

Application of Page Ranking Algorithm Based on Numbers of Link Visits in Web Recommendation System for Online Business

¹Mahesh Kumar Singh, ²Om Prakash Rishi,
and ³Sumit Wadhwa

¹Research Scholar, Department of Computer Science and Informatics,
University of Kota, Rajasthan, India

²Director Research, University of Kota, Rajasthan, India

³Assistant Professor, Department of Information technology & Management,
Greater Noida, U.P., India

maheshkrsg@gmail.com, omprakashrishi@yahoo.com, sumitwadhwa1988@gmail.com

Abstract

World Wide Web plays a vital role in global information service center. Online business is growing very rapidly by creating websites. Website is made of number of webpages. Webpage is dynamic collection of hyperlink and usage information, providing rich sources of web data for web mining. Due to exponential growth of dynamic information over the internet, information overload create big challenges for the researchers in this area. One of the key components which ensure the acceptance of web page search service is the web page ranker a component which is said to have been the main contributing factor of Google. This paper discusses the page ranking algorithms based on contents, structures and usages in web mining, proposed page ranking algorithm based on number of links visit. Since every product has its own unique web page for its decryption hence in selection of top most visited links can be applicable to create the recommendation list for the corresponding customer in online business.

Keywords- PageRank, Web Mining, Recommendation System, Information retrieval, Web Graph, Online business.

I. INTRODUCTION

Web Mining is defined as the application of data mining techniques on the World Wide Web to find hidden information, This hidden information i.e. knowledge could be contained in content of web pages or in link structure of WWW [8] or in web server logs. WWW is a vast resource of hyperlinked and heterogeneous information including text, image, audio, video, and metadata. With the rapid growth of information sources available on the WWW and growing needs of users, it is becoming difficult to manage the information on the web and satisfy the user needs. Actually, we are drowning in data but starving for knowledge. Therefore, it has become increasingly necessary for users to use some information retrieval techniques to find, extract, filter and order the desired information.

A. Search engine [2] receives users query, processes it, and searches into its index for relevant documents i.e. the documents that are likely related to query and supposed to be interesting then, search engine

ranks the documents found relevant and it shows them as results. This process can be divided in the following tasks:

- B. **Crawler** [11,12] is in charge of visiting as many pages and retrieves the information needed from them. The idea is that this information is stored for the use by the search engine afterwards.
- C. **Indexing** the information provided by a crawler has to be stored in order to be accessed by the search engine. As the user will be in front of his computer waiting for the answer of the search engine, time response becomes an important issue. That is why this information is indexed in order to decrease the time needed to look into it.
- D. **Searching** The web search engine represents the user interface needed to permit the user to query the information. It is the connection between the user and the information repository.
- E. **Sorting/Ranking** Due to the huge amount of information existing in the web, when a user sends a query about a general topic (e.g. java course), there exist an incredible number of pages related to this query but only a small part of such amount of information will be really interesting for the user. That is why the search engines incorporate ranking algorithms in order to sort the results.
- F. **Recommendation System** There is individual page for every product and services provided by online business. Page ranking algorithm rank every page which gives the popularity of the web page [9,10] that can be used to filter the most popular product. Today is the era of online businesses. There are billions of small and large companies that provide the online facilities of sale and purchase the product and services. Every product has the different links of its description. Hence PageRank is help to identify the popularity of the product.

This paper is categorized into seven parts. First part discusses the idea about the web mining and its types. Second part is the related study of page ranking algorithms; third part discusses the problems identification, forth part is implementation of page ranking algorithm based on number of link visits, fifth part discusses its application in online business, six part is comparison analysis, and last part discusses the conclusion of the paper.

II. WEB MINING

Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from large databases is called Data Mining. Web Mining is the application of data mining techniques to discover and retrieve useful information and patterns (knowledge) from the WWW documents and services web mining can be divided into three categories [1] as shown in figure 1.

- a. Web Content Mining,
- b. Web Structure Mining, &
- c. Web Usage Mining

- A. **Web Content Mining (WCM)** describes the automatic search of information resources available online, and involves mining web data content. It is emphasis on the content of the web page not its links. It can be applied on web pages itself or on the result pages obtained from a search engine. WCM is differentiated from two different points of view: Information Retrieval (IR) View and Database View. In IR view, most of the researches use bag of words, which is based on the statistics about single words in isolation, to represent unstructured text. For the semi-structured data, all the works utilize the HTML structures insides the documents. For database view, Web mining always tries to infer the structure of the Web site to transform a Web site to become a database.

