Volume 3, Issue 7, July 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Clustering Real Time Web Usage Data Using Fuzzy Approach

P.Sudheer[1], B. Srinivas[2], B. Raju[3]

Kakatiya Institute of Technology and Science, Warangal, India
Asst. Professor, Dept. of CSE, Kakatiya Institute of Technology and Science, Warangal, India
Asst. Professor, Dept. of CSE, Kakatiya Institute of Technology and Science, Warangal, India

Abstract— Categorization of Users and Web pages are the essential tasks associated with Web Personalization. In this project it is proposed any Matrix Based Fuzzy Clustering Approach MBFCA along with experimentally evaluated the approach to the effective breakthrough discovery of world-wide-web user groupings and website page clusters. The employment of MBFCA allows the generation of clusters which could capture the internet user's routing behaviour determined by their curiosity. In world-wide-web usage investigation, many any times you can find no razor-sharp boundary concerning clusters. Hence fuzzy clustering is much better suited with regard to Web Use Mining. The offered system could be enhanced through the use of web application mining that has been employed extensively with regard to Web customization. A quantity of personalized providers employ equipment learning procedures, particularly clustering methods, to examine Web application data along with extract useful knowledge to the recommendation associated with links that you follow within an affiliate site, or to the customization of Sites to this preferences of the users. A radical analysis of such methods, together using their benefits and drawbacks in this context associated with Web Personalization. PLSA may be used within the context associated with Collaborative Blocking and World-wide-web Usage Mining. In the first case, PLSA was used to construct any model-based composition that talks about user reviews.

Keywords— Web Usage, Clustering, Pattern Recognition

I. INTRODUCTION

Web usage mining is a very important part of web mining, and it tries to discover interesting web user access patterns or knowledge from the web log records. Web log files contain a huge amount of data about user access patterns. Hence, if properly exploited, they can reveal useful information about the browsing behavior of users in a site. Analyzing and exploring regularities in web log records can identify customers for e-commerce, enhance quality of IIS and improve web server system performance. Web usage mining approach applies Data Mining algorithms on Web usage data and among them clustering is an effective way to group users with common browsing behavior. In the choice of the clustering method for Web usage mining, one important constraint to be considered is the possibility to obtain overlapping clusters, so that a user can belong to more than one group. To deal with the ambiguity and the uncertainty underlying Web interaction data, as well as to derive overlapping clustering, fuzzy clustering appears to be an effective tool. Web mining has obvious fuzzy characteristic, so fuzzy clustering is better suited for the web mining. So the concept of Matrix Based Fuzzy Clustering Approach MBFCA is put forward for Web usage mining. The input data object of the Matrix Based Fuzzy Clustering is the web source matrix which represents the data objects and its attributes of the given web data set. But the processing data object of MBFCA is web fuzzy similarity matrix using which the web users and web pages clustering is done. So it is required to abstract web source data firstly and then transform it into web fuzzy matrix which is suitable for fuzzy clustering. In the end, fuzzy clustering method is applied on web fuzzy matrix to obtain the clustering results.

The proposed system can be enhanced by using web usage mining which has been used extensively for Web personalization. A number of personalized services employ machine learning methods, particularly clustering techniques, to analyse Web usage data and extract useful knowledge for the recommendation of links to follow within a site, or for the customization of Web sites to the preferences of the users. A thorough analysis of these methods, together with their pros and cons in the context of Web Personalization. PLSA has been used in the context of Collaborative Filtering and Web Usage Mining. In the first case, PLSA was used to construct a model-based framework that describes user ratings. Latent factors were employed to model unobservable motives, which were then used to identify similar users and items, in order to predict subsequent user ratings. In PLSA was used to identify and characterize user interests inside certain Web sites. The latent factors segmented user sessions to support a personalized recommendation process.

