A Novel Approach for conventional web spider results and log mining

C.SUSHAMA
Assistant Professor
Department of CSSE
Sree Vidyanikethan Engineering College, Tirupati
sushama05516@yahoo.co.in

P.NEELIMA
Assistant Professor
Dept of CSE,
Siddhartha Educational Academy Group of Institutions,
C.Golapalli, Tirupati
neelima.pannem@gmail.com

M. SUNIL KUMAR
Research Scholar, Dept of CSE
S V University
Tirupati
sunilmalchi@yahoo.co.in

Abstract— **Despite of the popularity of global search engines, people still suffer from low accuracy of site search. The primary reason lies in the differences of link structures and data scale between global web and websites, which leads to failures of traditional reranking methods such as HITS, PAGE RANK and DIRECT HIT. This paper proposes a novel reranking method based on user logs with in websites. With the help of website Taxonomy, We mine for generalized association rules and abstract access patterns of different levels. Mining results are subsequently used to re-rank the retrieved pages. One of the advantages of our mining algorithm is that it resolves the diversity problem of user's access behavior and discovers general patterns. Experiment shows that the proposed method outperforms keyword based method by 15% and Direct Hit by 13% respectively.**

1. INTRODUCTION

Web Mining is the extraction of interesting and potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web. There are roughly three knowledge discovery domains that pertain to web mining

- Web Content Mining.
- Web Structure Mining.
- Web Usage Mining.

Web content mining is the process of extracting knowledge from the content of documents or their descriptions. Web document text mining, resource discovery based on concepts indexing or agent based technology may also fall in this category. Web structure mining is the process of inferring knowledge from the Worldwide Web organization and links between references and referents in the Web. Finally, web usage mining, also known as Web Log Mining, is the process of extracting interesting patterns in web access logs.

1.1 WEB CONTENT MINING:

Web content mining is an automatic process that goes beyond keyword extraction. Since the content of a text document presents no machine-readable semantic, some approaches have suggested restructure the document content in a representation that could be exploited by machines.

The usual approach to exploit known structure in documents is to use wrappers to map documents to some data model. Techniques using lexicons for content interpretation are yet to come.

There are two groups of web content mining strategies: Those that directly mine the content of documents and those that improve on the content search of other tools like search engines.

1.2 WEB STRUCTURE MINING:

The goal of Web structure mining is to generate structural summary about the Web site and Web page. Technically, Web content mining mainly focuses on the structure of inner-document, while Web structure mining 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, such as the similarity and relationship between different Web sites.

Web structure mining can also have another direction discovering the structure of Web document itself. This type of structure mining can be used to reveal the structure (schema) of Web pages; this would be good for navigation purpose and make it possible to compare/integrate Web page schemes. This type of structure mining will facilitate introducing database techniques for accessing information in Web pages by providing a reference schema.

In general, if a Web page is linked to another Web page directly, or the Web pages are neighbors, we would like to discover the relationships among those Web pages. The relations maybe fall in one of the types, such as they related by synonyms or ontology, they may have similar contents, and both of them may sit in the same Web server created by the same person.

Another task of Web structure mining is to discover the nature of the hierarchy or network of hyperlink in the Web sites of a particular domain. This may help to generalize the

flow of information in Web sites that may represent some particular domain; therefore the query processing will be easier and more efficient.

I. 1.3 WEB USAGE MINING

Web usage mining is the type of Web mining activity that involves the automatic discovery of user access patterns from one or more Web servers.

As more organizations rely on the Internet and the World Wide Web to conduct business, the traditional strategies and techniques for market analysis need to be revisited in this context. Organizations often generate and collect large volumes of data in their daily operations. Most of this information is usually generated automatically by Web servers and collected in server access logs.

Web usage mining consists of three phases:

- Preprocessing
- Pattern discovery
- Pattern analysis.

These phases are described below.

Preprocessing consists of converting the usage, content, and structure information contained in the various available data sources into the data abstractions necessary for pattern discovery.

