An Efficient Web Recommender System for Web Logs

B. M. Vidyavathi, PhD
Visveshvaraya Technological University Department of Computer Science, BITM
BITM
Ballari, India

Haseena Begum
Visveshvaraya Technological University Department of Computer Science,
Ballari, India

ABSTRACT
Nowadays web sites have become an important means for communication. Every application is converted to internet based application. A huge amount of the data is added daily and huge amount of information is accessed from the web. The millions to billions of web users are accessing the interested data from the web making the network traffic. To overcome these problems, recommender systems which are the part of machine learning are introduced. At present, recommender systems play an important role in the e-commerce field. The shopping sites where the users are recommended with the interested products and resources based on their navigation behavior, profile and their interest on the products. The recommender systems are basically divided into three types, they are content based, collaborative and hybrid recommender systems. As the web users are increasing the network performance is affected. Thus it is required to enhance the performance of the network by using the mining and machine learning classification techniques. In this paper, we propose an efficient web recommender system, which uses the concept of preprocessing, clustering technique and expectation maximization naïve bayes as predictive model.

Keywords
Web access logs, Preprocessing, K-Means clustering, Expectation Maximization Naïve Bayes Classifier.

1. INTRODUCTION
Day by day the technology is advanced, the web users are increasing in rapid rate. Every application is internet based application and web applications produce large amount of data. The web users browse the data using different browsers. In this web based field, providing the needed information for the web users have become an important issue. The navigation information of the web user and the interacted information of web server are stored as web server logs. The collected information is irrelevant, incomplete and noisy data. But for today's technology and usage of web, the prefetching and caching techniques are not alone sufficient to handle this data.

Recommender systems are playing a vital role in recommending the interested resources for web users. Recommender systems are classified with content based filtering, collaborative filtering and hybrid recommender systems. In content based filtering, it

recommends based on the navigation and profile information including the vectors. In collaborative filtering, recommendation is based on the prior knowledge of navigation and profile information. The vectors are estimated in collaborative filtering. The hybrid recommender systems are union of content based filtering and collaborative filtering method.

The popular applications of recommender system product recommendation in e-commerce fields such as Amazon, Netflix, Flip kart and many others sites. To enhance the performance of the network, we propose an efficient web recommender system. The data cleaning technique is used for preprocessing stage, clustering technique for grouping the web users and a predictive model expectation maximization naïve bayes for classification.

The rest of the paper is organized as follows. Section 2 introduces the related study. Section 3 explains about the proposed work. Section 4 relates with the experimental analysis. Finally, we conclude our paper in section 5.

2. RELATED WORK
Learning algorithms are basically classified into supervised, unsupervised learning and semi supervised learning. Supervised learning is presented with the inputs to their desired outputs. Unsupervised learning is the method where input prior knowledge is not given to learning algorithms. Semi supervised learning is between supervised learning and unsupervised learning.

A recommendation system is developed for cache replacement which utilizes the proxy access logs for the analysis of data. The prefetching techniques and the three different algorithms are applied [1]. The preprocessing algorithm is used to refine the web access logs, to generate a cleaned and transformed data. The combined approach of clustering and decision tree algorithm is introduced. The decision tree algorithm ID3 takes the clustered input data to form the decision rules. The KNN algorithm is used to predict the URL's which are navigated by the web user. The computational complexity is not much adoptable. The performance of the system can be increased by applying other classification techniques.

A framework for prediction of web requests of users is developed [2]. The prefetching scheme and web usage mining concept is utilized to improve the performance. The Apriori algorithm is used to create the rules for prefetching the pages. Web usage mining is the method of extracting the patterns from the web server logs. Web server logs give the complete navigational information of the web user [3]. The data which is extracted from the server represents inconsistent data. Preprocessing is the method which filters the noisy web log data to consistent data. Preprocessing is comprehended with data cleaning, user's identification, session identification and path completion.

