Web Mining and its Methods

B. Rajdeepa, Dr. P. Sumathi

Abstract— The World Wide Web has becoming one of the most valuable resources for information retrievals and knowledge discoveries. The WWW is a fertile area for data mining research. From its very beginning, the potential of extracting valuable knowledge from the Web has been quite evident. Web mining - i.e. the application of data mining techniques to extract knowledge from Web content, structure, and usage - is the collection of technologies to fulfill this potential. In this we review the Web mining techniques. Then we focus on one of these techniques: the Web structure mining. Within this technique, we introduce link mining and review two popular methods applied in Web structure mining: HITS and PageRank.

Index Terms— Data Mining, HITS, Page Rank, Web Content Mining, Web Mining, Web Structure Mining, Web Usuage Mining

1 Introduction

Web mining is the application of data mining techniques to extract knowledge from web data. Web mining can be considered as the applications of the general data mining techniques to the Web. Even though Web contains huge volume of data, it is distributed on the internet. Before mining, we need to gather the Web document together.Web pages are semi-structured, in order for easy processing; documents should be extracted and represented into some format.

2 Web Mining Overview

Web mining has divided into the following tasks:
1. Resource finding: the task of retrieving intended Web documents.
2. Information selection and pre-processing: automatically selecting and pre-processing specific information from retrieved Web resources.
3. Generalization: automatically discovers general patterns at individual Web sites as well as across multiple sites.
4. Analysis: validation and/or interpretation of the mined patterns.
Web mining tasks can be classified into three categories: Web content mining, Web structure mining and Web usage mining. There are two different approaches to categorize Web mining. In both, the categories are reduced from three to two: Web content mining and Web usage mining. In one, Web structure is treated as part of Web Content; while in the other, Web usage is treated as part of Web Structure. All of the three categories focus on the process of knowledge discovery of implicit, previously unknown and potentially useful information from the Web. Each of them focuses on different mining objects of the Web. As follows, we provide a brief introduction about each of the categories.

- *B.Rajdeepa is currently pursuing Ph.D. in Computer Science in Chikkanna Government Arts College, Tirupur. E-mail: rajdeepab@gmail.com*
- *Dr.P.Sumathi ,Asst.Prof, PG & Research Department of Computer Science,Govt.Arts College,Coimbatore.*

Web content mining targets the knowledge discovery, in which the main objects are the traditional collections of text documents and, more recently, also the collections of multimedia documents such as images, videos, audios, which are embedded in or linked to the Web pages. Web content mining could be differentiated from two points of view: the agent-based approach or the database approach. The first approach aims on improving the information finding and filtering and could be placed into the following three categories:
1. *Intelligent Search Agents.* These agents search for relevant information using domain characteristics and user profiles to organize and interpret the discovered information.
2. *Information Filtering/ Categorization*. These agents use information retrieval techniques and characteristics of open hypertext Web documents to automatically retrieve, filter, and categorize them.
3. *Personalized Web Agents*. These agents learn user preferences and discover Web information based on these preferences, and preferences of other users with similar interest.
The second approach aims on modeling the data on the Web into more structured form in order to apply standard database querying mechanism and data mining applications to analyze it. The two main categories are multilevel databases and Web query systems. For further information about Web content mining.
Web structure mining focuses on the hyperlink structure of the Web. The different objects are linked in some way. Simply applying the traditional processes and assuming that the events are independent can lead to wrong conclusions. However, the appropriate handling of the links could lead to potential correlations, and then improve the predictive accuracy of the learned models. Two algorithms that have been proposed to lead with those potential correlations: HITS and PageRank, and Web structure mining itself will be discussed in the next section.
Web usage mining focuses on techniques that could predict the behavior of users while they are interacting with the WWW. Web usage mining collects the data from Web log records to discover user access patterns of Web pages. There are several available research projects and commercial prod-

ucts that analyze those patterns for different purposes. The applications generated from this analysis can be classified as personalization, system improvement, site modification, business intelligence and usage characterization.