PATTERN DISCOVERY

Pattern discovery draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. Here, some of mining activities that have been applied to the Web domain. Methods developed from other fields must take into consideration the different kinds of data abstractions and prior knowledge available for Web Mining.

PATTERN ANALYSIS

Pattern analysis is the last step in the overall Web Usage mining process. The motivation behind pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase. The exact analysis methodology is usually governed by the application for which Web mining is done. The most common form of pattern analysis consists of a knowledge query mechanism such as SQL.

Another method is to load usage data into a data cube in order to perform OLAP operations. Visualization techniques, such as graphing patterns or assigning colors to different values, can often highlight overall patterns or trends in the data. Content and structure information can be used to filter out patterns containing pages of a certain usage type, content type, or pages that match a certain hyperlink structure.

2.1 PROBLEM DEFINITION:

Now-a-days there are so many search engines available for people to find desired information on web. The main problem with search engines which uses Traditional reranking methods for site search are

i. Search time
ii. Search space
iii. Low accuracy

So, we propose a Novel reranking method to improve the time and space complexities. Novel reranking method is based on user logs within websites. Novel reranking method uses Log mining.

2.2 PROBLEM DESCRIPTION:

Global search engines such as Google, AltaVista have been great helpful to users to find desire information on ever growing web. given clear and unambiguous queries, they can return desire results most of time .however this is the always the case as point out users often pose unclear and general queries to find appropriate websites as good starting points ones at a site the user has the choice of following hyperlinks or using site search to get more specific information due to the low accuracy and efficiency of following hyperlinks there iis a tremendous need for site search techniques. In addition globle search engines can't index content with in dynamic websites or intranet , where site search is only way for users to find information .

Site search can be simply defined as search functionality specific to one web site however unlike global search engines site search engines are notoriously problematic at present.

Most site search engines merely use "full text search" methods which retrieves large amount of documents containing the same keywords inputted by user. Due to the shortness of query words and poor ranking mechanisms it's a time consuming job for the user to go through the results to find out their really desired information.

Many techniques which are very successful in web search seem directly applicable in site search such as link analysis and click thru –based rankings .But either of them can work well for the following reasons. First the link analysis techniques such as hits and page rank use hyperlinks among web pages to rank pages, where pages with more reference gets higher ranking score. However the link information within web site is not enough to reflect the page score. thus the most important web pages are not necessarily the highest referenced pages, high reference pages are often home pages index pages help pages which are not really wanted by users. Thus the failure of applying link analysis to web also demonstrates that link analysis does not work for the sub sequence of web.

Second click thru based ranking method such as direct hit are most problematic to be used in site search engines according to particular query direct hit utilizes previous sessions logs of same query to return pages that most users visited because of lack of previous query sessions and

diversity of users access patterns direct hit does not work for engines.

We propose a novel reranking method based on site logs every website keeps a set of access logs which embody browsing behaviors of its users and the time, duration and url . we obtain from this logs each page access frequency and the traversal pattern of information finding.

Generally the process of discovering useful patterns from web logs is called log mining. log mining includes straight forward statistics methods such as page access frequency as well as more sophisticated form of analysis such as association rules mining sequential pattern mining clustering.

A normal association rule mining algorithms may find to discover significant rules due to data diversity problems we use generalized association rules mining to utilize a predefined taxonomy and extract significant association rules at different abstract level of taxonomy this association rules are subsequently pruned and applied to page re-ranking.

3.MODEL FOR GENERALIZED ASSOCIATION RULE:

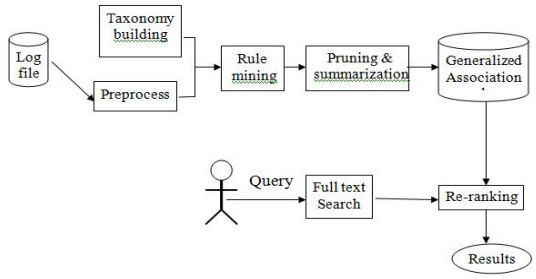


Fig: Flow of site search using generalized association rule

DESCRIPTION:

The user transactions are stored in log files which is present at the server side. The log file contains the time, session ids, user Ids and their corresponding url's. Each user is given a unique session Id.

