

SelARC with Content Mining

TO 7494



Developed by:

MUNAZZAH JABEEN
(282- FAS/MSCS)

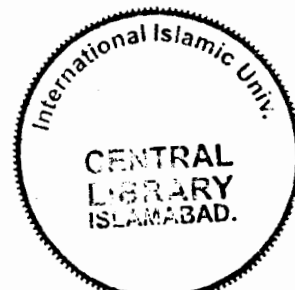
Supervised by:

Muhammad Imran Saeed

Co-supervised by:

Zubeda Khanum

Department of Computer Science
Faculty of Basic and Applied Sciences
International Islamic University Islamabad
2010



Accession No. TH 7494

MS
006.312
MUS

- 1 - Data mining
- 2 - web databases.

J. E.
J
2-3-11

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

In the Name of Allah The Most Beneficent

The Most Merciful

Department of Computer Science
International Islamic University, Islamabad

Date: 02-08-2010

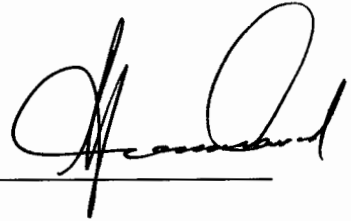
FINAL APPROVAL

It is certified that we have read the project titled "SelARC with Content Mining" submitted by **Miss Munazzah Jabeen Reg. No. 282-FAS/MSCS**. It is our judgment that this project is of sufficient standard to warrant its acceptance by International Islamic University, Islamabad, for the degree MS in Computer Science.

COMMITTEE

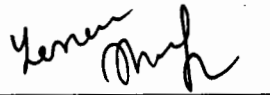
External Examiner:

Dr. Muhammad Younus Javed
HOD, Computer Engineering, College of EME (NUST)
Peshawar Road, Rawalpindi.



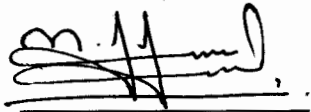
Internal Examiner:

Mrs. Zareen Sharf
Department of Computer Science,
International Islamic University, Islamabad

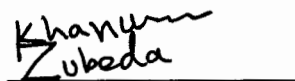


Supervisors:

Imran Saeed
Department of Computer Science,
International Islamic University Islamabad



Zubeda Khanum
Govt. Degree College, B-Block Rawalpindi.



**A dissertation submitted to the
Department of Computer Science,
International Islamic University, Islamabad
as a partial fulfillment of the requirements
for the award of the degree of
MS in Computer Science**

To My Loving Parents

“My Lord have Mercy on them (Parents) both as they did care for me when
I was little”

(AL-QURAN 17:24)

DECLARATION

I, hereby declare that “SelARC with Content Mining” software, neither as a whole nor as a part thereof has been copied out from any source. I have developed this software and the accompanied report entirely on the basis of my personal efforts made under the sincere guidance of my supervisor. No portion of the work presented in this report has been submitted in support of any application for any other degree or qualification of this or any other university or institution of learning.

Munazzah Jabeen
282-FAS/MSCS

ACKNOWLEDGEMENTS

I bow my head, in deep gratitude, before THE ALMIGHTY ALLAH for Blessing me with the wisdom and the capability and granting me the strength to accomplish this project.

I am very thankful to my kind, dynamic and able supervisors, Muhammad Imran Saeed and Zubeda Kanum for taking the time out of their ever busy schedule in providing me the guidance and direction. They always gave me their all out help and assistance right from the tricky and visionary stage of laying down the conceptual framework to the far more exacting stage of actual execution of this project successfully.

I adore and pray for all my teachers particularly, Dr. Abid Khan for guiding and assisting me in acquiring right type of knowledge and giving me hopes when I became hopeless. I am extremely grateful to my friends particularly Khadija al Madni, Nusrat and Shafaq. I shall ever remain grateful to all of them for their kind help.

I am thankful to my beloved parents for helping me grow mentally right from the day one in school besides catering to all my needs, affording every facility, consoling me at times when I felt dejected and fatigued, inspiring me when I started feeling that the project was getting beyond my control and grasp and praying day in and day out for my success.

I indeed also very appreciative of the contribution of my brothers, Arif Sohail, Imran, Irfan and Usman for their continuous solace and forbearance, which give me

knowledge and strength to complete my project successfully.

PROJECT IN BRIEF

Project Title:	SelARC with Content Mining
Objective:	TO blend ARC, SelHITS and Content Mining algorithms for topic distillation.
Undertaken By:	Munazzah Jabeen
Supervised By:	Muhammad Imran Saeed
Co- Supervised By:	Zubeda Khanum
Starting Date:	October 2008
Completion Date:	December 2009
Operating Systems:	Windows XP
Tool Used:	Matlab, Microsoft Access, C#.net
System Used:	Intel Pentium IV, 2.25 MHz processor

ABSTRACT

To search the most relevant and high quality web pages for some query topic on the World Wide Web is an important and challenging task. Many search systems are already developed and are being used by the people to get their required informative pages. But these commercial search engines are still lacking in providing the most relevant web pages to the users. As not all the web pages on the web are honest about their contents further there are billions of web pages containing text, images and other multimedia information, contained by the rapidly growing World Wide Web.

In this thesis, I have worked on a blend of Connectivity Analysis with the Content Analysis methods and presenting a system "SelARC with Content Mining" performs connectivity analysis as well as content analysis of a focused sub graph of the World Wide Web, to solve the problems of topic drift, topic contamination and extra time consumption. The presented system has improved results by applying the selective expansion on the root set and base set. Further I have implemented three regulatory content and connectivity algorithms on the augment set to prune the web pages. The important feature of this work is the way of calculating the hub and authority values, selective expansion up to two neighborhood and content analysis for calculating the relevancy weight. The results have shown that the above mentioned problems have overcome and users can get the most appropriate pages for their queries.

Table of Contents

<i>Chapter No.</i>	<i>Contents</i>	<i>Page No.</i>
1.	Introduction.....	1
	1.1 Data Mining.....	1
	1.1.1 Classification.....	2
	1.1.2 Clustering.....	2
	1.1.3 Association Rules.....	3
	1.1.4 Sequential Patterns.....	3
	1.2 Web Mining.....	3
	1.2.1 Web Content Mining.....	4
	1.2.2 Web Structure Mining.....	5
	1.2.3 Web Usage Mining.....	6
	1.3 Information Retrieval.....	7
	1.4 Topic Distillation.....	8
	1.5 World Wide Web and Web Browser.....	9
	1.6 Search Engine.....	9
	1.7 Connectivity Analysis.....	10
	1.7.1 Hyperlinks.....	10
	1.7.2 Significance of Hyperlinks.....	11
	1.8 Web as a Directed Graph.....	12
	1.9 Existing Link Analysis Algorithms	13
	1.9.1 Page rank.....	13
	1.9.2 Hypertext Induced Topic Search (HITS).....	14
	1.9.3 SALSA	14
	1.9.4 Trust Rank.....	15
	1.10 Content Analysis.....	15
	1.10.1 Conceptual Analysis	15
	1.10.2 Relational Analysis.....	16
2.	The Literature Survey.....	17
	2.1 Hypertext Induced Topic Search.....	17
	2.2 Page Rank, HITS and Unified Framework for Link Analysis.....	19
	2.3 Automatic Resource Compilation (ARC)	20
	2.4 Improved Algorithm for Topic Distillation in a Hyperlinked Environment.....	22

<i>Chapter No.</i>	<i>Contents</i>	<i>Page No.</i>
	2.4.1 Mutually Reinforcing Relationship between Hosts	22
	2.4.2 Automatically Generated Links.....	22
	2.4.3 Non-relevant Nodes.....	22
	2.5 Selective Hypertext Induced Topic Search.....	23
	2.6 Analysis of Previous Techniques	24
3.	Problem Statement	25
4.	Design.....	25
	4.1 ARC	27
	4.2 Content and Connectivity Analysis Algorithms.....	28
	4.3 SelHITS Algorithm.....	29
	4.4 System Architecture.....	30
5.	Implementation.....	34
	5.1 Tools and Technology.....	34
	5.1.1 MATLAB.....	34
	5.1.2 C#.net.....	35
	5.2 Important phases and Implementation functions.....	36
	5.2.1 Start set.....	36
	5.2.2 Candidate set.....	36
	5.2.3 Selective Base Set.....	36
	5.2.4 Aug Candidate Set.....	37
	5.2.5 Selective Augment Set.....	37
	5.2.6 Pruned Set.....	38
	5.2.7 Top hub and authority.....	38
	5.2.7.1 Top hub and authority through Medianr Algorithm	38
	5.2.7.2 Top hub and authority through Start Set Medianr algorithm ...	39
	5.2.7.3 Top hub and authority through Fraction of Maximum.....	39
6.	Testing and Results	40
	6.1 Testing Purpose.....	40
	6.2 Testing Principals.....	40
	6.3 Testing the Design	41
	6.4 Testing the Code.....	41

<i>Chapter No.</i>	<i>Contents</i>	<i>Page No.</i>
6.5	Testing Methods.....	41
6.5.1	Black Box Testing.....	42
6.5.2	White Box Testing.....	42
6.6	Testing of SelARC with Content mining.....	42
6.7	Top Authority Pages by medianr for query “windows”	43
6.8	Top Hub Pages by medianr for query “windows”	44
6.9	Top Authority Pages by Fractionr of Maximum for query “windows”	45
6.10	Top Hub Pages by Fractionr of Maximum for query “windows”	46
6.11	Top Authority Pages by start set median for query “windows”.....	47
6.12	Top Hub Pages by start set median for query “windows”	48
6.13	Analysis of Results	48
7.	Conclusion and Future Work.....	50
7.1	Conclusion.....	50
7.2	Future Work.....	51
	APPENDIX A – LIST OF ABBREVIATIONS	A-1
	APPENDIX B – SCREEN SHOTS	B-1
	References	

List of Figures

<i>Figure No.</i>	<i>Figure Name.</i>	<i>Page No.</i>
Figure 1.1	Taxonomy of web mining	4
Figure 1.2	Hyperlinked Pages	12
Figure 1.3	Directed Graph	13
Figure 2.1	How to get Root Set	18
Figure 2.2	Generating Base Set of Root Set	18
Figure 2.3	Hub Page Example (a), Authority Page Example (b)	19
Figure 2.4	The web page pk co-cites WebPages pi and pj (a) Pk is co-referenced by web pages pi and pj (b)	20
Figure 2.5	ARC	21
Figure 2.6	HITS	23
Figure 2.7	SelHITS	24
Figure 4.1	Automatic Resource Compilation	27
Figure 4.2	Content and Connectivity Algorithms	29
Figure 4.3	Selective Hypertext Induced Topic Search	30
Figure 4.4	SelARC with Content Mining	31
Figure 6.1	Top Authority Pages by medianr	43
Figure 6.2	Top Hub Pages by medianr	44
Figure 6.3	Top Authority Pages by Fractionr of Maximum	45
Figure 6.4	Top Hub Pages by Fractionr of Maximum	46
Figure 6.5	Top Authority Pages by start set medianr	47
Figure 6.6	Top Hub Pages by start set medianr	48

List of Tables

<i>Table No.</i>	<i>Table Name</i>	<i>Page No.</i>
------------------	-------------------	-----------------

Table 2.1	Comparison of existing Techniques	24
Table 6.1	Comparison of “SelARC with Content Mining” with other Techniques	49

1

Introduction



1. INTRODUCTION

This is really the age of clichéd Information. Information contents on the internet have grown extra ordinarily due to which internet has become the Global Information Infrastructure. This vast amount of information and the immense popularity of the Web have made the ordinary citizen to be a consumer of this information. Due to this ever expanding information the users have to face difficulties in finding and utilizing the information contents. Traditional web search engines do not seem to help the users much as hundreds and thousands of web pages are returned against a search query which requires the manual effort and time consumption by the user to find the information pieces one is looking for. That is why new tools and techniques are required to find the most relevant information against any user queries.

The data on the web is semi-structured as compared to the tabular data which is highly structured. Machine learning methods which are expected to operate on the structured data cannot be operated on the web data. The data on the web have many well defined structures but the order in which these structures will be appeared varies among the web documents. To develop more improved methods of applying machine learning techniques on the web data is very necessary. Exploiting all the features of the web like word contents, page structure, page URL, and link structures can give useful information while extracting important patterns. By assigning weights to these features different algorithms have been evolved. Still there are many deficiencies and new more improved techniques, creating server side and client side intelligent systems that can effectively mine for knowledge are needed.

1.1 Data mining

Data Mining can be defined as to find the useful patterns or hidden information from a large amount of data. It is also known as knowledge-discovery in databases (KDD), is the process of automatically searching large databases for meaningful

knowledge. With advances in the process of data collection and explosion in growth of data coupled with the advances in computing technologies really spruced up "Data Mining. Usually used in relation to analysis of data, data mining, like artificial intelligence, is an umbrella term and is used with different meanings in various contexts. The primary goals of data mining are prediction and description. Prediction involves using some variables or fields in the database to predict unknown or future values of other variables of interest, and description focuses on finding human-interpretable patterns describing the data. By using a variety of particular data-mining methods like, Classification, Clustering, Association Rules, Sequential Patterns, Regression, Summarization, Change and Deviation, Detection and many more the goals of data mining can be achieved . Information technology has evolved transaction and analytical systems separately but data mining links these two. The relationships and patterns in stored transaction data based on the user queries are analyzed by the data mining software. Many types of analytical software are available: statistical, machine learning, and neural networks. In general, data can be mined considering any of four types of relationships:

1.1.1 Classification

Classification is a data mining (machine learning) technique used to locate data instances in predetermined classes. A model is created automatically from a set of records that contain class labels. A function analyzes records that are already known to belong to a certain class, and creates a profile for a member of that class from the common characteristics of the records. Then the records that have not yet been classified can be classified using that model. It enables us to predict that if the new records belong to that particular class. Well known classification techniques are decision trees and neural networks.

1.1.2 Clustering

Clustering is a data mining technique in which data items are grouped into groups of similar data items (clusters), according to logical relationships without any advanced

knowledge of the group definition as in classification. In clustering the basic principle is to define a notion of similarity or dissimilarity between the data instances or you can say how similar the data items are and then partition them into k -clusters for example data can be mined to identify market segments or consumer likeness. From a machine learning perspective clusters correspond to hidden patterns so the search for clusters is unsupervised learning. K-means clustering and expectation maximization clustering are well known clustering techniques.

