American Journal of Applied Sciences 7 (6): 840-845, 2010
ISSN 1546-9239

Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval

P. Ravi Kumar and Ashutosh Kumar Singh
Department of Electrical and Computer Engineering,
Curtin University of Technology, Sarawak Campus, Miri, Malaysia

Abstract: Problem statement: A study on hyperlink analysis and the algorithms used for link analysis in the Web Information retrieval was done. **Approach:** This research was initiated because of the dependability of search engines for information retrieval in the web. Understand the web structure mining and determine the importance of hyperlink in web information retrieval particularly using the Google Search engine. Hyperlink analysis was important methodology used by famous search engine Google to rank the pages. **Results:** The different algorithms used for link analysis like PageRank (PR), Weighted PageRank (WPR) and Hyperlink-Induced Topic Search (HITS) algorithms are discussed and compared. PageRank algorithm was implemented using a Java program and the convergence of the PageRank values are shown in a chart form. **Conclusion:** This study was done basically to explore the link structure algorithms for ranking and compare those algorithms. The further research on this area will be problems facing PageRank algorithm and how to handle those problems.

Key words: Web mining, web content, web structure, web graph, information retrieval, hyperlink analysis, PageRank, weighted PageRank and HITS

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 (Da Gomes and Gong, 2005) 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) and machine learning 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 study is organized as follows. The basic problems in the Web structure mining and the mining categories are discussed below. After that our work on the link analysis and the PageRank computation is shown and then three popular hyperlink analysis algorithms namely PageRank, WPR and HITS are compared. At the end the study is concluded.

Web structure mining: According to Kosala and Blockeel (2000), 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 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),

Corresponding Author: P. Ravi Kumar, Department of Electrical and Computer Engineering, Curtin University of Technology, Sarawak Campus, Miri, Malaysia

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.

How big is web: A Google report says that there are 1 trillion (1,000,000,000,000) unique Universal Resource Locator (URLs) 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.

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*. (2008). 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.

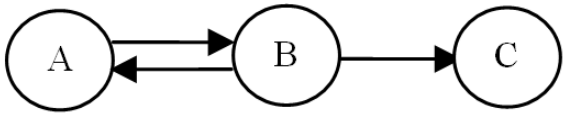

Fig. 1: A directed Graph (G)

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 graph in Fig. 1 is not strongly connected, as there is no path from vertex C to B. According to Broder *et al*. (2000), 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 paragraph explains the hyperlink analysis and the algorithms used in the hyperlink analysis for information retrieval.

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 (Chakrabarti *et al*., 1999; Haveliwala *et al*., 2002; Varlamis *et al*., 2004; Gibson *et al*., 1998; Kumar *et al*., 1999) 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 (Dean and Henzinger, 1999) and extended co-citation algorithm (Hou and Zhang, 2003) are proposed. These algorithms are simple and deeper relationships among the pages cannot be discovered. Three important algorithms PageRank (Brin and Page, 1998), Weighted PageRank (WPR) (Xing and Ghorbani, 2004) and Hypertext Induced Topic Search (HITS) (Kleinberg, 1999a) are discussed below in detail and compared.

MATERIALS AND METHODS

PageRank: We used this methodology to implement our link analysis algorithm. Brin and Page (1998) developed PageRank algorithm during their Ph.D. at Stanford University based on the citation analysis. 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.

RESULTS

Assume any arbitrary page A has pages T_1 to T_n pointing to it (incoming link). PageRank can be calculated by the following Eq. 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 are calculated manually by using Eq. 1. Then we implemented the PageRank program in Java and tested in an Intel Core 2 (2.40 Ghz) with 4GB RAM. The input i.e., the number of nodes and the number of outgoing links can be entered as input and the output PageRank value iterations for the 3 pages are produced in Excel CSV format. The PageRank iterations are shown as a chart on Fig. 3. The sample calculation is shown below.

