A Comparative Study of Various Page Ranking Algorithms

Hiral Y. Modi
Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

Prof. Meera Narvekar
Department of Computer Engineering
Dwarkadas J. Sanghvi College of Engineering
Mumbai, India

Abstract**—** *With the rapidly expanding Web, users get disoriented in the rich hyper structure. To provide the internet users with relevant information to satisfy their requirements is the main goal of website owners. Therefore, it is very crucial to find and retrieve the relevant data on the Internet and also find out the user's interests and needs from their behavior. When a user enters a query in a search engine, quite a large number of pages are generally referred in response to user's query. To aid users to navigate in the search result list, various ranking methods are applied on the search results. Most of the existing search engines use Page Rank algorithm to find the relevant documents to the given query. This paper deals with analysis and comparison of various page ranking algorithms to find out their pros and cons. Based on the analysis of different web page ranking algorithms, a comparative study is done to find out their relative advantages and limitations to determine the further scope of research in the field of web page ranking algorithm.*

Index Terms—Page Ranking, HITS, WPR, CARE, Weighted Page Rank Based on Visits of Links, KOMOS.

INTRODUCTION

The World Wide Web (Internet) is the most popular and interactive medium to disseminate information. The Web is huge, diverse, dynamic and widely distributed global information service center. As on today World Wide Web is the largest information achieve for knowledge reference. WWW consists millions of web pages and a huge amount of information is available within these web pages. In order to fetch required information from WWW, search engines perform number of tasks. In such a case, it is the moral responsibility of internet service provider to provide accurate, relevant and quality information to the user in response to their query submitted to the search engine.

Web Search Engines provides a means for information access on the web. On entering a query by the user, search engines will provide list of titles and snippets from the resultant pages based on the rank of each result. As the information available online is huge, seeking for the relevant document involves the user to go through many titles and snippets. Page Rank Algorithm performs ordering of these titles and snippets returned by existing search engine.

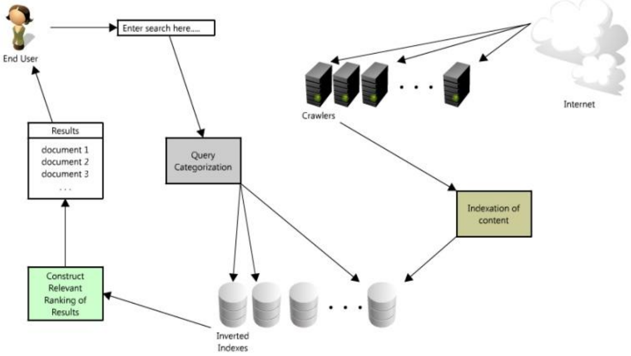


Figure 1: Working of Search Engine

To help users navigate the result list, various ranking methods are applied on the search results. The search engine deploys these ranking techniques to sort the results to be displayed to the user. In this manner an internet user can find the most useful and relevant result on the top of the list. There is a whole range of algorithms developed. This paper aims at analyzing the important page ranking algorithms for finding out their relative strengths, limitations and providing a future direction for the research in the field of web page ranking algorithms.

The remaining part of this paper is organized as follows: Related work is summarized in Section II. A tabular summary is presented in section III, which summarizes the techniques, advantages and limitations of some of the important web page ranking algorithms. A conclusion is given in section IV.

RELATED WORK

Web mining is an approach that classifies the web pages and internet users by considering the contents of the page and usage trends of internet user in the past. Web mining includes Web Content Mining (WCM), Web Structure Mining (WSM), and Web Usage Mining (WUM). WCM is the mining, extraction and integration of useful data, information and knowledge from Web page content. WSM discovers relationship between web pages by analyzing web hypertext structures.WUM is the process of extracting useful information from server logs. [1][2][3][4][5].

The various Page Ranking Algorithms are as given below.