II. PREVIOUS WORK

The World Wide Web is an immense source of data that can come either from the Web content, represented by the billions of pages publicly available, or from the Web usage, represented by the log information daily collected by all the servers around the world. Web Mining is that area of Data Mining which deals with the extraction of interesting knowledge from the World Wide Web. More precisely, Web Content Mining is that part of Web Mining which focuses on the raw information available in Web pages; source data mainly consist of textual data in Web pages (e.g., words, but also tags); typical applications are content-based categorization and content-based ranking of Web pages. Web Structure Mining is that part of Web Mining which focuses on the structure of Web sites; source data mainly consist of the structural information present in Web pages (e.g., links to other pages); typical applications are link-based categorization of Web pages, ranking of Web pages through a combination of content and structure, and reverse engineering of Web site models. Web Usage Mining is that part of Web Mining which deals with the extraction of knowledge from server log files; source data mainly consist of the (textual) logs that are collected when users access Web servers and might be represented in standard formats (e.g., Common Log Format, Extended Log Format, LogML); typical applications are those based on user modeling techniques, such as Web personalization, adaptive Web sites, and user modeling. Web Usage Mining applications are based on data collected from three main sources: (i) Web servers, (ii) proxy servers, and (iii) Web clients. Web servers are surely the richest and the most common source of data. They can collect large amounts of information in their log files and in the log files of the databases they use. These logs usually contain basic information e.g.: name and IP of the remote host, date and time of the request, the request line exactly as it came from the client, etc. This information is usually represented in standard format e.g.: Common Log Format, Extended Log Format, and LogML. Sometimes databases are used instead of text files to store log information so to improve querying of massive log repositories.

When exploiting log information from Web servers, the major issue is the identification of user's sessions, i.e., how to group all the users' page requests or click streams so to clearly identify the paths that users followed during navigation through the web site. This task is usually quite difficult and it depends on the type of information available in log files. The most common approach is to use cookies to track down the sequence of user's page requests for an overview of cookie standards. If cookies are not available, various heuristics can be employed to reliably identify user's sessions. Note however that, even if cookies are used, it is still impossible to identify the exact navigation paths since the use of the back button is not tracked at the server level. Apart from Web logs, user's behavior can also be tracked down on the server side by means of TCP/IP packet sniffers. Even in this case the identification of user's sessions is still an issue, but the use of packet sniffers provides some advantages. In fact: (i) data are collected in real time; (ii) information coming from different Web servers can be easily merged together into a unique log; (iii) the use of special buttons can be detected so to collect information usually unavailable in log files. Notwithstanding the many advantages, packet sniffers are rarely used in practice. Packet sniffers raise scalability issues on Web servers with high traffic, moreover they cannot access encrypted packets like those used in secure commercial transactions. Unfortunately, this limitation turns out to be quite severe when applying Web Usage Mining to e-businesses. Probably, the best approach for tracking Web usage consists of directly accessing the server application layer. Unfortunately, this is not always possible. First, there is issue related to the copyright of server applications. Most important, following this approach, Web Usage Mining applications must be tailored for the specific servers and have to take into account the specific tracking requirements. Many Internet Service Providers (ISPs) give to their customer proxy server services to improve navigation speed through caching. In many respects, collecting navigation data at the proxy level is basically the same as collecting data at the server level. The main difference in this case is that proxy servers collect data of groups of users accessing huge groups of web servers.

Even in this case, session reconstruction is difficult and not all users navigation paths can be identified. However, when there is no other caching between the proxy server and the clients, the identification of users sessions is easier. Usage data can be tracked also on the client side by using JavaScript, Java applets, or even modified browsers. These techniques avoid the problems of users sessions identification and the problems caused by caching. In addition, they provide detailed information about actual user behaviors.[1]

Data pre-processing has a fundamental role in Web Usage Mining applications. Notices that even if pre-processing techniques are widely used in Web Usage Mining, the literature on this topic is still quite limited, and that the most complete reference on pre-processing dates back to 1999. The pre-processing of Web logs is usually complex and time demanding. It comprises four different tasks: (i) the data cleaning, (ii) the identification and the reconstruction of user's sessions, (iii) the retrieving of information about page content and structure, and (iv) the data formatting. This step consists of removing all the data tracked in Web logs that are useless for mining purposes e.g.: requests for graphical page content (e.g., jpg and gif images); requests for any other file which might be included into a web page; or even navigation sessions performed by robots and Web spiders. While requests for graphical contents and files are easy to eliminate, robots and Web spider's navigation patterns must be explicitly identified.