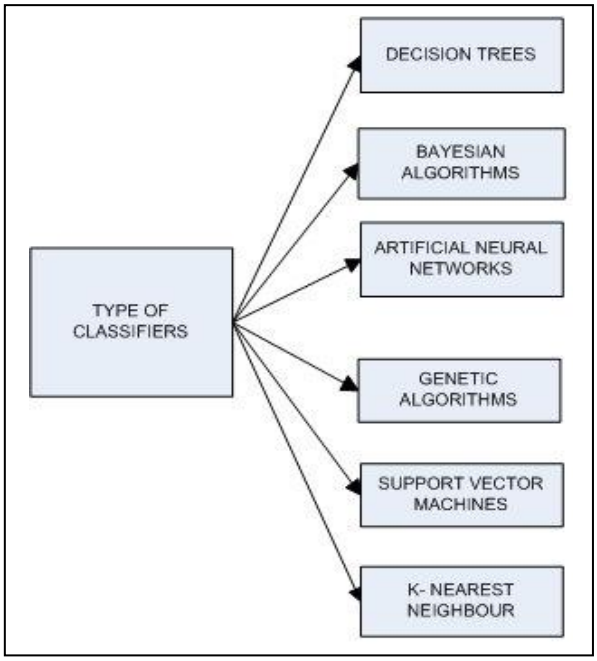

Fig 1: Type of Classifiers

Web usage mining is the mining technique which extracts the actual behavior of the web users. There are different applications where the navigation behavior of the users is analyzed. The most popular applications are e-commerce, recommender systems and web site designing [4]. Web servers keep all the log entries information of the web user. New techniques for preprocessing have been used to increase the performance. The Distinct user identification method along with time complexity is used to identify the distinct users [5]. This method identifies the fraud detection, counter terrorism and the detection of unusual access of the secured information. The data is classified into the content, structure, usage and user profile. Web mining is depicted into three mining areas that are web content mining, web structure mining and web usage mining.

Web usage mining relates with discovering the interesting patterns from web browsing data. To increase the quality of the data, the preprocessing steps are carried [6]. Preprocessing provides reliable and concise data by filtering the unreliable data. There are many problems faced in the preprocessing stage such as the accuracy of session identification and user identification. The various web mining algorithms are used to identify the transaction identification and mainly focus to the preprocessing steps.

Web data classification has become area where the classification techniques have been used. Web proxy caching techniques are used to improve the performance of the proxy server. The concept of iterative approach Expectation Maximization Naive Bayes classifier is introduced [7]. This technique mainly concentrates on the prediction of the web pages. The cache replacement policies such as least recently used and greedy dual size frequency are used with the combination of the expectation maximization naïve bayes classifier.

A framework for classification algorithms using web usage mining is presented [8]. The features and the limitations for each of the classification technique are tabulated. The comparative study between the decision tree algorithm and Naive Bayesian Classification algorithm are used to identify the interested users. The enhanced version of the Naive Bayesian Classification is experimented with the decision tree algorithm C4.5 [9]. The comparison of algorithms that is mainly related to web mining is discussed [10]. The algorithms include clustering, classification, frequent itemset mining and sequence analysis algorithms. The web usage mining is the part of the data mining which analyzes the patterns of the web user.

3. PROPOSED WORK

Classification algorithms are widely used to predict and classify the given data. The system includes major modules such as the Preprocessing, K-Means Clustering and Expectation Maximization Naïve Bayes Classifier.

The navigation information of web user between the browser and web user is stored in servers as web logs. The collected data is noisy, inaccurate and incomplete. These web logs require some amount of cleaning. Each record contains the attributes like timestamp, number of bytes transmitted, IP address, TCP information, status code, and size of the file, request methods, URL, analytics and the type of the file. The required attributes are selected from the records and stored in database. This information is provided as the input the preprocessing module. Preprocessing includes the data cleaning algorithm. Input web access logs are irrelevant and noisy. It requires some filtering, to remove the data which is incomplete and inaccurate. Data cleaning is important part of preprocessing to improve the efficiency. Cleaning process is carried by removing the all the images, the audio and video files, the request methods other than GET and POST, the status code files other than 200 and even the robots are cleaned. The cleaned data is given input to the clustering module. Data cleaning gives a complete, accurate and relevant data.

