Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval

Ravi Kumar P
Department of ECEC, Curtin University of Technology,
Sarawak Campus, Miri, Malaysia
ravi2266@gmail.com

Ashutosh Kumar Singh
Department of ECEC, Curtin University of Technology,
Sarawak Campus, Miri, Malaysia
ashutosh.s@curtin.edu.my

Abstract—**This paper focus on the Hyperlink analysis, the algorithms used for link analysis, compare those algorithms and the role of hyperlink analysis in Web searching. In the hyperlink analysis, the number of incoming links to a page and the number of outgoing links from that page will be analyzed and the reliability of the linking will be analyzed. Authorities and Hubs concept of Web pages will be explored. The different algorithms used for Link analysis like PageRank, HITS (Hyperlink-Induced Topic Search) and other algorithms will be discussed and compared. The formula used by those algorithms will be explored.**

Keywords - Web Mining, Web Content, Web Structure, Web Graph, Information Retrieval, Hyperlink Analysis, PageRank, Weighted PageRank and HITS.

I. INTRODUCTION

The Web is a massive, explosive, diverse, dynamic and mostly unstructured data repository, which delivers an incredible amount of information, and also increases the complexity of dealing with the information from the different perspectives of knowledge seekers, Web service providers and business analysts. The following are considered as challenges [1] in the Web mining:

- Web is huge and Web Pages are semi-structured
- Web information tends to be diversity in meaning
- Degree of quality of the information extracted
- Conclusion of the knowledge from the information extracted

Web mining techniques along with other areas like Database (DB), Information Retrieval (IR), Natural Language Processing (NLP), Machine Learning etc. can be used to solve the above challenges. Web mining is the use of data mining techniques to automatically discover and extract information from the World Wide Web (WWW). Web structure mining helps the users to retrieve the relevant documents by analyzing the link structure of the Web.

This paper is organized as follows. Next Section provides concepts of Web Structure mining and Web Graph. Section III provides Hyperlink analysis, algorithms and their comparisons. Paper is concluded in Section IV.

II. WEB STRUCTURE MINING

A. Overview

According to Kosala et al [2], Web mining consists of the following tasks:

- *Resource finding*: the task of retrieving intended Web documents.
- *Information selection and pre-processing*: automatically selecting and pre-processing specific information from retrieved Web resources.
- *Generalization*: automatically discovers general patterns at individual Web sites as well as across multiple sites.
- *Analysis*: validation and/or interpretation of the mined patterns.

There are three areas of Web mining according to the usage of the Web data used as input in the data mining process, namely, Web Content Mining (WCM), Web Usage Mining (WUM) and Web Structure Mining(WSM). Web content mining is concerned with the retrieval of information from WWW into more structured forms and indexing the information to retrieve it quickly. Web usage mining is the process of identifying the browsing patterns by analyzing the user's navigational behavior. Web structure mining is to discover the model underlying the link structures of the Web pages, catalog them and generate information such as the similarity and relationship between them, taking advantage of their hyperlink topology. Hyperlink analysis and the algorithms discussed here are related to Web Structure mining. Even though there are three areas of Web mining, the differences between them are narrowing because they are all interconnected.

B. How big is Web

A Google report [3] on 25th July 2008 says that there are 1 trillion (1,000,000,000,000) unique URLs (Universal Resource Locator) on the Web. The actual number could be more than that

and Google could not index all the pages. When Google first created the index in 1998 there were 26 million pages and in 2000 Google index reached 1 billion pages. In the last 9 years, Web has grown tremendously and the usage of the web is unimaginable. So it is important to understand and analyze the underlying data structure of the Web for effective Information Retrieval.

