ALGORITHM RESEARCH ON USER INTERESTS EXTRACTING VIA WEB LOG DATA

*T. Sumitra *, Shaik Shasha Ali***

**M.Tech Student of Bharath College of Engineering and Technology for Women, Kadapa, AP*

***Assistant Professor, Department of CSE, Bharath College of Engineering and Technology for Women, Kadapa, AP*

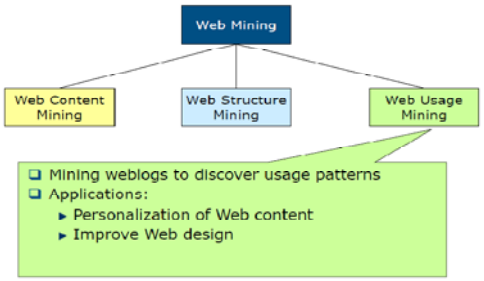
Abstract- The information on the web is growing dramatically. Without a recommendation system, the users may spend lots of time on the web in finding the information they are interested in. Today, many web recommendation systems cannot give users enough personalized help but provide the user with lots of irrelevant information. One of the main reasons is that it can't accurately extract user's interests. Therefore, analyzing users' Web Log Data and extracting users' potential interested domains become very important and challenging research topics of web usage mining. If users' interests can be automatically detected from users' Web Log Data, they can be used for information recommendation and marketing which are useful for both users and Web site developers. In this Project, we present some novel algorithms to mine users' interests. The algorithms are based on visit time and visit density which can be obtained from an analysis of web users' Web Log Data. Experimental results show that our new methods succeed in finding user's interested domains.

I. INTRODUCTION

Web mining - is the application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining.Web usage mining is the process of extracting useful information from server logs i.e. users history. Web usage mining is the process of finding out what users are looking for on internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data. This technology is basically concentrated upon the use of the web technologies which could help for betterment. Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. According to the type of web structural data, web structure mining can be divided a into two kinds:

1. Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location.

2. Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage. Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page contents. The heterogeneity and the lack of structure that permeates much of the ever expanding information sources on the World Wide Web, such as hypertext documents, makes automated discovery, organization, and search and indexing tools of the Internet and the World Wide Web We design our group analysis and publishing search logs with privacy related web mining. Search engine companies collect the database of intentions, the histories of their user’s search queries. These search logs are a gold mine for researchers.



Search engines play a crucial role in the navigation through the vastness of the Web. Today's search engines do not just collect and index web pages, they also collect and mine information about their users. They store the queries, clicks, IP-addresses, and other information about the interactions with users in what is called a search log .Search logs contain valuable information that search engines use to tailor their services better to their user's needs. They enable the discovery of trends, patterns, and anomalies in the search behavior of users, and they can be used in the development and testing of new algorithms to improve search performance and quality. Scientists all around the world would like to tap this gold mine for their own research search engine companies, however, do not release them because they contain sensitive information about their users, for example searches for diseases, lifestyle choices, personal tastes, and political affiliations. In this project to propose a novel approach to infer the user search goals by analyzing the search engine query logs. This approach to infer user search goals for a query by clustering our proposed user clicks. The User session is defined as the series of both clicked and unclicked URLs and ends with the last URL that was clicked in a session from user click-through logs.

In the early studies on personalized service, user's interest modeling techniques were not paid much attention to as what they are deserved. An amount of researches focused on personalized service to achieve the specific technology, such as the recommended technology, information retrieval, user clustering technology, but user modeling techniques are rarely mentioned. However, with the development and in depth study of personalized service, researchers gradually realize that the quality of personalized service not only depends on the specific recommendation technology, search technology, but also relies on user's preferences and other characteristics of interest, description of its computable, while the latter is particularly important. Therefore, in recent years, the user modeling techniques are separated from specific forms of personalization and serve as a basis technology research of personalized service Several researchers have presented their methods of building an implicit user interest model. In literature the user model was build according to the types of users with sample documents, through studying characteristics, types of paragraphs and the ability of classifying. Literature proposes a method based on multiple instances, which is combining more the user's information of interest to describe the user model together. A fine-grained client side user modeling method is proposed in literature.

In the last decade, many web personalization systems have been built based on different approaches. No matter what kind of approach they use, their data can be divided into two categories: usage data (the user's navigational behavior) and the user's profile data. Based on mining these data, the existing systems give the user a list of web pages that he or she might be interested in. None of them give the user a list of interested domains. The reason is interests extracting models of these systems only extract a list of web pages that the user is interested in, but don't extract a list of interested domains.