1.1.3 Association Rules

Association rules is an important data mining technique for finding frequent patterns, correlations, association among the data items in the database. Association rules are defined by considering two predefined factors Support and Confidence. The rule which satisfy the minimum support and confidence will be considered important. In other words the frequent patterns and their association which satisfy the constraint of minimum confidence are extracted from the database. The burger-drink example is an example of associative mining. Association rules are used in areas such as market and risk management, telecommunication networks, inventory control etc.

1.1.4 Sequential Patterns

Association rules defined on the dimension of time are considered as sequential patterns. The data having sequential nature like temporal data or in bioinformatics the amino-acid sequences are mined by defining finding sub-sequences that appear frequently and satisfy minimum support threshold. Sequential patterns highlight inter-transaction associations.

1.2 Web Mining

When data mining techniques are applied on the World Wide Web data it is called web mining. Web mining technologies have become very important research area. Initially two different approaches were taken to define web mining. One was “process-

centric view,” which defined web mining as a sequence of tasks (Etzioni 1996). The other was “data-centric view,” which defined web mining in terms of the types of web data that was being used in the mining process (Cooley, Srivastava, and Mobasher 1997). The data-centric view has become more acceptable, as is evident from the approach adopted in most research papers that have addressed the issue. We follow the data-centric view of web mining. The Web can be seen as a structure containing three main flavors of data: content, structure and usage. Web mining can be categorized into three following:

- Web content mining.
- Web structure mining.
- Web usage mining.

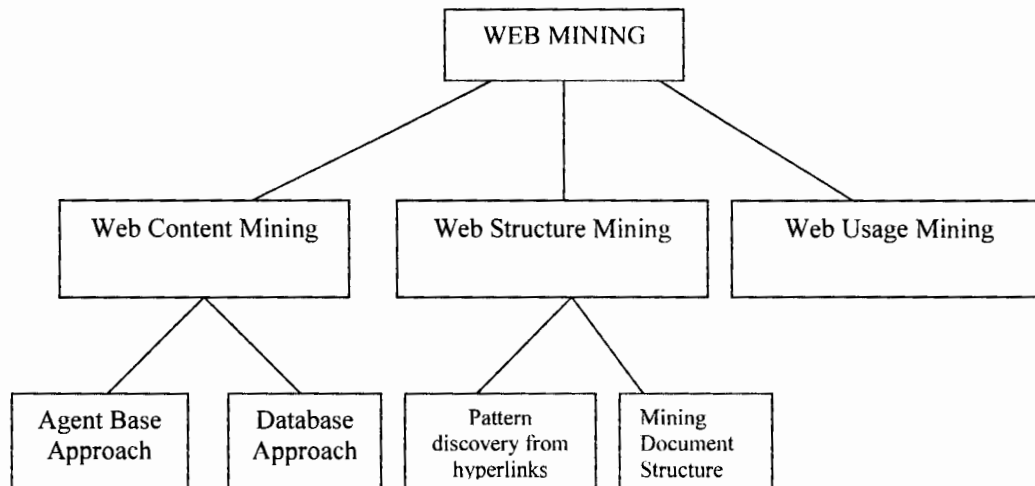


Figure 1.1: Taxonomy of web mining

1.2.1 Web Content Mining

There are many problems arising due to rapid expansion of the web and constant growth of information which causes increased difficulty to extract useful information.

Web content mining confronts this problem gathering explicit information from different web sites for its access and knowledge discovery.

Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It may consist of text, images, audio, video, or structured records such as lists and tables. Web content mining is different from text mining and data mining. It relates to text mining as much of the data on web is text but it is different because of the semi structure nature of the web but as we know text mining focuses on text which is unstructured. Web content mining is related to data mining as many data mining techniques can be applied in web content mining but it is different from data mining as it deals with semi structured data while in data mining structured data is dealt primarily. Web content mining essentially is an analog of data mining techniques for relational databases, since it is possible to find similar types of knowledge from the unstructured data residing in web documents.

1.2.2 Web Structure Mining

It is the process of discovering information from WWW organization and links between the reference web pages and referent web pages (hyperlinks at the inter-document level). In other words structural summary of the web site and web page is generated. In Web structure mining the graph theory is used which analyzes the node and connection structure of a web site. According to the type of web structural data, web structure mining is of two kinds. One is to extract patterns from the hyperlinks in the web pages. A hyperlink is a structural component that connects the web page to a different location. The other kind is mining the document structure. While analyzing the document structure it uses a tree like structure to analyze and describe the HTML and XML tags within the web page.

Hyperlinks provide additional information through the way in which different documents are connected to each other. We can view the web as a (directed) graph whose nodes are the documents and the edges are the hyperlinks between them. A model

underlying the link structures is discovered in Web structure mining which is based on the topology of the hyperlink with or without the link description. To categorize the Web pages and useful information such as similarity and relationships between Web sites can be generated using that model. Web pages can be filtered and ranked using important information contained in the link structure of web. Some new algorithms have been proposed that use link structure to search keywords and to identify web communities. So these algorithms perform better than Information Retrieval algorithms because they use more information than just the contents of the web pages. There are two major link-based search algorithms, HITS (Hypertext Induced Topic Search) and PageRank.

1.2.3 Web Usage Mining

To find the user navigation pattern by applying the data mining techniques on the web data is called Web Usage Mining. The path through which the user accesses the pages or visits any web site could be very important information. Most of the organizations now-a -days rely on the Internet and WWW to conduct business also collect and generate large volumes of data daily. Web servers usually generate and collect most of this information in access logs. Web usage mining is the discovery of user access patterns from web servers. Web usage data includes the pattern of usage of Web pages, such as IP addresses, page references, and the date and time of accesses and data that provides information about users of the Web site including registration data and customer profile information. This can help the organizations to determine cross marketing strategies, promotional campaigns, and life time value of customers. In order to create better structure web site so that the presence for the organization can be created affectively, the analysis of user registration data and the server access logs can provide very useful information. The organizations which use intranet technologies and sell their products on WWW, Web usage mining can provide help in more effective management of organizational infrastructure, workgroup communication and targeting ads to specific groups of users. For the discovery and analysis of patterns many techniques and systems have been emerged. Two main categories of such tools are Pattern Discovery Tools and Pattern Analysis Tools.

1.3 Information Retrieval

To search the web, relational databases for documents, information within documents and metadata about documents is called Information Retrieval. The terms text retrieval, data retrieval, information retrieval, document retrieval are overlapped but all these also have some specific technologies, theory and body of literature which make them different. Information retrieval system is based on Mathematics, Physics, library science, Information architecture, statistics, Information Science, Cognitive psychology and Computer science. To reduce the information overloading automated information retrieval systems are used now-a-days. There are many automated information retrieval systems provided by public libraries and universities to access books, documents and journals but the most popular are web search engine.

Most of the Information retrieval systems work well with finite and document controlled collections. In such finite collections documents are truthful about their contents and are self contained. For such collections relevancy of the document can be easily evaluated. Evaluation can be done in terms of recall and precision.

For World Wide Web the notion of recall has little meaning because it can not be measured as we cannot have complete picture of the web. Precision cannot be measured as user gives mostly short queries and the number of the documents containing query will be in thousands. Despite all this documents may not be true about their contents. Using hyperlinks, the documents or parts of the documents can be linked to each other and additional information can also be attached to the document. This supports to the manual browsing. However to find specific information or documents only hyperlinked browsing is not sufficient. Million of web pages are daily added to the web so there must be some retrieval mechanisms.

Generally when a user wants some specific information he uses either search service or browses. When a user gives a query to a search engine thousands of web pages are there to respond the query and few of them are returned by the search engine. These returned web pages may not satisfy the user's information needs even if they contain the

query terms. So besides the word matching some other ways and information have to be found to enhance the performance of IR systems.

1.4 Topic Distillation

To find the pages which are most relevant to the topic of the user query is called Topic Distillation. We can say if a broad topic is given the process of getting a small no of high quality web pages which represent the topic most, is called Topic distillation.

As web growth is increasing exponentially with the time and the number of end users cannot be a constant, web pages returned by search engines often contain exactly match with the query terms. Topic distillation does not satisfy the exact information needs of the users but it aims to return the web pages for the query topic and provides spectrum of information to the user.

An engine for distillation rather than search should return say thirty great pages in general about any broad topic, filtering out the large number of irrelevant pages and low-quality pages which would be appropriate responses to a more specific query.

Consider a query like Web structure mining. When the user is searching for web structure mining then it can be in several contexts such as web structure mining as a subject, books on web structure mining, conferences related to web structure mining, famous people in the web structure mining field, important research groups on web structure mining, companies providing tools for web structure mining etc. the broad interests of the user cannot be satisfied by exact query matches. Rather, the users have to search many times to find some aspect of web structure mining. e.g. "web structure mining conferences", "web structure mining tools". By topic distillation information about all aspects of web structure mining can be provided to the user with only performing one search. This is the main benefit of topic distillation over simple searching.

Different topic directories like Yahoo, Google get information about broad topics but considerable manual effort is required because topics are hard coded and arranged

hierarchically. To overcome this curb Topic Distillation can provide help. No manual effort is required when topic distillation is applied to answer any broad topic query. Chakrabarti et al. [3] have done experiments on constructing topic directories automatically.

1.5 World Wide Web and Web Browser

World Wide Web is a system that consists of a huge set of hypertext documents, videos, images and other resources which are interlinked by hyperlinks and can be accessed via the internet. The World Wide Web is commonly called as Web is one of the services accessible via the Internet, along with different other services like e-mail, file sharing, online gaming etc. To access the Web, Web browser is the Indispensable requirement. Web browser is a software that is used to view the web pages that may contain the text, videos, images, sounds etc and also to navigate between the web pages by using the hyperlinks.

To view a Web page on the World Wide Web the users have to type the URL into the web browser, or to follow a hyperlink to that page or resource. Then in order to fetch and display a web page a series of communication messages is initiated by the web browser behind the scenes. The URL of the page is resolved into the IP address using domain name system or DNS. This IP address is then used to locate the Web server. Then the browser requests the web server to send the page. First, the web browser requests the HTML text of the web page and parse it immediately then the requests for images and other files are made which make the parts of the page. When the required files are received by the web browser the page is then rendered onto the screen incorporating the images and other required resources so that the on-screen web page can be produced that the user will see.

1.6 Search Engine

Search engine is a tool to search information on the web and is a software program that searches a database for keywords and returns documents that are related to the query terms. The data presented by the search engines are in the form of lists called

hits and consists of web pages, images and other files. Some search engines also perform data mining. Search engines are usually general class of programs and operate algorithmically. Goggle, Alta vista, Yahoo, Excite are examples of search engine that enable the user to find the required information on the WWW and USENET newsgroups.

1.7 Connectivity Analysis

Connectivity analysis is an information retrieval technique. It identifies and utilizes the linkage information of the World Wide Web structure. Standard Information Retrieval techniques can not perform well due to the continuously increasing growth of the Web. On the other hand hyperlink connectivity analysis is used to identify the links between the documents so the connectivity analysis building blocks are increasing continuously and becoming more important.

1.7.1 Hyperlinks

A hyperlink is a navigational element in a web page which references to some other section of the same web page, some other web page or to some specific section of the other web page. Hyperlink is commonly called as link and can be a word, phrase or image that the user clicks and jumps automatically to the referred information on the web. Hyperlinks are found almost on all web pages. In IR Hyperlink Connectivity Analysis is done for two main purposes which are crawling and ranking.

All the links connecting the web pages are equivalent because the web does not give any priority to a link or a document more than any other. The connectivity patterns between web pages contain very important implicit information about the importance of the link. The links by the document or page author are concealed indications of human judgment. When a link is created it means that the referred information is recommended by the document author so the destination document gains some prestige. The documents containing clear and accurate information are pointed by many links as compared to the low quality documents. Although, there is no preference function attached to the link but the preference is implicit in the total links which point to a document. This implicit preference is produced by the group of web document authors.

Some mathematical techniques exist to get this preference information. There are two types of algorithms Page Rank (Brin & Page 1998) and HITS (Kleinberg 1998). These algorithms determine the quality of a web page by the number of web pages that link to it. These algorithms need much iteration to find the quality or “authority” of a web page. A web page will be of high quality if it is pointed to by many high quality web pages. It is equal to compute the eigenvectors if we represent the selected web as a matrix. In Page Rank algorithm the linkage matrix is used while in HITS the product of linkage matrix and its transpose is used. The two methods when combined with keyword search perform well and return much better quality web pages. The main disadvantage of these methods is the static nature as these can use only the links already exist and do not allow the web to adapt newly generated links.

1.7.2 Significance of Hyperlinks

The Hyperlinks are the fundamental part of the web structure. There is a general assumption that all the pages in the web are equally accessible. It is true that there is no formalized structure of web and no centralized organization to design web sites. Different web browsers interpret scripting and markup languages differently. It is a reality that some web sites have become more accessible than others. Search engines are used to index links, anchor text, keywords and paper content due to which the value of some web sites is regulated.

For example, Google, a very popular search engine uses links to determine the value and placement of a web site in search results. It indexes the links and ranks pages by interpreting the links between the pages in a way as a link from A to B as an endorsed of B by A. Page rank of a page is find in a way as if A has many links to it as compared to C, then a link from A to B will be of more worth than a link from C to B. Links are analyzed more correctly as compared to the page views or traffic and are measure of success and distributor of rank.

1.8 Web as a Directed Graph

Web can be represented as directed graph, the web pages as the nodes and the hyperlinks as the directed edges. This graph representation of the web can provide very useful information.

