ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.6

Rank the Search Result by Markov Chain Principle

Sumit Singh Siddhu[1], Dr. Pradeep Pant[2]
M.Tech Student[1], HOD[2]
Department of CS/IT
MIET, Meerut, India

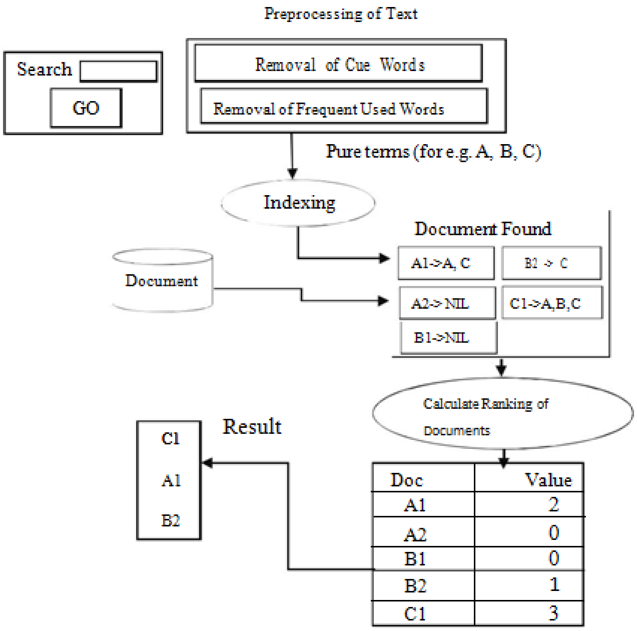
Abstract:
Now a days, search engines are been most widely used for extracting information from various resources throughout the world. PageRank (PR) is an algorithm used by Google Search to rank websites in their search engine results. PageRank was named after Larry Page, one of the founders of Google. PageRank is a way of measuring the importance of website pages. According to Google definition- PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. This thesis proposed an idea for ranking the web documents retrieved by search engine by using the probability based based known as Markov chain principle. This thesis proposes a new and efficient methodology for ranking of web documents. This technique provides relevant results to the user according to their query relevance wise. The methodology first lists the pages that are relevant according to the user query , then prepare a directed graph between the pages , where each page considers as a node and the hyperlink between them considers as a edge, then apply a probability based method to calculate the wattage of each page and presents the result to the user rank wise.

Keywords: search engine, Information Retrieval, indexes, Page Rank.

1. INTRODUCTION

Ranking is an annual performance evaluation method that grades documents on a simple best-to-worst scale to develop a quality work force. Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers). IR can also cover other kinds of data and information problems beyond that specified in the core definition above. The term “unstructured data” refers to data which does not have clear, semantically overt, easy-for-a-computer structure. Web indexing refers to various methods for indexing the contents of a website or of the internet as a whole. Individual websites may use a back-of-the-book index, while search engines usually use keywords to provide a more useful vocabulary for Internet. Advantage “of ranking is that it quickly identifies top performances. With the increasing number of web pages and users on the web, the number of queries Submitted the search engines is also increasing rapidly. Therefore, the search engine needs to be more efficient in its process. The search engines become very popular if they use efficient ranking mechanism. If the search results are not displayed according to the user interest then the search engine will lose its popularity. So the ranking algorithms become very important. Some of the ranking algorithms are Page Rank [PR], Weighted Page Rank (WPR) and Distance Rank”.

2. GENERAL ARCHITECTURE AND DESCRIPTION OF MODEL



Doc	Value
A1	2
A2	0
B1	0
B2	1
C1	3

Figure.1. General Architecture of Proposed Methodology

The system developed is a text based search engine which is capable of extracting the documents. Inputted text processing is initialized at this step which takes in the search text which is analyzed for keywords. Document retrieval is based on the occurrence of terminologies and keywords based on the user search text. Calculate Ranking of documents is based on the values of the Document Found.

1. Pre-processing:
It is done by the following steps-

i) Removal of Cue Words:
Cue words are like is, are, am, was, were, had, it, there etc. The reason behind for removing the cue words from the documents is to save the time during the processing time. When the user enter the keyword in search text box then the system starts processing to search the keyword in the document by ignoring all the cue words, this will help in fast searching.

ii) Removal of Frequent Used Words:
These are the words which are repeating more than once in the document. This is done because during the searching process, system does not access the same word at the same time. After removing these frequently used words, less time will be consumed while searching.

2. Pure Terms:
These are the keywords which we get after the pre-processing like A, B, C. Example- suppose a user enter a keyword like Denial Of Service ,then system starts its processing to find these keywords without accessing cue words and frequent used words. So this A, B, C are the keywords which user wants to search.

3. Document:
These are the documents which are extracted from database based on extracting keywords and terminologies from the documents and making a comparison. If match found then the document are been listed with matched keyword.

Suppose we have documents A1, A2, B1, B2, C1, C2.

