Improved Algorithm for Topic Distillation Using SelHITS

Developed by:

ZUBEDA KHANUM
(266- FAS/MSCS/F05)

Supervised by:

PROF. DR. M. SIKANDER HAYAT KHIYAL

Department of Computer Science
Faculty of Basic and Applied Sciences
International Islamic University Islamabad
2008

19-7-2018

See

T0 5256

1- Data mining

2- Algorithm for distillation

C2 of T05256

M.

27/12/10

بسم الله الرحمن الرحيم

In the Name of Allah The Most Beneficent
The Most Merciful

Department of Computer Science

International Islamic University, Islamabad

Date: 21-08-08

FINAL APPROVAL

It is certified that we have read the project titled “Improved Algorithm for Topic Distillation Using SelHITS” submitted by **Miss ZUBEDA KHANNUM Reg. No. 266-FAS/MSCS/F05.** 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. A. Sattar
Fmr. Director General
Pakistan Computer Bureau,
Islamabad.

Internal Examiner:

Dr. Muhammad Sher
Head of Department
Department of Computer Science
International Islamic University,
Islamabad.

Supervisors:

Dr. M. Sikandar Hayat Khiyal
Chair Person Department of Computer
Science/Software Engineering
Fatima Jinnah Women University,
The Mall, Rawalpindi.

Hayat

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

LITERATURE SURVEY

2.1 Hypertext Induced Topic Search

The **HITS** ("hypertext induced topic selection") algorithm is an algorithm for rating and rankingWeb pages Kleinberg [6]. HITS uses two values for each page, the *authority value* and the *hub value*. Authority and hub values are defined in terms of one another in a mutual recursion. An authority value is computed as the sum of the scaled hub values that point to that page. A hub value is the sum of the scaled authority values .Kleinberg 6] proposed a more refined notion for the importance of web pages. He suggested that web page importance should depend on the search query being performed. Furthermore, each page should have a separate "authority" rating (based on the links going *to* the page) and "hub" rating (based on the links going *from* the page). Kleinberg 6] proposed to use text-based web search engine (such as AltaVista) to get a "Root Set" consisting of a short list of web pages relevant to a given query. Second, the Root Set is augmented by pages which link to pages in the Root Set, and also pages which are linked to pages in the Root Set, to obtain a larger 'Base Set" of web pages.

- **Root Set:**

  For a given user query, we obtain a set of relevant documents using some existing search system e.g. Google, Yahoo! This set is called the root set. How to get a root set is shown in Figure2.1.

- **Base Set:**

  We expand root set by one link neighborhood to obtain the expanded root set or Base Set. How to generate Base Set from Root Set is shown in Figure 2.2.

- **Hub Page:**

  A page that doesn't provide information, but tell you where to find the information. Example of Hub page is shown in Figure2.3(a).

- **Authority page:**

  Authority page is a page which contains information about the topic of the query or is directly relevant to the topic of query. Example of an Authority page is shown in Figure2.3(b).

- **Co-Reference and Co-Citation:**

  The hub and authority matrices have interesting connection to two important concepts, co-citation and co-reference, which are fundamental metrics to characterize the similarity between two. The authority matrix is the sum of Co-citation and in degree. The fact that two distinct WebPages co-reference many other WebPages indicates that these have certain commonality. Co-reference measures the similarity between WebPages. Thus hub matrix is the sum of co-reference and out degree

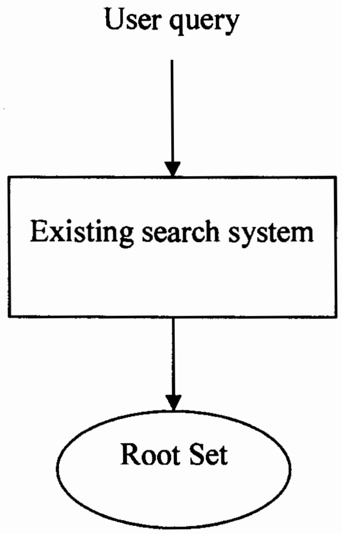


Fig2.1: How to get Root Set

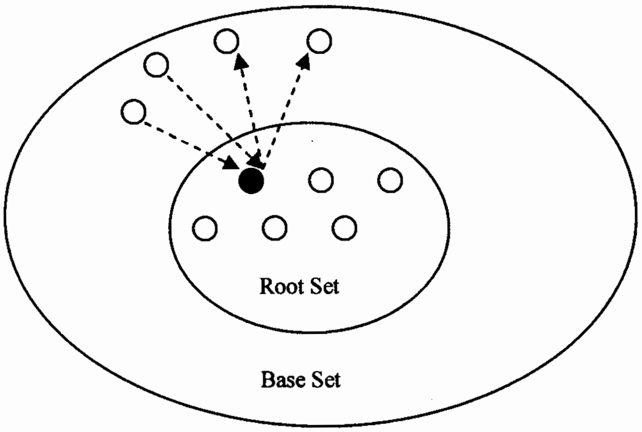
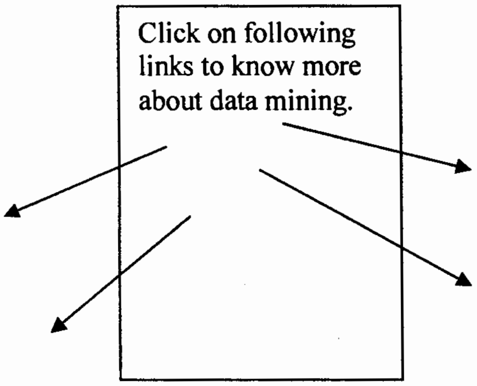


Fig2.2: Generating Base Set of Root Set



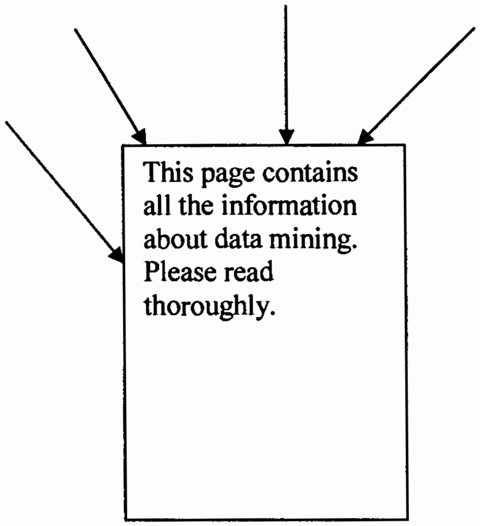

This page contains all the information about data mining. Please read thoroughly.

Figure2.3(a): Hub Page

Figure2.3(b): Authority Page

2.2 Page Rank, HITS and a Unified Framework for Link Analysis

Parry *et al* 12] discuss Page Rank and Hypertext Induced Topic Search (HITS) with mutual reinforcement of hub and authorities. Concept of Co-citation and Co-Reference is discussed. If two distinct web pages pi, pj are co-cited by many other web pages pk as in Figure2.4, pi, pj are likely to be related in some sense. The fact that two distinct webpage's pi; pj co-reference several other webpage's pk) indicates that pi; pj have certain commonality [11].

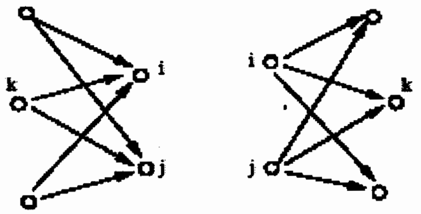


Figure 2.4: Left: web pages pi; pj are co-cited by webpage pk. Right: web pages pi; pj Co-reference webpage pk.

The most important feature of Hypertext Induced Topic Search (HITS) algorithm by Kleinberg *et al* (1998)[6] is the mutual reinforcement between hubs and authorities, while the most important feature of PageRank is the hyperlink weight normalization.

This paper combine concepts of mutual reinforcement and hyperlink weight normalization into a unified framework and introduce three new normalized ranking algorithms within this framework

1) **INORM Rank:** In this case inlinks are normalized using norm.
2) **ONORM Rank**: In this case outlinks are normalized by taking norm.
3) **SNORM Rank:** In this case in-links and out-links are normalized in symmetric fashion.

2.3 The Connectivity Server: fast access to Linkage information on the Web

Monika *et al* [11] describe a system called Connectivity Server that provides linkage information for all pages indexed by the AltaVista search engine. In its basic operation, the server accepts a query consisting of a set L of one or more URLs and returns a list of all pages that point to pages in L (predecessors) and a list of all pages that are pointed from pages in L (successors). More generally the server can produce the entire neighborhood (in the graph theory sense) of L up to a given distance and can include information about all links that exist among pages in the neighborhood. Although some of this information can be retrieved directly from Alta Vista or other search engines, these engines are not optimized for this purpose and the process of constructing the neighbourhood of a given set of pages is slow and laborious. In contrast Connectivity Server needs less than 0.1 ms per result URL. Two applications that use the Connectivity Server: a direct interface that permits fast navigation of the Web via the predecessor/successor relation, and a visualization tool for the neighbourhood of a given set of pages.

2.3.1 Internal organization

- **Initial data structures**

  Representing a small graph is trivial. Representing a graph with 100 millions nodes and close to a billion edges is an engineering challenge.

  We represent the Web as a graph consisting of nodes and directed edges. Each node represents a page and a directed edge from node *A* to node *B* means that page *A* contains a link to page *B*. The set of nodes is stored in an array, each element of the array representing a node. The array index of a node element is the node's *ID*. We represent the set of edges emanating from a node as an adjacency list that is for each node we maintain a list of its successors. In addition, for each node we also maintain an inverted adjacency list that is a list of nodes from which this node is directly accessible, namely its predecessors. Therefore a directed edge from node *A* to node B appears twice in our graph representation, in the adjacency list of *A* and the inverted adjacency

  list of B. This redundancy in representing edges simplifies both forward and backward traversal of edges. To minimize fragmentation, elements of all adjacency lists are stored together in one array called the Outlist. Similarly elements of all inverted adjacency lists are stored in another array called the Inlist. The adjacency and inverted adjacency lists stored in each node are represented as offsets into the Outlist and Inlist arrays respectively. The end of the adjacency list for a node is marked by an entry whose high order bit is set Figure2.5. Thus we can determine the predecessors and the successors of any node very quickly.

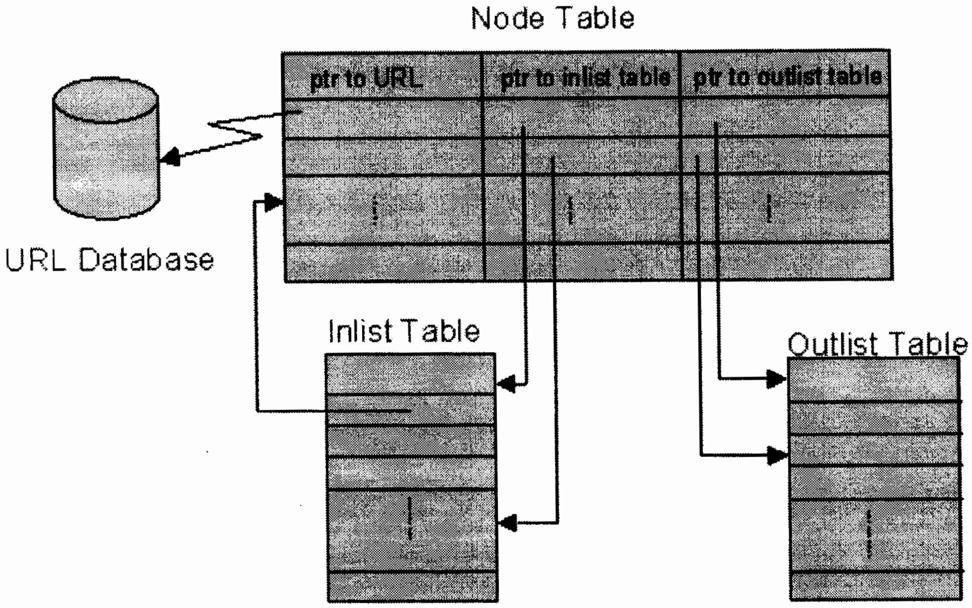


Figure2.5:. Representation of the graph.

A node in the Web-graph has an attached URL. Since URLs are rather long (about 80 bytes on average), storing the full URL within every node in the graph would be quite wasteful. (The storage requirement of a naive implementation would be about 8 gigabytes for 100 million URLs!) Instead the server maintains data structures that represent the *ID* to URL and URL to *ID* mappings.

After a full crawl of the Web, all the URLs that are to be represented in the server are sorted lexicographically. The index of a URL in this sorted list is its initial *ID* (see the discussion of updates below). Then the list of sorted URLs is stored as a delta-encoded text file, that is, each entry is stored as the difference (delta) between the current and previous URL. Since the common prefix between two URLs from the same server is often quite long, this scheme reduces the storage requirements significantly. With the 100 million URLs in our prototype we have seen a 70% reduction in size Figure2.6.

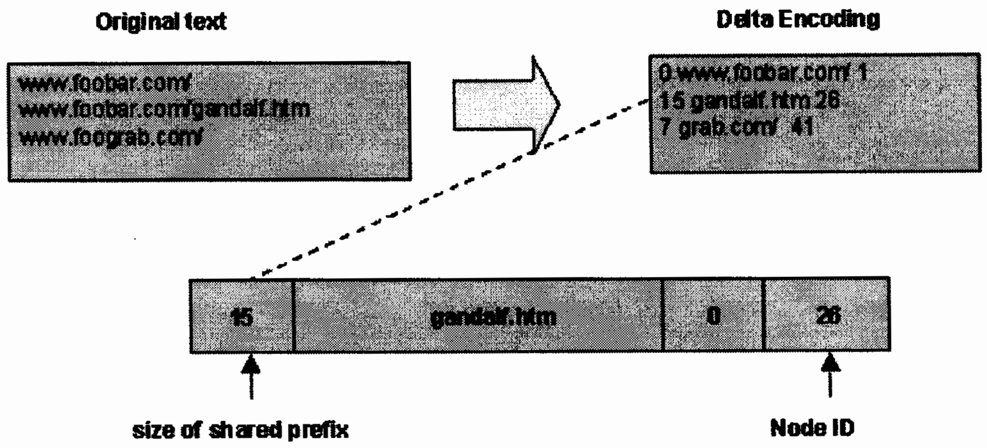


Figure2.6: Delta Encoding the URL's

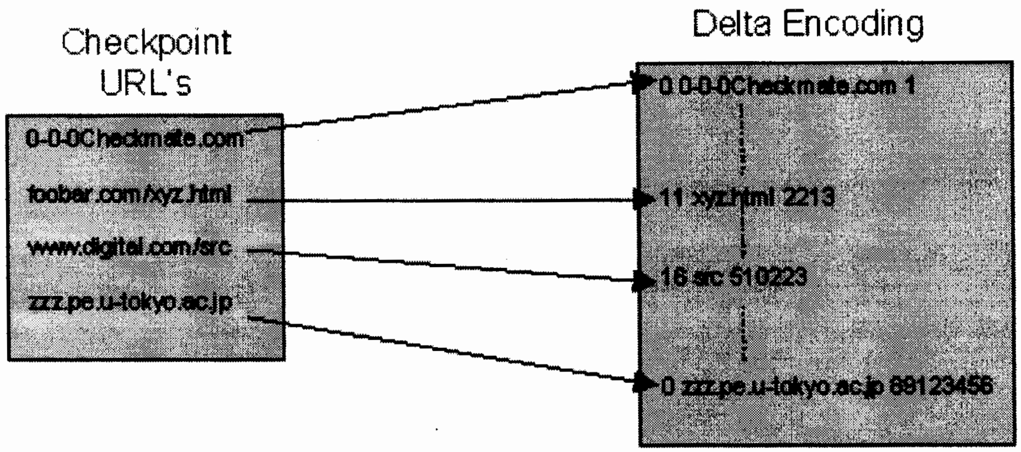


Figure 2.7: Indexing the Delta Encoding

This reduction in storage requirements comes at a price, namely the slowdown of the translation. In order to convert a delta encoded entry back to its complete URL, one needs to start at the first URL and apply all the deltas in sequence until arriving at the URL in question. Author avoids this problem by periodically storing the entire URL instead of the delta encoding. This entry is called a *checkpoint* URL. Therefore to translate a delta encoded URL, we need to apply the deltas starting from the last checkpoint URL rather than the first URL. The cost of the translation can be reduced by increasing the checkpoint frequency Figure2.7. To translate a URL to an internal ID we first search the sorted list of checkpoint URLs to find the closest checkpoint. Then the delta encoded list is searched linearly from that checkpoint URL until the relevant URL is reached. To speed up the reverse translation from internal ID to an URL, the relevant node points directly to the closest checkpoint URL. As before the URL is computed by searching linearly from the checkpoint URL.