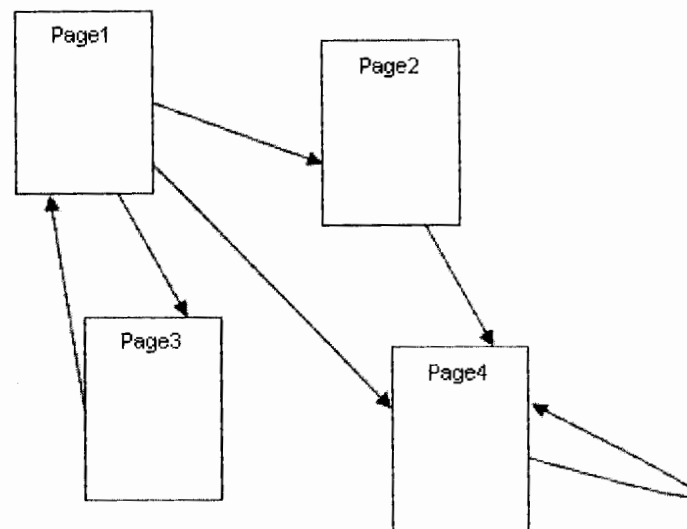


Figure 1.2: Hyperlinked pages

In figure 1.2 a hyperlinked structure is shown and in figure 1.3 this structure is converted into the directed graph. A node in the graph having a large number of links pointing to it is considered as a good quality web page having related information content to the query topic. Web page like this is called as Authority and the number of links directed to it is called in degree of that page. Similarly if a node having many directed edges going out from it to many authority web pages is also very useful resource. Such a web page is called as Hub and the large number of links going out from it is called out degree of a web page. For example in the context of literature citation, a review paper is a hub that has many links to the original research papers while an original research paper cited by many papers is called an authority.

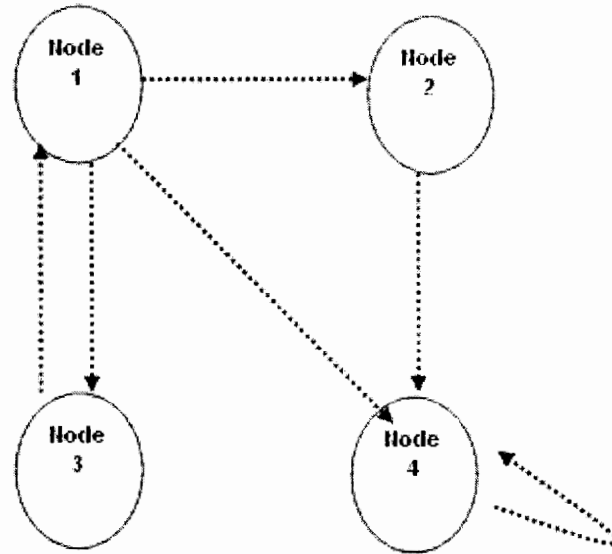


Figure 1.3: Directed graph

1.9 Existing Link Analysis Algorithms

The main algorithms for link analysis are:

- 1) Page Rank,
- 2) HITS
- 3) SALSA.
- 4) Trust Rank

HITS and Page Rank were proposed in 1998. SALSA was proposed in 2000. The main feature familiar to all of them is eigenvector computation.

1.9.1 Page Rank

It is a link analysis algorithm in which a numerical weight is assigned to each hyperlink to measure the relative importance. It can be applied to collection of entries with references and reciprocal quotations. The weight assigned to any element E is called page rank of that element denoted by $PR(E)$.

Now-a-days Google has become the most popular search engine because of its higher quality of search results. This superior quality of search results is due to Page Rank used to rank the web pages.

The Page Rank algorithm is based not only the number of incoming links but the fact that the document gains more importance as more the other documents link to it but the inbound links are not counted equally. A document is given high rank if other high ranked documents link to it.

1.9.2 Hypertext Induced Topic Search (HITS)

Hypertext Induced Topic Search (HITS) was developed by Jon Kleinberg. It's a link analysis algorithm. It calculates two values for each page to rate these are Authority and Hub. The authority value calculates the importance of the content of the web page while the hub value calculates the number of links it links to other pages. Authority and hub values can be defined in a mutual recursion as the hub value is the sum of the authority values of the pages it points to and the authority value is the sum of the hub values of the pages that point to that page. HITS like Brin's PageRank, is an iterative algorithm based on the linkage of the documents on the web. However it does have some major differences:

- It is executed at query time and not at indexing time.
- It is not commonly used by search engines.
- It computes two scores per document (hub and authority) as opposed to a single score.
- It is processed on a small subset of 'relevant' documents, not all documents as was the case with PageRank.

1.9.3 SALSA

SALSA stands for Stochastic Algorithm for Link Structure Analysis. The features of PageRank and HITS are combined in SALSA. Hub and authority values are calculated

per query as calculated by HITS. Markov chains are used to calculate these values as in PageRank.

1.9.4 Trust Rank

Trust Rank is a link analysis technique. For misleading the search engines many Web spam pages are created. These pages use various techniques to achieve higher-than-deserved rankings. While spam can be identified by human experts but to evaluate manually a large number of pages, is a difficult task.

Trust Rank method selects a small set of seed pages and evaluated by an expert. After identifying the reputable seed pages are manually, a crawl extending outward from the seed set seeks out trustworthy pages but as the documents are removed from the seed set Trust Rank's reliability diminishes.

1.10 Content Analysis

Content Analysis is the process of finding the presence of some words within the text. The meanings of these words or concepts are then analyzed and inferences are made to find the information underlying the relationships of these concepts or words. In the context of web mining content analysis can be used to extract the relevant information from the billions of web pages related to user query by analyzing the text for the presence of query terms in the documents. There are two general categories of content analysis.

1.10.1 Conceptual Analysis

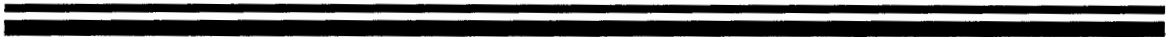
Conceptual analysis is to find the existence and frequency of the concepts and often represented by phrases or words in the text. For example say if your favorite poet writes mostly about the hunger. By conceptual analysis it can be determined that how many times the words like “hunger”, “starvation”, “famished” or “hungry” appear in that poet’s volume.

1.10.2 Relational Analysis

Relational analysis is one step further than the conceptual analysis. In this relationships among concepts are examined in the text. Returning the example of “hunger”, by relational analysis it identified that what other words are next to the words “hunger”, “hungry” or “famished” and then determine the meanings of these grouped words.

2

Literature Survey



2. LITERATURE SURVEY

2.1 Hypertext Induced Topic Search

Kleinberg [6] developed HITS algorithm for ranking and rating the web pages and uses two values *authority* and *hub* which can be defined in mutual recursion way. Authority value of a page is calculated as the sum of the hub values pointing it. A hub value of a page is calculated as the sum of the authority values. A more refined notion proposed by Kleinberg was that the importance of a web page is based on the search query performed by the user. To get the root set Kleinberg proposed to use text based search engine. Further he suggested augmenting the root set with the pages pointed to and pointed by the pages in the root set to form the base set.

Four basic components for Kleinberg approach are:

1. The amount of relevant information to some broad topic query is growing rapidly on the WWW. Due to that filtering the information resources is becoming more difficult for the users. The only way to deal this problem is to distill the broad topic query. Kleinberg proposed authoritative resources for this purpose.
2. The results produced by the Kleinberg approach are high in quality. The underlying domain is not restricted to the pages on a single web site or some focused set of pages.
3. A basic interface is required to any number of web search engines and the techniques to get enriched samples of pages to determine the quality that make sense globally.
4. Hub pages link densely to the authority pages in this approach. This equilibrium between hub and authority pages is that recurs in the context of variety of topics on web.

Some terms used by Kleinberg are as follows:

- **Root set**

A set of web pages obtained from some existing search engine against a user query is called root set. Figure 2.1 shows the procedure of getting the root set.

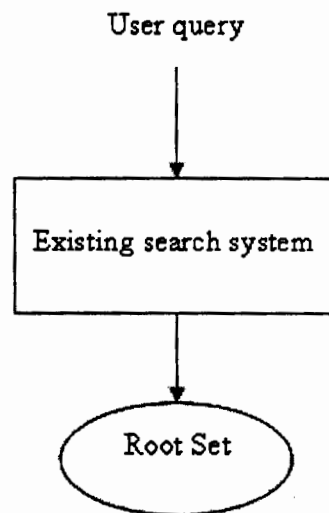


Figure 2.1: How to get Root Set

- **Base set**

The root set is expanded to 1-neighborhood to get the base set. All the pages pointed to and pointed by the pages in the root set are combined to form the base set. The figure 2.2 shows the method of obtaining the base set.

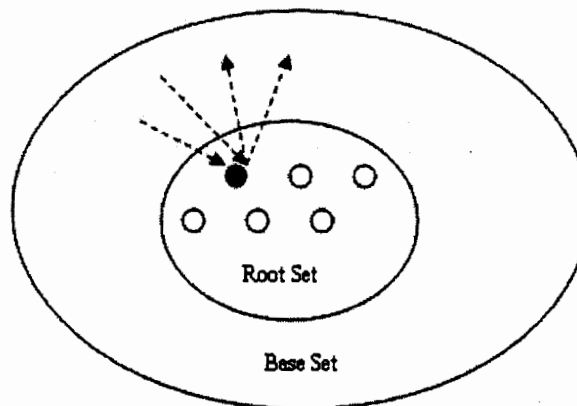


Figure2.2: Generating Base Set

- **Hub page**

A page that contains links to many informative pages or authoritative pages is called hub page. Figure 2.3 (a) shows the example of hub page.

- **Authority page**

A page that contains actual information on the query topic and is pointed to by many hub pages is called the authority page. Figure 2.3 (b) shows the example of the authority page.

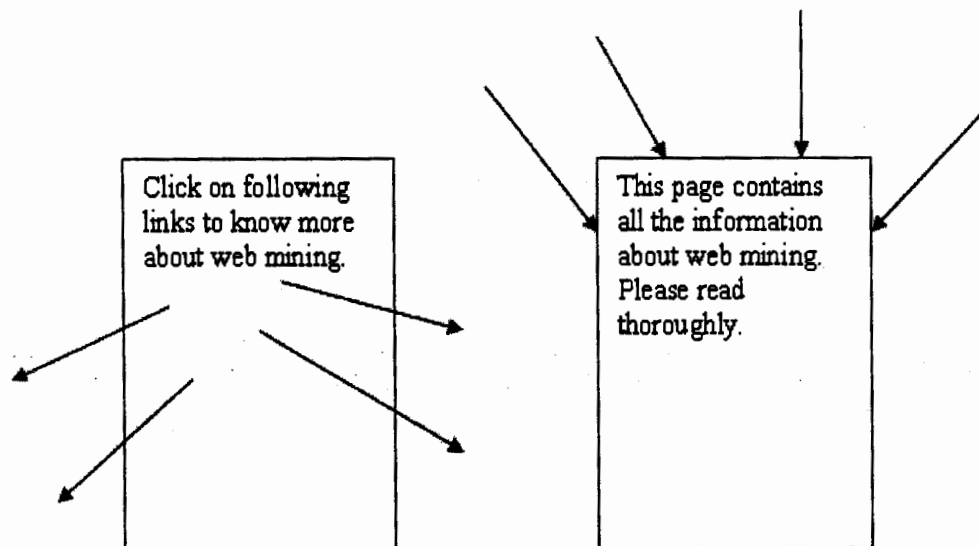


Figure 2.3 (a): Hub Page Example (b) Authority Page example

2.2 Page Rank, HITS and Unified Framework for Link Analysis

Parry *et al* [12] discusses Page Rank and Hyper Text Induced Search (HITS) with the problem of mutual reinforcement of hub and authorities. Concept of Co-citation and Co-Reference is discussed. If two pages co-cited by many other web pages are likely to be related in some sense. Similarly if two distinct web pages co-refer to several other web pages shows that there is some commonality [11].

The main feature of HITS by Kleinberg *et al* is the mutual reinforcement between hub and authority pages. On the other hand the main feature of the PageRank is normalization of the hyperlink weights.

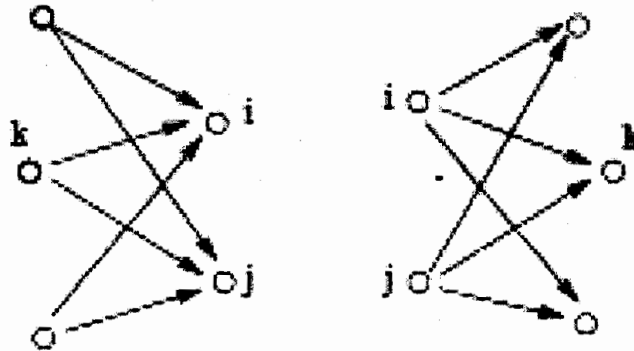


Figure 2.4 (a) the web page p_k co-cites WebPages p_i and p_j
 (b) p_k is co-referenced by web pages p_i and p_j

This paper also describes the concept of combining the mutual reinforcement with the hyperlink weight normalization into a unified framework and three new algorithms are introduced that perform normalized ranking in this new framework.

1. **INORM Rank:** inlinks are normalized using norm.
2. **ONORM Rank:** outlinks are normalized using norm.
3. **SNORM Rank:** inlinks and outlinks are normalized in symmetric fashion.

2.3 Automatic Resource Compilation (ARC)

Chakrabarti [3] presents an algorithm *Automatic Resource Compiler ARC*. ARC is also based on the extension of HITS algorithm with the anchor text analysis and performs 2-neighbourhood links analysis and weights the links. The goal of ARC is the automatic compilation of the resources and to list out the high quality informative pages.

This paper describes the use study and a comparison of the ARC compiler with the human compiled services. The users have to go on manual hierarchal taxonomy or

on the compilation assisted by the human like yahoo and info seeks. Such taxonomy provides help to find out the resource list of pages related to any broad topic.

This paper describes three phases of ARC.

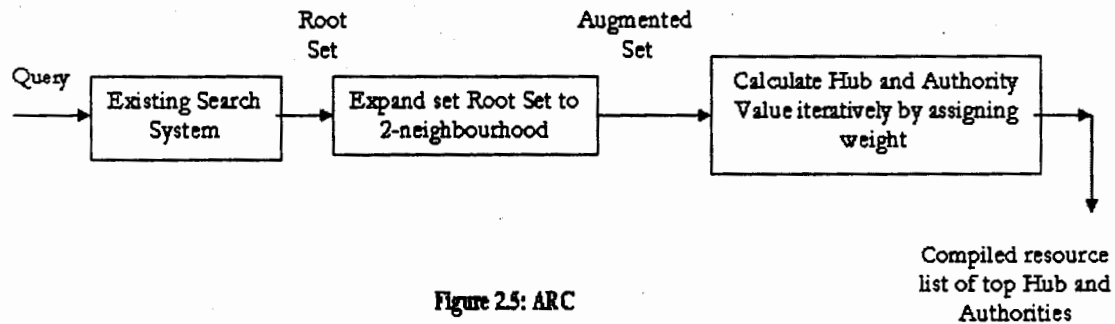


Figure 2.5: ARC

1. Search and Growth Phase

In this phase a set of 200 web pages are got and then augmented by the pages from 2_neighborhood.

