citeseer.nj.nec.com/cs
10. [pid:3] IEEE Conference on Computational Complexity URL: http://cs.utep.edu/longpre/complexity.html	10. [pid:3] IEEE Conference on Computational Complexity URL: http://cs.utep.edu/longpre/complexity.html
11. [pid:730] CiteSeer: The NEC Research Institute Scientific Literature URL: http://citeseer.org	11. [pid:730] CiteSeer: The NEC Research Institute Scientific Literature URL: http://citeseer.org
12. [pid:9] Computational Complexity Conference URL: http://computationalcomplexity.org	12. [pid:9] Computational Complexity Conference URL: http://computationalcomplexity.org
13. [pid:372] www.ncstrl.org URL: http://www.ncstrl.org	13. [pid:242] The Complexity Zoo URL: http://www.cs.berkeley.edu/ aaronson/zoo.html
14. [pid:242] The Complexity Zoo URL: http://www.cs.berkeley.edu/ aaronson/zoo.html	14. [pid:4] CC Published by Birkhäuser Verlag AG URL: http://www.birkhauser.ch/journals/3700/3700_tit.htm
15. [pid:1032] A compendium of NP optimization problems URL: http://www.nada.kth.se/ viggo/problemlist/compendium.html	15. [pid:372] www.ncstrl.org URL: http://www.ncstrl.org
Max	**SALSA**
1. [pid:1] ECCC - The Electronic Colloquium on Computational Complexity URL: http://eccc.uni-trier.de/eccc	1. [pid:1] ECCC - The Electronic Colloquium on Computational Complexity URL: http://eccc.uni-trier.de/eccc
2. [pid:2] Springer Link - Publication URL: http://link.springer-ny.com/link/service/journals/00037	2. [pid:5] My Computational Complexity Web Log URL: http://www.fortnow.com/lance/complog
3. [pid:256] ACM: Association for Computing Machinery URL: http://www.acm.org	3. [pid:2] Springer Link - Publication URL: http://link.springer-ny.com/link/service/journals/00037
4. [pid:557] The Electronic Journal of Combinatorics URL: http://www.combinatorics.org	4. [pid:256] ACM: Association for Computing Machinery URL: http://www.acm.org
5. [pid:384] Center for Discrete Mathematics and Theoretical Computer Science URL: http://dimacs.rutgers.edu	5. [pid:557] The Electronic Journal of Combinatorics URL: http://www.combinatorics.org
6. [pid:23] Complexity People URL: http://eccc.uni-trier.de/eccc/info/people.html	6. [pid:3] IEEE Conference on Computational Complexity URL: http://cs.utep.edu/longpre/complexity.html
7. [pid:766] IEEE Computer Society URL: http://computer.org	7. [pid:384] Center for Discrete Mathematics and Theoretical Computer Science URL: http://dimacs.rutgers.edu
8. [pid:328] European Association for Theoretical Computer Science URL: http://www.eatcs.org	8. [pid:729] Computer Science Papers NEC Research Institute CiteSeer Publicati URL: http://citeseer.nj.nec.com/cs
9. [pid:729] Computer Science Papers NEC Research Institute CiteSeer URL: http://citeseer.nj.nec.com/cs	9. [pid:766] IEEE Computer Society URL: http://computer.org
10. [pid:3] IEEE Conference on Computational Complexity URL: http://cs.utep.edu/longpre/complexity.html	10. [pid:9] Computational Complexity Conference URL: http://computationalcomplexity.org
11. [pid:9] Computational Complexity Conference URL: http://computationalcomplexity.org	11. [pid:23] Complexity People URL: http://eccc.uni-trier.de/eccc/info/people.html
12. [pid:730] CiteSeer: The NEC Research Institute Scientific Literature URL: http://citeseer.org	12. [pid:4] CC Published by Birkhäuser Verlag AG URL: http://www.birkhauser.ch/journals/3700/3700_tit.htm
13. [pid:242] The Complexity Zoo URL: http://www.cs.berkeley.edu/ aaronson/zoo.html	13. [pid:370] IEEE Conference on Computational Complexity URL: http://www.cs.utep.edu/longpre/complexity.html
14. [pid:4] CC Published by Birkhäuser Verlag AG URL: http://www.birkhauser.ch/journals/3700/3700_tit.htm	14. [pid:730] CiteSeer: The NEC Research Institute Scientific Literature URL: http://citeseer.org
15. [pid:372] www.ncstrl.org URL: http://www.ncstrl.org	15. [pid:328] European Association for Theoretical Computer Science URL: http://www.eatcs.org

Table B.13: *Top 15 results for query "computational complexity"*

HITS	**HubAvg**
1. [pid:4] Geometry in Action URL: http://www.ics.uci.edu/ eppstein/geom.html	1. [pid:1743] WebCT.com URL: http://www.webct.com
2. [pid:6] The former CGAL home page URL: http://www.cs.uu.nl/CGAL	2. [pid:1744] DREXEL UNIVERSITY: Philadelphia, PA URL: http://www.drexel.edu
3. [pid:237] The Geometry Junkyard URL: http://www.ics.uci.edu/ eppstein/junkyard	3. [pid:1745] A Virtual Math Community: Members helping Members URL: http://www.drexel.edu/ia/mathforum
4. [pid:155] David Eppstein URL: http://www.ics.uci.edu/ eppstein	4. [pid:4] Geometry in Action URL: http://www.ics.uci.edu/ eppstein/geom.html
5. [pid:3] Computational Geometry Resources URL: http://www.scs.carleton.ca/ csgs/resources/cg.html	5. [pid:3] Computational Geometry Resources URL: http://www.scs.carleton.ca/ csgs/resources/cg.html
6. [pid:161] Joseph O'Rourke URL: http://cs.smith.edu/ orourke	6. [pid:9] The CGAL Home Page URL: http://www.cgal.org
7. [pid:1] Springer Link - Publication URL: http://link.springer.de/link/service/journals/00454	7. [pid:161] Joseph O'Rourke URL: http://cs.smith.edu/ orourke
8. [pid:233] LEDA moved to Algorithmic Solutions Software GmbH URL: http://www.mpi-sb.mpg.de/LEDA/leda.html	8. [pid:26] Fast Robust Predicates for Computational Geometry URL: http://www.cs.cmu.edu/ quake/robust.html
9. [pid:318] The compgeom mailing lists URL: http://netlib.bell-labs.com/netlib/compgeom/readme.html	9. [pid:316] Computational Geometry Pages URL: http://www.uiuc.edu/ph/www/jeffe/compgeom
10. [pid:476] Günter M. Ziegler URL: http://www.math.tu-berlin.de/ ziegler	10. [pid:237] The Geometry Junkyard URL: http://www.ics.uci.edu/ eppstein/junkyard
11. [pid:448] Steven Skiena URL: http://www.cs.sunysb.edu/ skiena	11. [pid:219] GANG — Geometry Analysis Numerics Graphics URL: http://www.gang.umass.edu
12. [pid:446] Micha Sharir's Home Page URL: http://www.math.tau.ac.il/ sharir	12. [pid:12] Computational Geometry Code URL: http://compgeom.cs.uiuc.edu/ jeffe/compgeom/code.html
13. [pid:462] Seth Teller URL: http://graphics.lcs.mit.edu/ seth	13. [pid:1301] Computational Geometry Pages URL: http://compgeom.cs.uiuc.edu/ jeffe/compgeom
14. [pid:394] Paul Heckbert's Web Page URL: http://www.cs.cmu.edu/ ph	14. [pid:2099] ACM JEA Home Page URL: http://www.jea.acm.org
15. [pid:433] Franco P. Preparata's Home Page URL: http://www.cs.brown.edu/people/franco	15. [pid:1047] ACM: Association for Computing Machinery, the world's first URL: http://www.acm.org
AT-Avg	**Norm (2)**
1. [pid:4] Geometry in Action URL: http://www.ics.uci.edu/ eppstein/geom.html	1. [pid:4] Geometry in Action URL: http://www.ics.uci.edu/ eppstein/geom.html
2. [pid:1743] WebCT.com URL: http://www.webct.com	2. [pid:6] The former CGAL home page URL: http://www.cs.uu.nl/CGAL
3. [pid:1744] DREXEL UNIVERSITY: Philadelphia, PA URL: http://www.drexel.edu	3. [pid:237] The Geometry Junkyard URL: http://www.ics.uci.edu/ eppstein/junkyard
4. [pid:1745] A Virtual Math Community: Members helping Members URL: http://www.drexel.edu/ia/mathforum	4. [pid:3] Computational Geometry Resources URL: http://www.scs.carleton.ca/ csgs/resources/cg.html
5. [pid:6] The former CGAL home page URL: http://www.cs.uu.nl/CGAL	5. [pid:1743] WebCT.com URL: http://www.webct.com
6. [pid:237] The Geometry Junkyard URL: http://www.ics.uci.edu/ eppstein/junkyard	6. [pid:1744] DREXEL UNIVERSITY: Philadelphia, PA URL: http://www.drexel.edu
7. [pid:3] Computational Geometry Resources URL: http://www.scs.carleton.ca/ csgs/resources/cg.html	7. [pid:1745] A Virtual Math Community: Members helping Members URL: http://www.drexel.edu/ia/mathforum
8. [pid:1301] Computational Geometry Pages URL: http://compgeom.cs.uiuc.edu/ jeffe/compgeom	8. [pid:1301] Computational Geometry Pages URL: http://compgeom.cs.uiuc.edu/ jeffe/compgeom
9. [pid:161] Joseph O'Rourke URL: http://cs.smith.edu/ orourke	9. [pid:1] Springer Link - Publication URL: http://link.springer.de/link/service/journals/00454
10. [pid:1] Springer Link - Publication URL: http://link.springer.de/link/service/journals/00454	10. [pid:1047] ACM: Association for Computing Machinery, the world's first educa URL: http://www.acm.org
11. [pid:235] The Stony Brook Algorithm Repository URL: http://www.cs.sunysb.edu/ algorith	11. [pid:155] David Eppstein URL: http://www.ics.uci.edu/ eppstein
12. [pid:1047] ACM: Association for Computing Machinery, the world's first educa URL: http://www.acm.org	12. [pid:235] The Stony Brook Algorithm Repository URL: http://www.cs.sunysb.edu/ algorith
13. [pid:155] David Eppstein URL: http://www.ics.uci.edu/ eppstein	13. [pid:161] Joseph O'Rourke URL: http://cs.smith.edu/ orourke
14. [pid:9] The CGAL Home Page URL: http://www.cgal.org	14. [pid:1985] MathWorld URL: http://mathworld.wolfram.com
15. [pid:1985] MathWorld URL: http://mathworld.wolfram.com	15. [pid:7] Computational Geometry, Algorithms and Applications URL: http://www.cs.uu.nl/geobook
Max	**SALSA**
1. [pid:4] Geometry in Action URL: http://www.ics.uci.edu/ eppstein/geom.html	1. [pid:4] Geometry in Action URL: http://www.ics.uci.edu/ eppstein/geom.html
2. [pid:6] The former CGAL home page URL: http://www.cs.uu.nl/CGAL	2. [pid:6] The former CGAL home page URL: http://www.cs.uu.nl/CGAL
3. [pid:237] The Geometry Junkyard URL: http://www.ics.uci.edu/ eppstein/junkyard	3. [pid:1047] ACM: Association for Computing Machinery, the world's first educa URL: http://www.acm.org
4. [pid:3] Computational Geometry Resources URL: http://www.scs.carleton.ca/ csgs/resources/cg.html	4. [pid:1743] WebCT.com URL: http://www.webct.com
5. [pid:1301] Computational Geometry Pages URL: http://compgeom.cs.uiuc.edu/ jeffe/compgeom	5. [pid:1744] DREXEL UNIVERSITY: Philadelphia, PA URL: http://www.drexel.edu
6. [pid:155] David Eppstein URL: http://www.ics.uci.edu/ eppstein	6. [pid:1745] A Virtual Math Community: Members helping Members URL: http://www.drexel.edu/ia/mathforum
7. [pid:1047] ACM: Association for Computing Machinery, the world's first educa URL: http://www.acm.org	7. [pid:1] Springer Link - Publication URL: http://link.springer.de/link/service/journals/00454
8. [pid:1] Springer Link - Publication URL: http://link.springer.de/link/service/journals/00454	8. [pid:3] Computational Geometry Resources URL: http://www.scs.carleton.ca/ csgs/resources/cg.html
9. [pid:235] The Stony Brook Algorithm Repository URL: http://www.cs.sunysb.edu/ algorith	9. [pid:13] Computational Geometry URL: http://www.elsevier.nl/locate/comgeo
10. [pid:1985] MathWorld URL: http://mathworld.wolfram.com	10. [pid:237] The Geometry Junkyard URL: http://www.ics.uci.edu/ eppstein/junkyard
11. [pid:7] Computational Geometry, Algorithms and Applications URL: http://www.cs.uu.nl/geobook	11. [pid:7] Computational Geometry, Algorithms and Applications URL: http://www.cs.uu.nl/geobook
12. [pid:161] Joseph O'Rourke URL: http://cs.smith.edu/ orourke	12. [pid:9] The CGAL Home Page URL: http://www.cgal.org
13. [pid:233] LEDA moved to Algorithmic Solutions Software GmbH URL: http://www.mpi-sb.mpg.de/LEDA/leda.html	13. [pid:1301] Computational Geometry Pages URL: http://compgeom.cs.uiuc.edu/ jeffe/compgeom
14. [pid:788] Mesh Generation & Grid Generation on the Web URL: http://www-users.informatik.rwth-aachen.de/ roberts/meshgeneratio	14. [pid:1985] MathWorld URL: http://mathworld.wolfram.com
15. [pid:5] Computational Geometry on the WWW URL: http://www.dcc.unicamp.br/ guialbu/geompages.html	15. [pid:2275] Google URL: http://www.google.com

Table B.14: *Top 15 results for query "computational geometry"*

HITS	**HubAvg**
1. [pid:1] Death Penalty Information Center URL: http://www.deathpenaltyinfo.org	1. [pid:3858] CBS.SportsLine.com URL: http://cbs.sportsline.com
2. [pid:3] NCADP - National Coalition to Abolish the Death Penalty URL: http://www.ncadp.org	2. [pid:174] TDCJ - Statistics - Home Page URL: http://www.tdcj.state.tx.us/statistics/stats-home.htm
3. [pid:9] CUADP: For Alternatives to the Death Penalty. URL: http://www.cuadp.org	3. [pid:3857] CBSNews.com URL: http://www.cbsnews.com
4. [pid:87] Death Penalty : Sister Helen Prejean : Capital Punishment URL: http://www.moratorium2000.org	4. [pid:1] Death Penalty Information Center URL: http://www.deathpenaltyinfo.org
5. [pid:4] Death Penalty Information URL: http://sun.soci.niu.edu/ critcrim/dp/dp.html	5. [pid:2] Pro-death penalty.com URL: http://www.prodeathpenalty.com
6. [pid:6] Death Penalty Focus URL: http://www.deathpenalty.org	6. [pid:7] Campaign To End The Death Penalty URL: http://www.nodeathpenalty.org
7. [pid:7] Campaign To End The Death Penalty URL: http://www.nodeathpenalty.org	7. [pid:11] American Civil Liberties Union URL: http://www.aclu.org/death-penalty
8. [pid:5] Death Penalty Links URL: http://www.derechos.org/dp	8. [pid:87] Death Penalty : Sister Helen Prejean : Capital Punishment URL: http://www.moratorium2000.org
9. [pid:19] Virginians for Alternatives to the Death Penalty URL: http://www.vadp.org	9. [pid:10] LII: Law about...the Death Penalty URL: http://www.law.cornell.edu/topics/death_penalty.html
10. [pid:1214] Ohioans To Stop Executions URL: http://www.otse.org	10. [pid:230] ABCNEWS.com: Home URL: http://www.abcnews.com
11. [pid:797] Equal Justice USA, A Project of the Quixote Center URL: http://www.quixote.org/ej	11. [pid:3] NCADP - National Coalition to Abolish the Death Penalty URL: http://www.ncadp.org
12. [pid:1343] Murder Victims Families for Reconcilition URL: http://www.mvfr.org	12. [pid:4179] AP Digital - Providing Breaking News from The Associated Press URL: http://www.apdigitalnews.com
13. [pid:8] Death Penalty News & Updates URL: http://www.smu.edu/ deathpen	13. [pid:3646] Disney Online - The Official Home Page of The Walt Disney Company URL: http://www.disney.com
14. [pid:10] LII: Law about...the Death Penalty URL: http://www.law.cornell.edu/topics/death_penalty.html	14. [pid:3982] Salon.com URL: http://www.salon.com
15. [pid:25] Oklahoma Coalition to Abolish the Death Penalty URL: http://www.ocadp.org	15. [pid:235] AltaVista URL: http://www.altavista.com
AT-Avg	**Norm (2)**
1. [pid:1] Death Penalty Information Center URL: http://www.deathpenaltyinfo.org	1. [pid:1] Death Penalty Information Center URL: http://www.deathpenaltyinfo.org
2. [pid:3] NCADP - National Coalition to Abolish the Death Penalty URL: http://www.ncadp.org	2. [pid:3] NCADP - National Coalition to Abolish the Death Penalty URL: http://www.ncadp.org
3. [pid:9] CUADP: For Alternatives to the Death Penalty. URL: http://www.cuadp.org	3. [pid:2] Pro-death penalty.com URL: http://www.prodeathpenalty.com
4. [pid:2] Pro-death penalty.com URL: http://www.prodeathpenalty.com	4. [pid:9] CUADP: For Alternatives to the Death Penalty. URL: http://www.cuadp.org
5. [pid:87] Death Penalty : Sister Helen Prejean : Capital Punishment URL: http://www.moratorium2000.org	5. [pid:87] Death Penalty : Sister Helen Prejean : Capital Punishment URL: http://www.moratorium2000.org
6. [pid:7] Campaign To End The Death Penalty URL: http://www.nodeathpenalty.org	6. [pid:7] Campaign To End The Death Penalty URL: http://www.nodeathpenalty.org
7. [pid:5] Death Penalty Links URL: http://www.derechos.org/dp	7. [pid:5] Death Penalty Links URL: http://www.derechos.org/dp
8. [pid:4] Death Penalty Information URL: http://sun.soci.niu.edu/ critcrim/dp/dp.html	8. [pid:4] Death Penalty Information URL: http://sun.soci.niu.edu/ critcrim/dp/dp.html
9. [pid:6] Death Penalty Focus URL: http://www.deathpenalty.org	9. [pid:6] Death Penalty Focus URL: http://www.deathpenalty.org
10. [pid:1343] Murder Victims Families for Reconcilition URL: http://www.mvfr.org	10. [pid:1343] Murder Victims Families for Reconcilition URL: http://www.mvfr.org
11. [pid:11] American Civil Liberties Union URL: http://www.aclu.org/death-penalty	11. [pid:11] American Civil Liberties Union URL: http://www.aclu.org/death-penalty
12. [pid:8] Death Penalty News & Updates URL: http://www.smu.edu/ deathpen	12. [pid:8] Death Penalty News & Updates URL: http://www.smu.edu/ deathpen
13. [pid:1214] Ohioans To Stop Executions URL: http://www.otse.org	13. [pid:1214] Ohioans To Stop Executions URL: http://www.otse.org
14. [pid:797] Equal Justice USA, A Project of the Quixote Center URL: http://www.quixote.org/ej	14. [pid:797] Equal Justice USA, A Project of the Quixote Center URL: http://www.quixote.org/ej
15. [pid:594] American Civil Liberties Union URL: http://www.aclu.org	15. [pid:10] LII: Law about...the Death Penalty URL: http://www.law.cornell.edu/topics/death_penalty.html
Max	**SALSA**
1. [pid:1] Death Penalty Information Center URL: http://www.deathpenaltyinfo.org	1. [pid:1] Death Penalty Information Center URL: http://www.deathpenaltyinfo.org
2. [pid:3] NCADP - National Coalition to Abolish the Death Penalty URL: http://www.ncadp.org	2. [pid:3] NCADP - National Coalition to Abolish the Death Penalty URL: http://www.ncadp.org
3. [pid:2] Pro-death penalty.com URL: http://www.prodeathpenalty.com	3. [pid:2] Pro-death penalty.com URL: http://www.prodeathpenalty.com
4. [pid:9] CUADP: For Alternatives to the Death Penalty. URL: http://www.cuadp.org	4. [pid:3858] CBS.SportsLine.com URL: http://cbs.sportsline.com
5. [pid:87] Death Penalty : Sister Helen Prejean : Capital Punishment URL: http://www.moratorium2000.org	5. [pid:9] CUADP: For Alternatives to the Death Penalty. URL: http://www.cuadp.org
6. [pid:5] Death Penalty Links URL: http://www.derechos.org/dp	6. [pid:4] Death Penalty Information URL: http://sun.soci.niu.edu/ critcrim/dp/dp.html
7. [pid:7] Campaign To End The Death Penalty URL: http://www.nodeathpenalty.org	7. [pid:594] American Civil Liberties Union URL: http://www.aclu.org
8. [pid:4] Death Penalty Information URL: http://sun.soci.niu.edu/ critcrim/dp/dp.html	8. [pid:7] Campaign To End The Death Penalty URL: http://www.nodeathpenalty.org
9. [pid:1343] Murder Victims Families for Reconcilition URL: http://www.mvfr.org	9. [pid:87] Death Penalty : Sister Helen Prejean : Capital Punishment URL: http://www.moratorium2000.org
10. [pid:11] American Civil Liberties Union URL: http://www.aclu.org/death-penalty	10. [pid:5] Death Penalty Links URL: http://www.derechos.org/dp
11. [pid:594] American Civil Liberties Union URL: http://www.aclu.org	11. [pid:3590] WILPF - Women's International League for Peace and Freedom URL: http://www.wilpf.org
12. [pid:8] Death Penalty News & Updates URL: http://www.smu.edu/ deathpen	12. [pid:6] Death Penalty Focus URL: http://www.deathpenalty.org
13. [pid:6] Death Penalty Focus URL: http://www.deathpenalty.org	13. [pid:11] American Civil Liberties Union URL: http://www.aclu.org/death-penalty
14. [pid:10] LII: Law about...the Death Penalty URL: http://www.law.cornell.edu/topics/death_penalty.html	14. [pid:999] FindLaw : Constitutional Law Center URL: http://supreme.lp.findlaw.com
15. [pid:1214] Ohioans To Stop Executions URL: http://www.otse.org	15. [pid:8] Death Penalty News & Updates URL: http://www.smu.edu/ deathpen