- **Updates**

Since their structure is very tight, updates are not simple. Currently their design is to batch all the updates for a day. They view all the updates as a collection of nodes and edges to be added or deleted. All deletions can be done by marking the deleted edges and nodes in a straightforward manner. This requires an extra bit per edge and node. Additions are done as follows.

To allow for additions, They allocate initially larger tables than immediately necessary. For newly added nodes, they maintain a separate structure for the URL to id translation, organized as a string search tree. This tree contains all the newly added nodes and their assigned ID's in the main data structure. To update the Outlist table, the list of new edges is grouped by source. If the new Outlist associated to a node is longer than the old Outlist, space is allocated at the end of the current Outlist table. The update of the Inlist table is done similarly, except that edges are sorted by destination. Eventually the wasted gaps in tables consume too much space, and/or the additional node tree becomes too large and then the entire structure is rebuilt.

2.3.2Performance

The Connectivity Server performs three steps to process queries: translate the URLs in the query to node IDs, explore the Web graph around these nodes and translate the IDs in the result set back to URLs. Thus the time needed to process queries is proportional to the size of the result set. On a 300 MHz Digital Alpha with 4 GB memory, the processing time is approximately 0.1 ms/URL in the result set. Figure 4 shows the timings for 15 different queries where the answer size varies from 1192 to 5734 URLs. As the third step takes up most of the processing time, i.e. 80%. The remainder time is shared equally between steps one and two. Therefore, applications that can work with internal IDs can expect an even faster processing time of about 0.01 ms/URL.

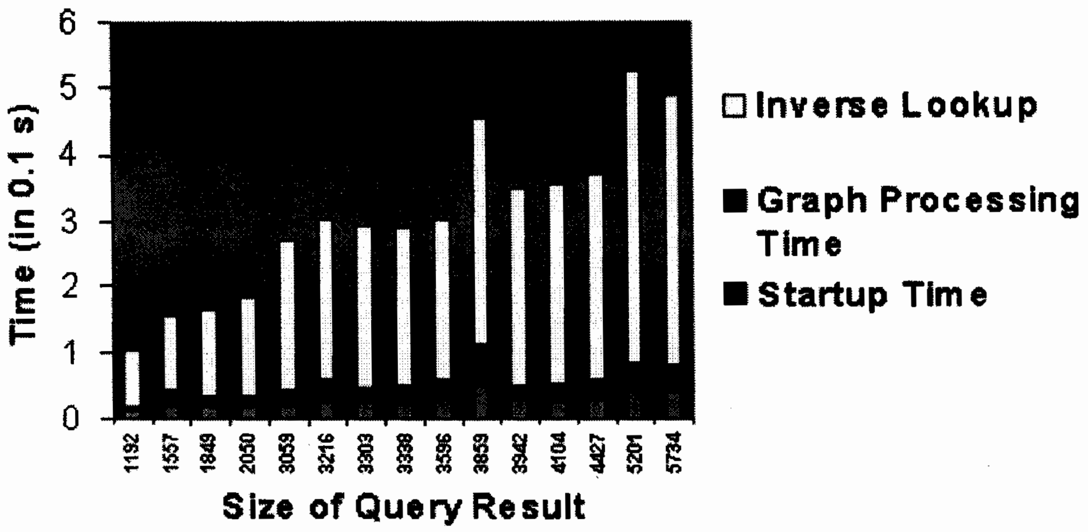


Figure 2.8 Query Processing Time

2.4 Selective Hypertext Induced Topic Search

Mitra *et al* [7]discussed SELHIT algorithm Figure 2.10. That is an improvement over Kleinberg *et al* [6] Hypertext Induced Topic Search algorithm (see Figure 2.9) for answering broad-topic queries and addresses some Hypertext Induced Topic Search (HITS) algorithm problems for example topic drift, Topic Contamination by selectively expanding the root set .

Basically the SELHIT algorithm first calculates the hub and authority scores on root set returned by the search engine and then selects top hubs and authorities for further expansion to get base set. This selective expansion procedure drastically reduces size of the base set to avoid topic drift, as irrelevant pages are not added to the root set. Therefore SELHITS algorithm indeed distills the most important and relevant pages for broad-topic queries.

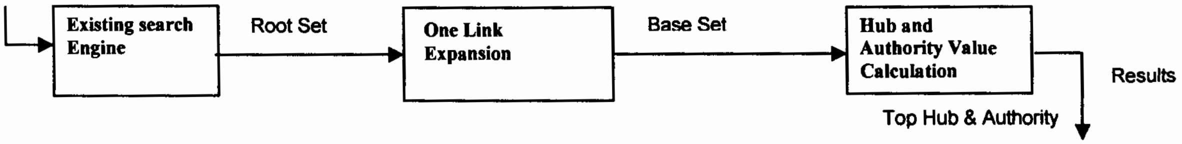


Figure 2.9: HITS Algorithm.

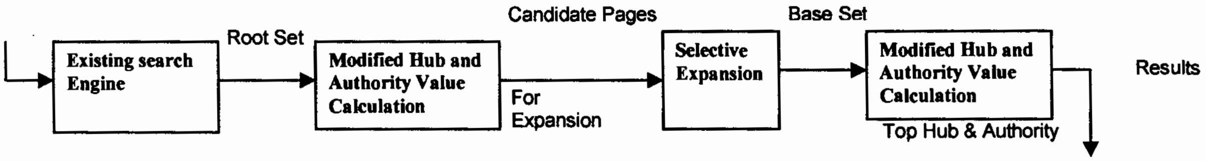


Figure 2.10: SelHit Algorithm.

2.5 Improved Algorithms for Topic Distillation in a Hyperlinked Environment

Bharat *et al* [1] described an approach to augment a connectivity analysis based algorithm with content analysis to find quality documents related to the query topic and described the three problems in Hypertext Induced Topic Search (HITS)[1] connectivity analysis algorithm and presented various algorithms to address those problems. Problems in Hypertext Induced Topic Search (HITS) connectivity analysis algorithm are as follows [1].

2.5.1. Mutually Reinforcing Relationships Between Hosts.

A set of documents on one host point to a single document on a second host. This drives up the hub scores of the documents on the first host and the authority score of the document on the second host. The reverse case, where there is one document on a first host pointing to multiple documents on a second host, creates the same problem. Since the set of documents on each host are authored by a single author or organization, these situations give undue weight to the opinion of one \person.

2.5.2 Automatically Generated Links:

Web documents generated by tools (e.g., Web authoring tools, database conversion tools) often have links that were inserted by the tool. For example, the Hyper news system, which turns USENET News articles into Web pages, automatically inserts a link to the Hypernews Web site. In such cases human's opinion is represented by the link, does not apply.

2.5.3. Non-relevant Nodes:

The neighborhood graph contains documents not relevant to the query topic. If these nodes are well connected, the topic drift problem arises: the most highly ranked authorities and hubs tend not to be about the original topic. For example, when running the algorithm on the query "jaguar and car" the computation drifted to the general topic "car" and returned the home pages of different car manufacturers as top authorities, and lists of car Non-relevant Nodes manufacturers as the best hubs.

This paper presented one connectivity based algorithms imp to address the first problem by giving fractional weights to each edge, basically the imp algorithms is an improvement over Hypertext Induced Topic Search (HITS) algorithm by Kleinberg *et al* (1997)[6] and described some other algorithms to address other two problems. These algorithms are the combination of content analysis using traditional Information Retrieval techniques with improved connectivity analysis algorithm "imp".

2.6 Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text

Chakrabarti *et al* [3] defines Automatic Resource Compilation (ARC). ARC is based on a combination of text and link analysis for distilling authoritative Web resources. The *ARC* algorithm also extends Hypertext Induced Topic Search (HITS) algorithm with textual analysis. ARC computes a distance-2 neighborhood graph and weights edges. The weight of each edge is based on the match between the query terms and the text surrounding the hyperlink in the source document. Both Chakrabarti *et al* [3] and Bharat *et al* [1] studies are different in three way . Bharat *et al*[1] uses an expanded query while Chakrabarti *et al* [3] uses the original query. Bharat *et al* [1] computed the relevancy using the whole document while Chakrabarti *et al* [3] computed the relevancy using the hyperlink text. The weight of an edge in Bharat *et al* [1] is either the relevance of the source document or the target document depending on whether authority or hub scores are being computed while in Chakrabarti *et al* [3] weight of each edge is based on the match between the query terms and the text surrounding the hyperlink .ARC is based on a combination of text and link analysis for distilling authoritative Web resources.

Chakrabarti *et.al*[3] discusses the design and evaluation of an *automatic resource compiler*. An automatic resource compiler is a system which, given a topic that is broad and well-represented on the web, will seek out and return a list of web resources that it considers the most authoritative for that topic. This system is built on an algorithm that performs a local analysis of both text and links to arrive at a "global consensus" of the best resources for the topic. This paper describes a user-study, comparing our resource compiler with commercial, human-compiled/assisted services. When web users seek definitive information on a broad topic, they frequently go to a hierarchical, manually-compiled taxonomy such as Yahoo!, or a human-assisted compilation such as Info seeks. The role of such taxonomy is to provide, for any broad topic, such a resource list with high-quality resources on the topic. The goal of ARC is to automatically compile a resource list on any topic that is broad and well-represented on the web. The ARC has three phases. Three phases of an ARC are shown in Figure 2.12.

1) **Search and growth phase:**

In this phase we get a set of 200 pages and then augment using links to 2-link neighbor hood.

2) **Weighting Phase:**

In this phase we assign to each link (from page *p* to page *q* of the augmented set) positive numerical *weight w (p,q)* that increases with the amount of topic-related text in the vicinity of the href from *p* to q.

3) **Iteration and Reporting Phase:**

This phase is to compute vectors **h** (for hub) and **a** (for authority), with one entry for each page in the augmented set. The entries of the first vector contain scores for the value of each page as a hub, and the second vector describes the value of each page as an authority. Then construct a matrix *W* that contains an entry corresponding to each ordered pair *p,q* of pages in the augmented set. This entry is *w (p,q)* (compute as below)when page *p* points to q, and *0* otherwise. Let *Z* be the matrix transpose of W. Then set the vector **h** equal to 1 initially and iteratively execute the following two steps *k* times.

$$\mathbf{a} = \mathbf{W}\,\mathbf{h}$$

$$\mathbf{h} = \mathbf{Z}\,\mathbf{a}$$

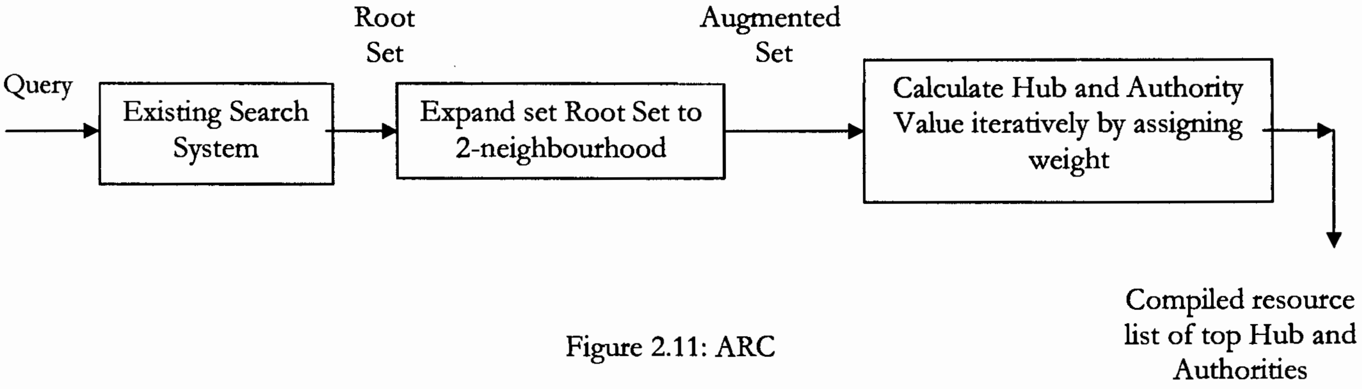

Query
Existing Search System
Root Set
Expand set Root Set to 2-neighbourhood
Augmented Set
Calculate Hub and Authority Value iteratively by assigning weight
Compiled resource list of top Hub and Authorities

Figure 2.11: ARC

3

Problem Definition

Problem Statement.

Bharat *et al* [1] discovered three problems with connectivity analysis as suggested by Kleinberg *et al* [6]

1) Mutually Reinforcing Relationships Between hosts,

2) Automatically Generated Links,

3) Non-relevant Documents and discussed the several techniques for tackling these three scenarios and achieved considerable improvements. However these techniques still contain the following problems.

3.1 Blindly Expansion of Root set

Existing search systems return thousands of results for broad queries, only top few are directly relevant and important for the topic of the query. After adding all pages in one link neighborhood, the size of the base set becomes of the order of a few thousand pages. Most of the pages added are useless and including them in the base set causes the **extra time consumption** because pages in the base are used for further processing for example content analysis.

3.2 Distilling Pure Topic.

Sometime 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 animal. So depending on meaning there will be different topics for the query.

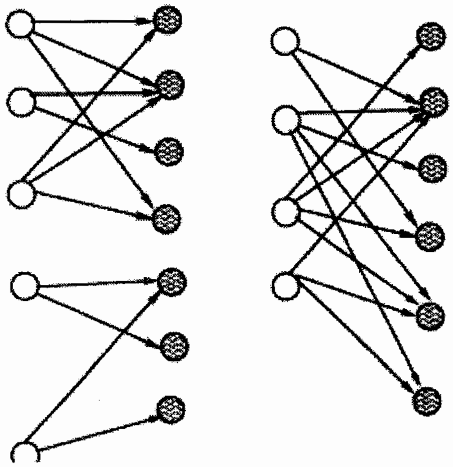


Fig3.1: Multiple Graphs Relating to Different Aspects Of A Single Broad Topic Query

4

Design

DESIGN

Bharat *et al* [1] discussed one purely connectivity based algorithms "imp" and nine content and connectivity based algorithms. We have considered the three content and connectivity base algorithms.

1) med.

2) startmed,

3) maxby10.

