I J C T A, 9(28) 2016, pp. 13-19

A WEB RECOMMENDATION SYSTEM BASED ON USER NEEDS AND INTERESTS

S.Umamaheswari* S.Bakiyalakshmi and S.K.Srivatsa*****

Abstract: Web is becoming an enormous storehouse of information and it will keep growing with improvements in internet technologies. But the human capability to read, access and understand content does not increase. Hence it becomes difficult to website owners to provide appropriate information to the users. This lead to personalized web services to provide personalized web experience to users. One of the well-liked approaches in providing web personalization is Web Usage Mining. In this paper, we will discuss preprocessing and it's various modules. We will also discuss Recommender systems; which makes use of Web personalization for providing tailored recommendations to the user. After that we will discuss system architecture of recommender system.

Keywords: Recommendation system, Personalization, Preprocessing

I. INTRODUCTION

Technological improvement has led to an explosive growth of recorded information, with the Web being a huge storehouse under no editorial control. Here, providing people with access to more information is not the problem; the problem is that more and more people navigate through large and complicated Web structures, find it difficult to access or get the information they want. Personalization can be the solution to this problem; since its objective is to provide users with information they want or need, without having to search for it explicitly. We meet cases of personalization in use in e-commerce applications, in information portals, in search engines and e learning applications. Web personalization can be defined as any action that personalizes the Web experience to a particular user, or a set of users. The experience can be something as casual as browsing a Website or as significant as trading stocks. Principal components of Web personalization include modeling of Web objects and subjects, categorization of objects and subjects, matching between and across objects and/or subjects, and determination of the set of actions to be recommended for personalization. Web mining is one of the main applications of data mining, Artificial Intelligence and so on to the Web data and it also forecasts the user's visiting behaviors and obtains his interests by investigating the samples [1].

The actions can range from simply making the presentation more pleasing and providing customized information. The ability of a web application to offer personalized content and to adapt is determined by its ability to anticipate users' needs and to provide them with the information and content they need. Adaptive web applications can do this only after analyzing data resulted from the users' current and former interaction with the system. Based upon the similarities discovered between different types of content and different user groups, one can make a series of recommendations enhancing the capacity of adaptation and personalization of web applications. Personalization systems based upon the user's surfing behavior analysis imply three phases, data collection and preparation, pattern discovery and content recommendation. Thus, a new research branch, called Web Usage Mining came into being, its goal being

* Research Scholar SCSVMV University Kancheepuram Tamil Nadu, India. **Email:** umarunn@gmail.com
** Assistant Professor, Dept. of IT, C.Abdul Hakeem College of Engineering and Technology, Melvisharam, Tamilnadu, India. **Email:**bakiyas90@gmail.com
*** Retired Senior Professor Anna University, Chennai, India **Email:** profsks@rediffmail.com

that of discovering useful information and knowledge as a result of analyzing these interactions. The WUM techniques use data extracted from log-files and provide information about activities undertaken by users during surfing sessions. In order to discover new useful information, WUM applies a series of diverse techniques, like classification, clustering, discovery of association rules or sequential patterns. In our research, we have used the technique of association rules, in order to discover correlations between the pages of a web application, based upon the analysis of the user's surfing sessions. Our efforts were channeled towards finding an efficient solution for implementing a recommendation [2] system within a web application capable to synthesize and to store only those data that are relevant for the recommendation process and a with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleansing differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at entry time, rather than on batches of data.

The actual process of data cleansing may involve removing typographical errors or validating and correcting values against a known list of entities. The results yielded by Web usage mining techniques can give guidelines for improving the design of Web applications [5].

Some data cleansing solutions will clean data by cross checking with a validated data set. Also data enhancement, where data is made more complete by adding related information, is a common data cleansing practice. For example, appending addresses with phone numbers related to that address. The interaction details of users with websites are recorded automatically in web servers in the form of weblogs [6].Web logs can be utilized in user profiling and similar image retrieval by tracing the visitor's on line behaviors for effective web usage mining [7].

Data cleansing may also involve activities like, harmonization of data, and standardization of data. For example, harmonization of short codes to actual words. Standardization of data is a means of changing a reference data set to a new standard, ex, use of standard codes.