The challenges involved in web usage mining could be divided in three phases:

1. Pre-processing. The data available tend to be noisy, incomplete and inconsistent. In this phase, the data available should be treated according to the requirements of the next phase. It includes data cleaning, data integration, data transformation and data reduction.

2. Pattern discovery. Several different methods and algorithms such as statistics, data mining, machine learning and pattern recognition could be applied to identify user patterns.

3. Pattern Analysis. This process targets to understand, visualize and give interpretation to these patterns.

Web usage mining depends on the collaboration of the user to allow the access of the Web log records. Due to this dependence, privacy is becoming a new issue to Web usage mining, since users should be made aware about privacy policies before they make the decision to reveal their personal data.

We should note that there is no clear boundary between the above categories. As we mentioned, the two or three category definitions are quite acceptable, showing that Web content mining, Web structure mining and Web usage mining could be used isolated or combined in an application.

3 WEB STRUCTURE MINING

The challenge for Web structure mining is to deal with the structure of the hyperlinks within the Web itself. Link analysis is an old area of research. However, with the growing interest in Web mining, the research of structure analysis had increased and these efforts had resulted in a newly emerging research area called Link Mining, which is located at the intersection of the work in link analysis, hypertext and web mining, relational learning and inductive logic programming, and graph mining. There is a potentially wide range of application areas for this new area of research, including Internet. The Web contains a variety of objects with almost no unifying structure, with differences in the authoring style and content much greater than in traditional collections of text documents. The objects in the WWW are web pages, and links are in, out- and co-citation (two pages that are both linked to by the same page). Attributes include HTML tags, word appearances and anchor texts. This diversity of objects creates new problems and challenges, since is not possible to directly made use of existing techniques such as from database management or information retrieval. Link mining had produced some agitation on some of the traditional data mining tasks. As follows, we summarize some of these possible tasks of link mining which are applicable in Web structure mining.

1. Link-based Classification.

Link-based classification is the most recent upgrade of a classic data mining task to linked domains. The task is to focus on the prediction of the category of a web page, based on words that occur on the page, links between pages, anchor text, html tags and other possible attributes found on the web page.

2. Link-based Cluster Analysis.

The goal in cluster analysis is to find naturally occurring sub-classes.

The data is segmented into groups, where similar objects are grouped together, and dissimilar object are grouped into different groups.

Different than the previous task, link-based cluster analysis is unsupervised and can be used to discover hidden patterns from data.

3. Link Type.

There are a wide range of tasks concerning the prediction of the existence of links, such as predicting the type of link between two entities, or predicting the purpose of a link.

4. Link Strength.

Links could be associated with weights.

5. Link Cardinality.

The main task here is to predict the number of links between objects. There are many ways to use the link structure of the Web to create notions of authority. The main goal in developing

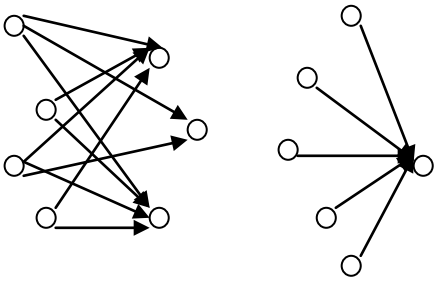
applications for link mining is to made good use of the understanding of these intrinsic social organization of the Web.

4 HITS AND PAGE RANK METHOD

In this section we review two approaches: HITS concept and PageRank method. Both approaches focus on the link structure of the Web to find the importance of the Web pages.

A. HITS: Computing Hubs and Authorities

In HITS concept, Kleinberg identifies two kinds of pages from the Web hyperlink structure: Authorities and hubs. For a given query, HITS will find authorities and hubs.



Hubs Authorities

Unrelated Page of large In-degree

Fig. 1 A densely linked set of Hubs and Authorities

Although HITS provides good search results for a wide range of queries, HITS did not work well in all cases due to the following three reasons:

1. Mutually reinforced relationships between hosts. Sometimes a set of documents on one host point to a single document on a second host, or sometimes a single document on one host point to a set of document on a second host. These situations could provide wrong definitions about a good hub or a good authority.

