ISSN: 2321-7782 (Online)
Volume 3, Issue 9, September 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Hybrid Clustering Methods for Web Usage Mining

N. Pushpalatha
Associate Professor in CSE
Marri Laxman Reddy Institute of Technology & Management
Hyderabad, India

Abstract: ***There is a rapid development of World Wide Web in its volume of traffic and the size and complexity of web sites. In this paper, a new approach is presented based on hybrid clustering methods for Web Usage Mining. The WUM process contains three steps: pre-processing, data mining and result analysis. First, it gives a brief description of the WUM process and Web data, then the presentation of the pre-processing step and the data warehouse that were employed. The hybrid clustering methods based on Fuzzy means clustering are used for analyzing and the Web logs taken from the real world Web servers. The results obtained after applying these methods and the corresponding interpretations are also presented. Furthermore, this paper also described web usage mining through cloud computing i.e. cloud mining. The Future work of web mining is to introduce a hierarchy on the information about the website.***

Keywords: Clustering, Web Usage Mining, Web logs, pre-processing, Data Warehouse, Cloud Computing.

I. INTRODUCTION

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services.

There are three general classes of information that can be discovered by web mining:

» Web activity, from server logs and Web browser activity tracking.

» Web graph, from links between pages, people and other data.

» Web content, for the data found on Web pages and inside of documents.

At Scale Unlimited we focus on the last one extracting value from web pages and other documents found on the web.

Note that there's no explicit reference to "search" in the above description. While search is the biggest web miner by far, and generates the most revenue, there are many other valuable end uses for web mining results. A partial list includes:

» Business intelligence

» Competitive intelligence

» Pricing analysis

» Events

» Product data

» Popularity

» Reputation

Four Steps in Content Web Mining

When extracting Web content information using web mining, there are four typical steps.

1. Collect – fetch the content from the Web
2. Parse – extract usable data from formatted data (HTML, PDF, etc)
3. Analyze – tokenize, rate, classify, cluster, filter, sort, etc.
4. Produce – turn the results of analysis into something useful patterns.

Web Mining versus Data Mining

When comparing web mining with traditional data mining, there are three main differences to consider:

1. **Scale** – In traditional data mining, processing 1 million records from a database would be large job. In web mining, even 10 million pages wouldn't be a big number.
2. **Access** – When doing data mining of corporate information, the data is private and often requires access rights to read. For web mining, the data is public and rarely requires access rights. But web mining has additional constraints, due to the implicit agreement with webmasters regarding automated access to this data. This implicit agreement is that a webmaster allows crawlers access to useful data on the website, and in return the crawler promises not to overload the site, and has the potential to drive more traffic to the website once the search index is published. With web mining, there often is no such index, which means the crawler has to be extra careful/polite during the crawling process, to avoid causing any problems for the webmaster.
3. **Structure** – A traditional data mining task gets information from a database, which provides some level of explicit structure. A typical web mining task is processing unstructured or semi-structured data from web pages. Even when the underlying information for web pages comes from a database, this often is obscured by HTML markup.

1) Web Usage Mining

This study of the users' navigations extracted from the web server's log files or proprietary traces may help the webmaster to understand the user behavior and then to rethink the structure and design of his/her website or to detect users' problems and improve the navigability. The WUM analysis, allows the webmaster to optimize the response of the Web server (Web caching) and to make recommendations to the user.

A Web Usage Mining process is commonly split in three phases: preprocessing, data mining and results analysis

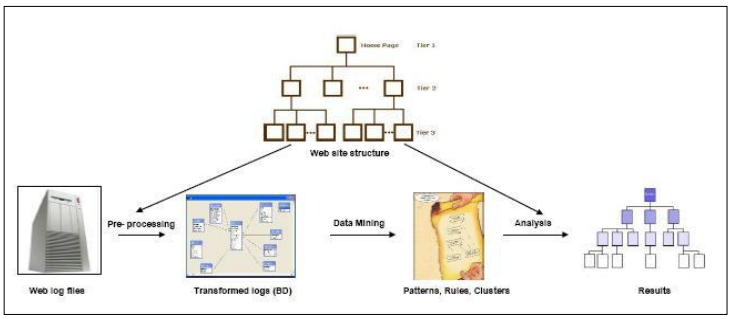

Fig1: Web Usage Mining Schema

2) Web Usage Data

The Web log file is the input data in the Web Usage Mining process. The Web site structure (hyperlinks graph) and the users' profiles may constitute supplementary data for such a process. To solve the problem of organizing these various and bulky data a database structure is necessary. More precisely, this should be a data warehouse as its characteristics (orientation,

subject, integration, history and non-volatility) are useful. The data warehouse is developed for decision purposes. The data warehouse is feed by mechanisms of extracting, transforming and loading data from the log files.

