International Journal of scientific research and management (IJSRM)
||Volume||2||Issue||7||Pages|| 1070-1080||2014||
Website: www.ijsrm.in ISSN (e): 2321-3418

An Efficient Approach For Indexing Web Pages Using Page Ranking Algorithm For The Enhancement Of Web Search Engine Results Results

Ms. Nilima V. Pardakhe[1], Prof. R. R. Keole[2]

[1]S.G.B.A.U., Amravati, H.V.P.M. College of Engg., Amravati,
McMahons Road, Frankston 3199, Australia
nilimapardakhe@gmail.com

[2]Department of Computer Science and Engineering, S.G.B.A.U.,Amravati,
H.V.P.M. College of Engg. Amravati,
ranjitkeole@gmail.com

Abstract: ***As web is the largest collection of information and plenty of pages or documents are newly added and deleted on frequent basis due to the dynamic nature of the web. The information present on the web is of great need, the world is full of questions and the web is serving as the major source of gaining information about specific query made by the user. Search engines generally return a large number of pages in response to user queries. To assist the users to navigate in the result list, ranking methods are applied on the search results. Most of the ranking algorithms proposed in the literature are either link or content oriented. Proposed methodology uses, a page ranking mechanism based on frequency of keywords found in the query made by user. This approach helps to rank most valuable and relevant search results on the top of the result list. Therefore tries to enhance the search engine results.***

Keywords: Information Retrieval, PageRank, Search Engine, Web Mining, World Wide Web.

1. Introduction

The World Wide Web (Web) is popular and interactive medium to propagate information today. The web is huge, diverse, dynamic, widely distributed global information service center. With the rapid growth of the web, users get easily lost in the rich hyperlink structure. Providing relevant information to the users to cater to their needs is the primary goal of website owners. Therefore, finding the content of the web and retrieving the users' interests and needs from their behavior have become increasingly important. When a user makes a query from search engine, it generally returns a large number of pages in response to user queries. This result-list contains many relevant and irrelevant pages according to user's query. As user impose more number of irrelevant pages in the search result-list, to assist the users to navigate in the result list, various ranking methods are applied on the search results. The search engine uses these ranking methods to sort the results to be displayed to the user. In that way user can find the most important and useful result first. [1] There are a variety of algorithms developed; few of them are PageRank, Weighted PageRank, and HITS etc.

1.1Web Search Engine

The World Wide Web consists billions of web pages and huge amount of information available within pages. To retrieve required information from World Wide Web, search engines perform number of task based on their respective architecture. The web search engine represents the user interface needed to permit the user to query the information. It is the connection between user and the information repository when user sends query to search engine, Web Search Engine is a tool enabling document search with respect to specified keywords in the web and returns a list of documents where the keywords were found. A web search engine is a software system that is designed to search for information on the World Wide Web. The search results are generally presented in a line of results often referred to as Search Engine Result Pages (SERPs). The information may be a specialist in web pages, images, information and other types of files. Some search engines also mine data available in databases or open directories. Unlike web directories, which are maintained only by human editors, search engines also maintain Real Time Computing information by running an algorithm on a web crawler.

1.1.1Components of Web Search Engine:

1. User Interface: The user interface, in the industrial design field of human machine interaction is the space where interaction between humans and machines occurs. The goal of this interaction is effective operation and control of the

machine on the user's end, and feedback from the machine, which aids the operator in making operational decisions. It is the part of Web Search Engine interacting with the users and allowing them to query and view query results.

2. Parser: It is the component providing term (keyword) extraction for both sides. The parser determines the keywords of the user query and all the terms of the Web documents which have been scanning by the crawler. Term extraction procedure includes the following subprocedures:

*Step 1: Tokenizing***.** As soon as a user inputs a query, the search engine — whether a keyword-based system or a full natural language processing (NLP) system — must tokenize the query stream, i.e., break it down into understandable segments. Usually a token is defined as an alpha-numeric string that occurs between white space and/or punctuation.

Step 2: Normalization. Since users may employ special operators in their query, including Boolean, adjacency, or proximity operators, the system needs to parse the query first into query terms and operators. These operators may occur in the form of reserved punctuation (e.g., quotation marks) or reserved terms in specialized format (e.g., AND, OR). In the case of an NLP system, the query processor will recognize the operators implicitly in the language used no matter how the operators might be expressed (e.g., prepositions, conjunctions, ordering).

At this point, a search engine may take the list of query terms and search them against the inverted file. In fact, this is the point at which the majority of publicly available search engines perform the search.

Steps 3: Stemming. Stemming algorithms are used to transform the words in texts into their grammatical root form, and are mainly used to improve the Information Retrieval Systems efficiency. To stem a word is to reduce it to a more general form, possibly its root. For example, stemming the term interesting may produce the term interest. Though the stem of a word might not be its root, want all words that have the same stem to have the same root. The effect of stemming on searches of English document collections has been tested extensively. Several algorithms exist with different techniques.

