An Effective Framework For Identifying Personalized Web Recommender System By Applying Web Usage Mining

T.SubMasthan Rao, Y.Ravindra,U.Satish Kumar,S.Sandeep,K.Srikanth
Department of Information Technology,KLUniversity,Vijayawada,Andhra Pradesh, India.

Abstract

The Internet is one of the fastest growing areas of intelligence gathering. During their navigation web users leave many records of their activity. This huge amount of data can be a useful source of knowledge. Sophisticated mining processes are needed for this knowledge to be extracted, understood and used. Web Usage Mini(WUM) systems are specifically designed to carry out this task by analyzing the data representing usage data about a particular Web Site. WUM can model user behavior and, therefore, to forecast their future movements. Online prediction is one web usage mining application. However, the accuracy of the prediction and classification in the current architecture of predicting users' future requests systems can not still satisfy users especially in Huge Web sites. To provide online prediction efficiently,we develop architecture for online recommendation for predicting in Web Usage Mining System .In this paper we propose architecture of online recommendation in Web usage mining(OLRWMS) for enhancing accuracy of classification by interaction between classifications, evaluation, and current user activates and user profile in online phase of this architecture.

1. Introduction

With the explosive growth of knowledge available on the World Wide Web, which lacks anintegrated structure or schema, it becomes much more difficult for users to access relevant information efficiently. Meanwhile, the substantial increase in the number of websites presents a challenging task for webmasters to organize the contents of the websites to cater to the needs of users. Modeling and analyzing web navigation behavior is helpful for understand what information online users demand. Following that, the analyzed results can be seens knowledge to be used in intelligent online applications, refining web site maps, and improving searching accuracy when seeking information.Nevertheless, an online navigation behavior grows each passing day, and thus extracting intelligently from it is a difficult issue. Web Mining has shown to be a viable technique to discover information "hidden" into Web-related data [1]. In particular, Web Usage Mining (WUM) is the process of extracting knowledge from Web user's access data by exploiting Data Mining (DM) technologies [2].It can be used for different purposes such as personalization, system improvement and site modification. Typically, the WUM prediction process is structured according to two components performed online and off-line with respect to the Web server activity [14], [3], [12], and [5]. The off-line component is aimed at building the knowledge base by analyzing historical data, such as server access log files, that is then used in the online component.

The main functions carried out by this component are Preprocessing, i.e. data cleaning and session identification, and Pattern Discovery, i.e. the application of DM techniques, like association rules, sequential patterns, clustering or classification. The online component is devoted to the generation of personalized content. On the basis of the knowledge extracted in the off-line component, it processes a request to the Web server by adding personalized content which can be expressed in several forms, such as links to pages, advertisements, and information relating to products or service estimated to be of interest for the current user. In the past, several WUM projects have been proposed to predict users' preference and their navigation behavior, as well as many recent results improved separately the quality of the recommendations or the user profiling phase [6], [7], [8].

2. Related Work

Recently, several WUM systems have been proposed to predicting user's preference and their navigation behavior. In the following we review some of the most significant WUM systems and architecture that can be compared with our system. Analog [9] is one of the first WUM systems. It is structured according to an off- line and an online component. The off-line component builds session clusters by analyzing past users activity recorded in server log files. Then the online component builds active user sessions which are then classified according to the generated model. The classification allows to identify pages related to the ones in the active session and to return the requested page with a list of suggestions. The geometrical approach used for clustering is affected by several limitations, related to scalability and to the effectiveness of the results found. Nevertheless, the architectural solution introduced was maintained in several other more recent projects.