II. EXISTING SYSTEM

We define user search goals as the information on different aspects of a query that user groups want to obtain. Information need is a user's particular desire to

obtain information to satisfy his/her need. User search goals can be considered as the clusters of information needs for a query. The inference and analysis of user search goals can have a lot of advantages in improving search engine relevance and user experience.Sometimes queries may not exactly represent user specific information needs since many ambiguous queries may cover a broad topic.

- Different users may want to get information on different aspects when they submit the same query.
- What users care about varies a lot for different queries, finding suitable predefined search goal classes is very difficult and impractical.
- Analyzing the clicked URLs directly from user click-through logs to organize search results. However, this method has limitations since the number of different clicked URLs of a query may be small. Since user feedback is not considered, many noisy search results that are not clicked by any users may be analyzed as well. Therefore, this kind of methods cannot infer user search goals precisely.
- Only identifies whether a pair of queries belongs to the same goal or mission and does not care what the goal is in detail.

B. DISADVANTAGES

- In web search applications, queries are submitted to search engines to represent the information needs of users.
- However, sometimes queries may not exactly represent user specific information needs since many ambiguous queries may cover a broad topic and different users may want to get information on different aspects when they submit the same query.
- For example, when the query “the sun” is submitted to a search engine, some user wants to locate the homepage of a United Kingdom newspaper, while some others want to learn the natural knowledge of the sun.

III. PROPOSED SYSTEM

In this Project, we will firstly introduce the original Web Log Data and its corresponding pretreatment technologies. Secondly, we will describe algorithms for extracting user's Long Term Interests and Short Term Interests based on visit time and visit density which can be obtained from an analysis of RwCs (records with category) generated from Web Log Data. Since a user visits his or her favorite Web sites routinely, the Category which is correspondingly a long term visited and has most steady visit densities represents his or her Long Term Interest Category, while short term visited but several steady visit densities existing represents his or her Short Term Interests. In this project, we aim at discovering the number of diverse user search goals for a query and depicting each goal with some keywords automatically. We first propose a novel approach to infer user search goals for a query by clustering our proposed user sessions. Then, we propose a novel optimization method to map user sessions to pseudo-documents which can efficiently reflect user information needs. At last, we cluster these pseudo documents to infer user search goals and depict them with some keywords. Our approaches are unique and different from the existing studies from the following aspects:

(1) The algorithms are unique and novel, they are based on lasting time of the visit behaviors of a domain and the visit density to judge whether the domain (category) is an interest. This idea, in accordance with the logic, is simple and effective.

(2) It not only extracts a list of web pages the user interested in, but also mines a list of interested domains, including Long Term Interests and Short Term Interests.

(3) Pretreatment is very important for extracting. It uses web mining and text mining technologies to preprocess

the original Web Log data, laying a good foundation for extracting, and uses vector model of weighted keywords to express user's interest. The keywords are the domains (categories) of the information on the web pages which are acquired by classify technologies but not cluster.

To sum up, our work has three major contributions as follows:

- We propose a framework to infer different user search goals for a query by clustering user sessions. We demonstrate that clustering user sessions is more efficient than clustering search results or clicked URLs directly. Moreover, the distributions of different user search goals can be obtained conveniently after user sessions are clustered.
- We propose a novel optimization method to combine the enriched URLs in a user session to form a pseudo-document, which can effectively reflect the information need of a user. Thus, we can tell what the user search goals are in detail.
- We propose a new criterion Algorithm for User Interests to evaluate the performance of user search goal inference based on restructuring web search results. Thus, we can determine the number of user search goals for a query.
- User sessions can be considered as a process of resembling.
- User session is also a meaningful combination of several URLs.
- When users submit one of the queries, the search engine can return the results that are categorized into different groups according to user search goals online. Thus, users can find what they want conveniently

IV.SURVEY ON EXISTING IBS

User Sessions

The inferring user search goals for a particular query. Therefore, the single session containing only one query is introduced, which distinguishes from the conventional session. Meanwhile, the user session in this project is based on a single session, although it can be extended to the whole session. The proposed user session consists of both clicked and unclicked URLs and ends with the last URL that was clicked in a single session. It is motivated that before the last click, all the URLs have been scanned and evaluated by users. Therefore, besides the clicked URLs, the unclicked ones before the last click should be a part of the user sessions. We Process the Analysis through given procedure:

- Individual System Web Log User Interests Extracting.
- Multiple Systems or Online Web Log User Interests Extracting.