HITS

Hyperlink-Induced Topic Search (HITS) . [15] (also known as collection of hubs and authorities) is based on concept of link analysis that ranks Web pages and is a predecessor to Page Rank Algorithm. HITS is a search query dependent algorithm that ranks the web page by processing its inlinks and outlinks. An ideal hub represents a page that points to many other pages, and an ideal authority represents a page that is in turn linked by many different hubs. The algorithm thus assigns two scores for each page: its authority value, which evaluates the value of the content of the page, and its hub value, which projects the value of its links to other pages. The hubs and authorities created by HITS are illustrated in the Figure 1 below. Authorities and hubs share a mutually emphasizing relationship: a better hub points to many good authorities, and a better authority is pointed to by many good hubs.

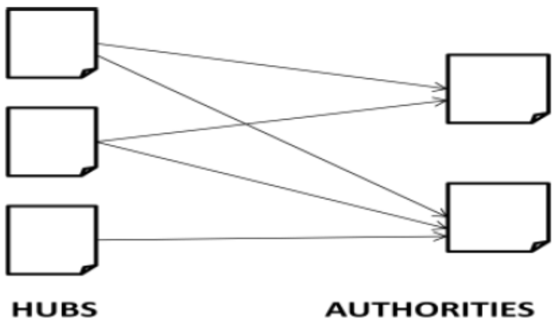


Figure 2: Hubs and Authorities

In the first step of the HITS algorithm the root set (most relevant pages to the query) is generated by considering the top n pages returned by a text-based search algorithm. A base set is generated by expanding the root set with all the web pages that are linked from it and some of the pages that link to it. The web pages in the base set and all hyperlinks among those pages form a focused sub graph. The HITS performs computation on this focused sub graph. According to Kleinberg, the reason for constructing a base set is to ensure that most of the strongest authorities are included. The following algorithm is used for calculating Hub score and Authority score for a node :

- Start by considering each node tha has both, hub score and authority score as 1.
- Apply the Authority Update Rule
- Apply the Hub Update Rule
- Normalize the values by dividing each Hub score by the sum of the squares of all Hub scores, and dividing each Authority score by the sum of the squares of all Authority scores.
- Repeat from the second step as necessary.[9]

A. Page Rank Algorithm

Page Rank algorithm is a widely used ranking algorithm by the Google search engine to rank web pages in its results. Page Rank was developed at Stanford University by Larry Page and Sergey Brin in 1996 as part of a research project which aimed at developing a new search engine. Page Rank uses the Link structure of the web to determine the importance of web pages. Page Rank orders the search results so that documents that are more relevant are placed on top of the search result list.[6] Page Rank is based on the concept that if a page contains important links towards it (inlinks) then the links of this page towards the other page (outlinks) are also to be considered as important pages. Apart from this, the Page Rank Algorithm also considers the back link for determining the rank score. A simplified version of Page Rank is given by:

$$\boldsymbol{PR(u) = c \sum_{v \in B(u)} \frac{PR(v)}{N_v}}$$

where u represents a web page, B(u) is the set of pages that point to u, PR(u) and PR(v) are rank scores of page u and v respectively, Nv denotes the number of outgoing links of page v, c is a factor used for normalization.
The modified version of PageRank is given by :

$$\boldsymbol{PR(u) = (1-d) + d \sum_{v \in B(u)} \frac{PR(v)}{N_v}}$$

Where d is a dampening factor that is frequently set to 0.85. d can be thought of as users following the direct links and (1 –d) as the page rank distribution from non- directly linked pages [7][8].