3.1.1 LOG FILE GENERATION:

The first module is log file generation. This log file contains the time,userid,sessionid and the web pages they have visited.

A sample log file is given below:

14/4/2007 11:25:28-430-admin-/logmine/pages/c/email.html

14/4/2007 11:25:32-430-admin-/logmine/pages/c/get.html

14/4/2007 11:25:36-430-admin-/logmine/pages/c/browse.html

14/4/2007 11:25:39-365-admin-/logmine/pages/c/music.html

14/4/2007 11:25:42-365-admin-/logmine/work.html

3.1.2 LOG PREPROCESSING:

Here, the log entries obtained from the log file is preprocessed. Preprocessing is nothing but eliminating the redundant accesses and any incomplete accesses of users.

3.1.3 RULE MINING:

After log preprocessing, we do Apriori algorithm. Apriori algorithm consists two steps:

i. It finds frequent itemsets.

ii. It generates association rules by using these frequent itemsets

3.1.4 ASSOCIATION RULES:

Let $I = \{i1, i2, \ldots, im\}$ be a set of *items*. An *association rule* is an implication of the form X_Y, where $X \grave{I} I$, $Y \grave{I} I$, and $X \Ç Y = \text{Æ}$.

For example, suppose users who accessed page *a* and page *b* also tend to access page *c*. The corresponding association rule is "*A^B _ C*". The antecedent of the rule *X* consists of *A* and *B*, and the consequent *Y* consists of *C*.

3.1.5 PRUNING:

Apriori algorithm generates so many association rules. In these rules some are interesting and some are uninteresting. Pruning is the process of eliminating these uninteresting rules.

3.2 GENERALIZED ASSOCIATION RULES:

Our method mines for generalized association rule instead of standard association rule to tackle the data diversity problems. Earlier works on association rule mining, only mine for relationships among distinct Web pages, which lead to three main problems:

- First, our statistics on a real Web log show that most of the pages have low hit rate. As can be seen from Figure 5, about 80% of the pages are visited less than 10 times. If we use standard association rule, those pages are always ignored. However, in most of cases, these pages may contain latent information about the user's access patterns which can't be discovered using standard association rule mining.

- Second, a website usually contains thousands, even millions of pages. It is not easy to find those users who access some common pages. This is because of the diversity of the users. Moreover, this also leads to the difficulty of finding the same access pattern.

- Third, there are some latent semantic topics in a website. For example, there exist two topics in Berkeley CS's website: AI and machine learning, each consisting of several pages. These two topics are frequently co-visited. However,

the standard association rule is unable to find the relationships between these two topics.

3.3 RE-RANKING:

We propose a novel algorithm to re-rank the results using generalized association rule. In general, the pages in an association rule are accessed frequently together and mostly the content of the pages are relative, so we can use the association rule to improve the performance of site search.

Our algorithm is similar to DirectHit. Both of them make use of the previous users' query sessions. DirectHit algorithm uses the click popularity to improve the performance of the search. The higher frequency a Web page is visited by user, the more important the Web page is. Compared to DirectHit, our algorithm utilizes the popular clicked pages of the same query and improves the associate pages' rank which is based on the association rules.

First we implement the DirectHit algorithm on a "full text search" engine. DirectHit uses the sessions of the users' queries and the pages which are relative to the users' queries.

1. Through the site search engine, the user inputs a query word *Q*, then the search engine returns a result set *D* with the score which is based on the similarity of the page *d* and the query *Q*.

2. The pages are re-ranked according to the similarity score and the click popularity;

Then we implement our algorithm using generalized association rules.