These algorithms pruned the irrelevant nodes by computing the Relevance Weights of the nodes in the Neighborhood graph.

The relevance weight of a node equals the similarity of document to the query topic. These algorithms use the documents in the root set to define a query and match every document in the Neighborhood graph against this query.and consider the concatenation of the first 1000 words from each document to be the query, *Q* and compute *similarity* (*Q,D*) by using following formula.

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

w_{iq} = $freq_{iq} \times IDF_i$.

w_{ij} = $freq_{ij} \times IDF_i$.

$freq_{iq}$ = the frequency of the term i in query Q.

$freq_{ij}$ = the frequency of the term i in document D_j.

IDF_i = an estimate of the inverse document frequency
of term i on the World Wide Web.

After getting the relevance weights threshold value is calculated by using relevance weights. In **med** algorithm threshold is the median of the relevance weights of the nodes in Neighborhood graph/base set. In **startmed** threshold is the median of the relevance weights of the nodes in the start/root set. and in **maxby10** threshold is a fixed fraction of the maximum weight of the nodes in Neighborhood graph/base set.. This threshold value is used to pruned irrelevant documents.

4.1 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 as follows Figure 4.1. A start/root set of documents matching the query is fetched from a search engine (say the top 200 matches), This set is blindly expanded by its neighborhood, which is the set of documents that either point to or are pointed by the documents in the root set. The documents in the root set and its neighborhood together form the nodes of the neighborhood graph, the number of nodes in neighborhood graph is called base set. After getting the base set, the algorithms perform content analysis to Pruned irrelevant Nodes from the Neighborhood Graph. Pruning is performed by Computing the Relevance Weights of the Nodes in neighborhood graph and use the relevance weight of a node to decide if it should be eliminated from the graph. This decision is dependent on the thresholds of the relevance weights. Thresholds are picked in one of three ways.

1. **med(Median Weight):** The threshold is the median of the relevance weights of node in Neighborhood graph/base set,
2. **startmed (Start Set Median Weight):** The threshold is the median of the relevance weights of the nodes in the start/root set.
3. **maxby10(Fraction of Maximum Weight):** The threshold is a fixed fraction of the maximum weight. Bharat *et al* [12] used *max*/10. The relevance weight of a node equals the similarity of its node/document to the query topic.

All nodes whose Weights are below a threshold are pruned from base set and resultant pruned set is used for further processing. On the pruned set/graph the connectivity based algorithm "imp" is applied to computes the hub and authority scores for all the nodes in pruned set/graph and call the corresponding algorithms: *med*, *startmed*, and *maxby*10. These algorithms report top hub and authority pages to the user.

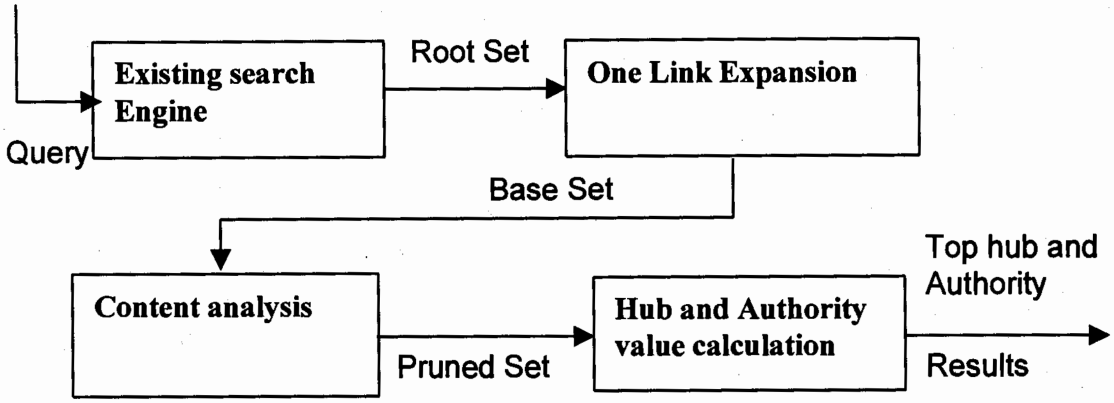


Figure 4.1: Architecture of three Content and Connectivity based Algorithms

Drawbacks:

- The blindly "One Link Expansion" procedure to get base set drastically increases size of the base set. Most of the pages added are useless and including them in the base set causes the **extra time consumption**, as the content analysis is also performed on useless pages as well as useful pages.
- In case of ambiguous queries and queries that have multiple meaning, Search Engine can return the results from multiple topics. Therefore blindly "one link expansion" procedure may causes the **topic contamination**, as one link expansion procedure is performed equally on multiple topic pages and resultant base set contains pages that are from multiple topic. Therefore content analysis is performed on multiple topic pages but the aim of topic distillation is to deliver the results from single topic only.

4.2 SelHITS Algorithm.

SelHITS algorithm by Mitra *et al* [7] begins with the user query. Then it gets a small root set from some existing search system against user query. The root set is of order of few hundred pages related to query topic. Then it calculates hub and authority values on the root set and select top hubs and top authorities pages as candidate Pages for further expansion. Refer to Figure 4.2.This selective expansion procedure of candidate pages drastically reduces size of the base set, as irrelevant pages are not added to the Candidate Pages to get the base set that avoids time consumption, topic contamination problems.

For base set SelHITS repeat the same process that it carried out on the root set. Then it reports top hub and authority pages to the user. Refer to Figure 4.2.

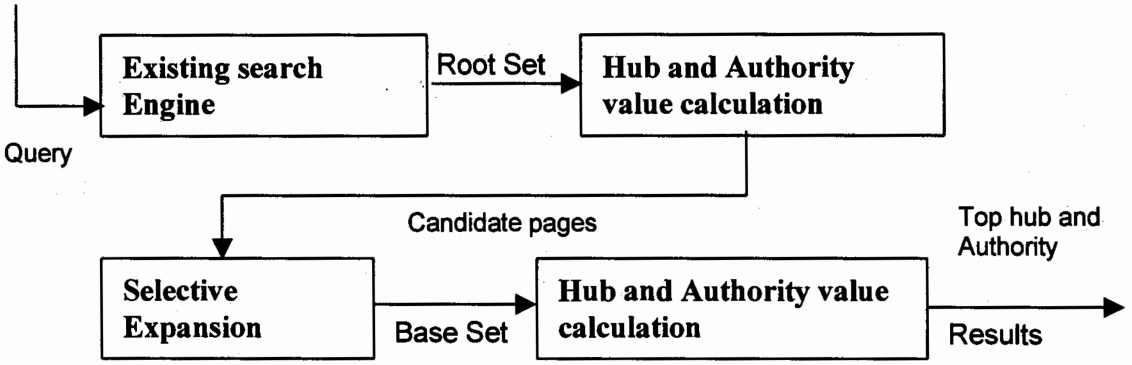


Figure4.2: Architecture of SelHit Algorithm

- The "Selective Expansion " procedure achieves considerable improvement over blindly "one link expansion" procedure, as only relevant pages are expanded instead of expanded all relevant and irrelevant pages Figure 4.1 to get base set. Therefore size of base set decreases. Most of pages added in base set are useful and about to user query topic.
- The "Selective Expansion "Figure 4.2 procedure also addresses problem **topic contamination that** accrues due to blindly" One Link Expansion" procedure to get base, as single topic pages are expanded instead of expanded multiple topic pages to get base set.

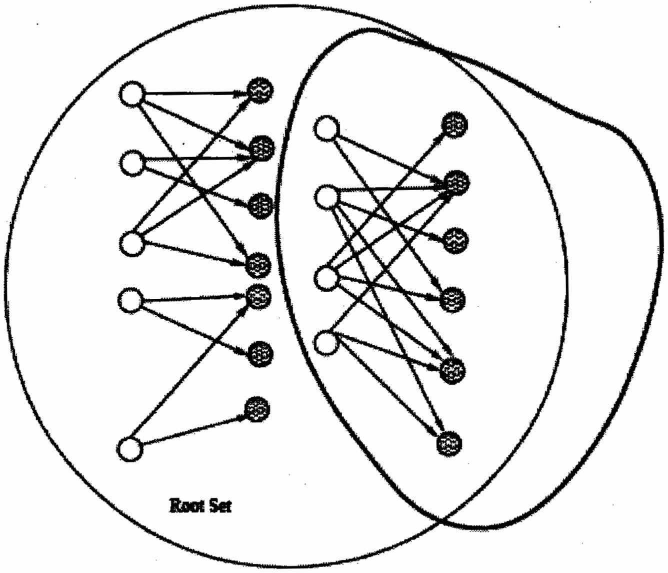


Fig 4.3: Most predominant community selected for expansion

4.3 System Architecture

Our aim is to blend method "SELHIT" algorithm by Mitra et al [7] with three content and connectivity based algorithms of Bharat et al [1]

1)med

2)startmed

3) maxby10 and hope that proposed system will further help to get better results.

In its basic operation, we have interchanged phase "one link expansion" with the two phases "Hub and Authority value calculation" and "Selective Expansion" (Fig 4.2) to address problems **topic contamination** and **extra time consumption.**

The "Hub and Authority value calculation" phase have calculated Hub and Authority values for all pages in root set and have selected top Hub and authority pages as candidate pages and "Selective Expansion" phase have selectively expanded the "candidate pages" to get the base set. This Selective Expansion procedure have decreased size of base set, as irrelevant pages have not been added in base set. Now the base set contains pages related to user query topic and content analysis have been performed only on query related pages that have removed the **extra time consumption.**

In case of ambiguous quires, “Selective Expansion” procedure have expanded candidate pages (user query related pages) from single topic to avoid **topic contamination** because “Hub and Authority calculation“ (Figure 4.3) phase have returned the candidate pages that are from the single topic.

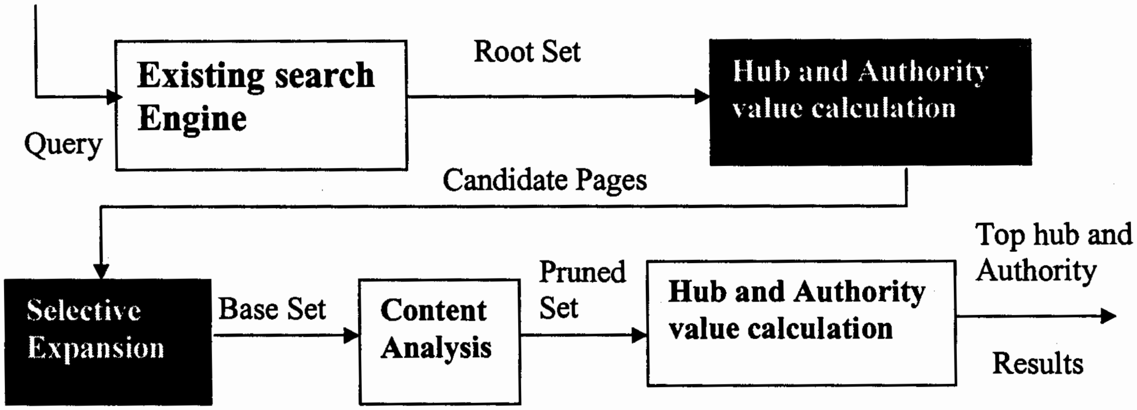


Figure4.4: Architecture of Proposed System.

Main phases of our system will be

- Root Set: Implemented system gets root set from existing search engine against user query.
- Candidate Pages: Implemented system calculates hub and authority values for each page in root set and selects top hub and authority pages as candidate pages.
- Base Set: Implemented system gets base set by the selective expansion of candidate pages. So instead of expanding all the pages in one link neighborhoods, Implemented system expands selective candidate pages only. Thus the resultant base set is much smaller and contains more relevant pages then total expanded root set.

5

Implementation

IMPLEMENTATION

Implementation is an important stage and phase of software lifecycle where the thoughts and ideas are given physical shape. Implementation is a summary description of the noteworthy organization of deliverables. A good implementation approach and strategy leads to successful application or system.

5.1 Technology

The technology infrastructure provides the foundation for the data and application architectures. The infrastructure encompasses the hardware and software that are used to support the application and data. This includes computers, operating systems, networks, telecommunication links, storage technologies and the architecture.

- Our implementation requires windows XP and internet.
- The tool used are Matlab-7 and C#
- This application can run on Pentium IV with at least 2.8GHz speed and high internet speed.

5.1.1 Matlab

MATLAB is a high-performance language for technical computing. It integrates computation, visualization, and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation. Typical uses include Math and computation Algorithm development Data acquisition Modeling, simulation, and prototyping Data analysis, exploration, and visualization Scientific and engineering graphics Application development, including graphical user interface building MATLAB is an interactive system whose basic data element is an array that does not require dimensioning. This allows you to solve many technical computing problems, especially those with matrix and vector formulations, in a fraction of the time it would take to write a program in a scalar no interactive language such as C or Fortran. The name MATLAB stands for matrix laboratory. MATLAB was originally written to provide easy access to matrix software developed by the LINPACK and EISPACK projects. Today, MATLAB engines incorporate the LAPACK and BLAS libraries, embedding the state of the art in software for matrix computation. MATLAB has evolved over a period of years with input from many users. In university environments, it is the standard instructional tool for introductory and advanced courses in mathematics, engineering, and science. In industry, MATLAB is the tool of choice for high-productivity research, development, and analysis. MATLAB features a family of add-on application-specific solutions called toolboxes. Very important to most users of MATLAB, toolboxes allow you to learn and apply specialized technology. Toolboxes are comprehensive collections of MATLAB functions (M-files) that extend the MATLAB environment to solve particular classes of problems. Areas in which toolboxes are available include signal processing, control systems, neural networks, fuzzy logic, wavelets, simulation, and many others.

We have used matlab for connecting search engine google and for extracting URl's from web pages and for designing the different functions that are helpful in getting results.

5.1.2 C#.Net C#.net

C# is an object-oriented programming language developed by Microsoft as part of the .NET initiative and later approved as a standard by ECMA.C# is intended to be a simple, modern, general-purpose, object-oriented programming language. C# is intended to be suitable for writing applications for both hosted and embedded systems, ranging from the very large that use sophisticated operating systems, down to the very small having dedicated functions. By design, C# is the programming language that most directly reflects the underlying Common Language Infrastructure (CLI).

We have used **C#.net** for text analysis of web pages.

5.2 Implementation

Following are the main phases of the system:

- Root Set.
- Candidate Pages.
- Base Set.
- Pruned Set
- Top hub and Authority.

Important functions used in each phase are as follows:

5.2.1 Root set:

The root set is of order of at least two hundred pages related to query topic. The root set is fetched from existing search engine (google) against user query.

- **Webbot().**

This function extracts links related to user query from a existing search engine Google.These links are used for further processing.

5.2.2 Candidate Pages:

The top hub and authority value pages in root set are called candidate pages.

- **match()**

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

- **auth()**

This function finds out authority value of each page in root set.

- **hub()**

This function finds out hub value of each page in root set.

- **Cond()**

This function selects top hub and authority pages as candidate pages.

5.2.3 Base Set:

Base set is the one link expansion of candidate pages.

- **root_exp_out()**

  This function finds out the out links of candidate pages.

- **root_exp_in().**

  This function finds out the in links of candidate pages.

- **Base()**

  This function combines the candidate pages and the pages resulted from the one link expansion of candidate pages as a base set.

5.2.4 Pruned Set

Pruned set is found by pruning irrelevant pages through the content analysis of pages in base set. Content analysis include the following steps.