3) HTTP Log Files

According to the HTTP client-server protocol, the client accessing a resource will send a request to the server containing this resource. At the other side, Web server interprets the HTTP request, accesses the requested resource and delivers it to the client. As most of the software programs, these operations are recorded in a log file. The log file allows having a detailed trace of the Web server activity. By using the ECLF log file format for the HTTP log files as described in (Luotonen, 1995) [5].

4) Pre-processing

The objective of the pre-processing step is to identify and structure user navigations. This step is based on two main processes: data cleaning and data transformation. At the end of this phase, the Web logs will be placed in a database warehouse. The preprocessing process consist of following steps

» Pre-processing of the Web log files Data Cleaning

» Data Transformation

» User/Session Identification

4.1. Pre-processing of the Web log files Data Cleaning:-

Log files have useful information about access of all users to a specific website. Extracting the information, reformatted log file which contains useful information such as “time, date, accessed URL and IP address” is formed and useless requests such as accesses to images are removed from log file in data cleaning process. Identifying Web robots and deleting the requests coming from these robots is another task of this process.

4.2 Data Transformation:-

To perform complex analysis, including clustering, grouping together several requests. All the requests made by a single user during the analyzed period constitute his/her session. A session is further split in several navigations, each navigation representing a single visit to the Web site. A navigation ends when a time threshold of at least 30 minutes exists between two consecutive requests.

4.3 User/Session Identification

Identifying users/sessions from the log file is difficult task because of several factors like: proxy servers, dynamic addresses, and the case of two or more users using the same computer or the same user that uses more than one browser or computer. In fact, by using the log file user can know only the computer's address and the User Agent of the user. This is not sufficient in most of the cases that is why there are other methods that can provide more information. The most used are: cookies, dynamic Web pages (with a session ID in the URL), registered users, modified browsers etc. In (Cooley, 2000) [6] the author differentiated users by their navigation.

5) Building a Web Usage Mining Data Warehouse

The Web Usage data can be used to populate the database. The DB is a lasting data warehouse where all the viewpoints are kept. It is different from all other DBs presented in the literature as they are more short-lived or more specialized**.**

a. The facts

The facts came from the log files filtered, ordered and enriched with data calculated as we have presented in the previous sections. The study is centered on the user, therefore the fields are mainly the duration and the size read when he/she accesses a Web page.

b. The selected dimensions can be split up into: Dimension related to the URL and the page viewed. It is the view Content. Many classifications are possible, just from the file extension but also from other information from the Web site. Dimension related to the date. It is the view Access regularity. The usual hierarchy second, minute, hour, day, month, and year can be used as well as any other hierarchy with more specialized periods. Dimension related to the session and the user. It is the view User. A session is issued from the user (when he/she stops requesting for a relatively long time). In our context the user belongs to the following hierarchy: domain → research unit → research team/service. But others classifications are allowed based on different criterions as we shall see further. The description may include the relevant class with this meaning. Dimension related to the referrer and the navigation. It is the view Navigation. The referrer allows mapping out the user navigation which may be linear or with many returns. To help the user in a collaborative view on his/her search in the web site it is essential to classify these navigations. Dimension related to the transaction status. It is the view Access efficiency. The server log files give the result as status success, failure, redirection, forbidden access. Here, once again, also define another status type within the context of the request emission rather than the request treatment.

6) Methodology used in Web Usage Mining.

After the pre-processing step, the data from the log file is structured in sessions and navigations and stored in a DB. This step objectify is to discover different types of user behaviors or categories of user behaviors using different approaches, sequences analysis algorithms, cluster analysis, predictive models, neural networks and automatic clustering. The objective is to develop a strategy which analyses the relations between the structure of the Web site and the log file. To reach it, applying hybrid clustering methods on different types of Web data.