The data which is the output of the data cleaning module is taken as input to clustering module. Here the clusters are formed, based on the IP address. It makes easier to find each user accessing the number of web pages and it is stored in the database. Clustering technique groups the web user accordingly to the number of accesses. The attributes such as the size of the webpage is taken from the web log information and clustered data is used to prepare the training data. The training dataset includes the time interval, frequency and the size of the web page. Each of these features is calculated.

EM-NB is learning module which is the fusion of both the supervised learning and unsupervised learning. EM-NB predicts the web pages that are cacheable or uncacheable. All the cacheable web pages are placed in cache. EM is an iterative method which iteratively compares the computed probability with current probability and Maximum likelihood hypothesis. EM-NB focuses on predicting the web pages that are cacheable or uncacheable based on the probability calculation.

4. EXPERIMENTAL ANALYSIS

Experimental analysis shows the comparison between expectation maximization naïve bayes classifier and naive bayes classifier. We received data from the web access logs which are downloaded from https://archieve.org. A web access logs generally consists of timestamp, number of bytes transmitted, IP address, TCP information, status code, and size of the file, request methods, URL, analytics and the type of the file. The log information is inconsistent and incomplete. It is transformed into structured format using preprocessing.

The cleaned web logs are clustered and the classification is based on expectation maximization naïve bayes classifier which is an iterative process. Naïve bayes classifier finds the posterior probability for each of the input data. In the experiments, for EM-NB the preprocessing and clustering techniques are applied before the classification of web pages as cacheable or uncacheable. The expectation maximization naïve bayes classifier computes the probability using the naïve bayes thereom. The web pages which have the highest probability, they are preferred as cacheable web pages.

4.1 Accuracy

The accuracy of the system is given as the amount of data that is recognized as cacheable during classification of input test data.

Accuracy = (Total number of classified cacheable records/ Total number of records)*100

Table 1. Accuracy

Dataset size	NB	EM-NB
100	69	77
300	72.3	77.33
500	76.8	86

The performance is evaluated in terms of accuracy in Table 1. The values of EM-NB are higher than NB classifier. For each dataset size in terms of number of records are given as input and the accuracy is evaluated in terms of percentage.

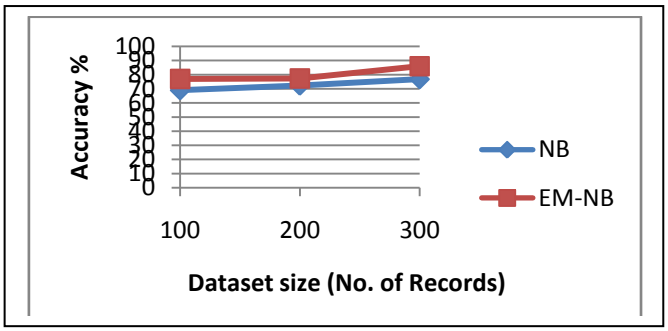

Fig 2: Accuracy

The comparative performance between both the techniques is shown in fig 2. The x-axis represents the dataset size, in terms of number of records and y-axis represents the accuracy in terms of percentage. The above graph depicts the accuracy of expectation maximization naïve bayes (EM-NB) is higher than naïve bayes (NB). The accuracy of the system is defined as the ratio of cacheable records to the total number of records.

4.2 Error rate

The error rate of the system is given as the amount of data that is not recognized during classification.

Error rate = (1 – accuracy)*100

Table 2. Error Rate

Dataset size	NB	EM-NB
100	31	23
300	27.7	22.67
500	23.2	14

The performance is evaluated in terms of error rate in Table 2. The values of EM-NB are lesser than NB classifier. For each dataset size, in terms of number of records are given as input and the error rate is evaluated in terms of percentage. The error rate defines the number of uncacheable records from the given input.