This is usually done for instance by referring to the remote hostname, by referring to the user agent, or by checking the access to the robots.txt file. However, some robots actually send a false user agent in HTTP request. In these cases, an heuristic based on navigational behavior can be used to separates robot sessions from actual users sessions. Is evidenced that search engine navigational paths are characterized by breadth first navigation in the tree representing the Web site structure and by unassigned referrer the referrer gives the site that the client reports having been referred from. The heuristic proposed is based on the previous assumption and a classification of navigations. Well known robots navigational paths are used to train the classifier, and the model obtained is used to classify further navigational sessions even if there is no a priori knowledge about the user agent that generated them.

The aim of Web usage mining is to discover patterns of user activities in order to better serve the needs of the users for example by dynamic link handling, by page recommendation, etc. The aim of a Web site or Web portal is to supply the user the information which is useful for him. There is a great competition between the different commercial portals and Web sites because every user means eventually money (through advertisements, etc.). Thus the goal of each owner of a portal is to make his site more attractive for the user. For this reason the response time of each single site have to be kept below 2s. Moreover some extras have to be provided such as supplying dynamic content or links or recommending pages for the user that are possible of interest of the given user. Clustering of the user activities stored in different types of log files is a key issue in the Web community. There are three types of log files that can be used for Web usage mining. Log files are stored on the server side, on the client side and on the proxy servers. By having more than one place for storing the information of navigation patterns of the users makes the mining process more difficult. Really reliable results could be obtained only if one has data from all three types of log file. The reason for this is that the server side does not contain records of those Web page accesses that are cached on the proxy servers or on the client side. Besides the log file on the server, that on the proxy server provides additional information. However, the page requests stored in the client side are missing. Yet, it is problematic to collect all the information from the client side. Thus, most of the algorithms work based only the server side data.

III. PROPOSED SYSTEM

A. *Session identification and reconstruction*

This step consists of (i) identifying the different users sessions from the usually very poor information available in log files and (ii) reconstructing the user's navigation path within the identified sessions. The complexity of this step can vary a lot depending on the quality and on the quantity of the information available in the Web logs. Most of the problems encountered in this phase are caused by the caching performed either by proxy servers either by browsers. Proxy caching causes a single IP address (the one belonging to the proxy Server) to be associated with different users sessions, so that it becomes impossible to use IP addresses as users identifiers. This problem can be partially solved by the use of cookies, by URL rewriting, or by requiring the user to log in when entering the Web site. A cookie is a piece of information sent by a Web server to a Web browser. This information is stored on the user's computer as a text file. Cookies may contain a lot of information about users; among them the one we are interested in is the session identifier. This information can be asked by the Web server every time a user asks for a Web page and stored in the Web log together with the page request. There are situations, however, where cookies will not work. Some browsers, for example, do not support cookies. Other browsers allow the user to disable cookie support. In such cases, URL rewriting can be used to track the users session by including the session ID in URLs. URL rewriting involves finding all links that will be written back to the browser, and rewriting them to include the session ID. For example, a link such as <a href="/store/catalogue"> can rewritten as <a href="/store/catalog; jsessionid=DA32242SSGE2"> so as to include the session ID information, i.e., DA32242SSGE2. Hence every time a user clicks on a link in the page, the rewritten form of the URL is sent to the server and stored in the Web log. Web browser caching is a more complex issue. Logs from Web servers cannot include any information about the use of the back button. This can generate inconsistent navigation paths in the user's sessions. However, by using additional information about the Web site structure is still possible to reconstruct a consistent path by means of heuristics. For example if a page request is made, and this page request is not directly linked to the previous page request, the referrer log can be checked to see from what page the request came from. If the page is in the user's recent history request is possible to assume that the user used the back button. And then based on this assumption is possible to reconstruct a complete and consistent navigational path. To solve both proxy and web caching issues, IBM has introduced within Surf Aid a JavaScript called Web Bug which has to be included in each Web page. Every time the Web page is loaded, Web Bug sends a request to the server asking for a 1 · 1 pixel image; the request is generated with parameters identifying the Web page containing the script and a numeric random parameter; the overall request cannot be cached neither by the proxy neither by the browser but it is logged by the Web server so as to solve caching problems. Because the HTTP protocol is stateless, it is virtually impossible to determine when a user actually leaves the Web site in order to determine when a session should be considered finished. This problem is referred to as sessionization.