C. *Web Data Structure*

The traditional information retrieval system basically focuses on information provided by the text of Web documents. Web mining technique provides additional information through hyperlinks where different documents are connected. The Web may be viewed as a directed labeled graph whose nodes are the documents or pages and the edges are the hyperlinks between them. This directed graph structure in the Web is called as *Web Graph.* A graph *G* consists of two sets *V* and *E*, Horowitz et al [4]. The set *V* is a finite, nonempty set of *vertices*. The set *E* is a set of pairs of vertices; these pairs are called *edges*. The notation *V*(*G*) and *E*(*G*) represent the sets of vertices and edges, respectively of graph *G*. It can also be expressed *G* = (*V*, *E*) to represent a graph. The graph in Fig. 1 is a directed graph with 3 Vertices and 3 edges.

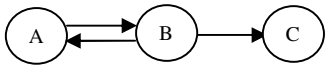

Figure 1 A Directed Graph G

The vertices *V* of *G, V*(*G*) = {*A*, *B*, *C*}. The Edges *E* of *G*, *E*(*G*) ={(*A*, *B*), (*B*, *A*), (*B*, *C*)}. In a directed graph with *n* vertices, the maximum number of edges is *n*(*n*-1). With 3 vertices, the maximum number of edges can be 3(3-1) = 6. In the above example, there is no link from (*C*, *B*), (*A*, *C*) and (*C*, *A*). A directed graph is said to be *strongly connected* if for every pair of distinct vertices *u* and *v* in *V*(*G*), there is a directed path from *u* to *v* and also from *v* to *u*. The above graph in Fig. 1 is not strongly connected, as there is no path from vertex *C* to *B*. According to Broader et al. [5], a Web can be imagined as a large graph containing several hundred million or billion of nodes or vertices, and a few billion arcs or edges. The following section explains the *hyperlink* analysis and the algorithms used in the *hyperlink* analysis for information retrieval.

III. HYPERLINK ANALYSIS

Many Web Pages do not include words that are descriptive of their basic purpose (for example rarely a search engine portal includes the word "search" in its home page), and there exist Web pages which contain very little text (such as image, music, video resources), making a text-based search techniques difficult. However, how others exemplify this page may be useful. This type of "characterization" is included in the text that surrounds the hyperlink pointing to the page.

Many researches [6, 7, 8, 11, 12] have done and solutions have suggested to the problem of searching, indexing or querying the Web, taking into account its structure as well as the meta-information included in the hyperlinks and the text surrounding them.

There are a number of algorithms proposed based on the Link Analysis. Using citation analysis, Co-citation algorithm [13] and Extended Co-citation algorithm [14] are proposed. These algorithms are simple and deeper relationships among the pages can not be discovered. Three important algorithms *PageRank*[15], *Weighted PageRank* (*WPR*)[16] and Hypertext Induced Topic Search *HITS*[17] are discussed below in detail and compared.

A. *PageRank*

Brin and Page developed *PageRank* [15] algorithm during their Ph D at Stanford University based on the citation analysis [9, 10]. *PageRank* algorithm is used by the famous search engine, Google. They applied the citation analysis in Web search by treating the incoming links as citations to the Web pages. However, by simply applying the citation analysis techniques to the diverse set of Web documents did not result in efficient outcomes. Therefore, *PageRank* provides a more advanced way to compute the importance or relevance of a Web page than simply counting the number of pages that are linking to it (called as "backlinks"). If a backlink comes from an "important" page, then that backlink is given a higher weighting than those backlinks comes from non-important pages. In a simple way, link from one page to another page may be considered as a vote. However, not only the number of votes a page receives is considered important, but the "importance" or the "relevance" of the ones that cast these votes as well.

Assume any arbitrary page A has pages T_1 to T_n pointing to it (incoming link). *PageRank* can be calculated by the following equation (1).

$$PR(A) = (1-d) + d(PR(T_1)/C(T_1) + .. + PR(T_n/C(T_n)) \quad (1)$$

The parameter *d* is a damping factor, usually sets it to 0.85 (to stop the other pages having too much influence, this total vote is "damped down" by multiplying it by 0.85). *C(A)* is defined as the number of links going out of page A. The *PageRanks* form a probability distribution over the Web pages, so the sum of all Web pages' *PageRank* will be one. *PageRank* can be calculated using a simple iterative algorithm, and corresponds to the principal eigenvector of the normalized link matrix of the Web.