Table B.15: *Top 15 results for query "death penalty"*

HITS	**HubAvg**
1. [pid:1226] NCBI HomePage URL: http://www.ncbi.nlm.nih.gov	1. [pid:1226] NCBI HomePage URL: http://www.ncbi.nlm.nih.gov
2. [pid:1231] The Genome Database URL: http://www.gdb.org	2. [pid:1849] National Institutes of Health (NIH) URL: http://www.nih.gov
3. [pid:1260] The Wellcome Trust Sanger Institute URL: http://www.sanger.ac.uk	3. [pid:1108] U.S. National Library of Medicine URL: http://www.nlm.nih.gov
4. [pid:1241] The Institute for Genomic Research URL: http://www.tigr.org	4. [pid:1495] OMIM Home Page – Online Mendelian Inheritance in Man URL: http://www3.ncbi.nlm.nih.gov/Omim
5. [pid:1495] OMIM Home Page – Online Mendelian Inheritance in Man URL: http://www3.ncbi.nlm.nih.gov/Omim	5. [pid:4984] www.genome.gov URL: http://www.nhgri.nih.gov
6. [pid:1210] Whitehead Institute/MIT Center for Genome Research URL: http://www-genome.wi.mit.edu	6. [pid:1231] The Genome Database URL: http://www.gdb.org
7. [pid:4984] www.genome.gov URL: http://www.nhgri.nih.gov	7. [pid:1] Genetic Alliance, Inc. URL: http://www.geneticalliance.org
8. [pid:1849] National Institutes of Health (NIH) URL: http://www.nih.gov	8. [pid:1260] The Wellcome Trust Sanger Institute URL: http://www.sanger.ac.uk
9. [pid:1224] European Bioinformatics Institute URL: http://www.ebi.ac.uk	9. [pid:1799] Entrez-PubMed URL: http://www4.ncbi.nlm.nih.gov/PubMed
10. [pid:1240] UK MRC HGMP-RC URL: http://www.hgmp.mrc.ac.uk	10. [pid:1210] Whitehead Institute/MIT Center for Genome Research URL: http://www-genome.wi.mit.edu
11. [pid:1189] MGI 2.96 - Mouse Genome Informatics URL: http://www.informatics.jax.org	11. [pid:1241] The Institute for Genomic Research URL: http://www.tigr.org
12. [pid:1253] NCGR Homepage URL: http://www.ncgr.org	12. [pid:1800] March of Dimes Home Page URL: http://www.modimes.org
13. [pid:1225] EMBL - Basic Research in Molecular Biology URL: http://www.embl-heidelberg.de	13. [pid:55] GeneTests Home Page URL: http://www.geneclinics.org
14. [pid:1235] GenomeNet URL: http://www.genome.ad.jp	14. [pid:9] National Society of Genetic Counselors, Inc. URL: http://www.nsgc.org
15. [pid:1262] Stanford Human Genome Center URL: http://www-shgc.stanford.edu	15. [pid:1224] European Bioinformatics Institute URL: http://www.ebi.ac.uk
AT-Avg	**Norm (2)**
1. [pid:1226] NCBI HomePage URL: http://www.ncbi.nlm.nih.gov	1. [pid:1226] NCBI HomePage URL: http://www.ncbi.nlm.nih.gov
2. [pid:1849] National Institutes of Health (NIH) URL: http://www.nih.gov	2. [pid:1849] National Institutes of Health (NIH) URL: http://www.nih.gov
3. [pid:1231] The Genome Database URL: http://www.gdb.org	3. [pid:1231] The Genome Database URL: http://www.gdb.org
4. [pid:1495] OMIM Home Page – Online Mendelian Inheritance in Man URL: http://www3.ncbi.nlm.nih.gov/Omim	4. [pid:1495] OMIM Home Page – Online Mendelian Inheritance in Man URL: http://www3.ncbi.nlm.nih.gov/Omim
5. [pid:4984] www.genome.gov URL: http://www.nhgri.nih.gov	5. [pid:4984] www.genome.gov URL: http://www.nhgri.nih.gov
6. [pid:1260] The Wellcome Trust Sanger Institute URL: http://www.sanger.ac.uk	6. [pid:1260] The Wellcome Trust Sanger Institute URL: http://www.sanger.ac.uk
7. [pid:1241] The Institute for Genomic Research URL: http://www.tigr.org	7. [pid:1241] The Institute for Genomic Research URL: http://www.tigr.org
8. [pid:1210] Whitehead Institute/MIT Center for Genome Research URL: http://www-genome.wi.mit.edu	8. [pid:1210] Whitehead Institute/MIT Center for Genome Research URL: http://www-genome.wi.mit.edu
9. [pid:1108] U.S. National Library of Medicine URL: http://www.nlm.nih.gov	9. [pid:1224] European Bioinformatics Institute URL: http://www.ebi.ac.uk
10. [pid:1224] European Bioinformatics Institute URL: http://www.ebi.ac.uk	10. [pid:1108] U.S. National Library of Medicine URL: http://www.nlm.nih.gov
11. [pid:1189] MGI 2.96 - Mouse Genome Informatics URL: http://www.informatics.jax.org	11. [pid:1189] MGI 2.96 - Mouse Genome Informatics URL: http://www.informatics.jax.org
12. [pid:1843] Entrez Home URL: http://www.ncbi.nlm.nih.gov/Entrez	12. [pid:1843] Entrez Home URL: http://www.ncbi.nlm.nih.gov/Entrez
13. [pid:1240] UK MRC HGMP-RC URL: http://www.hgmp.mrc.ac.uk	13. [pid:1240] UK MRC HGMP-RC URL: http://www.hgmp.mrc.ac.uk
14. [pid:1799] Entrez-PubMed URL: http://www4.ncbi.nlm.nih.gov/PubMed	14. [pid:1799] Entrez-PubMed URL: http://www4.ncbi.nlm.nih.gov/PubMed
15. [pid:55] GeneTests Home Page URL: http://www.geneclinics.org	15. [pid:1253] NCGR Homepage URL: http://www.ncgr.org
Max	**SALSA**
1. [pid:1226] NCBI HomePage URL: http://www.ncbi.nlm.nih.gov	1. [pid:1226] NCBI HomePage URL: http://www.ncbi.nlm.nih.gov
2. [pid:1849] National Institutes of Health (NIH) URL: http://www.nih.gov	2. [pid:1849] National Institutes of Health (NIH) URL: http://www.nih.gov
3. [pid:1231] The Genome Database URL: http://www.gdb.org	3. [pid:1495] OMIM Home Page – Online Mendelian Inheritance in Man URL: http://www3.ncbi.nlm.nih.gov/Omim
4. [pid:1495] OMIM Home Page – Online Mendelian Inheritance in Man URL: http://www3.ncbi.nlm.nih.gov/Omim	4. [pid:4984] www.genome.gov URL: http://www.nhgri.nih.gov
5. [pid:4984] www.genome.gov URL: http://www.nhgri.nih.gov	5. [pid:1231] The Genome Database URL: http://www.gdb.org
6. [pid:1260] The Wellcome Trust Sanger Institute URL: http://www.sanger.ac.uk	6. [pid:1] Genetic Alliance, Inc. URL: http://www.geneticalliance.org
7. [pid:1210] Whitehead Institute/MIT Center for Genome Research URL: http://www-genome.wi.mit.edu	7. [pid:1260] The Wellcome Trust Sanger Institute URL: http://www.sanger.ac.uk
8. [pid:1241] The Institute for Genomic Research URL: http://www.tigr.org	8. [pid:1108] U.S. National Library of Medicine URL: http://www.nlm.nih.gov
9. [pid:1224] European Bioinformatics Institute URL: http://www.ebi.ac.uk	9. [pid:2] The Genetic Algorithms Archive URL: http://www.aic.nrl.navy.mil/galist
10. [pid:1108] U.S. National Library of Medicine URL: http://www.nlm.nih.gov	10. [pid:1241] The Institute for Genomic Research URL: http://www.tigr.org
11. [pid:1843] Entrez Home URL: http://www.ncbi.nlm.nih.gov/Entrez	11. [pid:55] GeneTests Home Page URL: http://www.geneclinics.org
12. [pid:1189] MGI 2.96 - Mouse Genome Informatics URL: http://www.informatics.jax.org	12. [pid:1210] Whitehead Institute/MIT Center for Genome Research URL: http://www-genome.wi.mit.edu
13. [pid:1799] Entrez-PubMed URL: http://www4.ncbi.nlm.nih.gov/PubMed	13. [pid:1282] Adobe Reader - Download URL: http://www.adobe.com/products/acrobat/readstep2.html
14. [pid:1240] UK MRC HGMP-RC URL: http://www.hgmp.mrc.ac.uk	14. [pid:1799] Entrez-PubMed URL: http://www4.ncbi.nlm.nih.gov/PubMed
15. [pid:1225] EMBL - Basic Research in Molecular Biology URL: http://www.embl-heidelberg.de	15. [pid:9] National Society of Genetic Counselors, Inc. URL: http://www.nsgc.org

Table B.16: *Top 15 results for query "genetic"*

HITS
1. [pid:8] The Geometry Center Welcome Page URL: http://freeabel.geom.umn.edu
2. [pid:4] Geometry in Action URL: http://www.ics.uci.edu/ eppstein/geom.html
3. [pid:3] The Geometry Junkyard URL: http://www.ics.uci.edu/ eppstein/junkyard
4. [pid:3339] MathWorld URL: http://mathworld.wolfram.com
5. [pid:908] Euclid's Elements, Introduction URL: http://aleph0.clarku.edu/ djoyce/java/elements/elements.html
6. [pid:2303] A Gallery of Interactive On-Line Geometry URL: http://www.geom.umn.edu/apps/gallery.html
7. [pid:1] The Math Forum Home Page URL: http://mathforum.org
8. [pid:13] Geometry Formulas and Facts URL: http://www.geom.umn.edu/docs/reference/CRC-formulas
9. [pid:627] Wolfram Research, Inc. URL: http://www.wri.com
10. [pid:14] Directory of Computational Geometry Software URL: http://www.geom.umn.edu/software/cglist
11. [pid:23] GANG — Geometry Analysis Numerics Graphics URL: http://www.gang.umass.edu
12. [pid:823] National Council of Teachers of Mathematics - More and better URL: http://www.nctm.org
13. [pid:6] Geometry and Topology URL: http://www.maths.warwick.ac.uk/gt
14. [pid:204] WebCT.com URL: http://www.webct.com
15. [pid:26] Computational Geometry Resources URL: http://www.scs.carleton.ca/ csgs/resources/cg.html

HubAvg
1. [pid:204] WebCT.com URL: http://www.webct.com
2. [pid:8] The Geometry Center Welcome Page URL: http://freeabel.geom.umn.edu
3. [pid:3] The Geometry Junkyard URL: http://www.ics.uci.edu/ eppstein/junkyard
4. [pid:29] Connected Geometry Home Page URL: http://www.edc.org/LTT/ConnGeo
5. [pid:16] Cynthia Lanius' Lessons: Geometry Online URL: http://math.rice.edu/ lanius/Geom
6. [pid:90] Dynamic Geometry Home Page URL: http://www.edc.org/LTT/DG
7. [pid:116] C.a.R. URL: http://mathsrv.ku-eichstaett.de/MGF/homes/grothmann/car.html
8. [pid:4] Geometry in Action URL: http://www.ics.uci.edu/ eppstein/geom.html
9. [pid:46] Geometry Step by Step from the Land of the Incas, Intro. Antonio URL: http://agutie.homestead.com
10. [pid:66] The Interactive Geometry Software Cinderella - Redirection URL: http://www.cinderella.de
11. [pid:49] The Geometry of War URL: http://www.mhs.ox.ac.uk/geometry/content.htm
12. [pid:3339] MathWorld URL: http://mathworld.wolfram.com
13. [pid:1468] Mathematics Archives - K12 Internet Sites URL: http://archives.math.utk.edu/k12.html
14. [pid:18] Native American Geometry URL: http://www.earthmeasure.com
15. [pid:144] WebEQ has moved URL: http://www.webeq.com

AT-Avg
1. [pid:8] The Geometry Center Welcome Page URL: http://freeabel.geom.umn.edu
2. [pid:4] Geometry in Action URL: http://www.ics.uci.edu/ eppstein/geom.html
3. [pid:3] The Geometry Junkyard URL: http://www.ics.uci.edu/ eppstein/junkyard
4. [pid:3339] MathWorld URL: http://mathworld.wolfram.com
5. [pid:908] Euclid's Elements, Introduction URL: http://aleph0.clarku.edu/ djoyce/java/elements/elements.html
6. [pid:2303] A Gallery of Interactive On-Line Geometry URL: http://www.geom.umn.edu/apps/gallery.html
7. [pid:1] The Math Forum Home Page URL: http://mathforum.org
8. [pid:204] WebCT.com URL: http://www.webct.com
9. [pid:13] Geometry Formulas and Facts URL: http://www.geom.umn.edu/docs/reference/CRC-formulas
10. [pid:823] National Council of Teachers of Mathematics - More and better URL: http://www.nctm.org
11. [pid:23] GANG — Geometry Analysis Numerics Graphics URL: http://www.gang.umass.edu
12. [pid:627] Wolfram Research, Inc. URL: http://www.wri.com
13. [pid:14] Directory of Computational Geometry Software URL: http://www.geom.umn.edu/software/cglist
14. [pid:6] Geometry and Topology URL: http://www.maths.warwick.ac.uk/gt
15. [pid:830] ENC Online: A K-12 math and science teacher center. URL: http://www.enc.org

Norm (2)
1. [pid:8] The Geometry Center Welcome Page URL: http://freeabel.geom.umn.edu
2. [pid:4] Geometry in Action URL: http://www.ics.uci.edu/ eppstein/geom.html
3. [pid:3] The Geometry Junkyard URL: http://www.ics.uci.edu/ eppstein/junkyard
4. [pid:3339] MathWorld URL: http://mathworld.wolfram.com
5. [pid:908] Euclid's Elements, Introduction URL: http://aleph0.clarku.edu/ djoyce/java/elements/elements.html
6. [pid:204] WebCT.com URL: http://www.webct.com
7. [pid:2303] A Gallery of Interactive On-Line Geometry URL: http://www.geom.umn.edu/apps/gallery.html
8. [pid:1] The Math Forum Home Page URL: http://mathforum.org
9. [pid:23] GANG — Geometry Analysis Numerics Graphics URL: http://www.gang.umass.edu
10. [pid:13] Geometry Formulas and Facts URL: http://www.geom.umn.edu/docs/reference/CRC-formulas
11. [pid:823] National Council of Teachers of Mathematics - More and better URL: http://www.nctm.org
12. [pid:627] Wolfram Research, Inc. URL: http://www.wri.com
13. [pid:6] Geometry and Topology URL: http://www.maths.warwick.ac.uk/gt
14. [pid:830] ENC Online: A K-12 math and science teacher center. URL: http://www.enc.org
15. [pid:14] Directory of Computational Geometry Software URL: http://www.geom.umn.edu/software/cglist

Max
1. [pid:8] The Geometry Center Welcome Page URL: http://freeabel.geom.umn.edu
2. [pid:4] Geometry in Action URL: http://www.ics.uci.edu/ eppstein/geom.html
3. [pid:3] The Geometry Junkyard URL: http://www.ics.uci.edu/ eppstein/junkyard
4. [pid:3339] MathWorld URL: http://mathworld.wolfram.com
5. [pid:2303] A Gallery of Interactive On-Line Geometry URL: http://www.geom.umn.edu/apps/gallery.html
6. [pid:204] WebCT.com URL: http://www.webct.com
7. [pid:908] Euclid's Elements, Introduction URL: http://aleph0.clarku.edu/ djoyce/java/elements/elements.html
8. [pid:1] The Math Forum Home Page URL: http://mathforum.org
9. [pid:23] GANG — Geometry Analysis Numerics Graphics URL: http://www.gang.umass.edu
10. [pid:13] Geometry Formulas and Facts URL: http://www.geom.umn.edu/docs/reference/CRC-formulas
11. [pid:823] National Council of Teachers of Mathematics - More and better URL: http://www.nctm.org
12. [pid:627] Wolfram Research, Inc. URL: http://www.wri.com
13. [pid:830] ENC Online: A K-12 math and science teacher center. URL: http://www.enc.org
14. [pid:6] Geometry and Topology URL: http://www.maths.warwick.ac.uk/gt
15. [pid:14] Directory of Computational Geometry Software URL: http://www.geom.umn.edu/software/cglist

SALSA
1. [pid:8] The Geometry Center Welcome Page URL: http://freeabel.geom.umn.edu
2. [pid:204] WebCT.com URL: http://www.webct.com
3. [pid:4] Geometry in Action URL: http://www.ics.uci.edu/ eppstein/geom.html
4. [pid:3339] MathWorld URL: http://mathworld.wolfram.com
5. [pid:3] The Geometry Junkyard URL: http://www.ics.uci.edu/ eppstein/junkyard
6. [pid:6] Geometry and Topology URL: http://www.maths.warwick.ac.uk/gt
7. [pid:908] Euclid's Elements, Introduction URL: http://aleph0.clarku.edu/ djoyce/java/elements/elements.html
8. [pid:10] SpringerLink - Publication URL: http://link.springer.de/link/service/journals/00454
9. [pid:1] The Math Forum Home Page URL: http://mathforum.org
10. [pid:7] Empty title field URL: http://www.ams.org/ecgd
11. [pid:2303] A Gallery of Interactive On-Line Geometry URL: http://www.geom.umn.edu/apps/gallery.html
12. [pid:627] Wolfram Research, Inc. URL: http://www.wri.com
13. [pid:66] The Interactive Geometry Software Cinderella - Redirection URL: http://www.cinderella.de
14. [pid:1185] National Science Foundation (NSF) - Home Page URL: http://www.nsf.gov
15. [pid:13] Geometry Formulas and Facts URL: http://www.geom.umn.edu/docs/reference/CRC-formulas

Table B.17: *Top 15 results for query "geometry"*

HITS	**HubAvg**
1. [pid:2203] INDYMEDIA TIJUANA :: centro de medios independientes URL: http://www.tijuanaimc.org	1. [pid:2203] INDYMEDIA TIJUANA :: centro de medios independientes URL: http://www.tijuanaimc.org
2. [pid:2205] Baltimore Independent Media Center: home URL: http://baltimoreimc.org	2. [pid:571] Independent Media Center - URL: http://www.indymedia.org
3. [pid:2217] Empty title field URL: http://sdimc.org	3. [pid:2205] Baltimore Independent Media Center: home URL: http://baltimoreimc.org
4. [pid:2181] Melbourne Independent Media Center URL: http://www.melbourne.indymedia.org	4. [pid:2181] Melbourne Independent Media Center URL: http://www.melbourne.indymedia.org
5. [pid:2219] Urbana-Champaign Independent Media Center URL: http://www.ucimc.org	5. [pid:2219] Urbana-Champaign Independent Media Center URL: http://www.ucimc.org
6. [pid:2208] Danbury, CT Independent Media Center :: URL: http://www.madhattersimc.org	6. [pid:2180] Indymedia - news - Aotearoa Independent Media Centre URL: http://www.indymedia.org.nz
7. [pid:2204] IndyMedia Center - URL: http://indymedia.org.il	7. [pid:2217] Empty title field URL: http://sdimc.org
8. [pid:2180] Indymedia - news - Aotearoa Independent Media Centre (AIMC) URL: http://www.indymedia.org.nz	8. [pid:2208] Danbury, CT Independent Media Center :: URL: http://www.madhattersimc.org
9. [pid:2187] Vaikuttava Tietotoimisto (VAI) - URL: http://www.vaikuttava.net	9. [pid:2204] IndyMedia Center - URL: http://indymedia.org.il
10. [pid:2179] Adelaide indymedia - webcast news - page 1 URL: http://adelaide.indymedia.org.au	10. [pid:2179] Adelaide indymedia - webcast news - page 1 URL: http://adelaide.indymedia.org.au
11. [pid:571] Independent Media Center - URL: http://www.indymedia.org	11. [pid:2187] Vaikuttava Tietotoimisto (VAI) - URL: http://www.vaikuttava.net
12. [pid:2213] North Texas Independent Media Center :: Reclaim the Media URL: http://www.ntimc.org	12. [pid:2213] North Texas Independent Media Center :: Reclaim the Media URL: http://www.ntimc.org
13. [pid:2183] South Africa Independent Media Center URL: http://southafrica.indymedia.org	13. [pid:2198] indymedia uk URL: http://www.indymedia.org.uk
14. [pid:2196] Russia Indymedia - URL: http://russia.indymedia.org	14. [pid:2183] South Africa Independent Media Center URL: http://southafrica.indymedia.org
15. [pid:2190] Indymedia Italia - network di media indipendenti URL: http://italy.indymedia.org	15. [pid:2214] Empty title field URL: http://ohiovalleyimc.org

AT-Avg	**Norm (2)**
1. [pid:571] Independent Media Center - URL: http://www.indymedia.org	1. [pid:571] Independent Media Center - URL: http://www.indymedia.org
2. [pid:2203] INDYMEDIA TIJUANA :: centro de medios independientes URL: http://www.tijuanaimc.org	2. [pid:2203] INDYMEDIA TIJUANA :: centro de medios independientes URL: http://www.tijuanaimc.org
3. [pid:2205] Baltimore Independent Media Center: home URL: http://baltimoreimc.org	3. [pid:2205] Baltimore Independent Media Center: home URL: http://baltimoreimc.org
4. [pid:2181] Melbourne Independent Media Center URL: http://www.melbourne.indymedia.org	4. [pid:2181] Melbourne Independent Media Center URL: http://www.melbourne.indymedia.org
5. [pid:2180] Indymedia - news - Aotearoa Independent Media Centre (AIMC) URL: http://www.indymedia.org.nz	5. [pid:2217] Empty title field URL: http://sdimc.org
6. [pid:2219] Urbana-Champaign Independent Media Center URL: http://www.ucimc.org	6. [pid:2219] Urbana-Champaign Independent Media Center URL: http://www.ucimc.org
7. [pid:2217] Empty title field URL: http://sdimc.org	7. [pid:2180] Indymedia - news - Aotearoa Independent Media Centre URL: http://www.indymedia.org.nz
8. [pid:2208] Danbury, CT Independent Media Center :: URL: http://www.madhattersimc.org	8. [pid:2208] Danbury, CT Independent Media Center URL: http://www.madhattersimc.org
9. [pid:2204] IndyMedia Center - URL: http://indymedia.org.il	9. [pid:2204] IndyMedia Center - URL: http://indymedia.org.il
10. [pid:2179] Adelaide indymedia - webcast news - page 1 URL: http://adelaide.indymedia.org.au	10. [pid:2179] Adelaide indymedia - webcast news - page 1 URL: http://adelaide.indymedia.org.au
11. [pid:2198] indymedia uk URL: http://www.indymedia.org.uk	11. [pid:2187] Vaikuttava Tietotoimisto (VAI) - URL: http://www.vaikuttava.net
12. [pid:2187] Vaikuttava Tietotoimisto (VAI) - URL: http://www.vaikuttava.net	12. [pid:2198] indymedia uk URL: http://www.indymedia.org.uk
13. [pid:2213] North Texas Independent Media Center :: Reclaim the Media URL: http://www.ntimc.org	13. [pid:2213] North Texas Independent Media Center :: Reclaim the Media URL: http://www.ntimc.org
14. [pid:2183] South Africa Independent Media Center URL: http://southafrica.indymedia.org	14. [pid:2183] South Africa Independent Media Center URL: http://southafrica.indymedia.org
15. [pid:2196] Russia Indymedia - URL: http://russia.indymedia.org	15. [pid:2190] Indymedia Italia - network di media indipendenti URL: http://italy.indymedia.org

Max	**SALSA**
1. [pid:571] Independent Media Center - URL: http://www.indymedia.org	1. [pid:571] Independent Media Center - URL: http://www.indymedia.org
2. [pid:2198] indymedia uk URL: http://www.indymedia.org.uk	2. [pid:2198] indymedia uk URL: http://www.indymedia.org.uk
3. [pid:2203] INDYMEDIA TIJUANA :: centro de medios independientes URL: http://www.tijuanaimc.org	3. [pid:1] International Forum on Globalization URL: http://www.ifg.org
4. [pid:2205] Baltimore Independent Media Center: home URL: http://baltimoreimc.org	4. [pid:2025] WTO — Welcome to the WTO website URL: http://www.wto.org
5. [pid:2181] Melbourne Independent Media Center URL: http://www.melbourne.indymedia.org	5. [pid:2203] INDYMEDIA TIJUANA :: centro de medios independientes URL: http://www.tijuanaimc.org
6. [pid:2180] Indymedia - news - Aotearoa Independent Media Centre (AIMC) URL: http://www.indymedia.org.nz	6. [pid:2265] The Institute for Deep Ecology: Home Page URL: http://www.deep-ecology.org
7. [pid:2219] Urbana-Champaign Independent Media Center URL: http://www.ucimc.org	7. [pid:338] The World Bank Group URL: http://www.worldbank.org
8. [pid:2204] IndyMedia Center - URL: http://indymedia.org.il	8. [pid:2205] Baltimore Independent Media Center: home URL: http://baltimoreimc.org
9. [pid:2217] Empty title field URL: http://sdimc.org	9. [pid:2180] Indymedia - news - Aotearoa Independent Media Centre URL: http://www.indymedia.org.nz
10. [pid:2208] Danbury, CT Independent Media Center :: URL: http://www.madhattersimc.org	10. [pid:2181] Melbourne Independent Media Center URL: http://www.melbourne.indymedia.org
11. [pid:2179] Adelaide indymedia - webcast news - page 1 URL: http://adelaide.indymedia.org.au	11. [pid:2219] Urbana-Champaign Independent Media Center URL: http://www.ucimc.org
12. [pid:2187] Vaikuttava Tietotoimisto (VAI) - URL: http://www.vaikuttava.net	12. [pid:2204] IndyMedia Center - URL: http://indymedia.org.il
13. [pid:2183] South Africa Independent Media Center URL: http://southafrica.indymedia.org	13. [pid:2208] Danbury, CT Independent Media Center URL: http://www.madhattersimc.org
14. [pid:2213] North Texas Independent Media Center :: Reclaim the Media URL: http://www.ntimc.org	14. [pid:2217] Empty title field URL: http://sdimc.org
15. [pid:2214] Empty title field URL: http://ohiovalleyimc.org	15. [pid:2179] Adelaide indymedia - webcast news - page 1 URL: http://adelaide.indymedia.org.au