(i) Data editing is defined as the process involving the review and adjustment of collected survey data. The purpose is to control the quality of the collected data editing can be performed manually, with the assistance of a computer or a combination of both.

(ii) Data reduction is the transformation of numerical or alphabetical digital information derived empirically or experimentally into a corrected, ordered, and simplified form. The basic concept is the reduction of multitudinous amounts of data down to the meaningful parts.

When information is derived from instrument readings there may also be a transformation from analog to digital form. When the data are already in digital form the 'reduction' of the data typically involves some editing, scaling, coding, sorting, collating, and producing tabular summaries. When the observations are discrete but the underlying phenomenon is continuous then smoothing and interpolation are often needed. Often the data reduction is undertaken in the presence of reading or measurement errors. Some idea of the nature of these errors is needed before the most likely value may be determined.

An example in astronomy is the data reduction in the kepler satellite. This satellite records 95-megapixel images once every six seconds, generating tens of megabytes of data per second, which is orders of magnitudes more than the downlink bandwidth of 550 KBps. The on-board data reduction encompasses co-adding the raw frames for thirty minutes, reducing the bandwidth by a factor of 300. Furthermore, interesting targets are pre-selected and only the relevant pixels are processed, which is 6% of the total. This reduced data is then sent to Earth where it is processed further.

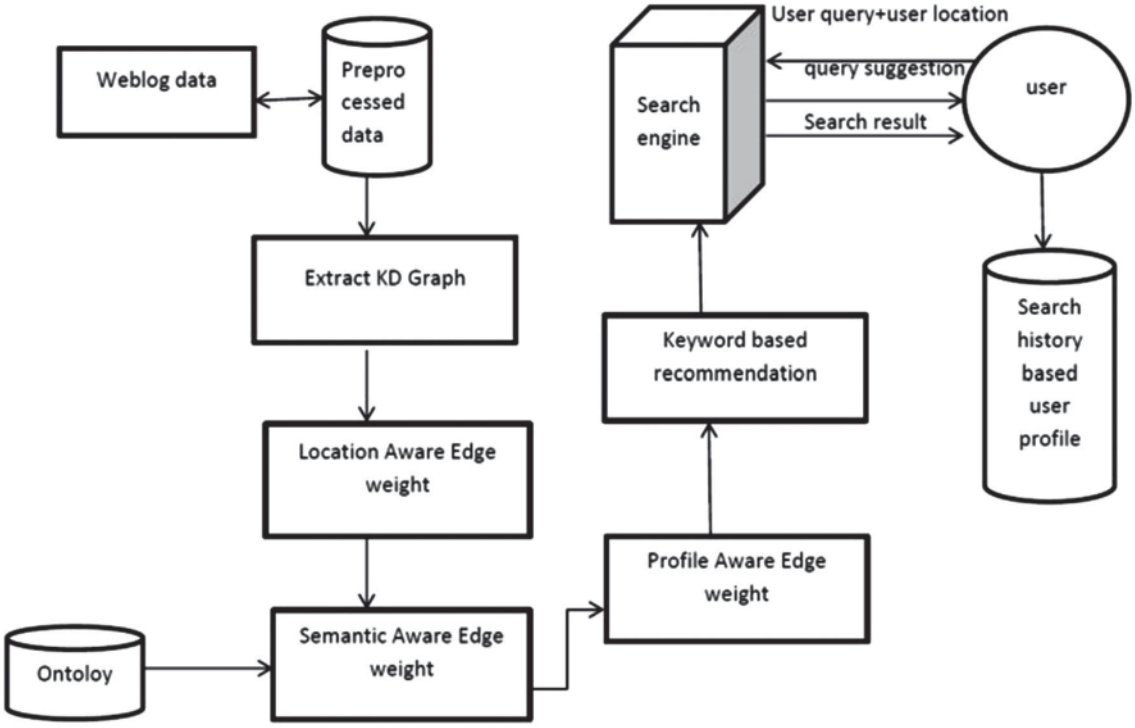


Figure 1. Architecture diagram for web page recommendation system

1.1 Extract KD Graph

KD graph is called as Keyword document graph. The keyword document graph is an edge weighted graph. It holds the keyword and its related document. Basically the KD graph is created with keyword at one end and the related documents at the other end. For same keyword there are 'n' numbers of documents.

