Information Retrieval By Analysing Hyperlinks –Web Structure Mining

B. Rajdeepa, Dr. P. Sumathi
Research Scholar Department of Computer Science, Chikkanna Government Arts College, Tirupur.
Asst.Prof, PG & Research Department of Computer Science,Govt.Arts College,Coimbatore.

Abstract

The World Wide Web is a fertile area for data mining research. From its very beginning, the potential of extracting valuable knowledge from the Web has been quite unmistakable. Web mining is the application of data mining techniques to extract knowledge from Web content, structure, and usage mining and this is the collection of technologies to fulfill this potential. This paper discusses about the techniques of web structure mining and various algorithms are discussed with examples. Web Structure Mining can be regarded as the process of discovering structure Information from the Web i.e. hyperlinks and document structure.

1. Introduction

Web mining is the application of data mining techniques to extract knowledge from Web data - including Web documents, hyperlinks between documents, usage logs of web sites, etc. The World Wide Web is a popular and interactive medium to disseminate today's information. The web is very huge, diverse, and dynamic and thus raises for the scalability, multimedia data, and temporal issues respectively. Web mining is the application of data mining techniques to extract knowledge from Web data, where at least one of structure or usage data is used in the mining process.

2. Web Mining Process

Following are various process of data mining

a. Finding Resources: - It is the task of retrieving intended web documents.

b. Information collection and pre-processing:- Automatically selecting and pre-processing specific from information retrieved web resources.

c. Simplification:-Automatically discovers common patterns from web sites.

d. Making Analysis:-Validation and interpretation of the mined patterns.

3. Categories of Web Mining

Web mining can be categorized as

Web Content Mining: is the application of data mining techniques to unstructured or semi-structured data, i.e. HTML-documents.

Web Structure Mining: use of the hyperlink structure of the Web as an additional information source.

Web Usage Mining: analysis of user interactions' patterns from a Web server.

4. Web Structure Mining Tasks

With the rapid growth of the web, there is an increasing volume of links and structure available in various web pages. Web structure mining refers to develop new techniques to effectively extract and mine useful information/knowledge from the web pages. There are various methods of web structure mining.

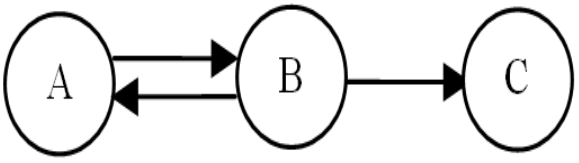
Hyperlinks: A Hyperlink is a structural unit that connects a Web page to different locations of web pages i.e. within the same page or to a different page. A hyperlink that connects to a different part of the same page is called an Intra-Document Hyperlink, and a hyperlink that connects two different pages is called an Inter-Document Hyperlink.

Document Structure: In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page.

These links are used to retrieve useful knowledge and helps users to get data in meaningful and quick manner.

5. Web Structure Mining Algorithms

The Web data Structure is available in different formats and one that format is through hyperlinks and Web structure mining uses this structure to provide additional information to users. Hyperlinks are one where different documents are connected. The Web may be viewed as a directed graph where nodes are identified as documents or pages and the edges as hyperlinks between them. This is called web graph. A graph G consists of two sets V and E. 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 is 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 graph in Fig. 1 is not strongly connected, as there is no path from vertex C to B.

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.

Hyperlink analysis:

Many web pages do not have text to explain the basic purpose of their data instead they may have as a hyperlink or images or videos here we face the difficulty to retrieve data here we use structure and meta-data included in the hyperlink to retrieve the data. Many algorithms are proposed based on the link analysis. Using citation analysis, co-citation algorithm and extended co-citation algorithm are proposed but these algorithms are simple and deeper relationships among the pages cannot be discovered. To overcome the problems of the above three more algorithms are proposed they are PageRank algorithm, Weighted PageRank (WPR) and Hypertext Induced Topic Search (HITS) and they are discussed in detail in this paper.