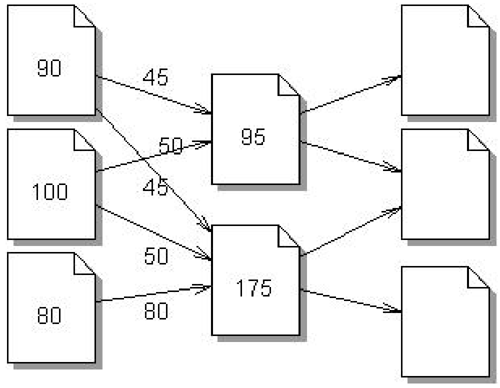


Figure 3 : Simplified Page Rank

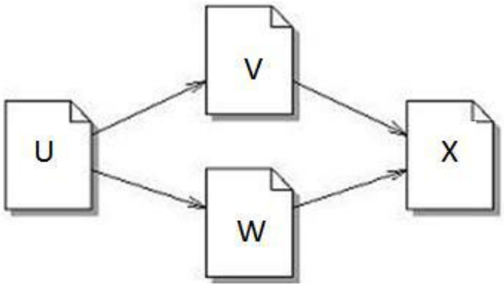


Figure 4: Backlinks

B. Weighted Page Rank Algorithm

The more prominent web pages are, the more interconnection other web pages tend to have with them or are linked to by them. Weighted Page Rank Algorithm– an extension to Page Rank algorithm-assigns larger rank values to more important (popular) pages. WPR considers both the inlinks and the outlinks of the pages and allocates rank scores based on the importance of the pages. WPR provides comparatively better performance than the traditional Page Rank algorithm in terms of returning larger number of relevant pages to a given query. Each outlink page gets a value proportional to its popularity (its number of inlinks and outlinks). The popularity from the number of inlinks and outlinks is recorded as W_{in} (v,u) and W_{out} (v,u), respectively. W_{in} (v,u) is the weight of link(v, u) calculated based on the number of inlinks of page u and the number of inlinks of all reference pages of page v.

$$W^{in}_{(v,u)} = \frac{I_u}{\sum_{p\in R(v)} I_p}$$

where I_u and I_p represent the number of inlinks of page u and page p, respectively. R(v) denotes the reference page list of page v. W_{out} (v,u) is the weight of link(v, u) calculated based on the number of outlinks of page u and the number of outlinks of all reference pages of page v.

$$W^{out}({}_{v,u}) = \frac{O_u}{\sum_{p\in R(v)} O_p}$$

where O_u and O_p represent the number of outlinks of page u and page p, respectively. R(v) denotes the reference page list of page v.
Considering the importance of pages, the conventional Page Rank formula is modified as : [10]

$$PR(u) = (1 - d) + d \sum_{v\in B(u)} PR(v) W^{in}_{(v,u)} W^{out}({}_{v,u})$$

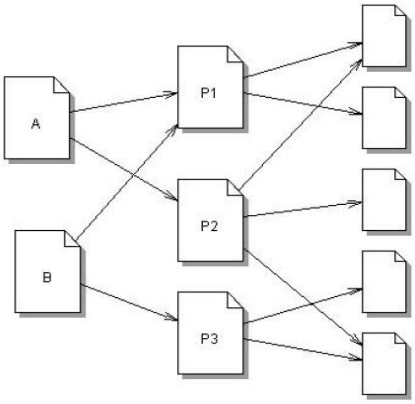


Figure 5 : Weighted Page Rank

C. *Weighted Page Rank Based On Visits Of Links (VOL)*

WPR based on VOL is a new algorithm in which user's usage trend is considered. Most of the ranking algorithms are either link or content oriented in which user usage trends (browsing behaviour) are not considered. Page Ranking based on Visits of Links (VOL) is developed for search engines, which works on the basic Page Rank algorithm of Google, and also takes into consideration number of visits of inbound links of web pages. This concept is very useful to display most relevant pages on the top of the result list on the basis of user's browsing behavior, which reduce the search space to a large extent. In this algorithm most visited outgoing links are assigned higher rank value. In this way a page rank value is computed based on visits of inbound links.
The modified version based on VOL is given as

$$PR(u) = (1 - d) + d \sum_{v\in B(u)} \frac{L_u \, PR(v)}{TL(v)}$$

Notations are :

- d is a dampening factor ,
- u represents a web page,
- B(u) is the set of pages that point to u,
- PR(u) and PR(v) are rank scores of page u and v respectively,
- Lu is the number of visits of link which is pointing page u from v.
- TL(v) denotes total number of visits of all links present on v. [10]

