International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:9, No:2, 2015

A Novel Approach to Improve Users Search Goal in Web Usage Mining

R. Lokeshkumar, P. Sengottuvelan

***Abstract*—**Web mining is to discover and extract useful Information. Different users may have different search goals when they search by giving queries and submitting it to a search engine. The inference and analysis of user search goals can be very useful for providing an experience result for a user search query. In this project, we propose a novel approach to infer user search goals by analyzing search web logs. First, we propose a novel approach to infer user search goals by analyzing search engine query logs, the feedback sessions are constructed from user click-through logs and it efficiently reflect the information needed for users. Second we propose a preprocessing technique to clean the unnecessary data's from web log file (feedback session). Third we propose a technique to generate pseudo-documents to representation of feedback sessions for clustering. Finally we implement k-medoids clustering algorithm to discover different user search goals and to provide a more optimal result for a search query based on feedback sessions for the user.

***Keywords*—**Data Preprocessing, Session Identification, Web log mining, Web Personalization.

I. Introduction

THE WW as a huge and dynamic information source is a Productive ground for data mining principles or Web Mining. Data mining is the study of data-driven techniques to discover patterns in large volumes of raw data. Web mining can be referred as the transformation of the data mining techniques to web data. Web mining is primarily aimed at deriving actionable knowledge from the Web through the application of various data mining techniques [1]-[17]. Web mining has three distinct phases involved – content, structure and usage mining of web data. Mining the content involves extracting the relevant information, structure mining studies the structure and prototype and usage mining is the analysis of the discovered patterns. Web usage mining (WUM) is all about identifying user browsing patterns over WWW, with the aid of knowledge acquired from web logs. The outcomes of the WUM can be used in web personalization, improving the performance of the system, modification of the site, business intelligence, usage characterization etc. Web Usage Mining is the discovery of user access patterns from Web server access logs [4].

A. Web Mining

Web mining is the application of data mining techniques to extract knowledge from Web data including Web documents, hyperlinks between one document to another document and usage logs of web sites [2]. A common taxonomy of web mining defines three main research lines:

1. Web content mining (WCM)
2. Web structure mining (WSM)
3. Web usage mining (WUM)

B. Web Content Mining

Web content mining (WCM) is the process to discover useful information from the content of a web page [15]. Basically, the Web content consists (WCM) of several types of data such as textual, image, audio, video, metadata as well as hyperlinks.

C. Web Structure Mining

Web Structure Mining (WSM) is the process of inferring knowledge from the World Wide Web organization and links between references and referents in the Web. The structure of a typical web graph consists of web pages as nodes and hyperlinks as edges connecting related pages [6]. Web Structure mining (WUM) is the process of using graph theory to analyze the node and connection structure of a web site.

D. Web Usage Mining

Web Usage mining (WUM) is the application of data mining techniques to discover usage patterns from web data. Data is usually collected from web log, e.g. web/proxy server logs, user queries, registration data. Usage mining tools discover and predict user behavior, in order to help the designer to improve the web site, to attract visitors, or to give regular users a personalized and adaptive service. In this paper we use the Web usage mining (WUM) data for the Discovery of meaningful patterns from data generated by client-server transactions for one or more Web localities. Analyzing and exploring regularities in Weblog records (consist of URL's, click sequence and etc.) for electronic commerce, enhance the quality and delivery of internet information services to the end user, and improve Web server system performance. Web mining studies analyzes and reveals useful information from the Web [5]. Web mining is a term used for applying data mining techniques to Web access logs.

E. Web Log Records

Based on the Weblog records, we have to construct the feedback session. Because Weblog data provide information about what kind of users will access what kind of Web pages. This session consists of URL's and click sequence and it focus on user search goals [16]. Only using a feedback session we do not understand the user search goals exactly.

R.Lokeshkumar and P.Sengottuvelan are with Department of Information Technology, Bannari Amman Institute of Technology, Sathyamangalam, Erode District Tamil Nadu, India. (Phone: 04295 - 226216; e-mail: rlokeshkumar@yahoo.com, sengottuvelan@rediffmail.com).

II. Related Work

A Log Data Preprocessor for mining Web browsing patterns [3], author develops a tool LODAP for the preprocessing of web log file. The developed Tool provides report summary at each step. It has there preprocessing steps.

Step1. In Data cleaning was performed based on access method, status code, multimedia objects, and request generated through robots.

Step2. In Data structuration, user sessions were identified based on IP Address, maximum timeout and set of resources within corresponding access time as given in (1):

$$S(i) = u(i), t(i), r(i) \quad (1)$$

Step3. In Data filtering, least requested pages were removed after defining threshold value.

For each resource *r*, total numbers of sessions *NSi*were calculated. Moreover, user defined threshold value ε, removes the each request $NSi < \varepsilon$. LODAP tool supports the preprocessing steps along with reports facility at each level of preprocessing in the form of data summarization.