B. *Content and structure retrieving*

The vast majority of Web Usage Mining applications use the visited URLs as the main source of information for mining purposes. URLs are however a poor source of information since; for instance, they do not convey any information about the actual page content. An additional categorization step in which Web pages is classified according to their content type; this additional information is then exploited during the mining of Web logs. If an adequate classification is not known in advance, Web Structure Mining techniques can be employed to develop one. As in search engines, Web pages are classified according to their semantic areas by means of Web Content Mining techniques; this classification information can then be used to enrich information extracted from logs. For instance, proposes to use Semantic Web for Web Usage Mining: Web pages are mapped onto ontology's to add meaning to the frequently observed paths. Given a page in the Web site, we must be able to extract domain-level structured objects as semantic entities contained in this page. This task may involve the automatic extraction and classification of objects of different types into classes based on the underlying domain ontologies. The domain ontologies themselves may be pre-specified, or may be learned automatically from available training data. Given this capability, the transaction data can be transformed into a representation which incorporates complex semantic entities accessed by users during a visit to the site. concept-based paths as an alternative to the usual user navigation paths; concept-based path are a high level generalization of usual path in which common concepts are extracted by means of intersection of raw user paths and similarity measures.

C. *Data formatting*

This is the final step of pre-processing. Once the previous phases have been completed, data are properly formatted before applying mining techniques. Stores data extracted from Web logs into a relational database using a click fact schema, so as to provide better support to log querying finalized to frequent pattern mining. Introduces a method based on signature tree to index log stored in databases for efficient pattern queries. A tree structure named WAP-tree is also introduced to register access sequence to Web pages; this structure is optimized to exploit the sequence mining algorithm developed by the same authors. Stores log data in another tree structure, the FBP-tree, to improve sequence pattern discovery. Uses a cube-like structure to store session information, to improve the extraction of cube slices used by clustering techniques. [1]

D. ***Clustering Algorithms***

The clustering is a process of discovering groups of objects such that the objects belonging to the same group are similar in a certain manner, and the objects belonging to different groups are dissimilar. The main problems one faces when creating a clustering algorithm are the following:

The objects can have hundreds of attributes that have to be taken into consideration for clustering. One of the key issues is how to reduce this number to achieve an efficient algorithm.The type of the attributes can be diverse, and not only numerical attributes has to be handled. Because of the first two problems defining a similarity function between the objects is not a trivial task. Many features and many types of attributes have to be handled efficiently. The main feature of clustering in a data mining application is the high number of objects that have to be clustered. Thus the processing time or the memory requirement of the algorithm can be huge, that has to be reduced using some heuristics. Validating the resulting clusters is also a hard task. In case of low dimensionality, when the clusters can be represented visually, the validation can be made by a human, but in case having large number of objects with high dimensionality statistical methods have to be used and indices have to be defined which can be computationally expensive.