*Steps 4: Stop word handling***.** After stemming it is necessary to remove unwanted words. There are 400 to 500 types of stop words such as of, and, the, etc., that provide no useful information about the documents topic. Stop-word removal is the process of removing these words. Stop-words account for about 20% of all words in a typical document. These techniques greatly reduce the size of the search engines index.

3. Web Crawler: A web crawler is a relatively simple automated program, or script that methodically scans or "crawls" through Internet pages to create an index of the data it is looking for. Alternative names for a web crawler include web spider, web robot, crawler, and automatic indexer.[16] When a web crawler visits a web page, it reads the visible text, the hyperlinks, and the content of the various tags used in the site, such as keyword rich meta tags. Using the information gathered from the crawler, a search engine will then determine what the site is about and index the information. Lastly, the website is included in the search engine's database and its page ranking process.

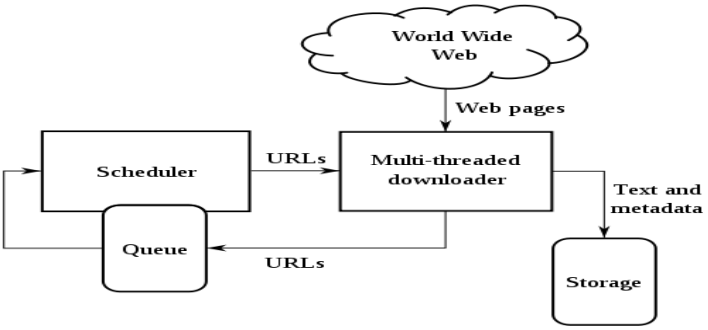


Figure 1 Web Crawler Architecture

4. Database: It is the component that all the text and metadata specifying the web documents scanned by the crawler.

5. Ranking Engine: The component is mainly the ranking algorithm operating on the current data, which is indexed by the crawler, to be able to provide some order of relevance, for the web documents, with respect to the user query.

Following figure shows the architecture of web search engine[33]

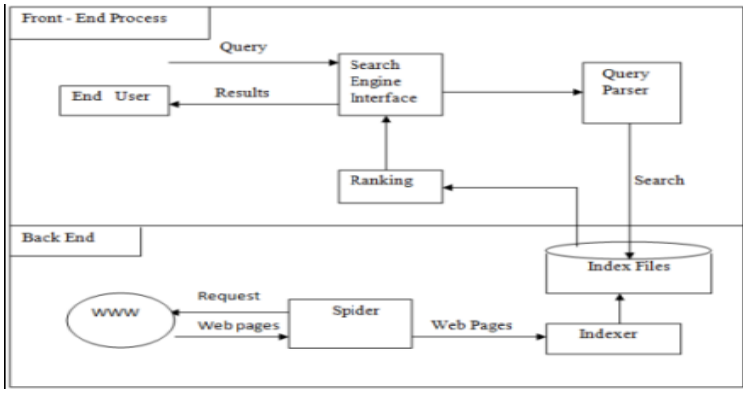


Figure 2 Web Search Engine Architecture

When we do web search we are actually search in the web made up of over 60 trillion individual pages and its constantly growing. This is done with the software programming called as Spiders. Spiders start with fetching of few web pages and forward the link on those pages and fetch the pages they link to and so on. Then the pages are sorted by their content and other factors and keep track of it in the index. When we search in the most basic level, google algorithm look up the search terms in the index to find the appropriate pages. There are many programs and formulas to deliver the best results. Algorithms get to work looking for clues (search methods, auto complete, spelling, synonyms, query understanding) to better understand what we mean. Based on these clues google pull relevant documents from the index using over 200 factors. Finally google combine all these factors together to produce each page over the score and sent back to the search about half a second after we submit our search.

1.2 Web Mining Overview

The most of the people use the internet for retrieving information. But most of the time, they suffer from insignificant and irrelevant document even after navigating several links. For retrieving information from the Web, Web mining techniques are used.

Web mining is an application of the data mining techniques to automatically discover and extract knowledge from the Web. Web mining consists of the following tasks:

1. Resource finding: the task of retrieving intended Web documents. [17]

2.Information selection and pre-processing: automatically selecting and pre-processing specific information from retrieved Web re-sources.
3. Generalization: automatically discovers general patterns at individual Web sites as well as across multiple sites.
4. Analysis: validation and/or interpretation of the mined patterns.