- **Extraction of plain text from web pages.**
- **Finding the relevance weight of web pages**

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

- **Med()**

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

- **Fraction()**

  This function find out fraction/10 of pages on the bases of relevance weights and pruned the pages whose relevance weight less than fraction/10. The remaining set is a pruned set.

- **Start_ Med()**

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

5.2.5 Top hub and Authority:

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

5.2.5.1 Top hub and Authority through *med* algorithm

- **Med match()**

  This function finds out the link relation among pruned set pages.
- **Med_auth()**

  This function finds out authority value of each page in pruned set and reports the top authority pages to user.
- **Med_hub()**

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

5.2.5.2 Top hub and Authority through *Start Set Median* algorithm

- **start_match()**

  This function finds out the link relation among pruned set pages.
- **start_auth()**

  This function finds out authority value of each page in pruned set and reports the top authority pages to user.
- **start_hub()**

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

5.2.5.3 Top hub and Authority through *Fraction of Maximum Weight* algorithm.

- **frac_match()**

  This function finds out the link relation among pruned set pages.
- **frac_auth()**

  This function finds out authority value of each page in pruned set and reports the top authority pages to user.
- **frac_hub()**

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

6

Testing and Results

TESTING AND RESULTS

Testing is the process of executing software and comparing the observed behavior to the desired behavior and uncovering issues so that hey can be addressed to validate that the team is doing the right thing.

Well-performed tests, initiated early in the software lifecycle, will significantly lower the cost of completing and maintaining the software. It will also greatly reduce the risks or liabilities associated with deploying poor quality software, such as poor user productivity, data entry and calculations errors, and unacceptable functional behavior.

6.1 Purpose

Testing is a process of executing a program with the intent of finding an error. A good test case is the one that has a high probability of finding an as-yet-undiscovered error. The objective should be to design tests that systematically uncover different classes of errors and do so with a minimum amount of time and effort. The purpose of testing is to achieve the following goals:

- ***Primary Goal*:** To discover errors in the software
- ***Secondary Goal*:** Building confidence in the proper operation of the software when testing does not discover errors.
- To verify the interaction between the objects.
- To verify the proper integration of all components of the software.
- To verify that all requirements have been implemented.

6.2 Testing Principles

Following are some of the principles to be kept in mind before doing the testing:

- All tests should be traceable to the requirements.
- Tests should be planned long before testing begins.
- Testing should gradually cover the whole project.
- Exhaustive testing is not possible.
- To be most effective, an independent third party should conduct testing.

6.3 Testing Specification Plan:

The purpose of Test Specification Plan is to give complete instructions on how to perform tests on the software so that they correspond to the requirements of an application.

For quality control to be effective, testing should follow the same pattern throughout. When test cases are changed the result becomes inconsistent with functionality of the software.

A test plan is simply a high level summary of the areas (functionality, elements, regions etc) to be tested, how often these areas are tested, and where in the development or publication process one will test them. A test plan also states the duration testing and list of required resources.

The purpose of software and the constraints under which it has been developed should be understood. The software should illustrate all the characteristics that were initially visualized before its creation and should follow the hoped-for "path" for its success.

6.4 Testing during Design:

It is very important that design document be tested and reviewed in order to develop a clear picture of how the system will work. Following issues were reviewed during design phase:

- Is the design healthy?
- Does the design meet the requirements?
- Is the design complete?
- Is the design implement able?
- How well the design handles error handling?

The design was reviewed several times and up to extent it is tried to make and implement an error free design.

6.5Testing during Coding

The testing strategy that would be used is that first the software as a whole will be tested against the specification to discover the "faults of omission", indicating the part of specification that has not been fulfilled. Then the software would be tested against the implementation to discover "faults of commission", indicating that part of implementation that is faulty.

Some programmers do it as they code, and others wait until the end. Either way, testing is a necessary part of any software development project. Without it, one cannot determine that the software functions correctly.

6.6Testing Of "*Improved algorithms for topic distillation using SelHITS*":

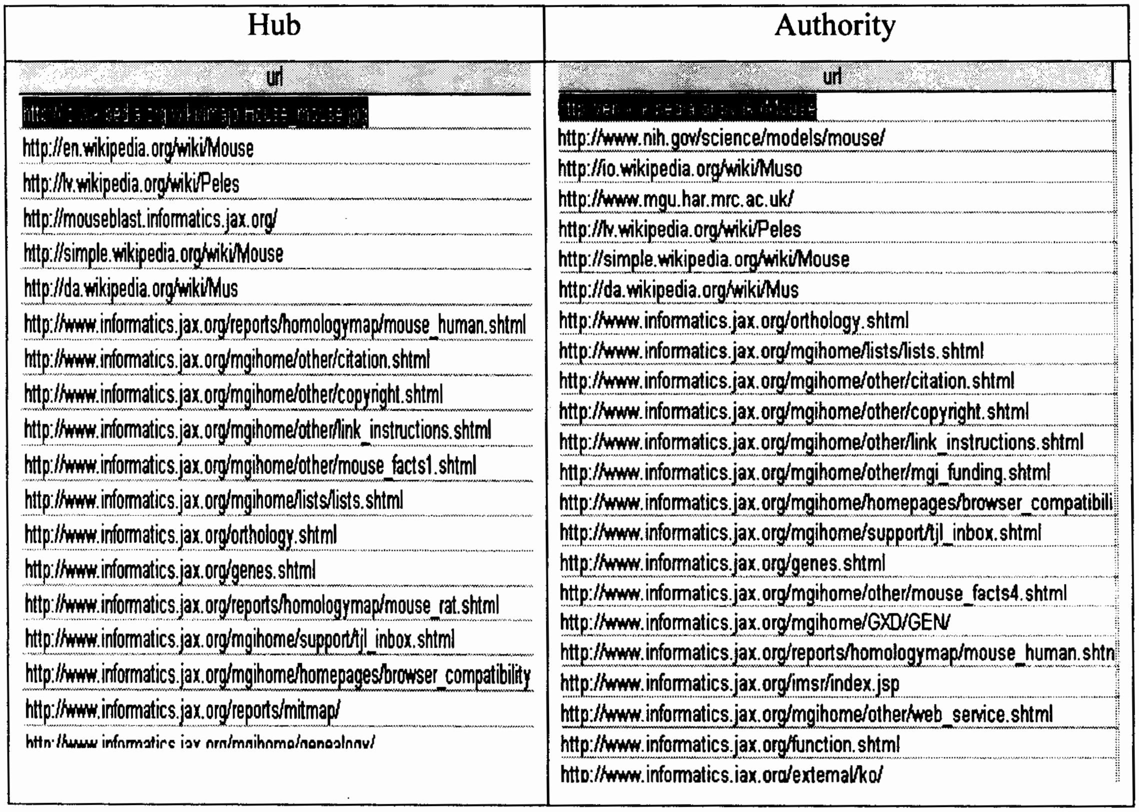
The testing of "Improved algorithms for topic distillation using SelHITS" is undergone through all stages of black box testing and to extent white box testing. The system is reviewed to see whether the objectives of the system are accomplished or not. A major factor considered during system evaluation is to evaluate the system with the perspective of queries entered by users.

The sample tests performed on our project "Improved algorithms for topic distillation using SelHITS" are performed on the following queries:

1) Mouse
2) Windows

6.6.1Mouse

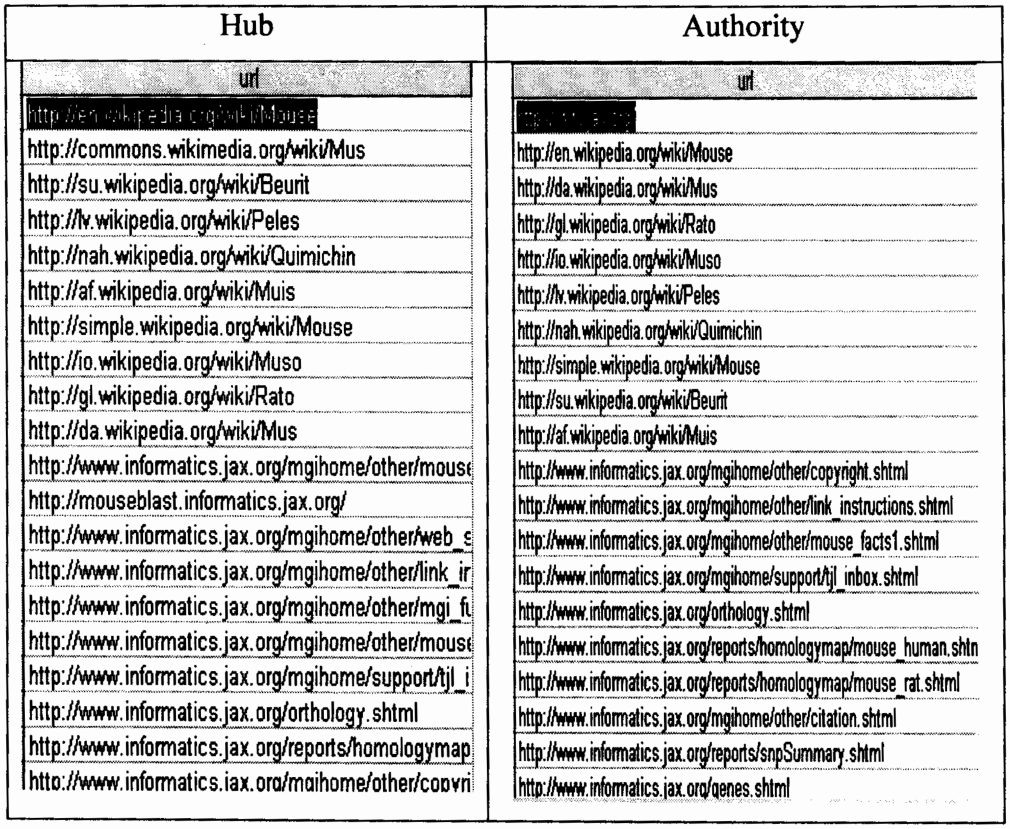
Table 6.1: Top Hub and Authority for Med algorithm



Hub	Authority
url	url
[illegible]	http://en.wikipedia.org/wiki/Mouse
http://en.wikipedia.org/wiki/Mouse	http://www.nih.gov/science/models/mouse/
http://lv.wikipedia.org/wiki/Peles	http://io.wikipedia.org/wiki/Muso
http://mouseblast.informatics.jax.org/	http://www.mgu.har.mrc.ac.uk/
http://simple.wikipedia.org/wiki/Mouse	http://lv.wikipedia.org/wiki/Peles
http://da.wikipedia.org/wiki/Mus	http://simple.wikipedia.org/wiki/Mouse
http://www.informatics.jax.org/reports/homologymap/mouse_human.shtml	http://da.wikipedia.org/wiki/Mus
http://www.informatics.jax.org/mgihome/other/citation.shtml	http://www.informatics.jax.org/orthology.shtml
http://www.informatics.jax.org/mgihome/other/copyright.shtml	http://www.informatics.jax.org/mgihome/lists/lists.shtml
http://www.informatics.jax.org/mgihome/other/link_instructions.shtml	http://www.informatics.jax.org/mgihome/other/citation.shtml
http://www.informatics.jax.org/mgihome/other/mouse_facts1.shtml	http://www.informatics.jax.org/mgihome/other/copyright.shtml
http://www.informatics.jax.org/mgihome/lists/lists.shtml	http://www.informatics.jax.org/mgihome/other/link_instructions.shtml
http://www.informatics.jax.org/orthology.shtml	http://www.informatics.jax.org/mgihome/other/mgi_funding.shtml
http://www.informatics.jax.org/genes.shtml	http://www.informatics.jax.org/mgihome/homepages/browser_compatibili
http://www.informatics.jax.org/reports/homologymap/mouse_rat.shtml	http://www.informatics.jax.org/mgihome/support/tjl_inbox.shtml
http://www.informatics.jax.org/mgihome/support/tjl_inbox.shtml	http://www.informatics.jax.org/genes.shtml
http://www.informatics.jax.org/mgihome/homepages/browser_compatibility	http://www.informatics.jax.org/mgihome/other/mouse_facts4.shtml
http://www.informatics.jax.org/reports/mitmap/	http://www.informatics.jax.org/mgihome/GXD/GEN/
http://www.informatics.jax.org/mgihome/genealogy/	http://www.informatics.jax.org/reports/homologymap/mouse_human.shtn
	http://www.informatics.jax.org/imsr/index.jsp
	http://www.informatics.jax.org/mgihome/other/web_service.shtml
	http://www.informatics.jax.org/function.shtml
	http://www.informatics.jax.org/external/ko/

The table 6.1 shows the top hub and authority URL's against the user query "Mouse" by using Med algorithm. These top hub and authority pages contains majority of the URL's related to the single aspect of "Mouse" query.

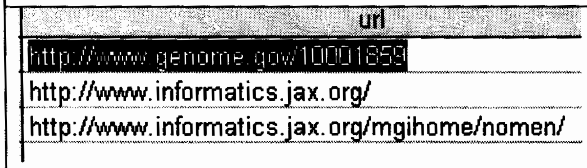
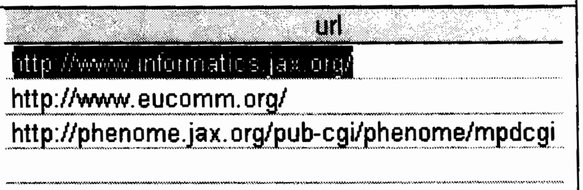
Table 6.2: Top Hub and Authority for Max by 10 algorithm



Hub	Authority
url	url
http://en.wikipedia.org/wiki/Mouse	[illegible]
http://commons.wikimedia.org/wiki/Mus	http://en.wikipedia.org/wiki/Mouse
http://su.wikipedia.org/wiki/Beurit	http://da.wikipedia.org/wiki/Mus
http://lv.wikipedia.org/wiki/Peles	http://gl.wikipedia.org/wiki/Rato
http://nah.wikipedia.org/wiki/Quimichin	http://io.wikipedia.org/wiki/Muso
http://af.wikipedia.org/wiki/Muis	http://lv.wikipedia.org/wiki/Peles
http://simple.wikipedia.org/wiki/Mouse	http://nah.wikipedia.org/wiki/Quimichin
http://io.wikipedia.org/wiki/Muso	http://simple.wikipedia.org/wiki/Mouse
http://gl.wikipedia.org/wiki/Rato	http://su.wikipedia.org/wiki/Beurit
http://da.wikipedia.org/wiki/Mus	http://af.wikipedia.org/wiki/Muis
http://www.informatics.jax.org/mgihome/other/mouse	http://www.informatics.jax.org/mgihome/other/copyright.shtml
http://mouseblast.informatics.jax.org/	http://www.informatics.jax.org/mgihome/other/link_instructions.shtml
http://www.informatics.jax.org/mgihome/other/web_s	http://www.informatics.jax.org/mgihome/other/mouse_facts1.shtml
http://www.informatics.jax.org/mgihome/other/link_ir	http://www.informatics.jax.org/mgihome/support/tjl_inbox.shtml
http://www.informatics.jax.org/mgihome/other/mgi_fu	http://www.informatics.jax.org/orthology.shtml
http://www.informatics.jax.org/mgihome/other/mouse	http://www.informatics.jax.org/reports/homologymap/mouse_human.shtn
http://www.informatics.jax.org/mgihome/support/tjl_i	http://www.informatics.jax.org/reports/homologymap/mouse_rat.shtml
http://www.informatics.jax.org/orthology.shtml	http://www.informatics.jax.org/mgihome/other/citation.shtml
http://www.informatics.jax.org/reports/homologymap	http://www.informatics.jax.org/reports/snpSummary.shtml
http://www.informatics.jax.org/mgihome/other/copyri	http://www.informatics.jax.org/genes.shtml

The table 6.2 shows the top hub and authority URL's against the user query "Mouse" by using Max by 10 algorithms. These top hub and authority pages contains majority of the URL's related to the single aspect of "Mouse" query.

