International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 16 (2015) pp36851-36855

Prediction Of Web Users Interest On Buying Behaviour For E-Commerce Solutions

B.Uma Maheswari
Reseach Scholar in Bharathiar University,
Coimbatore, TamilNadu, India
umasharan7@gmail.com

Dr. P.Sumathi
Assistant Professor,Govt. Arts College,
Coimbatore, TamilNadu, India
sumathirajes@hotmail.com

ABSTRACT :Web mining is used to extract the content of web site where web usage mining helps to define user pattern. But using this abundant and ambiguous in most efficient manner in useful decision making is still a big challenge. During our web surfing either it is online shopping or blogging or using tweets and chatting everything is recorded. Web servers and log files are used to collect all the activity information of a user accesses to a web server. Web usage mining consists of three phases:- Preprocessing, Pattern discovery and Pattern analysis. It is a fact that normal log files data is very huge, noisy, unclear and confusing with lots of redundancy. It is very important to preprocess the log data for efficient web usage mining process. Preprocessing results also influences the later phases of web usage mining. This makes the preprocessing of server log files a significant step in web usage mining. This research works on data preprocessing of web log file. The model enables the administrator to access the web log file and perform data preprocessing on it i.e. data cleaning on web log file and identifying user navigational pattern. By using classification algorithm we identify user interested web site.

Keywords: Web usage mining, Preprocessing, SVM classifier, ELM classifier, Data cleaning

1.INTRODUCTION

The World Wide Web is a repository of web pages that provides the lot of information to the internet users. For internet users the information available on web has become a vital source. Because of these reasons, there is increasing growth and complexity of websites available on internet, the size of web is large. A web site is the link the customer to company. The companies can study visitor's activities through web analysis, and find the patterns. Web mining is broadly defined as discovering and analysis of useful information from the World Wide Web. Web mining divided into three parts: web contents mining, web structure mining and web usage mining. Web Contents Mining can be as the automatic search and extraction of information and resources available from number of sites and on-line databases though search engines or web spiders. Web Usage Mining can be as the discovery and analysis of access patterns of user, through the mining of log files. The output of the WUM can be used in web personalization, improving the system performance, site modification, usage characterization etc

Web usage mining comprises of three phases:

- Preprocessing
- Pattern discovery
- Pattern Analysis

The data stored in the log files don't present an accurate picture of the user's accesses to the web server. Data preprocessing is the process to convert the raw data available in log files into the database tables for making it suitable for applying the data mining algorithm. Hence preprocessing of web log data is most essential and a pre-requisite phase before it can be used for the pattern discovery task. Due to large amount of irrelevant entries in the web log file, the original log files cannot be directly used in the web usage mining process. Therefore the preprocessing of web log file becomes significant and important. The research on data preprocessing of web usage mining is a field in focus nowadays. This paper attempts to present the process of data preprocessing in web usage mining

Web log file is a server log file which is a basic data sources in web usage mining, in which it contain - access logs of the web server. The important task in the web usage mining is data preprocessing phase. It consists of data cleaning, user identification, session identification, path completion. Data preprocessing is used to clean the irrelevant data from log file so it can be provide to the pattern discovery to identify the user pattern.

So, in this work explained the complete roadmap for the preprocessing stage of web usage mining for determining user behavior. For this, we have reviewed the work done in the preprocessing stage by various researchers in 2; we have given an overview of data preprocessing and related work in this area. In 3, define the research methodology. In 4, the complete results for preprocessing of web log files and detailed analysis of work done by various researchers is presented, In 5 the conclusion of the work.

2. LITERATURE SURVEY

Carmona et al (2012) presents the methodology used in an e-commerce website of extra virgin olive oil sale called www.OrOliveSur.com. They described the set of phases carried out including data collection, data preprocessing, extraction and analysis of knowledge. A complete preprocessing technique is being proposed by

Hussain et al (2010) to preprocess the web log for extraction of user patterns. Proposed hierarchical sessionization algorithm generates the hierarchy of sessions. We obtain unbiased hierarchical clusters from the web log file.Chandra,a et al (works on data preprocessing of web log file. The model enables the administrator to access the web log file and perform data preprocessing on it i.e. data cleaning on web log file and identifying user navigational pattern. By using classification algorithm we identify user interested web site. Huang et al (2012) shows that both LS-SVM and PSVM can be simplified further and a unified learning framework of LS-SVM, PSVM, and other regularization algorithms referred to extreme learning machine (ELM) can be built. Grace et al (2011) gives a detailed discussion about these log files, their formats, their creation, access procedures, their uses, various algorithms used and the additional parameters that can be used in the log files which in turn gives way to an effective mining.

