Page Ranking Algorithms for Web Mining

Rekha Jain
Department of Computer Science, Apaji Institute,
Banasthali University
C-62 Sarojini Marg, C-Scheme, Jaipur,Rajasthan

Dr. G. N. Purohit
Department of Computer Science, Apaji Institute,
Banasthali University

ABSTRACT

As the web is growing rapidly, the users get easily lost in the web's rich hyper structure. The primary goal of the web site owner is to provide the relevant information to the users to fulfill their needs. Web mining technique is used to categorize users and pages by analyzing users behavior, the content of pages and order of URLs accessed. Web Structure Mining plays an important role in this approach. In this paper we discuss and compare the commonly used algorithms i.e. PageRank, Weighted PageRank and HITS.

General Terms

PageRanking Algorithms

Keywords

Web Mining, Web Content Mining, Web Structure Mining, Web Usage Mining, PageRank, Weighted PageRank, HITS

1. INTRODUCTION

The World Wide Web is a rich source of information and continues to expand in size and complexity. Retrieving of the required web page on the web, efficiently and effectively, is becoming a challenge[1]. Whenever a user wants to search the relevant pages, he/she prefers those relevant pages to be at hand. Relevant web page is one that provides the same topic as the original page but it is not semantically identical to original page[1]. As the Web is unstructured data repository, which delivers the bulk amount of information and also increases the complexity of dealing information from different perspective of knowledge seekers, business analysts and web service providers[2]. According to Google report on 25th July 2008 there are 1 trillion unique URLs on the web[3]. Web has grown tremendously and the usage of web is unimaginable so it is important to understand the data structure of web. The bulk amount of information becomes very difficult for the users to find, extract, filter or evaluate the relevant information. This issue raises the necessity of some technique that can solve these challenges.

Web mining can be easily executed with the help of other areas like Database (DB), Information retrieval (IR), Natural Language Processing (NLP), Machine Learning etc. These can be used to discuss and analyze the useful information from WWW.

Following are some challenges [3]:

1) Web is huge. 2) Web pages are semi structured. 3) Web information stands to be diversity in meaning. 4) Degree of quality of the information extracted. 5) Conclusion of knowledge from information extracted.

This paper is organized as follows- Web Mining is introduced in Section 2. The areas of Web Mining i.e. Web Content Mining, Web Structure Mining and Web Usage Mining are discussed in Section 3. Section 4 describes the various Link analysis algorithms. Section 4.1 presented the PageRank algorithm and its functioning. Weighted PageRank algorithm which is an extension of PageRank is presented in Section 4.2. Section 4.3 includes combined analysis of both PageRank and Weighted PageRank algorithms. In Section 4.4 HITS algorithm and its constraints are discussed. Section 5 provides the comparison of these algorithms. Concluding remarks are given in Section 6.

2. WEB MINING

Web mining is the Data Mining technique that automatically discovers or extracts the information from web documents. It consists of following tasks[4]:

1. Resource finding: It involves the task of retrieving intended web documents. It is the process by which we extract the data either from online or offline text resources available on web.
2. Information selection and pre-processing: It involves the automatic selection and pre processing of specific information from retrieved web resources. This process transforms the original retrieved data into information. The transformation could be renewal of stop words, stemming or it may be aimed for obtaining the desired representation such as finding phrases in training corpus.
3. Generalization: It automatically discovers general patterns at individual web sites as well as across multiple sites. Data Mining techniques and machine learning are used in generalization
4. Analysis: It involves the validation and interpretation of the mined patterns. It plays an important role in pattern mining. A human plays an important role in information on knowledge discovery process on web.

3. WEB MINING CATEGORIES

There are three areas of Web Mining according to the web data used as input in Web Data Mining. Web Content Mining, Web Structure Mining and Web Usage Mining

3.1 Web Content Mining

It is the process of retrieving the information from WWW into more structured forms and indexing the information to retrieve it quickly. It focuses mainly on the structure within a document i.e. inner document level. Web Content Mining is related to Data Mining because many Data Mining techniques can be applied in Web Content Mining. It is also related with text mining because much of the web contents are text, but is also quite different from these because web data is mainly semi structured in nature and text mining focuses on unstructured text.