Let us take an example of hyperlink structure of three pages *A*, *B* and *C* as shown in Fig. 2. The *PageRank* for pages *A*, *B* and *C* can be calculated by using equation (1).

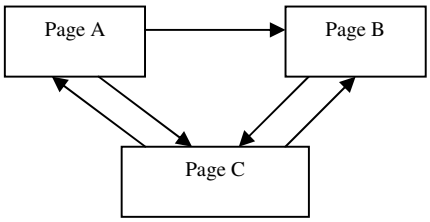

Figure 2 Hyperlink Structure for 3 pages

Let us assume the initial *PageRank* as 1.0 and do the calculation. The damping factor *d* is set to 0.85.

$$PR(A) = (1\text{-}d) + d\,(PR(C)/C(C)) = (1\text{-}0.85) + 0.85(1/2) = 0.15 + 0.425 = 0.575 \quad (1a)$$

$$PR(B) = (1\text{-}d) + d((PR(A)/C(A) + (PR(C)/C(C)) = 0.819 \quad (1b)$$

$$PR(C) = (1\text{-}d) + d((PR(A)/C(A) + (PR(B)/C(B)) = 1.091 \quad (1c)$$

Do the second iteration by taking the above *PageRank* values from (1a), (1b) and (1c).

$$PR(A) = 0.15 + 0.85(1.091/2)) = 0.614 \quad (1d)$$

$$PR(B) = 0.15 + 0.85((0.614/2)+(1.091/2)) = 0.875 \quad (1e)$$

$$PR(C) = 0.15 + 0.85((0.614/2)+(0.875/1)) = 1.155 \quad (1f)$$

After doing many more iterations of the above calculation, the following *PageRanks* arrived as shown in Table I.

TABLE I. ITERATIVE CALCULATION FOR PAGERANK

Iteration	PR(A)	PR(B)	PR(C)
0	1	1	1
1	0.575	0.819	1.091
2	0.614	0.875	1.155
...	...	...	...
15	0.701	0.999	1.297
16	0.701	0.999	1.297

For a smaller set of pages, it is easy to calculate and find out the *PageRank* values but for a Web having billions of pages, it is not easy to do the calculation like above. In the above Table I, you can notice that *PageRank* of C is higher than *PageRank* of B and A. It is because Page C has 2 incoming links and 2 outgoing links as shown in Fig. 2. Page B has 2 incoming links and 1 outgoing link . Page A has the lowest *PageRank* because Page A has only one incoming link and 2 outgoing links. From the Table I, after the iteration 15, the *PageRank* for the pages gets normalized. Previous experiments [18, 19] shows that the *PageRank* gets converged to a reasonable tolerance. The convergence of *PageRank* calculation for the Table I is shown as a graph in Fig. 3.

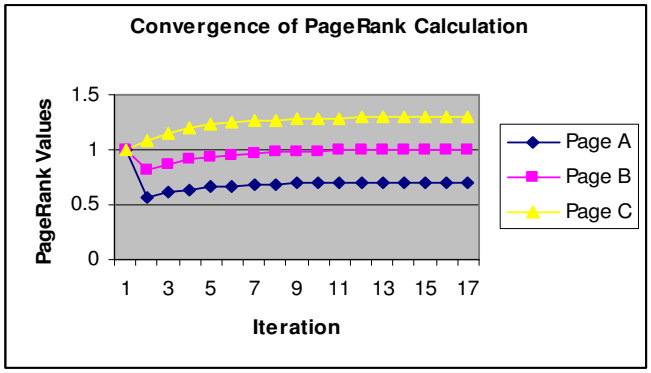

Figure 3 Convergence of PageRank Calculation

B. *Weighted PageRank Algorithm*

Wenpu Xing and Ali Ghorbani [16] proposed a *Weighted PageRank* (*WPR*) algorithm which is an extension of the *PageRank* algorithm. This algorithm assigns a larger rank values to the more important pages rather than dividing the rank value of a page evenly among its outgoing linked pages. Each outgoing link gets a value proportional to its importance. The importance is assigned in terms of weight values to the incoming and outgoing links and are denoted as $W^{in}(m, n)$ and $W^{out}(m, n)$ respectively. $W^{in}(m, n)$ as shown in equation (2) is the weight of $link(m, n)$ calculated based on the number of incoming links of page n and the number of incoming links of all reference pages of page m.