4. Document Found:

A1 → A, C

A2 → None

B1 → None

B2 → C

C1 → A, B, C

A B C are the keywords which are found in the documents after searching and preprocessing.
Like A1 document contain A and C keywords, Documents A2 and B1 do not contain any keyword, B2 contain C keyword, C1 contain all the three keywords A B C.

5. Ranking Table:
This table contains Documents and their values.
Example- value of A1 is 2, value of A2 and B1 is Null, value of B2 is 1, value of C1 is 3.
This table does not contain the value of C2.

6. Result:
All the documents will be shown that contain these keywords. Result is display according to the document priority. The document which has highest priority will display first and vice versa. Example-document C1 is displaying first in the result table because it has higher rank as compare to other documents, then A1 is displaying because its value is 2. In the end, B2 is displayed because its rank is lowest. A2 and B1 are not displaying because there is no value of them.

Filtering the search result
The final output of result is presented by applying the wattage fraction to the final outcome, the wattage of the web pages are calculated by applying Markov chain principle. The procedure is illustrated in the figure below-

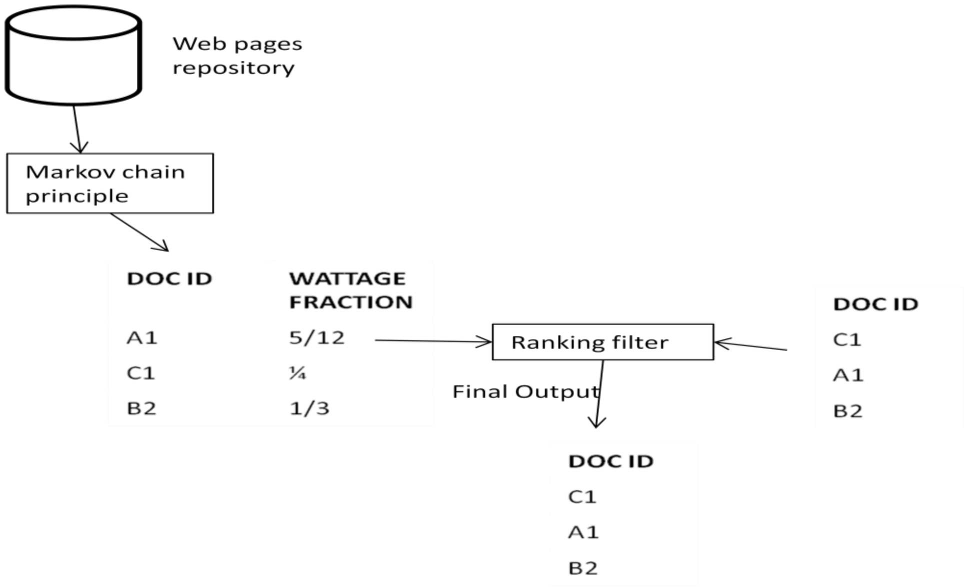


Figure.2. Filtration of search result

The web documents are stored in the web pages repository, after applying Markov principle a table is created containing a doc id and its corresponding wattage fraction, which is further utilized by the ranking filter to sort the documents according to their ranking as decided by the Markov principle.

4.3 Experimental Results and Analysis
In the table 4.1, there are some query terms. Based on these terms, the no of pages containing maximum no of query terms using Page ranking algorithm and the no of documents from where result will be fetched are calculated. Also it showing the

simulation of Page ranking algorithm and the performance of Page ranking. For example- Query Term IEEE found in 10 documents, 7 pages containing max no of query terms by using by using page ranking and 5 pages without using page ranking algorithm.

Table.1. Ranking of Documents

Query Term	No of documents	With using page ranking	Without using page ranking
		No of pages containing maximum no of query terms	
IEEE	10	7	5
Deadlock	20	9	6
Operating System	30	22	21
Distributed System	40	30	26
Computer network	50	42	42
Ethernet	60	50	45
DBMS	70	62	60
Information Technology	80	77	75
Mutual Exclusion	90	70	70
Wireless LAN	100	60	57

Graph shows that by using Page ranking algorithm, better result will be found as compare to without using Page ranking algorithm