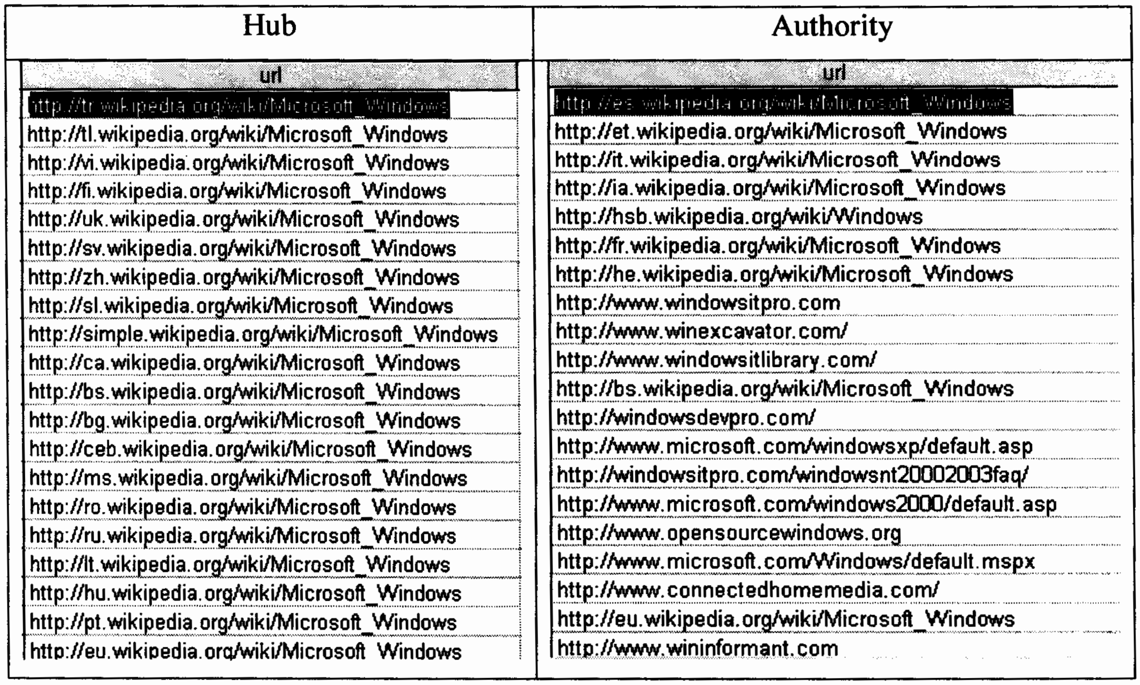
Table 6.3: Top Hub and Authority for Start Med algorithm

Hub	Authority
 url http://www.genome.gov/10001859 http://www.informatics.jax.org/ http://www.informatics.jax.org/mgihome/nomen/	 url http://www.informatics.jax.org/ http://www.eucomm.org/ http://phenome.jax.org/pub-cgi/phenome/mpdcgi

The table 6.3 shows the top hub and authority URL's against the user query "Mouse" by using Start Med algorithms. These top hub and authority pages contains majority of the URL's related to the single aspect of "Mouse" query.

6.6.2 Windows

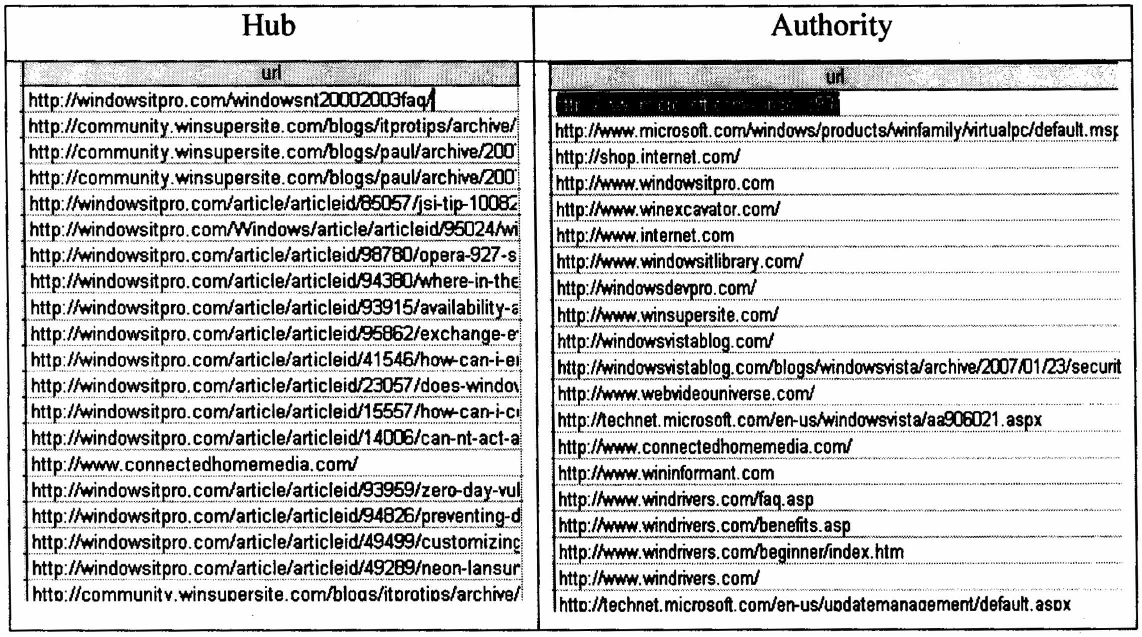
Table 6.4: Top Hub and Authority for med algorithm



Hub	Authority
url	url
http://tr.wikipedia.org/wiki/Microsoft_Windows	http://es.wikipedia.org/wiki/Microsoft_Windows
http://tl.wikipedia.org/wiki/Microsoft_Windows	http://et.wikipedia.org/wiki/Microsoft_Windows
http://vi.wikipedia.org/wiki/Microsoft_Windows	http://it.wikipedia.org/wiki/Microsoft_Windows
http://fi.wikipedia.org/wiki/Microsoft_Windows	http://ia.wikipedia.org/wiki/Microsoft_Windows
http://uk.wikipedia.org/wiki/Microsoft_Windows	http://hsb.wikipedia.org/wiki/Windows
http://sv.wikipedia.org/wiki/Microsoft_Windows	http://fr.wikipedia.org/wiki/Microsoft_Windows
http://zh.wikipedia.org/wiki/Microsoft_Windows	http://he.wikipedia.org/wiki/Microsoft_Windows
http://sl.wikipedia.org/wiki/Microsoft_Windows	http://www.windowsitpro.com
http://simple.wikipedia.org/wiki/Microsoft_Windows	http://www.winexcavator.com/
http://ca.wikipedia.org/wiki/Microsoft_Windows	http://www.windowsitlibrary.com/
http://bs.wikipedia.org/wiki/Microsoft_Windows	http://bs.wikipedia.org/wiki/Microsoft_Windows
http://bg.wikipedia.org/wiki/Microsoft_Windows	http://windowsdevpro.com/
http://ceb.wikipedia.org/wiki/Microsoft_Windows	http://www.microsoft.com/windowsxp/default.asp
http://ms.wikipedia.org/wiki/Microsoft_Windows	http://windowsitpro.com/windowsnt20002003faq/
http://ro.wikipedia.org/wiki/Microsoft_Windows	http://www.microsoft.com/windows2000/default.asp
http://ru.wikipedia.org/wiki/Microsoft_Windows	http://www.opensourcewindows.org
http://lt.wikipedia.org/wiki/Microsoft_Windows	http://www.microsoft.com/Windows/default.mspx
http://hu.wikipedia.org/wiki/Microsoft_Windows	http://www.connectedhomemedia.com/
http://pt.wikipedia.org/wiki/Microsoft_Windows	http://eu.wikipedia.org/wiki/Microsoft_Windows
http://eu.wikipedia.org/wiki/Microsoft_Windows	http://www.wininformant.com

The table 6.4 shows the top hub and authority URL's against the user query "Windows" by using Start Med algorithms. These top hub and authority pages contains majority of the URL's related to the single aspect of "Windows" query.

Table 6.5: Top Hub and Authority for Max by 10 algorithm



Hub	Authority
url	url
http://windowsitpro.com/windowsnt20002003faq/	[illegible]
http://community.winsupersite.com/blogs/itprotips/archive/	http://www.microsoft.com/windows/products/winfamily/virtualpc/default.msp
http://community.winsupersite.com/blogs/paul/archive/200	http://shop.internet.com/
http://community.winsupersite.com/blogs/paul/archive/200	http://www.windowsitpro.com
http://windowsitpro.com/article/articleid/85057/jsi-tip-1008	http://www.winexcavator.com/
http://windowsitpro.com/Windows/article/articleid/95024/wi	http://www.internet.com
http://windowsitpro.com/article/articleid/98780/opera-927-s	http://www.windowsitlibrary.com/
http://windowsitpro.com/article/articleid/94380/where-in-the	http://windowsdevpro.com/
http://windowsitpro.com/article/articleid/93915/availability-a	http://www.winsupersite.com/
http://windowsitpro.com/article/articleid/95862/exchange-e	http://windowsvistablog.com/
http://windowsitpro.com/article/articleid/41546/how-can-i-e	http://windowsvistablog.com/blogs/windowsvista/archive/2007/01/23/securit
http://windowsitpro.com/article/articleid/23057/does-windo	http://www.webvideouniverse.com/
http://windowsitpro.com/article/articleid/15557/how-can-i-c	http://technet.microsoft.com/en-us/windowsvista/aa906021.aspx
http://windowsitpro.com/article/articleid/14006/can-nt-act-a	http://www.connectedhomemedia.com/
http://www.connectedhomemedia.com/	http://www.wininformant.com
http://windowsitpro.com/article/articleid/93959/zero-day-vul	http://www.windrivers.com/faq.asp
http://windowsitpro.com/article/articleid/94826/preventing-d	http://www.windrivers.com/benefits.asp
http://windowsitpro.com/article/articleid/49499/customizing	http://www.windrivers.com/beginner/index.htm
http://windowsitpro.com/article/articleid/49289/neon-lansur	http://www.windrivers.com/
http://community.winsupersite.com/blogs/itprotips/archive/	http://technet.microsoft.com/en-us/updatemanagement/default.aspx

The table 6.5 shows the top hub and authority URL's against the user query "Windows" by using Start Med algorithms. These top hub and authority pages contains majority of the URL's related to the single aspect of "Windows" query.

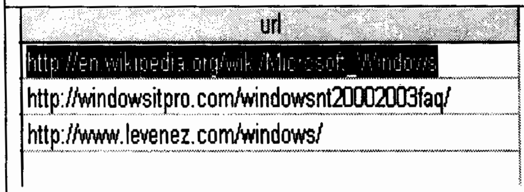
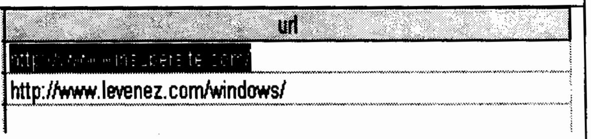
Hub	Authority
 url http://en.wikipedia.org/wiki/Microsoft_Windows http://windowsitpro.com/windowsnt20002003faq/ http://www.levenez.com/windows/	 url http://www.winsupersite.com/ http://www.levenez.com/windows/

Table 6.6: Top Hub and Authority for Start Med algorithm

The table 6.6 shows the top hub and authority URL's against the user query "Windows" by using Start Med algorithms. These top hub and authority pages contains majority of the URL's related to the single aspect of "Windows" query.

6.7Analysis of Results

From the obtained results, we have successfully achieved improved results. The analysis of results show that in case we blindly expand the root set of content and connectivity based algorithms, we get large amount of irrelevant pages .The majority of the pages obtained donot fulfill user needs.

In the implemented system we have selectively expand the root set, the selective expansion helps us in giving most appropriate and relevant pages.The selective expansion help us in solving the problem of topic drift in broad queries.

The first query that we select to test our system is Mouse. There are two interpretations for this broad query one is, animal mouse and other is, computer mouse. Our system has improved the results by giving top hub and top authorities as compare to previous algorithms, related to one aspect of mouse i.e animal mouse and remove the problem of topic contamination and topic drift to some extent than the previous ones. By running the same query using previous algorithm one will get the top hub and top authority related to different interpretations of the same query. Some top hubs and authorities are animal mouse and some are of computer mouse. The one who is firing the query 'Mouse'; it has more probability that he requires the informative pages related animal mouse. The existing search engine algorithms mostly consider it as computer mouse, which is not the Required interpretation for this query as the device use by computer users is basically computer mouse.

The achieved results have reduced the time consumption because we have selectively expanding the base set. When we have blindly expanded the root set, the base set drastically increased to 7000 pages, majority of which are irrelevant pages and give rise to the problem of topic contamination and topic drift.

Similar is the case , when we test 'Windows' and 'Gates' query.

7

Conclusion and Future Works

CONCLUSION AND FUTURE WORKS

7.1 Conclusion

In this thesis we have blended the "SELHIT" algorithm by Mitra et al[7] with three content and connectivity based algorithms by Bharat et al [11]:

1) Med
2) Startmed
3) Maxby10

The implemented system is successful in giving the improved desired results.

Previously the blind "One Link Expansion" procedure to get base set drastically increases size of the base set. Most of the pages added are useless and including them in the base set causes the extra time consumption, as the content analysis is also performed on useless pages as well as useful pages. By the blending of SelHITS with content and connectivity based algorithm we have distilled pure topic related to query to some extent as shown from our results. Selective expansion has also reduced extra time consumption and large amount of hyperlinks which are of no use to user are now not considered if they are not required.

The topic contamination is removed to some extent in case of broad topic queries. By implementing this technique now user will get results related to single aspect of query to some extent.

7.2 Future work

1) In future we want to design a search engine that is based on the improved algorithms and SelHITS.We hope that the proposed search engine will give better results than existing search engines.

2) We also want to further improve these connectivity and content based algorithms by expanding 2-neighbour hood. This is helpful in getting more number of relevant pages according to query.

3) We want to improve the remaining six algorithms by Bharat et.al [1] by applying SelHITS as we have done in these algorithms.