Hayatiet al (2010) describes an automated supervised machine learning solution which utilizes web navigation behavior to detect web spam bots. They proposed a new feature set (referred to as an action set) as a representation of user behavior to differentiate web spam bots from human users. Hayati et al (2010) referred to these web robots as spam bots that are capable of performing human tasks such as registering user accounts as well as browsing and posting content. Conventional content-based and link-based techniques are not effective in detecting and preventing web spam bots as their focus is on spam content identification rather than spam bots detection. Psaromiligkoset al (2011) described a new approach supported by a tool for analyzing learners' behavior in LMSs. Finally, described initial results arising from the use of the tool in two undergraduate courses. Priya and Vadivel (2012) build a tree using both frequent as well as non-frequent items and named as Revised PLWAP with Non-frequent Items RePLNI-tree in single scan. While mining sequential patterns, the links related to the non-frequent items are virtually discarded. Awadet al (2012) analyzed and studied Markov model and all- Kth Markov model in Web prediction. They proposed a new modified Markov model to alleviate the issue of scalability in the number of paths. In addition, they presented a new two-tier prediction framework that creates an example classifier EC, based on the training examples and the generated classifiers.

To help improve our modeling and understanding of this diversity, Kim et al (2012)applied automatic text classifiers, based on reading difficulty and topic prediction, to estimate a novel type of profile for important entities in Web search -- users, websites, and queries. Graepel (2010) described a new Bayesian click-through rate (CTR) prediction algorithm used for Sponsored Search in Microsoft's Bing search engine. The algorithm is based on a pro bit regression model that maps discrete or real-valued input features to probabilities. Web search engines have stored in their logs information about users since they started to operate. This information often serves many purposes. The primary focus of Silvestriet al (2010) is on introducing to the discipline of query mining by showing its foundations and by analyzing the basic algorithms and techniques that are used to extract useful knowledge from this (potentially) infinite source of information. Sharmaet al (2011) presents study about how to extract the useful information on the web and also give the superficial knowledge and comparison about data mining. They also described the current, past and future of web mining. Yang(2010) proposed a simple, yet powerful approach to profile users' web browsing behavior for the purpose of user identification. The importance of being able to identify users can be significant given a wide variety of applications in electronic commerce, such as product recommendation, personalized advertising, etc.

3. RESEARCH METHOFOLOGY

For identifying the user navigational pattern we proposed work for an efficient method of data preprocessing for data cleaning and identifying user navigational pattern has consist of different approaches and methods. To get better results of user category we used classification in this model.

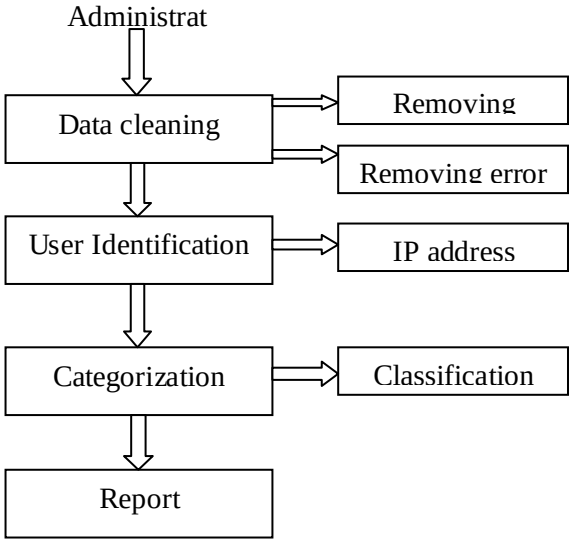

Fig. 1: System architecture

Here proxy server log file is taken as a data source which is collage proxy server log file as shown in fig.1. Our application is implementing for collage purpose to know the top most visited site and user navigational pattern. Admin is main user of our application. Firstly data preprocessing is done on log file and then categorized the file for generating user navigational pattern.