B. Web Structure Mining (WSM) is used to generate structural summary about the Web sites and Web pages. The structure of a typical Web graph consists of Web pages as nodes and hyperlinks as edges connecting two related pages. Technically, WCM mainly focuses on the structure of inner-document, while WSM tries to discover the link structure of the hyperlinks at the inter-document level. Web structure mining tries to discover the model underlying the link structures of the Web. The model is based on the topology of the hyperlink with or without the link description. This model can be used to categorize the Web pages and is useful to generate information such as similarity and relationships between Web sites. And the link structure of the Web contains important implied information, and can help in filtering or ranking Web pages. In particular, a link from page A to page B can be considered a recommendation of page B by the author of A. Some new algorithms have been proposed that exploit this link structure not only for keyword searching, but other tasks like automatically building a Yahoo-like hierarchy or identifying communities on the Web. The qualitative performance of these algorithms is generally better than the IR algorithms since they make use of more information than just the contents of the pages. While it is indeed possible to influence the link structure of the Web locally, it is quite hard to do so at a global level. So link analysis algorithms that work at a global level possess relatively robust defenses against spamming.

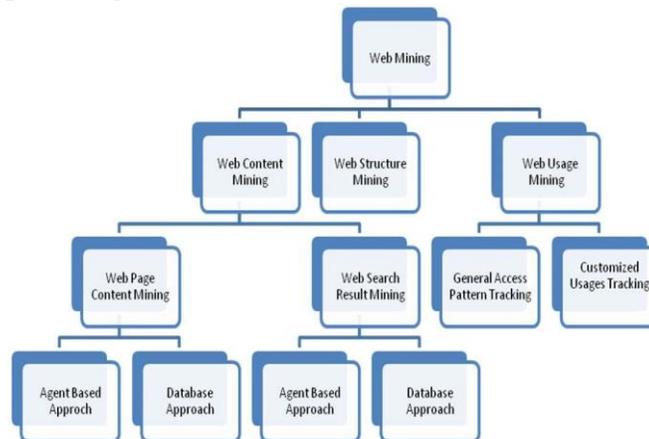


Fig. 1 Taxonomy of Web Mining

C. Web Usage Mining (WUM) tries to discover user navigation patterns from web data and the useful information from the secondary data derived from the interactions of the users while surfing on the Web. It focuses on the techniques that could predict user behavior while the user interacts with Web. This type of web mining allows for the collection of Web access information for Web pages. This usage data provides the paths leading to accessed Web pages. This information is often gathered automatically into access logs via the Web server. CGI scripts offer other useful information such as referrer logs, user subscription information and survey logs. This category is important to the overall use of data mining for companies and their internet/ intranet based applications and information access. The three categories of web mining described above have its own application areas including site improvement, business intelligence, Web personalization, site modification, usage characterization and page ranking etc. The search engines to find more important pages generally use the page ranking. Implement PRNLV method use web structure and web uses mining technique to rank web pages and its application in web recommendation system.

III. RELATED WORK OF RANKING ALGORITHMS

The web is very large and diverse and many pages could be related to a given query. That is why a method/algorithm is used to sort the entire pages subject to be interesting to a user's query. All the algorithms consider the web pages as a directed graph in which pages are denoted as nodes and links are denoted as edges.

A. PageRank Algorithm (PR)

Surgey Brin and Larry Page developed a ranking algorithm used by Google, named PageRank [2] after Larry Page (cofounder of Google search engine), that uses the link structure of the web to determine the importance of web pages [7]. It takes back links into account and propagates the ranking through links. Thus, a page has a high rank if the sum of the ranks of its back links is high. A simplified version of page rank is defined as follows

$$PR(p) = (1 - c) \sum_{q \in I(p)} \frac{PR(q)}{O(q)} \quad (3.1)$$

In the calculation of PageRank a factor c is used for normalization. Note that $0 < c < 1$ because there are pages without incoming links and their weight is lost. Later PageRank was modified observing that not all users follow the direct links on WWW.