In [18] and [11] B. Mobasher *et al., present WebPersonalizer* a system which provides dynamic recommendations, as a list of hypertext links, to users. The analysis is based on anonymous usage data combined with the structure formed by the hyperlinks of the site. Data mining techniques (i.e. clustering, association rules and sequential pattern discovery) are used in the preprocessing phase in order to obtain aggregate usage profiles. In this phase Web server logs are converted in clusters made up of sequences of visited pages, and cluster made up of set of pages with common usage characteristics. The online phase considers the active user session in order to find matches among the user's activities and the discovered usage profiles. Matching entries are then used to compute a set of recommendations which will be inserted into the last requested page as a list of hypertext links. Web Personalizer is a good example of two-tier architecture for Personalization systems. In [4] they proposed an architecture that named KLAS, base on customer's on-line navigation behaviors by analyzing their navigation patterns through pre-trained artificial neural networks. In [14] they have developed a recommendation system, termed Yoda that is designed to support large-scale Web-based applications requiring highly accurate recommendations in real-time. With Yoda, they introduced a hybrid approach that combines collaborative filtering (CF) and content-based querying to achieve higher accuracy. Yoda is structured as a tunable model that is trained online and employed for real-time recommendation on-line. The on-line process benefits from an optimized aggregation function with low complexity that allows real time weighted aggregation of the soft classification of active users to predefined recommendation sets. Liu and Keselj [2] proposed the automatic classification of web user navigation patterns and proposed a novel 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 retrieved web pages. They used character N-grams to represent the contents of web pages, and combined them with user navigation patterns by building user navigation profiles composed of a collection of N-grams. The approach is implemented as an experimental system, and itsperformance is evaluated based on two tasks: classification and prediction. The systemachieves the classification accuracy of nearly 70% and the prediction accuracy of about 65%, which is about 20% higher than the classification accuracy by mining Web server logs alone. In the session vectorization of this system for capturing the interest degree of a web page there is only two factors: Frequency and duration 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. In [15], Baraglia and Palmerini proposed a WUM system called SUGGEST, that provide useful information to make easier the web user navigation and to optimize the web server performance. The main goal of SUGGEST is to find useful information from the user access data collected in web server logs. SUGGEST adopts a two levels architecture composed by an offline creation of historical knowledge and an online engine that understands user's behavior. After a pre-processing of the data recorded in the web server log files, SUGGEST creates clusters of related pages based on users past activity, and then classifies new users by comparing pages in their active sessions with pages inside the clusters created. A set of suggestions is then obtained for each request. The main disadvantages of this system are: Online component and offline component work separately, how to maintain and update the knowledge extracted in the offline phase and how the system can exactly understand the differences between index page and content page.

In the new architecture of SUGGEST they put together the previous two components into a single online module performing the same operation [16, 17].As the requests arrive at this system module it incrementally updates a graph representation of the Web site based on the active user sessions and classifies the active session using a graph partitioning algorithm. This architecture was designed to be usable on Web sites made up of pages statically generated, i.e. Web sites with a fixed number of pages. A list containing all the information describing a Web site pages was required as input by this architecture at its start-up time.

Potential limitation of this architecture might be: a) the memory required to store Web server pages is quadratic in the number of pages. This might be a severe limitation in large sites made up of millions of pages; b) it does not permit us to manage Web sites made up of pages dynamically generated. The last contribution of SUGGEST architecture proposed by Baraglia et al. [9]. This version of SUGGEST introduce a novel solution to implement WP(Web Personalization) as a single online module that performs user profiling, model updating, and recommendation building. It is designed to dynamically generate personalized contents of potential interest for users of large Web sites made up of pages dynamically generated. It is based on an incremental personalization procedure tightly coupled with the Web server. It is able to update incrementally and automatically the knowledge base obtained from historical usage data and to dynamically generate a list of page links (suggestions). The suggestions are used to personalize the HTML page requested on-the-fly. The adoption of a LRU-based (Least Recently Used) algorithm handling the knowledge base makes it possible for SUGGEST to manage large Web sites. But in this system quality of recommendations is not better than previous version of this system.

3. Architecture of Online Recommendation in Web Usage Mining System

The OLRWUMS, shorting for the Online Recommendation for Predicting in Web Usage Mining system, is a Data Mining system that can be used for online predicting of users' request. According to different functions, the system can be partitioned into two main phases; offline phases and online phases. The architecture of this system is shown in Fig.1.

3.1 Offline Phase

This phase consists of two major modules: Data pretreatment and Navigation Patterns Mining. In this phase we startwith the primary Web-Log Preprocessing (Data pretreatment) toextract user navigation session from dataset and after that we will try to apply some algorithm to mining navigational patterns.

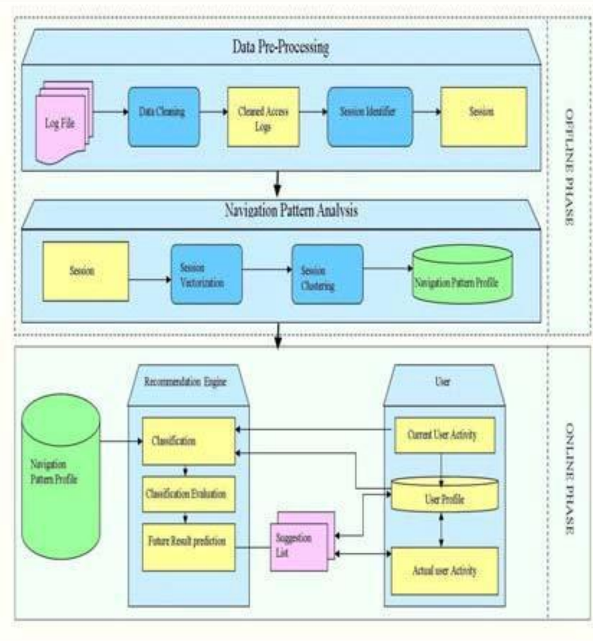


Fig 1: The Architecture of OLRWUMS

3.1.1 Data pretreatment

Data pretreatment in a web usage mining model (Web-Log preprocessing) aims to reformat the original web logs to identify all web access sessions. The Web server usually registers all users' access activities of the website as Web server logs. Due to different server setting parameters, there are many types of web logs, but typically the log files share the same basic information, such as: client IP address, request time, requested URL, HTTP status code, referrer, etc. Generally, several pretreatment tasks need to be done before performing web mining algorithms on the Web server logs. For our work, these include data cleaning, user differentiation and session identification. These preprocessing tasks are the same for any web usage mining problem and are discussed by Cooley et al. [19]. The original server logs are cleaned, formatted, and then grouped into meaningful sessions before being utilized by web usage mining.