3.1 Data Preprocessing

The data cleaning process is used here to removing the unnecessary and duplicate data from log file. The objective of WUM is to have a clear picture of the web user requirement or behavior; hence it required the removing of

files having the suffixes such as, jpeg. jpg, gif, css, cgi, etc, error codes like 401,404 are not relevant and not useful for mining.

3.2 Classification

After preprocessing web log file the text file generated, which is used for categorization. Classification is used for categorization. Categorization implements the objects that are grouped into categories for some specific purpose. Classification is the task of mapping a data item into one of several predefined classes. In the Web domain, one is interested in developing a profile of users belonging to a particular class or category. This requires extraction and selection of features that best describe the properties of given the class or category. Classification can be done by using supervised learning algorithms such as decision trees, naive Bayesian classifiers, k-nearest neighbor classifiers, and Support Vector Machines. Classification techniques play an important role in Web analytics applications for modeling the users according to various predefined metrics. For example, given a set of user transactions, the sum of purchases made by each user within a specified period of time can be computed. A classification model can then be built based on this enriched data in order to classify users into those Who have a high propensity to buy and those who do not, taking into account features such as users' demographic attributes, as well their navigational activities.

3.3 Extreme Learning Machine

Extreme Learning Machine (ELM) meant for Single Hidden Layer Feed-Forward Neural Networks (SLFNs) will randomly select the input weights and analytically determine the output weights of SLFNs. This algorithm tends to afford the best generalization performance at extremely fast learning speed. ELM contains an input layer, hidden layer and an output layer.

$$\beta = H^{\dagger}T \quad (5)$$

ELM was originally proposed for the single hidden-layer feed forward neural networks and was then extended to the generalized SLFNs where the hidden layer need not be neuron alike. In ELM, the hidden layer need not be tuned. The output function of ELM for generalized SLFNs (take one output node case as an example) is

$$f_L(x) = \sum_{i=1}^{L} \beta_i h_i(x) = h(x)\beta \quad (1)$$

where β=[β1,…, βL]T is the vector of the output weights between the hidden layer of L nodes and the output node and h(x) = [h1(x), . . . , hL(x)] is the output (row) vector of the hidden layer with respect to the input x. h(x) actually maps the data from the d-dimensional input space to the L-dimensional hidden-layer feature space (ELM feature space) H, and thus, h(x) is indeed a feature mapping. For the binary classification applications, the decision function of ELM is

$$f_L(x) = sign(h(x)\beta) \quad (2)$$

Different from traditional learning algorithms, ELM tends to reach not only the smallest training error but also the smallest norm of output weights. According to Bartlett's theory, for feed forward neural networks reaching smaller training error, the smaller the norms of weights are the better generalization performance the networks tend to have. We conjecture that this may be true to the generalized SLFNs where the hidden layer may not be neuron alike. ELM is to minimize the training error as well as the norm of the output weights

$$Minmize: \|H\beta - T\|^2 and \|\beta\| \quad (3)$$

Where H is the hidden layer output matrix

$$H = \begin{bmatrix} h(x_1) \\ \vdots \\ h(x_N) \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} h_1(x_1) & \cdots & h_L(x_1) \\ \vdots & \vdots & \vdots \\ h_1(x_N) & \vdots & h_L(x_N) \end{bmatrix} \end{bmatrix} \quad (4)$$

To minimize the norm of the output weights $\|\beta\|$ is actually to maximize the distance of the separatingmargins of the two different classes in the ELM feature space: $2/\|\beta\|$

The minimal norm least square method instead of the standard optimization method was used in the original implementation of ELM

Where$H^{\dagger}$is the Moore–Penrose generalized inverse of matrixH. Different methods can be used to calculate the Moore–Penrose generalized inverse of a matrix: orthogonal projection method, orthogonalization method, iterative method and singular value decomposition (SVD). The orthogonal projection method can be used in two cases: when HTH is nonsingular and H† = (HTH)−1HT, or when HHT is nonsingular and H† = HT(HHT)−1.According to the ridge regression theory, one can add a positive value to the diagonal of HTH or HHT; the resultant solution is stable and tends to have better generalization performance.