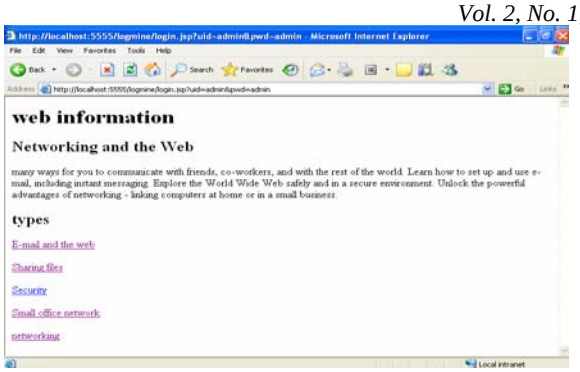
1. The user inputs a query word *Q*, and get a result set *D* with the score which is based on the similarity between the page and the query.

2. Similar to the pseudo relevance feedback approach, we get top n pages as a set P from the returned result.

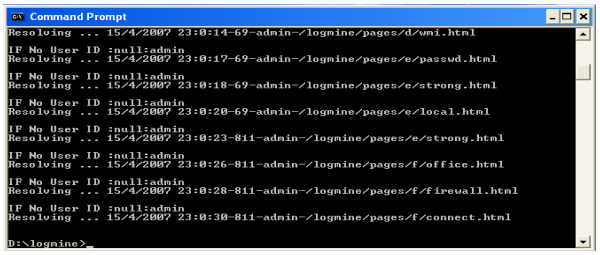
3. Then we get the rules whose antecedents contain the pages in *P* or the ancestors of the pages in *P*, and acquire the descendant of the rules as a set *R*.

4. According to *R*, we calculate the support and the confidence of each page *d* in the result set *D*, if the page *d* is in *R*, we calculate the support and the confidence of the page *d* directly, if a parent node of *d* is in *R*, we calculate the support and the confidence of the page *d* according to the ratio of the page to its parent.

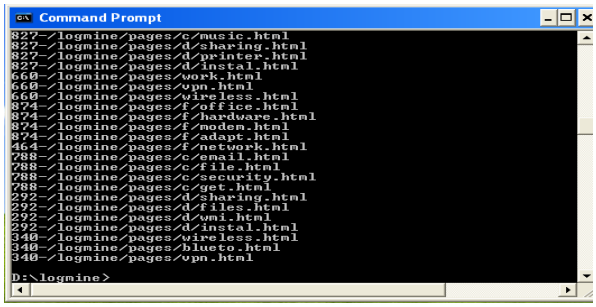
5. The result is re-ranked according to the similarity, the support and the confidence;



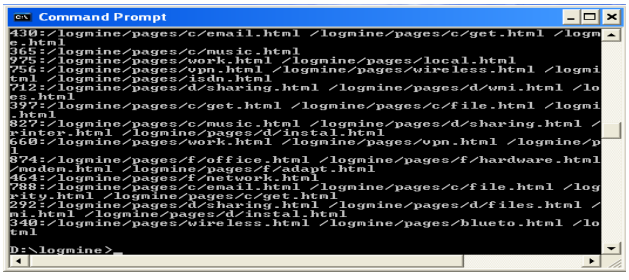
Here user can navigate from one page to another page by using hyperlinks among pages.



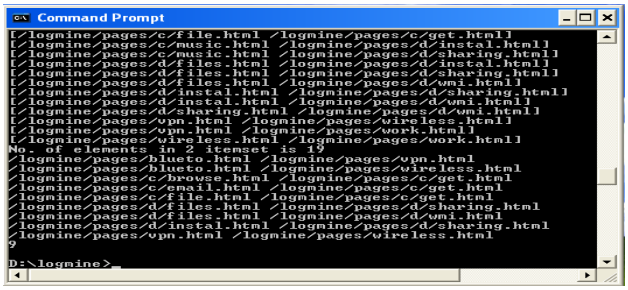
user which contains date, time, session Id, user Id and pages visited by admin.These outputs of individual users are stored in Log files.