Table B.18: *Top 15 results for query "globalization"*

HITS	**HubAvg**
1. [pid:1938] Coffee Club URL: http://www.batavia-rof.com	1. [pid:340] National Rifle Association - MyNRA URL: http://www.nra.org
2. [pid:1961] Hotel and Travel URL: http://www.bwdriftwood.com	2. [pid:1] The Brady Campaign to Prevent Gun Violence URL: http://www.handguncontrol.org
3. [pid:1937] Basement Writers URL: http://www.basement-writers.com	3. [pid:337] Gun Owners of America URL: http://www.gunowners.org
4. [pid:1942] Before Today URL: http://www.beforetoday.com	4. [pid:352] The Violence Policy Center URL: http://www.vpc.org
5. [pid:1945] Bennett Boxing URL: http://www.bennettboxing.com	5. [pid:336] CCRKBA Home Page URL: http://www.ccrkba.org
6. [pid:1950] Boeing Mail URL: http://www.boeingmail.com	6. [pid:42] Coalition to Stop Gun Violence/Educational Fund to Stop Gun Viole URL: http://www.gunfree.org
7. [pid:1960] Burdan USA URL: http://www.burdanusa.com	7. [pid:339] Jews for the Preservation of Firearms Ownership URL: http://www.jpfo.org
8. [pid:1964] British Jokes URL: http://www.callusforfun.com	8. [pid:2] Women Against Gun Control is a free to join group that supports URL: http://www.wagc.com
9. [pid:1944] Religious Happenings URL: http://www.bellbrook-umc.com	9. [pid:4] GunCite: gun control and Second Amendment issues URL: http://www.guncite.com
10. [pid:1949] Blade Liners URL: http://www.bladeliners.com	10. [pid:622] Second Amendment Sisters', Inc. URL: http://www.sas-aim.org
11. [pid:1951] Job Help URL: http://www.bossbid.com/hello.html	11. [pid:40] GunTruths: The truth about guns URL: http://www.guntruths.com
12. [pid:1952] Bowles Professional Roofing URL: http://www.bowlesroofing.com	12. [pid:410] Keep and Bear Arms - Gun Owners Home Page - 2nd Amendme URL: http://www.keepandbeararms.com
13. [pid:1955] Awesome Europe URL: http://www.brightonsearch.com	13. [pid:334] ATF Online - Bureau of Alcohol, Tobacco, Firearms and Explosives URL: http://www.atf.treas.gov
14. [pid:1956] Brian Littrell URL: http://www.brok-littrell.net	14. [pid:185] The Million Mom March united with the Brady Campaign to Prevent URL: http://www.millionmommarch.org
15. [pid:1959] Back To The Future Trilogy DVD URL: http://www.bttftrilogy.com	15. [pid:5] Gun Control vs. Gun Rights: The Issue URL: http://www.opensecrets.org/news/guns
AT-Avg	**Norm (2)**
1. [pid:340] National Rifle Association - MyNRA URL: http://www.nra.org	1. [pid:340] National Rifle Association - MyNRA URL: http://www.nra.org
2. [pid:337] Gun Owners of America URL: http://www.gunowners.org	2. [pid:1] The Brady Campaign to Prevent Gun Violence URL: http://www.handguncontrol.org
3. [pid:1] The Brady Campaign to Prevent Gun Violence URL: http://www.handguncontrol.org	3. [pid:337] Gun Owners of America URL: http://www.gunowners.org
4. [pid:336] CCRKBA Home Page URL: http://www.ccrkba.org	4. [pid:336] CCRKBA Home Page URL: http://www.ccrkba.org
5. [pid:339] Jews for the Preservation of Firearms Ownership URL: http://www.jpfo.org	5. [pid:352] The Violence Policy Center URL: http://www.vpc.org
6. [pid:352] The Violence Policy Center URL: http://www.vpc.org	6. [pid:339] Jews for the Preservation of Firearms Ownership URL: http://www.jpfo.org
7. [pid:42] Coalition to Stop Gun Violence/Educational Fund to Stop Gun Viole URL: http://www.gunfree.org	7. [pid:42] Coalition to Stop Gun Violence/Educational Fund to Stop Gun Viole URL: http://www.gunfree.org
8. [pid:2] Women Against Gun Control is a free to join group that supports URL: http://www.wagc.com	8. [pid:2] Women Against Gun Control is a free to join group that supports URL: http://www.wagc.com
9. [pid:622] Second Amendment Sisters', Inc. URL: http://www.sas-aim.org	9. [pid:622] Second Amendment Sisters', Inc. URL: http://www.sas-aim.org
10. [pid:410] Keep and Bear Arms - Gun Owners Home Page - 2nd Amendme URL: http://www.keepandbeararms.com	10. [pid:4] GunCite: gun control and Second Amendment issues URL: http://www.guncite.com
11. [pid:4] GunCite: gun control and Second Amendment issues URL: http://www.guncite.com	11. [pid:410] Keep and Bear Arms - Gun Owners Home Page - 2nd Amendme URL: http://www.keepandbeararms.com
12. [pid:40] GunTruths: The truth about guns URL: http://www.guntruths.com	12. [pid:40] GunTruths: The truth about guns URL: http://www.guntruths.com
13. [pid:334] ATF Online - Bureau of Alcohol, Tobacco, Firearms and Explosives URL: http://www.atf.treas.gov	13. [pid:334] ATF Online - Bureau of Alcohol, Tobacco, Firearms and Explosives URL: http://www.atf.treas.gov
14. [pid:1105] Welcome to 2ndlawlib.org! URL: http://www.2ndlawlib.org	14. [pid:185] The Million Mom March united with the Brady Campaign to Prevent G URL: http://www.millionmommarch.org
15. [pid:185] The Million Mom March united with the Brady Campaign to Prevent URL: http://www.millionmommarch.org	15. [pid:1105] Welcome to 2ndlawlib.org! URL: http://www.2ndlawlib.org
Max	**SALSA**
1. [pid:340] National Rifle Association - MyNRA URL: http://www.nra.org	1. [pid:340] National Rifle Association - MyNRA URL: http://www.nra.org
2. [pid:1] The Brady Campaign to Prevent Gun Violence URL: http://www.handguncontrol.org	2. [pid:1] The Brady Campaign to Prevent Gun Violence URL: http://www.handguncontrol.org
3. [pid:337] Gun Owners of America URL: http://www.gunowners.org	3. [pid:337] Gun Owners of America URL: http://www.gunowners.org
4. [pid:352] The Violence Policy Center URL: http://www.vpc.org	4. [pid:352] The Violence Policy Center URL: http://www.vpc.org
5. [pid:336] CCRKBA Home Page URL: http://www.ccrkba.org	5. [pid:339] Jews for the Preservation of Firearms Ownership URL: http://www.jpfo.org
6. [pid:339] Jews for the Preservation of Firearms Ownership URL: http://www.jpfo.org	6. [pid:42] Coalition to Stop Gun Violence/Educational Fund to Stop Gun URL: http://www.gunfree.org
7. [pid:42] Coalition to Stop Gun Violence/Educational Fund to Stop Gun Viole URL: http://www.gunfree.org	7. [pid:336] CCRKBA Home Page URL: http://www.ccrkba.org
8. [pid:2] Women Against Gun Control is a free to join group that supports URL: http://www.wagc.com	8. [pid:2] Women Against Gun Control is a free to join group that supports URL: http://www.wagc.com
9. [pid:622] Second Amendment Sisters', Inc. URL: http://www.sas-aim.org	9. [pid:40] GunTruths: The truth about guns URL: http://www.guntruths.com
10. [pid:4] GunCite: gun control and Second Amendment issues URL: http://www.guncite.com	10. [pid:4] GunCite: gun control and Second Amendment issues URL: http://www.guncite.com
11. [pid:410] Keep and Bear Arms - Gun Owners Home Page - 2nd Amendme URL: http://www.keepandbeararms.com	11. [pid:410] Keep and Bear Arms - Gun Owners Home Page - 2nd Amendme URL: http://www.keepandbeararms.com
12. [pid:40] GunTruths: The truth about guns URL: http://www.guntruths.com	12. [pid:622] Second Amendment Sisters', Inc. URL: http://www.sas-aim.org
13. [pid:334] ATF Online - Bureau of Alcohol, Tobacco, Firearms and Explosives URL: http://www.atf.treas.gov	13. [pid:185] The Million Mom March united with the Brady Campaign to Prevent G URL: http://www.millionmommarch.org
14. [pid:185] The Million Mom March united with the Brady Campaign to Prevent URL: http://www.millionmommarch.org	14. [pid:1295] TV Guide Online - [TV Guide] URL: http://www.tvguide.com
15. [pid:1105] Welcome to 2ndlawlib.org! URL: http://www.2ndlawlib.org	15. [pid:1297] Investor's Business Daily Stock Quotes Business News Stock Market URL: http://www.investors.com

Table B.19: *Top 15 results for query "gun control"*

HITS	**HubAvg**
1. [pid:3320] Google Search: URL: http://www.google.com/search	1. [pid:3320] Google Search: URL: http://www.google.com/search
2. [pid:2916] Moreover Technologies - Welcome URL: http://www.moreover.com	2. [pid:2916] Moreover Technologies - Welcome URL: http://www.moreover.com
3. [pid:712] ThePaperboy.com — Online Newspaper Directory URL: http://www.thepaperboy.com	3. [pid:712] ThePaperboy.com — Online Newspaper Directory URL: http://www.thepaperboy.com
4. [pid:3356] NewsLink URL: http://www.newslink.org/news.html	4. [pid:3356] NewsLink URL: http://www.newslink.org/news.html
5. [pid:3355] Kidon Media-Link URL: http://www.kidon.com/media-link	5. [pid:3355] Kidon Media-Link URL: http://www.kidon.com/media-link
6. [pid:3442] Welcome - Roam International URL: http://www.roamintl.com	6. [pid:3442] Welcome - Roam International URL: http://www.roamintl.com
7. [pid:166] Abu Dhabi News - current events and news. URL: http://www.abudhabi.com	7. [pid:990] Top Breaking News Headlines From 1stHeadlines URL: http://www.1stheadlines.com
8. [pid:365] Where is Raed ? URL: http://dear_raed.blogspot.com	8. [pid:2594] Google News URL: http://news.google.com
9. [pid:274] UNMOVIC URL: http://www.un.org/Depts/unmovic	9. [pid:1719] Yahoo! UK & Ireland News URL: http://uk.news.yahoo.com
10. [pid:215] Iraq Liberated - U.S. Department of State URL: http://usinfo.state.gov/regional/nea/iraq	10. [pid:1721] Yahoo! UK & Ireland URL: http://uk.yahoo.com
11. [pid:97] United for Peace URL: http://www.unitedforpeace.org	11. [pid:3706] Venezuela URL: http://venezuela.newstrove.com
12. [pid:327] DefenseLINK - Official Web Site of the U.S. Department of Defense URL: http://www.defenselink.mil	12. [pid:166] Abu Dhabi News - current events and news. URL: http://www.abudhabi.com
13. [pid:59] International A.N.S.W.E.R. (Act Now to Stop War & End Racism!) URL: http://www.internationalanswer.org	13. [pid:637] Yahoo! News - Front Page URL: http://news.yahoo.com
14. [pid:272] UN Office of the Iraq Program - Oil-for-Food URL: http://www.un.org/Depts/oip	14. [pid:1172] Fairness & Accuracy In Reporting: The National Media Watch Group URL: http://fair.org
15. [pid:168] ÇáÌÒíÑÉ äÊ / ÇáÒÝÍÉ ÇáÑÆíÓíÉ URL: http://www.aljazeera.net	15. [pid:30] CNN.com URL: http://www.cnn.com
AT-Avg	**Norm (2)**
1. [pid:3320] Google Search: URL: http://www.google.com/search	1. [pid:3320] Google Search: URL: http://www.google.com/search
2. [pid:712] ThePaperboy.com — Online Newspaper Directory URL: http://www.thepaperboy.com	2. [pid:2916] Moreover Technologies - Welcome URL: http://www.moreover.com
3. [pid:3356] NewsLink URL: http://www.newslink.org/news.html	3. [pid:712] ThePaperboy.com — Online Newspaper Directory URL: http://www.thepaperboy.com
4. [pid:2916] Moreover Technologies - Welcome URL: http://www.moreover.com	4. [pid:3356] NewsLink URL: http://www.newslink.org/news.html
5. [pid:3355] Kidon Media-Link URL: http://www.kidon.com/media-link	5. [pid:3355] Kidon Media-Link URL: http://www.kidon.com/media-link
6. [pid:3442] Welcome - Roam International URL: http://www.roamintl.com	6. [pid:3442] Welcome - Roam International URL: http://www.roamintl.com
7. [pid:990] Top Breaking News Headlines From 1stHeadlines URL: http://www.1stheadlines.com	7. [pid:2594] Google News URL: http://news.google.com
8. [pid:1719] Yahoo! UK & Ireland News URL: http://uk.news.yahoo.com	8. [pid:990] Top Breaking News Headlines From 1stHeadlines URL: http://www.1stheadlines.com
9. [pid:2594] Google News URL: http://news.google.com	9. [pid:1719] Yahoo! UK & Ireland News URL: http://uk.news.yahoo.com
10. [pid:3706] Venezuela URL: http://venezuela.newstrove.com	10. [pid:637] Yahoo! News - Front Page URL: http://news.yahoo.com
11. [pid:637] Yahoo! News - Front Page URL: http://news.yahoo.com	11. [pid:30] CNN.com URL: http://www.cnn.com
12. [pid:166] Abu Dhabi News - current events and news. URL: http://www.abudhabi.com	12. [pid:188] The New York Times on the Web URL: http://www.nytimes.com
13. [pid:1172] Fairness & Accuracy In Reporting: The National Media Watch Group URL: http://fair.org	13. [pid:3706] Venezuela URL: http://venezuela.newstrove.com
14. [pid:1721] Yahoo! UK & Ireland URL: http://uk.yahoo.com	14. [pid:166] Abu Dhabi News - current events and news. URL: http://www.abudhabi.com
15. [pid:30] CNN.com URL: http://www.cnn.com	15. [pid:1172] Fairness & Accuracy In Reporting: The National Media Watch Group URL: http://fair.org
Max	**SALSA**
1. [pid:3320] Google Search: URL: http://www.google.com/search	1. [pid:3320] Google Search: URL: http://www.google.com/search
2. [pid:2916] Moreover Technologies - Welcome URL: http://www.moreover.com	2. [pid:2916] Moreover Technologies - Welcome URL: http://www.moreover.com
3. [pid:712] ThePaperboy.com — Online Newspaper Directory URL: http://www.thepaperboy.com	3. [pid:712] ThePaperboy.com — Online Newspaper Directory URL: http://www.thepaperboy.com
4. [pid:3356] NewsLink URL: http://www.newslink.org/news.html	4. [pid:3356] NewsLink URL: http://www.newslink.org/news.html
5. [pid:3355] Kidon Media-Link URL: http://www.kidon.com/media-link	5. [pid:97] United for Peace URL: http://www.unitedforpeace.org
6. [pid:3442] Welcome - Roam International URL: http://www.roamintl.com	6. [pid:3355] Kidon Media-Link URL: http://www.kidon.com/media-link
7. [pid:97] United for Peace URL: http://www.unitedforpeace.org	7. [pid:3442] Welcome - Roam International URL: http://www.roamintl.com
8. [pid:2594] Google News URL: http://news.google.com	8. [pid:459] Welcome to the White House URL: http://www.whitehouse.gov
9. [pid:459] Welcome to the White House URL: http://www.whitehouse.gov	9. [pid:59] International A.N.S.W.E.R. (Act Now to Stop War & End Racism!) URL: http://www.internationalanswer.org
10. [pid:365] Where is Raed ? URL: http://dear_raed.blogspot.com	10. [pid:365] Where is Raed ? URL: http://dear_raed.blogspot.com
11. [pid:30] CNN.com URL: http://www.cnn.com	11. [pid:274] UNMOVIC URL: http://www.un.org/Depts/unmovic
12. [pid:1265] AlterNet: Top Stories URL: http://www.alternet.org	12. [pid:562] Corporation for Public Broadcasting (CPB)– Public TV, Public Rad URL: http://www.cpb.org
13. [pid:59] International A.N.S.W.E.R. (Act Now to Stop War & End Racism!) URL: http://www.internationalanswer.org	13. [pid:33] Not In Our Name URL: http://www.notinourname.net
14. [pid:1719] Yahoo! UK & Ireland News URL: http://uk.news.yahoo.com	14. [pid:560] ADM-Archer Daniels Midland Company URL: http://www.admworld.com
15. [pid:990] Top Breaking News Headlines From 1stHeadlines URL: http://www.1stheadlines.com	15. [pid:561] SBC Communications Inc. URL: http://www.sbc.com

Table B.20: *Top 15 results for query "iraq war"*

HITS	**HubAvg**
1. [pid:2110] Apple iPod Updater 1.3 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/12478	1. [pid:2684] Griffman's OS X Collection URL: http://homepage.mac.com/rgriff
2. [pid:2111] Apple iTunes 4.0 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/10227	2. [pid:2750] Amazon Honor System URL: http://s1.amazon.com/exec/varzea/pay/T1ZZV2ETFQHXW3
3. [pid:2112] VueScan 7.6.34 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/18495	3. [pid:180] Apple .Mac Welcome URL: http://www.mac.com
4. [pid:2113] VueScan 7.6.34 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/7610	4. [pid:214] Fink - Home URL: http://fink.sourceforge.net
5. [pid:2114] Apple iPod Updater 1.3 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/15793	5. [pid:1] Apple - Mac OS X URL: http://www.apple.com/macosx
6. [pid:2115] PHP 4.3.2RC2 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/14526	6. [pid:1180] Apple URL: http://www.apple.com
7. [pid:2116] Palm Desktop 4.1 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/12968	7. [pid:2727] Apple - Discussions - Welcome URL: http://discussions.info.apple.com
8. [pid:2126] Apple - Games - Trailers URL: http://www.apple.com/games/trailers	8. [pid:2686] LinuxPrinting.org URL: http://www.linuxprinting.org
9. [pid:2117] Dantz Retrospect 5.0 Driver Update 3.5 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/15085	9. [pid:2709] Jaguar & Gimp-Print URL: http://www.allosx.com/1030154694/index_html
10. [pid:2118] iView MediaPro 1.5.7 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/9659	10. [pid:2725] Xamba URL: http://xamba.sourceforge.net/ssp
11. [pid:2119] Acquisition 0.9 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/15375	11. [pid:1098] Apple URL: http://www.apple.com/legal
12. [pid:2120] Apple - Games - Alias URL: http://www.apple.com/games/articles/2002/12/alias	12. [pid:1440] Apple - Apple Customer Privacy Statement URL: http://www.apple.com/legal/privacy
13. [pid:2121] Apple - Games - Trailers - F1 Championship URL: http://www.apple.com/games/trailers/f1championship	13. [pid:84] OS X 10.2 Jaguar Troubleshooting URL: http://www.macattorney.com/tutorial.html
14. [pid:2122] Apple - Games - Trailers - Slots from Bally Gaming URL: http://www.apple.com/games/trailers/slots	14. [pid:1653] macosxhints - Get the most from X! URL: http://www.macosxhints.com
15. [pid:2123] Apple - Games - Trailers - WWII Online URL: http://www.apple.com/games/trailers/ww2online	15. [pid:2130] XDarwin is X11 for MacOS X URL: http://www.xdarwin.org
AT-Avg	**Norm (2)**
1. [pid:2110] Apple iPod Updater 1.3 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/12478	1. [pid:2110] Apple iPod Updater 1.3 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/12478
2. [pid:2111] Apple iTunes 4.0 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/10227	2. [pid:2111] Apple iTunes 4.0 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/10227
3. [pid:2112] VueScan 7.6.34 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/18495	3. [pid:2112] VueScan 7.6.34 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/18495
4. [pid:2113] VueScan 7.6.34 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/7610	4. [pid:2113] VueScan 7.6.34 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/7610
5. [pid:2114] Apple iPod Updater 1.3 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/15793	5. [pid:2114] Apple iPod Updater 1.3 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/15793
6. [pid:2115] PHP 4.3.2RC2 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/14526	6. [pid:2115] PHP 4.3.2RC2 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/14526
7. [pid:2116] Palm Desktop 4.1 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/12968	7. [pid:2116] Palm Desktop 4.1 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/12968
8. [pid:2126] Apple - Games - Trailers URL: http://www.apple.com/games/trailers	8. [pid:2126] Apple - Games - Trailers URL: http://www.apple.com/games/trailers
9. [pid:2117] Dantz Retrospect 5.0 Driver Update 3.5 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/15085	9. [pid:2117] Dantz Retrospect 5.0 Driver Update 3.5 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/15085
10. [pid:2118] iView MediaPro 1.5.7 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/9659	10. [pid:2118] iView MediaPro 1.5.7 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/9659
11. [pid:2119] Acquisition 0.9 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/15375	11. [pid:2119] Acquisition 0.9 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/15375
12. [pid:2120] Apple - Games - Alias URL: http://www.apple.com/games/articles/2002/12/alias	12. [pid:2120] Apple - Games - Alias URL: http://www.apple.com/games/articles/2002/12/alias
13. [pid:2121] Apple - Games - Trailers - F1 Championship URL: http://www.apple.com/games/trailers/f1championship	13. [pid:2121] Apple - Games - Trailers - F1 Championship URL: http://www.apple.com/games/trailers/f1championship
14. [pid:2122] Apple - Games - Trailers - Slots from Bally Gaming URL: http://www.apple.com/games/trailers/slots	14. [pid:2122] Apple - Games - Trailers - Slots from Bally Gaming URL: http://www.apple.com/games/trailers/slots
15. [pid:2123] Apple - Games - Trailers - WWII Online URL: http://www.apple.com/games/trailers/ww2online	15. [pid:2123] Apple - Games - Trailers - WWII Online URL: http://www.apple.com/games/trailers/ww2online
Max	**SALSA**
1. [pid:180] Apple .Mac Welcome URL: http://www.mac.com	1. [pid:180] Apple .Mac Welcome URL: http://www.mac.com
2. [pid:1180] Apple URL: http://www.apple.com	2. [pid:2684] Griffman's OS X Collection URL: http://homepage.mac.com/rgriff
3. [pid:1] Apple - Mac OS X URL: http://www.apple.com/macosx	3. [pid:1180] Apple URL: http://www.apple.com
4. [pid:1098] Apple URL: http://www.apple.com/legal	4. [pid:1] Apple - Mac OS X URL: http://www.apple.com/macosx
5. [pid:1440] Apple - Apple Customer Privacy Statement URL: http://www.apple.com/legal/privacy	5. [pid:1098] Apple URL: http://www.apple.com/legal
6. [pid:1089] The Apple Store (Japan) URL: http://www.apple.com/japanstore	6. [pid:2232] Empty title field URL: http://www.gamesarchiv.com/Layout/?id=grm-10114
7. [pid:214] Fink - Home URL: http://fink.sourceforge.net	7. [pid:1280] ThinkGeek :: O'Reilly Store URL: http://www.thinkgeek.com/oreilly
8. [pid:2684] Griffman's OS X Collection URL: http://homepage.mac.com/rgriff	8. [pid:2110] Apple iPod Updater 1.3 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/12478
9. [pid:1653] macosxhints - Get the most from X! URL: http://www.macosxhints.com	9. [pid:2111] Apple iTunes 4.0 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/10227
10. [pid:1097] Apple - ... URL: http://developer.apple.com/ja	10. [pid:2112] VueScan 7.6.34 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/18495
11. [pid:1652] MacFixIt - Troubleshooting Solution for the Macintosh URL: http://www.macfixit.com	11. [pid:2113] VueScan 7.6.34 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/7610
12. [pid:84] OS X 10.2 Jaguar Troubleshooting URL: http://www.macattorney.com/tutorial.html	12. [pid:2114] Apple iPod Updater 1.3 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/15793
13. [pid:1657] TidBITS Electronic Publishing URL: http://www.tidbits.com	13. [pid:2115] PHP 4.3.2RC2 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/14526
14. [pid:1618] OSXFAQ - Technical News and Support for Mac OS X URL: http://www.osxfaq.com	14. [pid:2116] Palm Desktop 4.1 - VersionTracker URL: http://www.VersionTracker.com/dyn/moreinfo/mac/12968
15. [pid:2750] Amazon Honor System URL: http://s1.amazon.com/exec/varzea/pay/T1ZZV2ETFQHXW3	15. [pid:2126] Apple - Games - Trailers URL: http://www.apple.com/games/trailers