There are many algorithms in the literature that deal with the problem of clustering large number of objects. The different algorithms can be classified regarding different aspects. One of the key issues, which also determine another features of the algorithm is the basic approach of the clustering algorithm. The aim of the partition-based algorithms is to decompose the set of objects into a set of disjoint clusters where the number of the resulting clusters is predefined by the user. The algorithm uses an iterative method, and based on a distance measure it updates the cluster of each object. It is done until any changes can be made. The most representative partition-based clustering algorithms are the k-means and the k-mediod, and in the data mining field the CLARANS. The advantage of the partition based algorithms that they use an iterative way to create the clusters, but the drawback is that the number of clusters has to be determined in advance and only spherical shapes can be determined as clusters.

IV. RESULTS

The concept of this paper is implemented and different results are shown below, The proposed paper is implemented in .Net technology on a Pentium-IV PC with minimum 20 GB hard-disk and 1GB RAM. The propose paper's concepts shows efficient results and has been efficiently tested on different Datasets.

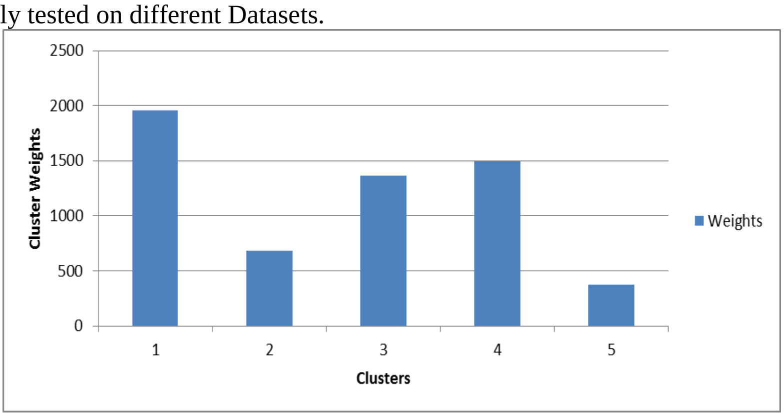

Fig. 1 Graph showing cluster weights for synthetic data.