Original Web Log Data

The primary source of data for this study was the anonym zed logs of URLs visited by users who opted in to provide data through a widely-distributed browser toolbar. These log entries include a unique identifier for the user, a timestamp for each page view, a unique browser of single system or various systems through window identifier (to resolve ambiguities in determining which browser a page was viewed), and the URL of the Web page visited. Intranet and secure (https) URL visits were excluded at the source.

Expression of user's interests

In this paper we describe a systematic, log-based study of numerous contextual sources for modeling user interests during Web interaction. The core task for any user modeling system is predicting future behavior, and we evaluate the informativeness of different sources of contextual evidence based on their informativeness for predicting users' future interests at different temporal durations. We assume that the user has browsed to a Web page and the task is to leverage context to predict their future interests. The use of the current page and

five distinct sources of context are evaluated: (i) interaction: recent interaction behavior preceding the current page; (ii) collection: pages with hyperlinks to the current page; (iii) task: pages related to the current page by sharing the same search engine queries; (iv) historic: the longterm interests for the current user, and; (v) social: the combined interests of other users that also visit the current page. We developed user interest models based on and the five sources of contextual information used in our study. The sources were chosen based on elements of a nested model of context stratification proposed. The dimensions of that model represent the main contextual influences affecting users engaged in information behavior:

Collection context: The interest model for the collection context was created using Web pages containing hyperlinks that refer to . We obtained the set of in-links for each from the index of a large commercial Web search engine. An ODP label was assigned to each in-link, and in a similar way to other contexts, we created a ranked list of the labels based on their frequency.

Social context: The interest model for social context was created by combining the historic contexts of users that also visit . Note that this differs from the task context in that we focus on other users' long-term interests rather than only leveraging common querying behavior to find related URLs. From the browse trails in we found users who have also visited , and combined their interest models (historic contexts) to create a ranked list of ODP labels based on label frequency.

Long Term Interests Extracting

A Long Term Interest is a category which is visited for a long term (such as one year, it can be designated by client user) and most of the visited densities in the long term are correspondingly steady.

Historic context: The interest model for the historic context was created for each user based on their long-term interaction history. To create each user's historic context, we classified all Web pages they visited in , and created a ranked list of ODP labels based on label frequency. This list represents the interest model for the historic context for all visited by that user.

1) Definitions and Criteria: Some related criteria and definitions for Long Term Interest are introduced in this subsection.

a) Lasting time criterion (lastingTimemin): Lasting time criterion of a Long Term Interest Category. For example, if lasting time that the user visits a certain category is larger than lastingTimemin, the category is a Long Term Interest Category. This criterion is determined experimentally or it can be designated by client user.

b) Day interval (daygap): the time interval (three days, five days and so on) that is used in counting Density. It can be determined by client user.

c) Visit density (Density): the visiting frequency per day of a user visiting a category c. When the user's visit records of which the values of Category are c can be sorted in a time sequence

Short Term Interests Extracting

A Short Term Interest is a category which is visited for a correspondingly short term (such as one month, it can be designated by the client user) and existing several correspondingly high visited densities in the short term.

1) Definitions and Criteria: We will introduce some related criteria and definitions for a Short Term Interest in this subsection.

a) Lasting time (day) criterion (lastingTimemax): Lasting time criteria of a Short Term Interest Category. For example, if the lasting time of the user visiting a certain category less than lastingTimemax, the category is a Short Term Interest Category. This criterion may be determined experimentally or it can be designated by client user.

V. RELATED WORK

Current Internet includes billions of pages consist of drowned data information layout. Whether to convert existing sites or sites semantics for intelligent use embedded data, the definition of data mining techniques is of great interest. For that reason, the extraction of data from the Internet has been and continues to be the subject of much research. Related works can be grouped into two categories. The automatic extraction and rules handcrafted techniques. The main focus of automatic extraction techniques is inference through features extracted from HTML .Handcrafted rules is mostly used to extract information from HTML through string manipulation functions [2]. Godoy, Schiaffino, and Amandi [13] demonstrated that the use of Web Mining can be used to extract knowledge from observed actions. Crescenzi and al. [14], Baumgartner and al. [15], and Liu and al. [16] are based on the HTML markup generated automatically or semi Automatically extracting useful data modules. Each extraction module is used for extracting data of pages whose information content and structure are homogeneous. Adelberg [17] draw on the definition of a target structure for the data to be extracted. This structure is created by analyzing a sample document. According to this structure, an algorithm defines extraction rules based on delimiters (constant punctuation, text), and browsing other documents of the same type in order to extract the data in a format conforming to the target structure. Chung and al. [19] Propose a mixed method (HTML markup and ontologies) to integrate homogeneous HTML documents on the informational level but heterogeneous in terms of structure and presentation. Rules to restructure documents based on structural and visual information of HTML markup are used to transform the source XML documents. To give names to represent XML elements, the user defines a first set of concepts of application domain, and examples of instances (keyword) or models of instances for these concepts. These models and keywords are compared to textual information met during the restructuring. From XML documents, a DTD file describing common structures is derived. JIANG Chang-Bin Chen and Li [21] provide a log file preprocessing algorithm of Web data based on collaborative filtering. It can identify the user session fast and flexibly, even if the statistics are not sufficient and the historical records of visits of the user is absent.