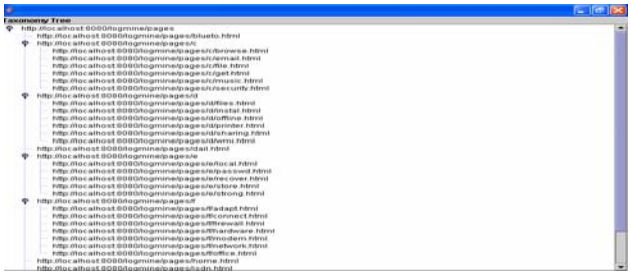
Log Process which contains only session Id's and URL's.



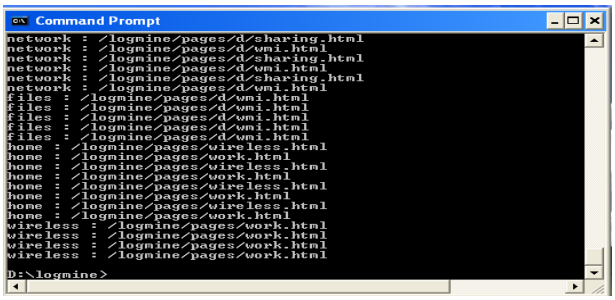
Shows transactions generated for log file



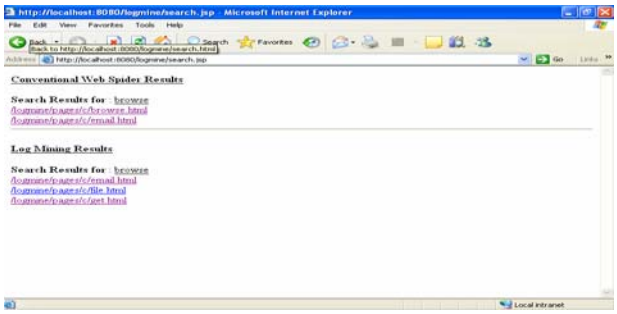
Shows association rules generated by using transactions.



Shows the Taxonomy tree for whole website.



Generalized association rules, in above screen keyword is present at left side and the corresponding URL for that keyword is present at right side**.** This updates the novel rules in database.



The conventional web spider results and log mining results. In above screen red link indicates that user visited that page and blue link indicates that user doesn't visited that page.

4. CONCLUSION

This paper discusses a log mining method for improving search functionality inside the website. We propose a generalized association rule mining method, which utilizes taxonomy of website to mine for association rules at different levels. We also propose a novel re-ranking method using these generalized association rules. Our experiments show that the method is efficient and feasible for site search.

The construction of the taxonomy is based on the URLs of the pages, which implies that the underlying page organization reflects the semantics of the pages. In case this cannot be assumed, the taxonomy should be constructed according to the semantics of the pages, e.g., using content-based hierarchical clustering method, which is one of our on-going works.

Our algorithm use generalized association rules to re-rank Web pages based on previous pages in the same user session. This search scheme is much restrictive in some cases. Intuitively, association rules among Web pages can be seen as additional hyperlinks (or implicit links) among Web pages. Thus they could be utilized in link analysis algorithm such as HITS and Page Rank.

How to calculate this kind of implicit links, combine implicit links and explicit hyperlinks together, and apply them to link analysis algorithms, are main lines of our future researches.

REFERENCES

[1] C.H Yun and M.S. Chen, "Mining Web Transaction Patterns in an Electronic Commerce Environment", in Proceedings of the 4th Pacific-Asia Conf. on Knowledge Discovery and DataMining,April 2000.

[2] D. Hawking, E. Voorhees, P. Bailey, and N. Craswell, "Overview of TREC-8 Web Track", In Proceedings of TREC-8, pp. 131–150,Gaithersburg MD, November 1999.

[3] DirectHit: http://www.directhit.com.

[4] H. Mannila, H. Toivonen, and A. I. Verkamo, "Discovery of frequent episodes in event sequences", Data Mining and Knowledge Discovery,1(3): 259 - 289, November 1997.