D. *KOMOS-A Keyword Occurrence Method for Ordering Search Results*

The Web Pages ordered by Page Rank algorithm creates a problem of finding the relevant documents in a returned list of document. Since the algorithm computes ranks based on the links and frequency of access, it may happen that for some documents the frequency of access may be small since its occurrence is at the end of the returned search result documents list. But in contrast it may be quite relevant to the given query. Such a problem arises since the user of the search engine usually tends to go through only the first 10-20 web pages from the returned results. To mitigate this problem the documents should be ranked based on the occurrence of a keyword in domain name, URL, title tag, Meta tag and inside the document in addition to the popularity of the documents.
Normally, document relevant to the given query will have the keyword in the root domain name, URL, TITLE tag and META tag and have high term frequency. This technique sorts the search results so that most relevant documents are placed at the beginning of the list. The user can select the needed document quickly so that the search time can be reduced.

Algorithm :

- Extract keywords from the user's query
- Create a list L1 of documents that will contain the keyword in its root domain name

- Create a list L2 of documents that will contain the keyword in its URL
- Create a list L3 of documents that will contain the keyword in its title tag
- Create a list L4 of documents that will contain the keyword in its meta tag
- Create a list L5 of documents that are not included in the above lists L1 to L4.
- Assign ranks to the documents in each list based on the frequency of a keyword inside the documents [12]

E. Case and relation Based (CARE) Page Rank algorithm

The Web is a large collection of data but computers on their own cannot understand or make any decisions with this data. To make this problem easier Semantic Web is introduced. The Semantic Web is an extension of the World Wide Web, in which information is given well defined meaning, better enabling computers and people to work in cooperation.

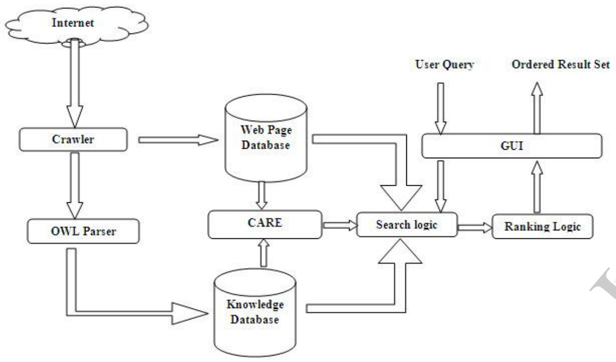


Figure 6: System Architecture of CARE based Search Engine

A web-crawler gathers the Web pages on the Internet in conjunction with its semantic mark (RDF label) and corresponding ontology, which is stated in a Web Ontology Language (OWL) document in the Internet. The collected Web pages are stored in a Web page database for retrieving URLs and corresponding Web pages in future. The ontology, OWL document, is conveyed to an OWL parser. The mapping of the ontology into CBR and then to a relational database is carried out by the OWL Parser. The RDF label uses a formal method to annotate Web pages. First the "CARE" algorithm will analyze the keyword combination input by the user. "CARE" will then search the CBR for the past assembled relations from the ontology. If not found then it will assemble the concepts to some concept pairs and send these pairs to the ontology database to retrieve all relations defined by ontology between concept pairs. After all relations between concept pairs are retrieved from the ontology database, it generates a concept-relation graph based on these relations and concepts. In the concept-relation graph, ellipses represent concepts, words near the ellipses represent the corresponding keywords, arcs depicts relations between two concepts, and numbers accompanying the arcs represent how many relations are present in the ontology database, as shown in Figure 7.

In Figure 7, the ellipses represents the concepts "destination" and "accommodation," and the nearby words "Blue Mountains" and "Europe" represent the corresponding keywords of the concepts "destination" and "accommodation," respectively. And, the number "3" near the arc between the two concepts represents that there are three relations in the ontology database.