Document Clustering In this work, content mining is used approach for document clustering. Assume G = { g1, g2,..,gn }is the set of n website's pages. Applying the clustering algorithm shown in following Document clustering steps, pages were grouped in content based clusters.

1. Clear each document from stop words such as: about, all , am, almost, as, be, by, but, do and any other word which have not any key role in determining the content of document.
2. Identify document keywords by TF-IDF technique.
3. Assign each document keyword list as a document to a single cluster.
4. Merge primary clusters based on the Jaccard coefficient similarity measure.
5. The second step repeated until all documents being clustered into a pre defined number of clusters. DC = {DC_1, DC_2,…, DC_n} is the result set. Each DC_i represents a set of URLs with similar content.

7) Fuzzy Means Algorithm

The fuzzy means algorithm requires good initialization. These initial values are provided by the ant based algorithm. The result will be small homogenous heaps that will be merged by repeating the steps. By increasing the number of iterations the number of heaps decreases. The User clustering algorithm shows the hybrid algorithm used in this study to cluster users in appropriate groups:

6. Scatter the users randomly on the board
7. Use the cluster centers obtained in step 3 to initialize cluster centers for the fuzzy C-means algorithm
8. Cluster the data using the fuzzy C-means algorithm
9. Harden the data obtained from the Fuzzy C-means algorithm, using the maximum membership criterion, to form new heaps
10. Repeat step 1-6 by considering each heap as a single object.

Experimental Results

Data Summary for Proposed Experiment

Web servers	Reputed web server
Period	1-15 January 2009
Number of requests	6 040 312
Requests after pre-processing	673 389
Number of sessions	115 825
Number of navigation	174 015

II. CONCLUSION

In this paper proposed methodologies used for classifying the user using Web Usage data. This model analysis the users behaviors and depend on the interests of similar patterns provides appropriate recommendations for active user. The model uses the benefits of both content based and collaborative based recommender systems. The results of evaluations shows that using more efficient algorithms for finding similar users lead to recommender system that provides more interesting recommendations for website users. Proposed work can be extended by considering the effect of users feedback for increasing the quality of recommendation. This can be done, eventually, by introducing new parameters for the characterization of the Web Usage data. Future plan to introduce a hierarchy on the semantic topics.

References

1. Bamshad Mobasher, "Data Mining for Web Personalization," LCNS, Springer-Verleg Berlin Heidelberg, 2007.
2. Jaideep Srivastave, Robert Cooley, Mukund Deshpande, Pang-Ning Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," SIGKDD Explorations. ACM SIGKDD,2000.
3. Pierrakos. D, "Web usage mining as a tool for personalization: a survey", User Modeling and User-Adapted Interaction, 13(4), pp. 311-372.
4. Ralph Kimball. Entrepots de donnees. Editions Vuibert, 2001.
5. Luotonen. The common log file format. http://www.w3.org/Daemon/User/Config/ Logging.html, 1995.
6. R. Cooley. Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data. PhD thesis, University of Minnesota, May 2000
7. M. Spiliopoulou, L. C. Faulstich, and K. Winkler. A data miner analyzing the navigational behaviour of web users. In Proc. of the Workshop on Machine Learning in User Modeling of the ACAI'99 Int. Conf., Creta, Greece, July 1999.
8. M. Perkowitz and O. Etzioni. Adaptive web sites: Automatically synthesizing web pages. In AAAI/IAAI, pages 727 to732, 1998.
9. V. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White. Visualization of navigation patterns on a web site using model-based clustering. In In Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 280{284, Boston, Massachusetts, 2000.
10. Mobasher, H. Dai, T. Luo, and M. Nakagawa. Discovery and evaluation of aggregate usage profiles for web personalization. Data Mining and Knowledge Discovery, 61-82, January 2002.

AUTHOR(S) PROFILE

Mrs. N. Pushpa latha. working as a Assoc.professor in Marri Laxman Reddy Institute of Technology and Management, Hyderabad. She has 8+ years teaching experience and good knowledge in computer subjects. She completed master degree in computer science and engineering dept. from University College of Engg, JNTU Campus, Kakinada Presently pursuing Ph. D from JNTU, Hyderabad.

