Web Mining Using Topic Sensitive Weighted PageRank

Shesh Narayan Mishra, Alka Jaiswal, Asha Ambhaikar

Abstract— The World Wide Web contains the large amount of information sources. While searching the web for particular topics, users usually fetch irrelevant and redundant information causing a waste in user time and accessing time of the search engine. So narrowing down this problem, user's interests and needs from their behavior have become increasingly important. Web structure mining plays an effective role in this approach. Some page ranking algorithms PageRank, Weighted PageRank are commonly used in web structure mining. The original PageRank algorithm search-query results independent of any particular search query. To yield more specific and accurate search results against a particular topic, we proposed a new algorithm Topic Sensitive Weighted PageRank based on web structure mining that will show the relevancy of the pages of a given topic is better determined, as compared to the existing PageRank, Topic sensitive PageRank and Weighted PageRank algorithms. For ordinary keyword search queries, Topic Sensitive Weigted PageRank scores will satisfy the topic of the query.

Index Terms— Web structure mining; Weighted PageRank; Topic sensitive PageRank; TSWPR.

1 INTRODUCTION

TTODAY, the World Wide Web is the popular and interactive medium to disseminate information. The Web is huge, diverse and dynamic. The Web contains vast amount of information and provides an access to it at any place at any time. The most of the people use the internet for retrieving information. But most of the time, they gets lots of insignificant and irrelevant document even after navigating several links. For retrieving information from the Web, Web mining techniques are used.

2 WEB MINING OVERVIEW

Web mining is an application of the data mining techniques to automatically discover and extract knowledge from the Web.

According to Kosala et al [3], Web mining consists of the following tasks:

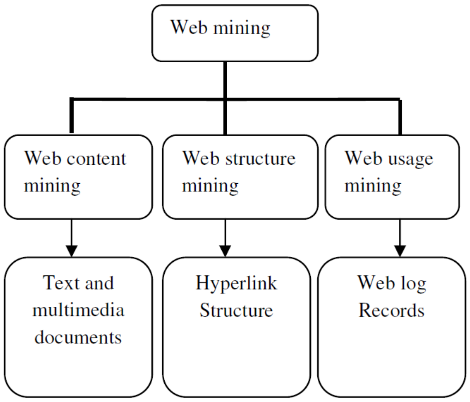
Resource finding: the task of retrieving intended Web documents.

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 usage mining, Web structure mining, and Web content mining. Web usage mining refers to the discovery of user access patterns from Web usage logs. Web structure mining tries to discover useful knowledge from the structure of hyperlinks which helps to investigate the node and connection structure of web sites. According the type of web structural data, web structure mining can be divided into two kinds 1) extracting the documents from hyperlinks in the web 2) analysis of the tree-like structure of page structure. Based on the topology of the hyperlinks, web structure mining will categorize the web page and generate the information,

- Mr. Shesh Narayan Mishra is presently studying in the final semester of M. Tech (Computer Technology) at Rungta College of Engineering and Technology, Bhilai, C.G., India. Has obtained degree of MCA from CSVTU, Bhilai University in 2009. Research Interest includes Web Mining Techniques
- *Alka Jaiswal presently working as Asst. Prof. in the Department of Information Technology at Rungta College of Engineering and Technology, Bhilai, C.G., India. Has been awarded degree of M.Tech (Information Security) with honors from National Institute of Technology, Bhopal (M.P.), India in 2010. Research Interest includes Database Security, Data Mining. 1 Research paper in international Journal has been published.*
- *Prof. Asha Ambhaikar is currently working as Associate Professor in Computer Science Engineering in RCET Bhilai, India, PH-09229655211. E-mail: asha31.a@rediffmail.com*

such as the similarity and mining is concerned with the retrieval of information from WWW into more structured form 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. Web classification is shown in Fig 1.

2.1 Web Content Mining (WCM)

Web Content Mining is the process of extracting useful information from the contents of web documents. The web documents may consists of text, images, audio, video or structured records like tables and lists. Mining can be applied on the web documents as well the results pages produced from a search engine. There are two types of approach in content mining called agent based approach and database based approach. The agent based approach concentrate on searching relevant information using the characteristics of a particular domain to interpret and organize the collected information. The database approach is used for retrieving the semi-structure data from the web.

2.2 Web Usage Mining (WUM)

Web Usage Mining is the process of extracting useful information from the secondary data derived from the interactions of the user while surfing on the Web. It extracts data stored in server access logs, referrer logs, agent logs, client-side cookies, user profile and meta data.

2.3 Web Structure Mining

The goal of the Web Structure Mining is to generate the structural summary about the Web site and Web page. It tries to discover the link structure of the hyperlinks at the inter-document level. Based on the topology of the hyperlinks, Web Structure mining will categorize the Web pages and generate the information like similarity and relationship between different Web sites. This type of mining can be performed at the document level (intra-page) or at the hyperlink level (inter-page). It is important to understand the Web data structure for Information Retrieval.

3 Related Work

3.1 PageRank

Brin and Page 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 "back links").

If a back link comes from an "important" page, then that back link is given a higher weighting than those back links 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 *T1* to *Tn* pointing to it (incoming link). *PageRank* can be calculated by the following.

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

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.