$$PR(p) = (1 - d) + d \sum_{q \in I(p)} \frac{PR(q)}{O(q)} \quad (3.2)$$

Where d is a dampening factor that is usually set to 0.85 (any value between 0 and 1), d can be thought of as the probability of users' following the links and could regard $(1 - d)$ as the page rank distribution from non-directly linked pages. Consider the following directed graph [4]

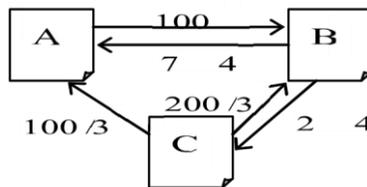


Fig. 2. Example Graph

The PageRanks for pages A, B, C in figure 2 are calculated by using (3.2) with $d=0.5$, the page ranks of pages A, B and C becomes: $PR(A)=1.2$, $PR(B)=1.2$, $PR(C)=0.8$

B. Weighted Page Rank Algorithm (WPR)

Wenpu Xing and Ali Ghorbani [5] proposed an extension to standard PageRank called Weighted PageRank (WPR). It rank pages according to their importance not only consider link structure of web graph. This algorithm assigns larger rank values to more important pages instead of dividing the rank value of a page evenly among its outgoing linked pages. Each outlink page gets a value proportional to its popularity. The popularity is measured by its number of inlinks and outlinks.[6]

$$PR(p) = (1 - d) + d \sum_{q \in I(p)} PR(q) W_{(q,p)}^{in} W_{(q,p)}^{out} \quad (3.3)$$

Where Win (q,p) and Wout (q,p), for inlinks and outlinks is given as

$$W_{(q,p)}^{in} = \frac{I_p}{\sum_{v \in R(q)} I_v} \quad (3.3.1)$$

$$W_{(q,p)}^{out} = \frac{O_p}{\sum_{v \in R(q)} O_v} \quad (3.3.2)$$

Where I_v, I_p and O_v, O_p represent the number of inlinks and outlinks of page v and page p respectively.

The Page Ranks for pages A, B, C are calculated by using (3.3) with $d=0.5$, the page ranks of pages A, B and C are $PR(A)=0.65$, $PR(B)=0.93$, $PR(C)=0.60$.

C. Page Content Rank Algorithm (PCR)

Jaroslav Pokorny and Jozef Smizansky [3] gave a new ranking method of page relevance ranking employing WCM technique, called Page Content Rank (PCR). This method combines a number of heuristics that seem to be important for analyzing the content of web pages. The page importance is determined on the basis of the importance of terms, which the page contains. The importance of a term is specified with respect to a given query q . PCR uses a neural network as its inner classification structure. The importance of a page P in PCR is calculated as an aggregate value of the importance of all terms that P contains. For a promotion of the significant term and a suppression of the others, the second moment is again used as an aggregate function [2] $Page_importance(P) = sec_moment(\{importance(t): t \in P\})$ (3.4)

D. Hyperlinked Induced Topic Search Algorithm (HITS) [13, 24]

This algorithm assumes that for every query topic, there is a set of "authoritative" or "authority" pages/sites that are relevant and popular focusing on the topic and there are "hub" pages/sites that contain useful links to relevant sites including links to many related authorities as shown in figure 3.

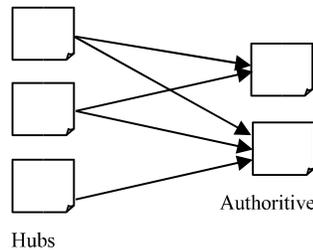


Fig 3. Hubs and Authorities

E. Working of HITS

The HITS works in two phases Sampling and Iterative in the Sampling phase a set of relevant pages for the given query are collected i.e. a sub-graph S of G is retrieved which is high in authority pages. The Iterative phase finds hubs and authorities using the output of the sampling phase using following equations.

$$H_p = \sum_{q \in I(p)} A_q \quad (3.4)$$

$$A_p = \sum_{q \in B(p)} H_q \quad (3.5)$$

Where H_p is the hub weight, A_p is the Authority weight, $I(p)$ and $B(p)$ denotes the set of reference and referrer pages of page p .