3.1.2 Navigation Pattern Mining

After the data pretreatment step, we will perform navigation pattern mining on the derived user access sessions. As an important operation of navigation pattern mining, clustering aims to group sessions into 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 process of our system. In this system module, we will introduce how we perform the session clustering and how we identify the optimal number of clusters from clustering results.

3.1.3 Session vectorization

For the session clustering we should assign a weight to web page visited in a session. The weight needs to be appropriately determined to capture a user's interest in a web page. In general, all the accessed page can be considered interesting to various degrees because users visited them. In reference [50] they proposed a weight measure for approximating the interest degree of a web page to a user. In this research for representing the interest degree of a web page to a user in the session, they measured "Frequency" and "Duration" of a page in the session. But we are interested in incorporation more influencing factors into the weight measure of the session vectorization, for example sequence of accessed web pages. After interest measuring every user access session is successfully transformed into an m-dimensional vector of weights of web pages, where m is the number of web pages visited in all user access sessions. For reducing dimensions, we can use a frequency threshold *f*min as a constraint to filter out web pages that are accessed less than *f*min time sin all access sessions.

3.1.4 Session Clustering

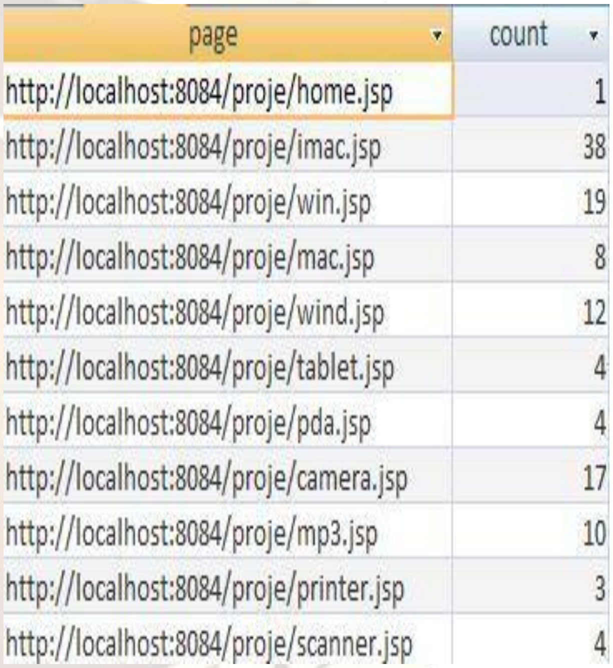
In this step, standard clustering algorithm can partition user access sessions. The result of session clustering is used to represent the set of user navigation patterns. Given the transformation of user access sessions into a multi-dimensional space as vectors of web pages, standard clustering algorithms can partition this space into groups of sessions that are close to each other based on a distance measure. The results of sessions clustering will save in navigational pattern profile.

3.2 Online Phase

During the online phase, when a new request arrives at the server, the URL requested and the session to which the user belongs are identified, the underlying knowledge base is updated, and a list of suggestion is appended to the requested page. This phase consists of some functions that are discussed below.

3.2.1 Recommendation engine

The objectives of this part of OLRWUMS are to classify user navigation patterns and predict users' future requests.we use simplest mechanism for recommendations.Unlike all algorithms purely based on weblog for recommendation process.Usage of weblog is complex mechanishm and consumes lots of cpu time .so we minimize the use of weblog.In our recommendation process we create separate for every user who registered in the website.so that we could track down the pages Visited by a particular user,the table created for register ed users consists of details about the pages visited by the user and the frequency of the page visited by the user i.e., the count of how many times that particular user visited that particular page.By this we can easily detect that which page is visited by user frequently.And The page with maximum count is given as recommenda tion to the particular user.



