Study and Analysis of Page Ranking Algorithms in Web Structure Mining

RenuKumari, MamtaYadav
[#1]M. Tech. Scholar,[#2]Assistant Prof., CSE Department

Abstract
The World Wide Web consists billions of web pages and huge amount of information available within the web pages. To retrieve required information from World Wide Web, search engines perform number of tasks based on their respective architecture and use various ranking algorithms for getting the desired result. To support the users to navigate in the result list, various ranking methods are applied on the search results. Some basic algorithms are Page Rank Algorithm, Weighted Page Rank Algorithm, and HITS. All these algorithms are used to discover more relevant pages on the top of the result-list This new ranking mechanism is known as Weighted Page Rank Algorithm based on Visits of Links (VOL).The original Weighted Page Rank algorithm (WPR) takes into account the importance of both the in links and out links of the web pages and distributes rank scores based on the popularity of the pages. The original Weighted Page Rank Algorithm is based on the popularity (importance) of in links and out links of a hyperlinked web graph. It calculates relevancy of a web page higher than the standard Page Rank algorithm, which is used by famous search engine Google. This technique increases the relevancy score than the existing one. So, this concept is very useful to display most valuable pages on the top of the result list on the basis of user's browsing behavior, which reduce the search space to a large scale.

Keywords: *WWW, PAGE, Techniques, PRA, WPR, HITS.*

I. INTRODUCTION

World Wide Web (Web) is popular and interactive medium to disseminate information today. The Web is huge, diverse, dynamic, widely distributed global information service center. As on today WWW is the largest information repository for knowledge reference.
Users could encounter following problems when interacting with the Web:

a) Finding relevant information
Most people use some search service when they want to find specific information on the Web. A user usually inputs a simple keyword query and a result is a list of ranked pages. This ranking is based on their similarity to the query. Today's search tools have some problems: Low precision and low recall, mainly because of wrong or incomplete keyword query. This leads to irrelevance of many search results.

b) Creating new knowledge
This problem is data-triggered process that presumes that we have a collection of Web data and we want to extract potentially useful knowledge from these data.

c) Personalization of information
People differ in the contents and presentations they prefer while interacting with the Web.

d) Learning about consumers or individual users
This is a group of sub-problems such as mass customizing information to intended consumers, problems related to effective Website design and management, problems related to marketing and others.
Web mining techniques could be used to solve information overload problems above. With the rapid growth of WWW and the user's demand on knowledge, it is becoming more difficult to manage the information on WWW and satisfy the user needs. Therefore, the users are looking for better information retrieval techniques and tools to locate, extract, filter and find the necessary information. Most of the users use information retrieval tools like search engines to find information from the WWW. There are tens and hundreds of search engines available but some are popular like Google, Yahoo, Bing etc., because of their crawling and ranking methodologies. The search engines download, index and store hundreds of millions of web pages. They answer tens of millions of queries every day. So Web mining and ranking mechanism becomes very important for effective information retrieval.

(i)Overview of Web Mining
Web mining is the Data Mining technique that automatically discovers or extracts the information from web documents. It is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web.

(ii)Web Mining Process

The complete process of extracting knowledge from Web data [44] is follows in Fig.1

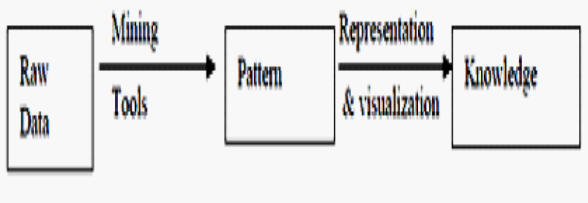


Fig. 1 Web Mining Process

The various steps are explained as follows.

1. Resource finding: It is the task of retrieving intended web documents.

2. Information selection and pre-processing: Automatically selecting and pre- processing specific from information retrieved Web resources.

3. Generalization: Automatically discovers general patterns at individual Web site as well as multiple sites.

4. Analysis: Validation and interpretation of the mined patterns.

In general web mining tasks are:

I. Mining web search engine data
II. Analyzing the web's link structures
III. Classifying web document automatically
IV. Mining web page semantic structure and page contents
V. Mining web dynamics
VI. Personalization.