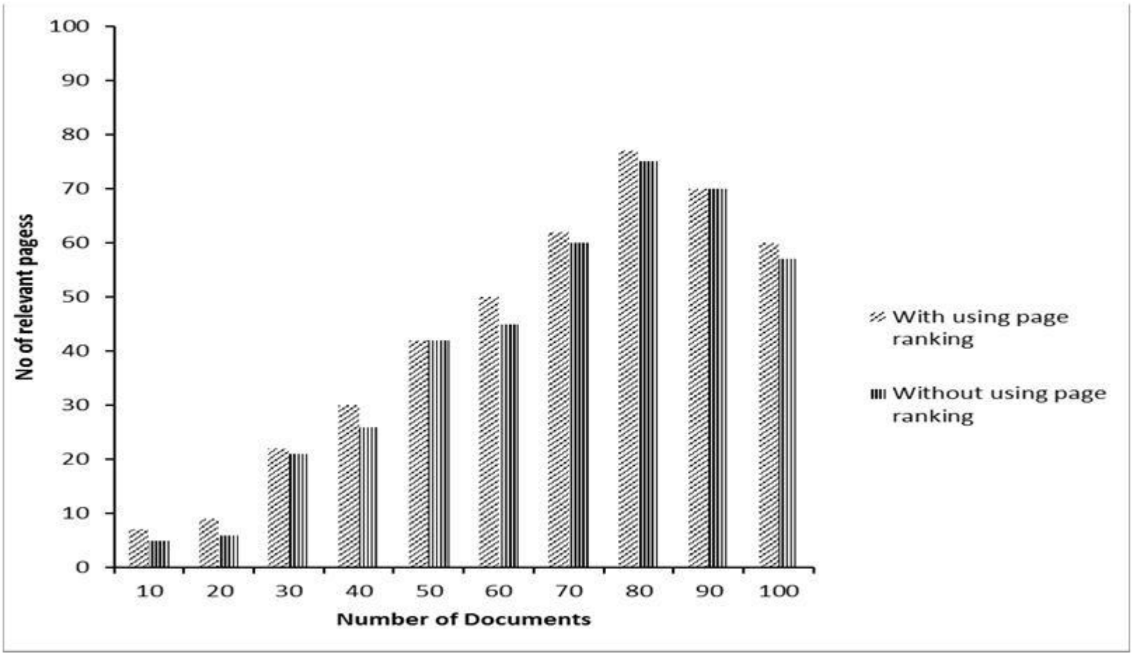


Figure.3. Graph of Ranking of Page

5. CONCLUSIONS & FUTURE WORK

This paper suggests the efficient technique for ranking the documents to show the relevant result according to the query. This is a term based ranking algorithm which uses the important keywords for ranking the documents. Ranking can be used by a search engine to better estimate the quality and relevance of the web page. The proposed model possesses a number of advantages in easier and faster ranking of the document after removing the frequent used words and cue words. This system finds the query terms in the documents and also improve the ability of users to locate relevant information on high quality Web documents become increasingly important. Ranking of the documents is based on the priority of the keywords in the documents. The document having the highest priority will display first and vice versa. The preprocessing of text end the indexing of the documents can be done in parallel for fast execution of the system. Images are also used for further suggesting ranking.

6. REFERENCES

[1]. Jayanthi Manicassamy et al /International Journal on Computer Science and Engineering Vol.1(2),2009,111-115

[2]. R. Agrawal, C. Aggarwal, and V. V. V. Prasad. Depth-first generation of large item sets for association rules. Technical Report RC21538, IBM Technical Report, October 1999.

[3]. R. Agrawal, C. Aggarwal, and V. V. V. Prasad. A tree projection algorithm for generation of frequent item sets. Journal of Parallel and Distributed Computing, 61(3):350–371, 2001.

[4]. M .Ankerst , M. Breunig, H. Kriegel, and J. Sander. Optics: Ordering points to identify the clustering structure. In 1999 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'99), pages 49–60, Philadelphia, PA, June 1999.

[5]. F. Beil, M. Ester, and X. Xu. Frequent term-based text clustering. In Proc. 8th Int. Conf. on Knowledge Discovery and Data Mining (KDD)'2002, Edmonton, Alberta, Canada, 2002. http://www.cs.sfu.ca/˜ ester/publications.html.

[6]. S. Chakrabarti. Data mining for hypertext: A tutorial survey. SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM, 1:1–11, 2000.

[7]. M. Charikar, C. Chekuri, T. Feder, and R. Motwani. Incremental clustering and dynamic information retrieval. In Proceedings of the 29th Symposium on Theory Of Computing STOC 1997, pages 626–635, 1997.

[8]. P. Domingos and G. Hulten. Mining high-speed data streams. In Knowledge Discovery and Data Mining, pages 71–80, 2000.

[9]. R. C. Dubes and A. K. Jain. Algorithms for Clustering Data. Prentice Hall College Div, Englewood Cli®s, NJ, March 1998.

[10]. S. Guha, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams. In IEEE Symposium on Foundations of Computer Science, pages 359–366, 2000.

[11]. M. G. da Gomes Jr. and Z.Gong, "Web Structure Mining: An Introduction", Proceedings of the IEEE International Conference on Information Acquisition, 2005.

[12]. N. Duhan, A. K. Sharma and K. K. Bhatia, "Page Ranking Algorithms: A Survey, Proceedings of the IEEE International Conference on Advance Computing, 2009.

[13]. R. Kosala, H. Blockeel, "Web Mining Research: A Survey", SIGKDD Explorations, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining Vol. 2, No. 1 pp 1-15, 2000.

[14]. R. Cooley, B. Mobasher and J. Srivastava, "Web Minig: Information and Pattern Discovery on the World Wide Web". Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence, pp. (ICTAI'97), 1997.