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

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

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

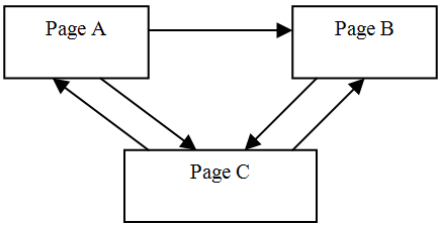

Fig. 2: Hyperlink structure for 3 pages

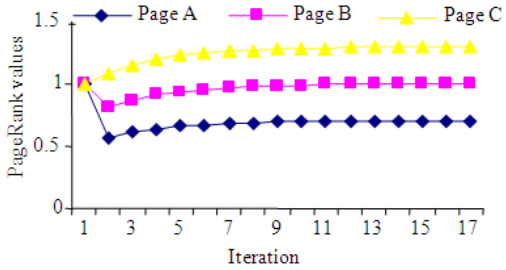

Fig. 3: Convergence of PageRank calculation

Table 1: Iterative calculation for PageRank

Iteration	PR (A)	PR (B)	PR (C)
0	1.000	1.000	1.000
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

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

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

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

$$PR(B) = 0.15 + 0.85(1.614/2)) = (1.091/2)) = 0.875 \quad (1e)$$

$$PR(C) = 0.15 + 0.85(1.614/2)) = (0.875/1)) = 1.155 \quad (1f)$$

After doing many more iterations of the above calculation, the PageRanks arrived as shown in Table 1.

For a smaller set of pages, the computation is easier but for a Web having billions of pages, the computation becomes more complex above.

Result analysis: In the Table 1, 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. So the link analysis becomes very important in the PageRank. From the Table 1, after the iteration 15, the PageRank for the pages gets normalized. The PageRank gets converged to a reasonable tolerance. The convergence of PageRank computation for the Table 1 is shown as a chart in Fig. 3.

Weighted PageRank algorithm: Xing and Ghorbani (2004) 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 Eq. 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 = The number of incoming links of page n and page p respectively

R(m) = Denotes the reference page list of page m

$W^{out}_{(m,n)}$ = As shown in Eq. 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

O_n and O_p = The number of outgoing links of page n and p respectively

The formula as proposed by Xing and Ghorbani (2004) for the WPR is as shown in Eq. 4 which is a modification of the PageRank formula (Eq. 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 Eq. 4d, 4e and 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 Eq. 4f-4h respectively. In this, WPR(A)>WPR(C)>WPR (B). This results shows that the PageRank order is different from PageRank.

The HITS algorithm-hubs and authorities: Kleinberg (1999a) 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 (1999a) 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 Hyperlink Induced Topic Search (HITS). 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.

Table 2: Comparison of hyperlink algorithms

	Algorithm		
Criteria	PageRank	Weighted PageRank	HITS
Mining technique used	WSM	WSM	WSM and 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 and 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

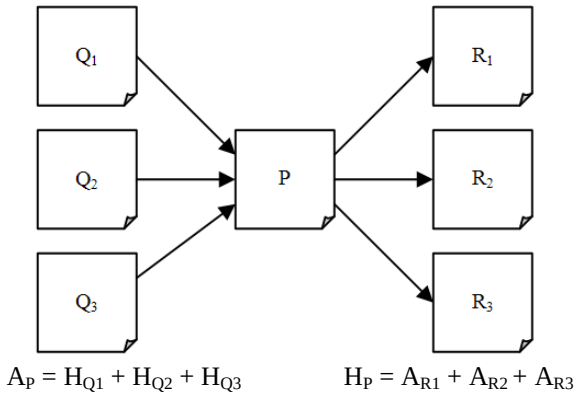

Fig. 4: Calculation of hubs and authorities

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 Eq. 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 = The hub weight
A_p = 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 (1999b). Similarly, a page's hub weight is proportional to the sum of the authority weights of pages that it links to. Figure 4 shows an example of the calculation of authority and hub scores.

The following are the constraints of HITS algorithm (Chakrabarti *et al*., 1999):

- 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 2 shows the comparison (Duhan *et al*., 2009) of all the algorithms discussed above

CONCLUSION

This study covers the basics of Web mining. The importance of the Web structure mining in Information retrieval is explained. The main purpose of this study is to explore the hyperlink structure and understand the Web graph in a simple way. The PageRank computation results shows that the incoming links and the outgoing links play an important role in ranking of Web pages using link analysis. This study also focuses on the important algorithms used for hyperlink analysis, explore those algorithms and compare them. This study is done basically to explore the link structure algorithms for ranking and compare those algorithms. The further work on this area will be problems facing PageRank algorithm and how to handle those problems.

