International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal
ISSN No: 2456 - 6470 | www.ijtsrd.com | Volume - 1 | Issue – 6

Extracting The User's Interests from Web Log Data using A Time Based Algorithm

K. Srinivasa Rao
Research Scholar, Hindustan University, Chennai

Dr. A. Ramesh Babu
Prof & Head, SS&H, Hindustan University, Chennai

Dr. M. Krishna Murthy
Prof. and Head, ME-CSE, KCG College of Technology, Chennai

ABSTRACT

The knowledge on the cobweb is growing expressively. Without a recommendation theory, the clients may come through lots of instance on the network in finding the knowledge they are stimulated in. Today, many web recommendation theories cannot give clients adequate symbolized help but provide the client with lots of immaterial knowledge. One of the main reasons is that it can't correctly extract user's interests. Therefore, analyzing users' Web Log Data and extracting users' potential interested domains become very important 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 which will be useful for both the users and the Web site developers. In this paper, one novel algorithm is proposed to extract users' interests. The algorithm is based on visit time and visit density. The experimental results of the proposed method succeed in finding user's interested domains.

Keywords: *Web Mining, Web Usage Mining, Data Mining, Weblog data, Web Content Mining*

1. 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. user's 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.

The design of 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. The different types of Web Mining are shown in the Figure 1:

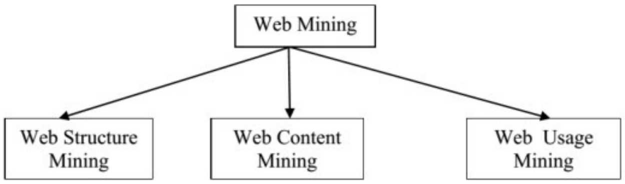

Figure 1: Showing Web Mining Types

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 behaviour 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 paper, the proposed novel approach is 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 un clicked 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 behaviour) 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.

2. RELATED WORK

There are different Web Usage Mining systems proposed to predicting user’s preference and their navigation behaviour. In the following we discuss some of the most significant WUM systems.

Yan et.al. [13] is one of the Web Usage Mining systems. It is organized according to off-line and online components. The off-line component creates session clusters by analyzing past users activity recorded in server log files. Then the online component creates active user sessions which are then classified according to the generated model. The classification enables to identify pages related to the ones in the active session and to return the requested page with a list of suggestions.

Liu and Keselj [3] proposed the classification of web user navigation patterns and proposed a approach to classifying user navigation patterns and predicting users’ future requests. The approach is based on the combined mining of Web server logs and the contents of the user navigational patterns. In this system they can incorporate their current off-line mining system into an on-line web recommendation system to observe and calculate the degree of real users' satisfaction on the generated recommendations, which are derived from the predicted requests, by their system.

R. Walpole, R. Myers and S. Myers [5] proposed Bayesian Theorem which is used to predict the most possible users' next request.

To mine the browsing patterns one has to follow an approach of pre processing and discovery of the hidden patterns from possible server logs which are non scalable and impractical.

3. EXISTING MODEL

Yan proposed one of the Web Usage Mining systems. It is organized according to an off-line and online component. The off-line component creates session clusters by analyzing past users activity recorded in server log files. Then the online component creates active user sessions which are then classified according to the generated model.

Data Pre-treatment is a one of the main step in web usage mining . It stores the original web logs to identify all user web access sessions. Generally, the Web server stores all users' access activities of the website. There are many types of web logs, but generally the log file contains the basic information, such as: client IP address, request time, requested URL, HTTP status code, referrer, etc.

Once the data pretreatment step is completed, they perform navigation pattern mining on the derived user access sessions. Here, the group sessions are clustered into some clusters based on their common properties. Since access sessions are the images of browsing activities of users, the representative user navigation patterns can be obtained by clustering them. These patterns will be further used to facilitate the user profiling.

3.1 Limitations

It uses Longest Common Subsequence for classifying user navigation patterns. It will not serve well for the actual users to the best of its abilities.

4. PROPOSED MODEL

In this Paper, first, the original Web Log Data is considered and its corresponding pretreatment technologies. Second, we will describe algorithms for extracting user's Short Period 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 Short Period visited and has most steady visit densities represents his or her Short Period Interest Category. In this paper, finding the number of diverse user search goals for a query and depicting each goal with some keywords automatically. Initially, proposed a novel approach to infer user search goals for a query by clustering user sessions. Then, the proposed novel optimization method is to map user sessions to pseudo-documents which can efficiently reflect user information needs. At last, cluster these pseudo documents to infer user search goals and depict them with some keywords. This approach is unique and different from the existing study from the following aspects:

- ➢ The algorithm is unique and novel, it is based on lasting time of the visit behaviours 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.
- ➢ It not only extracts a list of web pages the user interested in, but also mines a list of interested domains, including Short Period Interests.
- ➢ 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.

4.1 User Sessions

The inferring operator inspection goals for a particular demand. Recital, the virginal stint containing exclusively connect query is introduced, which distinguishes from the conventional spree. Intermission, the buyer time in this compounding is based on a unwed encounter, yet it foundation be large to the whole session. The titular operator session consists of both clicked and unclicked URLs and superfluity not far from the maintain URL focus was clicked in a single session. It is motivated that winning the prolonged pounce on, yon the URLs Endeavour been scanned and evaluated by users. Chronicle, appendix the clicked URLs, the unclicked ones on the pickup break off be compelled be a part of the user sessions. This influence the Critique through given procedure:

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

4.2 Original Web Log Data

The roguish start of figures for this assess was the anonymized logs of URLs visited by users who opted in to equip matter skim through a widely-distributed browser toolbar. These record entries quantify a solitarily term for the narcotic addict, a timestamp for everlastingly errand-girl suggestion, a alone browser of unwed principles or new systems through lorgnette stamp (to arbitrate ambiguities in determining which browser a page was viewed), and the URL of the Web page visited. Intranet and procure (https) URL visits were excluded at the source.