We come to the conclusion that it is a good effort by authors to support their work by proper MS Access based tool. In summing up it can be said that if authors were able to perform some sort of classification through LODAP, it would be an effective tool for preprocessing of Web usage mining. Preprocessing step is complex and laborious task

In this paper authors proposed two algorithms, first algorithm is to read the log file from any of the three given log file formats and convert the log file data into a database. Second algorithm is to filter out the all the un-interested attributes of web log file [7], [8].

Only *"URL"* attributes was declared interested. *Date, Time, IP Address,* and *User Agent* are some other useful attributes were also dropped. By dropping out such important attributes, the reliability of later phases of Web usage mining cannot be secured. We come to the conclusion that proposed algorithms for data cleaning and data filtering techniques are very weak and needs to be modified.

There are three main types of web server log file formats to capture the activities of user on web site. Log file act as health monitor for the web sites and are the main source of user access data and user feedback [11].

Web log file is a simple plain text file which records information about each user, display of log files data in three different forms namely, web server log formats namely Common log file format (NCSA), Extended log format (W3C), IIS log format.

W3C log format is default log file format on ASCII server. Field are separated by space, time is recorded as GMT (Green which Mean Time) [14]. It can be customized that is administrator can add or remove fields depending on the web server manager.

NCSA (National Centre for Supercomputing Application Format) is to record basic information about the user request such as user name, record host name, date, time, request type, HTTP status code and number of bytes send by server showed in Fig. 2. NCSA is a fixed format, it cannot be changed.

```
#software: Microsoft Internet Information Server 7.5
#Version:1.0
#Date:2014-01-22 09:55:33
#Fields: date time cs-method cs-uri-stem c-ipsc-version sc-
status example:
2014-01-13 5:34:11 GET/website/::1HTTP/1.1301
```

Fig. 1 Example for W3C log file format

```
::1—[20/feb/2014:11:02:44+05300
"GET/website/HTTP/1.1"200 1107
```

Fig. 2 Example for NCSA log file format

IIS log file format IIS format is not customized, it is fixed ASCII format. Fields are separated by comma, easy to read. Time recorded in local time, it records more information than NCSA format. IIS consist of Attributes which cannot be modified.

```
::1, -, 1/25/2012, 9:57:42, W3SVC1, JAY-PC, ::1, 3,
965, 1153, 200, 0, POST, /WebSite/default.aspx, -,
```

Fig. 3 Example for ASCII log file format

Data clustering is the process of dividing data elements into classes or clusters so that items in the same class are as similar as possible, and items in different classes are as dissimilar as possible [9]. Clustering are of different types, which is used depends upon the needs [12]. Main three methods of clustering are

i. Partitioning Methods
ii. Hierarchical Methods
iii. Density-based Methods

Depending on the nature of the data and the purpose for which clustering is being used, different measures of similarity may be used to place items into classes. After constructing the pseudo-document the Web search results are restructured based on the documents collection detail, and that pseudo-document is converted into groups using clustering algorithm.

III. Proposed System

In this paper, we aim at discovering the number of diverse user search goals for a query and depicting each goal with some keywords automatically. First we propose of preprocessing technique to clean the feedback session from web log file to extract data without any error or unwanted data. Second, we propose a novel optimization method to map feedback 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 using K-medoids algorithm. Modules are

A. Capturing Feedback Sessions,
B. Preprocessing Feedback Session,
C. Building Pseudo-Documents,

D. Clustering Pseudo-Documents.

A. Capturing Feedback Sessions

The feedback 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 [13]. A session for web search is a series of successive queries to satisfy a single information need and some clicked search results.

For inferring user search goals it is more efficient to analyze the feedback sessions than to analyze search results or clicked URL's directly because there are different feedback sessions in user click-through logs.

In our paper, we focus on inferring user search goals for a particular query. The representation of feedback session is based on number of clicks. In this project the feedback session is based on a single session, although it can be extended to the whole session.

B. Per-processing Web User Log

Preprocessing is a process of cleaning the unwanted and erroneous data. In this project we used a preprocessing technique applied on feedback session to cleaning and to identify the users search goals based on users clicks. Preprocessing involves four steps they are,

i. Data fusion
ii. Data cleaning
iii. Data summarization
iv. Data structuration.

C. Building Pseudo-Documents

We used pseudo-documents to infer user search goals. Users have some unclear words for representing the interests. They use these keywords to determine whether a document can satisfy their needs. These keywords are known as "goal texts". Goal texts can reflect user information needs; they are hidden and not expressed explicitly. So, pseudo-documents are used as surrogates to approximate goal texts.