(iii)Web Mining Categories

Web content mining (WCM), Web structure mining (WSM), and Web Usage Mining (WUM). Web content mining refers to the discovery of useful information from web contents, including text, image, audio, video, etc. Web structure mining studies the web's hyperlink structure. It usually involves analysis of the in-links and out-links of a web page, and it has been used for search engine result ranking. Web usage mining focuses on analyzing search logs or other activity logs to find interesting patterns. One of the main applications of web usage mining is to learn user profiles.

(iv)Benefits of Web Mining

Web Mining enables e-tailors to leverage their on-line customer data by understanding and predicting the behavior of their customers. For the first time e-tailors now have access to detailed marketing intelligence on the visitors to their web sites. The business benefits that web mining afford to digital service providers include - personalization, collaborative filtering, enhanced customer support, product and service strategy definition, particle marketing and fraud detection. In short, the ability to understand their customers' needs and to deliver the best and most appropriate service to those individual customers at any given moment.

CONCEPT PAGE RANKING ALGORITHM

Page Rank ConceptSince the early stages of the World Wide Web, search engines have developed different methods to rank web pages The occurrence of a search phrase can thereby be weighted by the length of a document (ranking by keyword density)or by its accentuation within a document by HTMLtags. Following this concept, the number of inbound links for a document measures its general importance.

Hence, a web page is generally more important, if many other web pages link to it. The concept of link popularity often avoids good rankings for pages which are only created to deceive search engines and which don't have any significance within the web, but numerous webmasters avoid it by creating masses of inbound links for doorway pages from just as insignificant other web pages. The basic approach of Page Rank is that a document is in fact considered the more important the more other documents link to it, but those inbound links do not count equally. First of all, a document ranks high in terms of Page Rank, if other high ranking documents link to it.Their rank again is given by the rank of documents which link to them. Hence, the Page Rank of a document is always determined recursively by the Page Rank of other documents. Since - even if marginal and via many links - the rank of any document influences the rank of any other, Page Rank is, in the end, based on the Linkingstructure of the whole web.

The Page Rank Algorithm

The original Page Rank algorithm was described by Lawrence Page and Sergey Brin [43] in several publications. It is given by

PR (A) = (1-d) + d (PR (T1)/C (T1) + ... + PR (Tn)/C (Tn))

Where

I. PR(A) is the Page Rank of page A,
II. PR(Ti) is the Page Rank of pages Ti which link to page A,
III. C(Ti) is the number of outbound links on page Ti and
IV. d is a dampening factor which can be set between 0 and 1.

A Different Notation of the Page Rank Algorithm Page and Sergey Brin have published two different

versions of their Page Rank algorithm in different papers. In the second version of the Algorithm, the Page Rank of page A is given as

PR(A) = (1-d) / N + d (PR(T1)/C(T1) + ... + PR (Tn)/C (Tn))

Where N is the total number of all pages on the web.

A **simplified version** of Page Rank [41] is defined in Eq. 1:

$$PR(u) = c \sum_{v \in B(u)} \frac{PR(v)}{N_v} \quad (1)$$

whereu represents a web page, B(u) is the set of pages that point to u, PR(u) and PR(v) are rank scoresof page u and v respectively, N_v denotes the numberof outgoing links of page v, c is a factor used for normalization.

Later Page Rank was modified observing that not all users follow the direct links on WWW. The modified version is given in Eq. 2.

$$PR(u) = (1-d) + d \sum_{v \in B(u)} \frac{PR(v)}{N_v} \quad (2)$$

Where d is a dampening factor that is usually set to 0.85. d can be thought of as the probability of users' following the direct links and (1 -d) as the page rank distribution from non- directly linked pages.

