MANTECH Publications

Journal of Data Mining and Knowledge Engineering
Volume 1, Issue 1, January-June 2016

Long-Period Extracting Algorithm to Mine User Interests from Web Log Data

K. Srinivasa Rao[1], M. Krishna Murthy[2]

[1]KSRM College of Engineering, Kadapa, Andhra Pradesh, India [2]

ME-CSE, KCG College of Technology, Chennai, Tamil Nadu, India

Email ID. *srinu532@gmail.com, mkrish@kcgcollege.com*

Abstract

As the knowledge available on the World Wide Web is increasing every day, many users find it difficult to get the data they need as they don't use any recommendation system to scan the web. A recommendation system can significantly reduce the time taken for finding the data on the web. However, even some recommendation systems don't give the precise information users need as these systems don't take into account the user's interests. As analyzing user interests and identifying the right domain is essential to in Web Usage Mining. Detecting user's interests is possibly by scanning their Web Log Data, and the information obtained from there will be useful for website developers and users alike. In this paper, a unique algorithm is presented that can extract users' interests accurately as it is based on visit density and time. This method is very useful in finding a user's interested domains.

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

INTRODUCTION

Web mining is the process of extracting specific data or information patterns from the Web. It is divided into three different types: 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. It involves a process of finding out what users are looking for on the internet. Some users might be looking at only textual data, whereas others might be interested in multimedia data. Usage mining is basically concentrated on the use

Page 15-25 © MANTECH PUBLICATIONS 2016. All Rights Reserved

of the web technologies that help in scanning data that enable a better understanding of user behavior. Web Structure Mining is the process of extracting the interested patterns on the Web from the available hyperlinks. Based on the web structure data, Web Structure Mining is divided into two types:

1. It is the process of extracting the patterns on the Web from the hyperlinks.

2. It is the process of analyzing the tree structure of page structure.

Web Content Mining is the process of extracting useful information or knowledge from the Web based on the content given by the user. Structured Data Extraction is the process of getting the formatted output data from the Web page. It also creates a database using this formatted output data. Information integration is the process of integrating the information from different websites.

The design of our group analysis is all about publishing search logs with privacy related web mining. Search engine companies collect the database of intentions and the histories of their user's search queries. These search logs are a gold mine for researchers.

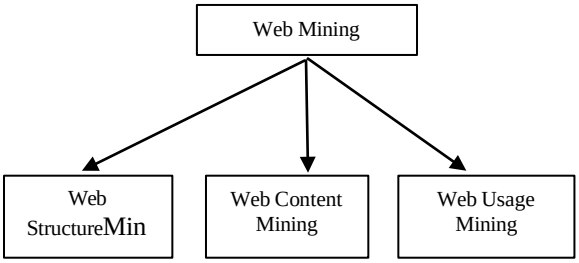


Fig. 1: *Showing Web Mining Types.*

Search engines play an essential role in the direction-finding through the immensity of the Web. Today's search engines are not just gathering the web pages and guides of web pages; they also gather user's historical data and extract useful information about the users. They stock up the queries given by the users, their clicks, corresponding IP-address, and other information about the communications with users. This is known as search log. A Search log contains valuable information about the users. They facilitate the discovery of trend pattern and anomaly in the search activities of users and these findings are used in the development and

Page 15-25 © MANTECH PUBLICATIONS 2016. All Rights Reserved

test of different new algorithms to improve search quality and presentation. Scientists from all the countries of the World are doing their research work on search engines. But, they do not make their researches public because they contain valuable information about the users.

In this paper, the proposed unique 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 involves clustering the 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 research on personalized web search, user's interest model technique was not given the importance it deserves. Most researches were done on personalized web search was to aid in development of new technologies such as recommendation system, retrieving the information from the Web and extracting the user's interests based on the user's past history. But, with the gradual development and detailed analysis of personalized web search, the researchers or scientists found that the quality of the personalized web search is not only dependent on the recommendation systems, but also on the user's preference and interests. Therefore, the user model technique is separated from specific forms of personalization. So, it became a basic technology research topic of personalized web search. Different researchers have given different methods for construction of user interest model. This user interest model was constructed according to the types of users with sample documents through study of characteristics. Our work discusses the classification of web user navigation patterns and proposes an 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.