2. Weighting phase

In this phase all the links are assigned positive numerical values as weights which show how related a page is, to the query topic.

3. Iteration and reporting phase

This phase iteratively computes **hub** and **authority** vectors. The values of the vector **a** contains the scores of each page as authority. The second vector **h** contains the entries for hub value of each page. Then a matrix **w** is calculated that contains entries like $w(p,q)$ computed as below, for every link of page **p** to page **q** and if there is no link from page **p** to page **q** the entry is 0. The vector **h** is set to 1 initially and the following steps are performed **k** times.

$$a = w h$$

$$h = Z a$$

Here **Z** is the matrix transpose of the vector **W**.

2.4 Improved Algorithm for Topic Distillation

Bharat et al [1] combined the connectivity analysis with the content analysis for finding the quality pages related to the query topic. He described three problems with the Hypertext Induced Topic Search (HITS) algorithm and also proposed solution and some algorithms to solve these problems. The main problems are as follows [6].

2.4.1 Mutual Reinforcing Relationships between Hosts.

There can be many possibilities that when more than one pages point to a single page on another host. This can increase the hub score of the pages on the first host and the authority score of the page on the other host. Similarly if a single page on one host points to many pages on the other host can cause the same problem. These situations give undue importance to the set of documents on a single host or give undue importance to the authority or opinion of the single author or organization of that host.

2.4.2 Automatically Generated Links

A link from one document to the other shows that the document pointed to by, on the other host has some prestige in the sight of the author of the first document but there are many links which are automatically generated by tools like web authoring and database conversion tools, cause problems while calculating the hub and authority scores. For example, Usenet News articles are converted to web pages by the Hyper News System and links to the Hyper News web site are automatically inserted. Such automatically generated links must not represent the opinion of humans as they are created by tools not by some human.

2.4.3 Non-relevant Nodes

For a query topic the neighborhood graph can have many irrelevant nodes. These irrelevant nodes can cause topic drift problem because the highly ranked hubs and

authorities may not have the information about the query topic. For example if a user enters the query “jaguar and car”, the computation can drift to the common topic “car” and the web pages returned, will be of the home pages different manufacturers of the cars. So some non-relevant pages can get high scores of authority and hubs.

In this paper Bharat et al also presented an algorithm named **imp** which is an improvement over HITS by Kleinberg et al [6], to solve the first problem. Fractional weights are assigned to the links so that no link can get undue weight. To solve the other two problems this paper also presents some other algorithms which combine the content analysis with the improved connectivity analysis algorithm.

2.5 Selective Hypertext Induced Topic Search

To answer the broad topic queries Mitra et al [7] developed SelHITS algorithm which is an improvement over HITS algorithm. This paper also describes the problems for example topic drift, topic contamination and extra time consumption. To solve these problems Mitra et al presented the selective expansion procedure by which root set is expanded selectively. This selective expansion ignores the irrelevant pages as it calculates hub and authority values of the root set pages and then only top hub and top authorities are expanded to form the base set. So by avoiding the non-relevant pages SelHITS distills the most important web pages for any broad topic query.

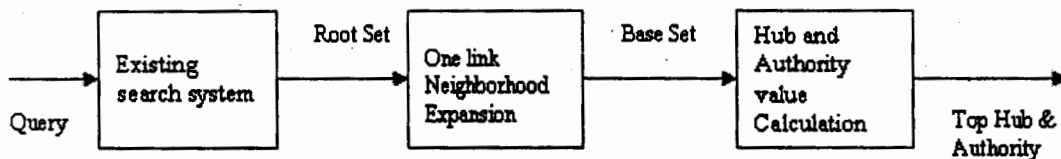


Figure 2.6 HITS

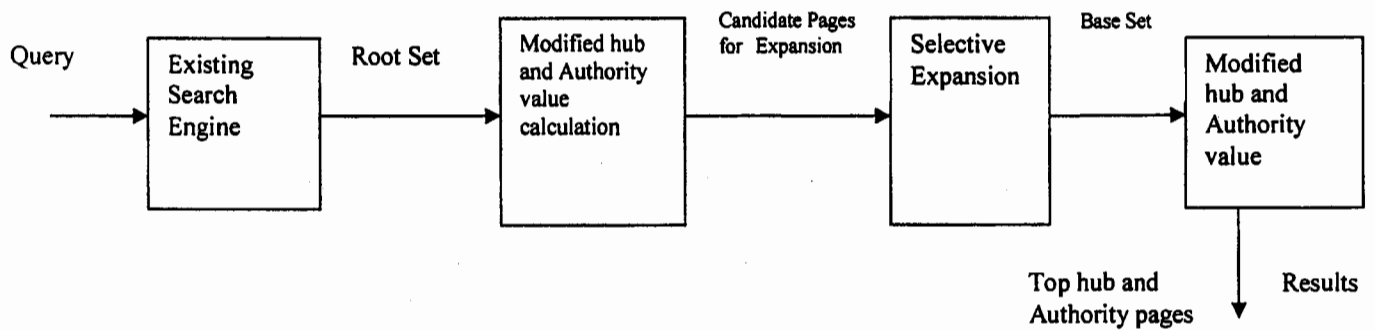


Figure 2.7 SelHITS

2.6 Analysis of exiting techniques

The comparison of existing content and connectivity analysis techniques against different parameters is shown below in table 2.1 which clearly shows the defficiencies of the prievous techniques.

Table 2.1: Comparison of existing techniques

Existing Techniques	One Link Expansion	2-Link Expansion	Selective Expansion	Anchor Text Mining	Content Analysis	Hub & Authority Calculation
Kleinberg et al. (1997)	✓	x	x	x	x	✓
Bharat et al. (1998)	✓	x	x	x	✓	✓
Chakrabarti et al. (2002)	✓	✓	x	✓	x	✓
Awekar et al. (2007)	✓	x	✓	x	x	✓

3

Problem Definition

3. PROBLEM STATEMENT

Searching the web can simply be considered as a process of querying and getting the most relevant pages through some search system. Here are some problems found in most of the search systems:

- **Topic Drift**

The problem of Topic drift is also occurred when there are more than one interpretations of a query term and the computation drifts to the more common topic.

- **Distilling the pure topic**

Sometimes users type ambiguous queries and search engine returns results from multiple topics, that causing topic contamination but the aim of topic distillation process is to deliver results for a single topic only. For example user fires the query "mouse". This query is ambiguous and has multiple meanings. Meanings of mouse can be device of computer or can be animal. So depending on meaning there will be different topics for the query.

- **Drawback of Anchor text mining**

The problem encountered in ARC algorithm is when Anchor text mining is performed on Augment set a large amount of information is lost for example in case of user entering the query "mouse" there are large no of chances that the URLs that do not contain the query term mouse may contain the rich information about mouse.

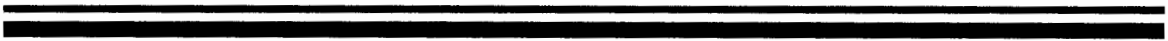
In case of query mouse the URL "http:informatix.jax" will be ignored when anchor text mining is performed. But this page contains very valuable information about mouse which will not be returned to the user and will be unconsidered. Similarly, the URLs which contain the query term mouse may not contain relevant and important information about mouse will be considered important.

- **Extra time consumption due to Blind Expansion**

Most of the search systems return results for broad queries only few are relevant to the topic of the query and are important for the user. Blind expansion of the root set adds a large no of nodes in the base set which are useless and not fulfilling the information needs of the user and also causing the extra time consumption.

4

Design



4. DESIGN

The techniques being used for web searching still lacking in returning most relevant results to the user query as discussed in literature survey. The problems like Topic contamination, topic drift, extra time consumption due to blindly expansion of the root set and drawbacks of Anchor text mining are highlighted in problem definition chapter.

This chapter presents an approach “SelARC with Content Mining” for improving the efficiency of web search systems and getting more relevant results against a user query. The basic aim is to combine the content analysis with the connectivity analysis methods to distil the pure topic and most relevant information. Section 4.1 describes the Automatic Resource Compilation (ARC) by chakrabaarti *et al* [3] also points towards the main drawback of this approach. Section 4.2 describes the Content and Connectivity Algorithms by Bharat *et al* [1] and also describes the main drawback of the algorithm. Section 4.3 gives brief detail of SelHITS by Mitra *et al* [7]. Section 4.4 describes the system architecture and also the main phases of the SelARC with Content Mining.

4.1 Automatic Resource Compilation (ARC).

Chakrabarti *et al* [3] designed an *automatic resource compiler* (ARC). An automatic resource compiler as discussed in detail in literature survey is a system to which a broad topic that is well represented on the World Wide Web is given and it will return a list of resources which it considers the most relevant. ARC is based on link and text analysis combination to distill out most relevant pages against the query topic. ARC computes a 2-neighborhood graph instead of one-neighborhood graph. Three phases of an ARC are shown in figure 4.1. ARC first gets a set of pages from some existing search engine and then augment using the links to 2-neighborhood. Then each link is assigned a weight $w(p, q)$. The weight of a link increases as topic related text increases in the anchor text of that link. Then in iteration and reporting phase iteratively calculates the values of Hub and Authority and reports the top hub and authority pages. The procedure “Expand root set to 2-neighborhood” gives a broader graph as compared to the “1-link expansion” procedure in other algorithms. So to find related pages to the query in a bigger graph of nodes will definitely improve the chances to get more relevant pages to the query topic.

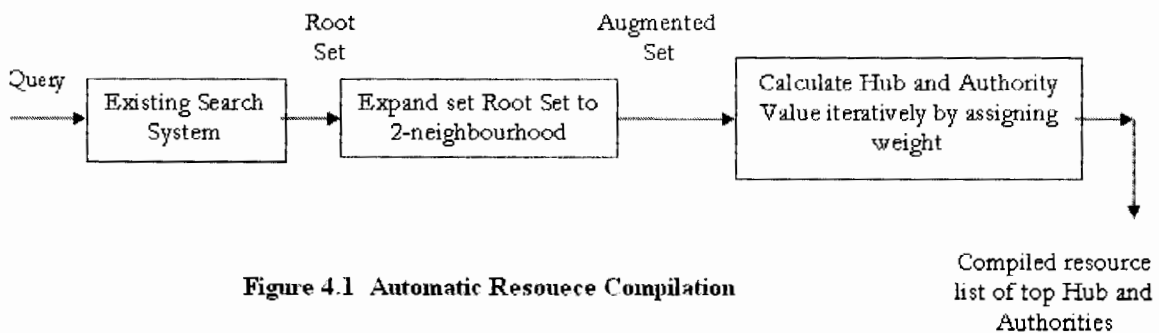


Figure 4.1 Automatic Resource Compilation

Drawback

The main drawbacks of ARC as discussed in the problem definition chapter are the problems due to the blind expansion and the text analysis of the href only and not the full contents of the page are analyzed. So many important pages can be ignored and many irrelevant pages can be added in the augment set.

4.2 Content and Connectivity Algorithms

Bharat *et al* [1] discussed nine content and connectivity based algorithms. These algorithms first construct a query specific graph against the user query whose nodes are documents. The graph is constructed by getting the root set of the web pages from some existing search engine by entering a user query. Then blind expansion is done by adding web pages from its neighborhood, which is the set of web pages either point to web pages in the root set or are pointed to by the web pages in the root set. The documents in the root set and its neighborhood documents collectively form the neighborhood graph's nodes. The number of nodes in the neighborhood graph is called the base set. After getting the base set, content analysis is performed to prune the irrelevant nodes from the neighborhood graph as shown in figure 4.2.

Pruning is performed by computing the relevance weights of the nodes in neighborhood graph and make use of the relevance weight of each node to make a decision if it should be removed from the graph. This decision is dependent on the thresholds of relevance weights. Relevance weight of any node is equal to the similarity of the query topic with the content in its document. All nodes whose weights are under the threshold value are pruned from base set and resultant pruned set is used then for further processing. Thresholds can be picked in one of three ways.

1. Median Weight

In this way the median of relevance weights of all the nodes in the graph is set as threshold.

2. Root Set Median Weight

In this way the median of relevance weights of all the documents in the root set is set as threshold.

3. Fraction of Maximum Relevance Weight

In this way a fixed fraction of the maximum relevance weight is set as threshold. Bharat *et al* used $\text{max}/10$.

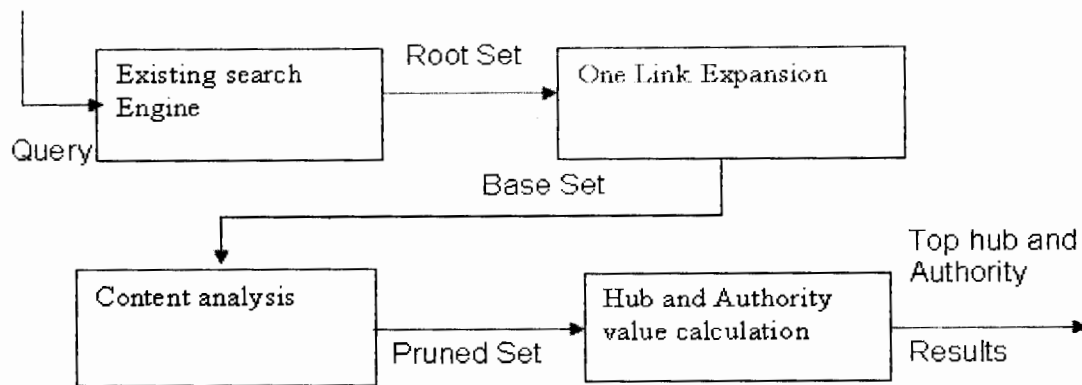


Figure 4.2 content and connectivity algorithms

On the pruned set the connectivity based algorithm “imp” is applied to compute the hub and authority scores for all the nodes in the pruned set and corresponding algorithms are named as *med*, *startmed*, and *maxby10*. These algorithms then report the top hub and authority pages to the user.

Drawback

The main drawback of the algorithms by Bharat *et al* is the blind expansion as discussed in the Problem definition chapter, due to which topic contamination and extra time consumption problems occur.