III. PAGERANKING REDUCTION TECHNIQUES

The World Wide Web consists billions of web pages and huge amount of information available within the pages. To retrieve required information from World Wide Web, search engine perform number of task based on their respective architecture. These can be complicated and time consuming processes. Every search engine process goes from Crawling, Indexing, Searching and sorting/ranking of information.A crawler visits and downloads all the webpage of the website and retrieve information needed from themSo search engine uses ranking algorithm in order to sort the results to be displayed. In that way user will have the most important and useful result first. Most of the ranking algorithms proposed in literature are either link or content oriented, which do not consider user usage trends. In this existing technique the user usage trends is considered.

(A)Page Ranking Based on Number of Visits of Links

In this paper, a page ranking mechanism called Page Ranking based on Visits of Links(VOL) is being devised for search engines, which works on the basic ranking algorithm of Google, i.e. Page Rank and takes number of visits of inbound links of web pages into account. This concept is very useful to display most valuable pages on the top of the result list on the basis of user browsing behavior, which reduce the search space to a large scale.$PR(u) = (1-d) + d \sum_{v \in B(u)} L_u \, \mathrm{PR(v)}/TL(v)$

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,
- L_u 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.

Example:

To explain the working of Page Rank, let us take an example hyperlinked structure shown in fig 4.1, we regard a small web consisting of three pages A, B and C. where page A links to the page B and C, page B links to page C and page C links to page A and each link has its corresponding visits.

The Page Rank for pages A, B and C are calculated by using equation (1):

PR (A) = (1-d) +d (PR(C)*2/2)

PR (B) = (1-d) +d (PR(A)*1/3)

PR (C) = (1-d) +d ((PR(B)*2/2)+ (PR(A)*2/3))

These equations can easily be solved. We get the following Page Rank values for the single pages.

PR (A) =1.10
PR (B) =0.68
PR(C) =1.21

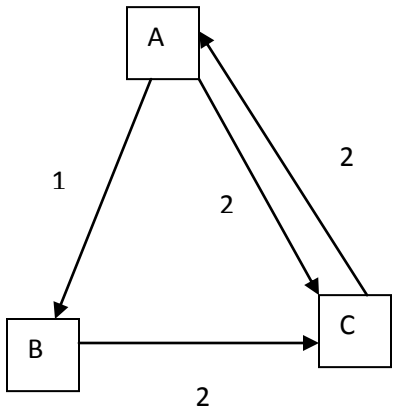
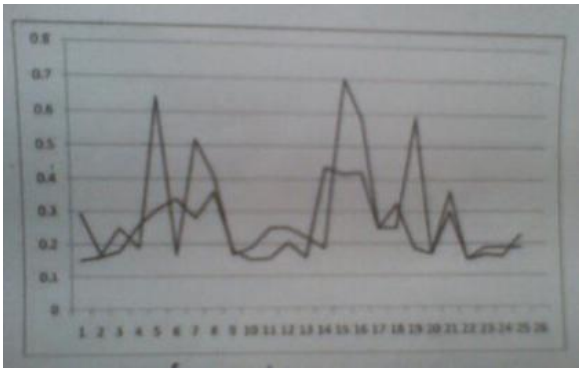


Fig .2 Hyper linked web graph with link visits

When we calculate page rank value of above graph using original Page Rank algorithm, then rank of pages will be 1.077, 0.769 and 1.154 for A, B and C respectively and it will, not change till link structure of the web graph will be same.

RESULT



(B)Weighted Page Rank Algorithm

This algorithm is also an extension of Page Rank algorithm. WPR takes into account the importance of both the in links and the out links of the pages and distributes rank scores based on the popularity of the pages. WPR performs better than the conventional Page Rank algorithm in terms of returning larger number of relevant pages to a given query.

The popularity from the number of in links and out links 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 in links of page *u* and the number of in links of all reference pages of page *v*.

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

$W^{out}_{(v,u)}$ is the weight of *link*(*v*, *u*) calculated based on the number of out links of page *u* and the number of out links of all reference pages of page *v*

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

Considering the importance of pages, the original Page Rank formula is modified as

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

(C) HITS (Hyperlink-Induced Topic Search)

The HITS algorithm considers the WWW as a directed graph G (V, E), where V is a set of vertices representing pages and E is a set of edges that correspond to links. A directed edge (p, q) indicates a link from page p to page q. The search engine may not retrieve all relevant pages for the query; therefore the initial pages retrieved by the search engine are a good starting point to move further. But relying only on the initial pages does not guarantee that authority and hub pages are also retrieved efficiently. To remove this problem, HITS uses a proper method to find the relevant information regarding the user query.