Appendix-A

List of Abbreviations

Appendix A

DEFINITION OF TERMS

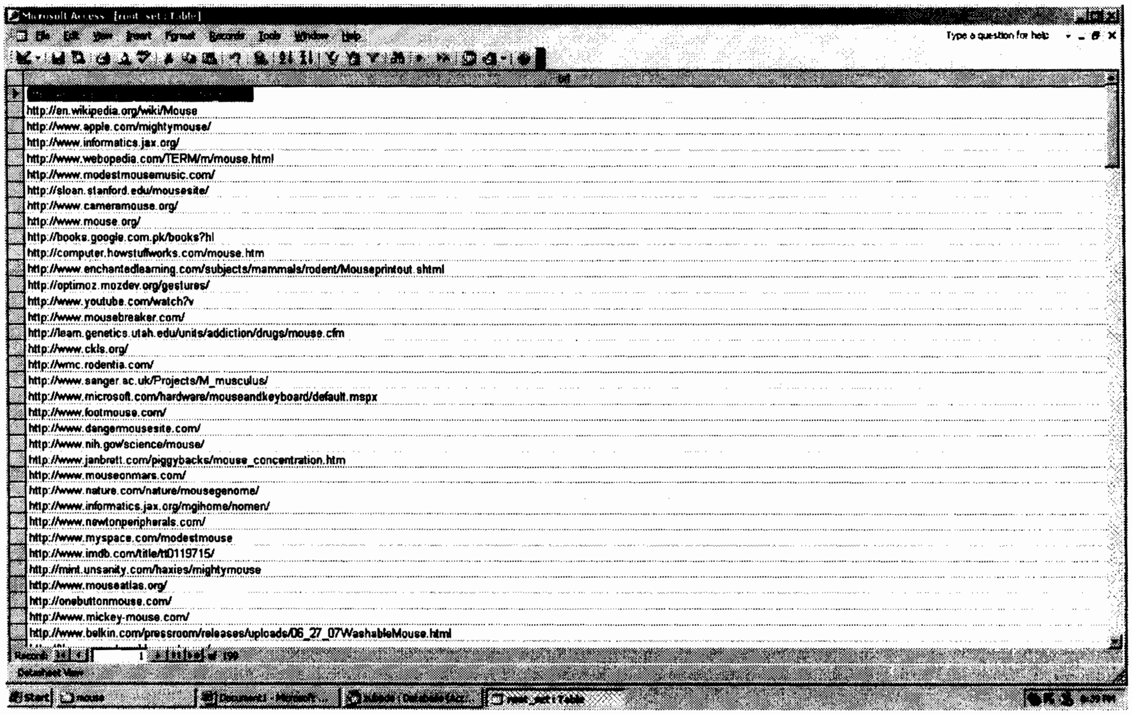
Abbreviations	**Full Form**
HITS	Hypertext Induced Topic Search
SelHITS	Selective Hypertext Induced Topic Search
H	Hub
A	Authority
SALSA	Stochastic Algorithm for Link Structure Analysis
URL	Universal Resource Locator
HTML	Hypertext Induced Topic Search
WWW	World Wide Web
N/W	Network
ARC	Automatic Resource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

Appendix-B

Screen Shots

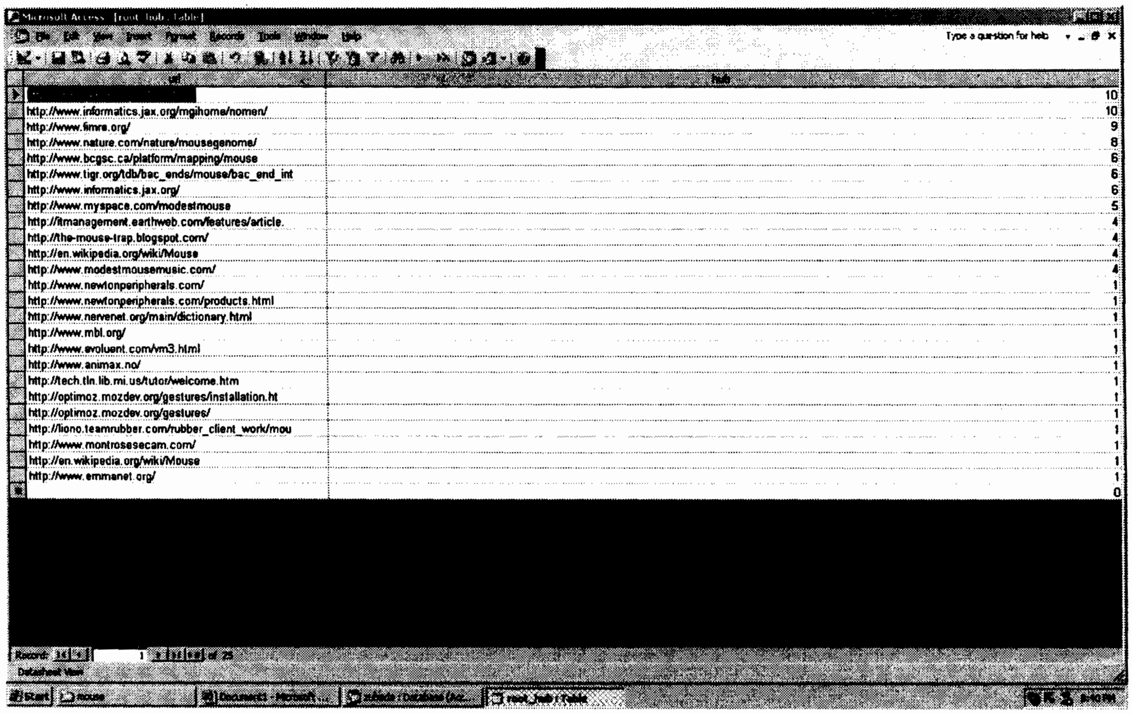
Appendix B

SCREEN SHOTS


Microsoft Access - [root_set : Table]
File Edit View Insert Format Records Tools Window Help
Type a question for help
Datasheet View
Start
mouse

```
http://en.wikipedia.org/wiki/Mouse
http://www.apple.com/mightymouse/
http://www.informatics.jax.org/
http://www.webopedia.com/TERM/m/mouse.html
http://www.modestmousemusic.com/
http://sloan.stanford.edu/mousesite/
http://www.cameramouse.org/
http://www.mouse.org/
http://books.google.com.pk/books?hl
http://computer.howstuffworks.com/mouse.htm
http://www.enchantedlearning.com/subjects/mammals/rodent/Mouseprintout.shtml
http://optimoz.mozdev.org/gestures/
http://www.youtube.com/watch?v
http://www.mousebreaker.com/
http://learn.genetics.utah.edu/units/addiction/drugs/mouse.cfm
http://www.ckls.org/
http://wmc.rodentia.com/
http://www.sanger.ac.uk/Projects/M_musculus/
http://www.microsoft.com/hardware/mouseandkeyboard/default.mspx
http://www.footmouse.com/
http://www.dangermousesite.com/
http://www.nih.gov/science/mouse/
http://www.janbrett.com/piggybacks/mouse_concentration.htm
http://www.mouseonmars.com/
http://www.nature.com/nature/mousegenome/
http://www.informatics.jax.org/mgihome/nomen/
http://www.newtonperipherals.com/
http://www.myspace.com/modestmouse
http://www.imdb.com/title/tt0119715/
http://mint.unsanity.com/haxies/mightymouse
http://www.mouseatlas.org/
http://onebuttonmouse.com/
http://www.mickey-mouse.com/
http://www.belkin.com/pressroom/releases/uploads/06_27_07WashableMouse.html
Record: 1 of 199
```

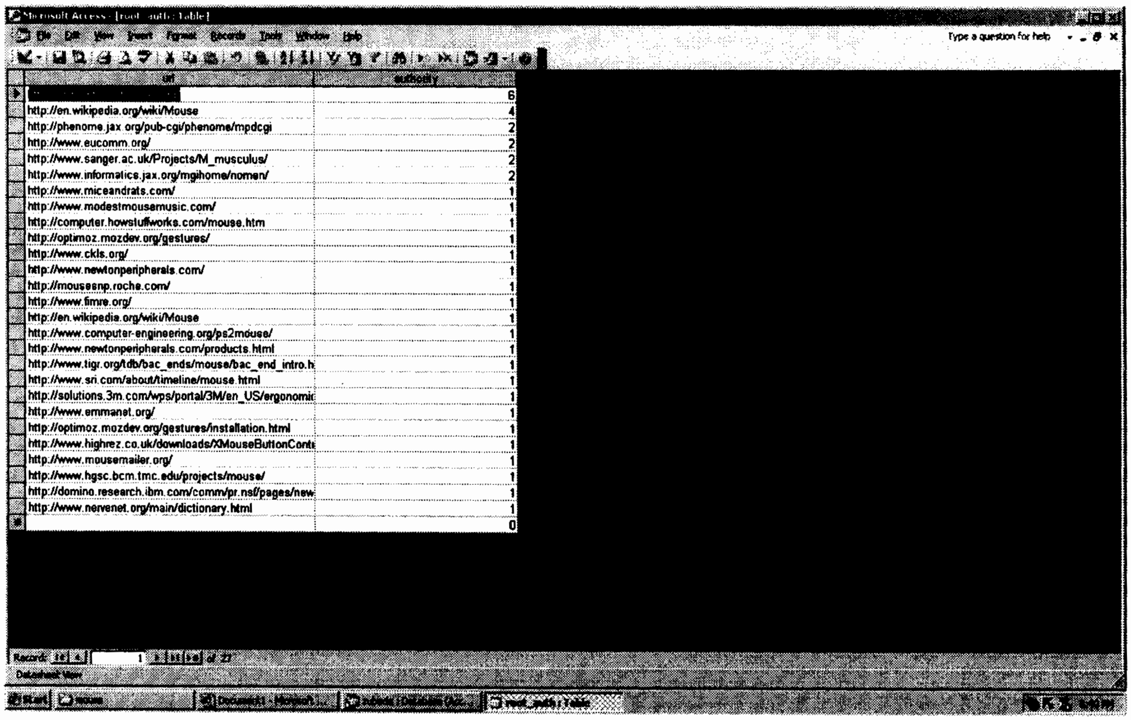
Fig B-1: Root Set



Microsoft Access - [root_hub : Table]
File Edit View Insert Format Records Tools Window Help
Type a question for help
Record: 1 of 25
Datasheet View
Start
mouse
Document1 - Microsoft ...
root_hub : Table

url	hub
[illegible]	10
http://www.informatics.jax.org/mgihome/nomen/	10
http://www.fimre.org/	9
http://www.nature.com/nature/mousegenome/	8
http://www.bcgsc.ca/platform/mapping/mouse	6
http://www.tigr.org/tdb/bac_ends/mouse/bac_end_int	6
http://www.informatics.jax.org/	6
http://www.myspace.com/modestmouse	5
http://itmanagement.earthweb.com/features/article.	4
http://the-mouse-trap.blogspot.com/	4
http://en.wikipedia.org/wiki/Mouse	4
http://www.modestmousemusic.com/	4
http://www.newtonperipherals.com/	1
http://www.newtonperipherals.com/products.html	1
http://www.nervenet.org/main/dictionary.html	1
http://www.mbl.org/	1
http://www.evoluent.com/vm3.html	1
http://www.animax.no/	1
http://tech.tln.lib.mi.us/tutor/welcome.htm	1
http://optimoz.mozdev.org/gestures/installation.ht	1
http://optimoz.mozdev.org/gestures/	1
http://liono.teamrubber.com/rubber_client_work/mou	1
http://www.montrosesecam.com/	1
http://en.wikipedia.org/wiki/Mouse	1
http://www.emmanet.org/	1
	0

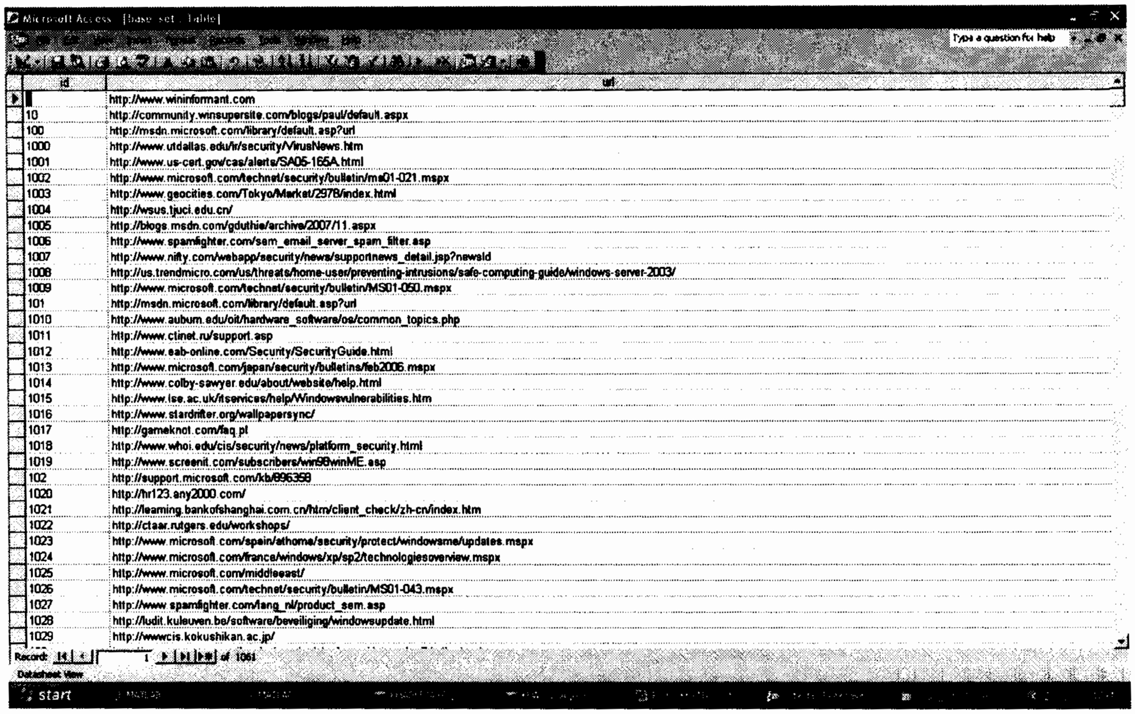
Fig B-2: Root_Hub



Microsoft Access - [root_auth : Table]
File Edit View Insert Format Records Tools Window Help
Type a question for help
Record: 1 of 27
Datasheet View
Start
mouse
Document1 - Microsoft ...
root_auth : Table