Table B.21: *Top 15 results for query "jaguar"*

HITS	HubAvg
1. [pid:2947] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/00	1. [pid:1900] Yahoo! Directory URL: http://us.rd.yahoo.com/dir/yahoo/*http://dir.yahoo.com
2. [pid:2948] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/01	2. [pid:1901] Yahoo! URL: http://us.rd.yahoo.com/dir/yahoo/*http://www.yahoo.com
3. [pid:2949] Empty title field URL: http://g.msn.com/0nwenus0/AK/02	3. [pid:1902] Yahoo! Help - URL: http://us.rd.yahoo.com/dir/help/*http://help.yahoo.com/help/us/di
4. [pid:2950] MSN Search – More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/03	4. [pid:1903] Yahoo! Advanced Directory Search URL: http://search.yahoo.com/dir/advanced
5. [pid:2951] Welcome to MSN Shopping URL: http://g.msn.com/0nwenus0/AK/04	5. [pid:1904] Yahoo! Suggest a Site URL: http://us.rd.yahoo.com/dir/suggest/*http://add.yahoo.com/fast/add
6. [pid:2952] MSN Money - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/05	6. [pid:1905] Empty title field URL: http://us.rd.yahoo.com/dir/email/*http://mtf.news.yahoo.com/mailt
7. [pid:2953] MSN People and Chat - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/06	7. [pid:9] Jordan Tourism Board URL: http://www.see-jordan.com
8. [pid:2954] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/14	8. [pid:106] The Royal Autombile Club of Jordan - RACJ URL: http://www.racj.com
9. [pid:2979] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/07	9. [pid:3] National Information Center URL: http://www.nic.gov.jo
10. [pid:2980] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/08	10. [pid:3884] rowaq.com hosted at HostSave, the affordable way to web at $7.95 URL: http://www.rowaq.com
11. [pid:2981] Empty title field URL: http://g.msn.com/0nwenus0/AK/09	11. [pid:179] Multitasking - multitasking.com URL: http://www.multitasking.com
12. [pid:2982] MSN Search – More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/10	12. [pid:8] Jordan Embassy - U.S.A. URL: http://www.jordanembassyus.org
13. [pid:2983] Welcome to MSN Shopping URL: http://g.msn.com/0nwenus0/AK/11	13. [pid:109] jordanzed.com URL: http://www.jordanzed.com
14. [pid:2984] MSN Money - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/12	14. [pid:7] Central Bank of Jordan Home Page URL: http://www.cbj.gov.jo
15. [pid:2985] MSN People and Chat - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/13	15. [pid:1359] Empty title field URL: http://www.bigbearvalleygallery.com

AT-Avg	Norm (2)
1. [pid:9] Jordan Tourism Board URL: http://www.see-jordan.com	1. [pid:2947] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/00
2. [pid:106] The Royal Autombile Club of Jordan - RACJ URL: http://www.racj.com	2. [pid:2948] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/01
3. [pid:3] National Information Center URL: http://www.nic.gov.jo	3. [pid:2949] Empty title field URL: http://g.msn.com/0nwenus0/AK/02
4. [pid:8] Jordan Embassy - U.S.A. URL: http://www.jordanembassyus.org	4. [pid:2950] MSN Search – More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/03
5. [pid:7] Central Bank of Jordan Home Page URL: http://www.cbj.gov.jo	5. [pid:2951] Welcome to MSN Shopping URL: http://g.msn.com/0nwenus0/AK/04
6. [pid:2444] Welcome to HIS ROYAL HIGHNESS PRINCE HASSAN BIN TALAL's WEB SITE URL: http://www.princehassan.gov.jo	6. [pid:2952] MSN Money - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/05
7. [pid:41] Department Of Statistics URL: http://www.dos.gov.jo	7. [pid:2953] MSN People and Chat - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/06
8. [pid:23] ÍÇÆÑÉ ÇáÎãÇÑß – Jordan Customs Department URL: http://www.customs.gov.jo	8. [pid:2954] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/14
9. [pid:6] The University of Jordan's homepage URL: http://www.ju.edu.jo	9. [pid:2979] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/07
10. [pid:2466] RJ HOME URL: http://www.rja.com.jo	10. [pid:2980] Welcome to MSN.com URL: http://g.msn.com/0nwenus0/AK/08
11. [pid:3884] rowaq.com hosted at HostSave, the affordable way to web at $7.95 URL: http://www.rowaq.com	11. [pid:2981] Empty title field URL: http://g.msn.com/0nwenus0/AK/09
12. [pid:179] Multitasking - multitasking.com URL: http://www.multitasking.com	12. [pid:2982] MSN Search – More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/10
13. [pid:99] Jordan Export Development & Commercial Centres Corporation (J URL: http://www.jedco.gov.jo	13. [pid:2983] Welcome to MSN Shopping URL: http://g.msn.com/0nwenus0/AK/11
14. [pid:22] Islamic Republic News Agency (I R N A) URL: http://www.jordan-online.com	14. [pid:2984] MSN Money - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/12
15. [pid:40] Petra News -Home Page- URL: http://www.petra.gov.jo	15. [pid:2985] MSN People and Chat - More Useful Everyday URL: http://g.msn.com/0nwenus0/AK/13

Max	SALSA
1. [pid:9] Jordan Tourism Board URL: http://www.see-jordan.com	1. [pid:9] Jordan Tourism Board URL: http://www.see-jordan.com
2. [pid:106] The Royal Autombile Club of Jordan - RACJ URL: http://www.racj.com	2. [pid:179] Multitasking - multitasking.com URL: http://www.multitasking.com
3. [pid:3] National Information Center URL: http://www.nic.gov.jo	3. [pid:3884] rowaq.com hosted at HostSave, the affordable way to web at $7.95 URL: http://www.rowaq.com
4. [pid:8] Jordan Embassy - U.S.A. URL: http://www.jordanembassyus.org	4. [pid:106] The Royal Autombile Club of Jordan - RACJ URL: http://www.racj.com
5. [pid:7] Central Bank of Jordan Home Page URL: http://www.cbj.gov.jo	5. [pid:3] National Information Center URL: http://www.nic.gov.jo
6. [pid:2444] Welcome to HIS ROYAL HIGHNESS PRINCE HASSAN BIN TALAL's WEB SITE URL: http://www.princehassan.gov.jo	6. [pid:109] jordanzed.com URL: http://www.jordanzed.com
7. [pid:2466] RJ HOME URL: http://www.rja.com.jo	7. [pid:739] Home Page URL: http://www.lawtownmusic.8k.com
8. [pid:41] Department Of Statistics URL: http://www.dos.gov.jo	8. [pid:139] SheilaJordanJazz.com URL: http://www.sheilajordanjazz.com
9. [pid:6] The University of Jordan's homepage URL: http://www.ju.edu.jo	9. [pid:1359] Empty title field URL: http://www.bigbearvalleygallery.com
10. [pid:23] ÍÇÆÑÉ ÇáÎãÇÑß – Jordan Customs Department URL: http://www.customs.gov.jo	10. [pid:1984] stamps-by-year URL: http://stamps-of-jordan.tripod.com
11. [pid:46] ...::::: Ministry of Tourism & Antiquities :::::... URL: http://www.mota.gov.jo	11. [pid:1669] ESPN.com URL: http://www.espn.com
12. [pid:99] Jordan Export Development & Commercial Centres Corporation (J URL: http://www.jedco.gov.jo	12. [pid:8] Jordan Embassy - U.S.A. URL: http://www.jordanembassyus.org
13. [pid:40] Petra News -Home Page- URL: http://www.petra.gov.jo	13. [pid:7] Central Bank of Jordan Home Page URL: http://www.cbj.gov.jo
14. [pid:84] Official Website of Her Majesty Queen Noor URL: http://www.noor.gov.jo	14. [pid:50] Jordan Electric Violins URL: http://www.jordanmusic.com
15. [pid:22] Islamic Republic News Agency (I R N A) URL: http://www.jordan-online.com	15. [pid:2444] Welcome to HIS ROYAL HIGHNESS PRINCE HASSAN BIN TALAL's WEB SITE URL: http://www.princehassan.gov.jo

Table B.22: *Top 15 results for query "jordan"*

HITS	HubAvg
1. [pid:675] Long Distance Rate Finder .com - Best telephone calling URL: http://www.longdistanceratefinder.com	1. [pid:675] Long Distance Rate Finder .com - Best telephone calling URL: http://www.longdistanceratefinder.com
2. [pid:717] Cognigen: Worldwide Telecommunications Long Distance URL: http://longdist.net/?apl	2. [pid:717] Cognigen: Worldwide Telecommunications Long Distance URL: http://longdist.net/?apl
3. [pid:738] BILLZilla - The Best Long Distance Rate Calculator URL: http://www.billzilla.com/apl	3. [pid:738] BILLZilla - The Best Long Distance Rate Calculator URL: http://www.billzilla.com/apl
4. [pid:741] Talk America Local And Long Distance Bundle Service URL: http://cognigen.net/talkamerica/?apl	4. [pid:741] Talk America Local And Long Distance Bundle Service URL: http://cognigen.net/talkamerica/?apl
5. [pid:748] OneStar Communications - Long Distance And Local Telephone URL: http://cognigen.net/onestar/?apl	5. [pid:748] OneStar Communications - Long Distance And Local Telephone URL: http://cognigen.net/onestar/?apl
6. [pid:752] CogniDial Discount International Long Distance URL: http://www.cognidial.com/dial-around/?apl	6. [pid:752] CogniDial Discount International Long Distance URL: http://www.cognidial.com/dial-around/?apl
7. [pid:757] Speakeasy High Speed Internet Service Offered by Cognigen URL: http://www.cognigen.net/speakeasy/?apl	7. [pid:757] Speakeasy High Speed Internet Service Offered by Cognigen URL: http://www.cognigen.net/speakeasy/?apl
8. [pid:761] DISH Network e-Store URL: http://cognigen.net/dish/?apl	8. [pid:761] DISH Network e-Store URL: http://cognigen.net/dish/?apl
9. [pid:769] Cognigen: Worldwide Telecommunications Long Distance URL: http://ld.net/?apl	9. [pid:769] Cognigen: Worldwide Telecommunications Long Distance URL: http://ld.net/?apl
10. [pid:774] Exchange-it - Free Banner Exchange URL: http://www.exchange-it.com/link.go?b107780	10. [pid:701] Name a Star - International Star Registry (R) Pick name of your c URL: http://click.linksynergy.com/fs-bin/stat?id=sZV71WSrLU0&offerid=2
11. [pid:701] Name a Star - International Star Registry (R) URL: http://click.linksynergy.com/fs-bin/stat?id=sZV71WSrLU0&offerid=2	11. [pid:774] Exchange-it - Free Banner Exchange URL: http://www.exchange-it.com/link.go?b107780
12. [pid:724] Dish Network Satellite TV! URL: http://www.vmcsatellite.com/?aid=35056	12. [pid:714] Empty title field URL: http://www.adreporting.com/dir.php?a=669219&p=166&w=i_1340
13. [pid:677] Name a Star - International Star Registry (R) URL: http://click.linksynergy.com/fs-bin/stat?id=sZV71WSrLU0&offerid=2	13. [pid:677] Name a Star - International Star Registry (R) URL: http://click.linksynergy.com/fs-bin/stat?id=sZV71WSrLU0&offerid=2
14. [pid:700] ShopAETV.com - Product Detail URL: http://www.qksrv.net/. . . /url=http%3A%2F%2Fstore	14. [pid:700] ShopAETV.com - Product Detail URL: http://www.qksrv.net/. . . /url=http%3A%2F%2Fstore
15. [pid:679] Books-A-Million Online Bookstore : Buy Discount Books Magazines URL: http://www.qksrv.net/click-310374-35140	15. [pid:724] Dish Network Satellite TV! URL: http://www.vmcsatellite.com/?aid=35056

AT-Avg	Norm (2)
1. [pid:675] Long Distance Rate Finder .com - Best telephone calling URL: http://www.longdistanceratefinder.com	1. [pid:675] Long Distance Rate Finder .com - Best telephone calling URL: http://www.longdistanceratefinder.com
2. [pid:717] Cognigen: Worldwide Telecommunications Long Distance URL: http://longdist.net/?apl	2. [pid:717] Cognigen: Worldwide Telecommunications Long Distance URL: http://longdist.net/?apl
3. [pid:738] BILLZilla - The Best Long Distance Rate Calculator URL: http://www.billzilla.com/apl	3. [pid:738] BILLZilla - The Best Long Distance Rate Calculator URL: http://www.billzilla.com/apl
4. [pid:741] Talk America Local And Long Distance Bundle Service URL: http://cognigen.net/talkamerica/?apl	4. [pid:741] Talk America Local And Long Distance Bundle Service URL: http://cognigen.net/talkamerica/?apl
5. [pid:748] OneStar Communications - Long Distance And Local Telephone URL: http://cognigen.net/onestar/?apl	5. [pid:748] OneStar Communications - Long Distance And Local Telephone URL: http://cognigen.net/onestar/?apl
6. [pid:752] CogniDial Discount International Long Distance URL: http://www.cognidial.com/dial-around/?apl	6. [pid:752] CogniDial Discount International Long Distance URL: http://www.cognidial.com/dial-around/?apl
7. [pid:757] Speakeasy High Speed Internet Service Offered by Cognigen URL: http://www.cognigen.net/speakeasy/?apl	7. [pid:757] Speakeasy High Speed Internet Service Offered by Cognigen URL: http://www.cognigen.net/speakeasy/?apl
8. [pid:761] DISH Network e-Store URL: http://cognigen.net/dish/?apl	8. [pid:761] DISH Network e-Store URL: http://cognigen.net/dish/?apl
9. [pid:769] Cognigen: Worldwide Telecommunications Long Distance Services URL: http://ld.net/?apl	9. [pid:769] Cognigen: Worldwide Telecommunications Long Distance Services URL: http://ld.net/?apl
10. [pid:701] Name a Star - International Star Registry (R) Pick name of your c URL: http://click.linksynergy.com/fs-bin/stat?id=sZV71WSrLU0&offerid=2	10. [pid:701] Name a Star - International Star Registry (R) Pick name of your c URL: http://click.linksynergy.com/fs-bin/stat?id=sZV71WSrLU0&offerid=2
11. [pid:774] Exchange-it - Free Banner Exchange URL: http://www.exchange-it.com/link.go?b107780	11. [pid:774] Exchange-it - Free Banner Exchange URL: http://www.exchange-it.com/link.go?b107780
12. [pid:724] Dish Network Satellite TV! URL: http://www.vmcsatellite.com/?aid=35056	12. [pid:724] Dish Network Satellite TV! URL: http://www.vmcsatellite.com/?aid=35056
13. [pid:677] Name a Star - International Star Registry (R) Pick name of your c URL: http://click.linksynergy.com/fs-bin/stat?id=sZV71WSrLU0&offerid=2	13. [pid:677] Name a Star - International Star Registry (R) Pick name of your c URL: http://click.linksynergy.com/fs-bin/stat?id=sZV71WSrLU0&offerid=2
14. [pid:700] ShopAETV.com - Product Detail URL: http://www.qksrv.net/. . . /url=http%3A%2F%2Fstore	14. [pid:700] ShopAETV.com - Product Detail URL: http://www.qksrv.net/. . . /url=http%3A%2F%2Fstore
15. [pid:714] Empty title field URL: http://www.adreporting.com/dir.php?a=669219&p=166&w=i_1340	15. [pid:714] Empty title field URL: http://www.adreporting.com/dir.php?a=669219&p=166&w=i_1340

Max	SALSA
1. [pid:675] Long Distance Rate Finder .com - Best telephone calling URL: http://www.longdistanceratefinder.com	1. [pid:675] Long Distance Rate Finder .com - Best telephone calling URL: http://www.longdistanceratefinder.com
2. [pid:717] Cognigen: Worldwide Telecommunications Long Distance Services URL: http://longdist.net/?apl	2. [pid:223] Empty title field URL: http://www.nasa.gov
3. [pid:738] BILLZilla - The Best Long Distance Rate Calculator URL: http://www.billzilla.com/apl	3. [pid:1236] American Express URL: http://www.americanexpress.com/cards/online_guarantee
4. [pid:741] Talk America Local And Long Distance Bundle Service URL: http://cognigen.net/talkamerica/?apl	4. [pid:544] Real Estate Australia - Property for sale lease and rent online URL: http://www.realestate.com.au
5. [pid:748] OneStar Communications - Long Distance And Local Telephone URL: http://cognigen.net/onestar/?apl	5. [pid:2] Moon Hoax Index URL: http://www.redzero.demon.co.uk/moonhoax
6. [pid:752] CogniDial Discount International Long Distance URL: http://www.cognidial.com/dial-around/?apl	6. [pid:717] Cognigen: Worldwide Telecommunications Long Distance URL: http://longdist.net/?apl
7. [pid:757] Speakeasy High Speed Internet Service Offered by Cognigen URL: http://www.cognigen.net/speakeasy/?apl	7. [pid:738] BILLZilla - The Best Long Distance Rate Calculator URL: http://www.billzilla.com/apl
8. [pid:761] DISH Network e-Store URL: http://cognigen.net/dish/?apl	8. [pid:741] Talk America Local And Long Distance Bundle Service URL: http://cognigen.net/talkamerica/?apl
9. [pid:769] Cognigen: Worldwide Telecommunications Long Distance URL: http://ld.net/?apl	9. [pid:748] OneStar Communications - Long Distance And Local Telephone URL: http://cognigen.net/onestar/?apl
10. [pid:701] Name a Star - International Star Registry (R) Pick name URL: http://click.linksynergy.com/fs-bin/stat?id=sZV71WSrLU0&offerid=2	10. [pid:752] CogniDial Discount International Long Distance URL: http://www.cognidial.com/dial-around/?apl
11. [pid:774] Exchange-it - Free Banner Exchange URL: http://www.exchange-it.com/link.go?b107780	11. [pid:757] Speakeasy High Speed Internet Service Offered by Cognigen URL: http://www.cognigen.net/speakeasy/?apl
12. [pid:714] Empty title field URL: http://www.adreporting.com/dir.php?a=669219&p=166&w=i_1340	12. [pid:761] DISH Network e-Store URL: http://cognigen.net/dish/?apl
13. [pid:677] Name a Star - International Star Registry (R) Pick name of your c URL: http://click.linksynergy.com/fs-bin/stat?id=sZV71WSrLU0&offerid=2	13. [pid:769] Cognigen: Worldwide Telecommunications . . . URL: http://ld.net/?apl
14. [pid:700] ShopAETV.com - Product Detail URL: http://www.qksrv.net/. . . /url=http%3A%2F%2Fstore	14. [pid:8] Phil Plait's Bad Astronomy: Bad TV URL: http://www.badastronomy.com/bad/tv/foxapollo.html
15. [pid:724] Dish Network Satellite TV! URL: http://www.vmcsatellite.com/?aid=35056	15. [pid:549] World Vision Australia My World Vision: . . . URL: http://svc003.bne104p.server-web.com/worldvision/news/lhn_redirec

Table B.23: *Top 15 results for query "moon landing"*

HITS	**HubAvg**
1. [pid:2] The Internet Movie Database (IMDb). URL: http://www.imdb.com	1. [pid:219] Amazon.com–Earth's Biggest Selection URL: http://www.amazon.com/. . . /internetmoviedat
2. [pid:6238] DHTML Lab: HierMenus CENTRAL - dhtmlab.com URL: http://www.hiermenuscentral.com	2. [pid:2] The Internet Movie Database (IMDb). URL: http://www.imdb.com
3. [pid:6214] internet.com: the Internet and IT Network from Jupitermedia Corp. URL: http://www.internet.com	3. [pid:682] Google URL: http://www.google.com
4. [pid:6246] WebReference.com - The Webmaster's Reference Library - Web Author URL: http://webreference.com	4. [pid:5793] CNI Newspapers: News Front Page URL: http://www.cninewsonline.com
5. [pid:6205] Welcome to internet.com's Developer Channel URL: http://www.internet.com/sections/webdev.html	5. [pid:5932] JS Online: General Information URL: http://graphics.jsonline.com/adsections
6. [pid:6212] Jupitermedia Corporation Web Site User Agreement URL: http://www.internet.com/corporate/legal.html	6. [pid:77] Signs on DVD URL: http://www.signs.movies.com
7. [pid:6213] Jupitermedia Privacy Policy URL: http://www.internet.com/corporate/privacy/privacypolicy.html	7. [pid:3] Hollywood.com - Your entertainment source for movies, movie URL: http://www.hollywood.com
8. [pid:6233] internet.com Commerce Partners Program URL: http://www.internet.com/partners	8. [pid:34] Empty title field URL: http://www.film.com
9. [pid:6215] Search Internet.com URL: http://search.internet.com	9. [pid:164] ROTTEN TOMATOES: Movie Reviews & Previews URL: http://www.rottentomatoes.com
10. [pid:6216] internet.com Media Kit URL: http://www.internet.com/mediakit	10. [pid:3566] Get Wild - GetWild - getwild.com URL: http://www.getwild.com
11. [pid:6217] Corporate Information URL: http://www.internet.com/corporate	11. [pid:4005] Movie Review Query Engine URL: http://www.mrqe.com
12. [pid:6218] e-newsletters.internet.com: Free mailing lists, email newsletters URL: http://e-newsletters.internet.com	12. [pid:31] All Movie Guide URL: http://www.allmovie.com
13. [pid:6219] Free Opt-In Email Announcements URL: http://e-newsletters.internet.com/mailinglists.html	13. [pid:128] Greatest Films URL: http://www.filmsite.org
14. [pid:6220] Welcome to internet.com's News Channel URL: http://www.internet.com/sections/news.html	14. [pid:472] The New York Times on the Web URL: http://www.nytimes.com
15. [pid:6221] Welcome to internet.com's Internet Investing Channel URL: http://www.internet.com/sections/stocks.html	15. [pid:169] American Film Institute URL: http://www.afionline.org