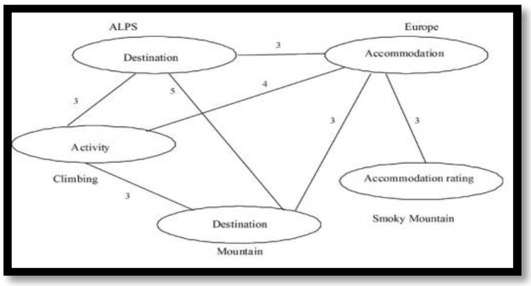


Figure 7 : A Concept-Relation Graph

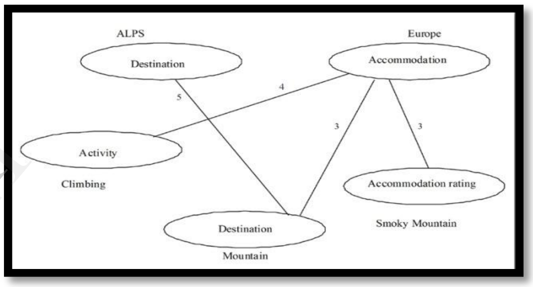
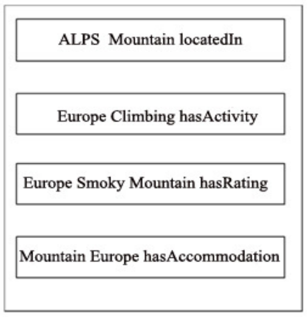


Figure 8 : A Concept-Relation Sub graph

In the next step, some arcs are eliminated from the graph and thus diversified sub graphs are formed, as shown in Figure. 8. In each sub graph, there are some quantitative relations between the concepts. The larger the number neighboring the arc the more is the relation existing between the concepts. Finally, the system fetches the relation and its corresponding keyword pair from each arc in sub graphs to form a property-keyword candidate set.


ALPS Mountain locatedIn

Europe Climbing hasActivity

Europe Smoky Mountain hasRating

Mountain Europe hasAccommodation

Figure 9 : Property-keyword candidate set

Then, the property-keyword candidate set is transferred to the database to get a relevant result set for the users.

Figure. 9 shows a property keyword candidate set formed from a sub graph shown in Figure 8.

The CARE algorithm uses Textual Case Based Reasoning (TCBR) [14] and Relation-based Page Ranking algorithms. The textual case based reasoning helps reduce the number of irrelevant result pages during the search process. Textual case based reasoning uses previous knowledge of the search results. By using this it fetches the results which are more relevant to search query. The results are promising in terms of relevancy, time complexity and accuracy [13].

III. COMPARISON OF VARIOUS WEB PAGE RANKING ALGORITHMS

TABLE 1 : COMPARING VARIOUS PAGE RANKING ALGORITHMS

ALGORITHM	**HITS**	**PAGE RANK**	**WEIGHTED PAGE RANK**	**WPR BASED ON VISITS OF LINKS**	**KOMOS**	**CARE**
MAIN TECHNIQUE	Web Structure Mining, Web Content Mining	Web Structure Mining	Web Structure Mining	Web Structure Mining, Web Content Mining, Web Usage Mining	Web Content Mining	Web Structure Mining
METHODOLOGY	The hubs and authority of the relevant pages are determined . It provides relevant as well as important page as the result.	The score for Pages is computed at the time of indexing of the pages.	Weight of web page is calculated on the basis of inlinks and outlinks and on the basis of weight the importance of page is decided.	a page rank value is calculated based on visits of inbound links. more weightage is given to those web pages which is most visited by users, i.e. those web page which have higher popular inlinks.	documents are ranked according to the occurrence of a keyword in domain name, URL, title tag, meta tag and inside the document in addition to the popularity of the documents.	Ranking of web pages for semantic search engine. It uses the information extracted from the queries of the user and annotated resources.
INPUT PARAMETER	Content, Back and Forward links	Back links	Back links and Forward links.	Content, Back and Forward Links	Keywords	User search query, Keywords
RELEVANCY	More (this algo. Uses the hyperlinks so it will give good results and also consider the content of the page)	Less (this algo. ranks the pages on the indexing time)	Less (as ranking is based on the calculation of weight of the web page at the time of indexing.)	more (it consider the relative position of the pages)	More (As this algorithm works with keywords, relevant documents are displayed before other less relevant documents.)	High (as it is keyword based algorithm so it only returns the result if the keyword entered by the user match with the page.)
QUALITY OF RESULTS	Less than PR	Medium	Higher than PR	Higher than PR	High	High
ADVANTAGES	Returned pages have high relevancy and importance.	Rank is calculated on the basis of the importance of pages.	It gives higher accuracy in terms of ranking because it uses the content of the pages. The pages are sorted according to the importance.	It displays most valuable pages on the top of the result list on the basis of user browsing behavior. rank of the page depends on the probability of visits	Time taken to search the relevant document is reduced.	Well defined semantics with clear interpretation. Efficiently provide answer to quantitative bibliometric questions. It provides promising results in terms of both time complexity and accuracy
LIMITATIONS	Topic drift and efficiency problem	Results come at the time of indexing and not at the query time	It is based only on the popularity of the web page. Relevancy is ignored.	periodic crawling of web servers so as to collect the accurate and up to date visit count of pages. Specialized crawlers need to be designed	This work concentrates on only queries with single keyword.	In this ranking algorithm every page is to be annotated with respect to some ontology.