The weight is generated based on the keyword relevancy to the document in particular location. For each and every document the edge weight[3] is calculated by user clicks, without user interaction there will be no edge weight assigned to the keyword.

So initially we assign weight based on keyword frequency in document. The weight of edge is same and equal to the number of clicks on document given keyword query. Therefore, the direct relevance between a keyword query and a clicked document is captured by the edge weight.

A. Location Aware Edge weight

After generating the keyword document graph, the location based edge weight is calculated. Location aware keyword relevancy is extracted by getting clicks of the user in particular location over the document. Initially the user completes the registration by proving their user id, password and the location. Then the user logged in with their authorized user id and password.

The user may search for any details in the search engine depending upon their need. The information collected from the tracking tools is stored in the log files [8].The location aware query suggestion gives priorities to the documents which depend to the particular location. This location aware edge weight is generated only based on the number of clicks done by the user at their location registered initially. In the keyword document graph there are 'n' numbers of documents for a keyword with their edge weight. There are three phases in Web usage mining: (i) Preprocessing does a series of activities such as data cleaning, user identification, session identification, path completion and transaction identification [4].

The weight of particular keyword and its related document edge increases when the click or selection of document for particular keyword increases for particular location made by the user. The location based edge weight is not calculated without the user interaction. So we have graph which holds different weight edge for different location for same document and same keyword. So the calculated location aware edge is added to the keyword document graph.

B. Semantic Aware Edge Weight

Semantic search is used to generate more relevant results to the user. It improves the search accuracy by understanding the searcher's intent and the contextual meaning of the keywords which the user types in the search engine. Semantic search system consider various points including context of search, location, intent, and variation of words, synonyms, generalized and specialized queries, concept matching and natural language queries to provide relevant search results.

The semantic aware edge weight module optimizes our graph by reducing the keyword document graph by finding their semantic relationship. For this, the search based on semantic relationship between the keywords is also included along with location based search. Keywords with semantic relationship with another keyword are merges to one keyword and their edge weight are added together.

For a single keyword there are many semantically related keywords. So a separate database called ontology database is created. In this database the semantic relation between the keywords is included. So the user typed keyword is compared with the keyword in our semantic database to find the semantic relationship.

In the keyword document graph the location aware edge weight and the semantic aware edge weight are added together. This provides more relevancies to the user's search result.

C. User profile Aware Edge Weight

The user profile aware edge weight module considers the user's search history from their profile. A registration form with user name, email id, password along with the user's location is created. For every user separate user profile is created. They can log in to their own profile using the user name and password. Without the help of users' content ratings, user sessions are clustered on a semantic level by Yilmaz H. and Senkul P [10] to know the different behavioral groups. Every cluster refers to a behavior group instead of simple data groups. The user profile is a user detail created based on their previous search history which may explain their search behavior. From user's search profile we can extract keyword weight based on his/her preferences which may lead to more relevant results. This edge weight is called as user profile aware edge weight.

The keyword document graph edge weight will be updated based on this relevancy during runtime. Because the keyword document graph made for individual so we can change the graph completely before and storing this graph for every users need extra space.

So the user can search their needs in the search engine only when they are logged in to their profile.

D. Keyword query suggestion

The keyword query suggestion module provides suggestions to user by adding all the edge weights calculated previously.

When a user login to our system we extract their user profile and extract their keyword and frequency. So the location aware edge weight, semantic aware edge weight are added along with user profile

aware edge weight and our system extract query keyword and auto suggest query while user types the keyword in the search textbox based on extracted weights. Profile based weight works only when a user is logged in, and it is processed during the runtime without affecting the keyword document graph

In recent days, the web usage mining has great importance and is often employed for the tasks like Web personalization, Web pages pre-fetching Website reorganization, etc [3].

Finally the query suggestion based on user's location, user's search history and the semantic relation between the keywords provides efficient and the relevant suggestions to the user.