3.2 Web Structure Mining

It is the process by which we discover the model of link structure of the web pages. We catalog the links, generate the information such as the similarity and relations among them by taking the advantage of hyperlink topology. PageRank and hyperlink analysis also fall in this category. The goal of Web Structure Mining is to generate structured summary about the website and web page. It tries to discover the link structure of hyper links at inter document level. As it is very common that the web documents contain links and they use both the real or primary data on the web so it can be concluded that Web Structure Mining has a relation with Web Content Mining. It is quite often to combine these two mining tasks in an application.

3.3 Web Usage Mining

It is the process by which we identify the browsing patterns by analyzing the navigational behavior of user. It focuses on techniques that can be used to predict the user behavior while the user interacts with the web. It uses the secondary data on the web. This activity involves the automatic discovery of user access patterns from one or more web servers. Through this mining technique we can ascertain what users are looking for on Internet. Some might be looking for only technical data, where as some others might be interested in multimedia data. Table 1 gives an overview of above mining categories [4].

TABLE 1: Web Mining Categories

	Web Mining			
	Web Content Mining		**Web Structure Mining**	**Web Usage Mining**
	IR view	**DB View**		
View of Data	-Unstructured -Structured	-Semi Structured -Web Site as DB	-Link Structure	-Interactivity
Main Data	- Text documents -Hypertext documents	-Hypertext documents	-Link Structure	-Server Logs -Browser Logs
Representation	-Bag of words, n-gram Terms, -phrases, Concepts or ontology -Relational	-Edge labeled Graph, -Relational	-Graph	-Relational Table -Graph
Method	-Machine Learning -Statistical (including NLP)	-Proprietary algorithms -Association rules	-Proprietary algorithms	-Machine Learning -Statistical -Association rules
Application Categories	-Categorization -Clustering -Finding extract rules -Finding patterns in text	-Finding frequent sub structures -Web site schema discovery	-Categorization -Clustering	-Site Construction -adaptation and management -Marketing, -User Modeling

4. LINK ANALYSIS ALGORITHMS

Web Mining technique provides the additional information through hyperlinks where different documents are connected[2]. We can view the web as a directed labeled graph whose nodes are the documents or pages and edges are the hyperlinks between them. This directed graph structure is known as web graph.

There are number of algorithms proposed based on link analysis. Three important algorithms PageRank[5], Weighted PageRank[6] and HITS (Hyper-link Induced Topic Search)[7] are discussed below.

4.1 PageRank

This algorithm was developed by Brin and Page at Stanford University which extends the idea of citation analysis[5]. In citation analysis the incoming links are treated as citations but this technique could not provide fruitful results because this gives some approximation of importance of page. So PageRank provides a better approach that can compute the importance of web page by simply counting the number of pages that are linking to it. These links are called as backlinks.

If a backlink comes from an important page than this link is given higher weightage than those which are coming from non-important pages. The link from one page to another is considered as a vote. Not only the number of votes that a page receives is important but the importance of pages that casts the vote is also important.

Page and Brin proposed a formula to calculate the PageRank of a page A as stated below-

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

here *PR(Ti)* is the PageRank of the Pages Ti which links to page A, *C(Ti)* is number of outlinks on page Ti and *d* is damping factor. It is used to stop other pages having too much influence. The total vote is “damped down” by multiplying it to 0.85.

The PageRank forms a probability distribution over the web pages so the sum of PageRanks of all web pages will be one. The PageRank of a page can be calculated without knowing the final value of PageRank of other pages. It is an iterative algorithm which follows the principle of normalized link matrix of web. PageRank of a page depends on the number of pages pointing to a page.

4.2 Weighted Page Rank

This algorithm was proposed by Wenpu Xing and Ali Ghorbani which is an extension of PageRank algorithm[7]. 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.