. There are three areas of Web mining namely, Web Content Mining (WCM), Web Usage Mining (WUM) and Web Structure Mining (WSM).[32]

1.2.1 Web Content Mining (WCM)

Web Content Mining is the process of extracting useful information from the contents of web documents.[6] The web documents may consists of text, images, audio, video or structured records like tables and lists. Mining can be applied on the web documents as well the results pages produced from a search engine. Web content mining could be differentiated from two points of view: the agent-based approach or the database approach. The first approach aims on improving the information finding and filtering and could be placed into the following three categories [19]:
1. Intelligent Search Agents. These agents search for relevant information using domain characteristics and user profiles to organize and interpret the discovered information.
2. Information Filtering/ Categorization. These agents use information retrieval techniques and characteristics of open hypertext Web documents to automatically retrieve, filter, and categorize them.
3. Personalized Web Agents. These agents learn user preferences and discover Web information based on these preferences, and preferences of other users
with similar interest.

The second approach aims on modeling the data on the Web into more structured form in order to apply standard database querying mechanism and data mining applications to analyze it. The two main categories are Multilevel databases and Web query systems.

1.2.2Web Usage Mining (WUM)

Web Usage Mining is the process of extracting useful information from the secondary data derived from the interactions of the user while surfing on the Web. It extracts data stored in server access logs, referrer logs, agent logs, client-side cookies, user profile and metadata. [7]
The challenges involved in web usage mining could be divided in three phases:
1. Pre-processing: The data available tend to be noisy, incomplete and inconsistent. In this phase, the data available should be treated according to the requirements of the next phase. It includes data cleaning, data integration, data transformation and data reduction.
2. Pattern discovery: Several different methods and algorithms such as statistics, data mining, machine learning and pattern recognition could be applied to identify user patterns.
3. Pattern Analysis: This process targets to understand, visualize and give interpretation to these patterns. [13]

1.2.3 Web Structure Mining (WSM)

The goal of the Web Structure Mining is to generate the structural summary about the Web site and Web page.[18] It tries to discover the link structure of the hyperlinks at the inter-document level. Based on the topology of the hyperlinks, Web Structure mining will categorize the Web pages and generate the information like similarity and relationship between different Web sites. This type of mining can be performed at the document level (intra-page) or at the hyperlink level (inter-page). It is important to understand the Web data structure for Information Retrieval. The Web contains a variety of objects with almost no unifying structure, with differences in the authoring style and content much greater than in traditional collections of text documents.

The objects in the WWW are web pages, and links are in, out and co-citation i.e. two pages that are both linked to the same page. There are some possible tasks of link mining which are applicable in Web structure mining and are described as follows: [18]
1. Link-based Classification: It is the most recent upgrade of a classic data mining task to linked Domains. The task is to focus on the prediction of the category of a web page, based on words that occur on the page, links between pages, anchor text, html tags and other possible attributes found on the web page.
2. Link-based Cluster Analysis: The goal in cluster analysis is to find naturally occurring sub-classes. The data is segmented into groups, where similar objects are grouped together, and dissimilar objects are grouped into different groups. Different than the previous task, link-based cluster analysis is unsupervised and can be used to discover hidden patterns from data.
3. Link Type: There are a wide range of tasks concerning the prediction of the existence of links, such as predicting the type of link between two entities, or predicting the purpose of a link.
4. Link Strength: Links could be associated with weights.
5. Link Cardinality: The main task here is to predict the number of links between objects.

Due to the heterogeneity of network resources and the lack of structure of web data, automated discovery of targeted knowledge retrieval mechanism is still facing many research challenges. Effective organization of search results is critical for improving the utility of any search engine. The utility of a search engine is affected by multiple factors. While the primary factor is the soundness of the underlying retrieval model and ranking function. Search engines generally return a large number of pages in response to user queries. To assist the users to navigate in the result list, ranking methods are applied on the search results. The proposed methodology try to overcome the problem of ranking by using the page ranking based on the frequency of the keyword entered by the user, which then helps to display most relevant and valuable pages on the top of the result list.

2. Literature Review

The information present on the web is huge in amount, distributed, heterogeneous and dynamic. Due to the heterogeneity of network resources and the lack of structure of web data, automated discovery of targeted knowledge retrieval mechanism is still facing many research challenges. With the huge amount of information available online, the World Wide Web is a fertile area for data mining research. The Web mining research is at the cross road of research from several research communities, such as database, information retrieval, and within AI, especially the sub-areas of machine learning and natural language processing. However, there are a lot of confusions when comparing research efforts from different point of views.