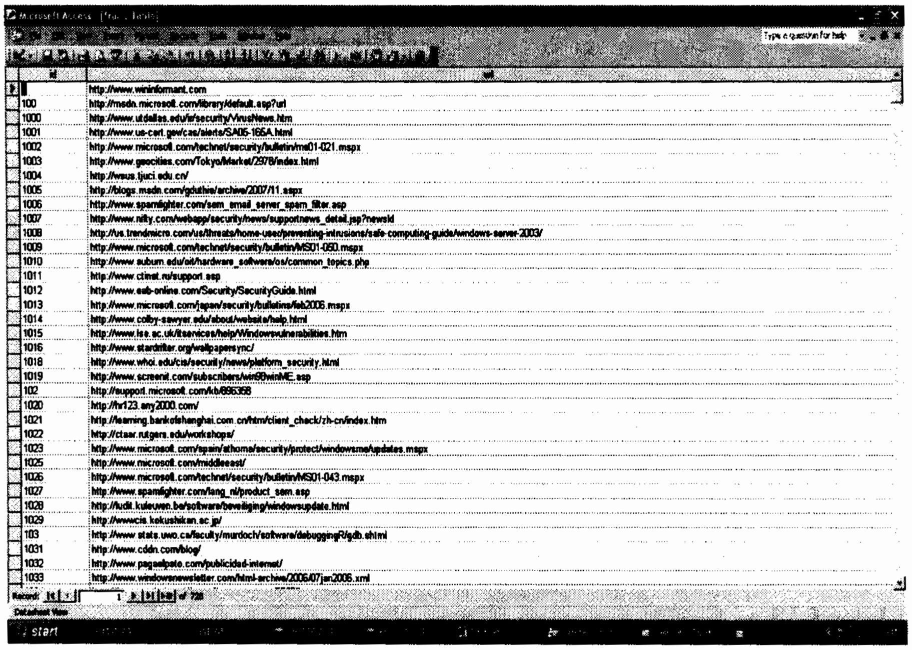
url	authority
[illegible]	6
http://en.wikipedia.org/wiki/Mouse	4
http://phenome.jax.org/pub-cgi/phenome/mpdcgi	2
http://www.eucomm.org/	2
http://www.sanger.ac.uk/Projects/M_musculus/	2
http://www.informatics.jax.org/mgihome/nomen/	2
http://www.miceandrats.com/	1
http://www.modestmousemusic.com/	1
http://computer.howstuffworks.com/mouse.htm	1
http://optimoz.mozdev.org/gestures/	1
http://www.ckls.org/	1
http://www.newtonperipherals.com/	1
http://mousesnp.roche.com/	1
http://www.fimre.org/	1
http://en.wikipedia.org/wiki/Mouse	1
http://www.computer-engineering.org/ps2mouse/	1
http://www.newtonperipherals.com/products.html	1
http://www.tigr.org/tdb/bac_ends/mouse/bac_end_intro.h	1
http://www.sri.com/about/timeline/mouse.html	1
http://solutions.3m.com/wps/portal/3M/en_US/ergonomic	1
http://www.emmanet.org/	1
http://optimoz.mozdev.org/gestures/installation.html	1
http://www.highrez.co.uk/downloads/XMouseButtonConti	1
http://www.mousemailer.org/	1
http://www.hgsc.bcm.tmc.edu/projects/mouse/	1
http://domino.research.ibm.com/comm/pr.nsf/pages/new	1
http://www.nervenet.org/main/dictionary.html	1
	0

Fig B-3:Root_Auth



id	url
	http://www.wininformant.com
10	http://community.winsupersite.com/blogs/paul/default.aspx
100	http://msdn.microsoft.com/library/default.asp?url
1000	http://www.utdallas.edu/ir/security/VirusNews.htm
1001	http://www.us-cert.gov/cas/alerts/SA05-165A.html
1002	http://www.microsoft.com/technet/security/bulletin/ms01-021.mspx
1003	http://www.geocities.com/Tokyo/Market/2978/index.html
1004	http://wsus.tjuci.edu.cn/
1005	http://blogs.msdn.com/gduthie/archive/2007/11.aspx
1006	http://www.spamfighter.com/sem_email_server_spam_filter.asp
1007	http://www.nifty.com/webapp/security/news/supportnews_detail.jsp?newsId
1008	http://us.trendmicro.com/us/threats/home-user/preventing-intrusions/safe-computing-guide/windows-server-2003/
1009	http://www.microsoft.com/technet/security/bulletin/MS01-050.mspx
101	http://msdn.microsoft.com/library/default.asp?url
1010	http://www.auburn.edu/oit/hardware_software/os/common_topics.php
1011	http://www.ctinet.ru/support.asp
1012	http://www.eab-online.com/Security/SecurityGuide.html
1013	http://www.microsoft.com/japan/security/bulletins/feb2006.mspx
1014	http://www.colby-sawyer.edu/about/website/help.html
1015	http://www.lse.ac.uk/itservices/help/Windowsvulnerabilities.htm
1016	http://www.stardrifter.org/wallpapersync/
1017	http://gameknot.com/faq.pl
1018	http://www.whoi.edu/cis/security/news/platform_security.html
1019	http://www.screenit.com/subscribers/win98winME.asp
102	http://support.microsoft.com/kb/896358
1020	http://hr123.any2000.com/
1021	http://learning.bankofshanghai.com.cn/htm/client_check/zh-cn/index.htm
1022	http://ctaar.rutgers.edu/workshops/
1023	http://www.microsoft.com/spain/athome/security/protect/windowsme/updates.mspx
1024	http://www.microsoft.com/france/windows/xp/sp2/technologiesoverview.mspx
1025	http://www.microsoft.com/middleeast/
1026	http://www.microsoft.com/technet/security/bulletin/MS01-043.mspx
1027	http://www.spamfighter.com/lang_nl/product_sem.asp
1028	http://ludit.kuleuven.be/software/beveiliging/windowsupdate.html
1029	http://wwwcis.kokushikan.ac.jp/

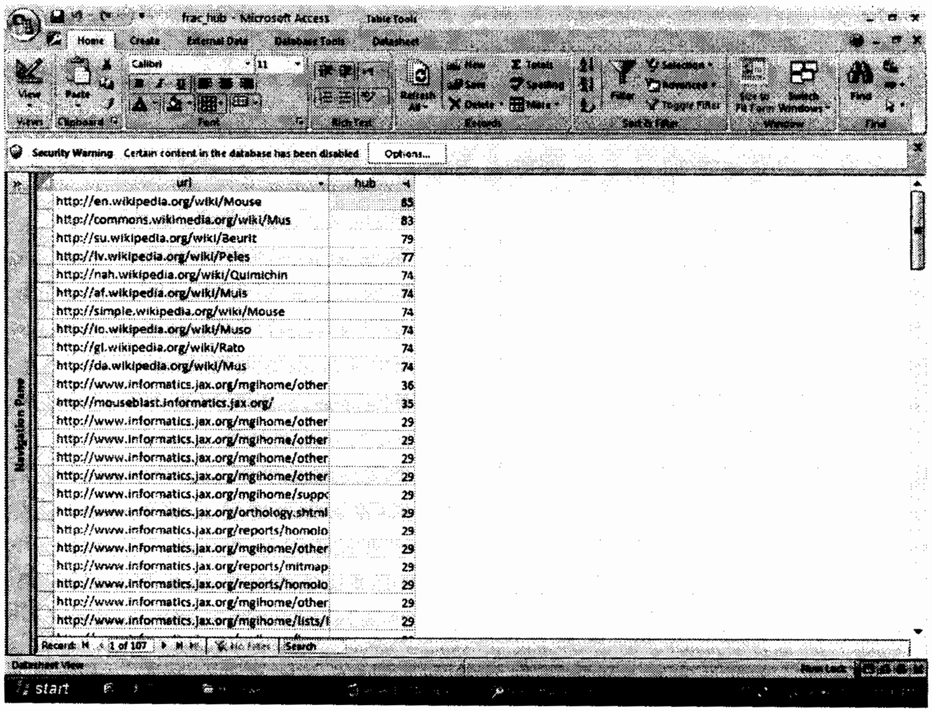
Fig B-4: Base_Set



id	url
	http://www.wininformant.com
100	http://msdn.microsoft.com/library/default.asp?url
1000	http://www.utdallas.edu/ir/security/VirusNews.htm
1001	http://www.us-cert.gov/cas/alerts/SA06-165A.html
1002	http://www.microsoft.com/technet/security/bulletin/ms01-021.mspx
1003	http://www.geocities.com/Tokyo/Market/2978/index.html
1004	http://wsus.tjuci.edu.cn/
1005	http://blogs.msdn.com/gduthie/archive/2007/11.aspx
1006	http://www.spamfighter.com/sem_email_server_spam_filter.asp
1007	http://www.nifty.com/webapp/security/news/supportnews_detail.jsp?newsId
1008	http://us.trendmicro.com/us/threats/home-user/preventing-intrusions/safe-computing-guide/windows-server-2003/
1009	http://www.microsoft.com/technet/security/bulletin/MS01-050.mspx
1010	http://www.auburn.edu/oit/hardware_software/os/common_topics.php
1011	http://www.ctinet.ru/support.asp
1012	http://www.eab-online.com/Security/SecurityGuide.html
1013	http://www.microsoft.com/japan/security/bulletins/feb2006.mspx
1014	http://www.colby-sawyer.edu/about/website/help.html
1015	http://www.lse.ac.uk/itservices/help/Windowsvulnerabilities.htm
1016	http://www.stardrifter.org/wallpapersync/
1018	http://www.whoi.edu/cis/security/news/platform_security.html
1019	http://www.screenit.com/subscribers/win98winME.asp
102	http://support.microsoft.com/kb/896358
1020	http://hr123.any2000.com/
1021	http://learning.bankofshanghai.com.cn/htm/client_check/zh-cn/index.htm
1022	http://ctaar.rutgers.edu/workshops/
1023	http://www.microsoft.com/spain/athome/security/protect/windowsme/updates.mspx
1025	http://www.microsoft.com/middleeast/
1026	http://www.microsoft.com/technet/security/bulletin/MS01-043.mspx
1027	http://www.spamfighter.com/lang_nl/product_sem.asp
1028	http://ludit.kuleuven.be/software/beveiliging/windowsupdate.html
1029	http://wwwcis.kokushikan.ac.jp/
103	http://www.stats.uwo.ca/faculty/murdoch/software/debuggingR/gdb.shtml
1031	http://www.cddn.com/blog/
1032	http://www.pagaelpato.com/publicidad-internet/
1033	http://www.windowsnewsletter.com/html-archive/2006/07jan2006.xml

Record: 1 of 728

Fig B-5: Max by 10

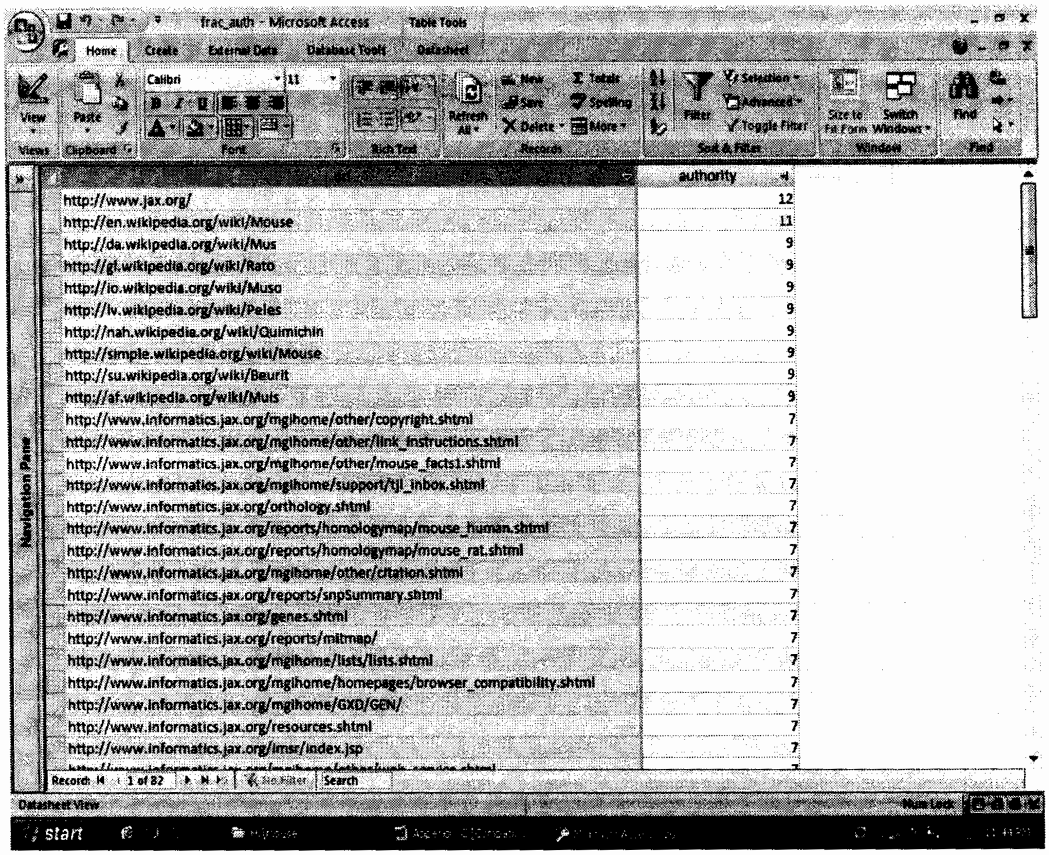


Security Warning Certain content in the database has been disabled Options...

url	hub
http://en.wikipedia.org/wiki/Mouse	85
http://commons.wikimedia.org/wiki/Mus	83
http://su.wikipedia.org/wiki/Beurit	79
http://lv.wikipedia.org/wiki/Peles	77
http://nah.wikipedia.org/wiki/Quimichin	74
http://af.wikipedia.org/wiki/Muis	74
http://simple.wikipedia.org/wiki/Mouse	74
http://io.wikipedia.org/wiki/Muso	74
http://gl.wikipedia.org/wiki/Rato	74
http://da.wikipedia.org/wiki/Mus	74
http://www.informatics.jax.org/mgihome/other	36
http://mouseblast.informatics.jax.org/	35
http://www.informatics.jax.org/mgihome/other	29
http://www.informatics.jax.org/mgihome/other	29
http://www.informatics.jax.org/mgihome/other	29
http://www.informatics.jax.org/mgihome/other	29
http://www.informatics.jax.org/mgihome/suppc	29
http://www.informatics.jax.org/orthology.shtml	29
http://www.informatics.jax.org/reports/homolo	29
http://www.informatics.jax.org/mgihome/other	29
http://www.informatics.jax.org/reports/mitmap	29
http://www.informatics.jax.org/reports/homolo	29
http://www.informatics.jax.org/mgihome/other	29
http://www.informatics.jax.org/mgihome/lists/l	29