Building pseudo-documents includes two steps

i. Representing the URL's in feedback session.
ii. Forming pseudo-documents based on URL representations

D. Clustering

Pseudo-documents are clustered by using k-medoids clustering algorithm [10]. An optimal value will be determined through the evaluation. After clustering of all pseudo-documents, each cluster is considered as one key word search goal of a user.

IV. Implementation And Result

We captured a data set from web log file using four systems IP address collected over a period of two weeks. Mouse is a key word which we chosen for our experiment.

A. Dataset Collection and Preprocessing

The dataset we have used is based on the users URL clicks. Dataset consist on 3000 URL's in which, we applied a preprocessing technique to remove noisy data's. Noisy data refers to error pages on web.

In those 3000 URL's after applying preprocessing all other URL's result except mouse has been removed from the dataset. At last we had 500 datasets for a single Query.

B. Clustering Algorithm

We applied a K-Medoids clustering algorithm to for clusters based on three criteria they are Mickey Mouse, electrical mouse and animal mouse. Steps involved in applying K-medoids Clustering algorithm are

Input
K: the number of clusters
D: a data set containing n objects
Output: A set of k clusters Method:
1) Arbitrary choose k objects from D as representative Objects (seeds)
2) Repeat
3) Assign each remaining object to the cluster with the Nearest representative object
4) For each representative object Oj
5) Randomly select a non-representative object Orandom
6) Compute the total cost S of swapping representative object Oj with Orandom
7) if S<0 then replace Oj with Orandom
8) Until no change

C. Screenshots

The experimental result of our project is explained with screen shots. We have done our experiment for queries like computer mouse, animal mouse and mickey mouse. But we have shown our screen shots results only for computer mouse, in the same way we have implemented for animal mouse and computer mouse too.

D. Screen Shot Description

Here we have given an explanation for each and every screen shots. Search box page of our project is shown in Fig. 4 in which a query input is given and clicks search. It will display dataset of that particular query before preprocessing is shown in Fig. 4.

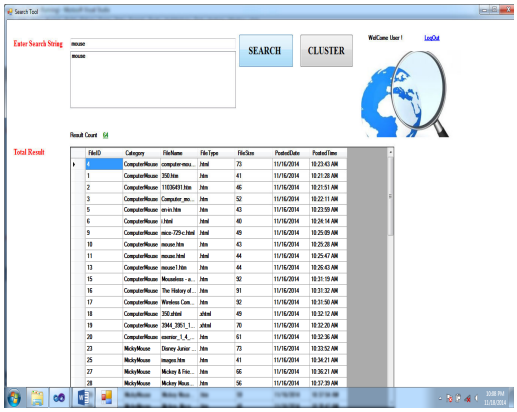

Fig. 4 Search box

Number of cluster formed with the dataset is shown in Fig. 5.

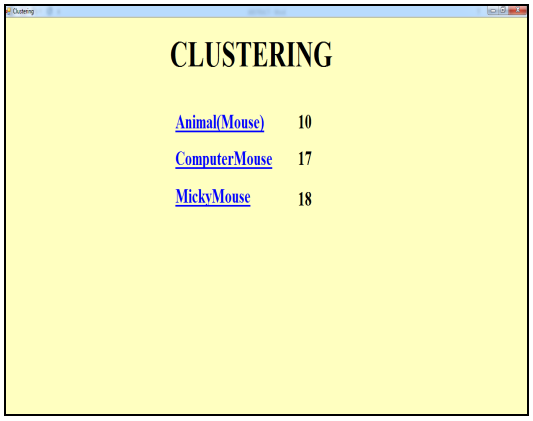

Fig. 5 Clustered result for Computer mouse

Clustering result for the query entered by the user with all attribute details are shown in Fig. 6.

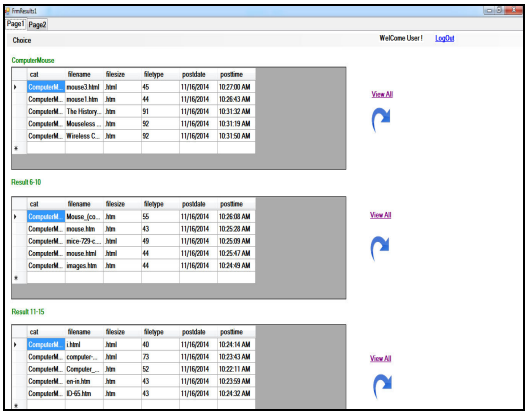

Fig. 6 Clustered result for Computer mouse

Result obtained for the search keyword after preprocessing is shown in Fig. 7.