4.3 SelHITS Algorithm.

SelHITS algorithm by Mitra *et al* [7] begins with the user query and a small start set from existing search engine. The start set will be of few hundred pages related to the query. Hub and Authority values are then calculated of the nodes in the root set. Then top Hub and Authority pages are selected as candidate pages for further expansion. This selective expansion procedure of candidate pages drastically reduces the size of base set, as irrelevant nodes are not added to the candidate pages to get the base set. This avoids the problem of time consumption and topic contamination problem. SelHITS repeats the

TH 7494

same process on the base set as performed on the root set. Then top Hub and Authority pages are reported to the user. Refer the figure 4.3.

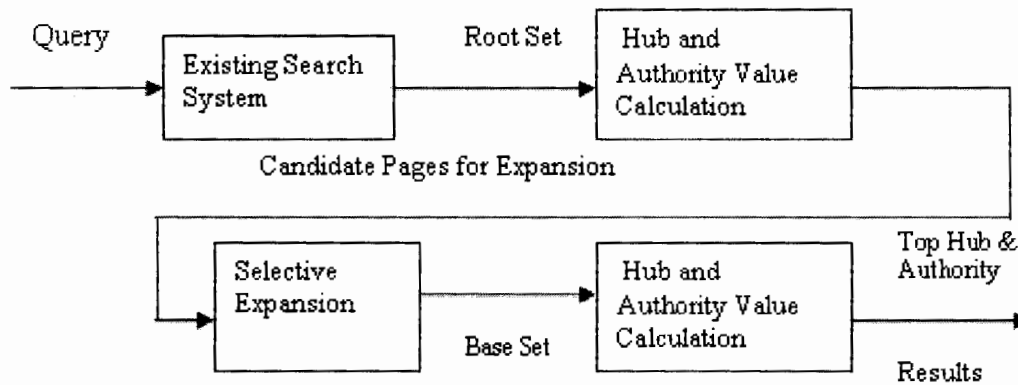


Figure 4.3: SelHITS

- The “Selective Expansion” procedure achieves considerable improvement over blindly “One Link Expansion” procedure as in figure 4.3. As only relevant pages are expanded instead of the expanding all relevant or irrelevant pages to get base set. Therefore the size of base set decreases drastically as compared to the blind expansion and the resultant pages added in the base set are useful and about to the user query.
- The “Selective Expansion” procedure also addresses the problem of topic contamination which occurs due to the blind expansion to get the base set. So the pages from the single topic are extended instead of the multiple topics.

4.4 System Architecture

The Proposed System “**SelARC with Content Mining**” is a combination of Content Analysis and Connectivity analysis methods. The proposed system handles the problems like Topic Drift, Topic Contamination and Extra Time Consumption. The main goal is not to provide heavy performance like search engines but to return the most relevant pages to only one of the main query topic.

The system starts with the user query given to some search engine and root set is got. The Hub and Authority value calculation phase calculates the Hub and Authority values for each page in the root set and selects top hub and authority pages as candidate pages. Then these candidate pages are gone through the “selective expansion” phase. This phase first expands the candidate pages to 1-neighbourhood and the pages with top hub and authority values are selected as base set. Then the base set is expanded up to 2-neighbourhood level by selective expansion process and we get the Augment set. This selective expansion will decrease the size of augment set as compared to the augment set by ARC algorithm. By selective expansion irrelevant pages will not be added first to the base set and then to the augment set. Now for further content analysis we will have only the pages related to the query topic so no extra time will be expanded on the irrelevant pages. As the augment set will contain only the pages from the single topic so in the case of ambiguous queries there will be no topic contamination problems. The system design of SelARC with Content Mining is shown in the figure 4.4.

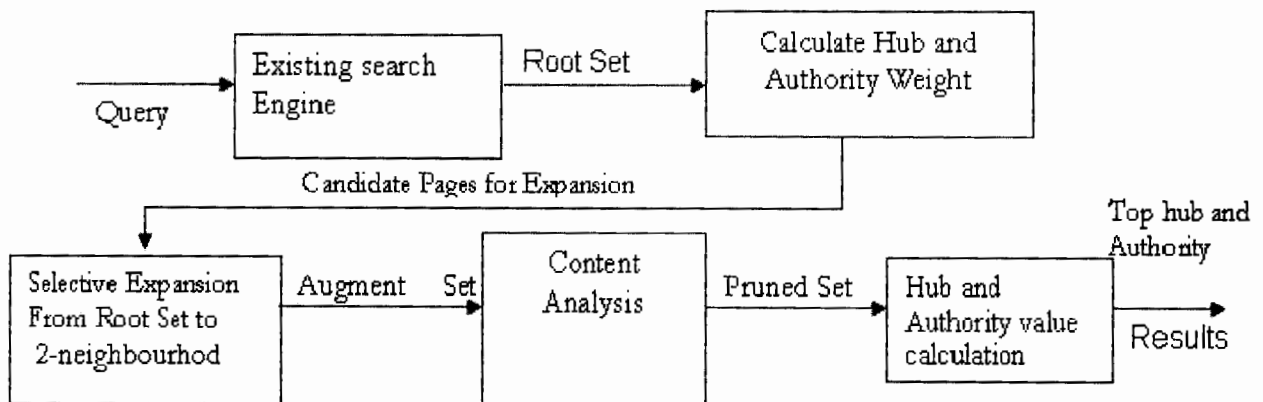


Figure 4.4 SelARC with Content Mining

The main phases of SelARC with Content Mining are:

- **Root Set:** user gives the query to the implemented system and the root set is got from some existing search engine.

- **Candidate Pages:** The implemented system performs the authority and hub values calculation for every web page in the root set and the top hub and authority pages are selected as the candidate pages.
- **Base Set:** The candidate pages are expanded selectively to 1-neighbourhood and the resultant set of pages is called the base set.
- **Augment Set:** The implemented system again calculates hub and authority values of each page in the base set and the top hub and authority pages are the candidate for further expansion. These candidate pages are then selectively expanded up to 2-neighborhood and the resultant pages form the augment setas shown in figure 4.5.

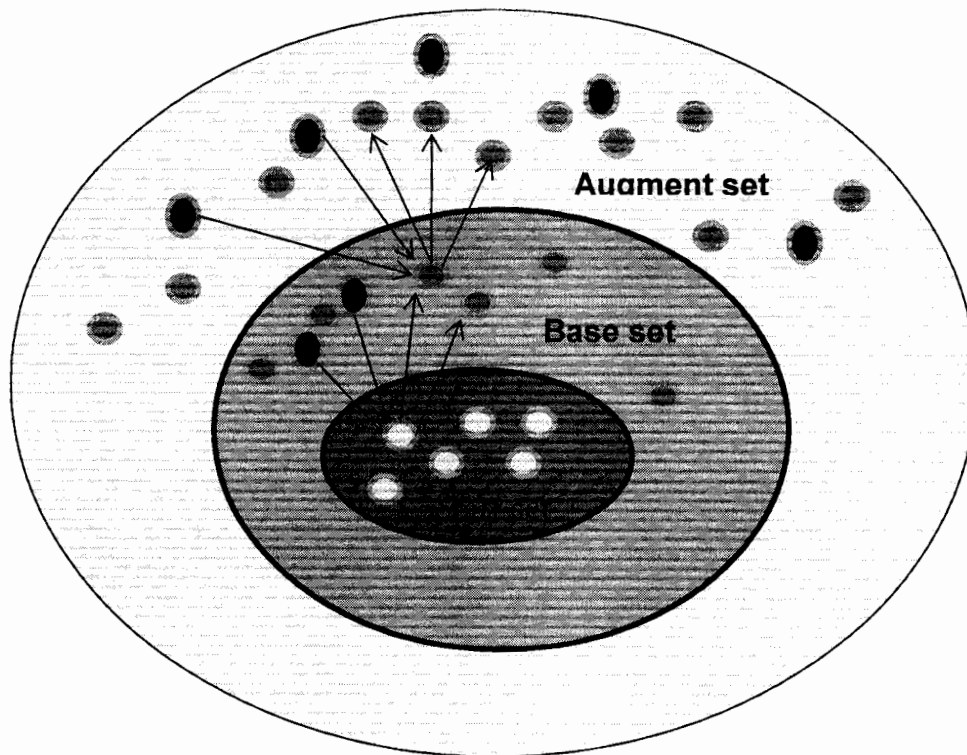


Figure 4.5 Augment Set

-
- **Pruned set:** Implemented system gets pruned set by further pruning the pages by applying the content analysis on the pages in the augment set.

 - **Content Analysis:** Content analysis is performed in one of three ways using Content analyzer:
 1. Start Setr
 2. Medianr
 3. Fraction of MaximumThe pruned set contains the pages which have relevance weight more or equal to the threshold value.

 - **Top Hub and Authority pages:** Finally the system calculates the top hub and authority values for the pruned set and reports the most relevant pages to the user query.

5

Implementation

5. IMPLEMENTATION

In the software lifecycle implementation phase is the stage where the realization of some idea takes place. It is the phase when thoughts, some plan, design or some specification are given some physical shape. A good implementation describes the summery of the noteworthy organization of deliverables. In computer science field implementation is a realization of some specifications as software component. For successful application or system a good implementation approach and strategy is very important.

5.1 Tools and Technology

For the implementation of SelARC with Content Mining the minimum requirements are as follows:

- Windows XP.
- High speed Internet.
- MATLAB7
- Visual Studio 2008 with C#
- MS Access
- To run this application Pentium IV with at least 3GHz processor speed is required.

5.1.1 MATLAB

MATLAB stands for matrix laboratory created by Cleve Moler in the late 1970s. It is a fourth generation programming language and provides us an environment for numerical computing. It is used for implementing algorithms, creation of interfaces, matrix manipulation, interfacing with programs in other programming languages and to plot different functions and data. MATLAB is numeric only but some optional toolbox also provides the capabilities of computer algebra. Graphical simulation and model-based

designing can also be done for dynamic and embedded systems. MATLAB integrates programming, computation and visualization in an environment which is easy to use. MATLAB was originally developed for matrix software but now it has been evolved over a period of years. It has become the standard instructional tool in the universities. It is an interactive system and its basic data element is an array that does not require dimensioning. This helps in solving many technical computing problems that have matrix and vector formulations. It has many add-on application specific features called toolboxes. These toolboxes are the collection of m-files and functions that can be extended to resolve particular classes of problems. The toolboxes available in the areas are control systems, signal processing, fuzzy logic, simulations, wavelets, neural networks and many others.

I used MATLAB for connecting with search engine google and for extracting URL's from web pages. I also used matlab for implementing different functions required to get the overall system's functionality.

5.1.2 C#.net

C# pronounced as see sharp, is an object oriented programming language. It is developed as part of the .NET initiative by Microsoft. It was then approved by ECMA and ISO. It encompasses many disciplines like object oriented and functional. C# also provides disciplines which allow to code generic and component oriented modules. Anders Hejlsberg leaded the development of C#. He already had designed the Borland's Turbo Pascal. It is a multi-paradigm programming language suitable for writing general purpose programs. One can code from very small dedicated systems to very large systems and can be used not only to code for embedded systems but also for hosted systems. C# is designed for Common Language Infrastructure (CLI) and the object oriented syntax of c# is similar to C++.

I used C# .net for content analysis of web pages by developing the SelARC Parser and Content Analyzer

5.2 Important phases and implementation functions

5.2.1 Start set

The start set is the order of at least 200 web pages related to the query topic. This start set is fetched from existing search engine (Google) against the user query. To fetch start set the function used is `web_root()`.

- **`web_root()`**

This function extracts the URL's from an existing search engine related to a user query. These links make the start set which is used for further processing.

5.2.2 Candidate set

The candidate pages are the pages selected from the start set whose hub and authority values are high. For calculation of top hub and authority pages four following functions are used.

- **`url_call()`**

This function finds out the links among the pages of the start set. This mapping of URL's helps in finding the top hub and top authority pages.

- **`start_authority()`**

This function calculates the authority value for each page in the start set.

- **`start_hub()`**

This function calculates the hub value for each page in the start set.

- **`candidates()`**

This function selects the top hub and top authority pages and forms the candidate set.

5.2.3 Selective Base set

The expansion of the candidate set up to the `1_neighborhood` level forms the selective base set. The pages which points to or pointed to by the hyperlinks in the pages

of the candidate set are added to the selective base set. The following functions fulfill this task of one link expansion.

- **cand_in()**
This function finds out the in links of the pages in the candidate set.
- **cand_out()**
The out links of the candidate set pages are founded by this function.
- **base_cand()**
The URL's of the candidate set pages and their in links found by `cand_in()`, out links found by `cand_out()` are combined by this function which forms the selective base set.

5.2.4 Aug Candidate set

The Aug Candidate pages are the pages selected from the selective candidate set whose hub and authority values are high. For calculation of top hub and authority pages four functions used are similar to the functions used to form the candidate set. The functions are `aug_url_call()`, `aug_top_authority()`, `aug_top_hub()`, `aug_candidates()`.

5.2.5 Selective Augment set

The Selective Augment set is formed by including all the out and in links of the aug candidate set up to `1_neighbourhood` and due to this twice one link expansion we get expansion of start set pages up to `2_neighbourhood`. Two functions perform this task are:

- **aug_out()**
This function finds all the out links of the pages in the aug candidate set.
- **augment()**
This function combines the URL's of aug candidate set and the URL's find out by `aug_out()` to form the selective augment set.

5.2.6 Pruned Set

Pruned set is formed by performing the content analysis due to which irrelevant pages are removed.

- **Plain text Extraction from web pages.**
- **Relevance weight calculation of web pages**

$$\text{similarity}(Q, D_j) = \frac{\sum_{i=1}^t (w_{iq} \times w_{ij})}{\sqrt{\sum_{i=1}^t (w_{iq})^2 \times \sum_{i=1}^t (w_{ij})^2}}$$

This formula is used to find the similarity of the query and the document by Bharat et al [1].

- **Medr()**

This function finds out the median of the relevance weights of the selective augment set. Then this median is used to prune the pages whose relevance weight is less than the median. The remaining set is the prune set.

- **Fracr()**

This function finds out the fraction/10 of the relevance weight of the pages in the selective augment set. The remaining set is a pruned set.

- **Start_medr()**

This function finds out the median of root set pages on the bases of relevance weights and pruned the pages on the bases of relevance weight less than median. The remaining set is pruned set.

5.2.7 Top hub and authority

Finally top hub and the authority pages in pruned set are reported to the user.

5.2.7.1 Top hub and authority through Medianr Algorithm

- **medr_match()**

This function finds out the link relation among pruned set pages.

- **medr_auth()**

This function finds out the top authority value of each page in the pruned set and reports the top authority pages.