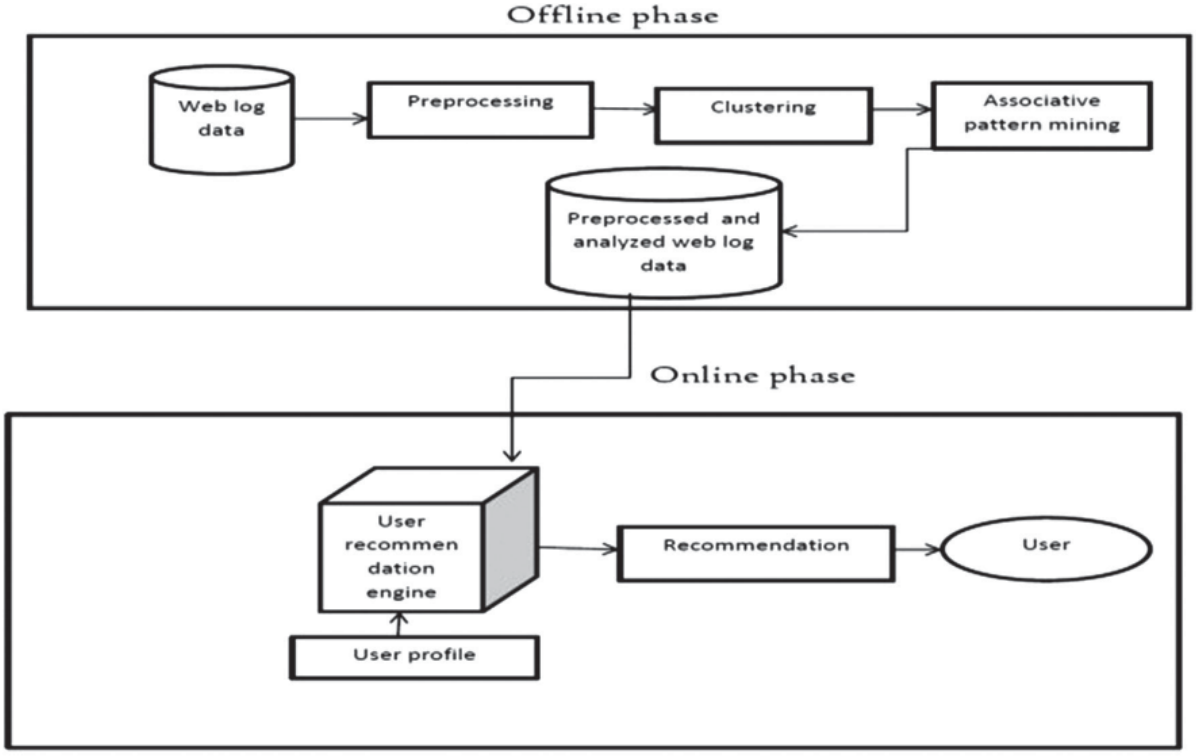


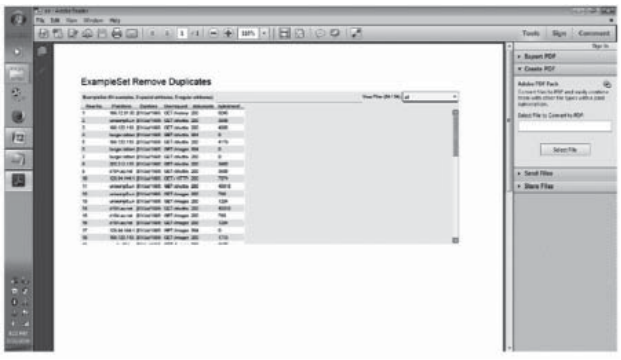
Figure 2. Overall architecture for proposed system

