ISSN:0975-9646

Web Recommender System using EM-NB Classifier

Haseena Begum[1], Dr.B.M Vidyavathi[2]

Department of CSE, Visveswaraiah Technological University
BITM, Bellary, India

***Abstract*— Recommender system is an important application of machine learning. Over the last few years, the recommender systems have changed the way of communication between websites and web users. The popular websites like Amazon, Netflix, eBay and Apple try to recommend the products making the information search easier. The recommendation system sorts huge amounts of data to identify the interest of web users. There are two types of approaches to develop a recommendation system. They are Content-based filtering method and Collaborative filtering method. The internet based applications are also growing rapidly. Every user uses the web differently for different purposes. Due to this network performance is affected. Thus, it is required to find out from the web that, what kind of activities the users are doing and in which resource they are interested. This is done in the web usage mining with recommender system. In this paper, we propose to improve the performance of web proxy server using data cleaning algorithm for preprocessing, to improve the quality of the web log data and Expectation Maximization Naive Bayes classifier for prediction of widely used web pages.**

***Keywords*— Web usage Mining, Preprocessing, K-Means Clustering, Expectation Naive Bayes Classifier.**

I. INTRODUCTION

Nowadays everyone is using the web, the internet applications are growing tremendously, the data is increasing, to handle this data, data mining is used. Data mining deals with the extraction of actual data from the raw data. Web users use the web for different purposes, due to this usage of the web, the network traffic is getting increased and the performance is decreased. To handle this kind of problem we go for web usage mining with recommender system. A Web usage mining deals with the interested knowledge of the web logs from the web server. To improve performance of the network, the web caching and web prefetching schemes are used. But for today's web usage this is not sufficient alone to handle the network traffic, so the machine learning techniques are used.

Recommender systems can be defined as the system which recommends some information to users based on the activity, behavior and the prior knowledge or navigation history of the user. The most popular applications of recommender systems are movies, music, research articles, books, news, social tags and products. But, in today's the web recommender systems are playing a vital role.

Recommender systems have mainly three approaches. They are Collaborative Filtering, Content based Filtering and Hybrid recommender System. The Collaborative filtering approach requires prior knowledge of user's behavior and the activities. This approach estimates the feature vectors along with prediction. In Content based filtering the prediction is based on a profile of the user and a history of the user's interaction with the web and the feature vectors. The Hybrid approach is both the combination of the collaborative filtering and content based filtering. In this paper, we propose to improve the performance of web proxy server using data cleaning algorithm for preprocessing, to improve the quality of the web log data and Expectation Maximization Naïve Bayes classifier for prediction of widely used web pages.

The rest of the paper is organized as follows. Section II introduces the related study in preprocessing and classification algorithms. Section III explains about the proposed work. Finally, we conclude our paper in Section IV.

II. RELATED WORK

A data model, which utilizes the proxy access log for data analysis is modified by using K-Means algorithm which is applied for user data personalization [1]. The pre-fetching techniques for web URL's and a recommendation system is developed for cache replacement. The three different algorithms are applied in sequence. First, the input web access log of the proxy server is pre-processed and stored in an intermediate storage. Second, a K-means clustering algorithm is applied and IP address based clusters are produced. Third, user based clustered data is used to develop ID3 based decision rules and the current user navigation sequences based next URLs is predicted KNN algorithm. The performance of the proposed system is enhanced more by applying the other classification techniques. We extend their methodology by using Expectation Maximization Naïve Bayes Classifier.

The web usage mining and pre--fetching scheme were used to improve the performance of web proxy servers [2]. Pre-fetching fetches objects that are likely to be accessed in the near future and store them in advance, thus the response time of the user request is reduced. The Apriori algorithm is used to find the patterns and generate the rules for the pre-fetching.

Web usage mining deals with discovering usage patterns from server logs in order to understand and serve the needs of web users [3]. The raw data contains the irrelevant and noisy data, which requires preprocessing. Preprocessing

includes cleaning, user identification, session identification, path completion and structurization.

Web Usage Mining applies mining techniques in log data to extract the behavior of users which is used in various applications like adaptive web sites, customer profiling, personalized services, prefetching, creating attractive web sites etc. Web usage mining includes three phases preprocessing, pattern discovery and pattern analysis. This paper presents various works done by different researchers on preprocessing methods [4]. Log files are the best source to know user behavior. But the raw log files contain unnecessary detail like image access, videos, failed entries etc., which will affect the accuracy of pattern discovery and analysis. So preprocessing stage is an important work in mining to make efficient pattern analysis.
The Distinct user identification technique enhances the pre-processing steps of web log usage data in data mining [5]. The authors have used two pre-processing techniques to combine within one pre-processing step time of user identification and which finds out distinct user, based on their attended session time. DUI algorithm is very efficient as compared to other identification techniques. This algorithm improves the design of WebPages.

The data preparation techniques that preprocess web log files in order to identify unique users are proposed [6]. The User data session, which is part of preprocessing used to improve the performance features of web mining. However, many problems remain such as data collection, applications of some heuristics in some phases of data pre-processing and the accuracy of user identification and session identification.

The web proxy caching approaches, namely EMNB-LRU, and EMNB-GDSF for improving the operation of the conventional World Wide Web proxy caching are discussed [7]. The Expectation Maximization Naive Bayes classifier discovers from World Wide Web proxy log file to forecast the categories of objects to be revisited or not. The traditional Web proxy cache replacement policies such as LRU (Least Recently Used) and GDSF (Greedy Dual Size Frequency) are used to assimilate the semi supervised machine learning technique for raising the operation of the Web proxy cache. Most of the studies show that EMNB is better when compared to other classification techniques.
A frame for web usage mining based on classification algorithms, including their features and limitations [8]. The Naive Bayesian Classification algorithm for classifying the interested users and also a comparison study of using the enhanced version of the decision tree algorithm C4.5 and Naive Bayesian Classification algorithm for identifying interested users [9]. Web usage mining is a very useful area in terms of analysis of users and their behavior regarding the web contents. The authors discuss about the survey of web usage mining techniques [10]. This paper shows comparisons of algorithms that are related to frequent itemset mining algorithms, clustering algorithms, classification algorithm and finally the sequence analysis algorithms.