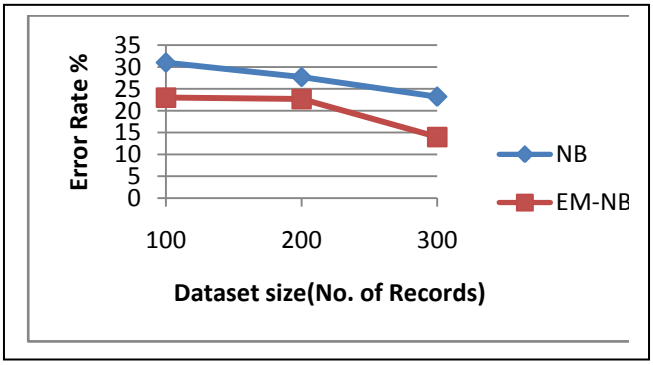

Fig 3: Error rate

The error rate between both the techniques is shown in fig 3. The x-axis represents the dataset size in terms of number of records and y-axis represents the error rate in terms of percentage. According to the results, the EM-NB classifier has less error rate than naïve bayes algorithm. Hence we conclude that EM-NB is more promising than naïve bayes classifier.

5. CONCLUSION

In today's life, the technology is advanced day by day. The numbers of web users are increased to use the web. Web users are facing many problems to access the interested web pages and resources from the web server. The pages are delayed due to network traffic. To overcome these problems, the combined approach using the data mining and the machine learning techniques of classification are used. Recommender systems are the part of the machine learning which help to recommend the interested resource to the web users. The preprocessing technique gives the output as consistent data, which improves the performance of the system. The K-Means clustering groups the web pages according to the IP address, which makes easier to identify the unique web users. EM-NB is used to predict the cacheable and uncacheable URL's from the given dataset. The output of the system is the recommended data which are web patterns. By placing the web URL's in the proxy server, the performance is improved. The experimental result shows that EM-NB is efficient, when compared to naïve bayes classification. In future the performance can be increased by using other classification techniques. The proposed system can be enhanced by applying other algorithms of preprocessing and improved clustering techniques.

6. REFERENCES

[1] Priyansha Bangar and Kedar Nath Singh. 2015. Investigation and Performance Improvement of Web Cache Recommender System, International Conference on Futuristic trend in Computational Analysis and Knowledge Management.

[2] Nanhay Singh, Arvind Panwar and Ram Shringar Raw. 2013. Enhancing the Performance of Web Proxy Server through Cluster Based Prefetching Techniques, International Conference on Advances in Computing, Communications and Informatic.

[3] Saritha Vemulapalli and M. Shashi. 2012. Design and Implementation of an Effective Web Server Log Preprocessing System, Proceedings of the InConINDIA Springer Verlag Berlin Heidelberg.

[4] V.Chitraa and Dr. Antony Selvdoss Davamani. 2010. A Survey on Preprocessing Methods for Web Usage Data, International Journal of Computer Science and Information Security.

[5] Sheetal A Raiyani and Shailendra jain. 2012. Efficient Preprocesing technique using Web log mining, International Journal of Advancements in Research & Technology.

[6] Preeti Gupta. 2014. Pre-Processsing E-Commerce Web Log Files for Web usage Mining, International Journal of Advanced Research in Computer Science and Software Engineering.

[7] P. Julian Benadit, F. Sagayaraj Francis and U. Muruganantham. 2015. Improving the Performance of a Proxy Cache Using Expectation Maximization with Naïve Bayes Classifier, Computational Intelligence in Data Mining, Springer India.

[8] Supreet Dhillon and Kamaljit Kaur. 2014. Comparative Study of Classification Algorithms for Web Usage Mining, International Journal of Advanced Research in Computer Science and Software Engineering.

[9] A. K. Santra and S. Jayasudha. 2012. Classification of Web Log Data to Identify Interested Users Using Naïve Bayesian Classification, International Journal of Computer Science Issues.

[10] Parth Suthar and Prof. Bhavesh Oza. 2015. A Survey of Web Usage Mining Techniques, International Journal of Computer Science and Information Technologies.