Raymond Kosala and Hendrik Blockeel [17] survey the research in the area of Web mining, point out some confusion regarded the usage of the term Web mining and suggest three Web mining categories. Then situate some of the

research with respect to these three categories and explore the connection between the web mining categories and the related agent paradigm. For the survey, focus is made on representation issues, on the process, on the learning algorithm, and on the application of the recent works as the criteria. The challenge for Web structure mining [18] is to deal with the structure of the hyperlinks within the Web itself. Link analysis is an old area of research. However, with the growing interest in Web mining, the research of structure analysis had increased and these efforts had resulted in a newly emerging research area called Link Mining, which is located at the intersection of the work in link analysis, hypertext and web mining, relational learning and inductive logic programming, and graph mining. There is a potentially wide range of application areas for this new area of research, including Internet. Moreover, the semi structured and unstructured nature of web data creates the need for web content mining. . In paper [7] the author differentiates web content mining from two different points of view. Information retrieval view and database view. In paper [21] research area of web mining and different categories of web mining are discussed briefly. They also summarized the research works done for unstructured data and semi structured data from information retrieval view. In IR view, the unstructured text is represented by bag of words and semi-structured words are represented by HTML structure and hyperlink structure [8]. In Database (DB) view, the mining always tries to infer the structure of the web site to transform a web site into a database. A new method for relevance ranking of web pages with respect to given query was determined in paper[5]. Various problem of identifying content such as a sequence labeling problem, a common problem structure in machine learning and natural language processing is identified in [12]. A survey of web content mining plays as an efficient tool in extracting structured and semi structured data and mining them into useful knowledge is presented in [21].

2.1 Ranking in web search

Nowadays searching on the internet is most widely used operation on the World Wide Web. The amount of information is increasing day by day rapidly that creates the challenge for information retrieval. There are so many tools to perform efficient searching. Due to the size of web and requirements of users creates the challenge for search engine page ranking [22]. Ranking is the main part of any information retrieval system Today's search engines may return million of pages for a certain query It is not possible for a user to preview all the returned results So, page ranking is helpful in web searching. Rankers are classified into two groups: - Content-based rankers and Connectivity-based rankers. Content-based rankers works on the basis of number of matched terms, frequency of terms, location of terms, etc. Connectivity-based rankers work on the basis of link analysis technique; links are edges that point to different web pages. There are two famous link analysis methods:- 1)PageRank Algorithm[9] and 2) HITS Algorithm[11].

PageRank has been developed by Google and is named after Larry Page, Google's co-founder and president [9]. PageRank ranks pages based on the web structure. PageRank uses global link information and is stated to be the primary link recommendation scheme employed in the Google search engine and search appliance. PageRank is designed to simulate the behavior of a "random web surfer" [9] who navigate a web by randomly following links. If a page with no outgoing links is reached, the surfer jumps to a randomly chosen bookmark. In addition to this normal surfing behavior, the surfer occasionally spontaneously jumps to a bookmark instead of following a link. The PageRank of a page is the probability that the web surfer will be visiting that page at any given moment. PageRank is a static algorithm for measure the global importance of one page which only considers the link relation among the pages. We can calculate the PageRank value for all pages off-line. This value is irrelevant to user-specific-query.

Wenpu Xing and Ali Ghorbani [10] introduce an extended PageRank algorithm called the Weighted PageRank algorithm (WPR). This Algorithm assigns rank values to pages according to their importance rather than dividing it evenly. The importance is assigned in terms of weight values to incoming and outgoing links. HITS was used for the first time in the Clever search engine from IBM, and PageRank is used by Google combined with other several features such as anchor text, IR measures, and proximity. HITS provide an innovative methodology for Web searching and topics distillation. According to the definition by Google, a web page is an authority on a topic if it provides good information and is a hub if it provides links to good authorities. HITS uses the mutual reinforcement operation to propagate hub and authority values to represent the linking characteristic [10]. HITS is a real time dynamic algorithm for each query. We need to calculate the authority of each page on-line. This prevent HITS algorithm from popularly using in real search engines. All these algorithms consider the importance of information quality for ranking, but with the development of network, the recommendation among pages is becoming less important.

3. Proposed Methodology

The proposed methodology aims to provide the results to the users which are more relevant to the user query. It tries to overcome the problem of page ranking, in which an approach of relevant search which ranks the web pages based on the frequency or count of keywords (searched by user) is proposed. The web page containing maximum frequency or counts of keyword searched by the user is more relevant and displayed first in the list of web page links on the user screen. Every result is individually analyzed based on frequency of keywords and thus based on the user query, search results are obtained.