III. PROPOSED WORK

In the field of data mining, machine learning algorithms are being used to discover valuable knowledge from the large databases. Machine learning will play an important role in computer science and computer technology. Our proposed system uses the concept of web usage mining including the classification technique of machine learning. The below figure illustrates the proposed work.

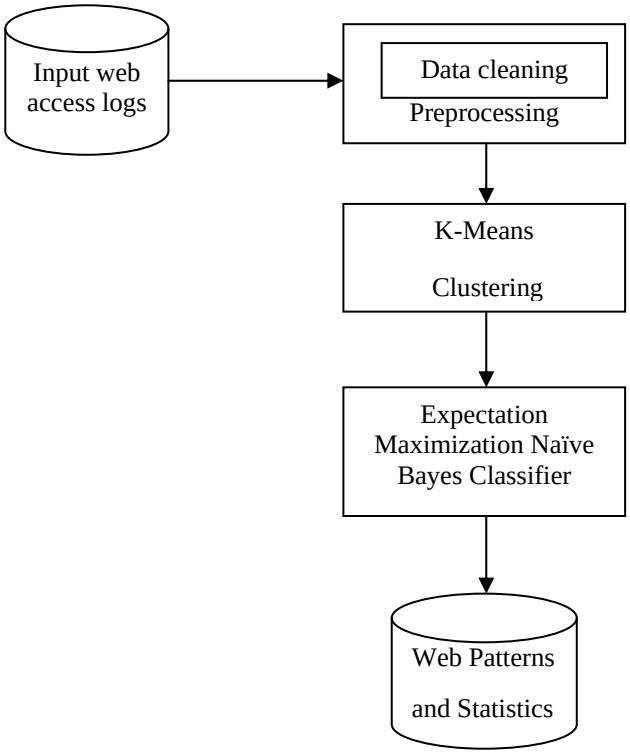

Fig 1: Proposed System

1. *Input web access logs:* These are the web logs extracted from the web server. Web logs are irrelevant, inconsistent, incomplete noisy data. This data is provided as the input the preprocessing module.
2. *Data cleaning:* This paper relates only with data cleaning algorithm in preprocessing. Data cleaning algorithm removes the image, audio, video, script, style sheet file, crawler request, error request and other than get or post request. This stage is effective to improve the performance of the system.
3. *K-Means Clustering:* The cleaned data is taken as input to the clustering module. The input data is clustered based on the IP source of the web user.
4. *Expectation Maximization Naive Bayes Classifier:* The Clustered data is provided as the input to the classifier. EM-NB is a semi supervised learning module which predicts the web pages that may be accessed by the web user in future. EM mainly focuses on the missing information for the current approximation and Maximum likelihood hypothesis. Naive Bayes is the most popular learning technique to estimate the probabilities and the prediction of web patterns.

5. *Web patterns and Statistics:* Web patterns are the web urls which are recommended to the web users. These urls are stored in proxy cache to reduce the network traffic. Web statistics mainly relates with the web analytics that is the extraction of busy time, popular sites and popular user.

A. Features

1. The proposed system utilizes the web usage mining with machine learning. The web logs are filtered using the data cleaning algorithm.
2. The cleaned data is clustered using K-Means clustering. An Expectation Maximization Naïve Bayes Classifier is used in the proposed system for prediction of web pages as recommended web urls to the web users.
3. Expectation Maximization Naïve Bayes Classifier performs better when compared to other classification techniques.

IV. CONCLUSION

Web sites are the most important means of communication where the tremendous amount of data is added daily. The millions of web users get accessed with the information which they actually need, facing with the problems in downloading the page, resulting increment in network traffic and response time. To handle this problem, in this paper, we proposed, the most efficient classification technique, Expectation Maximization Naive Bayes Classifier. Expectation Maximization Naïve Bayes Classifier is used to predict the web pages which are widely used. The recommended data that is the web patterns which is the output the EM-NB can be placed in the proxy cache which improves the performance of the web server. The predicted web pages can be accessed from the proxy server instead of accessing from web server, which in turn decreases the response time and increases the performance. The proposed system gives the efficient results when compared to other classifiers.

REFERENCES

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

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

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

[4]. V.Chitraa and Dr. Antony Selvdoss Davamani, *"A Survey on Preprocessing Methods for Web Usage Data"*, International Journal of Computer Science and Information Security, Volume 7, 2010.

[5]. Sheetal A Raiyani and Shailendra jain, *"Efficient Preprocesing technique using Web log mining"*, International Journal of Advancements in Research & Technology, Volume 1, 2012.

[6]. Preeti Gupta, *"Pre-Processsing E-Commerce Web Log Files for Web usage Mining"*, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, 2014.

[7]. P. Julian Benadit, F. Sagayaraj Francis and U. Muruganantham, *"Improving the Performance of a Proxy Cache Using Expectation Maximization with Naïve Bayes Classifier"*, Computational Intelligence in Data Mining, Volume 2, Springer India 2015.

[8]. Supreet Dhillon and Kamaljit Kaur, *"Comparative Study of Classification Algorithms for Web Usage Mining"*, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, 2014.

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

[10]. Parth Suthar and Prof. Bhavesh Oza, *"A Survey of Web Usage Mining Techniques"*, International Journal of Computer Science and Information Technologies, Volume 6, 2015.