AT-Avg	**Norm (2)**
1. [pid:2] The Internet Movie Database (IMDb). URL: http://www.imdb.com	1. [pid:2] The Internet Movie Database (IMDb). URL: http://www.imdb.com
2. [pid:77] Signs on DVD URL: http://www.signs.movies.com	2. [pid:77] Signs on DVD URL: http://www.signs.movies.com
3. [pid:682] Google URL: http://www.google.com	3. [pid:682] Google URL: http://www.google.com
4. [pid:3] Hollywood.com - Your entertainment source for movies URL: http://www.hollywood.com	4. [pid:3] Hollywood.com - Your entertainment source for movies URL: http://www.hollywood.com
5. [pid:34] Empty title field URL: http://www.film.com	5. [pid:34] Empty title field URL: http://www.film.com
6. [pid:3566] Get Wild - GetWild - getwild.com URL: http://www.getwild.com	6. [pid:3566] Get Wild - GetWild - getwild.com URL: http://www.getwild.com
7. [pid:31] All Movie Guide URL: http://www.allmovie.com	7. [pid:31] All Movie Guide URL: http://www.allmovie.com
8. [pid:164] ROTTEN TOMATOES: Movie Reviews & Previews URL: http://www.rottentomatoes.com	8. [pid:164] ROTTEN TOMATOES: Movie Reviews & Previews URL: http://www.rottentomatoes.com
9. [pid:4005] Movie Review Query Engine URL: http://www.mrqe.com	9. [pid:4005] Movie Review Query Engine URL: http://www.mrqe.com
10. [pid:2528] Paramount Pictures URL: http://www.paramount.com	10. [pid:128] Greatest Films URL: http://www.filmsite.org
11. [pid:472] The New York Times on the Web URL: http://www.nytimes.com	11. [pid:2528] Paramount Pictures URL: http://www.paramount.com
12. [pid:128] Greatest Films URL: http://www.filmsite.org	12. [pid:472] The New York Times on the Web URL: http://www.nytimes.com
13. [pid:1047] New Line Cinema URL: http://www.newline.com	13. [pid:1047] New Line Cinema URL: http://www.newline.com
14. [pid:27] Home page of the About Classic Movies site URL: http://classicfilm.about.com	14. [pid:27] Home page of the About Classic Movies site URL: http://classicfilm.about.com
15. [pid:639] A Beautiful Mind URL: http://abeautifulmind.com	15. [pid:639] A Beautiful Mind URL: http://abeautifulmind.com

Max	**SALSA**
1. [pid:2] The Internet Movie Database (IMDb). URL: http://www.imdb.com	1. [pid:2] The Internet Movie Database (IMDb). URL: http://www.imdb.com
2. [pid:77] Signs on DVD URL: http://www.signs.movies.com	2. [pid:682] Google URL: http://www.google.com
3. [pid:682] Google URL: http://www.google.com	3. [pid:77] Signs on DVD URL: http://www.signs.movies.com
4. [pid:3] Hollywood.com - Your entertainment source for movies URL: http://www.hollywood.com	4. [pid:3566] Get Wild - GetWild - getwild.com URL: http://www.getwild.com
5. [pid:34] Empty title field URL: http://www.film.com	5. [pid:1981] Gannett Company, Inc. URL: http://www.gannett.com
6. [pid:3566] Get Wild - GetWild - getwild.com URL: http://www.getwild.com	6. [pid:34] Empty title field URL: http://www.film.com
7. [pid:31] All Movie Guide URL: http://www.allmovie.com	7. [pid:3] Hollywood.com - Your entertainment source for movies URL: http://www.hollywood.com
8. [pid:4005] Movie Review Query Engine URL: http://www.mrqe.com	8. [pid:6681] Knight Ridder Corporate Web site URL: http://www.knightridder.com
9. [pid:164] ROTTEN TOMATOES: Movie Reviews & Previews URL: http://www.rottentomatoes.com	9. [pid:639] A Beautiful Mind URL: http://abeautifulmind.com
10. [pid:128] Greatest Films URL: http://www.filmsite.org	10. [pid:3637] USATODAY.com - News & Information Homepage URL: http://www.usatoday.com
11. [pid:2528] Paramount Pictures URL: http://www.paramount.com	11. [pid:1047] New Line Cinema URL: http://www.newline.com
12. [pid:169] American Film Institute URL: http://www.afionline.org	12. [pid:5793] CNI Newspapers: News Front Page URL: http://www.cninewsonline.com
13. [pid:472] The New York Times on the Web URL: http://www.nytimes.com	13. [pid:27] Home page of the About Classic Movies site URL: http://classicfilm.about.com
14. [pid:27] Home page of the About Classic Movies site URL: http://classicfilm.about.com	14. [pid:472] The New York Times on the Web URL: http://www.nytimes.com
15. [pid:658] Welcome to The Academy of Motion Picture Arts and Sciences URL: http://www.oscar.org	15. [pid:4005] Movie Review Query Engine URL: http://www.mrqe.com

Table B.24: *Top 15 results for query "movies"*

HITS	**HubAvg**
1. [pid:3261] E Business Solutions,Website Promotion Services URL: http://www.intermesh.net/advertis.html	1. [pid:1] National Park Service - Experience Your America URL: http://www.nps.gov
2. [pid:3377] Empty title field URL: http://www.indiangiftsportal.com	2. [pid:211] Privacy Statement, National Park Service URL: http://www.nps.gov/privacy.htm
3. [pid:3411] Business Solutions,Ecommerce Business Solutions URL: http://www.intermesh.net	3. [pid:15] Park Geology Tour - Geologic Features URL: http://www.nature.nps.gov/grd/tour
4. [pid:3403] Empty title field URL: http://news.indiamart.com	4. [pid:2497] USGS Western Earth Surface Processes Team Home URL: http://wrgis.wr.usgs.gov/wgmt
5. [pid:3373] Empty title field URL: http://www.indiamart.com	5. [pid:379] NPS Search Portal URL: http://www.nps.gov/search.htm
6. [pid:3374] Empty title field URL: http://apparel.indiamart.com	6. [pid:28] NatureNet: The National Park Service's natural resource website URL: http://www.nature.nps.gov
7. [pid:3400] Empty title field URL: http://handicraft.indiamart.com	7. [pid:24] GORP.com-adventure travel-hiking, national parks URL: http://www.gorp.com
8. [pid:3401] India Finance and Investment Guide, India URL: http://finance.indiamart.com	8. [pid:2] National Park Guide URL: http://www.nps.gov/parks.html
9. [pid:3402] Empty title field URL: http://health.indiamart.com	9. [pid:5] National Parks Conservation Association URL: http://www.npca.org
10. [pid:3404] Empty title field URL: http://auto.indiamart.com	10. [pid:25] L.L.Bean - Park Search URL: http://www.llbean.com/parksearch
11. [pid:3394] Anniversary Gifts,Wedding Anniversary Gift URL: http://www.indiangiftsportal.com/india-shopping/occasions/anniver	11. [pid:176] NPS Freedom of Information Act URL: http://www.nps.gov/refdesk/npsfoia.html
12. [pid:3395] Jewelry Box, Jewelry Gift Box, Jewelry Box Shopping Online URL: http://www.indiangiftsportal.com/india-shopping/exclusives/jewell	12. [pid:175] Take Pride In America - Home URL: http://www.takepride.gov
13. [pid:3396] Mixed Bag, Exclusives, Indian Gifts Portal URL: http://www.indiangiftsportal.com/india-shopping/exclusives/mixed-	13. [pid:4] U.S. National Parks - Welcome to the U.S. National Parks Net URL: http://www.us-national-parks.net
14. [pid:3397] Wedding Gifts,Wedding Anniversary Gift,Wedding Gift Idea URL: http://www.indiangiftsportal.com/india-shopping/occasions/wedding	14. [pid:45] Recreation.gov URL: http://www.recreation.gov
15. [pid:3398] Birthday Gifts,Birthday Gift Idea,Send Birthday Gifts URL: http://www.indiangiftsportal.com/india-shopping/occasions/birthda	15. [pid:19] Sw Parks - SwParks - swparks.com URL: http://www.swparks.com
AT-Avg	**Norm (2)**
1. [pid:3261] E Business Solutions,Website Promotion Services URL: http://www.intermesh.net/advertis.html	1. [pid:3261] E Business Solutions,Website Promotion Services URL: http://www.intermesh.net/advertis.html
2. [pid:3403] Empty title field URL: http://news.indiamart.com	2. [pid:3403] Empty title field URL: http://news.indiamart.com
3. [pid:3373] Empty title field URL: http://www.indiamart.com	3. [pid:3377] Empty title field URL: http://www.indiangiftsportal.com
4. [pid:3374] Empty title field URL: http://apparel.indiamart.com	4. [pid:3411] Business Solutions,Ecommerce Business Solutions URL: http://www.intermesh.net
5. [pid:3400] Empty title field URL: http://handicraft.indiamart.com	5. [pid:3373] Empty title field URL: http://www.indiamart.com
6. [pid:3401] India Finance and Investment Guide, India URL: http://finance.indiamart.com	6. [pid:3374] Empty title field URL: http://apparel.indiamart.com
7. [pid:3402] Empty title field URL: http://health.indiamart.com	7. [pid:3400] Empty title field URL: http://handicraft.indiamart.com
8. [pid:3404] Empty title field URL: http://auto.indiamart.com	8. [pid:3401] India Finance and Investment Guide, India URL: http://finance.indiamart.com
9. [pid:3377] Empty title field URL: http://www.indiangiftsportal.com	9. [pid:3402] Empty title field URL: http://health.indiamart.com
10. [pid:3411] Business Solutions,Ecommerce Business Solutions,Business Applicat URL: http://www.intermesh.net	10. [pid:3404] Empty title field URL: http://auto.indiamart.com
11. [pid:3394] Anniversary Gifts,Wedding Anniversary Gift,Anniversary Gift URL: http://www.indiangiftsportal.com/india-shopping/occasions/anniver	11. [pid:3394] Anniversary Gifts,Wedding Anniversary Gift,Anniversary Gift URL: http://www.indiangiftsportal.com/india-shopping/occasions/anniver
12. [pid:3395] Jewelry Box, Jewelry Gift Box, Jewelry Box Shopping Online, Woode URL: http://www.indiangiftsportal.com/india-shopping/exclusives/jewell	12. [pid:3395] Jewelry Box, Jewelry Gift Box, Jewelry Box Shopping Online, Woode URL: http://www.indiangiftsportal.com/india-shopping/exclusives/jewell
13. [pid:3396] Mixed Bag, Exclusives, Indian Gifts Portal URL: http://www.indiangiftsportal.com/india-shopping/exclusives/mixed-	13. [pid:3396] Mixed Bag, Exclusives, Indian Gifts Portal URL: http://www.indiangiftsportal.com/india-shopping/exclusives/mixed-
14. [pid:3397] Wedding Gifts,Wedding Anniversary Gift,Wedding Gift Idea URL: http://www.indiangiftsportal.com/india-shopping/occasions/wedding	14. [pid:3397] Wedding Gifts,Wedding Anniversary Gift,Wedding Gift Idea URL: http://www.indiangiftsportal.com/india-shopping/occasions/wedding
15. [pid:3398] Birthday Gifts,Birthday Gift Idea,Send Birthday Gifts URL: http://www.indiangiftsportal.com/india-shopping/occasions/birthda	15. [pid:3398] Birthday Gifts,Birthday Gift Idea,Send Birthday Gifts URL: http://www.indiangiftsportal.com/india-shopping/occasions/birthda
Max	**SALSA**
1. [pid:1] National Park Service - Experience Your America URL: http://www.nps.gov	1. [pid:1] National Park Service - Experience Your America URL: http://www.nps.gov
2. [pid:211] Privacy Statement, National Park Service URL: http://www.nps.gov/privacy.htm	2. [pid:4408] Car Insurance America: Instant Quotes Online URL: http://www.carinsuranceamerica.com
3. [pid:24] GORP.com-adventure travel-hiking, national parks URL: http://www.gorp.com	3. [pid:3261] E Business Solutions,Website Promotion Services URL: http://www.intermesh.net/advertis.html
4. [pid:15] Park Geology Tour - Geologic Features URL: http://www.nature.nps.gov/grd/tour	4. [pid:3403] Empty title field URL: http://news.indiamart.com
5. [pid:2497] USGS Western Earth Surface Processes Team Home URL: http://wrgis.wr.usgs.gov/wgmt	5. [pid:3373] Empty title field URL: http://www.indiamart.com
6. [pid:379] NPS Search Portal URL: http://www.nps.gov/search.htm	6. [pid:3374] Empty title field URL: http://apparel.indiamart.com
7. [pid:2] National Park Guide URL: http://www.nps.gov/parks.html	7. [pid:3377] Empty title field URL: http://www.indiangiftsportal.com
8. [pid:28] NatureNet: The National Park Service's natural resource website URL: http://www.nature.nps.gov	8. [pid:3400] Empty title field URL: http://handicraft.indiamart.com
9. [pid:5] National Parks Conservation Association URL: http://www.npca.org	9. [pid:3401] India Finance and Investment Guide, India URL: http://finance.indiamart.com
10. [pid:25] L.L.Bean - Park Search URL: http://www.llbean.com/parksearch	10. [pid:3402] Empty title field URL: http://health.indiamart.com
11. [pid:19] Sw Parks - SwParks - swparks.com URL: http://www.swparks.com	11. [pid:3404] Empty title field URL: http://auto.indiamart.com
12. [pid:4] U.S. National Parks - Welcome to the U.S. National Parks Net URL: http://www.us-national-parks.net	12. [pid:3411] Business Solutions,Ecommerce Business Solutions URL: http://www.intermesh.net
13. [pid:2115] One stop shopping for residential, commercial and property manage URL: http://www.jasperrealestate.ab.ca	13. [pid:3394] Anniversary Gifts,Wedding Anniversary Gift,Anniversary Gift URL: http://www.indiangiftsportal.com/india-shopping/occasions/anniver
14. [pid:1547] What You Need to Know About^{tm} URL: http://about.com	14. [pid:3395] Jewelry Box, Jewelry Gift Box, Jewelry Box Shopping Online, Woode URL: http://www.indiangiftsportal.com/india-shopping/exclusives/jewell
15. [pid:64] Empty title field URL: http://www.americanparknetwork.com	15. [pid:3396] Mixed Bag, Exclusives, Indian Gifts Portal URL: http://www.indiangiftsportal.com/india-shopping/exclusives/mixed-

Table B.25: *Top 15 results for query "national parks"*

HITS	**HubAvg**
1. [pid:1811] movabletype.org URL: http://www.movabletype.org	1. [pid:1291] ArticleCentral - Content and Articles for Webmasters URL: http://articlecentral.com
2. [pid:1831] Boing Boing: A Directory of Wonderful Things URL: http://boingboing.net	2. [pid:2139] Microsoft Corporation URL: http://www.microsoft.com
3. [pid:1828] Metafilter — Community Weblog URL: http://www.metafilter.com	3. [pid:27] EFF: Homepage URL: http://www.eff.org
4. [pid:1827] Wired News URL: http://www.wired.com	4. [pid:1854] NabaviOnline URL: http://www.nabavionline.com
5. [pid:1838] The Doc Searls Weblog : Sunday, July 6, 2003 URL: http://doc.weblogs.com	5. [pid:15] Vtw Directory Page URL: http://www.vtw.org
6. [pid:1839] what's in rebecca's pocket? URL: http://www.rebeccablood.net	6. [pid:254] Internet Free Expression Alliance URL: http://www.ifea.net
7. [pid:1836] InstaPundit.Com URL: http://www.instapundit.com	7. [pid:526] The Center for Democracy and Technology URL: http://www.cdt.org
8. [pid:1825] blogdex - the weblog diffusion index URL: http://blogdex.media.mit.edu	8. [pid:498] American Civil Liberties Union URL: http://www.aclu.org
9. [pid:1835] kottke.org :: home of fine hypertext products URL: http://www.kottke.org	9. [pid:304] EFF Blue Ribbon Campaign Home Page URL: http://www.eff.org/blueribbon.html
10. [pid:1830] kuro5hin.org —— technology and culture, from the trenches URL: http://www.kuro5hin.org	10. [pid:5] P E A C E F I R E URL: http://www.peacefire.org
11. [pid:1826] BBC News Front Page URL: http://news.bbc.co.uk	11. [pid:1811] movabletype.org URL: http://www.movabletype.org
12. [pid:1837] Scripting News URL: http://www.scripting.com	12. [pid:4] HotWired: Cyber Rights Under Attack! URL: http://www.hotwired.com/special/indecent
13. [pid:1854] NabaviOnline URL: http://www.nabavionline.com	13. [pid:256] Global Internet Liberty Campaign Home Page URL: http://www.gilc.org
14. [pid:1829] FARK.com: Drew Curtis' FARK.com URL: http://www.fark.com	14. [pid:1812] Validation Results URL: http://validator.w3.org/check?uri=http%3A//www.ordinary-life.net
15. [pid:1853] Iranian.com - Today URL: http://www.iranian.com/today.html	15. [pid:156] libertus.net: about censorship and free speech URL: http://libertus.net
AT-Avg	**Norm (2)**
1. [pid:27] EFF: Homepage URL: http://www.eff.org	1. [pid:27] EFF: Homepage URL: http://www.eff.org
2. [pid:254] Internet Free Expression Alliance URL: http://www.ifea.net	2. [pid:254] Internet Free Expression Alliance URL: http://www.ifea.net
3. [pid:526] The Center for Democracy and Technology URL: http://www.cdt.org	3. [pid:526] The Center for Democracy and Technology URL: http://www.cdt.org
4. [pid:498] American Civil Liberties Union URL: http://www.aclu.org	4. [pid:498] American Civil Liberties Union URL: http://www.aclu.org
5. [pid:15] Vtw Directory Page URL: http://www.vtw.org	5. [pid:15] Vtw Directory Page URL: http://www.vtw.org
6. [pid:5] P E A C E F I R E URL: http://www.peacefire.org	6. [pid:5] P E A C E F I R E URL: http://www.peacefire.org
7. [pid:256] Global Internet Liberty Campaign Home Page URL: http://www.gilc.org	7. [pid:256] Global Internet Liberty Campaign Home Page URL: http://www.gilc.org
8. [pid:156] libertus.net: about censorship and free speech URL: http://libertus.net	8. [pid:156] libertus.net: about censorship and free speech URL: http://libertus.net
9. [pid:304] EFF Blue Ribbon Campaign Home Page URL: http://www.eff.org/blueribbon.html	9. [pid:304] EFF Blue Ribbon Campaign Home Page URL: http://www.eff.org/blueribbon.html
10. [pid:559] The Freedom Forum URL: http://www.freedomforum.org	10. [pid:559] The Freedom Forum URL: http://www.freedomforum.org
11. [pid:2010] CPSR Home Page URL: http://www.cpsr.org	11. [pid:2010] CPSR Home Page URL: http://www.cpsr.org
12. [pid:517] Going to the Dogs URL: http://www.liberty.org.uk/cacib	12. [pid:517] Going to the Dogs URL: http://www.liberty.org.uk/cacib
13. [pid:1739] ALA — Home URL: http://www.ala.org	13. [pid:4] HotWired: Cyber Rights Under Attack! URL: http://www.hotwired.com/special/indecent
14. [pid:540] Yaman Akdeniz CV, Cyber-Rights & Cyber-Liberties (UK) URL: http://www.leeds.ac.uk/law/pgs/yaman/yaman.htm	14. [pid:1739] ALA — Home URL: http://www.ala.org
15. [pid:557] FACT URL: http://w3.trib.com/FACT	15. [pid:540] Yaman Akdeniz CV, Cyber-Rights & Cyber-Liberties (UK) URL: http://www.leeds.ac.uk/law/pgs/yaman/yaman.htm
Max	**SALSA**
1. [pid:27] EFF: Homepage URL: http://www.eff.org	1. [pid:27] EFF: Homepage URL: http://www.eff.org
2. [pid:254] Internet Free Expression Alliance URL: http://www.ifea.net	2. [pid:254] Internet Free Expression Alliance URL: http://www.ifea.net
3. [pid:526] The Center for Democracy and Technology URL: http://www.cdt.org	3. [pid:498] American Civil Liberties Union URL: http://www.aclu.org
4. [pid:498] American Civil Liberties Union URL: http://www.aclu.org	4. [pid:526] The Center for Democracy and Technology URL: http://www.cdt.org
5. [pid:15] Vtw Directory Page URL: http://www.vtw.org	5. [pid:5] P E A C E F I R E URL: http://www.peacefire.org
6. [pid:5] P E A C E F I R E URL: http://www.peacefire.org	6. [pid:15] Vtw Directory Page URL: http://www.vtw.org
7. [pid:256] Global Internet Liberty Campaign Home Page URL: http://www.gilc.org	7. [pid:1811] movabletype.org URL: http://www.movabletype.org
8. [pid:156] libertus.net: about censorship and free speech URL: http://libertus.net	8. [pid:156] libertus.net: about censorship and free speech URL: http://libertus.net
9. [pid:304] EFF Blue Ribbon Campaign Home Page URL: http://www.eff.org/blueribbon.html	9. [pid:304] EFF Blue Ribbon Campaign Home Page URL: http://www.eff.org/blueribbon.html
10. [pid:559] The Freedom Forum URL: http://www.freedomforum.org	10. [pid:256] Global Internet Liberty Campaign Home Page URL: http://www.gilc.org
11. [pid:4] HotWired: Cyber Rights Under Attack! URL: http://www.hotwired.com/special/indecent	11. [pid:4] HotWired: Cyber Rights Under Attack! URL: http://www.hotwired.com/special/indecent
12. [pid:2010] CPSR Home Page URL: http://www.cpsr.org	12. [pid:1291] ArticleCentral - Content and Articles for Webmasters URL: http://articlecentral.com
13. [pid:517] Going to the Dogs URL: http://www.liberty.org.uk/cacib	13. [pid:13] Going to the Dogs URL: http://goingtothedogs.blogspot.com
14. [pid:894] Google URL: http://www.google.com	14. [pid:894] Google URL: http://www.google.com
15. [pid:1739] ALA — Home URL: http://www.ala.org	15. [pid:1515] Anonymizer - Online Privacy and Security URL: http://anonymizer.com

Table B.26: *Top 15 results for query "net censorship"*

HITS	**HubAvg**
1. [pid:338] SpringerLink: Lecture Notes in Computer Science URL: http://link.springer.de/link/service/series/0558/tocs/t2161.htm	1. [pid:68] Computational Geometry, Algorithms and Applications URL: http://www.cs.uu.nl/geobook
2. [pid:342] SpringerLink: Lecture Notes in Computer Science 2141 URL: http://link.springer.de/link/service/series/0558/tocs/t2141.htm	2. [pid:549] Directory of Computational Geometry Software URL: http://www.geom.umn.edu/software/cglist
3. [pid:346] SpringerLink: Lecture Notes in Computer Science URL: http://link.springer.de/link/service/series/0558/tocs/t2149.htm	3. [pid:547] The former CGAL home page URL: http://www.cs.uu.nl/CGAL
4. [pid:332] ALGO 2002 URL: http://www.dis.uniroma1.it/ algo02	4. [pid:552] Welcome to Springer, springer-verlag URL: http://www.springer.de
5. [pid:552] Welcome to Springer, springer-verlag URL: http://www.springer.de	5. [pid:548] LEDA moved to Algorithmic Solutions Software GmbH URL: http://www.mpi-sb.mpg.de/LEDA/leda.html
6. [pid:553] Mark Overmars Homepage URL: http://www.cs.uu.nl/people/markov	6. [pid:577] CMSC 754 - Comp Geom URL: http://www.cs.umd.edu/ mount/754
7. [pid:184] Pankaj K. Agarwal's Home Page URL: http://www.cs.duke.edu/ pankaj	7. [pid:553] Mark Overmars Homepage URL: http://www.cs.uu.nl/people/markov
8. [pid:697] Thomas H. Cormen URL: http://www.cs.dartmouth.edu/ thc	8. [pid:712] Cormen/Leiserson/Rivest/Stein: Introduction to Algorithms URL: http://theory.lcs.mit.edu/ clr
9. [pid:64] Algorithms Courses on the WWW URL: http://www.cs.pitt.edu/ kirk/algorithmcourses	9. [pid:64] Algorithms Courses on the WWW URL: http://www.cs.pitt.edu/ kirk/algorithmcourses
10. [pid:279] WAE '98 URL: http://www.mpi-sb.mpg.de/ wae98	10. [pid:574] Computational Geometry, Algorithms and Applications URL: http://www.cs.uu.nl/geobook/geom.html
11. [pid:6] HTML redirection URL: http://cui.unige.ch/tcs/random-approx	11. [pid:589] Google Directory - Science > Math > Geometry > Computati URL: http://directory.google.com/Top/Science/Math/Geometry/Computation
12. [pid:68] Computational Geometry, Algorithms and Applications URL: http://www.cs.uu.nl/geobook	12. [pid:592] Fortune's Voronoi algorithm, implemented visually URL: http://www.diku.dk/hjemmesider/studerende/duff/Fortune
13. [pid:353] ARACNE 2000 URL: http://www.dia.unisa.it/ uv/ARACNE2000.html	13. [pid:551] Otfried Cheong URL: http://www.cs.ust.hk/ otfried
14. [pid:673] Carleton Scientific URL: http://www.carleton-scientific.com	14. [pid:156] Computer Science Papers NEC Research Institute CiteSeer URL: http://citeseer.nj.nec.com/cs
15. [pid:475] ANALYSIS of ALGORITHMS HOME PAGE URL: http://pauillac.inria.fr/algo/AofA	15. [pid:281] David Eppstein URL: http://www.ics.uci.edu/ eppstein