ACKNOWLEDGEMENT

Authors would like to acknowledge Alex Goh Kwang Leng and Billy Lau Pik Lik, Computer Science students of Curtin University of Technology, Sarawak

for their contribution in the PageRank program implementation.

REFERENCES

Brin, S. and L. Page, 1998. The anatomy of a large scale hypertextual web search engine. Comput. Network ISDN Syst., 30: 107-117. DOI: 10.1016/S0169-7552(98)00110-X

Broder, A., R. Kumar, F. Maghoul, P. Raghavan and S. Rajagopalan *et al*., 2000. Graph structure in the web. Comput. Networks: Int. J. Comput. Telecommun. Network., 33: 309-320. DOI: 10.1016/S1389-1286(00)00083-9

Chakrabarti, S., B. Dom, D. Gibson, J. Kleinberg and R. Kumar *et al*., 1999. Mining the link structure of the world wide web. IEEE Comput., 32: 60-67. DOI: 10.1.1.62.546

Da Gomes Jr., M.G. and Z. Gong, 2005. Web structure mining: An introduction. Proceeding of the IEEE International Conference on Information Acquisition, June 27-July 3, IEEE Xplore Press, Hong Kong and Macau, China, pp: 6. DOI: 10.1109/ICIA.2005.1635156

Dean, J. and M. Henzinger, 1999. Finding related pages in the world wide web. Comput. Networks: Int. J. Comput. Telecommun. Network., 31: 1467-1479. DOI: 10.1016/S1389-1286(99)00022-5

Duhan, N., A.K. Sharma and K.K. Bhatia, 2009. PageRanking algorithms: A survey. Proceeding of the IEEE International Conference on Advance Computing, Mar. 6-7, IEEE Xplore Press, Patiala, India, pp: 1-1.

Gibson, D., J. Kleinberg and P. Raghavan, 1998. Inferring web communities from link topology. Proceeding of the of the 9th ACM Conference on Hypertext and Hypermedia, June 20-24, ACM Press, PA., USA., pp: 225-234. DOI: 10.1145/276627.276652

Haveliwala, T.H., A. Gionis, D. Klein and P. Indyk, 2002. Evaluating strategies for similarity search on the web. Proceeding of the 11th International Conference on WWW, May 7-11, ACM Press, Hawaii, USA, pp: 432-442. DOI: 10.1145/511446.511502

Horowitz, E., S. Sahni and S. Rajasekaran, 2008. Fundamentals of Computer Algorithms. Galgotia Publications Pvt. Ltd., ISBN: 81-7515-257-5, pp: 112-118.

Hou, J. and Y. Zhang, 2003. Effectively finding relevant web pages from linkage information. IEEE Trans. Knowl. Data Eng., 15: 940-951. DOI: 10.1109/TKDE.2003.1209010

Kleinberg, J., 1999a. Authoritative sources in a hyper-linked environment. J. ACM, 46: 604-632. DOI: 10.1145/324133.324140

Kleinberg, J., 1999b. Hubs, authorities and communities. ACM Comput. Surveys, 31: 1-3. DOI: 10.1145/345966.345982

Kosala, R. and H. Blockeel, 2000. Web mining research: A survey. Newsletter ACM Spec. Interest Group Knowl. Discov. Data Min., 2: 1-15. DOI: 10.1145/360402.360406

Kumar, R., P. Raghavan, S. Rajagopalan and A. Tomkins, 1999. Trawling the web for emerging cyber-communities. Comput. Networks: Int. J. Comput. Telecommun. Network., 31: 1481-1493. DOI: 10.1016/S1389-1286(99)00040-7

Varlamis, I., M. Vazirgiannis, M. Halkidi, B. Nguyen and Thesus, 2004. A closer view on web content management enhanced with link semantics. IEEE Trans. Knowl. Data Eng. J., 16: 685-700. DOI: 10.1109/TKDE.2004.16

Xing, W. and A. Ghorbani, 2004. Weighted PageRank algorithm. Proceeding of the 2nd Annual Conference on Communication Networks and Services Research, May 19-21, IEEE Computer Society, Washington DC., USA., pp: 305-314. DOI: 10.1109/DNSR.2004.1344743