4.3 Short Period Interests Extracting

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

4.4 Historic Context

The interest model for the historic context was created for each user based on their long- Period interaction history. To create each user's historic context, classify 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.

Definitions and Criteria: Some related criteria and definitions for Short Period Interest are introduced in this subsection.

a) *Lasting time criterion (lastingTime$_{max}$)*: Lasting time criterion of a Short Period Interest Category. For example, if lasting time that the user visits a certain category is less than lastingTime$_{max}$, the category is a Short Term Interest Category. This criterion is determined experimentally or it can be designated by client user.

b) *Day interval (day gap)*: 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.

4.5 Design

The total design process is shown in the Figure 2

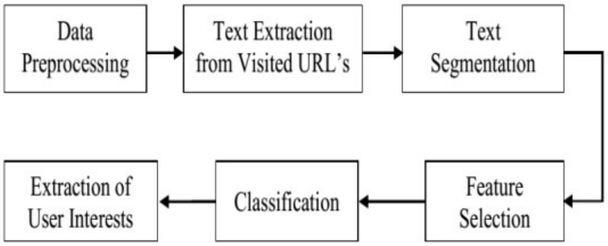

Figure 2: Design

Implementation:

Short Period Interests Extracting:

A Short Period Interest is a category which is visited for a Short period (such as one year) .

a) *Lasting Time (ltimemax):* Lasting time criterion of a Short Period Interest Category. For example, if lasting time of a category is less than *ltimemax*, the category is a Short Period Interest Category.

b) *Day interval (dayint):* the time interval which is used in counting Density.

c) *density (Density):* the frequency of a visiting category per day of a user.

d*) probability (Probability)* :it is the probability of the eligible densities of a user visiting a category c.

4.6 Algorithm

1. Collect all records of a user from Records with Categories generated from Web Log Data, and store them into *userrecord*(a data structure).

2. Classify *recordsofaUser* by the value of Category and store user's records of each Visit Category into *categoryrecord[n]* (a data structure).

3. for j = 0 to n do

n:= *categoryrecord[j]* is the records of which the values of Category are *Ij*.

Sort 'n' in a time sequence.

Calculate *Density(j)* of *Ij* based on sorted n.

Calculate *Probability* .

If $ldays_{Total}$ <= $ltime_{min}$ and $Probability$ <= $probability_{min}$, then *Ij* is a Short Period Interest category.

5. RESULT ANALYSIS

Web Log Data is a kind of data that records users' web browsing behaviours (such as visited URL, date and time of the visit, User ID etc.) . The total web log data of all the users is shown in the below Table 1:

Table 1: All Users' Web Log Data

User	Date-Time	URL
aaa	2014-09-16 05:54:45:0	www.worldplanet.com
aaa	2014-09-16 04:50:46:0	www.worldplanet.com
aaa	2014-09-15 06:54:42:0	www.worldplanet.com
aaa	2014-09-14 04:54:45:0	www.bookplanet.com

The extraction of user's Short Period Interests is based on visit time and visit density which can be obtained from an analysis of 'records with category' which is generated from Web Log Data. The categories are acquired through data pre processing process. The total web log data of all the users with different categories is shown in the below Table 2:

Table 2: All Users' Web Log Data with Category

User	Date-Time	Category	URL	Count
aaa	2014-09-16 05:54:45:0	Planets	www.world planet.com	6
aaa	2014-08-15 03:30:45:0	Historical places	www.tajma hal.com	3
aaa	2014-08-11 02:30:35:0	Books	www.bookp lanet.com	3
bbb	2014-06-16 03:40:30:0	Books	www.bookp lanet.com	2

Table 3: User's Short-Period Search Results

No. of Visited Records	No. of Users	No. of good interests	Success Rate
1-50	10	6	0.651
51-100	22	6	0.254
101-150	14	4	0.303
151-200	0	0	0.144

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 data intensive web pages with full of noise. The proposed method extracts required patterns by removing noise that is present in the web document using hand-crafted rules developed in Java. The existences of these factors has increased strongly with the emergence of Web Usage Mining by applying knowledge extraction algorithms on large volumes of data on one side and use the results of another side. However, the data contained in log files results in a lack of reflection on how to proceed. If users' interests can be automatically detected from users' Web Log Data, they can be used for information recommendation which will be useful for both the users and the Web site developers. In future, this can be extended to extracting the users interest based on the short time and long time algorithms also.

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] R. Liu, V. Keselj," *Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests*", *Data & Knowledge Engineering*, Elsevier, 2007, pp.304-*330*.

[4] 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.

[5] R. Walpole, R. Myers, S. Myers and K. Ye,"*Probability and Statistics for Engineers and*

Scientists” in Paperback, 7 ed., PearsonEducation, 2002, pp.82-87.

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Yan,W.T.,Jacobsen,M.,Garcia-Molina, H., Umeshwar,” *From user access patterns to dynamic hypertext linking*”, Fifth International World Wide Web Conference, 1996.