AT-Avg	**Norm (2)**
1. [pid:64] Algorithms Courses on the WWW URL: http://www.cs.pitt.edu/ kirk/algorithmcourses	1. [pid:64] Algorithms Courses on the WWW URL: http://www.cs.pitt.edu/ kirk/algorithmcourses
2. [pid:68] Computational Geometry, Algorithms and Applications URL: http://www.cs.uu.nl/geobook	2. [pid:68] Computational Geometry, Algorithms and Applications URL: http://www.cs.uu.nl/geobook
3. [pid:475] ANALYSIS of ALGORITHMS HOME PAGE URL: http://pauillac.inria.fr/algo/AofA	3. [pid:475] ANALYSIS of ALGORITHMS HOME PAGE URL: http://pauillac.inria.fr/algo/AofA
4. [pid:548] LEDA moved to Algorithmic Solutions Software GmbH URL: http://www.mpi-sb.mpg.de/LEDA/leda.html	4. [pid:548] LEDA moved to Algorithmic Solutions Software GmbH URL: http://www.mpi-sb.mpg.de/LEDA/leda.html
5. [pid:225] IEEE Computer Society URL: http://computer.org	5. [pid:225] IEEE Computer Society URL: http://computer.org
6. [pid:411] Center for Discrete Mathematics and Theoretical Computer Science URL: http://dimacs.rutgers.edu	6. [pid:411] Center for Discrete Mathematics and Theoretical Computer Science URL: http://dimacs.rutgers.edu
7. [pid:31] MFCS'98 home page URL: http://www.fi.muni.cz/mfcs98	7. [pid:31] MFCS'98 home page URL: http://www.fi.muni.cz/mfcs98
8. [pid:549] Directory of Computational Geometry Software URL: http://www.geom.umn.edu/software/cglist	8. [pid:549] Directory of Computational Geometry Software URL: http://www.geom.umn.edu/software/cglist
9. [pid:156] Computer Science Papers NEC Research Institute CiteSeer URL: http://citeseer.nj.nec.com/cs	9. [pid:156] Computer Science Papers NEC Research Institute CiteSeer URL: http://citeseer.nj.nec.com/cs
10. [pid:157] CiteSeer: The NEC Research Institute Scientific Literature URL: http://citeseer.org	10. [pid:157] CiteSeer: The NEC Research Institute Scientific Literature URL: http://citeseer.org
11. [pid:552] Welcome to Springer, springer-verlag URL: http://www.springer.de	11. [pid:552] Welcome to Springer, springer-verlag URL: http://www.springer.de
12. [pid:6] HTML redirection URL: http://cui.unige.ch/tcs/random-approx	12. [pid:474] Complexity results for scheduling problems URL: http://www.mathematik.uni-osnabrueck.de/research/OR/class
13. [pid:474] Complexity results for scheduling problems URL: http://www.mathematik.uni-osnabrueck.de/research/OR/class	13. [pid:6] HTML redirection URL: http://cui.unige.ch/tcs/random-approx
14. [pid:547] The former CGAL home page URL: http://www.cs.uu.nl/CGAL	14. [pid:547] The former CGAL home page URL: http://www.cs.uu.nl/CGAL
15. [pid:181] David B. Shmoys URL: http://www.orie.cornell.edu/ shmoys	15. [pid:181] David B. Shmoys URL: http://www.orie.cornell.edu/ shmoys

Max	**SALSA**
1. [pid:64] Algorithms Courses on the WWW URL: http://www.cs.pitt.edu/ kirk/algorithmcourses	1. [pid:64] Algorithms Courses on the WWW URL: http://www.cs.pitt.edu/ kirk/algorithmcourses
2. [pid:68] Computational Geometry, Algorithms and Applications URL: http://www.cs.uu.nl/geobook	2. [pid:68] Computational Geometry, Algorithms and Applications URL: http://www.cs.uu.nl/geobook
3. [pid:549] Directory of Computational Geometry Software URL: http://www.geom.umn.edu/software/cglist	3. [pid:31] MFCS'98 home page URL: http://www.fi.muni.cz/mfcs98
4. [pid:548] LEDA moved to Algorithmic Solutions Software GmbH URL: http://www.mpi-sb.mpg.de/LEDA/leda.html	4. [pid:6] HTML redirection URL: http://cui.unige.ch/tcs/random-approx
5. [pid:475] ANALYSIS of ALGORITHMS HOME PAGE URL: http://pauillac.inria.fr/algo/AofA	5. [pid:81] MHHE: INTRODUCTION TO ALGORITHMS, Second Edition URL: http://www.mhhe.com/catalogs/0070131511.mhtml
6. [pid:225] IEEE Computer Society URL: http://computer.org	6. [pid:294] The Digital Object Identifier URL: http://www.doi.org
7. [pid:411] Center for Discrete Mathematics and Theoretical Computer Science URL: http://dimacs.rutgers.edu	7. [pid:231] Masaryk University Brno URL: http://www.muni.cz
8. [pid:31] MFCS'98 home page URL: http://www.fi.muni.cz/mfcs98	8. [pid:332] ALGO 2002 URL: http://www.dis.uniroma1.it/ algo02
9. [pid:156] Computer Science Papers NEC Research Institute CiteSeer URL: http://citeseer.nj.nec.com/cs	9. [pid:338] SpringerLink: Lecture Notes in Computer Science URL: http://link.springer.de/link/service/series/0558/tocs/t2161.htm
10. [pid:552] Welcome to Springer, springer-verlag URL: http://www.springer.de	10. [pid:342] SpringerLink: Lecture Notes in Computer Science 2141 URL: http://link.springer.de/link/service/series/0558/tocs/t2141.htm
11. [pid:547] The former CGAL home page URL: http://www.cs.uu.nl/CGAL	11. [pid:346] SpringerLink: Lecture Notes in Computer Science URL: http://link.springer.de/link/service/series/0558/tocs/t2149.htm
12. [pid:157] CiteSeer: The NEC Research Institute Scientific Literature Digita URL: http://citeseer.org	12. [pid:232] Empty title field URL: http://www.brno-city.cz
13. [pid:474] Complexity results for scheduling problems URL: http://www.mathematik.uni-osnabrueck.de/research/OR/class	13. [pid:84] Title Details - Cambridge University Press URL: http://www.cup.org/Reviews&blurbs/RanAlg/RanAlg.html
14. [pid:553] Mark Overmars Homepage URL: http://www.cs.uu.nl/people/markov	14. [pid:156] Computer Science Papers NEC Research Institute CiteSeer URL: http://citeseer.nj.nec.com/cs
15. [pid:181] David B. Shmoys URL: http://www.orie.cornell.edu/ shmoys	15. [pid:552] Welcome to Springer, springer-verlag URL: http://www.springer.de

Table B.27: *Top 15 results for query "randomized algorithms"*

HITS	**HubAvg**
1. [pid:4934] HonoluluAdvertiser.com > HawaiisClassifieds.com URL: http://www.hawaiisclassifieds.com	1. [pid:4710] Le Web des îles URL: http://www.chez.com/zanozile
2. [pid:4936] Gannett Company, Inc. URL: http://www.gannett.com	2. [pid:4709] Please stand by.. URL: http://www.sofcom.com.au
3. [pid:4917] AP MoneyWire URL: http://apmoneywire.mm.ap.org	3. [pid:4560] Sign in - Yahoo! Groups URL: http://groups.yahoo.com/group/mauritianrecipes/post
4. [pid:4893] e.thePeople : Honolulu Advertiser : What makes the Republi- can URL: http://www.e-thepeople.com/affiliates/honoluluadvertiser	4. [pid:4556] Microsoft bCentral - FastCounter URL: http://fastcounter.bcentral.com/fc-join
5. [pid:4868] News From The Associated Press URL: http://customwire.ap.org/dynamic/fronts/HOME?SITE=HIHAD	5. [pid:2155] Recipes are Cooking at NetCooks! URL: http://www.netcooks.com
6. [pid:4892] Honolulu Traffic Cameras, City and County of Honolulu URL: http://www.co.honolulu.hi.us/cameras/traffic.htm	6. [pid:161] Mauritian cuisine, cooking and recipes from Mauritius URL: http://ile-maurice.tripod.com
7. [pid:4906] News From The Associated Press URL: http://customwire.ap.org/dynamic/fronts/SPORTS?SITE=HIHAD	7. [pid:4706] Mauritius Australia Connection URL: http://www.cjp.net
8. [pid:4911] News From The Associated Press URL: http://customwire.ap.org/. . . /ENTERTAINMENT?SITE=HIHAD	8. [pid:4557] Mauritius Australia Connection URL: http://www.users.bigpond.com/clancy/travmau.htm
9. [pid:4918] News From The Associated Press URL: http://customwire.ap.org/. . . /BUSINESS?SITE=HIHAD	9. [pid:4561] SleepAngel.com - Are you snoring yourself to death? Sleep Apnea? URL: http://wcpsecure.com/app/aftrack.asp?afid=2139
10. [pid:4924] News From The Associated Press URL: http://customwire.ap.org/dynamic/fronts/TECHNOLOGY?SITE=HIHAD	10. [pid:3108] Chef Jobs Foodservice Culinary Institute Cooking School URL: http://chef2chef.net
11. [pid:4935] HonoluluAdvertiser.com > HawaiisClassifieds.com URL: http://www.hawaiisclassifieds.com/placead.html	11. [pid:4559] Golden Web Awards.com - 2002/2003 URL: http://www.goldenwebawards.com/officialawardwinner.shtml
12. [pid:4937] Hawaii's People2People URL: http://www.people2people.com/?connect=honolulu	12. [pid:4617] Mauritian Cuisine URL: http://members5.boardhost.com/cuisine46
13. [pid:4858] The Honolulu Advertiser - Hawaii's Newspaper URL: http://www.honoluluadvertiser.com	13. [pid:4707] alapage.com : livres, cd, dvd, vidéo, cédéro URL: http://www.alapage.com/?donnee_appel=CJPMA
14. [pid:4876] HonoluluAdvertiser.com > Subscriber Services URL: http://www.honoluluadvertiser.com/subscribe	14. [pid:3] Top Secret Recipes on the Web URL: http://www.topsecretrecipes.com
15. [pid:167] Simple spring green - The Honolulu Advertiser - Hawaii's News- pap URL: http://the.honoluluadvertiser.com/current/il/taste	15. [pid:26] Favorite Spanish Food Recipes URL: http://www.xmission.com/ dderhak/recipes.html
AT-Avg	**Norm (2)**
1. [pid:2] EPICURIOUS: WORLD'S GREATEST RECIPE COLLECTION URL: http://www.epicurious.com	1. [pid:4934] HonoluluAdvertiser.com > HawaiisClassifieds.com URL: http://www.hawaiisclassifieds.com
2. [pid:1] All Recipes — Recipes URL: http://www.allrecipes.com	2. [pid:4936] Gannett Company, Inc. URL: http://www.gannett.com
3. [pid:16] Food Network URL: http://www.foodtv.com	3. [pid:4917] AP MoneyWire URL: http://apmoneywire.mm.ap.org
4. [pid:5] RecipeSource: Your Source for Recipes on the Internet URL: http://www.recipesource.com	4. [pid:4893] e.thePeople : Honolulu Advertiser : What makes the Republi- can URL: http://www.e-thepeople.com/affiliates/honoluluadvertiser
5. [pid:3] Top Secret Recipes on the Web URL: http://www.topsecretrecipes.com	5. [pid:4868] News From The Associated Press URL: http://customwire.ap.org/dynamic/fronts/HOME?SITE=HIHAD
6. [pid:2283] Find Lost Recipes at Recipelink.com - Cooking on the Net Si URL: http://www.recipelink.com	6. [pid:4892] Honolulu Traffic Cameras, City and County of Honolulu URL: http://www.co.honolulu.hi.us/cameras/traffic.htm
7. [pid:30] www.BettyCrocker.com URL: http://www.bettycrocker.com	7. [pid:4906] News From The Associated Press URL: http://customwire.ap.org/. . . /SPORTS?SITE=HIHAD
8. [pid:38] FATFREE: The Low Fat Vegetarian Recipe Archive URL: http://www.fatfree.com	8. [pid:4911] News From The Associated Press URL: http://customwire.ap.org/. . . /entertainment?SITE=HIHAD
9. [pid:23] VegWeb - Vegan/Vegetarian Info URL: http://www.vegweb.com	9. [pid:4918] News From The Associated Press URL: http://customwire.ap.org/. . . /business?SITE=HIHAD
10. [pid:79] Meals For You - Thousands Of Delicious Recipes And Meals URL: http://www.mealsforyou.com	10. [pid:4924] News From The Associated Press URL: http://customwire.ap.org/. . . /technology?SITE=HIHAD
11. [pid:2491] What You Need to Know About^{tm} URL: http://www.about.com	11. [pid:4935] HonoluluAdvertiser.com > HawaiisClassifieds.com URL: http://www.hawaiisclassifieds.com/placead.html
12. [pid:2275] Mimi's Cyber-Kitchen Recipes - Recipes Recipes Recipes - Your Fir URL: http://www.cyber-kitchen.com	12. [pid:4937] Hawaii's People2People URL: http://www.people2people.com/?connect=honolulu
13. [pid:69] Meals.com - Recipes, Cooking & Meal Planning URL: http://www.my-meals.com	13. [pid:4858] The Honolulu Advertiser - Hawaii's Newspaper URL: http://www.honoluluadvertiser.com
14. [pid:6] CopyKat.com - Your home for recipes you'd normally find AWAY from URL: http://www.copykat.com	14. [pid:4876] HonoluluAdvertiser.com > Subscriber Services URL: http://www.honoluluadvertiser.com/subscribe
15. [pid:62] Free Recipes from iChef URL: http://www.ichef.com	15. [pid:167] Simple spring green - The Honolulu Advertiser - Hawaii's News- pap URL: http://the.honoluluadvertiser.com/current/il/taste
Max	**SALSA**
1. [pid:2] EPICURIOUS: WORLD'S GREATEST RECIPE COLLECTION URL: http://www.epicurious.com	1. [pid:2] EPICURIOUS: WORLD'S GREATEST RECIPE COLLECTION URL: http://www.epicurious.com
2. [pid:16] Food Network URL: http://www.foodtv.com	2. [pid:16] Food Network URL: http://www.foodtv.com
3. [pid:1] All Recipes — Recipes URL: http://www.allrecipes.com	3. [pid:1] All Recipes — Recipes URL: http://www.allrecipes.com
4. [pid:5] RecipeSource: Your Source for Recipes on the Internet URL: http://www.recipesource.com	4. [pid:5] RecipeSource: Your Source for Recipes on the Internet URL: http://www.recipesource.com
5. [pid:3] Top Secret Recipes on the Web URL: http://www.topsecretrecipes.com	5. [pid:2491] What You Need to Know About^{tm} URL: http://www.about.com
6. [pid:2283] Find Lost Recipes at Recipelink.com - Cooking on the Net Si URL: http://www.recipelink.com	6. [pid:23] VegWeb - Vegan/Vegetarian Info URL: http://www.vegweb.com
7. [pid:30] www.BettyCrocker.com URL: http://www.bettycrocker.com	7. [pid:38] FATFREE: The Low Fat Vegetarian Recipe Archive URL: http://www.fatfree.com
8. [pid:23] VegWeb - Vegan/Vegetarian Info URL: http://www.vegweb.com	8. [pid:1795] Cutting-edge natural health treatments for Aging, Weight Loss URL: http://www.youngagain.com
9. [pid:38] FATFREE: The Low Fat Vegetarian Recipe Archive URL: http://www.fatfree.com	9. [pid:3] Top Secret Recipes on the Web URL: http://www.topsecretrecipes.com
10. [pid:2491] What You Need to Know About^{tm} URL: http://www.about.com	10. [pid:4710] Le Web des îles URL: http://www.chez.com/zanozile
11. [pid:79] Meals For You - Thousands Of Delicious Recipes And Meals URL: http://www.mealsforyou.com	11. [pid:4709] Please stand by.. URL: http://www.sofcom.com.au
12. [pid:1795] Cutting-edge natural health treatments for Aging, Weight Loss, Pr URL: http://www.youngagain.com	12. [pid:2283] Find Lost Recipes at Recipelink.com - Cooking on the Net Si URL: http://www.recipelink.com
13. [pid:2275] Mimi's Cyber-Kitchen Recipes - Recipes Recipes Recipes - Your Fir URL: http://www.cyber-kitchen.com	13. [pid:4560] Sign in - Yahoo! Groups URL: http://groups.yahoo.com/group/mauritianrecipes/post
14. [pid:69] Meals.com - Recipes, Cooking & Meal Planning URL: http://www.my-meals.com	14. [pid:30] www.BettyCrocker.com URL: http://www.bettycrocker.com
15. [pid:6] CopyKat.com - Your home for recipes you'd normally find AWAY from URL: http://www.copykat.com	15. [pid:79] Meals For You - Thousands Of Delicious Recipes And Meals URL: http://www.mealsforyou.com

Table B.28: *Top 15 results for query "recipes"*

HITS	HubAvg
1. [pid:2472] Site Meter - Counter and Statistics Tracker URL: http://www.sitemeter.com/stats.asp?site=freya	1. [pid:166] Fan Forum: Entertainment 4 Fans URL: http://www.fanforum.com
2. [pid:2469] Dreambook - Camarila's Image Galleries of Fortune Cities URL: http://books.dreambook.com/camarila/alternate.sign.html	2. [pid:179] Forums 4 Fans: Roswell (1) URL: http://www.forums4fans.com/ultimatebb.php?ubb=forum&f=3
3. [pid:2470] Camarila's Image Galleries's Dreambook URL: http://books.dreambook.com/camarila/main.html	3. [pid:1] Crashdown.com URL: http://www.crashdown.com
4. [pid:2471] Camarila's Image Galleries of Fortune Cities's Dreambook URL: http://books.dreambook.com/camarila/alternate.html	4. [pid:12] Welcome to Roswell Rods.com URL: http://www.roswellrods.com
5. [pid:2460] Camarila's Fantasy Image Galleries URL: http://members.fortunecity.com/camarila/fantasy.html	5. [pid:190] William Sadler - Wild on the Web (The Official Web Site) URL: http://www.williamsadler.com
6. [pid:2461] Camarila's Sci-Fi Image Galleries URL: http://members.fortunecity.com/camarila/scifi.html	6. [pid:1575] Adobe Acrobat Reader - Download URL: http://www.adobe.com/products/acrobat/readstep.html
7. [pid:2462] Camarila's Horror Image Galleries URL: http://members.fortunecity.com/camarila/horror.html	7. [pid:91] ——————————Welcome to Roswell land - Take a seat —— URL: http://roswell_land.tripod.com
8. [pid:2466] Camarila's Artists Gallery List URL: http://members.fortunecity.com/camarila/artistlist.html	8. [pid:2605] Roswell Movie [Campaign] URL: http://www.roswellmovie.net
9. [pid:2473] MIDI PAGE URL: http://rivendell.fortunecity.com/redguard/636/midi.html	9. [pid:54] Roswell: Crashdown (Episodes) URL: http://www.crashdown.com/episodes
10. [pid:2475] Camarila's X-Files Pages -9 seasons-complete URL: http://members.fortunecity.com/camarila/xfiles.html	10. [pid:127] general5 URL: http://www.roswellproof.homestead.com
11. [pid:2488] CAMARILA'S FOREVER KNIGHT PAGE URL: http://members.fortunecity.com/camarila/foreverknight.html	11. [pid:777] LeavingNormal.Net URL: http://www.leavingnormal.net
12. [pid:2459] Camarila's Lord of the Rings Image Gallery URL: http://rivendell.fortunecity.com/rhydin/959/lordofrings.html	12. [pid:38] Roswell Screen Grab Galleries URL: http://www.fortunecity.co.uk/roswell/philosophy/63
13. [pid:2463] Camarila's Kindred Clan Pages URL: http://members.fortunecity.com/camarila/kindred.html	13. [pid:504] Dreambook - Roswell Mp3s - downloadable music from the WB TV show URL: http://books.dreambook.com/phoebelove/roswell.sign.html
14. [pid:2465] Camarila's H.R.Giger Galleries URL: http://members.fortunecity.com/lioncourt77/gigerintro.html	14. [pid:505] Roswell Mp3s Discussion Board @ www.ezboard.com URL: http://pub19.ezboard.com/broswellmp3s
15. [pid:2474] Camarila's Sci-Fi and Horror TV Episode Guides URL: http://members.fortunecity.com/camarila/tvshows.html	15. [pid:2230] The Rittenhouse Review URL: http://rittenhouse.blogspot.com

AT-Avg	Norm (2)
1. [pid:2472] Site Meter - Counter and Statistics Tracker URL: http://www.sitemeter.com/stats.asp?site=freya	1. [pid:2472] Site Meter - Counter and Statistics Tracker URL: http://www.sitemeter.com/stats.asp?site=freya
2. [pid:2469] Dreambook - Camarila's Image Galleries of Fortune Cities URL: http://books.dreambook.com/camarila/alternate.sign.html	2. [pid:2469] Dreambook - Camarila's Image Galleries of Fortune Cities URL: http://books.dreambook.com/camarila/alternate.sign.html
3. [pid:2470] Camarila's Image Galleries's Dreambook URL: http://books.dreambook.com/camarila/main.html	3. [pid:2470] Camarila's Image Galleries's Dreambook URL: http://books.dreambook.com/camarila/main.html
4. [pid:2471] Camarila's Image Galleries of Fortune Cities's Dreambook URL: http://books.dreambook.com/camarila/alternate.html	4. [pid:2471] Camarila's Image Galleries of Fortune Cities's Dreambook URL: http://books.dreambook.com/camarila/alternate.html
5. [pid:1379] Ampira Hosting - web hosting, domain name, email address service URL: http://www.ampira.com	5. [pid:1379] Ampira Hosting - web hosting, domain name, email address service URL: http://www.ampira.com
6. [pid:2468] Edward Gorey, amphigorey URL: http://www.geocities.com/SoHo/Canvas/9700/gorey1.html	6. [pid:2468] Edward Gorey, amphigorey URL: http://www.geocities.com/SoHo/Canvas/9700/gorey1.html
7. [pid:2464] Camarila's Angels and Fairies Pages URL: http://www.geocities.com/gabriella66/angel.html	7. [pid:2464] Camarila's Angels and Fairies Pages URL: http://www.geocities.com/gabriella66/angel.html
8. [pid:2476] Smallville Pages, Episode Guide,Photos and Links URL: http://www.geocities.com/gabriella66/smallville.html	8. [pid:2476] Smallville Pages, Episode Guide,Photos and Links URL: http://www.geocities.com/gabriella66/smallville.html
9. [pid:2486] Camarila's Total Recall-2070 Episode Guide URL: http://www.geocities.com/camarilasouth/totalrecall2070.html	9. [pid:2486] Camarila's Total Recall-2070 Episode Guide URL: http://www.geocities.com/camarilasouth/totalrecall2070.html
10. [pid:2487] Highlander Pages-6 seasons episode guide URL: http://www.geocities.com/akasha7_7/highlander.html	10. [pid:2487] Highlander Pages-6 seasons episode guide URL: http://www.geocities.com/akasha7_7/highlander.html
11. [pid:2460] Camarila's Fantasy Image Galleries URL: http://members.fortunecity.com/camarila/fantasy.html	11. [pid:2460] Camarila's Fantasy Image Galleries URL: http://members.fortunecity.com/camarila/fantasy.html
12. [pid:2461] Camarila's Sci-Fi Image Galleries URL: http://members.fortunecity.com/camarila/scifi.html	12. [pid:2461] Camarila's Sci-Fi Image Galleries URL: http://members.fortunecity.com/camarila/scifi.html
13. [pid:2462] Camarila's Horror Image Galleries URL: http://members.fortunecity.com/camarila/horror.html	13. [pid:2462] Camarila's Horror Image Galleries URL: http://members.fortunecity.com/camarila/horror.html
14. [pid:2466] Camarila's Artists Gallery List URL: http://members.fortunecity.com/camarila/artistlist.html	14. [pid:2466] Camarila's Artists Gallery List URL: http://members.fortunecity.com/camarila/artistlist.html
15. [pid:2473] MIDI PAGE URL: http://rivendell.fortunecity.com/redguard/636/midi.html	15. [pid:2473] MIDI PAGE URL: http://rivendell.fortunecity.com/redguard/636/midi.html