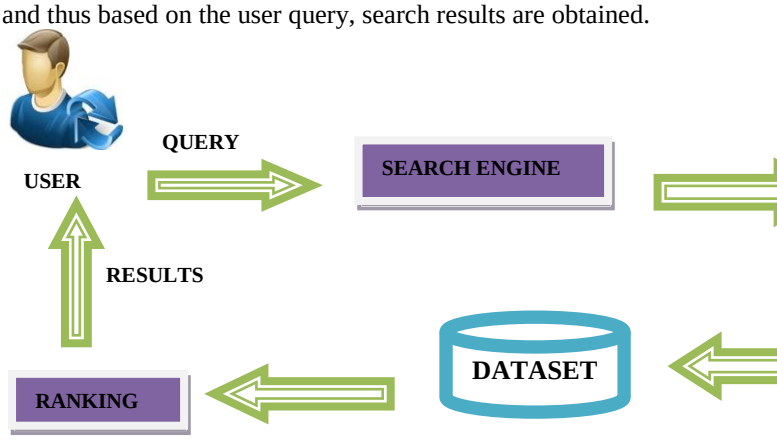


Figure 3 Architecture of Proposed Methodology

Proposed methodology works as follows:

It involves user request to search for the particular query to obtain the search results according to the user query. In Proposed methodology user has to first enter the query. Then preprocessing is performed on that entered query which involves three steps such as text filtering, stemming, stop words

removal. After preprocessing of the query keywords are obtained. Then web snippets related to that keyword are fetched from the dataset and frequency of that particular keyword is calculated and finally on the basis of that frequency of keyword, search results are ranked that is the search results are displayed in descending order of frequency of keyword to the user.

3.1Dataset used:

Dataset is created by collecting web snippets for some particular keywords. So, here for implementation of ranking rather than fetching the snippets from any search engine AMBIENT [13] data set is used, in which numbers of snippets already has stored. This means that this work considers that, the work had already done for extraction of top 200-500 snippets from top search engine and can be stored in text file. The implementation is done with the AMBIENT dataset.

AMBIENT

It is a dataset designed for evaluating the subtopic information retrieval. It consists of 44 topics which are selected from Wikipedia disambiguation page. Each topic has a set of subtopics. Each subtopic has a set of documents that comprises of URL, title and snippet, retrieved from a Web search engine as of January 2012. They are annotated with subtopic relevance judgments. The AMBIENT dataset has 44 topics with an average of 17 subtopics under each topic. The topics and its subtopics which do not have any appropriate terms within the search result are considered to be noise and are removed from the dataset.

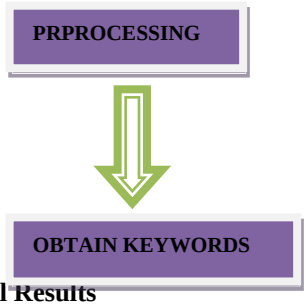
3.2 Specification and Techniques Used

Implementation is the stage of project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective. The implementation stage involves careful planning, investigation of the existing system and its constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods.

Algorithm for Mining Web Content

Algorithm: Relevancy and keyword frequency based approach
Input: User query
Output: Reordered search results

Step 1: Enter the user query.
Step 2: Perform preprocessing of user query.
Step 3: Obtain keywords from processed query.
Step 4: Extract the web snippets from the dataset related to the specified keyword.
Step 5: Find frequency of the keyword.
Step 6: Display the search results in descending order of keyword frequency.



Experimental Results

Experiment is conducted with a user query “B-52” against specific search-engine. Top 20 web snippets from that search-engine are taken as an input dataset which contains their id, url, title and ranking are listed in Table I.

TABLE I.
INPUT DATA SET

Id	url	Title	Search Engine Rank
1	http://www.boeing.com/defense-space/military/b52-strat	Boeing: Integrated Defense Systems - B-52 Stratofortress	1

2	http://www.fas.org/nuke/guide/usa/bomber/b-52.htm	B-52 Stratofortress - United States Nuclear Forces	**2**
3	http://www.theb52s.com/	The B-52's : World's Greatest Party Band	**3**
4	http://en.wikipedia.org/wiki/B-52_Stratofortress	B-52 Stratofortress - Wikipedia, the free encyclopedia	**4**
5	http://www.af.mil/factsheets/factsheet.asp?fsID=83	Factsheets : B-52 Stratofortress :	**5**
6	http://www.b-52pro.com/	b-52	**6**
7	http://www.globalsecurity.org/wmd/systems/b-52.htm	B-52 Stratofortress	**7**
8	http://www.stratofortress.org/	B52 Stratofortress Association	**8**
9	http://www.dfrc.nasa.gov/Gallery/Photo/B-52	NASA Dryden B-52 Photo Collection	**9**
10	http://en.wikipedia.org/wiki/B-52_(cocktail)	B-52 (cocktail) - Wikipedia, the free encyclopedia	**10**
11	http://www.rottentomatoes.com/m/1111762-b52/articlegate.php	B-52 - Rotten Tomatoes	**11**
12	http://www.britannica.com/eb/article-9011572/B-52	B-52 Encyclopaedia Britannica	**12**
13	http://www.amazon.com/B-52s/dp/B000002KKD	Amazon.com: The B-52's: Music: The B-52's	**13**
14	http://www.globalsecurity.org/wmd/systems/b-52-upgrade.htm	B-52 Stratofortress	**14**
15	http://youtube.com/?v=exdjbeHn-_8	YouTube - B-52	**15**
16	http://www.boeing.com/defense-space/military/b52-strat/b52_50th/index.html	Boeing: News Feature - B-52 50th Anniversary -	**16**
17	http://www.b-52pro.com/start.html	Welcome to B-52 Professional	**17**
18	http://www.af.mil/photos/index.asp?galleryID=15	Air Force Link – Photos	**18**
19	http://commons.wikimedia.org/wiki/B-52_Stratofortress	B-52 Stratofortress - Wikimedia Commons	**19**
20	http://us.imdb.com/Title?0278954	B-52 (2001) B-52 on IMDb:	**20**

Now Keyword Frequency based ranking approach is applied and again search results containing id, url, title and ranking are listed in TABLE II.