- **Medr_hub()**

The hub value of each page of the pruned set is calculated by this function and the top hub pages are returned to the user.

5.2.7.2 Top hub and authority through Start Set Medianr algorithm

- **Startr_match()**

This function finds out the link relation among pruned set pages.

- **Startr_auth()**

This function finds out the top authority value of each page in the pruned set and reports the top authority pages.

- **Startr_hub()**

This hub value of each page of the pruned set is calculated by this function and the top hub pages are returned to the user.

5.2.7.3 Top hub and authority through Fractionr of Maximum algorithm

- **fracr_match()**

This function finds out the link relation among pruned set pages.

- **fracr_auth()**

This function finds out the top authority value of each page in the pruned set and reports the top authority pages.

- **fracr_hub()**

This hub value of each page of the pruned set is calculated by this function and the top hub pages are returned to the user.

6

Testing and Results

6. TESTING AND RESULTS

Testing is the process of investigation to check the quality of the product. Software testing is the independent view of the software to verify that it is satisfying the required functionality and behavior according to the needs of the user.

Well performed testing of the software can decrease the cost of maintaining the software. It also reduces the risks associated with the poor quality software like poor user productivity, calculation and data entry errors, and unacceptable functional behavior.

6.1 Testing Purpose

The main purpose of testing is to execute the program or application and find the bugs. A test is as good as much it has the probability to find the undiscovered errors in the software. Tests are designed that can uncover the bugs in minimum time and effort.

There are some main goals that have to be achieved:

- Discovering errors in the software.
- To build confidence in the proper operation of the software when there are no errors discovered by testing.
- Verify that all the requirements of the user are implemented.
- Verify that all the components of the software are integrated properly.
- Interaction between the objects must be verified.

Software Testing is the process of validating and verifying that a software product meets the business and technical requirements that guided its design and development, works as expected, and can be implemented with the same characteristics.

6.2 Testing Principals

Following are the principals to be kept in mind before testing.

- All tests must be traceable to the user requirements.

- As testing starts the tests must be planned long before.
- The whole software must be tested gradually.
- An independent third party must conduct the testing.

6.3 Testing the Design

Testing the design document is the key for a successful implementation. To discover design defects in the software is difficult due its complexity as software and any digital system are not continuous. So it is not sufficient to test only the boundary values to guaranty the correctness of software. It is necessary to test and verify all the values. Following points must be considered during the design testing.

- Is the design healthy?
- Design meets all the requirements?
- Is the design complete?
- Can the design be implemented?

6.4 Testing the Code

The testing strategy that will be used is that first the whole software will be tested against the specification to discover the part of specification that is not completed. Then the software will be tested against the implementation to discover those parts which are fault.

6.5 Testing methods

Software testing methods are of two types, black box testing and white box testing. These two approaches are used when the test engineer designs the test cases.

6.5.1 Black box testing

Software is treated as black box in this type of testing. No internal implementation detail is considered. The functionality of the software is tested according to the requirements. So the input data is given to the test object and is checked either the output data is or is not according to the specifications.

6.5.2 White box testing

Software is tested like a white box, or viewed as glass box. The full code every data structure and algorithm is tested and is visible to the tester. According to the implementation details all the tests are designed considering the programming language, styles and data structures. All the tests are derived by following the structure and flow of the program. Control-flow testing, data-flow testing, and loop testing, all correspond flow structure of the software.

6.6 Testing of SelARC with Content Mining

The testing of SelARC with Content Mining is undergone through all stages of black box testing and to some extent white box testing. The system is reviewed to see whether the objectives of the system are accomplished or not. A major factor considered during system evaluation is to evaluate the system with the perspective of the quires entered by the users.

The sample tests performed on SelARC with Content Mining are performed on the following queries:

1. Data mining
2. Apple
3. Windows

6.7 Top Authority Pages by medianr for query “windows”

The screenshot shows a Microsoft Access window titled "Microsoft Access - [medr_auth : Table]". The window displays a table with two columns: "url" and "authority". The table contains 20 rows of data, with the first row highlighted. The "authority" column shows values ranging from 5 to 1. The "url" column contains various web addresses related to Microsoft Windows and related services.

url	authority
http://www.microsoft.com/windowsserver2003/default.msp	5
http://www.microsoft.com/windowsmobile/en-us/help/default.msp	5
http://www.microsoft.com/windowsvista/features/foreveryone/sidebar.msp	5
http://www.microsoft.com/windowsxp/mediacenter/default.asp	5
http://www.microsoft.com/windows/catalog/server/default-v1.aspx	5
http://windowslive.com/	4
http://www.winsupersite.com/	4
http://messenger-support.spaces.live.com/blog/cns	2
http://www.ubcd4win.com/	2
http://officesharepointpro.com/	2
http://donavon.com	2
http://www.cnet.com/windows-vista.html	2
http://www.commonssensesecurity.info/	1
http://store.digitalriver.com/servlet/ControllerServlet?Action	1
http://social.zune.net/reportAbuse.aspx	1
http://mccain.livegadgets.net/	1
http://ad.doubleclick.net/clk	1
http://josh.corn/tiny/	1
http://280.photobucket.com/albums/kk189/donavonwest/LiveGadgets/WrittenByLiveGadget	1
http://forums.techguy.org/	1
http://blogs.msdn.com/pix/pages/Plug_2000_ins.aspx	1
http://blog.getpaint.net/	1
http://blog.donavon.com/2008/09/daily-gallup-poll-gadget.html	1
http://msdn.microsoft.com/subscriptions/	1

Record: 1 of 40
 Datasheet View CAPS NUM

Figure 6.1: Top Authority Pages by medianr

Figure 6.1 shows the top authority URL's against the user query “windows” by medianr algorithm. These top hub and authority web pages contain only one aspect of “windows” query. There is not a single page from other aspect of “windows” like house windows. The URL's show that even the anchor text don't contain the query term “windows” also are considered important which were ignored due to anchor text mining.

6.8 Top Hub Pages by medianr for query “windows”

The screenshot shows a Microsoft Access window titled "Microsoft Access - [medr_hub : Table]". The window displays a table with two columns: "url" and "hub". The table contains 20 rows of data, each representing a URL and its corresponding hub value. The hub values range from 11 down to 4. The interface includes a menu bar (File, Edit, View, Insert, Format, Records, Tools, Window, Help) and a toolbar with various icons. At the bottom, there is a status bar showing "Record: 1 of 196" and "Datasheet View".

url	hub
http://en.wikipedia.org/wiki/Microsoft_Windows	11
http://www.microsoft.com/en/us/default.aspx	11
http://windowslivewire.spaces.live.com/blog/cns	10
http://windowslivewire.spaces.live.com/?_c11_BlogPart_pagedir	8
http://livegadgets.net	7
http://www.kellys-korner-xp.com/xp_w.htm	6
http://www.xxcopy.com/	5
http://blogs.msdn.com/e7/archive/2008/11/30/accessibility-in-windows-7.aspx	5
http://www.windowsservercatalog.com/marketplace	5
http://bhandler.spaces.live.com/blog/cns	5
http://clk.atdmt.com/MRT/go/157262048/direct/01/	4
http://eu.microsoft.com/windowsxp/expertzone/tips/kellytheriot/kelly92.asp	4
http://www.microsoft.com/windows/windows-vista/features/easy-transfer.aspx	4
http://www.microsoft.com/windowsvista/features/foreveryone/sidebar.mspix	4
http://www.microsoft.com/windowsxp/mediacenter/default.asp	4
http://www.microsoft.com/windows/business/windows-7.aspx	4
http://clk.atdmt.com/MRT/go/157262048/direct/01/	4
http://www.microsoft.com/windowsxp/mediacenter/default.asp	4
http://www.microsoft.com/windows/products/winfamily/default.mspix	4
http://keznews.com/	4
http://www.microsoft.com/windows/windows-vista/features/easy-transfer.aspx	4
http://www.microsoft.com/windows/products/winfamily/default.mspix	4
http://www.microsoft.com/windows/business/windows-7.aspx	4
http://blogs.msdn.com/e7/archive/2009/01/30/our-next-engineering-milestone.aspx	4

Figure 6.2: Top Hub Pages by medianr

Figure 6.2 shows the top hub URL’s against the user query “windows” by medianr algorithm. These top hub and authority web pages contain only one aspect of “windows” query. There is not a single page from other aspect of “windows” like house windows. The URL’s show that even the anchor text don’t contain the query term “windows” also are considered important which were ignored due to anchor text mining.

6.9 Top Authority Pages by Fractionr of Maximum for query “windows”

The screenshot shows a Microsoft Access window titled "Microsoft Access - [frac_auth : Table]". The window displays a table with two columns: "url" and "authority". The table contains the following data:

url	authority
http://ad.doubleclick.net/jump/windowseur.click.com/adtarget	2
http://wincowslivewire.spaces.live.com/	1
http://www.msn.com/	1
http://www.sqlmag.com/Issues/Index.cfm?Action	3
http://www.windowstpro.com/Authors/AuthorID/1789/1789.html	1
http://www.windowstpro.com/Blog/	1
http://www.windowstpro.com/go/1Newsletters_top_nav_WIN_PUB	1
http://www.windowstpro.com/Issues/	1
http://www.windowstpro.com/Publications/	1
http://www.windowstpro.com/Topics/	1
*	0

Figure 6.3: Top Authority Pages by Fractionr of Maximum

Figure 6.3 shows the top authority URL’s against the user query “windows” by Fractionr of Maximum algorithm. These top hub and authority web pages contain only one aspect of “windows” query. There is not a single page from other aspect of “windows” like house windows. The URL’s show that even the anchor text don’t contain the query term “windows” also are considered important which were ignored due to anchor text mining.

6.10 Top Hub Pages by Fractionr of Maximum for query “windows”

The screenshot shows a Microsoft Access window titled "Microsoft Access - [frac_hub : Table]". The window displays a table with two columns: "url" and "hub". The table contains 20 rows of data, each representing a URL and its corresponding hub value. The hub values range from 3 to 5. The table is displayed in Datasheet View, and the status bar at the bottom indicates "Record: 1 of 140" and "Sunday, January 31, 2010".

url	hub
http://emea.windowsitpro.com/emea/	5
http://www.windowsitpro.com/emea/Article/ArticleID/102376/EMEA_102376.ht	5
http://www.windowsitpro.com/emea/Article/ArticleID/102391/EMEA_102391.ht	5
http://www.windowsitpro.com/emea/Article/ArticleID/102547/EMEA_102547.ht	5
http://www.windowsitpro.com/emea/Article/ArticleID/102549/EMEA_102549.ht	5
http://www.windowsitpro.com/emea/Article/ArticleID/102548/EMEA_102548.ht	5
http://www.windowsitpro.com/Downloads/	3
http://www.windowsitpro.com/windowspaulthurrott/	3
http://forums.windowsitpro.com/web/forum/messageview.aspx?catid	3
http://www.windowsitpro.com/AboutUs/Index.cfm?Action	3
http://www.windowsitpro.com/Blog/	3
http://www.windowsitpro.com/email	3
http://www.windowsitpro.com/Email/Index.cfm?promocode	3
http://www.windowsitpro.com/Publications/	3
http://www.windowsitpro.com/Authors/AuthorID/1789/1789.html	3
http://www.windowsitpro.com/EMEA/Issues/IssueID/942/Index.html	3
http://www.windowsitpro.com/Registration/Index.cfm?Action	3
http://www.windowsitpro.com/Windows/Issues/Index.html	3
http://www.windowsitpro.com/mediakit/	3
http://www.windowsitpro.com/Issues/	3
http://www.windowsitpro.com/go/1/Newsletters_top_nav_WIN_PUB	3
http://www.windowsitpro.com/Forums/	3
http://www.windowsitpro.com/essential	3
http://www.windowsitpro.com/Topics/	3

Figure 6.4: Top Hub Pages by Fractionr of Maximum

Figure 6.4 shows the top hub URL’s against the user query “windows” by Fractionr of Maximum algorithm. These top hub and authority web pages contain only one aspect of “windows” query. There is not a single page from other aspect of “windows” like house windows. The URL’s show that even the anchor text don’t contain the query term “windows” also are considered important which were ignored due to anchor text mining.

6.11 Top Authority Pages by start set median for query “windows”

The screenshot shows a Microsoft Access window titled "Microsoft Access - [start_r_auth : Table]". The window displays a table with two columns: "url" and "authority". The table contains 20 rows of data, sorted by authority in descending order. The first row has an authority of 8, and the last row has an authority of 1. The status bar at the bottom indicates "Record: 1 of 37" and "Datasheet View".

url	authority
http://www.microsoft.com/	8
http://www.winsupersite.com/	7
http://www.windowsitlibrary.com/	3
http://www.ubcd4win.com/	3
http://www.windowslive.com/	3
http://keznews.com/	2
http://leoville.tv/podcasts/www.xml	2
http://www.opera.com/download/	2
http://www.gamesforwindows.com/	2
http://www.windowsmarketplace.com/	2
http://www.skype.com/download/	2
http://twit.tv/www	2
http://www.cnet.com/windows-vista.html	2
http://www.activewin.com/	1
http://www.live.com/	1
http://www.annoyances.org/	1
http://www.dougknox.com/	1
http://audacity.sourceforge.net/download/windows	1
http://www.windweaver.com/w95man.htm	1
http://www.divx.com/divx/	1
http://windowsupdate.microsoft.com/	1
http://www.winistaclub.com/Ultimate_Windows_Tweaker.html	1
http://v4.windowsupdate.microsoft.com/	1
http://blogs.msdn.com/e7/	1

Figure 6.5: Top Authority Pages by start set medianr

Figure 6.5 shows the top authority URL's against the user query “windows” by start set medianr algorithm. These top hub and authority web pages contain only one aspect of “windows” query. There is not a single page from other aspect of “windows” like house windows. The URL's show that even the anchor text don't contain the query term “windows” also are considered important which were ignored due to anchor text mining.