This is denoted as $W^{in}_{(m,n)}$ and $W^{out}_{(m,n)}$ respectively. $W^{in}_{(m,n)}$ is the weight of link(m,n) as given in (2). It is calculated on the basis of number of incoming links to page n and the number of incoming links to all reference pages of page m.

$$W^{in}_{(m,n)} = \frac{I_n}{\sum_{p \in R(m)} I_p} \qquad \text{....(2)}$$

I_n is number of incoming links of page n, I_p is number of incoming links of page p, R(m) is the reference page list of page m.

$W^{out}_{(m,n)}$ is the weight of link(m,n)as given in (3). It is calculated on the basis of the number of outgoing links of page n and the number of outgoing links of all the reference pages of page m.

$$W^{out}_{(m,n)} = \frac{O_n}{\sum_{p \in R(m)} O_p} \qquad \text{....(3)}$$

O_n is number of outgoing links of page n, O_p is number of outgoing links of page p,

Then the weighted PageRank is given by formula in (4)

$$WPR(n) = (1-d) + d \sum_{m \in B(n)} WPR(m) W^{in}_{(m,n)} W^{out}_{(m,n)} \qquad \text{...(4)}$$

4.2.1 PageRank VS Weighted PageRank

In order to compare the WPR with the PageRank, the resultant pages of a query are categorized into four categories based on their relevancy to the given query.[8] They are

1. Very Relevant Pages (VR): These are the pages that contain very important information related to a given query.
2. Relevant Pages (R): These Pages are relevant but not having important information about a given query.
3. Weakly Relevant Pages (WR): These Pages may have the query keywords but they do not have the relevant information.
4. Irrelevant Pages (IR): These Pages are not having any relevant information and query keywords.

The PageRank and WPR algorithms both provide ranked pages in the sorting order to users based on the given query. So, in the resultant list, the number of relevant pages and their order are very important for users. Relevance Rule is used to calculate the relevancy value of each page in the list of pages. That makes WPR different from PageRank.

Relevancy Rule: The Relevancy Rule is as given in (5). The Relevancy of a page to a given query depends on its category and its position in the page-list. The larger the relevancy value, the better is the result.

$$k = \sum_{i \in R(p)} (n-i) * W_i \qquad \text{...(5)}$$

where *i* denotes the i^{th} page in the result page-list *R(p)*, *n* represents the first n pages chosen from the list R(p), and W_i is the weight of i^{th} page as given in (6).

$$W_i = (v1, v2, v3, v4) \qquad \text{...(6)}$$

where *v1, v2, v3* and *v4* are the values assigned to a page if the page is VR, R, WR and IR respectively. The values are always v1>v2>v3>v4. Experimental studies shows that WPR produces larger relevancy values than the PageRank.

4.3 HITS (Hyper-link Induced Topic Search)

Klienberg gives two forms of web pages called as hubs and authorities. Hubs are the pages that act as resource lists. Authorities are pages having important contents. A good hub page is a page which is pointing to many authoritative pages on that content and a good authority page is a page which is pointed by many good hub pages on the same content. A page may be a good hub and a good authority at the same time[8,9].

The HITS algorithm treats WWW as directed graph G(V,E), where V is a set of vertices representing pages and E is set of edges corresponds to link. Figure 1 shows the hubs and authorities in web [2].

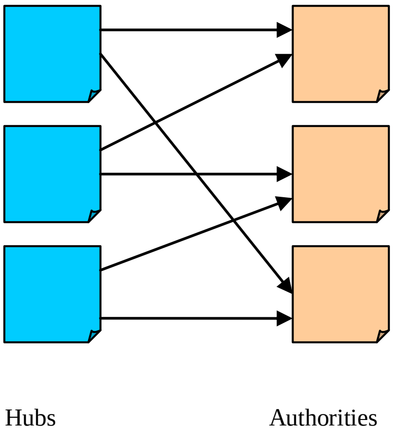

Figure 1: Hubs and Authorities

It has two steps:

1. Sampling Step:- In this step a set of relevant pages for the given query are collected.
2. Iterative Step:- In this step Hubs and Authorities are found using the output of sampling step.