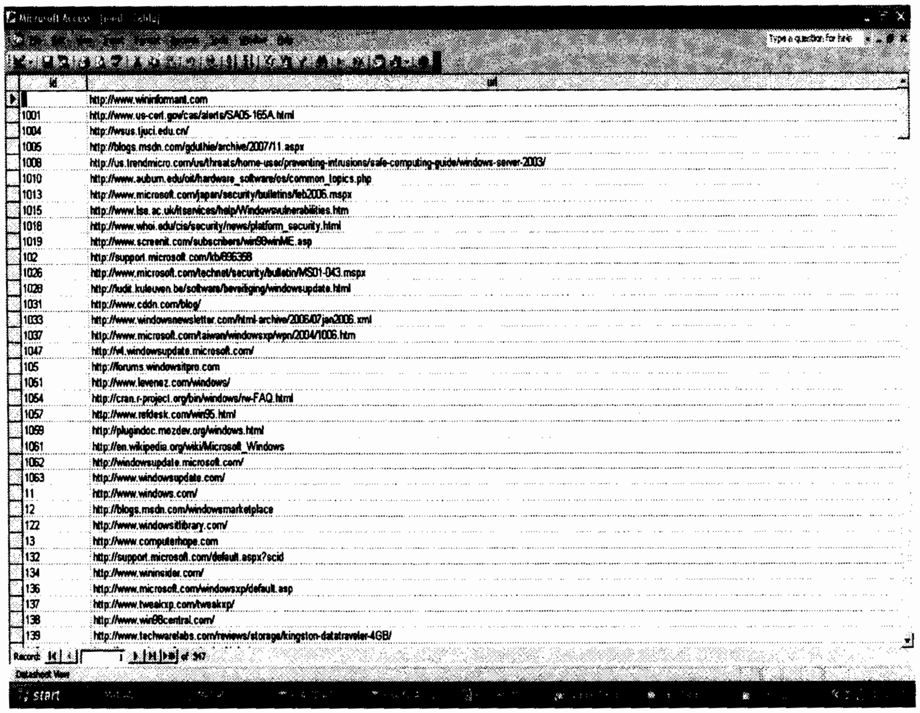
Record: 1 of 107 No Filter Search

Fig B-6: Max by 10_hub



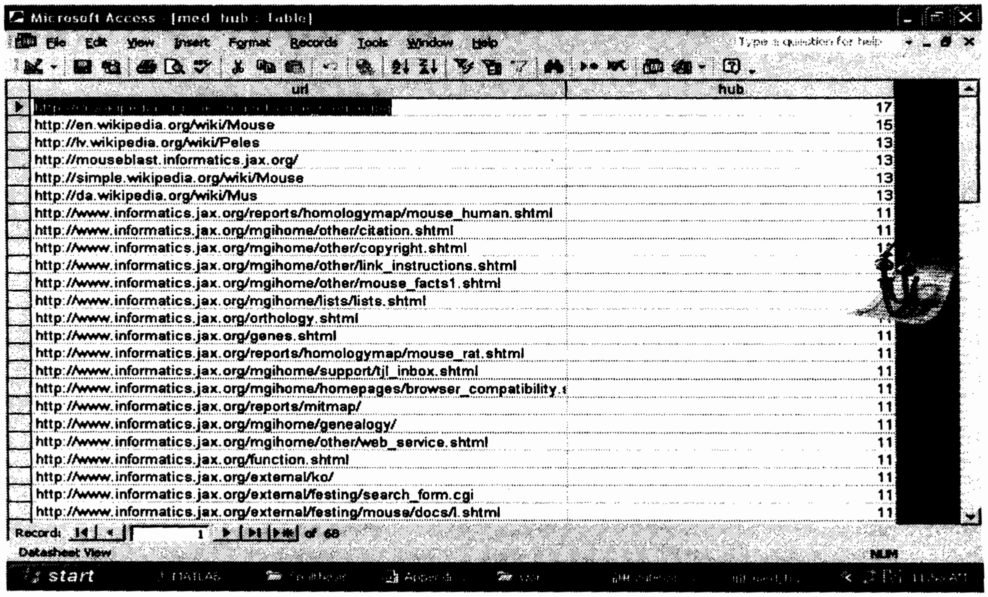
[illegible]	authority
http://www.jax.org/	12
http://en.wikipedia.org/wiki/Mouse	11
http://da.wikipedia.org/wiki/Mus	9
http://gl.wikipedia.org/wiki/Rato	9
http://io.wikipedia.org/wiki/Muso	9
http://lv.wikipedia.org/wiki/Peles	9
http://nah.wikipedia.org/wiki/Quimichin	9
http://simple.wikipedia.org/wiki/Mouse	9
http://su.wikipedia.org/wiki/Beurit	9
http://af.wikipedia.org/wiki/Muis	9
http://www.informatics.jax.org/mgihome/other/copyright.shtml	7
http://www.informatics.jax.org/mgihome/other/link_instructions.shtml	7
http://www.informatics.jax.org/mgihome/other/mouse_facts1.shtml	7
http://www.informatics.jax.org/mgihome/support/tjl_inbox.shtml	7
http://www.informatics.jax.org/orthology.shtml	7
http://www.informatics.jax.org/reports/homologymap/mouse_human.shtml	7
http://www.informatics.jax.org/reports/homologymap/mouse_rat.shtml	7
http://www.informatics.jax.org/mgihome/other/citation.shtml	7
http://www.informatics.jax.org/reports/snpSummary.shtml	7
http://www.informatics.jax.org/genes.shtml	7
http://www.informatics.jax.org/reports/mitmap/	7
http://www.informatics.jax.org/mgihome/lists/lists.shtml	7
http://www.informatics.jax.org/mgihome/homepages/browser_compatibility.shtml	7
http://www.informatics.jax.org/mgihome/GXD/GEN/	7
http://www.informatics.jax.org/resources.shtml	7
http://www.informatics.jax.org/imsr/index.jsp	7

Fig B-7: Max by 10_auth



id	url
	http://www.wininformant.com
1001	http://www.us-cert.gov/cas/alerts/SA05-165A.html
1004	http://wsus.tjuci.edu.cn/
1005	http://blogs.msdn.com/gduthie/archive/2007/11.aspx
1008	http://us.trendmicro.com/us/threats/home-user/preventing-intrusions/safe-computing-guide/windows-server-2003/
1010	http://www.auburn.edu/oit/hardware_software/os/common_topics.php
1013	http://www.microsoft.com/japan/security/bulletins/feb2006.mspx
1015	http://www.lse.ac.uk/itservices/help/Windowsvulnerabilities.htm
1018	http://www.whoi.edu/cis/security/news/platform_security.html
1019	http://www.screenit.com/subscribers/win98winME.asp
102	http://support.microsoft.com/kb/896358
1026	http://www.microsoft.com/technet/security/bulletin/MS01-043.mspx
1028	http://ludit.kuleuven.be/software/beveiliging/windowsupdate.html
1031	http://www.cddn.com/blog/
1033	http://www.windowsnewsletter.com/html-archive/2006/07jan2006.xml
1037	http://www.microsoft.com/taiwan/windowsxp/wpn/2004/1006.htm
1047	http://v4.windowsupdate.microsoft.com/
105	http://forums.windowsitpro.com
1051	http://www.levenez.com/windows/
1054	http://cran.r-project.org/bin/windows/rw-FAQ.html
1057	http://www.refdesk.com/win95.html
1059	http://plugindoc.mozdev.org/windows.html
1061	http://en.wikipedia.org/wiki/Microsoft_Windows
1062	http://windowsupdate.microsoft.com/
1063	http://www.windowsupdate.com/
11	http://www.windows.com/
12	http://blogs.msdn.com/windowsmarketplace
122	http://www.windowsitlibrary.com/
13	http://www.computerhope.com
132	http://support.microsoft.com/default.aspx?scid
134	http://www.wininsider.com/
136	http://www.microsoft.com/windowsxp/default.asp
137	http://www.tweakxp.com/tweakxp/
138	http://www.win98central.com/
139	http://www.techwarelabs.com/reviews/storage/kingston-datatraveler-4GB/

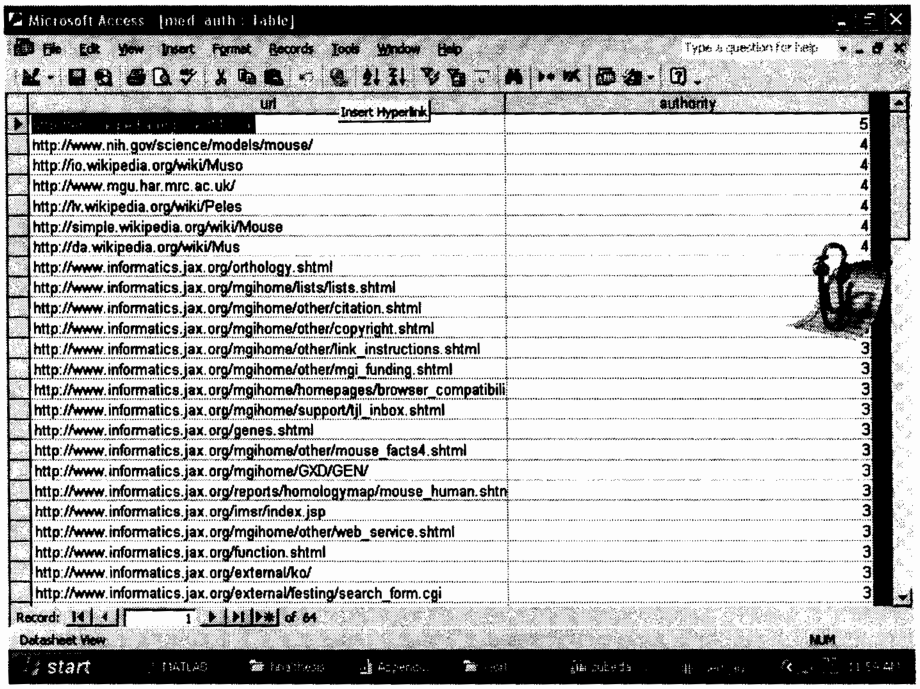
Fig B-8: Median



Microsoft Access - [med_hub : Table]
File Edit View Insert Format Records Tools Window Help
Type a question for help
Record: 1 of 68
Datasheet View
NUM
start

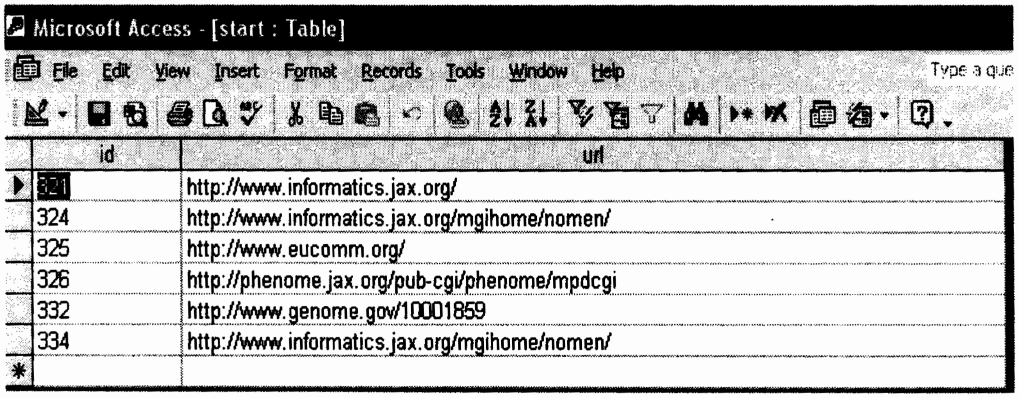
url	hub
[illegible]	17
http://en.wikipedia.org/wiki/Mouse	15
http://lv.wikipedia.org/wiki/Peles	13
http://mouseblast.informatics.jax.org/	13
http://simple.wikipedia.org/wiki/Mouse	13
http://da.wikipedia.org/wiki/Mus	13
http://www.informatics.jax.org/reports/homologymap/mouse_human.shtml	11
http://www.informatics.jax.org/mgihome/other/citation.shtml	11
http://www.informatics.jax.org/mgihome/other/copyright.shtml	[illegible]
http://www.informatics.jax.org/mgihome/other/link_instructions.shtml	[illegible]
http://www.informatics.jax.org/mgihome/other/mouse_facts1.shtml	[illegible]
http://www.informatics.jax.org/mgihome/lists/lists.shtml	[illegible]
http://www.informatics.jax.org/orthology.shtml	[illegible]
http://www.informatics.jax.org/genes.shtml	11
http://www.informatics.jax.org/reports/homologymap/mouse_rat.shtml	11
http://www.informatics.jax.org/mgihome/support/tjl_inbox.shtml	11
http://www.informatics.jax.org/mgihome/homepages/browser_compatibility.s	11
http://www.informatics.jax.org/reports/mitmap/	11
http://www.informatics.jax.org/mgihome/genealogy/	11
http://www.informatics.jax.org/mgihome/other/web_service.shtml	11
http://www.informatics.jax.org/function.shtml	11
http://www.informatics.jax.org/external/ko/	11
http://www.informatics.jax.org/external/festing/search_form.cgi	11
http://www.informatics.jax.org/external/festing/mouse/docs/l.shtml	11

Fig B-9: Med_Hub


Microsoft Access [med auth : Table]
File Edit View Insert Format Records Tools Window Help
Type a question for help
Insert Hyperlink
Record: 1 of 64
Datasheet View
NUM
start

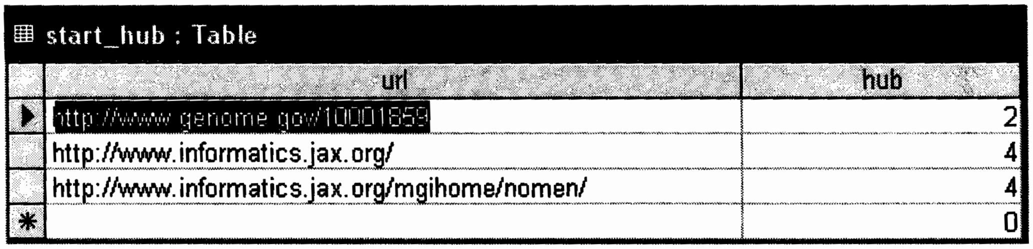
url	authority
[illegible]	5
http://www.nih.gov/science/models/mouse/	4
http://io.wikipedia.org/wiki/Muso	4
http://www.mgu.har.mrc.ac.uk/	4
http://lv.wikipedia.org/wiki/Peles	4
http://simple.wikipedia.org/wiki/Mouse	4
http://da.wikipedia.org/wiki/Mus	4
http://www.informatics.jax.org/orthology.shtml	[illegible]
http://www.informatics.jax.org/mgihome/lists/lists.shtml	[illegible]
http://www.informatics.jax.org/mgihome/other/citation.shtml	[illegible]
http://www.informatics.jax.org/mgihome/other/copyright.shtml	[illegible]
http://www.informatics.jax.org/mgihome/other/link_instructions.shtml	3
http://www.informatics.jax.org/mgihome/other/mgi_funding.shtml	3
http://www.informatics.jax.org/mgihome/homepages/browser_compatibili	3
http://www.informatics.jax.org/mgihome/support/tjl_inbox.shtml	3
http://www.informatics.jax.org/genes.shtml	3
http://www.informatics.jax.org/mgihome/other/mouse_facts4.shtml	3
http://www.informatics.jax.org/mgihome/GXD/GEN/	3
http://www.informatics.jax.org/reports/homologymap/mouse_human.shtn	3
http://www.informatics.jax.org/imsr/index.jsp	3
http://www.informatics.jax.org/mgihome/other/web_service.shtml	3
http://www.informatics.jax.org/function.shtml	3
http://www.informatics.jax.org/external/ko/	3
http://www.informatics.jax.org/external/festing/search_form.cgi	3

Fig B-10:Med_Auth


Microsoft Access - [start : Table]
File Edit View Insert Format Records Tools Window Help
Type a que

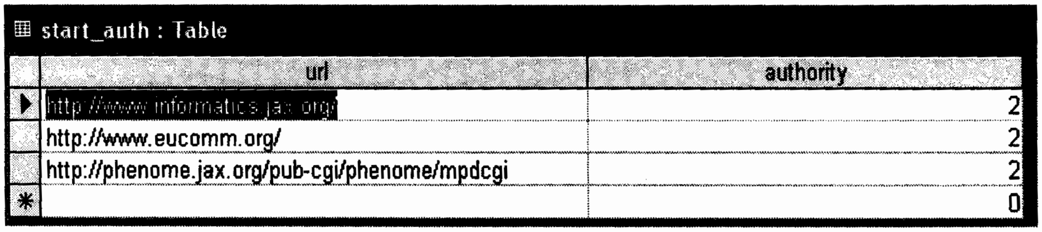
id	url
321	http://www.informatics.jax.org/
324	http://www.informatics.jax.org/mgihome/nomen/
325	http://www.eucomm.org/
326	http://phenome.jax.org/pub-cgi/phenome/mpdcgi
332	http://www.genome.gov/10001859
334	http://www.informatics.jax.org/mgihome/nomen/
*	

Fig B-11:Start



	url	hub
▶	http://www.genome.gov/10001859	2
	http://www.informatics.jax.org/	4
	http://www.informatics.jax.org/mgihome/nomen/	4
*		0

Fig B-12:Start_hub



	url	authority
▶	http://www.informatics.jax.org/	2
	http://www.eucomm.org/	2
	http://phenome.jax.org/pub-cgi/phenome/mpdcgi	2
*		0

Fig B-13:Start_auth