Max	SALSA
1. [pid:5] Roswell, NM URL: http://www.roswellnm.org	1. [pid:5] Roswell, NM URL: http://www.roswellnm.org
2. [pid:12] Welcome to Roswell Rods.com URL: http://www.roswellrods.com	2. [pid:12] Welcome to Roswell Rods.com URL: http://www.roswellrods.com
3. [pid:2] Welcome to Adobe GoLive 6 URL: http://www.roswell.org	3. [pid:166] Fan Forum: Entertainment 4 Fans URL: http://www.fanforum.com
4. [pid:127] general5 URL: http://www.roswellproof.homestead.com	4. [pid:127] general5 URL: http://www.roswellproof.homestead.com
5. [pid:91] ——————————Welcome to Roswell land - Take a seat —— URL: http://roswell_land.tripod.com	5. [pid:91] ——————————Welcome to Roswell land - Take a seat —— URL: http://roswell_land.tripod.com
6. [pid:441] Roswell UFO Crash of July 1947 / Nazi UFOs & Operation Paperc URL: http://www.roswellufocrash.com	6. [pid:1241] San Antonio Express-News Archives URL: http://archives.newsbank.com/saenews
7. [pid:1975] The Chaparral Rockhounds Gem And Mineral Society: Main page URL: http://www.chaparralrockhounds.com	7. [pid:2472] Site Meter - Counter and Statistics Tracker URL: http://www.sitemeter.com/stats.asp?site=freya
8. [pid:101] MP3.com: Lights Over Roswell URL: http://www.lightsoverroswell.com	8. [pid:1277] Rackspace Managed Hosting - Dedicated Hosting with Fanatical Supp URL: http://www.rackspace.com/?supbid=mysa15
9. [pid:662] Empty title field URL: http://www.mufon.com	9. [pid:2469] Dreambook - Camarila's Image Galleries of Fortune Cities URL: http://books.dreambook.com/camarila/alternate.sign.html
10. [pid:78] Empty title field URL: http://www.roswellsearch.com	10. [pid:2470] Camarila's Image Galleries's Dreambook URL: http://books.dreambook.com/camarila/main.html
11. [pid:1981] Church Christ - ChurchChrist - churchchrist.org URL: http://www.churchchrist.org	11. [pid:2471] Camarila's Image Galleries of Fortune Cities's Dreambook URL: http://books.dreambook.com/camarila/alternate.html
12. [pid:1009] Geography Home Page URL: Http://geography.miningco.com	12. [pid:101] MP3.com: Lights Over Roswell URL: http://www.lightsoverroswell.com
13. [pid:1401] BUGSTUFF – Home URL: http://www.vwbugstuff.com	13. [pid:78] Empty title field URL: http://www.roswellsearch.com
14. [pid:7] Roswell Report: Case Closed URL: http://www.af.mil/lib/roswell	14. [pid:441] Roswell UFO Crash of July 1947 / Nazi UFOs & Operation Paperc URL: http://www.roswellufocrash.com
15. [pid:360] UFO books by Beverly Fox, Roswell NM URL: http://www.uforanks.com	15. [pid:1975] The Chaparral Rockhounds Gem And Mineral Society: Main page URL: http://www.chaparralrockhounds.com

Table B.29: *Top 15 results for query "roswell"*

HITS	**HubAvg**
1. [pid:135] AltaVista URL: http://www.altavista.com	1. [pid:135] AltaVista URL: http://www.altavista.com
2. [pid:1401] Ego Surf - EgoSurf - egosurf.com URL: http://www.egosurf.com	2. [pid:745] Yahoo! URL: http://www.yahoo.com
3. [pid:5225] Yahoo! Danmark URL: http://www.yahoo.dk	3. [pid:732] Google URL: http://www.google.com
4. [pid:3390] AltaVista Text-Only Search URL: http://ragingsearch.altavista.com	4. [pid:762] Lycos Home Page URL: http://www.lycos.com
5. [pid:4539] Euroseek URL: http://euroseek.net	5. [pid:82] My Excite URL: http://www.excite.com
6. [pid:1721] Your Search Engine Internet directory, information, search engine URL: http://www.searchpalm.com	6. [pid:760] Homepage HotBot Web Search URL: http://www.hotbot.com
7. [pid:143] About Web Search - Guide to Search Engine Optimization & Online S URL: http://websearch.about.com	7. [pid:1] Dogpile. Unleash the power of meta-search! URL: http://www.dogpile.com
8. [pid:4979] Abacho - THE POWERFUL SEARCHENGINE!! URL: http://www.abacho.co.uk	8. [pid:902] WebCrawler Index - WebCrawler URL: http://www.webcrawler.com
9. [pid:1844] careerhighway.com URL: http://www.careerhighway.com	9. [pid:840] Northern Light URL: http://www.northernlight.com
10. [pid:5766] Ananzi South Africa - Search Engine URL: http://www.ananzi.co.za	10. [pid:124] AlltheWeb.com URL: http://www.alltheweb.com
11. [pid:1303] Empty title field URL: http://www.portalhub.com	11. [pid:907] GO.com URL: http://www.infoseek.com
12. [pid:5283] DINO-Online Suchmaschine Webkatalog Linkliste Internet Verzeichni URL: http://www.dino-online.de	12. [pid:6] Search.com URL: http://www.search.com
13. [pid:5296] Lycos.de URL: http://www.lycos.de	13. [pid:773] LookSmart URL: http://www.looksmart.com
14. [pid:5264] Excite France URL: http://www.excite.fr	14. [pid:758] Ask Jeeves - Ask.com URL: http://www.askjeeves.com
15. [pid:5285] Fireball - Die Suchmaschine URL: http://www.fireball.de	15. [pid:2] Search Engine Watch: Tips About Internet Search Engines & Search URL: http://searchenginewatch.com
AT-Avg	**Norm (2)**
1. [pid:135] AltaVista URL: http://www.altavista.com	1. [pid:135] AltaVista URL: http://www.altavista.com
2. [pid:745] Yahoo! URL: http://www.yahoo.com	2. [pid:1401] Ego Surf - EgoSurf - egosurf.com URL: http://www.egosurf.com
3. [pid:732] Google URL: http://www.google.com	3. [pid:5225] Yahoo! Danmark URL: http://www.yahoo.dk
4. [pid:762] Lycos Home Page URL: http://www.lycos.com	4. [pid:143] About Web Search - Guide to Search Engine Optimization & Online S URL: http://websearch.about.com
5. [pid:760] Homepage HotBot Web Search URL: http://www.hotbot.com	5. [pid:3390] AltaVista Text-Only Search URL: http://ragingsearch.altavista.com
6. [pid:1] Dogpile. Unleash the power of meta-search! URL: http://www.dogpile.com	6. [pid:1303] Empty title field URL: http://www.portalhub.com
7. [pid:82] My Excite URL: http://www.excite.com	7. [pid:4539] Euroseek URL: http://euroseek.net
8. [pid:902] WebCrawler Index - WebCrawler URL: http://www.webcrawler.com	8. [pid:1721] Your Search Engine Internet directory, information, search engine URL: http://www.searchpalm.com
9. [pid:840] Northern Light URL: http://www.northernlight.com	9. [pid:1844] careerhighway.com URL: http://www.careerhighway.com
10. [pid:6] Search.com URL: http://www.search.com	10. [pid:4958] CIA - The World Factbook 2002 URL: http://www.odci.gov/cia/publications/factbook
11. [pid:124] AlltheWeb.com URL: http://www.alltheweb.com	11. [pid:5415] IraqiOasis.com Web's leading source to Iraqi Info. URL: http://www.iraqioasis.com
12. [pid:773] LookSmart URL: http://www.looksmart.com	12. [pid:3368] Empty title field URL: http://www.anzwers.com.au
13. [pid:758] Ask Jeeves - Ask.com URL: http://www.askjeeves.com	13. [pid:4979] Abacho - THE POWERFUL SEARCHENGINE! URL: http://www.abacho.co.uk
14. [pid:907] GO.com URL: http://www.infoseek.com	14. [pid:4378] FirstGov - Your First Click to the US Government URL: http://www.firstgov.gov
15. [pid:5] mamma URL: http://www.mamma.com	15. [pid:5264] Excite France URL: http://www.excite.fr
Max	**SALSA**
1. [pid:135] AltaVista URL: http://www.altavista.com	1. [pid:135] AltaVista URL: http://www.altavista.com
2. [pid:745] Yahoo! URL: http://www.yahoo.com	2. [pid:745] Yahoo! URL: http://www.yahoo.com
3. [pid:732] Google URL: http://www.google.com	3. [pid:732] Google URL: http://www.google.com
4. [pid:762] Lycos Home Page URL: http://www.lycos.com	4. [pid:762] Lycos Home Page URL: http://www.lycos.com
5. [pid:760] Homepage HotBot Web Search URL: http://www.hotbot.com	5. [pid:760] Homepage HotBot Web Search URL: http://www.hotbot.com
6. [pid:1] Dogpile. Unleash the power of meta-search! URL: http://www.dogpile.com	6. [pid:1] Dogpile. Unleash the power of meta-search! URL: http://www.dogpile.com
7. [pid:82] My Excite URL: http://www.excite.com	7. [pid:82] My Excite URL: http://www.excite.com
8. [pid:902] WebCrawler Index - WebCrawler URL: http://www.webcrawler.com	8. [pid:902] WebCrawler Index - WebCrawler URL: http://www.webcrawler.com
9. [pid:840] Northern Light URL: http://www.northernlight.com	9. [pid:840] Northern Light URL: http://www.northernlight.com
10. [pid:124] AlltheWeb.com URL: http://www.alltheweb.com	10. [pid:124] AlltheWeb.com URL: http://www.alltheweb.com
11. [pid:6] Search.com URL: http://www.search.com	11. [pid:2] Search Engine Watch: Tips About Internet Search Engines & Search URL: http://searchenginewatch.com
12. [pid:907] GO.com URL: http://www.infoseek.com	12. [pid:6] Search.com URL: http://www.search.com
13. [pid:773] LookSmart URL: http://www.looksmart.com	13. [pid:907] GO.com URL: http://www.infoseek.com
14. [pid:2] Search Engine Watch: Tips About Internet Search Engines & Search URL: http://searchenginewatch.com	14. [pid:1401] Ego Surf - EgoSurf - egosurf.com URL: http://www.egosurf.com
15. [pid:758] Ask Jeeves - Ask.com URL: http://www.askjeeves.com	15. [pid:773] LookSmart URL: http://www.looksmart.com

Table B.30: *Top 15 results for query "search engines"*

HITS	**HubAvg**
1. [pid:11] The Oregon Shakespeare Festival URL: http://www.orshakes.org	1. [pid:1174] Mr. William Shakespeare and the Internet URL: http://daphne.palomar.edu/shakespeare
2. [pid:34] Shakespeare & Company URL: http://www.shakespeare-company.org	2. [pid:682] The Complete Works of William Shakespeare URL: http://the-tech.mit.edu/Shakespeare/works.html
3. [pid:35] Idaho Shakespeare Festival URL: http://www.idahoshakespeare.org	3. [pid:2] shakespeare.com home URL: http://www.shakespeare.com
4. [pid:111] Welcome to the Utah Shakespearean Festival URL: http://www.bard.org	4. [pid:6] Shakespeare's Globe Theatre URL: http://www.rdg.ac.uk/globe
5. [pid:15] The Shakespeare Theatre URL: http://www.shakespearedc.org	5. [pid:1] The Complete Works of William Shakespeare URL: http://the-tech.mit.edu/Shakespeare
6. [pid:28] Alabama Shakespeare Festival URL: http://www.asf.net	6. [pid:8] RSC - Royal Shakespeare Company URL: http://www.rsc.org.uk
7. [pid:1174] Mr. William Shakespeare and the Internet URL: http://daphne.palomar.edu/shakespeare	7. [pid:13] Shakespeare Oxford Society Home Page URL: http://www.shakespeare-oxford.com
8. [pid:21] Shakespeare's Globe Theatre, Bankside, Southwark, London URL: http://www.shakespeares-globe.org	8. [pid:17] Shakespeare Magazine URL: http://www.shakespearemag.com
9. [pid:39] Welcome to Georgia Shakespeare Festival URL: http://www.gashakespeare.org	9. [pid:10] Shakespeare: Internet Editions URL: http://web.uvic.ca/shakespeare
10. [pid:66] Kentucky Shakespeare Festival Welcomes You! URL: http://www.kyshakes.org	10. [pid:21] Shakespeare's Globe Theatre, Bankside, Southwark, London URL: http://www.shakespeares-globe.org
11. [pid:18] Shakespeare & Company - Lunatics, Lovers, Madmen and Clowns URL: http://www.shakespeare.org	11. [pid:7] The Shakespeare Birthplace Trust URL: http://www.shakespeare.org.uk
12. [pid:8] RSC - Royal Shakespeare Company URL: http://www.rsc.org.uk	12. [pid:9] Folger Shakespeare Library URL: http://www.folger.edu
13. [pid:44] Colorado Shakespeare Festival URL: http://www.coloradoshakes.org	13. [pid:3394] Internet Public Library: Shakespeare Bookshelf URL: http://www.ipl.org/reading/shakespeare/shakespeare.html
14. [pid:53] Michigan Shakespeare Festival URL: http://www.michshakefest.org	14. [pid:24] Shakespeare Resource Center URL: http://www.bardweb.net
15. [pid:46] The Shakespeare Theatre of New Jersey URL: http://www.njshakespeare.org	15. [pid:729] Surfing with the Bard URL: http://www.ulen.com/shakespeare
AT-Avg	**Norm (2)**
1. [pid:1174] Mr. William Shakespeare and the Internet URL: http://daphne.palomar.edu/shakespeare	1. [pid:1174] Mr. William Shakespeare and the Internet URL: http://daphne.palomar.edu/shakespeare
2. [pid:2] shakespeare.com home URL: http://www.shakespeare.com	2. [pid:682] The Complete Works of William Shakespeare URL: http://the-tech.mit.edu/Shakespeare/works.html
3. [pid:6] Shakespeare's Globe Theatre URL: http://www.rdg.ac.uk/globe	3. [pid:2] shakespeare.com home URL: http://www.shakespeare.com
4. [pid:682] The Complete Works of William Shakespeare URL: http://the-tech.mit.edu/Shakespeare/works.html	4. [pid:6] Shakespeare's Globe Theatre URL: http://www.rdg.ac.uk/globe
5. [pid:10] Shakespeare: Internet Editions URL: http://web.uvic.ca/shakespeare	5. [pid:1] The Complete Works of William Shakespeare URL: http://the-tech.mit.edu/Shakespeare
6. [pid:1] The Complete Works of William Shakespeare URL: http://the-tech.mit.edu/Shakespeare	6. [pid:10] Shakespeare: Internet Editions URL: http://web.uvic.ca/shakespeare
7. [pid:13] Shakespeare Oxford Society Home Page URL: http://www.shakespeare-oxford.com	7. [pid:13] Shakespeare Oxford Society Home Page URL: http://www.shakespeare-oxford.com
8. [pid:17] Shakespeare Magazine URL: http://www.shakespearemag.com	8. [pid:17] Shakespeare Magazine URL: http://www.shakespearemag.com
9. [pid:24] Shakespeare Resource Center URL: http://www.bardweb.net	9. [pid:21] Shakespeare's Globe Theatre, Bankside, Southwark, London URL: http://www.shakespeares-globe.org
10. [pid:7] The Shakespeare Birthplace Trust URL: http://www.shakespeare.org.uk	10. [pid:7] The Shakespeare Birthplace Trust URL: http://www.shakespeare.org.uk
11. [pid:9] Folger Shakespeare Library URL: http://www.folger.edu	11. [pid:24] Shakespeare Resource Center URL: http://www.bardweb.net
12. [pid:21] Shakespeare's Globe Theatre, Bankside, Southwark, London URL: http://www.shakespeares-globe.org	12. [pid:9] Folger Shakespeare Library URL: http://www.folger.edu
13. [pid:8] RSC - Royal Shakespeare Company URL: http://www.rsc.org.uk	13. [pid:8] RSC - Royal Shakespeare Company URL: http://www.rsc.org.uk
14. [pid:729] Surfing with the Bard URL: http://www.ulen.com/shakespeare	14. [pid:729] Surfing with the Bard URL: http://www.ulen.com/shakespeare
15. [pid:3] Mr. William Shakespeare and the Internet URL: http://shakespeare.palomar.edu	15. [pid:11] The Oregon Shakespeare Festival URL: http://www.orshakes.org
Max	**SALSA**
1. [pid:1174] Mr. William Shakespeare and the Internet URL: http://daphne.palomar.edu/shakespeare	1. [pid:1174] Mr. William Shakespeare and the Internet URL: http://daphne.palomar.edu/shakespeare
2. [pid:682] The Complete Works of William Shakespeare URL: http://the-tech.mit.edu/Shakespeare/works.html	2. [pid:682] The Complete Works of William Shakespeare URL: http://the-tech.mit.edu/Shakespeare/works.html
3. [pid:6] Shakespeare's Globe Theatre URL: http://www.rdg.ac.uk/globe	3. [pid:2] shakespeare.com home URL: http://www.shakespeare.com
4. [pid:2] shakespeare.com home URL: http://www.shakespeare.com	4. [pid:6] Shakespeare's Globe Theatre URL: http://www.rdg.ac.uk/globe
5. [pid:1] The Complete Works of William Shakespeare URL: http://the-tech.mit.edu/Shakespeare	5. [pid:1] The Complete Works of William Shakespeare URL: http://the-tech.mit.edu/Shakespeare
6. [pid:10] Shakespeare: Internet Editions URL: http://web.uvic.ca/shakespeare	6. [pid:8] RSC - Royal Shakespeare Company URL: http://www.rsc.org.uk
7. [pid:13] Shakespeare Oxford Society Home Page URL: http://www.shakespeare-oxford.com	7. [pid:13] Shakespeare Oxford Society Home Page URL: http://www.shakespeare-oxford.com
8. [pid:17] Shakespeare Magazine URL: http://www.shakespearemag.com	8. [pid:21] Shakespeare's Globe Theatre, Bankside, Southwark, London URL: http://www.shakespeares-globe.org
9. [pid:24] Shakespeare Resource Center URL: http://www.bardweb.net	9. [pid:129] Shakespeare Navigators URL: http://www.clicknotes.com
10. [pid:7] The Shakespeare Birthplace Trust URL: http://www.shakespeare.org.uk	10. [pid:10] Shakespeare: Internet Editions URL: http://web.uvic.ca/shakespeare
11. [pid:21] Shakespeare's Globe Theatre, Bankside, Southwark, London URL: http://www.shakespeares-globe.org	11. [pid:7] The Shakespeare Birthplace Trust URL: http://www.shakespeare.org.uk
12. [pid:9] Folger Shakespeare Library URL: http://www.folger.edu	12. [pid:24] Shakespeare Resource Center URL: http://www.bardweb.net
13. [pid:729] Surfing with the Bard URL: http://www.ulen.com/shakespeare	13. [pid:11] The Oregon Shakespeare Festival URL: http://www.orshakes.org
14. [pid:8] RSC - Royal Shakespeare Company URL: http://www.rsc.org.uk	14. [pid:3608] Yahoo! URL: http://www.yahoo.com
15. [pid:129] Shakespeare Navigators URL: http://www.clicknotes.com	15. [pid:34] Shakespeare & Company URL: http://www.shakespeare-company.org