4. EXPERIMENTAL RESULTS

In this section describe the efficient results for our time varying queries present in the ELM. It evaluates the concurrent results for every customers present in the data base. Experimental result is implemented by using MATLAB.

Table 1: Attributes used for e commerce

Attributes	Description
A1	Look for product offers
A2	Price details
A3	Read sub pages
A4	Product benefits
A5	Visit all pages for few minutes
A6	visit web page regularly

The characteristics of the attributes are discussed in table 1. It contains six attributes to find the interested and non interested buyers.

Table 2:Rate for interested and non-interested customer

Customer Class	Attributes	Rate
With purchase interest	A1	0,0,1,0,1,1
	A2	1,1,1,1,1,0
	A3	0,2,1,3,0,1
	A4	1,0,1,0,1,1
Without purchase interest	A5	0,1,0,0,1,0
	A6	0,0,0,0,1,0

Table 2 shows the values for with purchasing and without purchasing interest. The interested buyers may have maximum value of above one for all the attributes, but non interest buyers have maximum value as zero. Table 2 taken a six customer from huge amount of customers and show the result.

Number of records for training	Classifier Rate (%)				
	Subsequences length	With purchase interest		Without purchase interest	
		SVM	ELM	SVM	ELM
100	2	32.6	20.2	19.3	20.1
	3	35.8	33.5	21.7	28.6
	4	36.3	65.8	33.5	40.9
	5	32.8	68.1	41.1	52.8
	6	37.7	70.4	45.6	65.7
500	2	38.2	73.7	43.8	65.9
	3	35.7	71.3	45.2	66.2
	4	39.8	73.8	45.8	67.4
	5	41.3	75.6	46.5	70.5
	6	43.8	75.9	49.3	71.3
1000	2	44.6	69.2	50.8	73.8
	3	44.9	75.0	51.2	74.6
	4	45.6	79.5	55.3	77.5
	5	46.2	82.7	56.8	79.3
	6	46.8	85.6	57.4	81.6

Table 3: Classification Rate (Percent) of Different Classifiers in the customer reading interest using Different Subsequence Lengths

According to these data, ELM performs better than the SVM classifiers in terms of accuracy. The percentages of users correctly classified by ELM are higher to the results obtained and lower than the percentages obtained by SVM. In general, the difference between ELM and the SVM is considerable for small subsequence lengths, but this difference decreases when this length is longer. These results show that using an appropriate subsequence length, the proposed classifier can compete well with offline approaches. Nevertheless, the proposed environment needs a classifier able to process streaming data in online and in real time. In addition, the learning in ELM is performed in single pass and a significantly smaller memory is used. Spending too much time for training is clearly not adequate for this purpose. Then the number of attributes is very large in the proposed environment and it changes frequently, ELM is the most suitable alternative. Finally, the ELM structure is simple and interpretable.

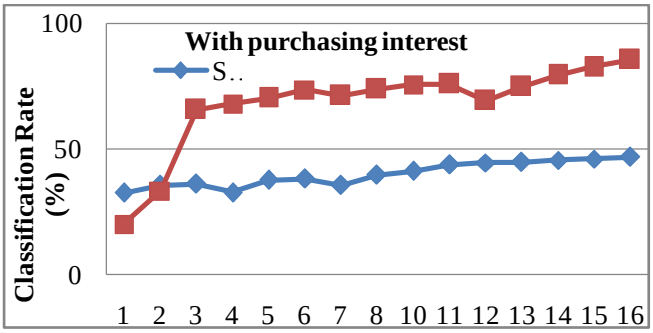

Figure 2: Classification rate for interested customer

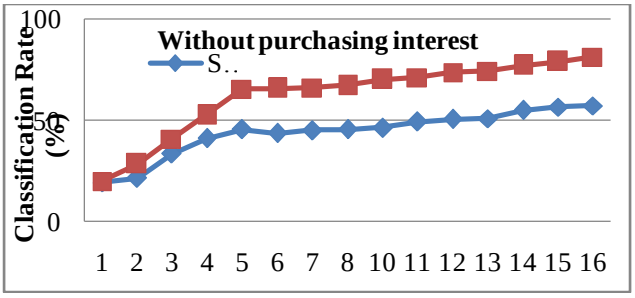

Figure 3: Classification rate for non-interested customer

Figure 2 and figure 3 show the interested and non-interested customers for purchasing in e-commerce. Thus the proposed method of ELM has better performance in classification rate when compare with the SVM.