page	count
http://localhost:8084/proje/home.jsp	1
http://localhost:8084/proje/imac.jsp	38
http://localhost:8084/proje/win.jsp	19
http://localhost:8084/proje/mac.jsp	8
http://localhost:8084/proje/wind.jsp	12
http://localhost:8084/proje/tablet.jsp	4
http://localhost:8084/proje/pda.jsp	4
http://localhost:8084/proje/camera.jsp	17
http://localhost:8084/proje/mp3.jsp	10
http://localhost:8084/proje/printer.jsp	3
http://localhost:8084/proje/scanner.jsp	4

fig:user tables with page and count columns

The table consists of two columns, page and count.The page consists of page visited by the user and count consists of the frequency of the page visited by the user.And for the unregistered users the page that is frequently visited by all users is given as recommendations. If the user register then a table is created for the particular user by his username, therefore the username must be unique. This mecha nism is very simple to implement but very effective though as we didn't use complex algorithms for recommendation process. Recommendations is also based on certain priorities such as the time spent by the user on a particular product or item is added to the cart and purchase is stopped due to certain errors these type of items are of higher priority than those items which are just visited by the user.

3.2.2 Session Identification

The objective of session identification is to detect the time spent by the user on a particular product and to detect the interest of the user towards the product by this priorities of the items can be distinguished i.e., the item whose purchase stopped at cart stage has to be given more priority than those items which are just viewed by the user.

4.Conclusion

In this paper, the architecture is proposed to classify user navigation pattern and Online Recommendation to users' for predication of future request by mining of web server logs. This architecture will be used in the Web Usage Mining System named as ORWUMS. In this architecture, a recommendation engine that works in online phase predicts next user request by interacting user profile and classification module. After classification part, accuracy of classification will be evaluated by evaluation part. If this accuracy doesn't satisfy the user, classification part will operate again based on latest user activity and user profile until the needed accuracy is satisfied.

5. References:

[1]. R. Kosala, H. Blockeel, Web mining research: a survey,SIGKDD: SIGKDDexplorations: newsletter of the special interest group (SIG) on knowledge discovery & data mining, ACM 2 (1), pages 1-15, 2000.

[2]. H. Liu, V. Keselj, Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests, Elsevier, 2007.

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

[4]. Shuchih Ernest Changa, S. Wesley Changchiena ,Ru-Hui Huangb ,Assessing users' productspecific knowledgefor personalization in electronic commerce, Expert Systems with Applications 30 ,pages 682–693,2006.

[5]. R. Baraglia and P. Palmerini. Suggest: A web usage mining system. In Proc. of IEEE Int'l Conf. on InfoTech: Coding and Computing, April 2002.

[6.] O. Nasraoui and C. Petenes. Combining web usage mining and fuzzy inference for website personalization. In Proc. Of WebKDD, 2003.

[7]. M. Nakagawa and B. Mobasher. A hybrid web personalization model based on site connectivity. In Proc. of WebKDD, pages 59–70, 2003.

[8]. E. Frias-Martinez and V. Karamcheti. Reduction of user perceived latency for a dynamic and personalized site using web-mining techniques. In Proc. of WebKDD, pages 47–57, 2003.

[9]. T. W. Yan, M. Jacobsen, H. Garcia-Molina, and D. Umeshwar. From user access patterns to dynamic hypertext linking. Fifth International WorldWide Web Conference, May 1996.

[10]. R. Baraglia, F. Silvestri, Dynamic Personalization of Web Sites Without User Intervention. Communication of the ACM,February 2007.

[11]. M. Nakagawa and B. Mobasher. A hybrid web personalization model based on site connectivity. In Proc. of WebKDD, pages 59–70, 2003.

[12]. B. Mobasher, N. Jain, E.-H. S. Han, and J. Srivastava. Web mining: Pattern discovery from world wide web transactions.TR 96-050, University of Minnesota, 1996.

[13]. C. Shahabi, F.B. Kashani, Y.-S. Chen, D. McLeod,Yoda: An accurate and scalable web-based recommendation system, in: Proceedings of the 9th International Conference on Cooperative Information Systems, Springer-Verlag, pages 418–432, 2001

[14]. B. Mobasher, R. Cooley, and J. Srivastava. Automatic personalization based on web usage mining. Communications of the ACM, 43(8):142– 151, august 2000.

[15]. R. Baraglia and P. Palmerini. Suggest: A web usage mining system. In Proc. of IEEE Int'l Conf. on I.T: Coding and Computing,April 2002.

[16]. F. Silvestri, R. Baraglia, P. Palmerini and M.Serrano. On-line generation of suggestions for web

users. In Proc. of IEEE Int'l Conf. on Info. Technology: Coding and Computing, 2004

[17]. R. Baraglia, F. Silvestri, An online recommender system for large Web sites. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pages 20–24, 2004.

[18]. B. Mobasher, R. Cooley, and J. Srivastava. Automatic personalization based on web usage mining. Communications of the ACM, 43(8):142–151, august 2000.

[19]. R. Cooley, B. Mobasher, J. Srivastava. Data Preparation for Mining World Wide Web Browsing Patterns. Journal of Knowledge and Information Systems. Vol. 1.No. 1. Pages 5–32.1999.