Following expressions (7,8) are used to calculate the weight of Hub *(H_p)* and the weight of Authority *(A_p)*.

$$H_p = \sum_{q \in I(p)} A_q \qquad \text{... (7)}$$

$$A_p = \sum_{q \in B(p)} H_q \qquad \text{... (8)}$$

here H_q is Hub Score of a page, A_q is authority score of a page, *I(p)* is set of reference pages of page p and *B(p)* is set of referrer pages of page p,
the authority weight of a page is proportional to the sum of hub weights of pages that link to it. Similarly a hub of a page is proportional to the sum of authority weights of pages that it links to.

4.3.1 Constraints with HITS algorithm

Following are some constraints of HITS algorithm[10]

- Hubs and authorities: It is not easy to distinguish between hubs and authorities because many sites are hubs as well as authorities.

- Topic drift: Sometime HITS may not produce the most relevant documents to the user queries because of equivalent weights.
- Automatically generated links: HITS gives equal importance for automatically generated links which may not have relevant topics for the user query
- Efficiency: HITS algorithm is not efficient in real time.

HITS was used in a prototype search engine called Clever for an IBM research project. Because of the above constraints HITS could not be implemented in a real time search engine.

5. COMPARISON

Table 2 shows comparison of all the three algorithms discussed above[12].

Table 2: Comparison of algorithms

Algorithm	PageRank	Weighted PageRank	HITS
Mining technique used	WSM	WSM	WSM & WCM
Working	Computes scores at indexing time. Results are sorted according to importance of pages.	Computes scores at indexing time. Results are sorted according to Page importance.	Computes hub and authority scores of n highly relevant pages on the fly.
I/P Parameters	Backlinks	Backlinks, Forward links	Backlinks, Forward Links & content
Complexity	O(log N)	<O(log N)	<O(log N)
Limitations	Query independent	Query independent	Topic drift and efficiency problem
Search Engine	Google	Research model	Clever

6. CONCLUSION

Web Mining is powerful technique used to extract the information from past behavior of users. Web Structure Mining plays an important role in this approach. Various algorithms are used in Web Structure Mining to rank the relevant pages. PageRank, Weighted PageRank and HITS treat all links equally when distributing the rank score. PageRank and Weighted PageRank are used in Web Structure Mining. HITS is used in both structure Mining and Web Content Mining. PageRank and Weighted PageRank calculates the score at indexing time and sort them according to importance of page where as HITS calculates the hub and authority score of n highly relevant pages. The input parameters used in Page Rank are BackLinks, Weighted PageRank uses Backlinks and Forward Links as Input Parameter, HITS uses Backlinks, Forward Link and Content as Input Parameters. Complexity of PageRank algorithm is O(log N) where as complexity of Weighted PageRank and HITS algorithms are <O(log N).

As part of our future work, we are planning to carry out performance analysis of PageRank and Weighted PageRank and working on finding the ways to categorize the users and web pages to obtain the better PageRank results.

7. REFERENCES

[1] J. Hou and Y. Zhang, Effectively Finding Relevant Web Pages from Linkage Information, IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 4, 2003.

[2] P Ravi Kumar, and Singh Ashutosh kumar, Web Structure Mining Exploring Hyperlinks and Algorithms for Information Retrieval, American Journal of applied sciences, 7 (6) 840-845 2010.

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

[4] R. Kosala, and 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.

[5] S. Brin, and L. Page, The Anatomy of a Large Scale Hypertextual Web Search Engine,, Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117, 1998.

[6] Wenpu Xing and Ali Ghorbani, Weighted PageRank Algorithm, Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.

[7] J. Kleinberg, Authoritative Sources in a Hyper-Linked Environment, Journal of the ACM 46(5), pp. 604-632, 1999.

[8] C. Ding, X. He, P. Husbands, H. Zha, and H. Simon, Link analysis: Hubs and authorities on the world. Technical report: 47847, 2001.

[9] J. M. Klienberg, Authoritative sources in a hyperlinked environment. Journal of the ACM, 46(5):604-632, September 1999.

[10] S. Chakrabarti, B.Dom, D.Gibson, J. Kleinberg, R. Kumar, P. Raghavan,S. Rajagopalan, and A. Tomkins, Mining the Link Structure of the World Wide Web, IEEE Computer, Vol. 32, pp. 60-67, 1999.

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