TABLE II.
FREQUENCY BASED RANKING

Id	url	Title	Proposed System Rank
1	http://www.boeing.com/defense-space/military/b52-strat	Boeing: Integrated Defense Systems - B-52 Stratofortress	**4**
2	http://www.fas.org/nuke/guide/usa/bomber/b-52.htm	B-52 Stratofortress - United States Nuclear Forces	**13**
3	http://www.theb52s.com/	The B-52's : World's Greatest Party Band	**5**
4	http://en.wikipedia.org/wiki/B-52_Stratofortress	B-52 Stratofortress - Wikipedia, the free encyclopedia	**10**
5	http://www.af.mil/factsheets/factsheet.asp?fsID=83	Factsheets : B-52 Stratofortress :	**15**
6	http://www.b-52pro.com/	b-52	**16**
7	http://www.globalsecurity.org/wmd/systems/b-52.htm	B-52 Stratofortress	**1**
8	http://www.stratofortress.org/	B52 Stratofortress Association	**8**
9	http://www.dfrc.nasa.gov/Gallery/Photo/B-52	NASA Dryden B-52 Photo Collection	**9**
10	http://en.wikipedia.org/wiki/B-52_(cocktail)	B-52 (cocktail) - Wikipedia, the free encyclopedia	**11**
11	http://www.rottentomatoes.com/m/1111762-b52/articlegate.php	B-52 - Rotten Tomatoes	**12**
12	http://www.britannica.com/eb/article-9011572/B-52	B-52 Encyclopaedia Britannica	**18**
13	http://www.amazon.com/B-52s/dp/B000002KKD	Amazon.com: The B-52's: Music: The B-52's	**19**
14	http://www.globalsecurity.org/wmd/systems/b-52-upgrade.htm	B-52 Stratofortress	**3**
15	http://youtube.com/?v=exdjbeHn-_8	YouTube - B-52	**7**
16	http://www.boeing.com/defense-space/military/b52-strat/b52_50th/index.html	Boeing: News Feature - B-52 50th Anniversary -	**14**
17	http://www.b-52pro.com/start.html	Welcome to B-52 Professional	**20**

18	http://www.af.mil/photos/index.asp?galleryID=15	Air Force Link – Photos	**2**
19	http://commons.wikimedia.org/wiki/B-52_Stratofortress	B-52 Stratofortress - Wikimedia Commons	**6**
20	http://us.imdb.com/Title?0278954	B-52 (2001) B-52 on IMDb:	**17**

The same set of document is given to different users to compare the system results against user ranking. The details of the results are as follows.

4.5 Performance Evaluation

Performance evaluation of the proposed approach is done based on classification context scenario.[29] Precision plays a major role in performance evaluation. Precision measure is calculated based on the formula Precision = $\frac{tp}{tp+fp}$

Where

tp – True Positive (Correct result)
fp – False Positive (Unexpected Result)

In this proposed work sample Dataset TABLE I. is consider for evaluation purpose and top 20 documents that are more relevant to the user based on user decision is classified manually with different users . Now the same relevant dataset is evaluated against retrieved dataset. Comparison results of the proposed approach are given in the TABLE III.

TABLE III.
RANKING COMPARISION

Id	url	Search Engine Rank	Proposed System Rank	Manual Rank
1	http://www.boeing.com/defense-space/military/b52-strat	**1**	**4**	**4**
2	http://www.fas.org/nuke/guide/usa/bomber/b-52.htm	**2**	**13**	**13**
3	http://www.theb52s.com/	**3**	**5**	**3**
4	http://en.wikipedia.org/wiki/B-52_Stratofortress	**4**	**10**	**10**
5	http://www.af.mil/factsheets/factsheet.asp?fsID=83	**5**	**15**	**15**
6	http://www.b-52pro.com/	**6**	**16**	**16**
7	http://www.globalsecurity.org/wmd/systems/b-52.htm	**7**	**1**	**1**
8	http://www.stratofortress.org/	**8**	**8**	**8**
9	http://www.dfrc.nasa.gov/Gallery/Photo/B-52	**9**	**9**	**9**
10	http://en.wikipedia.org/wiki/B-52_(cocktail)	**10**	**11**	**11**
11	http://www.rottentomatoes.com/m/1111762-b52/articlegate.php	**11**	**12**	**18**
12	http://www.britannica.com/eb/article-9011572/B-52	**12**	**18**	**12**