6.12 Top Hub Pages by start set median for query “windows”

The screenshot shows a Microsoft Access window titled 'Microsoft Access - [start...hub - Table]'. The window displays a table with two columns: 'url' and 'hub'. The table contains 22 records, with the 'hub' column values ranging from 6 down to 0. The records are sorted by the 'hub' value in descending order. The first record has a 'hub' value of 6 and a URL starting with 'http://emea.windowsitpro.com/emea/'. The last record has a 'hub' value of 0 and a URL starting with 'http://windowslivewire.spaces.live.com/'.

url	hub
http://emea.windowsitpro.com/emea/	6
http://www.windowsitpro.com/emea/Article/ArticleID/102376/EMEA_102376.html	5
http://en.wikipedia.org/wiki/Microsoft_Windows	5
http://www.windowsitpro.com/Publications/	4
http://www.levenez.com/windows/	4
http://forums.windowsitpro.com/web/forum/messageview.aspx?catid	3
http://www.windowsitpro.com/AboutUs/Index.cfm?Action	3
http://www.windowsitpro.com/Authors/AuthorID/1789/1789.html	3
http://www.windowsitpro.com/Downloads/	3
http://www.windowsitpro.com/email	3
http://www.windowsitpro.com/Email/Index.cfm?promocode	3
http://www.microsoft.com/	3
http://www.windowsitpro.com/EMEA/Issues/IssueID/942/Index.html	3
http://www.windowsitpro.com/windowspaulthurrott/	3
http://www.windowsitpro.com/essential	3
http://www.windowsitpro.com/Forums/	3
http://www.windowsitpro.com/mediakit/	3
http://www.windowsitpro.com/Registration/Index.cfm?Action	3
http://www.windowsitpro.com/Topics/	3
http://www.windowsitpro.com/Windows/Issues/Index.html	3
http://www.windowsitpro.com/emea/Article/ArticleID/102391/EMEA_102391.html	3
http://windowslivewire.spaces.live.com/	1
*	0

Record: 1 of 22
 Datasheet View
 Sunday, January 31, 2010

Figure 6.6: Top Hub Pages by start set medianr

Figure 6.6 shows the top hub URL's against the user query “windows” by start set medianr algorithm. These top hub and authority web pages contain only one aspect of “windows” query. There is not a single page from other aspect of “windows” like house windows. The URL's show that even the anchor text don't contain the query term “windows” also are considered important which were ignored due to anchor text mining.

6.13 Analysis of Results

The analysis of results shows that in case of performing content analysis rather than only the anchor text mining as in ARC [4], has improved the results as many pages which were ignored are considered important now. On the other hand some irrelevant pages

added due to blind expansion are now ignored and the problem of extra time consumption is solved now to some extent as the root set and base set is selectively expanded. So we can see from results that the augment set is much smaller as it would be in case of blind expansion.

The first query that I selected to test this system is “windows”. There are two interpretations for this broad query. One is the window in the wall and the second is windows operating system. Due to these interpretations there was the problem of topic contamination as the top hub and authority pages resulted by previous algorithms were from different interpretations instead of only one. So by selective expansion of root set up to 2-neighborhood and then finally performing content analysis has proved the better results as compared to just performing the anchor text mining of the previous algorithms.

Similarly I got improved performance in the case if the queries “apple” and “data mining”. The table 6.1 shows the comparison of SelARC with Content Mining system with the previous techniques.

Table 6.1: Comparison of “SelARC with Content Mining” with other Techniques

Existing Techniques	One Link Expansion	2-Link Expansion	Selective Expansion	Anchor Text Mining	Content Analysis	Hub & Authority Calculation
Kleinberg et al. (1997)	✓	x	x	x	x	✓
Bharat et al. (1998)	✓	x	x	x	✓	✓
Chakrabarti et al. (2002)	✓	✓	x	✓	x	✓
Awekar et al. (2007)	✓	x	✓	x	x	✓
Our Approach	✓	✓	✓	x	✓	✓

7

Conclusion and Future Works

7. Conclusion and Future Works

7.1 Conclusion

The Proposed System “**SelARC with Content Mining**” is a combination of Content Analysis and Connectivity analysis methods. The implemented system is successful in giving the improved desired results. Previously the blind expansion of ARC and other content analysis algorithms drastically increases the size of augment set. Due to which useless pages added to the augment set takes extra time consumption for further processing.

The selective one link expansion of SelARC causes this time consumption reduced but the pages added are only from one neighborhood so the area is limited in the sense that pages added to the base set are less than the no of pages in the augment set which contains pages from 2-neighborhood. So the SelARC with Content Mining by performing selective expansion of root set to 2-neighborhood gets more pages related to the query and the problem of topic drift is also solved as the chances of adding pages which are of no use to the user are distilled.

Further the content analysis is performed on the selected augment set by which the pages ignored due to the anchor text mining of ARC, are not ignored and added to the pruned set. Finally the distilled set of pages purely related to a single aspect of the query topic is got. So in case of broad topic queries the problem of topic contamination is solved to some extent.

7.2 Future Works

1. To get the root set dynamically from the web sources instead of getting it from some existing search engine can be an important future direction of this research work.
2. To improve the same system by expanding the root set to n-neighborhood.
3. Further to make a search engine using the improved algorithms of SelARC with Content Mining is another future direction of this research.
4. To combine the Anchor Text Mining with the Content mining of the web pages can also be a future work of this research.

Appendix-A

List of Abbreviations

Appendix A**DEFINITION OF TERMS**

Abbreviations	Full Form
HITS	Hypertext Induced Topic Search
SelHITS	Selective Hypertext Induced Topic Search
H	Hub
A	Authority
SALSA	Stochastic Algorithm for Link Structure Analysis
URL	Universal Resource Locator
HTML	Hypertext Induced Topic Search
WWW	World Wide Web
N/W	Network
ARC	Automatic Resource Compilation
DOM	Document Object Model
IR	Information Retrieval

Appendix-B

Screen Shots

Appendix B

SCREEN SHOTS

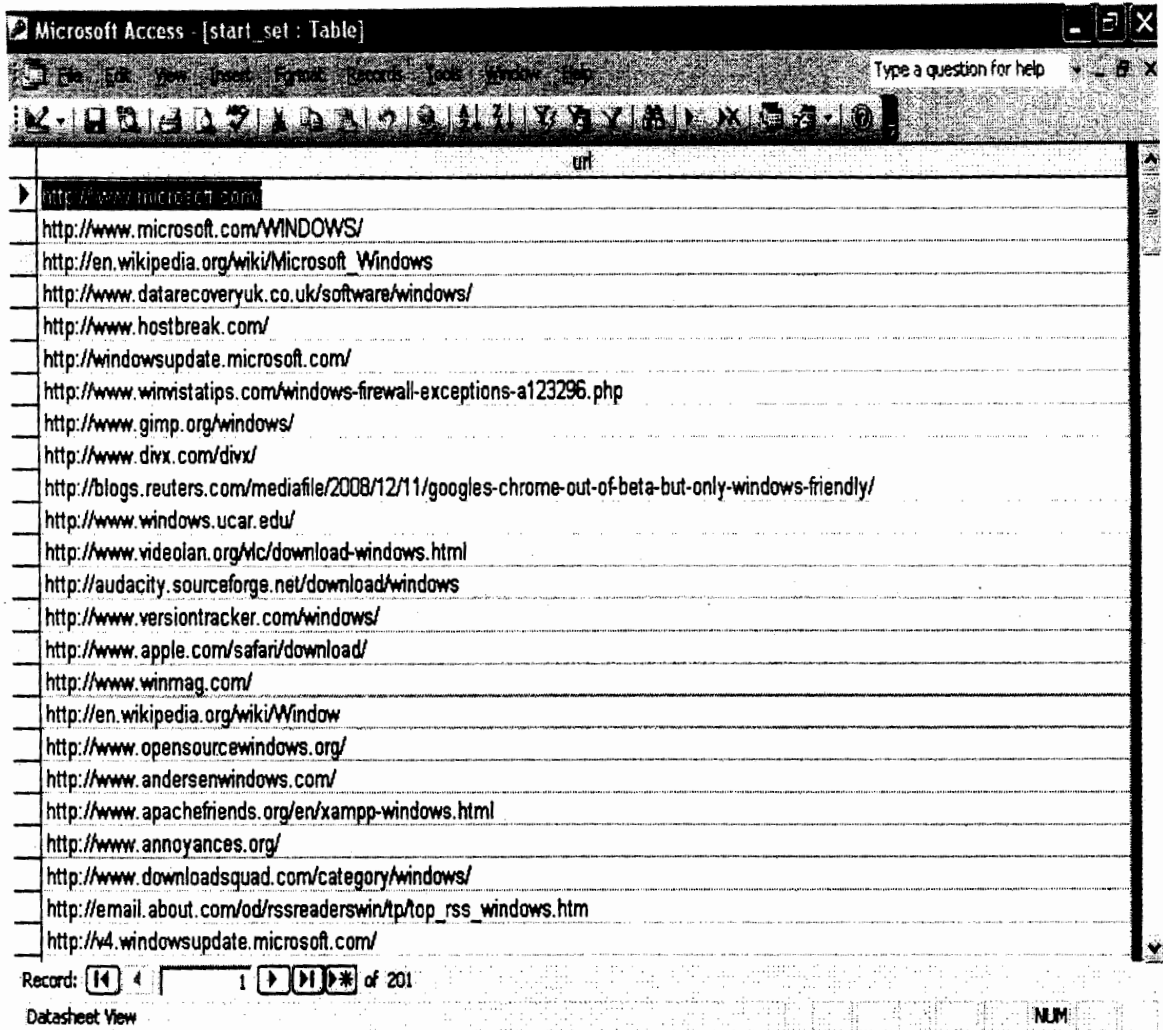


Fig B-1: Start/Root Set

url	hub
http://blogs.msdn.com/powershell/	45
http://blogs.msdn.com/powershell/	44
http://blogs.technet.com/markrussinovich/archive/2	44
http://blogs.technet.com/virtualization/	44
http://cran.r-project.org/bin/windows/base/rw-FAQ	44
http://en.wikipedia.org/wiki/Microsoft_Windows	51
http://h20341.www2.hp.com/integrity/cache/497701-0	44
http://httpd.apache.org/docs/1.3/windows.html	44
http://httpd.apache.org/docs/2.0/platform/windows	44
http://kottke.org/08/12/does-the-broken-windows-lh	1
http://labmics.techtarget.com/windowsxp/default.ht	45
http://leoville.tv/podcasts/ww.xml	60
http://lifehacker.com/tag/windows/	3
http://neosmart.net/blog/2008/windows-vista-recove	44
http://plugindoc.mozdev.org/windows.html	44
http://support.microsoft.com/kb/126449	44
http://twit.tv/ww	2
http://windowsclient.net/	44
http://windowsitpro.com/windowsnt20002003faq/	10
http://windowslivewriter.spaces.live.com/	44
http://windowsonecare.spaces.live.com/	44
http://windowsteamblog.com/blogs/windowsvista/arch	44
http://windowsupdate.microsoft.com/	45
http://www.3ivx.com/download/windows.html	44

Record: 14 of 65
 Datasheet View

Fig B-2: Start_Hub

url	authority
http://www.winsupersite.com/	44
http://www.winsupersite.com/	7
http://www.windowstlibrary.com/	3
http://www.ubcd4win.com/	3
http://www.windowslive.com/	3
http://www.gamesforwindows.com/	3
http://keznews.com/	2
http://leoville.tv/podcasts/ww.xml	2
http://www.cnet.com/windows-vista.html	2
http://www.windowsmarketplace.com/	2
http://www.skype.com/download/	2
http://twit.tv/ww	2
http://www.opera.com/download/	2
http://www.activewin.com/	1
http://www.linuxrsp.ru/win-lin-soft/table-eng.html	1
http://www.annoyances.org/	1
http://www.nu2.nu/pebuilder/	1
http://www.dougknox.com/	1
http://audacity.sourceforge.net/download/windows	1
http://www.windweaver.com/w95man.htm	1
http://www.divx.com/divx/	1
http://windowsupdate.microsoft.com/	1
http://www.winvistaclub.com/Ultimate_Windows_Tweaker.html	1
http://v4.windowsupdate.microsoft.com/	1

Record: 14 of 39
 Datasheet View

Fig B-3: Start_Auth

Microsoft Access - [candidates : Table]

Type a question for help

url
http://www.microsoft.com/
http://leoville.tv/podcasts/www.xml
http://keznews.com/
http://www.windowsmarketplace.com/
http://www.windowslive.com/
http://www.opera.com/download/
http://www.windowstlibrary.com/
http://www.cnet.com/windows-vista.html
http://www.gamesforwindows.com/
http://www.skype.com/download/
http://twit.tv/www
http://www.winsupersite.com/
http://blogs.msdn.com/e7/
http://blogs.msdn.com/powershell/
http://blogs.technet.com/markrussinovich/archive/2
http://blogs.technet.com/virtualization/
http://cran.r-project.org/bin/windows/base/rw-FAQ
http://en.wikipedia.org/wiki/Microsoft_Windows
http://h20341.www2.hp.com/integrity/cache/497701-0
http://httpd.apache.org/docs/1.3/windows.html
http://httpd.apache.org/docs/2.0/platform/windows
http://labmice.techtarget.com/windowsxp/default.ht
http://leoville.tv/podcasts/www.xml

Record: 14 of 69

Datasheet View

Fig B-4: Candidates

Microsoft Access - [cand_in : Table]

Type a question for help

url_in
http://www.jkdefrag.fr/
http://blog.rootshell.be/category/grm/
http://unexpectedcy.com/
http://www.blogaholic.de/blog/2007/01/
http://soaringbear.com/comp/software.html
http://www.broadbandreports.com/forum/21534853-HELP-My-PC-got-screwed-big-timeVont-Start
http://www.vcit.ca/wordpress/2008/01/28/best-ever-windows-xp-how-to-tips-maintenance-freeware/
http://ezhey.com/
http://www.ultimatebootcd.com/
http://unixadm.blogspot.com/2007_11_01_archive.html
http://www.libellules.ch/dotclear/index.php?category/Livecd
http://igorbrejc.net/category/fridaygoodies
http://blog.rootshell.be/tag/tools/
http://www.kommunikationsforum.dk/kristian-risager-larsen
http://www.aeromondo.com/2007_05_01_archive.html
http://ubcd4win.com/forum/index.php?showtopic
http://www.recuperationdedonneesperdues.com/
http://www.miketechshow.com/etom.xml
http://www.idiotvox.com/Languages/PodCast_Review_Comunicando_podcast_programa_de_radio_tecnologia_Internet_y_juegos
http://www.tinyapps.org/blog/2005_05_01_archive.html
http://www.supersatellite.com/category/tech/hardware/
http://www.mefn.org/board/lofversion/index.php/122773.html
http://news.cnet.com/8301-13554_3-9988099-33.html