VI. CONCLUSION

Web page content extraction is extremely useful in search engines, web page classification and clustering process, it is the basis of many other technologies about data mining, which aims to extract the worthiest information from dataintensive web pages full of noise. In the proposed method we extract required patterns by removing noise that is present in the web document using hand-crafted rules developed in Java. In future we plan to extend our work to the Web usage Mining. In the introduction, we marked the considerable character figures for the use of the Internet and the number of pages available. It can be considered in parallel need, what the website owners to understand their users. The existences of these factors has increased strongly the emergence of Web Usage Mining by applying knowledge extraction algorithms on large volumes of data on one side and use the results of an other side. However, the data contained in log files results in a lack of reflection on how to proceed. The step data mining itself deserves further work to be adapted to the needs of the analysis of log files.

REFERENCES

[1] Berkhin, P., Becher, J. D., and Randall, D. J., “Interactive Path Analysis of Web Site Traffic”, proceedings, Seventh International Conference on Knowledge Discovery and Data Mining (KDD01), 2001, pp.414-419.

[2] Z. Ma, G. Pant, and S. Liu, "Interest-based personalized search," ACM Trans. Inform. Syste., vol. 25, no. 1, article 5, 2007.

[3] Pazzani, M., Muramatsu J., and Billsus, D., "Syskill & Webert: Identifying interesting web sites", In the Proceedings of the National Conference on Artificial Intelligence, Portland, 1996.

[4] Pei, J., Han, J., Mortazavi-asl, B., and Zhu, H., "Mining Access Patterns Efficiently from Web Logs", Proceedings of PAKDD Conference, LNAI 1805, 2000, pp.396-407.

[5] Srivastava, J., Cooley, R., Deshpande, M., and Tan, P.-N., Web Usage Mining: "Discovery and Applications of Usage Patterns from Web Data", ACM SIGKDD Explorations, Vol.1, No.2, 2000, pp.12-23.

[6] Zhu, T., Greiner, R., and Haubl, G.: "Learning a model of a web user's interests". In: User Modeling (UM), 2003 pp.65-75.

[7] Minxiao Lei, and Lisa Fan., "A Web Personalization System Based on Users' Interested Domains", Proc. 7th IEEE Int. Conf. on Cognitive Informatics (ICCI'08), 2008.

[8] Murata, T., "Discovery of User Communities from Web Audience Measurement Data", Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence (WI2004), 2004, pp.673-676.

[9] T. Van and M. Beigbeder, "Hybrid method for personalized search in scientific digital libraries" Computational Linguistics and Intelligent Text Processing. Berlin, Germany: Springer, 2008, pp. 512-521.

[10] J. Cervantes, X.Li and W.Yu, "Support vector machine classification for large data sets via minimum enclosing ball clustering" Neurocomputing, 2008, pp.611-619.

[11] C. Ling, Q. Yang, J. Wang, and S. Zhang. "Decision trees with minimal costs", In Proc. of ICML04, 2004.

[12] G. Ou, Y.L. Murphey, and L. Feldkamp. "Multiclass pattern classification using neural networks". In Proceeding of the International conference on Pattern Recognition, 2004.

[13] S. Kotsiantis, and P. Pintelas, "Logiboost of simple Bayesian classifer," Informatica, 2005, pp. 53-59.

BIOGRAPHY

Author Details: T. Sumitra, Student of M.Tech, Bharath College of Engineering and Technology for Women, Kadapa, AP. Email: sumitratella@gmail.com

Guide Details: *Shaik Shasha Ali,* Assistant Professor, Dept. of CSE, Bharath College of Engineering and Technology for Women, Kadapa, AP