Table B.31: *Top 15 results for query "shakespeare"*

HITS	**HubAvg**
1. [pid:2] ITTF URL: http://www.ittf.com	1. [pid:2] ITTF URL: http://www.ittf.com
2. [pid:1] Empty title field URL: http://www.usatt.org	2. [pid:1] Empty title field URL: http://www.usatt.org
3. [pid:9] ETTU - European Table Tennis Union URL: http://www.ettu.org	3. [pid:9] ETTU - European Table Tennis Union URL: http://www.ettu.org
4. [pid:3] ETTA: English Table Tennis Association URL: http://www.etta.co.uk	4. [pid:3] ETTA: English Table Tennis Association URL: http://www.etta.co.uk
5. [pid:5] Denis' Table Tennis / Ping-Pong World URL: http://www.tabletennis.gr	5. [pid:1017] Tibhar-HomePage URL: http://www.tibhar.de
6. [pid:246] What You Need to Know About^{tm} URL: http://www.about.com	6. [pid:20] Table Tennis and Ping Pong Accessories - Butterfly URL: http://www.butterflyonline.com
7. [pid:25] Welcome to Sweden Table Tennis URL: http://www.tabletennis.se	7. [pid:7] TABLE TENNIS CANADA TENNIS DE TABLE URL: http://www.ctta.ca
8. [pid:669] BS Table Tennis URL: http://bstt.cjb.net	8. [pid:25] Welcome to Sweden Table Tennis URL: http://www.tabletennis.se
9. [pid:8] Table Tennis / Ping-Pong — megaspin.net URL: http://www.megaspin.net	9. [pid:5] Denis' Table Tennis / Ping-Pong World URL: http://www.tabletennis.gr
10. [pid:7] TABLE TENNIS CANADA TENNIS DE TABLE URL: http://www.ctta.ca	10. [pid:246] What You Need to Know About^{tm} URL: http://www.about.com
11. [pid:20] Table Tennis and Ping Pong Accessories - Butterfly URL: http://www.butterflyonline.com	11. [pid:59] Aktuelle News - Deutscher Tischtennis-Bund URL: http://www.tischtennis.de
12. [pid:14] The Table Tennis Association Of Wales (TTAW) Homepage URL: http://www.btinternet.com/ ttaw	12. [pid:8] Table Tennis / Ping-Pong — megaspin.net URL: http://www.megaspin.net
13. [pid:88] BDTTA Home Page URL: http://www.bristoltt.demon.co.uk	13. [pid:143] Paddle Palace Table Tennis / Ping Pong equipment & supplies URL: http://www.paddlepalace.com
14. [pid:143] Paddle Palace Table Tennis / Ping Pong equipment & supplies URL: http://www.paddlepalace.com	14. [pid:36] Site Web FFTT URL: http://www.fftt.com
15. [pid:52] Welcome to tabletennisvideos.com URL: http://www.tabletennisvideos.com	15. [pid:669] BS Table Tennis URL: http://bstt.cjb.net
AT-Avg	**Norm (2)**
1. [pid:2] ITTF URL: http://www.ittf.com	1. [pid:2] ITTF URL: http://www.ittf.com
2. [pid:1] Empty title field URL: http://www.usatt.org	2. [pid:1] Empty title field URL: http://www.usatt.org
3. [pid:9] ETTU - European Table Tennis Union URL: http://www.ettu.org	3. [pid:9] ETTU - European Table Tennis Union URL: http://www.ettu.org
4. [pid:3] ETTA: English Table Tennis Association URL: http://www.etta.co.uk	4. [pid:3] ETTA: English Table Tennis Association URL: http://www.etta.co.uk
5. [pid:5] Denis' Table Tennis / Ping-Pong World URL: http://www.tabletennis.gr	5. [pid:5] Denis' Table Tennis / Ping-Pong World URL: http://www.tabletennis.gr
6. [pid:246] What You Need to Know About^{tm} URL: http://www.about.com	6. [pid:246] What You Need to Know About^{tm} URL: http://www.about.com
7. [pid:25] Welcome to Sweden Table Tennis URL: http://www.tabletennis.se	7. [pid:25] Welcome to Sweden Table Tennis URL: http://www.tabletennis.se
8. [pid:8] Table Tennis / Ping-Pong — megaspin.net URL: http://www.megaspin.net	8. [pid:8] Table Tennis / Ping-Pong — megaspin.net URL: http://www.megaspin.net
9. [pid:7] TABLE TENNIS CANADA TENNIS DE TABLE URL: http://www.ctta.ca	9. [pid:20] Table Tennis and Ping Pong Accessories - Butterfly URL: http://www.butterflyonline.com
10. [pid:20] Table Tennis and Ping Pong Accessories - Butterfly URL: http://www.butterflyonline.com	10. [pid:7] TABLE TENNIS CANADA TENNIS DE TABLE URL: http://www.ctta.ca
11. [pid:669] BS Table Tennis URL: http://bstt.cjb.net	11. [pid:669] BS Table Tennis URL: http://bstt.cjb.net
12. [pid:143] Paddle Palace Table Tennis / Ping Pong equipment & supplies URL: http://www.paddlepalace.com	12. [pid:143] Paddle Palace Table Tennis / Ping Pong equipment & supplies URL: http://www.paddlepalace.com
13. [pid:14] The Table Tennis Association Of Wales (TTAW) Homepage URL: http://www.btinternet.com/ ttaw	13. [pid:52] Welcome to tabletennisvideos.com URL: http://www.tabletennisvideos.com
14. [pid:52] Welcome to tabletennisvideos.com URL: http://www.tabletennisvideos.com	14. [pid:14] The Table Tennis Association Of Wales (TTAW) Homepage URL: http://www.btinternet.com/ ttaw
15. [pid:88] BDTTA Home Page URL: http://www.bristoltt.demon.co.uk	15. [pid:88] BDTTA Home Page URL: http://www.bristoltt.demon.co.uk
Max	**SALSA**
1. [pid:2] ITTF URL: http://www.ittf.com	1. [pid:2] ITTF URL: http://www.ittf.com
2. [pid:1] Empty title field URL: http://www.usatt.org	2. [pid:1] Empty title field URL: http://www.usatt.org
3. [pid:9] ETTU - European Table Tennis Union URL: http://www.ettu.org	3. [pid:9] ETTU - European Table Tennis Union URL: http://www.ettu.org
4. [pid:3] ETTA: English Table Tennis Association URL: http://www.etta.co.uk	4. [pid:20] Table Tennis and Ping Pong Accessories - Butterfly URL: http://www.butterflyonline.com
5. [pid:5] Denis' Table Tennis / Ping-Pong World URL: http://www.tabletennis.gr	5. [pid:3] ETTA: English Table Tennis Association URL: http://www.etta.co.uk
6. [pid:246] What You Need to Know About^{tm} URL: http://www.about.com	6. [pid:1017] Tibhar-HomePage URL: http://www.tibhar.de
7. [pid:25] Welcome to Sweden Table Tennis URL: http://www.tabletennis.se	7. [pid:669] BS Table Tennis URL: http://bstt.cjb.net
8. [pid:8] Table Tennis / Ping-Pong — megaspin.net URL: http://www.megaspin.net	8. [pid:5] Denis' Table Tennis / Ping-Pong World URL: http://www.tabletennis.gr
9. [pid:20] Table Tennis and Ping Pong Accessories - Butterfly URL: http://www.butterflyonline.com	9. [pid:7] TABLE TENNIS CANADA TENNIS DE TABLE URL: http://www.ctta.ca
10. [pid:7] TABLE TENNIS CANADA TENNIS DE TABLE URL: http://www.ctta.ca	10. [pid:25] Welcome to Sweden Table Tennis URL: http://www.tabletennis.se
11. [pid:669] BS Table Tennis URL: http://bstt.cjb.net	11. [pid:143] Paddle Palace Table Tennis / Ping Pong equipment & supplies URL: http://www.paddlepalace.com
12. [pid:143] Paddle Palace Table Tennis / Ping Pong equipment & supplies URL: http://www.paddlepalace.com	12. [pid:246] What You Need to Know About^{tm} URL: http://www.about.com
13. [pid:14] The Table Tennis Association Of Wales (TTAW) Homepage URL: http://www.btinternet.com/ ttaw	13. [pid:112] Newgy Table Tennis (Ping Pong) Robots & Equipment URL: http://www.newgy.com
14. [pid:52] Welcome to tabletennisvideos.com URL: http://www.tabletennisvideos.com	14. [pid:36] Site Web FFTT URL: http://www.fftt.com
15. [pid:88] BDTTA Home Page URL: http://www.bristoltt.demon.co.uk	15. [pid:88] BDTTA Home Page URL: http://www.bristoltt.demon.co.uk

Table B.32: *Top 15 results for query "table tennis"*

HITS	**HubAvg**
1. [pid:1490] Yahoo! URL: http://www.yahoo.com	1. [pid:1490] Yahoo! URL: http://www.yahoo.com
2. [pid:1574] Yahoo! Terms of Service URL: http://docs.yahoo.com/info/terms	2. [pid:1574] Yahoo! Terms of Service URL: http://docs.yahoo.com/info/terms
3. [pid:61] Yahoo! Autos URL: http://autos.yahoo.com	3. [pid:1562] Yahoo! Autos Sell Your Car URL: http://classifieds.autos.yahoo.com/class/submit/start.html?refsrc
4. [pid:1491] Yahoo! - Autos URL: http://help.yahoo.com/help/autos	4. [pid:1565] Sign in - Yahoo! Companion URL: http://us.edit.companion.yahoo.com/config/getbutton_companion?.bt
5. [pid:1562] Yahoo! Autos Sell Your Car URL: http://classifieds.autos.yahoo.com/class/submit/start.html?refsrc	5. [pid:61] Yahoo! Autos URL: http://autos.yahoo.com
6. [pid:1565] Sign in - Yahoo! Companion URL: http://us.edit.companion.yahoo.com/config/getbutton_companion?.bt	6. [pid:1491] Yahoo! - Autos URL: http://help.yahoo.com/help/autos
7. [pid:1572] Yahoo! Privacy URL: http://privacy.yahoo.com	7. [pid:1572] Yahoo! Privacy URL: http://privacy.yahoo.com
8. [pid:1566] Yahoo! Classifieds URL: http://classifieds.yahoo.com	8. [pid:1573] Yahoo! Media Relations URL: http://docs.yahoo.com/info/copyright/copyright.html
9. [pid:1571] Yahoo! Shopping URL: http://shopping.yahoo.com	9. [pid:1566] Yahoo! Classifieds URL: http://classifieds.yahoo.com
10. [pid:1573] Yahoo! Media Relations URL: http://docs.yahoo.com/info/copyright/copyright.html	10. [pid:1571] Yahoo! Shopping URL: http://shopping.yahoo.com
11. [pid:1567] HotJobs, a Yahoo! Company - the premier job search engine on the URL: http://careers.yahoo.com	11. [pid:1567] HotJobs, a Yahoo! Company - the premier job search engine on the URL: http://careers.yahoo.com
12. [pid:1568] Discover great singles near you at Yahoo! Personals. You'll love URL: http://personals.yahoo.com	12. [pid:1568] Discover great singles near you at Yahoo! Personals. You'll love URL: http://personals.yahoo.com
13. [pid:1569] Yahoo! Pets URL: http://pets.yahoo.com	13. [pid:1569] Yahoo! Pets URL: http://pets.yahoo.com
14. [pid:1570] Yahoo! Auctions: Auctions URL: http://auctions.yahoo.com	14. [pid:1570] Yahoo! Auctions: Auctions URL: http://auctions.yahoo.com
15. [pid:1559] Yahoo! Autos - Research URL: http://autos.yahoo.com/research.html	15. [pid:1559] Yahoo! Autos - Research URL: http://autos.yahoo.com/research.html

AT-Avg	**Norm (2)**
1. [pid:1490] Yahoo! URL: http://www.yahoo.com	1. [pid:1490] Yahoo! URL: http://www.yahoo.com
2. [pid:1574] Yahoo! Terms of Service URL: http://docs.yahoo.com/info/terms	2. [pid:1562] Yahoo! Autos Sell Your Car URL: http://classifieds.autos.yahoo.com/class/submit/start.html?refsrc
3. [pid:61] Yahoo! Autos URL: http://autos.yahoo.com	3. [pid:1565] Sign in - Yahoo! Companion URL: http://us.edit.companion.yahoo.com/config/getbutton_companion?.bt
4. [pid:1562] Yahoo! Autos Sell Your Car URL: http://classifieds.autos.yahoo.com/class/submit/start.html?refsrc	4. [pid:61] Yahoo! Autos URL: http://autos.yahoo.com
5. [pid:1565] Sign in - Yahoo! Companion URL: http://us.edit.companion.yahoo.com/config/getbutton_companion?.bt	5. [pid:1574] Yahoo! Terms of Service URL: http://docs.yahoo.com/info/terms
6. [pid:1491] Yahoo! - Autos URL: http://help.yahoo.com/help/autos	6. [pid:1491] Yahoo! - Autos URL: http://help.yahoo.com/help/autos
7. [pid:1572] Yahoo! Privacy URL: http://privacy.yahoo.com	7. [pid:1572] Yahoo! Privacy URL: http://privacy.yahoo.com
8. [pid:1573] Yahoo! Media Relations URL: http://docs.yahoo.com/info/copyright/copyright.html	8. [pid:1573] Yahoo! Media Relations URL: http://docs.yahoo.com/info/copyright/copyright.html
9. [pid:1566] Yahoo! Classifieds URL: http://classifieds.yahoo.com	9. [pid:1566] Yahoo! Classifieds URL: http://classifieds.yahoo.com
10. [pid:1571] Yahoo! Shopping URL: http://shopping.yahoo.com	10. [pid:1571] Yahoo! Shopping URL: http://shopping.yahoo.com
11. [pid:1567] HotJobs, a Yahoo! Company - the premier job search engine on the URL: http://careers.yahoo.com	11. [pid:1567] HotJobs, a Yahoo! Company - the premier job search engine on the URL: http://careers.yahoo.com
12. [pid:1568] Discover great singles near you at Yahoo! Personals. You'll love URL: http://personals.yahoo.com	12. [pid:1568] Discover great singles near you at Yahoo! Personals. You'll love URL: http://personals.yahoo.com
13. [pid:1569] Yahoo! Pets URL: http://pets.yahoo.com	13. [pid:1569] Yahoo! Pets URL: http://pets.yahoo.com
14. [pid:1570] Yahoo! Auctions: Auctions URL: http://auctions.yahoo.com	14. [pid:1570] Yahoo! Auctions: Auctions URL: http://auctions.yahoo.com
15. [pid:1559] Yahoo! Autos - Research URL: http://autos.yahoo.com/research.html	15. [pid:1559] Yahoo! Autos - Research URL: http://autos.yahoo.com/research.html

Max	**SALSA**
1. [pid:1720] MCSCCpix URL: http://www.mcsccpix.homestead.com	1. [pid:1720] MCSCCpix URL: http://www.mcsccpix.homestead.com
2. [pid:2244] Awesome Dodge Chargers for sale immediately! Mopar muscle! URL: http://martinpacker.com	2. [pid:2244] Awesome Dodge Chargers for sale immediately! Mopar muscle! URL: http://martinpacker.com
3. [pid:2832] VINTAGE POSTCARDS & EPHEMERA URL: http://vintagepostcards1.tripod.com	3. [pid:2832] VINTAGE POSTCARDS & EPHEMERA URL: http://vintagepostcards1.tripod.com
4. [pid:868] Home,die-cast-models, vintage die cast, die cast merchandise, vin URL: http://www.vintagediecast.com	4. [pid:868] Home,die-cast-models, vintage die cast, die cast merchandise, vin URL: http://www.vintagediecast.com
5. [pid:1901] F1 is Web F1 - Formula One News and more... URL: http://www.webf1.net	5. [pid:759] Classic Car Classifieds from Hemmings Motor News URL: http://www.hemmings.com
6. [pid:759] Classic Car Classifieds from Hemmings Motor News URL: http://www.hemmings.com	6. [pid:1490] Yahoo! URL: http://www.yahoo.com
7. [pid:1689] infoclassic - auto d'epoca - classic cars - voitures collection URL: http://www.infoclassic.net	7. [pid:1901] F1 is Web F1 - Formula One News and more... URL: http://www.webf1.net
8. [pid:2265] Welcome to Barn Hill Minis URL: http://www.barnhillminisusa.com	8. [pid:5] Classic Car - ClassicCar.com URL: http://www.classicar.com
9. [pid:1712] . . . _BuyDomains.com_. . . - Discount domain registration, URL: http://www.vintageracing.net	9. [pid:12] AutoWeek - Premier Source for Auto News, Reviews, Motorsports Inf URL: http://www.autoweek.com
10. [pid:1755] Pacific Coast Alfa Romeo Owners Assc. URL: http://alfaowners.cjb.net	10. [pid:1562] Yahoo! Autos Sell Your Car URL: http://classifieds.autos.yahoo.com/class/submit/start.html?refsrc
11. [pid:1372] Owens Export Services, Inc. URL: http://www.militaryjeep.com	11. [pid:1565] Sign in - Yahoo! Companion URL: http://us.edit.companion.yahoo.com/config/getbutton_companion?.bt
12. [pid:50] Classic Cars at Fossil Cars URL: http://www.fossilcars.com	12. [pid:13] Vintage Cars Classifieds URL: http://adlistings.vintagecars.about.com
13. [pid:1718] Shared Top Border URL: http://www.calgaryvintageracing.com	13. [pid:393] Crain Communications, Inc. URL: http://www.crain.com
14. [pid:13] Vintage Cars Classifieds URL: http://adlistings.vintagecars.about.com	14. [pid:50] Classic Cars at Fossil Cars URL: http://www.fossilcars.com
15. [pid:5] Classic Car - ClassicCar.com URL: http://www.classicar.com	15. [pid:61] Yahoo! Autos URL: http://autos.yahoo.com

Table B.33: *Top 15 results for query "vintage cars"*

HITS	**HubAvg**
1. [pid:6] NOAA - National Weather Service URL: http://www.nws.noaa.gov	1. [pid:9] NOAA Home Page URL: http://www.noaa.gov
2. [pid:9] NOAA Home Page URL: http://www.noaa.gov	2. [pid:6] NOAA - National Weather Service URL: http://www.nws.noaa.gov
3. [pid:1090] NOAA - National Weather Service - Forecast Products Disclaimer URL: http://www.nws.noaa.gov/disclaimer.html	3. [pid:1090] NOAA - National Weather Service - Forecast Products Disclaimer URL: http://www.nws.noaa.gov/disclaimer.html
4. [pid:138] National Hurricane Center / Tropical Prediction Center URL: http://www.nhc.noaa.gov	4. [pid:1858] NOAA Home Page - Privacy & Security Notice URL: http://www.noaa.gov/privacy.html
5. [pid:385] Climate Prediction Center URL: http://www.cpc.ncep.noaa.gov	5. [pid:385] Climate Prediction Center URL: http://www.cpc.ncep.noaa.gov
6. [pid:1] weather.com URL: http://www.weather.com	6. [pid:60] NWS page URL: http://www.wrh.noaa.gov/wrhq/nwspage.html
7. [pid:1858] NOAA Home Page - Privacy & Security Notice URL: http://www.noaa.gov/privacy.html	7. [pid:138] National Hurricane Center / Tropical Prediction Center URL: http://www.nhc.noaa.gov
8. [pid:5] Intellicast - Weather For Active Lives URL: http://www.intellicast.com	8. [pid:391] Department of Commerce Home Page URL: http://www.doc.gov
9. [pid:60] NWS page URL: http://www.wrh.noaa.gov/wrhq/nwspage.html	9. [pid:433] NOAA - National Weather Service - Public Affairs URL: http://www.nws.noaa.gov/pa
10. [pid:3] Weather Underground: Welcome to The Weather Underground URL: http://www.wunderground.com	10. [pid:1091] NOAA - National Weather Service - Privacy Notice URL: http://www.nws.noaa.gov/notice.html
11. [pid:10] UM Weather URL: http://cirrus.sprl.umich.edu/wxnet	11. [pid:434] National Weather Service Doppler Radars URL: http://weather.noaa.gov/radar/mosaic/DS.p19r0/ar.us.conus.shtml
12. [pid:433] NOAA - National Weather Service - Public Affairs URL: http://www.nws.noaa.gov/pa	12. [pid:1] weather.com URL: http://www.weather.com
13. [pid:187] GEOSTATIONARY SATELLITE SERVER URL: http://www.goes.noaa.gov	13. [pid:187] GEOSTATIONARY SATELLITE SERVER URL: http://www.goes.noaa.gov
14. [pid:434] National Weather Service Doppler Radars URL: http://weather.noaa.gov/radar/mosaic/DS.p19r0/ar.us.conus.shtml	14. [pid:6463] NOAA GOES East DATA - EASTERN US IR URL: http://www.goes.noaa.gov/ECIR4.html
15. [pid:1927] NCDC: * National Climatic Data Center (NCDC) * URL: http://www.ncdc.noaa.gov	15. [pid:2569] Redirect URL: http://205.156.54.206/pa
AT-Avg	**Norm (2)**
1. [pid:9] NOAA Home Page URL: http://www.noaa.gov	1. [pid:9] NOAA Home Page URL: http://www.noaa.gov
2. [pid:6] NOAA - National Weather Service URL: http://www.nws.noaa.gov	2. [pid:6] NOAA - National Weather Service URL: http://www.nws.noaa.gov
3. [pid:1090] NOAA - National Weather Service - Forecast Products Disclaimer URL: http://www.nws.noaa.gov/disclaimer.html	3. [pid:1090] NOAA - National Weather Service - Forecast Products Disclaimer URL: http://www.nws.noaa.gov/disclaimer.html
4. [pid:1858] NOAA Home Page - Privacy & Security Notice URL: http://www.noaa.gov/privacy.html	4. [pid:138] National Hurricane Center / Tropical Prediction Center URL: http://www.nhc.noaa.gov
5. [pid:385] Climate Prediction Center URL: http://www.cpc.ncep.noaa.gov	5. [pid:1858] NOAA Home Page - Privacy & Security Notice URL: http://www.noaa.gov/privacy.html
6. [pid:138] National Hurricane Center / Tropical Prediction Center URL: http://www.nhc.noaa.gov	6. [pid:385] Climate Prediction Center URL: http://www.cpc.ncep.noaa.gov
7. [pid:1] weather.com URL: http://www.weather.com	7. [pid:1] weather.com URL: http://www.weather.com
8. [pid:60] NWS page URL: http://www.wrh.noaa.gov/wrhq/nwspage.html	8. [pid:60] NWS page URL: http://www.wrh.noaa.gov/wrhq/nwspage.html
9. [pid:433] NOAA - National Weather Service - Public Affairs URL: http://www.nws.noaa.gov/pa	9. [pid:5] Intellicast - Weather For Active Lives URL: http://www.intellicast.com
10. [pid:5] Intellicast - Weather For Active Lives URL: http://www.intellicast.com	10. [pid:433] NOAA - National Weather Service - Public Affairs URL: http://www.nws.noaa.gov/pa
11. [pid:434] National Weather Service Doppler Radars URL: http://weather.noaa.gov/radar/mosaic/DS.p19r0/ar.us.conus.shtml	11. [pid:3] Weather Underground: Welcome to The Weather Underground URL: http://www.wunderground.com
12. [pid:187] GEOSTATIONARY SATELLITE SERVER URL: http://www.goes.noaa.gov	12. [pid:187] GEOSTATIONARY SATELLITE SERVER URL: http://www.goes.noaa.gov
13. [pid:3] Weather Underground: Welcome to The Weather Underground URL: http://www.wunderground.com	13. [pid:434] National Weather Service Doppler Radars URL: http://weather.noaa.gov/radar/mosaic/DS.p19r0/ar.us.conus.shtml
14. [pid:10] UM Weather URL: http://cirrus.sprl.umich.edu/wxnet	14. [pid:10] UM Weather URL: http://cirrus.sprl.umich.edu/wxnet
15. [pid:1091] NOAA - National Weather Service - Privacy Notice URL: http://www.nws.noaa.gov/notice.html	15. [pid:1091] NOAA - National Weather Service - Privacy Notice URL: http://www.nws.noaa.gov/notice.html
Max	**SALSA**
1. [pid:9] NOAA Home Page URL: http://www.noaa.gov	1. [pid:9] NOAA Home Page URL: http://www.noaa.gov
2. [pid:6] NOAA - National Weather Service URL: http://www.nws.noaa.gov	2. [pid:6] NOAA - National Weather Service URL: http://www.nws.noaa.gov
3. [pid:1090] NOAA - National Weather Service - Forecast Products Disclaimer URL: http://www.nws.noaa.gov/disclaimer.html	3. [pid:1] weather.com URL: http://www.weather.com
4. [pid:138] National Hurricane Center / Tropical Prediction Center URL: http://www.nhc.noaa.gov	4. [pid:138] National Hurricane Center / Tropical Prediction Center URL: http://www.nhc.noaa.gov
5. [pid:1] weather.com URL: http://www.weather.com	5. [pid:1090] NOAA - National Weather Service - Forecast Products Disclaimer URL: http://www.nws.noaa.gov/disclaimer.html
6. [pid:385] Climate Prediction Center URL: http://www.cpc.ncep.noaa.gov	6. [pid:5] Intellicast - Weather For Active Lives URL: http://www.intellicast.com
7. [pid:1858] NOAA Home Page - Privacy & Security Notice URL: http://www.noaa.gov/privacy.html	7. [pid:3] Weather Underground: Welcome to The Weather Underground URL: http://www.wunderground.com
8. [pid:5] Intellicast - Weather For Active Lives URL: http://www.intellicast.com	8. [pid:385] Climate Prediction Center URL: http://www.cpc.ncep.noaa.gov
9. [pid:60] NWS page URL: http://www.wrh.noaa.gov/wrhq/nwspage.html	9. [pid:1858] NOAA Home Page - Privacy & Security Notice URL: http://www.noaa.gov/privacy.html
10. [pid:3] Weather Underground: Welcome to The Weather Underground URL: http://www.wunderground.com	10. [pid:10] UM Weather URL: http://cirrus.sprl.umich.edu/wxnet
11. [pid:10] UM Weather URL: http://cirrus.sprl.umich.edu/wxnet	11. [pid:3544] Gannett Company, Inc. URL: http://www.gannett.com
12. [pid:433] NOAA - National Weather Service - Public Affairs URL: http://www.nws.noaa.gov/pa	12. [pid:38] CNN.com URL: http://www.cnn.com
13. [pid:187] GEOSTATIONARY SATELLITE SERVER URL: http://www.goes.noaa.gov	13. [pid:250] Google URL: http://www.google.com
14. [pid:434] National Weather Service Doppler Radars URL: http://weather.noaa.gov/radar/mosaic/DS.p19r0/ar.us.conus.shtml	14. [pid:60] NWS page URL: http://www.wrh.noaa.gov/wrhq/nwspage.html
15. [pid:1927] NCDC: * National Climatic Data Center (NCDC) * URL: http://www.ncdc.noaa.gov	15. [pid:187] GEOSTATIONARY SATELLITE SERVER URL: http://www.goes.noaa.gov

Table B.34: *Top 15 results for query "weather"*