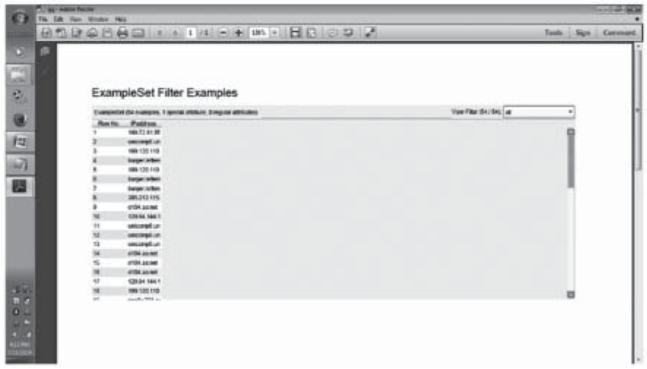
Offline phase: Web log data file is given as input and this web log data is taken for preprocessing[4] techniques using the rapid miner tool. Based in the preprocessed weblog data clustering has to be done. After that associative pattern mining are done by using the de blog analyzer tool. Finally in the offline phase preprocessed analyzed weblog data are stored in the database. ***Online phase:***

In this phase by using the preprocessed analyzed weblog data based on the user search efficient recommendation are given to the user through the user recommendation engine.

2. PERFORMANCE AND RESULT ANALYSIS

Weblog data server logs is increased. This proposed work can be enhanced for improving the quality aspects for effective web pages recommendation to the users. Also, the analysis of navigational patterns of users interacting with one or more web sites is considered as an important task to solve the problem of finding desirable, required and accurate information from the web due to the reasons of low precision & recall.





In this snapshot all the duplicates are removed by using the analyzer tool and filtering can be done by using the rapid miner tool.

3. CONCLUSION

Log files usually contain noisy and irrelevant data. So, preprocessing is required for removal of unnecessary data from the log file and formats the data to be ready for further processing. Pattern discovery techniques such as association rule mining and clustering are applied on the reduced log file. The proposed recommendation system will be useful for recommending the most suitable and relevant URLs to the web users according to their needs and preferences. The results obtained show a significant improvement in the effectiveness of the recommendation of the proposed system. The quality of an algorithm used for web page recommendations has been decided on the basis of its efficiency. The efficiency of an association rule mining technique when used for discovery of patterns from web.

References

1. Nithya P. and Sumathi P. (2012) 'An enhanced pre-processing technique for web log mining by removing web robots', Proceedings of IEEE International conference on computational intelligence and computing research., pp 1-4.
2. Hussain T, Asghar S. and Masood N. (2010) 'Web usage mining: a survey on preprocessing of web log file', Proceedings of IEEE International conference on information and emerging technologies., pp 1-6.
3. Sumathi C.P, Padmaja Valli R. and Santhanam T. (2010) 'Automatic recommendation of web pages in web usage mining', International journal on Computer Science and Engineering., Vol 2(9), pp 46-52.
4. Maheswara Rao V.V.R. and Valli Kumari V. (2011) 'An enhanced pre processing research framework for web log data using a learning algorithm', Netcom., pp 01–15.
5. Narendra Sharma, Aman Bajpai and Ratnesh Litoriya (2012) 'Comparison of the various clustering algorithms of Weka tools', International journal of emerging technology and advanced engineering., Vol 2, pp 73-80.
6. Sanjay Babu Thakare, Sangram Z. and Gawali (2010) 'A effective and complete preprocessing for web usage mining', International journal on Computer Science and Engineering., Vol 02, pp 848-851.

7. Suguna R. and Sharmila D. (2013) 'User interest level based preprocessing algorithms using web usage mining', International journal on Computer Science and Engineering., Vol 5(9), pp 815-822.

8. Sadhna Mishra K, Vineet Richaria and Vivek Sharma. (2013) 'Recognition of interested Web users behavior', International journal of Computer applications., Vol 61, pp 14-17.

9. Mahmoud Naghibzadeh and Mehrdad Jalali. (2012) 'Web page recommendation based on semantic web usage mining', Springer., Vol 7710, pp 293-405.

10. Yilmaz H. and Senkul P. (2010) 'Using ontology and sequence information for extracting behavior patterns from web navigation logs', Proceedings of IEEE International conference on data mining workshops.,pp549–556.

11. Ganesan S, Sivaneri A.I.U. and Selvaraju S. (2014) 'Interest based user groups using PSO algorithm', Proceedings of IEEE International conference on recent trends in Information Technology., pp 1–6.

12. Poorna Latha G. and Raghavendra Prakash S. (2011) 'Clustering web page sessions using sequence alignment method', Springer., Vol 250, pp 479-483.