IV. PROBLEMS & ISSUES OF THE PAGE RANKING ALGORITHMS

The main problems and issues of discussed page ranking algorithms are summarized as:

- A. **Rank quality of PageRank:** The discussed ranking algorithms have shown a really high quality and the proof is that success of Google (or they are still being used) successfully. However, some improvements can be done on it.
- B. **Data Mining Technique of PageRank:** PageRank algorithm used only Web Structure Mining and Web Content Mining technique; it does not use Web Usage Mining, which may significantly improve the quality of rank of web pages according to users information needs.
- C. **PageRank is Static in Nature:** In PageRank algorithm, the importance or rank score of each page are static in nature. The rank changes only with link structure of web.

V. PAGE RANKING ALGORITHMS BASED ON NUMBERS OF LINK VISITS (PRNLV)

PRNLV (Page Ranking based on Link Visits) based on Web Structure Mining and Usage Mining; it takes the user visits of pages/links into account with the aim to determine the importance and relevance score of the web pages. To accomplish the complete task from gathering the usage characterization till final rank determination many subtasks are performed such as

- A. Storage of user's access information (hits) on an outgoing link of a page in related server log files.
- B. Fetching of pages and their access information by the targeted web crawler.
- C. For each page link, computation of weights based on the probabilities of their being visited by the users.
- D. Final rank computation of pages based on the weights of their incoming links.
- E. Retrieval of ranked pages corresponding to user queries.

A. Calculation of Visits (hits) of links

If p is a page with outgoing-link set $O(p)$ and each outgoing link is associated with a numerical integer indicating visitcount (VC), then the weight of each outgoing link connecting to page p to page o is calculated by

$$Weight_{link}(p, q) = \frac{VC(p, q)}{\sum_{q' \in O(p)} VC(p, q')} \quad (5.1)$$

B Page Rank based on Numbers Link Visit (PRNLV)

If p is a page having inbound-linked pages in set $B(p)$, then the rank (PRNLV) is given equation (5.2)

$$PRLV(p) = (1-d) + d \left(\sum_{b \in B(p)} PRLV(b) Weight_{link}(b, p) \right) \quad (5.2)$$

VI. APPLICATION OF PAGE RANKING IN RECOMMENDATION SYSTEM

Every on line commercial company uses the specified pages for specified product that contains the specification of the product on their web site. Every page has its own unique URL hence the popularity of the product can be computed by the popularity of the links that is page rank of the page. Recommendation System tries to identify the user's interest in the specific domain of contents based on their previous experiences. When a user interact with the E-commercial site he/she offers a set of implicit or explicit information like clicks, rating, comments etc. about his/her taste. Association Rule mining play a vital to associate the e-mail ID with the page with its rank that can be helpful in creating the recommended list for the specified customer.

VII. COMPARISION OF PRNLV WITH PR & WPR

From the table 1 it is clear that PRNLV is much better than PR and WPR.

Table 1: Comparison of Page ranking Algorithms

Algorithm Parameter	PageRank (PR)	Weighted Page Rank (WPR)	Page Rank based on Numbers Link-Visit (PRNLV)
Description	Computes scores at indexing time. Results are sorted according to importance of pages.	Computes scores at indexing time. Results are sorted according to importance of pages.	Computes scores at indexing time. Pages are sorted according to importance and relevance.
Mining Technique Used	Web Structure Mining	Web Structure Mining	Web Structure Mining, Web Usage Mining
Rank	Ranks are	Ranks are	Ranks are unequally
Distribution	equally distributed to outgoing links.	equally distributed to outgoing links.	distributed among outgoing links according to their probabilities of visit.
I/P Parameters	Inbound links of pages	Inbound links and Outbound links of pages	Inbound links, Outbound links, Visit Counts of links.
Working levels	n^*	n^*	N
Complexity	$O(\log n)$	$O(\log n)$	$> O(\log n)$
Nature of Rank	Less dynamic (rank changes with link structure)	Less dynamic (rank changes with link structure)	More dynamic (rank changes with visit counts & structure of links)
Relevancy of pages	No	No	Yes

Importance of pages	Yes	Yes	Yes
Quality of result	Low	High	High
Advantages	Computation of ranks with minimum effort and less complexity.	Computation of ranks with minimum effort and less complexity.	Pages returned are of high quality and relevancy as user feedbacks are taken into account. Search space can be very much pruned as pages are sorted according to users' information needs.
Limitations	No relevancy of pages is considered in rank computation. All links are considered equally important.	No relevancy of pages is considered in rank computation. All links are considered equally important.	Extra effort on crawlers to fetch the visit counts of pages from web servers. Extra calculations to find the weights of links.