Fig. 7 Links obtained for Computer mouse

Search result for Computer mouse which is clicked by the user or a particular cluster is shown in Fig. 8

Fig. 8 Search Result for Computer mouse

V. Conclusion

As the Web and its usage continues to grow, the opportunity to analyze Web data and extract all useful knowledge from it because very tedious. In the past five years we have seen the emergence of Web mining as a rapidly growing area, due to the efforts of the research community as well as various organizations that are practicing it to bring more thrust areas. In this project, we proposed a technique for user search goals using feedback session and pseudo document. First we captured a feedback session to analysis the user search goal from the Weblog record and preprocessing technique is applied on feedback session to clean the noise data. It cannot provide the accurate result by directly creating clusters. So we introduce a pseudo document to provide the accurate results. Based on the pseudo document we restructured the Web search results. Pseudo-document is the initial step of clustering. We also used K-medoids clustering Algorithm to cluster the users search goals. At last the web users receive an expected and clustered result for a searched query.

VI. Future Enhancements

In future we can implement the project for more number of keywords and use some improved clustering algorithm which creates a number of clusters themselves (automatically) based on data set. We can use a criterion “Classified Average Precision (CAP)” to evaluate the performance of inferring user search goals. Experimental results can be presented using user click-through logs from a commercial search engine to validate the effectiveness of our proposed methods.

TABLE I
Comparison Table

Method I	Displays first 100 results
Method II	Displays recently viewed Url’s
Method III	Displays based on feedback Session through clicked url using binary number.
Our Method	Displays based on feedback session through clicked url using number of clicks.

REFERENCES

[1] Weiyao Lin, Member, IEEE, and Zhaohui Zheng, "A New Algorithm for Inferring User Search Goals with Feedback Sessions." IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 3, March 2013.

[2] B. Uma Maheswari, Dr. P.Sumathi,"A New Clustering and Preprocessing for Web Log Mining", World Congress on Computing and Communication Technologies 2014.

[3] Castellano, G., A. M. Fanelli, "A Log Data Preprocessor for mining Web browsing patterns". Proceedings of the 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Corfu Island, Greece, February 16-19 , 2007.

[4] Khasawneh, N. and Chan, "Active User-Based and Ontology-Based Web Log Data Preprocessing for Web Usage Mining". Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings) (WI'06) 0-7695-2747- 7/06© 2006.

[5] Li. X, Y.-Y Wang, and A. Acero, "Learning Query Intent from Regularized Click Graphs," Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '08), pp. 339-346, 2008.

[6] Murata, T. and K. Saito. "Extracting Users' Interests from Web Log Data", Proceedings of IEEE/WIC/ACM Conference on Web Intelligence (WI 2006 Main Conference Proceedings) (WI'06) 0-7695-2747-7/06, 2006.

[7] Nicolas Labroche, "Learning Web Users Profiles with Relational Clustering Algorithms" In IADIS International WWW/Internet Conference, 503–510, 2010.

[8] Noor Kamal Kaur, Usvir Kaur, Dr. Dheerendra Singh, "K-Medoid Clustering Algorithm- A Review", International Journal of Computer Application and Technology (IJCAT) Volume 1 Issue 1 ISSN: 2349-1841, April 2014.

[9] O. A. Mohamed Jafar, R. Sivakumar, "A Study on Possibilistic and Fuzzy Possibilistic C-Means Clustering Algorithms for Data Clustering" - International Conference on Emerging Trends in Science, Engineering and Technology2012.

[10] Pabarskaite, Z, "Implementing Advanced Cleaning and End - User Interpretability Technologies in Web Log Mining". 24th Int. Conf. information Technology Interfaces/ TI, Cavtat, Croatia, June 24-27, 2002.

[11] Raghavi Chouhan, Abhishek Chauhan "An Ameliorated Partitioning Clustering Algorithm for Large Data Sets", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 7, July 2014.

[12] Rupinder Kaur, Simarjeet Kaur, "A Review: Techniques for Clustering of Web Usage Mining", International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064, 2012.

[13] Suneetha, K. R. and D. R. Krishnamoorthi, "Identifying User Behaviour by Analyzing Web Server Access Log File." Published in IJCSNS International Journal of Computer Science and Network Security, vol.9 No.4, April 2009.

[14] Uichin Lee, Zhenyu Liu, Junghoo Cho "Automatic Identification of User Goals in Web Search" In Proceedings of VLDB '04, 2004.

[15] Wahab, M. H. A., M. N. H. Mohd, et al, "Data Preprocessing on Web Server Logs for Generalized Association Rules Mining Algorithm". World Academy of Science, Engineering and Technology 48 2008.

[16] Xuanhui Wang, Cheng Xiang Zhai, "Learn from Web Search Logs to Organize Search Results" Journal of Graph Algorithms and Applications, 2010.

[17] Yuan, F., L.-J. Wang, et al. Study on "Data Preprocessing Algorithm in Web Log Mining". Proceedings of the Second International Conference on Machine Learning and Cybernetics Wan, 2-5 November 2003.