5. CONCLUSION

To get user interested web site or user behavior data preprocessing for data cleaning and defining the user pattern is most useful. This model helps the for data preprocessing and categorizing the website so understand the user interested web site and per user need or request. By using ELM algorithm of classification it helps easily to categorize users visited web site and provide better efficiency and performance as compare to other algorithm. ELM is a learning mechanism for the generalized SLFNs, where learning is made without iterative tuning. The essence of ELM is that the hidden layer of the generalized SLFNs should not be tuned. As verified by the experimental results, compared to SVM and ELM achieves similar or better generalization performance for classification cases, and much better generalization performance for multiclass classification cases. ELM has better scalability and runs at much faster learning speed (up to thousands of times) than traditional SVM.

REFERENCES

[1] Carmona, Cristóbal J., S. Ramírez-Gallego, F. Torres, E. Bernal, M. Jose del Jesus, and Salvador García. "Web usage mining to improve the design of an e-commerce website: OrOliveSur. com." *Expert Systems with Applications* 39, no. 12 (2012): 11243-11249.

[2] Hussain, Tasawar, Sohail Asghar, and Nayyer Masood. "Hierarchical sessionization at preprocessing level of WUM based on swarm intelligence." In*Emerging Technologies (ICET), 2010 6th International Conference on*, pp. 21-26. IEEE, 2010.

[3] Chandrama, Wasvand, P. R. Devale, and Ravindra Murumkar. "Data Preprocessing Method of Web Usage Mining for Data Cleaning and Identifying User navigational Pattern."

[4] Huang, Guang-Bin, Hongming Zhou, Xiaojian Ding, and Rui Zhang. "Extreme learning machine for regression and multiclass classification." *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 42, no. 2 (2012): 513-529.

[5] Grace, L. K., V. Maheswari, and Dhinaharan Nagamalai. "Analysis of web logs and web user in web mining." *arXiv preprint arXiv:1101.5668* (2011).

[6] Hayati, Pedram, Vidyasagar Potdar, Kevin Chai, and Alex Talevski. "Web spambot detection based on web navigation behaviour." In *Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on*, pp. 797-803. IEEE, 2010.

[7] Hayati, Pedram, Kevin Chai, Vidyasagar Potdar, and Alex Talevski. "Behaviour-Based web spambot detection by utilising action time and action frequency." In*Computational Science and Its Applications–ICCSA 2010*, pp. 351-360. Springer Berlin Heidelberg, 2010.

[8] Psaromiligkos, Yannis, Maria Orfanidou, Christos Kytagias, and Evmorfia Zafiri. "Mining log data for the analysis of learners' behaviour in web-based learning management systems." *Operational Research* 11, no. 2 (2011): 187-200.

[9] Priya, Vishnu, and A. Vadivel. "User behaviour pattern mining from WebLog."*International Journal of Data Warehousing and Mining (IJDWM)* 8, no. 2 (2012): 1-22.

[10] Awad, Mamoun, and Issa Khalil. "Prediction of user's web-browsing behavior: Application of markov model." *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 42, no. 4 (2012): 1131-1142.

[11] Kim, Jin Young, Kevyn Collins-Thompson, Paul N. Bennett, and Susan T. Dumais. "Characterizing web content, user interests, and search behavior by reading level and topic." In *Proceedings of the fifth ACM international conference on Web search and data mining*, pp. 213-222. ACM, 2012.

[12] Graepel, Thore, Joaquin Q. Candela, Thomas Borchert, and Ralf Herbrich. "Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine." In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 13-20. 2010.

[13] Silvestri, Fabrizio. "Mining query logs: Turning search usage data into knowledge." *Foundations and Trends in Information Retrieval* 4, no. 1—2 (2010): 1-174.

[14] Sharma, Kavita, Gulshan Shrivastava, and Vikas Kumar. "Web mining: Today and tomorrow." In *Electronics Computer Technology (ICECT) 2011 3rd International Conference on*, vol. 1, pp. 399-403. IEEE, 2011