Step I-Sampling Step

Input: Root set R; Output: Base set S

Let S=R

1. For each page p ϵ S, do Step 3 to 5.
2. Let T be the set of all pages S points to.
3. Let F be the set of all pages that point to S.
4. Let S=S +T+ some or all of F.
5. Delete all links with the same domain name.
6. Return S.

IV. Conclusion

Two commonly used algorithms in web structure mining are HITS and Page Rank, which are used to rank the relevant pages. Both algorithms treat all links equally when distributing rank scores. Several algorithms have been developed to improve the performance of these methods. This paper introduces the WPR algorithm, an extension to the Page Rank algorithm. WPR takes into account the importance of both the in links and the out links of the pages and distributes rank scores based on the popularity of the pages.

References

[1] ShotaHatakenaka, Takao Miura, "*Ranking Documents using Similarity- based Page Ranks*", 2011 IEEE, 978-1-4577-0251-8.

[2] Gyanendra Kumar, NeelamDuahn, 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.

[3] Lili Yan, Yingbin Wei, ZhanjiGui, and Chen Yizhuo, "*Research on Page Rank and Hyperlink-Induced Topic Search in Web Structure Mining*", 2011 IEEE, 978-1-4244-7255-0.

[4] Chongchong Zhao, Zhiqiang Zhang, Hualong Li, and XieXiaoqin, "*A Search Result Ranking Algorithm Based on Web Pages and Tags Clustering",* 978-1-4244-8728-8, 2011 IEEE.

[5] M. Sathya, J. Jayanthi, and Basker N., "*Link Based K-Means Clustering Algorithm for Information Retrieval*", 2011 IEEE, 978-1-4577-0590-8.

[6] P Ravi Kumar, and Singh Ashutoshkumar, "*Web Structure Mining Exploring Hyperlinks and Algorithms for Information Retrieval*", American Journal of Applied Sciences, 7 (6) 840-845 2010.

[7] D.K. Sharma, and Sharma A.K., "*A Comparative Analysis of Web Page Ranking Algorithms*", International Journal on Computer Science and Engineering Vol. 02-08, 2010, pp. 2670-2676

[8] N. Duhan, A. K. Sharma and Bhatia K. K., "*Page Ranking Algorithms: A Survey*, Proceedings of the IEEE International Conference on Advance Computing, 2009, 978-1-4244-1888-6.

[9] C.D. Manning, P. Raghavan, and Schutze, H., " *Introduction to Information Retrieval*", Cambridge University Press, 2008
[10] Wen-ChihPeng and Lin Yu-Chin, "*Ranking Web Search Results from Personalized Perspective*", Proceedings of the 8th IEEE International Conference on E-Commerce Technology and the 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (CEC/EEE'06), 0-7695-2511-3.
[11] Michael Brinkmeier," *Page Rank revisited*", [J], ACM Transactions on Internet Technology, 2006, l6 (3): 282 - 301
[12] A.N. Langville, and Meyer, C.D., "*Google's Page Rank and Beyond: The Science of Search Engine Rankings*", Princeton University Press, June 2006
[13] Yiqun Liu, Min Zhang, and RuLiyun, "*Automatic Query Type Identification Based on Click Through Information*", Asia Information Retrieval Symposium(AIRS), 2006.
[14] M. G. da, Gomes Jr. and Gong Z., "*Web Structure Mining: An Introduction*", Proceedings of the IEEE International Conference on Information Acquisition, 2005.
[15] Zhang M, Ma SP, and Song RH, "DF *or IDF: On the use of primary feature model for Web information retrieval*". Journal of Software, 2005, 5(16):1012-1020.
[16] Guo Yan, BaiShuo, Yang Zhi-feng, and Zhang Kai, "*Analyzing Scale of Web Logs and Mining Users' Interests*", [J], Chinese Journal of Computers, 2005, 9(28):1483-1496.