13	http://www.amazon.com/B-52s/dp/B000002KKD	**13**	**19**	**19**
14	http://www.globalsecurity.org/wmd/systems/b-52-upgrade.htm	**14**	**3**	**5**
15	http://youtube.com/?v=exdjbeHn-_8	**15**	**7**	**7**
16	http://www.boeing.com/defense-space/military/b52-strat/b52_50th/index.html	**16**	**14**	**14**
17	http://www.b-52pro.com/start.html	**17**	**20**	**20**
18	http://www.af.mil/photos/index.asp?galleryID=15	**18**	**2**	**2**
19	http://commons.wikimedia.org/wiki/B-52_Stratofortress	**19**	**6**	**6**
20	http://us.imdb.com/Title?0278954	**20**	**17**	**17**

TABLE III represents the matching of manual ranking against proposed approach mismatching of manual ranking against proposed approach. From the table, it is understood that precision of the proposed system is 0.8 out of 1 where as search-engine precision is 0.2 out of 1.

TABLE III contains result for evaluating the proposed approach against performance measure like Precision. The results of the performance measure are plotted in Fig.2.

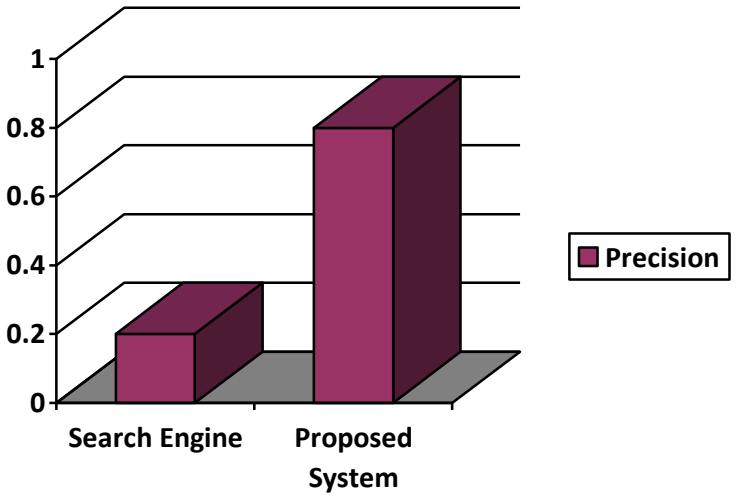


Figure 4 Performance of Proposed System

4. Conclusion

How to improve the quality of search results, making the web pages order of search engine returns meet the user requirements, become a recent research topic of today's SEO (search engine optimization) experts. Therefore in proposed technique an approach is developed in which ranking is made based on the frequency of keywords (search by user) so that it tries to ranks the web pages from most relevant to least relevant web sites or web pages. The ordering of pages in this way increases the relevancy of pages and therefore provides the user with quality search results. As a result, user may find the desired content in the top few pages, Therefore helps to enhance the web search engine results. Proposed methodology focus only on text based mining to rank the relevancy of the web pages where nowadays relevant information may be available in any format like images, audio and video files. Future work will focus on all types of data sets.

References

[1] Ashish Jain, Rajeev Sharma, Gireesh Dixit, Varsha Tomar ," Page Ranking Algorithms in Web Mining, Limitations of Existing methods and a New Method for Indexing Web