MATERIALS AND METHODS

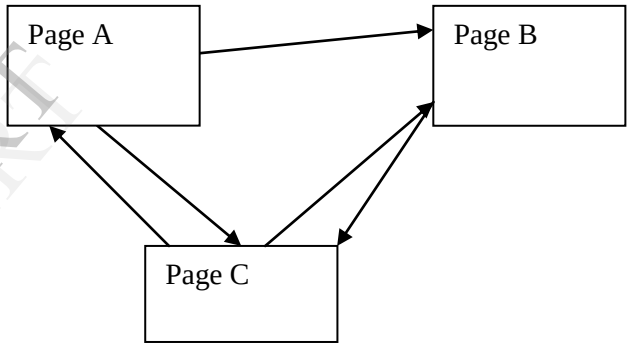
There are various techniques available through which we can mine useful information. Here, in this paper, I am describing various algorithms used to fetch information. Various algorithms are described as below.

PageRank: We used this methodology to implement our link analysis algorithm. Brin and Page (1998) developed PageRank algorithm based on the citation analysis. Google search engine uses this PageRank algorithm. 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 but it did not result in efficient outcomes. Therefore, PageRank algorithm provides a more efficient 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 web page A has pages T1 to Tn pointing to it (incoming link). PageRank can be calculated by the following Eq. 1:

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

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.



"Figure 2. Hyperlink structure for 3 pages"

After doing many more iterations of the above equation, the PageRanks arrived.For a smaller set of pages, the computation is easier but for a Web having billions of pages, the computation becomes more complex above. 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. After some iteration the PageRank for the pages gets normalized. The PageRank gets converged to a reasonable tolerance.

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 Win(m,n) and Wout(m,n) respectively. Win(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:

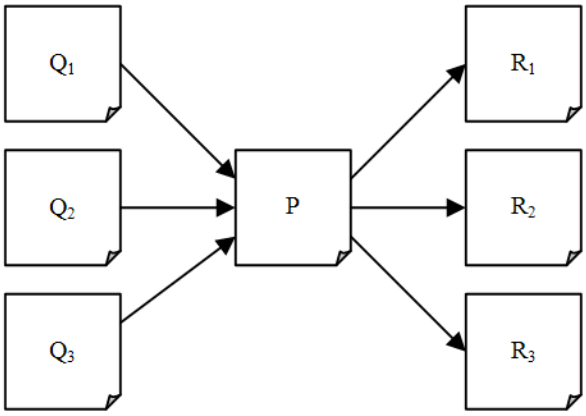
$$W \text{ in } (m,n) = I\ n / \sum_{p \in R(m)} I\ p \qquad (2)$$

Similarly we can calculate for outgoing links and

$$WPR(n) = (1-d) + d / \sum_{m \in B(n)} WPR(m)\ W \text{ in } (m,n)\ Wout\ (m,n) \qquad (3)$$

The results shows that WPR(A) is greater than WPR(C) greater than WPR(B) and the results shows that the result of page rank and weighted page rank algorithm are different.

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. 3. 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.



"Figure 3. 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. 4 :

$$Hp = \sum_{q \in I(p)} Aq \qquad (4)$$

Hp is hub weight and Similary we can calculate for authority weight

The page's authority weight is proportional to the sum of the hub weights of pages that it links to it. 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 allowing are the constraints of HITS algorithm.

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.

6. Conclusion

This paper covers about Web mining and the importance of the Web structure mining in Information extraction. The main aim of this paper is to explore the hyperlink structure and to understand the Web graph in a simple way. In Page Rank computation results shows that the incoming links and the outgoing links play an important role in ranking of Web pages using link analysis and this paper also discusses about weighted page rank algorithm and HITS algorithm. HITS mainly concentrated on hubs and authorities. The further work on this area will be to overcome the difficulties faced by the algorithms and makes the search efficient by minimizing the time taken for retrieval and to retrieve the exact content through our search.

7. 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:

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 hyperlinked 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 cybercommunities.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