VIII. CONCLUSION

Mining of knowledgeable data from a huge amount of data is very complex task, World Wide Web information play a vital role for information collection and sharing. The ranking algorithms are used to search the relevance information in very efficient manner. Different page ranking algorithms are used in different techniques. The PRNLV uses the user's browsing information in consideration to calculate rank of a documents rather than link structure. Due to browsing information in consideration PRNLV system is more dynamic than other ranking algorithms. The popularity of the web pages (products) computed by the page ranking algorithm can be used for performance of the products in the on line business.

REFERENCES

- [1] N. Duhan, A. K. Sharma, K. K. Bhatia, "Page Ranking Algorithms: A Survey", 2009 IEEE International Advance Computing Conference (IACC 2009) Patiala, India, 6-7 March 2009.
- [2] A Broder. A taxonomy of web search. Technical report, IBM Research, 2002.
- [3] Jaroslav Pokorny, Jozef Smizansky, Page Content Rank: An Approach to the Web Content Mining.
- [4] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, R. Stata. Graph structure in the web. In In Proceedings of the 9th International World Wide Web Conference, 2000.
- [5] W Xing,A. Ghorbani, "Weighted PageRank Algorithm", Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR'04), 2004 IEEE
- [6] G. Salton . C. Buckley. "Term Weighting Approaches in Automatic Text Retrieval. In Information Processing and Management. Vol. 24, No. 5, pp. 513– 523 1998.
- [7] S. Chakrabarti, B. E. Dom, S. R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, D. Gibson, and J. Kleinberg, Mining the Web's link structure. Computer, 32(8):60–67, 1999.
- [8] T. B.-Lee, R. Cailliau, A. Luotonen, H.F. Nielsen, A. Secret. "The World-Wide Web. Communications of the ACM", 37(8):76–82, 1994.
- [9] M.G Campana, F. Delmastro:" Recommender Systems for online and mobile Social Networks: A survey", Online Social Networks and Media3-4(2017) , pp 75-97

- [10] Q. Su , L. Chem:” A Method for discovering cluster of e-Commerce in terest patterns using Click- stream data”. Electronic Commerce Research and Applications. 1567-4223 Elsevier-2014 <http://dx.doi.org/10.1016/j.elerap.2014.10.002>
- [11] B. Pinkerton.” Finding what people want: Experiences with the web crawler”. In The second Internation WWW Conference Chicago, 1994.
- [12] J. Cho, H. Garc´ya-Molina, L. Page. “Efficient crawling through URL ordering”. Computer Networks and ISDN Systems, 30(1–7):161–172, 1998.
- [13] C. Ding, X. He, P. Husbands, H. Zha, H. Simon. “Link analysis: Hubs and authorities on the world”, Technical report:47847, 2001.
- [14] A. Borodin, G. O. Roberts, Jeffrey S. Rosenthal, P. Tsaparas. Finding authorities and hubs from link structures on the world wide web. In World Wide Web, pages 415–429, 2001.
- [15] S. L. Yong, M. Hagenbuchner Ah Chung Tsoi, “Ranking Web Pages using Machine Learning Approaches”, IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008.

Author’s Profile



Mahesh Kumar Singh received his B.Tech from UPTU Lucknow in 2004 , M.Tech. from Jamia Hamdard University, New Delhi in 2012 and Perusing Ph D in Computer Science, Kota University on web mining and its application to web business intelligence.



Dr. Om Prakash Rishi is Director of Research University of Kota and Head of Department of Computer Science & Informatics. His Research areas are Artificial Intelligence, Intelligent Systems & CBR, Software Systems and Agent technology, Cloud Computing, Information Security.



Sumit Wadhwa received his B.Tech from Kurukshetra University, Kurukshetra in 2009 , M.Tech. from Kurukshetra University, Kurukshetra, in 2013 and research areas are Wireless Sensor Network & Artificial intelligence.