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

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

Where I_n and I_p are the number of incoming links of page n and page p respectively. $R(m)$ denotes the reference page list of page m. $W^{out}(m, n)$ is as shown in equation (3) is the weight of $link(m, n)$ calculated based on the number of outgoing links of page n and the number of outgoing links of all reference pages of m. Where O_n and O_p are the number of outgoing links of page n and p respectively. The formula as proposed by Wenpu et al for the *WPR* is as shown in equation (4) which is a modification of the *PageRank* formula (equation 1).

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

Use the same hyperlink structure as shown in Fig. 2 and do the *WPR* Calculation. The *WPR* equations for Pages *A*, *B* and *C* are as follows.

$$WPR(A) = (1-d) + d(WPR(C).W^{in}_{(C,A)}.W^{out}_{(C,A)}) \quad (4a)$$

$$WPR(B) = (1-d) + d(WPR(A).W^{in}_{(A,B)}.W^{out}_{(A,B)} + WPR(C).W^{in}_{(C,B)}.W^{out}_{(C,B)}) \quad (4b)$$

$$WPR(C) = (1-d) + d(WPR(A).W^{in}_{(A,C)}.W^{out}_{(A,C)} + WPR(B).W^{in}_{(B,C)}.W^{out}_{(B,C)}) \quad (4c)$$

The incoming link and outgoing link weights are calculated as follows:

$$W^{in}_{(C,A)} = I_A/(I_A + I_B) = 1/(1+2) = 1/3 \quad (4d)$$

$$W^{out}_{(C,A)} = O_A/(O_A + O_B) = 2/(2+1) = 2/3 \quad (4e)$$

By substituting the values of equations (4d) and (4e) to equation (4a), you will get the *WPR* of Page *A* by taking a value of 0.85 for *d* and the initial value of *WPR*(*C*) = 1.

$$WPR(A) = (1-0.85) + 0.85(1*1/3*2/3) = 0.69 \quad (4f)$$

$$WPR(B) = (1-0.85 + 0.85((0.69*1/2*1/3) = 0.44 \quad (4g)$$

$$WPR(C) = (1-0.85) + 0.85((0.69*1/2*2/3) = 0.47 \quad (4h)$$

The values of *WPR*(*A*), *WPR*(*B*) and *WPR*(*C*) are shown in equations (4f), (4g) and (4h) respectively. In this, *WPR*(*A*)>*WPR*(*C*)>*WPR*(*B*). This results shows that the page rank order is different from *PageRank*.

C. The HITS Algorithm - Hubs & Authorities

Kleinberg [17] identifies two different forms of Web pages called *hubs* and *authorities*. Authorities are pages having important contents. Hubs are pages that act as resource lists, guiding users to authorities. Thus, a good hub page for a subject points to many authoritative pages on that content, and a good authority page is pointed by many good hub pages on the same subject. Hubs and Authorities and their calculations are shown in Fig. 4. Kleinberg says that a page may be a good hub and a good authority at the same time. This circular relationship leads to the definition of an iterative algorithm called HITS (Hyperlink Induced Topic Search). The HITS algorithm treats 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.

There are two major steps in the HITS algorithm. The first step is the *Sampling Step* and the second step is the *Iterative step*. In the *Sampling step*, 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. This algorithm starts with a root set *R*, a set of *S* is obtained, keeping in mind that *S* is relatively small, rich in relevant pages about the query and contains most of the good authorities. The second step, *Iterative step,* finds hubs and authorities using the output of the sampling step using equations (5) and (6).

$$H_p = \sum_{q \in I(p)} A_q \quad (5)$$

$$A_p = \sum_{q \in B(p)} H_q \quad (6)$$

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. The page's authority weight is proportional to the sum of the hub weights of pages that it links to it, Kleinberg [20]. Similarly, a page's hub weight is proportional to the sum of the authority weights of pages that it links to. Fig. 4 shows an example of the calculation of authority and hub scores.