Record: 14 of 68

Datasheet View

Fig B-5: cand_in

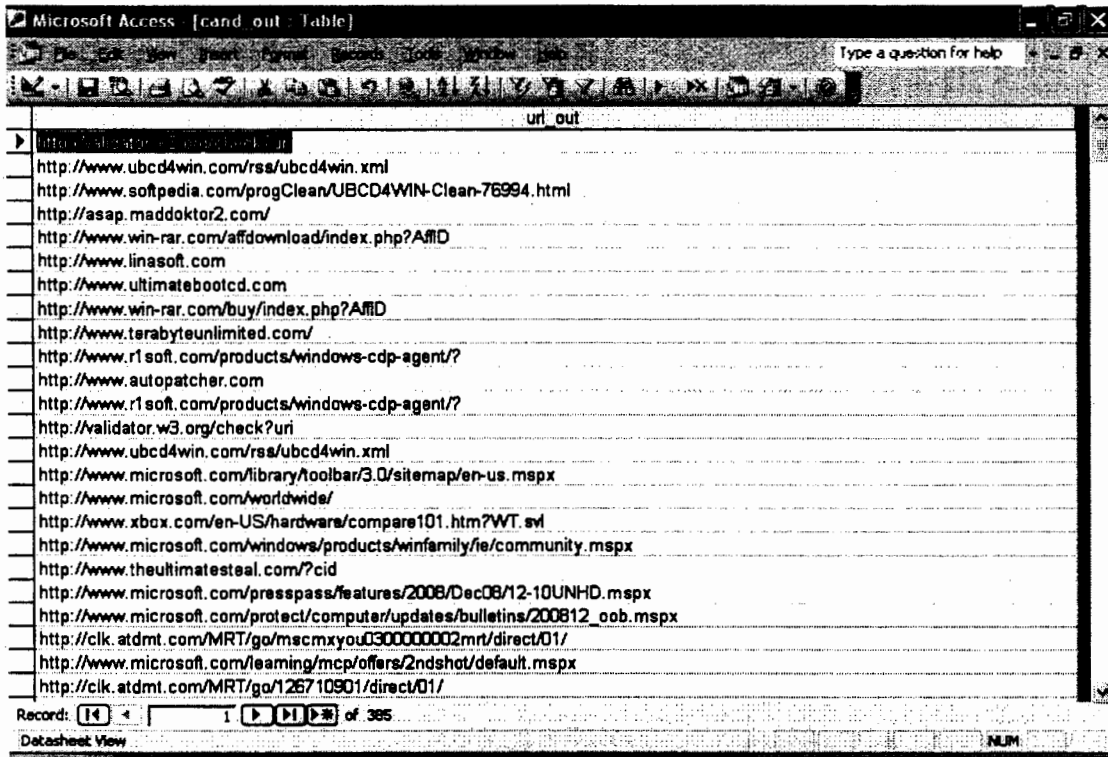


Fig B-6: cand_out

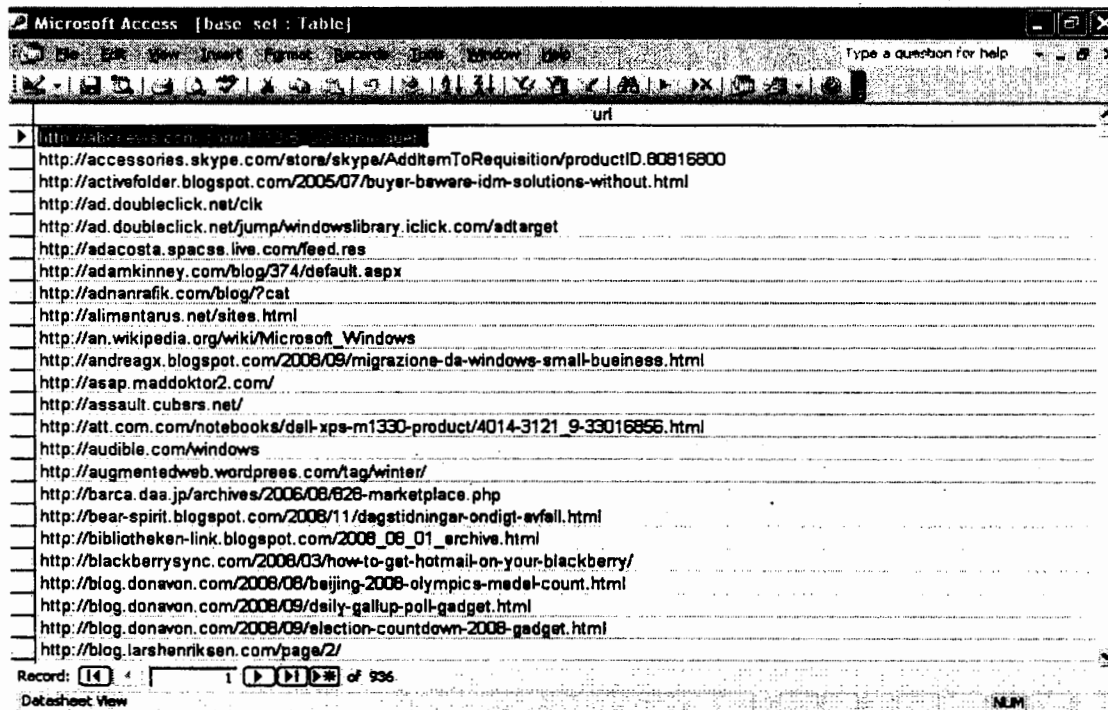


Fig B-7: base_set

Microsoft Access [out - Table]

Type a question for help

	url
	http://65.2.146.90/ash-central/
	http://66.129.1.101/top20.htm
	http://66.33.126.34/php/home.php?lang
	http://72andsunny.com/
	http://99-bottles-of-beer.is-la.net/
	http://9pregnancy.com/
	http://a.collective-media.net/jump/ldgt.greendatacenter/
	http://a.collective-media.net/jump/ldgt.thegamereviews/home_above
	http://a.tribalfusion.com/i.click?site
	http://a9.com/optical?a
	http://aa.usno.navy.mil/data/docs/RS_OneDay.html
	http://aaai.org/AITopics/html/faq.html
	http://aaai.org/AITopics/html/quotes.html
	http://aaai.org/AITopics/html/reference.html
	http://aabestcomfort.com/
	http://aalc.rkfilms.com/
	http://aanr.com/
	http://abbicebanding.wordpress.com/
	http://abe.midco.net/baika/600x600project/
	http://abel.math.harvard.edu/
	http://abigailreviews.blogspot.com
	http://aboutus.enterprise.com/files/181/ERAC_Thin_Client_Terminal_Apri08.pdf
	http://abstrusegoose.com/
	http://abu.cnam.fr/

Record: 1 of 21634

Datasheet View

Fig B-8: base out links

Microsoft Access [set - Table]

Type a question for help

id	url
	http://www.wininformant.com
10	http://www.devconnections.com/shows/FALL2009WIN/default.asp?c
100	http://www.windowsitpro.com/Web/
1000	http://jeffmcneill.com/category/voip/
10000	http://www.gamerzines.com/pc/overview.html
10001	http://www.gamerzines.com/psp/overview.html
10002	http://www.gamerzines.com/ds/overview.html
10003	http://www.gamerzines.com/support/about-us.html
10004	http://www.gamerzines.com/support/contact-us.html
10005	http://www.gamerzines.com/support/privacy-policy.html
10006	http://www.gamerzines.com/support/terms-and-conditions.html
10007	http://www.gamerzines.com/support/competition-terms-and-conditions.html
10008	http://www.gamerzines.com/support/site-map.xml
10009	http://www.cranberrypublishing.com/
1001	http://jeffmcneill.com/category/weather/
10010	http://www.gamerzines.com/xbox-360/magazine/section-8-magazine.html
10011	http://www.gamerzines.com/pc/news/buy-champ-man-for-a-few-quid.html
10012	http://www.gamerzines.com/wii/magazine/cursed-mountain-magazine.html
10013	http://www.gamerzines.com/pc/news/grin-to-close.html
10014	http://www.gamerzines.com/ps3/previews/12-sturmovik-handson.html
10015	http://www.gamerzines.com/pc/news/sid-meier-xcom.html
10016	http://www.gamerzines.com/pc/magazine/champions-online-beta-keys.html
10017	http://www.gamerzines.com/pc/news/fallout-3-mothership-zeta.html
10018	http://www.gamerzines.com/ps3/news/mw2-prestige-night-vision.html

Record: 1 of 22203

Datasheet View

Fig B-9: augment set

id	weight
100	0.100631596091826
1000	0.0811250166648056
1001	0.23217947950967
1002	0.174135964616616
1003	0.392475254013612
1004	0.105808912847842
1005	0.313073050627348
1006	0.237107999109134
1007	0.266662605911073
101	0.170027872929104
1010	0.225494783436744
1011	0.250528071322377
1012	0.192056374566669
1013	0.0962493740406737
1014	0.0912161498830737
1015	0.0773511305288661
1016	0.678572127315118
1017	0.265341204691469
1018	0.451399921089904
1019	0.426885045947905
102	0.262384338722983
1020	0.269421924325019
1021	0.368913133818957
	0.124959894713367

Fig B-10: median weight

id	weight
10	0.847474137276359
100	0.823698181123166
101	0.665227793387839
102	0.878288748846198
103	0.854829571002378
104	0.626637391530304
105	0.866815682867661
106	0.235824196200973
107	0.762870755568149
108	0.822096084814456
109	0.809755240955071
11	0.786454850663326
111	0.871487863699792
112	0.828157341968321
113	0.892494804194718
114	0.721333028753064
115	0.86034834703939
116	0.776035044499745
117	0.826138172124381
118	0.809668635442429
119	0.597270588766507
12	0.631753214970559
120	0.78810855695493
	0.874769748311008

Fig B-11: start median weights

id	url
1000	http://jeffmcneill.com/category/web/
1004	http://jeffmcneill.com/category/wiki/
1015	http://jeffmcneill.com/distance-education-and-virtual-worlds-some-issues/
1017	http://jeffmcneill.com/we-dont-build-websites-anymore/
1018	http://jeffmcneill.com/mathml-tex-latex-texvc-mimetex-oh-my/
102	http://www.windowsitpro.com/Blog/
1022	http://jeffmcneill.com/2006/11/
1023	http://jeffmcneill.com/category/opensource/page/2/
1024	http://jeffmcneill.com/category/entrepreneurship/
1035	http://jeffmcneill.com/blogs-blogging-bloggerfic-blogtastic/
1036	http://jeffmcneill.com/mit-launches-center-for-collective-intelligence/
1037	http://jeffmcneill.com/2006/10/
1039	http://jeffmcneill.com/2009/05/
1042	http://jeffmcneill.com/tag/culture/
1045	http://jeffmcneill.com/tag/open/
1049	http://jeffmcneill.com/2006/09/
105	http://www.sqlmag.com/Issues/Index.cfm?Action
1050	http://jeffmcneill.com/impact-of-the-web-on-elearning-and-the-blackboard-patent-dispute/
1051	http://jeffmcneill.com/2006/08/
1057	http://jeffmcneill.com/2008/07/
1064	http://jeffmcneill.com/2009/07/
1068	http://jeffmcneill.com/tag/calendar/
107	http://ad.doubleclick.net/jump/windowseur.iclick.com/edtarget
1070	http://jeffmcneill.com/tag/google/

Record: 1 of 312
Datasheet View

Fig B-12: median links

url	authority
http://www.microsoft.com/windows/catalog/server/default-v1.aspx	4
http://www.microsoft.com/windows/mobile/en-us/help/default.mspix	5
http://www.microsoft.com/windows/server/2003/default.mspix	5
http://www.microsoft.com/windows/vista/features/foreveryone/sidebar.mspix	5
http://www.microsoft.com/windows/xp/mediacenter/default.asp	5
http://windowslive.com/	4
http://messenger-support.spaces.live.com/blog/cns	2
http://www.ubcd4win.com/	2
http://donavon.com	2
http://store.digitalriver.com/service/ControllerServlet?Action	1
http://social.zune.net/reportAbuse.aspx	1
http://obama.livegadgets.net/	1
http://msdn.microsoft.com/subscriptions/	1
http://josh.com/tiny/	1
http://www.friendfeed.com/leolaporte	1
http://1280.photobucket.com/albums/kk189/donavonwest/LiveGadgets/WrittenByLiveGadget	1
http://hillary.livegadgets.net/	1
http://friendfeed.com/leolaporte	1
http://forums.techguy.org/	1
http://blogs.msdn.com/pix/pages/Plug_2000_ins.aspx	1
http://blog.getpaint.net/	1
http://blog.donavon.com/2008/09/daily-gallup-poll-gadget.html	1
http://mccain.livegadgets.net/	1
http://www.pctorium.com/	1

Record: 1 of 40
Datasheet View

Fig B-13: Top Medianr Authority

url	hub
http://asap.maddoktor2.com/charter.html	11
http://windowslivewire.spaces.live.com/?c11_BlogPart_pagedir	10
http://windowslivewire.spaces.live.com/blog/cns	7
http://livegadgets.net	5
http://www.xxcopy.com/	5
http://www.windowsservercatalog.com/marketplace	4
http://www.microsoft.com/windowsxp/mediacenter/default.asp	4
http://www.microsoft.com/windows/business/windows-7.aspx	4
http://www.microsoft.com/windows/products/winfamily/default.mspx	4
http://www.microsoft.com/windows/windows-vista/features/easy-transfer.aspx	4
http://www.microsoft.com/windowsvista/features/foreveryone/sidebar.mspx	4
http://eu.microsoft.com/windowsxp/expertzone/tips/kellytheriot/kelly92.asp	4
http://clk.atdmt.com/MRT/go/157262048/direct/01/	3
http://www.windowssitpro.com/Blog/	2
http://leoville.com/category/video	2
http://blog.donavon.com/2008/09/daily-gallup-poll-gadget.html	2
http://leoville.com/category/news	2
http://www.windowssitpro.com/Downloads/	2
http://www.windowssitpro.com/Authors/AuthorID/1789/1789.html	2
http://www.voder.com/nwdsk/	2
http://leoville.com/category/blog	2
http://leoville.com/category/appearances	2
http://messengersays.spaces.live.com/	2
http://josh.com/tiny/	1

Record: 1 of 92
 Datasheet View

Fig B-14: Top Medianr Hub