3.2 Weighted PageRank

Wenpu Xing and Ali Ghorbani [1] 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 $W^{in}_{(m,n)}$ as and $W^{out}_{(m,n)}$ respectively. $W^{in}_{(m,n)}$ as shown in (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*.

$$W^{in}_{(m,n)} = \frac{In}{\sum_{p \in R(m)} Ip} \qquad (2)$$

$$W^{out}_{(m,n)} = \frac{On}{\sum_{p \in R(m)} Op} \qquad (3)$$

Where and *Ip* 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 (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 *On* and *Op* 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 (4) which is a modification of the *PageRank* formula.

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

3.3 Topic Sensitive PageRank

In Topic Sensitive PageRank, several scores are computed: multiple importance scores for each page under several topics that form a composite PageRank score for those pages matching the query. During the offline crawling process, 16 topic-sensitive PageRank vectors are generated, using as a guideline the top-level category from Open Directory Project (ODP). At query time, the similarity of the query is compared to each of these vectors or topics; and subsequently, instead of using a single global ranking vector, the linear combination of the topic-sensitive vectors is weighed using the similarity of the query to the topics. This method yields a very accurate set of results relevant to the context of the particular query.

3.3.1 ODP Biasing

The first step in this approach is performed once during the offline preprocessing of the Web crawl. It generates a set of biased PageRank vectors using a set of basis topics. The specific methodology of creating the biased PageRank vectors is to use the URLs listed below the 16 top-level categories in ODP. These then, constitute the personalization vectors. A single, non-biased PageRank vector is also computed for comparison purposes. Third, a set of 16 class term-vectors is also computed, which consists of the terms in the documents below each of the 16 top-level categories.

3.3.2 Query-Time Importance Score

The second step is performed at query time. If a query q was issued by highlighting the term q in some web page u, then q (the context of q) consists of the terms in u. For ordinary queries not done in context, let $q = q$. using a unigram language model, with parameters set to their maximum-likelihood estimates, the class probabilities for each of the 16 top-level classes in ODP conditioned on q are computed. Using a text index, the URLs for all documents containing the original query terms q are retrieved. Finally, the query-sensitive importance score of each of these retrieved URLs is computed. The results are then ranked according to the composite score.

4 METHODOLOGY

We are using Weighted PageRank to implement our Topic Sensitive Weighted PageRank, we precompute importance scores offline using Weighted PageRank. For each page, we compute multiple importance scores, with respect to various topics. At time of query, the composite weighted page rank will be formed by combining these importance scores.

5 CONCLUSION

The rapid proliferation of World Wide Web has led the web content to increase tremendously. Hence, there is a great requirement to have algorithms that could list relevant web pages accurately and efficiently on the top of few pages. Mostly search engines used PageRank, Weighted PageRank but users may not get required documents easily. With a view to resolve the existing problems, a new algorithm called Topic Sensitive Weighted PageRank has been proposed which employs Web Structure mining. This algorithm will improve the order of web pages in the result list so that user may get the most relevant pages easily.

REFERENCES

[1] W. Xing and Ali Ghorbani, “Weighted PageRank Algorithm”, *Proc. ofthe Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE*, 2004.

[2] Taher H. Haveliwala. Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No4, July/August 2003, 784-796.

[3] 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.

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

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

[6] 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.

[7] X. Wang, T. Tao, J. T. Sun, A. Shakery and C. Zhai, “DirichletRank: Solving the Zero-One Gap Problem of PageRank”. *ACM Transaction on Information Systems*, Vol. 26, Issue 2, 2008.

[8] Z. Gyongyi and H. Garcia-Molina, “Web Spam Taxonomy”. *Proc. of the First International Workshop on Adversarial Information Retrieval on the Web”*, 2005.

[9] M. Bianchini, M.. Gori and F. Scarselli, “Inside PageRank”. *ACM Transactions on Internet Technology*, Vol. 5, Issue 1, 2005

[10] C.. H. Q. Ding, X. He, P. Husbands, H. Zha and H. D. Simon, “PageRank: HITS and a Unified Framework for Link Analysis”. *Proc. of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002.

[11] J. Cho and S. Roy, “Impact of Search Engines on Page Popularity”. *Proc. of the 13th International Conference on WWW*, pp. 20-29, 2004.

[12] J. Cho, S. Roy and R. E. Adams, “Page Quality: In search of an unbiased web ranking”. *Proc. of ACM International Conference on Management of Data”*. Pp. 551-562, 2005.

[13] A. M. Zareh Bidoki and N. Yazdani, “DistanceRank: An intelligent ranking algorithm for web pages” *Information Processing and Management*, Vol 44, No. 2, pp. 877-892, 2008.

[14] 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 Society Press*, Vol 32, Issue 8 pp. 60 – 67, 1999.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tompkins and E. Upfal, “Web as a Graph”, *Proceedings of the Nineteenth ACM SIGMOD-SIGACT-SIGART symposium on Database systems*, 2000.

[20] R. Cooley, B. Mobasher and J. Srivastava, “Web Minig: Information and Pattern Discovery on the World Wide Web”. *Proceedings of the 9th*

IEEE International Conference on Tools with Artificial Intelligence, pp. (ICTAI'97), 1997.

[21] Sung Jin Kim and Sang Ho Lee, “An Improved Computation of the PageRank Algorithm”, In proceedings of the European Conference on Information Retrieval (ECIR), 2002.

[22] Ricardo Baeza-Yates and Emilio Davis ,"Web page ranking using link attributes" , In proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, PP.328-329, 2004.