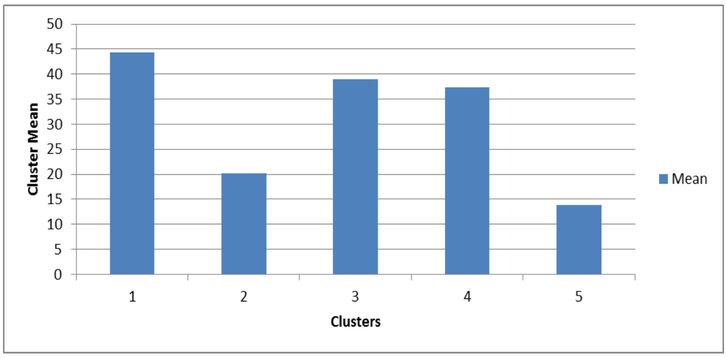

Fig. 1 Graph showing cluster means for synthetic data

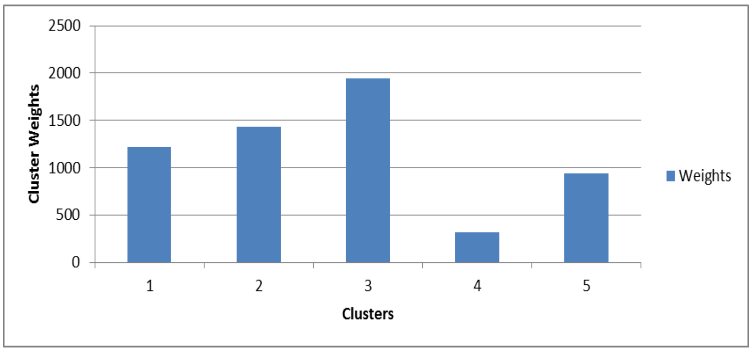

Fig. 3 Graph showing cluster weights for real time data

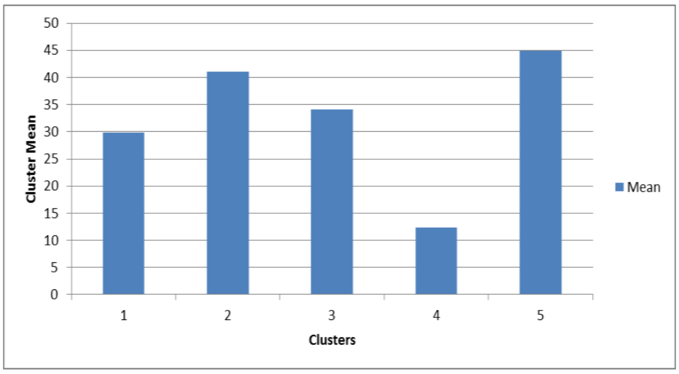

Fig. 4 Graph showing cluster means for real time data

V. CONCLUSIONS

In this paper, the concept and processing model of Matrix Based Fuzzy Clustering Approach MBFCA is put forward and discussed. The experimental results are shown to prove that this approach can be used for effective web user clustering and web page clustering. This is a simple to implement approach. This approach will produce efficient clusters of web users and web pages with less run time and with reduced memory usage. The resultant web page and web user clusters are found to match the existing web page and web user clusters in the given sample data. One future work on this could be to create matrices that reflect the time each user spends on each page and come up with a different set of web user and web page clusters based on this data. Another future scope for this approach is that this can be automated using any programming language and can be used to measure the efficiency of this approach for many web sites.

REFERENCES

[1] Facca, F. M., &Lanzi, P. L. (2005). Mining Interesting Knowledge from Weblogs: A Survey, 53, 225–241.

[2] RENATA IVANCSY and FERENC KOVACS, “Clustering Techniques Utilized in Web Usage Mining,” in Proceedings of the 5th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain, February 15-17, 2006, pp. 237-242.

[3] Chu-Hui Lee, Yu-Hsiang Fu “Web Usage Mining Based on Clustering of Browsing Features”, IEEE Eighth International Conference on Intelligent Systems Design and Applications, 2008, p. 281-286.

[4] Xinlin Zhang, Xiangdong Yin “Design of an Information Intelligent System based on Web Data Mining”, IEEE International Conference on Computer Science and Information Technology, 2008, p. 88-91.

[5] A. Vakali, J. Pokorný and T. Dalamagas, “An Overview of Web Data Clustering Practices,” EDBT Workshops, 2004, pp. 597-606.

[6] Han, Q., Gao, X., Wu, W.: Study on Web Mining Algorithm Based on Usage Mining. In: 9th International Conference on Computer-Aided Industrial Design and Conceptual Design, CAID/CD November 2008.

[7] Sudhamathy, G. 2010. Mining web logs: an automated approach. In Proceedings of the 1st Amrita ACM-W Celebration on Women in Computing in india (Coimbatore, India, September 16 - 17, 2010). A2CWiC '10. ACM, New York, NY, 1-4. DOI= http://doi.acm.org/10.1145/1858378.1858435

[8] Thorleuchter, D., Poel, D. V. D., &Prinzie, A. (2012). Analysing existing customers’ websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing. Expert Systems with Applications, 39, 2597–2605

AUTHORS BIOGRAPHY

. Sudheer P is pursuing his M.tech (Software Engineering) at Kakatiya Institute of Technology & Science, Warangal, A.P, INDIA. He has completed his B.tech from JNTU in 2010 and his main research includes database administration and programming in MS.NET

B. Srinivas is currently working as Asst. Professor at Kakatiya Institute of Technology & Science, Warangal, A.P, INDIA. He has completed his M.Tech from JNTU. His main research includes Data Mining, Algorithm Analysis and Design, Operating Systems. He has been involved in the organization of a number of conferences and workshops.

B. Raju is currently working as Asst. Professor at Kakatiya Institute of Technology & Science, Warangal, A.P, INDIA. He has completed his M.Tech from JNTU .His main research includes Data Mining, Compiler Design; Theory of Computation. He has been involved in the organization of a number of conferences and workshops.