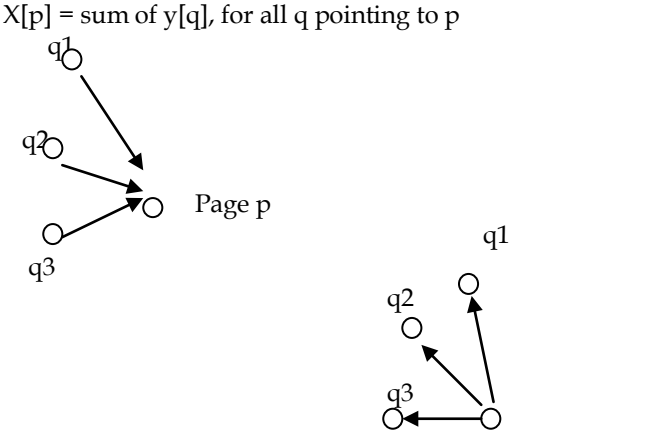


Fig. 2 The basic operations of HITS

2. Automatically generated links. Web document generated by tools often have links that were inserted
by the tool.
3. Non-relevant nodes. Sometimes pages point to other pages with no relevance to the query topic.

5 PageRank Model

L. Page and S. Brin proposed the Page Rank algorithm to calculate the importance of web pages using the link structure of the web. In their approach Brin and Page extends the idea of simply counting in-links equally, by normalizing by the number of links on a page. The Page Rank algorithm is defined as: "We assume page A has pages T1...Tn which point to it (i.e., are citations). The parameter d is a damping factor, which can be set between 0 and 1. We usually set d to 0.85. There are more details about d in the next section. Also C (A) is defined as the number of links going out of page A. The Page Rank of a page A is given as follows:

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

Note that the Page Ranks form a probability distribution over web pages, so the sum of all web pages' Page Ranks will be one." And "The d damping factor is the probability at each page the "random surfer" will get bored and request another random page."
Note that the rank of a page is divided evenly among its outlinks to contribute to the ranks of the pages they point to. The equation is recursive, but starting with any set of ranks and iterating the computation until it converges may compute it. Page Rank can be calculated using a simple iterative algorithm, and corresponds to the principal eigen vector of the normalized link matrix of the web. Page Rank algorithm needs a few hours to calculate the rank of millions of pages.

Applications

HITS was used for the first time in the Clever search engine from IBM, and PageRank is used by Google combined with other several features such as anchor text, IR measures, and proximity.

The notion of authoritativeness comes from the idea that we wish not only to locate a set of relevant pages, but rather the relevant pages of the highest quality. However, the Web consists not only of pages but also of links that connect one page to another. This structure contains a large amount of information that should be exploited.
PageRank and HITS belong to a class of ranking algorithms, where the scores can be computed as a fixed point of a linear equation. Bianchini noted that HITS and PageRank are used as starting points for new solutions, and there are some extensions of theses two approaches. There are other link-based approaches to be applied on the Web.
Besides being used for weighting Web pages, link resource can also be used for clustering or classifying Web pages. The principle is based on the assumption that (1) if page p1 has a link to page p2, p1 should be similar to p2 in content, and (2) if p1 and p2 are co-cited by some common pages, p1 and p2 should also similar. Web pages can be clustered into a lot of connected page communities with respect to their citation and co-citation strengths among the pages. In fact, Ziv Bar-Yossef and Sridhar Rajagopalan put all the algorithms, which uses links, to three categories.
1. Relevant Linkage Principle: Links points to relevant resources.
2. Topical Unity Principle: Documents often co-cited are related, as are those with extensive bibliographic overlap. This idea is previous addressed by Kesseler for bibliographic information retrieval in.
3. Lexical Affinity Principle: Proximity of text and links within a page is a measure of the relevance of one to another.
Even though those link algorithms can always provide a good support for Web information retrievals, clustering and knowledge discoveries on the Web, authors also find problems associated with those technologies.