Earlier, many personalization systems have been constructed based on different methods. In all the methods we use for our work, the user data can be divided into two types: usage data and user profile data. Usage data means the user navigational activities while user profile data means the information about the user. By extracting this type of data, the existing models can give the user a set of web pages that he is interested. None of the existing models give the user a list of domains the user is interested in. Extracting of data under

Page 15-25 © MANTECH PUBLICATIONS 2016. All Rights Reserved

these models gets only a list of web pages that the user is interested. It does not extract the interested domains.

RELATED WORK

Different Web Usage Mining (WUM) systems are proposed to predict a user's preference and their navigation behavior. Here, we discuss some of the most significant WUM systems.

Yan *et al.* is one of the Web Usage Mining systems [1]. It is organized according to two components: an off-line and an online. 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 identification of pages related to the ones in the active session and to return the requested page with a list of suggestions.

Liu and Keselj proposed the classification of web user navigation patterns and proposed an approach to classifying user navigation patterns and predicting users' future requests [2]. 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 online 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 proposed Bayesian Theorem which is used to predict the users' most probable next request [3].

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. This is impractical.

EXISTING MODEL

Yan proposed one of the Web Usage Mining systems. It is organized according to one each off-line and an 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.

Data Pretreatment is a one of the main steps in web usage mining. It stores the original web logs to identify all user web

Page 15-25 © MANTECH PUBLICATIONS 2016. All Rights Reserved

access sessions. Generally, Web server stores all the users' access data pertaining to different websites. There are many types of web logs, but generally these log files contain 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 do navigation pattern mining on the derived user access sessions. Here, the group sessions are clustered 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 user profiling.

LIMITATIONS

It uses Longest Common Subsequence for classifying user navigation patterns. This will not serve the actual users well as some other methods are able to.

PROPOSED MODEL

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

1) The algorithm is new and unique, as it is based on lasting time of the visit behaviors in a domain and its visit density to judge whether the domain (category) is an interest. This idea is much more logical as it is simple and effective.
2) It not only extracts a list of web pages the user is interested in, but also mines

Page 15-25 © MANTECH PUBLICATIONS 2016. All Rights Reserved

a list of interested domains, including Long Period 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 information on the web pages which are acquired by classify technologies.

USER SESSIONS

The inferring operator inspects goals for a particular demand. Recital, the virginal stint containing exclusively connect query is introduced, which distinguishes it from the conventional spree. Intermission, the buyer time in this compounding is based on an unwed encounter, yet it is the foundation of the whole session. The titular operator session consists of both clicked and unclicked URLs and superfluity was not far from the maintained URL focus as it was clicked in a single session. Clicked URLs and the unclicked ones on the pickup break off and become a part of the user sessions. This influences the Critique through the given procedure:

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

ORIGINAL WEB LOG DATA

The start of this review was initiated by the logs of URLs visited by the users. These documentation entries calculate a solitary term for the tranquilizer devotee, a timestamp for eternally errand-girl suggestion. The original web log data contains information about the previous history of the user, like the URL of the visited web page, the time of visiting web page, the date of visiting web page and how many times the web page was visited, etc.

LONG PERIOD INTERESTS EXTRACTING

A Long Period 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- period interaction history. To create each user’s historic context, all Web pages they visited are classified, and a ranked list of ODP labels is created based on label frequency. This list represents the interest

Page 15-25 © MANTECH PUBLICATIONS 2016. All Rights Reserved

model for the historic context for all the visited web pages by that user.

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

Lasting Time Criterion (Lasting Time$_{min}$): Lasting time criterion of a Long Period Interest Category need to be evaluated. For example, if lasting time that the user visits a certain category is larger than Lasting Time$_{min}$, the category is a Long Term Interest Category. This criterion is determined experimentally or it can be designated by client user.

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.