Pages", 2013 IEEE International Conference on Communication Systems and Network Technologies.
[2]P.Sudhakar, G.Poonkuzhali, R.Kishore Kumar," Content Based Ranking for Search Engines",Proceedings of the International MultiConference of Engineers and Computer Scientists 2012 Vol I, Hong Kong.
[3] Gyanendra Kumar, Neelam Duhan, A. K. Sharma ," Page Ranking Based on Number of Visits of Links of Web Page" , International Conference on Computer & Communication Technology (ICCCT), 978-1-4577-1386-611, 2011 IEEE.
[4] J. Gou, "Web Content Mining & Structured Data Extraction & Integration", University of Illinois at Urbana- Champign.
[5] Chakrabarti, S., Berg, M., and Dom, B. Focused Crawling: A New Approach to Topic-specific Web Resource Discovery. Computer Networks, Amsterdam, Netherlands, 1999.
[6] Bing Liu, Kevin Chen- Chuan Chang ," Editorial: Special issue on Web Content Mining" , SIGKDD Explorations, Volume 6, Issue 2.
[7] Gibson, J., Wellner, B., Lubar, S, "Adaptive web-page content identification", In WIDM '07: Proceedings of the 9th annual ACM international workshop on Web information and data management. New York, USA,2007.
[8] Georgies Lappas, An overview of web mining in societal benefit areas, The 9th IEEE International Conference on E-Commerce Technology, IEEE 2007.
[9] L. Page, S. Brin, R. Motwani, and T. Winograd. The Pagerank citation ranking: Bring order to the web. Technical report, Stanford University,1998.
[10] Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm" Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR'04) 2004 IEEE.
[11] Xianchao Zhang, Hong Yu, Cong Zhang, and Xinyue Liu "An Improved Weighted HITS Algorithm Based on Similarity and Popularity", 2007IEEE.
[12] Brin, S., and Page, L., 1998. "The anatomy of a large-scale hyper textual Web search engine",Computer Networks and ISDN Systems, Vol. 1-7, pp: 107-117.
[13] Carpineto C., Romano G. Ambient dataset, http://credo.fub.it/ambient/, (2012)
[14] ODP dataset, http://credo.fub.it/odp239/, (2009)
[15] Yiqun Liu, Yupeng Fu, Min Zhang, Shaoping Ma, Liyun Ru. Automatic Search Engine Performance Evaluation with Click-through Data Analysis. In Proceedings of WWW, 2007.
[16] A. Gupta, A. Dixit, A. K. Sharma, "Relevant Document Crawling with Usage Pattern and Domain Profile Based Page Ranking", IEEE, 2013.
[17] Raymond Kosala, Hendrik Blockeel," Web Mining Research: A Survey", SIGKDD Explorations, Volume 2, Issue 1 ACM SIGKDD, July 2000.
[18] Miguel Gomes da Costa ,Júnior Zhiguo Gong," Web Structure Mining: An Introduction", Proceedings of the 2005. IEEE, International Conference on Information Acquisition ,June 27 - July 3, 2005, Hong Kong and Macau, China.
[19] Cooley, R.; Mobasher, B.; Srivastava, J.; "Web mining: information and pattern discovery on the World Wide Web. Tools with Artificial Intelligence", 1997. Proceedings., Ninth IEEE International Conference. Page(s):558 – 567 -3-8 Nov. 1997.
[20] Jianli Duan , Shuxia Liu ,"Research on web log mining analysis",2012 International Symposium on Instrumentation & Measurement, Sensor Network and Automation (IMSNA), 2012 IEEE.
[21] Cheng Wang, Ying Liu, Liheng Jian, Peng Zhang, "A Utility based Web Content Sensitivity Mining Approach", International Conference on Web Intelligent and Intelligent Agent Technology (WIIAT), IEEE/WIC/ACM 2008.
[22] B. Liu. "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data", Springer, 2006.
[23] Andrei Broder , "A taxonomy of web search", In SIGIR Forum, 2002.
]24] Diane Kelly, Jaime Teevan, " Implicit Feedback for Inferring User Preference: A Bibliography", In SIGIR Forum, 2003.
[25] Chongchong Zhao, Zhiqiang Zhang , Hualong Li, Xiaoqin Xie , "A Search Result Ranking Algorithm Based on Web Pages and Tags Clustering", 978-1-4244-8728-8/11/- 2011 IEEE.
[26] Chenhui Wang, Wenchen Wang , Qifan Wu, Xin Wang ," The Optimization of Search Engines Ranking Technology Based On Grey System", 2012 Fourth International Conference on Computational Intelligence and Communication Networks, DOI 10.1109/CICN.2012.41 , 2012 IEEE.
[27] Wei Huang, Bin Li , "An Improved Method for the Computation of PageRank", 2011 International Conference on Mechatronic Science, Electric Engineering and Computer August 19-22, 2011, Jilin, China, 978-1-61284-722-1/11/- 2011 IEEE.
[28] J.Kleinberg,"Authoritative Source in a Hyperlinked Environment,"Proc.ACM-SIAM Symposium on Discrete Algorithm,1998, pp. 668-677.
[29] R.Lempel and S. Moran,"SALSA: The Stochastic Approach for Link- Structure Analysis," ACM Tracsactions on Information Systems,Vol. 19,April 2001,pp. 131-160.
[30] D. Cohn and H. Chang, "Learning to probabilistically identify Authoritative documents," In Proceedings of 17th International Conference on Machine Learning, pages 167–174, Morgan Kaufman , San Francisco, CA, 2000.
[31] Ranveer Singh, Dilip Kumar Sharma," RatioRank: Enhancing the Impact of Inlinks and Outlinks ", 978-1-4673-4529-3/12/- 2012 IEEE.
[32] Faustina Johnson , Santosh Kumar Gupta," Web Content Mining Techniques: A Survey", International Journal of Computer Applications (0975 – 888) Volume 47– No.11, June 2012.
[33] http://en.wikipedia.org/wiki/Web_search_engine.