6 Conclusion

In this paper we survey the research area of Web mining, focusing on the category of Web structure mining. We had introduced Web mining. Later in the paper when we had discussed Web structure mining, and introduced Link mining, as well as block-level link mining issues. We had also reviewed two popular algorithms to have an idea about their application and effectiveness. Since this is a huge area, and there a lot of work to do, we hope this paper could be a useful starting point for identifying opportunities for further research.

References

[1] O. Etzioni. "The World Wide Web: Quagmire or gold mine." Com-

munications of the ACM, 39(11):65-68, 1996.

[2] Raymond Kosala, Hendrik Blockeel, "Web Mining Research: A Survey", ACM SIGKDD Explorations Newsletter, June 2000, Volume 2 Issue 1.

[3] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pag-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", ACM SIGKDD Explorations Newsletter, January 2000, Volume 1 Issue 2.

[4] Jidong Wang, Zheng Chen, Li Tao, Wei-Ying Ma, Liu Wenyin, "Ranking User's Relevance to a Topic through Link Analysis on Web Logs", WIDM' 02, November 2002.

[5] G. Piatetsky-Shapiro, and W.J. Frawley, Knowledge Discovery in Databases. AAAI/MIT Press, 1991

[6] Q. Lu, and L. Getoor. Link-based classification. In Proceedings of ICML-03, 2003.

[7] .L. Getoor, "Link Mining: A New Data Mining Challenge". SIGKDD Explorations, vol. 4, issue 2, 2003.

[8] S. Chakrabarti, B. E. Dom, D, Gibson, J. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Mining the Link Structure of the World Wide Web. February, 1999.

[9] L. Page, S. Brin, R. Motwani, and T. Winograd. The Pagerank citation ranking: Bring order to the web. Technical report, Stanford University, 1998.

[10] Han, J., Kamber, M. Kamber. "Data mining: concepts and techniques". Morgan Kaufmann Publishers, 2000.

[11] Cooley, R.; Mobasher, B.; Srivastava, J.,"Web mining: information and pattern discovery on the World Wide Web." Tools with Artificial Intelligence, 1997. Proceedings., Ninth IEEE International Conference. Page(s):558 – 567 - 3-8 Nov. 1997.

[12] Kleinberg, J.M., Authoritative sources in a hyperlinked environment. In Proceedings of ACM-SIAM Symposium on Discrete Algorithms, 1998, pages 668-677 – 1998.

[13] Bianchini, M.; Gori, M.; Scarselli, F.; Inside PageRank. ACM Transaction on Internet Technology (TOIT), Volume 5 Issue 1 – February, 2005.

[14] Bianchini, M.; Gori, M.; Scarselli, F.; Inside PageRank. ACM Transaction on Internet Technology (TOIT), Volume 5 Issue 1 – February, 2005.

[15] Brin, S.; Page,L. The Anatomy of a Large-scale Hypertextual Web Search Engine. Proceedings of the Seventh International World Wide Web Conference, 1998.

[16] Ziv Bar-Yossef and Sridhar Rajagopalan., "Template Detection via Data Mining and its Applications." In: Proceedings of WWW2002, May 7-11, 2002, Honolulu, Hawaii, USA. 580-591.

[17] M. Kessler. Bibliographic coupling between scientific papers. American Documentation, 14:10-25, 1963.

[18] 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.

[19] Deng Cai, Shipeng Yu, and et al. "Block-based Web Search." In Proceedings of ACM SIGIR'04, July 25-29, 2004, Sheffield, South Yorkshire, UK. 465-463.

[20] Sian-Hua Lin and Jan-Ming Ho. "Discovering Information Content Blocks from Web Documents." In: Proceedings of ACM SIGKDD'02, July 23-26, 2002, Edmonton, Alberta, Canada. 588-59