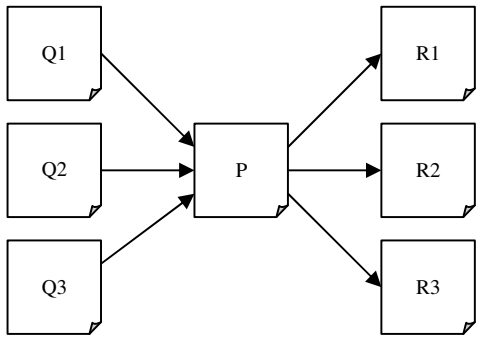

Figure 4. Calculation of hubs and Authorities

The following are the constraints of HITS algorithm [6]:

- *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 produce relevant topics for the user query.
- *Efficiency*: HITS algorithm is not efficient in real time.

Table II shows the comparison [21] of all the algorithms discussed above.

TABLE II. COMPARISON OF HYPERLINK ALGORITHMS

Algorithm / Criteria	*PageRank*	*Weighted PageRank*	*HITS*
Mining technique used	*WSM*	*WSM*	*WSM & WCM*
Working	Computes scores at index time. Results are sorted on the importance of pages.	Computes scores at index time. Results are sorted on the Page importance.	Computes 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

IV. CONCLUSION

This paper covers the basics of Web mining. The importance of the Web structure mining in Information retrieval is explained. The main purpose of this paper is to explore the hyperlink structure and understand the *Web graph* in a simple way. This paper also focuses on the important algorithms used for hyperlink analysis, explore those algorithms and compare them.

REFERENCES

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

[2] 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.

[3] http://googleblog.blogspot.com/2008/07.

[4] E. Horowitz, S. Sahni and S. Rajasekaran, "Fundamentals of Computer Algorithms", *Galgotia Publications Pvt. Ltd*., pp. 112-118, 2008.

[5] A. Broder, R. Kumar, F Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, "Graph Structure in the Web", *Computer Networks : The International Journal of Computer and telecommunications Networking*, Vol. 33, Issue 1-6, pp 309-320, 2000.

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

[7] T.H. Haveliwala, A. Gionis, D. Klein, P. Indyk, "Evaluating Strategies for Similarity Search on the Web", *Proc. of the WWW 11, Hawaii, USA*, 2002.

[8] I. Varlamis, M. Vazirgiannis, M. Halkidi, B. Nguyen, THESUS, "A Closer View on Web Content Management Enhanced with Link Semantics", *IEEE Transaction on Knowledge and Data Engineering Journal*.

[9] Eugene Garfield, "Citation Analysis as a tool in journal evaluation", *Science* 178, pp. 471-479, 1972.

[10] G. Pinski and F. Narin, "Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics", in *Information Processing and Management*, 1976.

[11] D. Gibson, J. Kleinberg, P. Raghavan, "Inferring Web Communities from Link Topology", *Proc. of the 9th ACM Conference on Hypertext and Hypermedia*, 1998.

[12] R. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, "Trawling the Web for Emerging Cyber-Communities", *Proc. of the 8th WWW Conference(WWW8),* 1999.

[13] J. Dean and M. Henzinger, "Finding Related Pages in the World Wide Web", *Proc. Eight Int'l World Wide Web Conf.*, pp. 389-401, 1999.

[14] 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.

[15] S. Brin, 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.

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

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

[18] 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.

[19] C. Ridings and M. Shishigin, "PageRank Convered". *Technical Report*, 2002.

[20] J. Kleinberg, " Hubs, Authorities, and Communities", *ACM Computing Surveys*, 31(4), 1999.

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