[5] J. Han, Jian Pei, and Yiwen Yin, "Mining Frequent Patterns without Candidate Generation", ACM SIGMOD Intl. Conference on Management of Data, 2000.

[6] J. Kleinberg, "Authoritative Sources in Hyperlinked Environment", in Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithm, 1998.

[7] J. Pei, J. Han, B. Mortazavi-Asl, and H. Zhu, "Mining Access Pattern efficiently from web logs", in Proceedings of 2000 Pacific-Asia Conf. on Knowledge Discovery and Data Mining, April 2000.

[8] M. Chen, J. Park, and P. S. Yu, "Efficient Data Mining for Path Traversal Patterns", IEEE Trans. on Knowledge and Data Engineering, Vol. 10, No. 2, pp. 209-221, April 1998.

[9] M. Hearst, "Next Generation Web Search: Setting Our Sites",IEEE Data Engineering Bulletin, Special issue on Next Generation Web Search, Luis Gravano(Ed.), September 2000.

[10] M. Spiliopoulou and C. Pohle, "Data mining for measuring and improving the success ofWeb sites", Data Mining and Knowledge Discovery, 5:85–14, 2001.

[11] M. Spiliopoulou, C. Pohle, and L. Faulstich, "Improving the effectiveness of a Web site with Web usage mining", In Advances in Web Usage Analysis and User Profiling, Berlin,Springer, pp. 142–162, 2000.

[12] M.Chen, M.Hearst, J. Hong, and J.Lin, "Cha-Cha: A System for Organizing Intranet Search Results", In Proceedings of the 2nd USITS, Boulder, CO, October 1999.

[13] M.Levene and R.Wheeldon, "A Web Site Navigation Engine",in Proceedings 10th International WWW Conference, 2001.

[14] P. Hagen, H. Manning, and Y. Paul, "Must search stink? The Forrester report", Forrester, June 2000.

[15] Q. Yang, H. Hanning Zhang, and I.Tianyi Li, "Mining Web Logs for Prediction Models in WWW Caching and Prefetching",In The Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD'01, Industry Applications Track,August 2001.

[16] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules", in Proceedings of the 20th Int'lConference on VLDB, Santiago,Chile, September 1994.

[17] R. Agrawal, T. Imielinski, and A. Swami, "Mining Associations between Sets of Items in Massive Databases", in Proceedings of the ACM-SIGMOD 1993 Int'l Conference on Management of Data, Washington D.C., May 1993.

[18] R. Baeza-Yates and B.Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley, 1999.

[19] R. Cooley, B. Mobasher, and J. Srivastava, "Data Preparation for Mining World Wide Web Browsing Patterns", Knowledge and Information Systems V1(1),1999.

[20] R. Cooley, P. Ning Tan, and J. Srivastava, "Discovery of Interesting Usage Patterns from Web Data", To appear in Springer-Verlag LNCS/LNAI series, 2000.

AUTHORS PROFILE

Mrs. C.Sushama has completed B.Tech in Electronics from JNT University and M.Tech in Computer Science from JNT University. Presently she is pursuing Ph.D in Computer Science and EngineeringS.V.University, TIRUPATI. She is currently working as Assistant Professor in the Department of CSE, Sree Vidyanikethan Engineering College, A. Rangampet, Tirupati, A.P. His main research interest includes Software Engineering, Software Architecture.

P.Neelima has completed B.Tech in Computer Science and Engineering from JNT University. She is currently working as Assistant Professor in the Department of CSE, Siddhartha Educational Academy Group of Institutions, C.Golapalli, Tirupati, A.P. Her main research interest includes Software Engineering, Software Architecture, Information Retrieval and Database Management Systems.

Mr. M Sunil Kumar has completed B.Tech in Computer Science & Information Technology from JNT University and M.Tech in Computer Science from JNT University. Presently he is pursuing Ph.D in Computer Science and Engineering, S.V.University, TIRUPATI. His main research interest includes Software Engineering, Software Architecture, Information Retrieval and Database Management Systems.