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

DESIGN

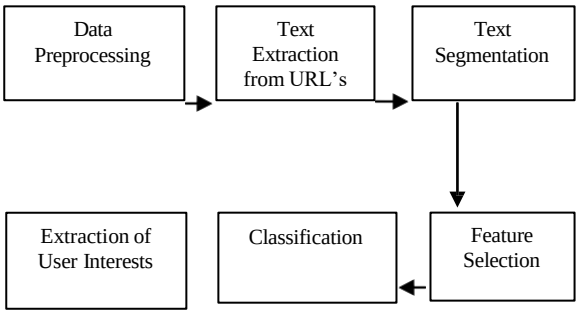


Fig. 2: *Design.*

IMPLEMENTATION

Long Period Interests Extracting

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

Lasting Time (ltimemin): Lasting time criterion of a Long Period Interest Category need to be measured. For example, if lasting time of a category is larger than *ltimemin*, the category is a Long Period Interest Category.

Page 15-25 © MANTECH PUBLICATIONS 2016. All Rights Reserved

Day Interval (Dayint): The time interval which is used in counting density.

Density: The frequency of a visiting category per day of a user.

Probability: It is *the* probability of the eligible densities of a user visiting a category c.

ALGORITHM

1. Collect all records of a user from Records with Categories specified that were generated from Web Log Data, and store them into *user record* (a data structure).
2. Classify records of a user 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 the sorted *N*.

   Calculate *Probability.*

   If $ldays_{Total} >= ltime_{min}$ and *Probability* >= $probability_{min}$, then *Ij* is a Long Period Interest Category.

RESULT ANALYSIS

Web Log Data is a kind of data that records users' web browsing behaviors (such as visited URL, date and time of the visit, User ID etc.)

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

Fig. 3: *All Users' Web Log Data.*

Page 15-25 © MANTECH PUBLICATIONS 2016. All Rights Reserved

The extraction of user's Long 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 preprocessing process.

User	Date-Time	Category	URL	Count
aaa	2014-09-16	Planets	www.worldplanet.com	6
aaa	2014-08-15	Historical	www.tajmahal.com	3
aaa	2014-08-11	Books	www.bookplanet.com	3
bbb	2014-06-16	Books	www.bookplanet.com	2

Fig. 4: *All Users' Web Log Data with Category.*

No. of Visits	No. of Users	No. of good interests	Success Rate
1–50	6	6	0.851
51–100	27	8	0.654
101–150	21	9	0.603
151–200	26	6	0.544

Fig. 5: *User's Long-Period Search Results.*

Page 15-25 © MANTECH PUBLICATIONS 2016. All Rights Reserved

CONCLUSION

Web page content extraction is extremely useful for search engines that rely on web page classification and clustering process. It is the basis of many other technologies used in data mining, which aim to extract the most useful information from data intensive web pages. The proposed method extracts required patterns by removing noise that is present in the web document using hand-crafted rules developed in Java. The presence of these factors has increased strongly with the emergence of Web Usage Mining by applyingknowledge extraction algorithms on large volumes of data on one hand, and use of extracted results on the other. However, the data contained in log files does not reflect 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 website developers.

REFERENCES

1. Yan W.T.,Jacobsen M.,Garcia- Molina, H.Umeshwar," *From user access patterns to dynamic hypertext linking* ", Fifth International WWW Conference, 1996.

2. 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; 304– 330p.

3. R. Walpole, R. Myers, S. Myers and K. Ye. *Probability and Statistics for Engineers and Scientists*" in Paperback, 7 ed., Pearson Education, 2002; 82–87p.

   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; 396–407Srivastava 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; 12–23p. Zhu T., Greiner R., and Haubl G.: "*Learning a model of a web user's interests*". In: User Modeling (UM), 2003; 65–75p.

Page 15-25 © MANTECH PUBLICATIONS 2016. All Rights Reserved

5. 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.

6. 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; 673–676p.

7. T. Van and M. Beigbeder, "*Hybrid method for personalized search in scientific digital libraries*" Computational Linguistics and Intelligent Text Processing. Berlin, Germany: Springer, 2008; 512–521p.

8. J. Cervantes, X.Li and W.Yu, "*Support vector machine classification for large data sets via minimum enclosing ball clustering*" Neurocomputing, 2008; 611–619p.

9. Berkhin P., Becher J. D., Randall D. J. "*Interactive Path Analysis of Web Site Traffic*", proceedings, Seventh International Conference on Knowledge Discovery and Data Mining (KDD01), 2001; 414–419p .

Page 15-25 © MANTECH PUBLICATIONS 2016. All Rights Reserved