CONCLUSION

Based on the ranking algorithm used, the algorithms provides a definite rank to resultant web pages. Based on the specific needs of the users, a typical search engine should employ web page ranking techniques. After going through detailed analysis of algorithms for ranking of web pages against the various parameters such as methodology, input parameters, relevancy of results and quality of the results, it is concluded that some of the existing techniques have limitations particularly in terms of time response, accuracy of results and relevancy of results. As a future guidance, the algorithms which equally consider the relevancy as well as importance of a page should be developed which are compatible with global standards of web technology, so that the quality of search results can be improved.

REFERENCES

Wenpu Xing and Ali Ghorbani, "Weighted PageRank Algorithm", In proceedings of the 2rd Annual Conference on Communication Networks & Services Research, PP. 305-314, 2004.

Neelam Duhan, A. K. Sharma and Komal Kumar Bhatia, "Page Ranking Algorithms: A Survey", In proceedings of the IEEE International Advanced Computing Conference (IACC), 2009.

R. Kosala and H. Blockeel, "Web Mining Research: A survey", In ACM SIGKDD Explorations, 2(1), PP.1–15, 2000.

S. Madria, S. S. Bhowmick, W. K. Ng, and E.-P. Lim, "Research Issues in Web Data Mining", In Proceedings of the Conference on Data Warehousing and Knowledge Discovery, PP. 303–319, 1999.

S. Pal, V. Talwar, and P. Mitra, "Web Mining in Soft Computing Framework : Relevance, State of the Art and Future Directions:, In IEEE Trans. Neural Networks, 13(5), PP.1163–1177,2002.

L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", Technical Report, Stanford Digital Libraries SIDL-WP-1999-0120, 1999.

C. Ridings and M. Shishigin, "Pagerank Uncovered", Technical Report, 2002.

Sergey Brin and Larry Page, "The anatomy of a Large-scale Hypertextual Web Search Engine", In Proceedings of the Seventh International World Wide Web Conference, 1998.

Nidhi Grover and Ritika Wason, "Comparative Analysis Of Pagerank And HITS Algorithms", International Journal of Engineering Research & Technology (IJERT), 2012.

Neelam Tyagi and Simple Sharma, "Weighted Page Rank Algorithm Based on Number of Visits of Web Page", International Journal of Soft Computing and Engineering (IJSCE), 2012

Gyanendra Kumar, Neelam Duahn, and Sharma A. K., "Page Ranking Based on Number of Visits of Web Pages", International Conference on Computer & Communication Technology (ICCCT)-2011, 978-1-4577-1385-9.

Poomagal S and Hamsapriya T, "KOMOS-A Keyword Occurrence Method for Ordering Search Results", International Journal of Computer Applications, 2010

MS.N.Preethi and Dr.T.Devi, " New Integrated Case and Relation based (CARE) Page Rank Algorithm", International Conference on Computer Communication and Informatics (ICCCI), 2013

A. Aamodt and E. Plaza, "Case-based reasoning: Foundational issues, methodological variations, and systemapproach," AI Communicatons , vol. 7, no.1, pp. 39–59, 1994

Jon Kleinberg, "Authoritative Sources in a Hyperlinked Environment", In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms, 1998.