AN EFFICIENT MINING TECHNIQUE FOR WEB CACHE OF SERVER LOG FILES

K. Suguna[1], K. Nandhini[2]
[1]Research Scholar, Bharathiar University,
Assistant Professor, Department of Computer Applications,
Dr.N.G.P Arts and Science College, Coimbatore-641046, India.
[2]Assistant Professor, Department of Computer Science,
Chikkanna Government Arts College, Tirupur-641602, India.

ABSTRACT

Data mining contemplates on non-trivial extraction of implicit previously unknown and useful information from the huge amount of data. Web mining is the one of the application of data mining which becomes a significant area of research due to huge amount of World Wide Web services in recent years. The World Wide Web is an interactive and popular way to transfer information. The prefetching approaches based on data mining techniques are popularly applied on caching problems. The data mining techniques classification and clustering are applied to find the interesting unknown data patterns in large data sets. This approach is used to analyze web access pattern of user by using information present in the proxy server log files. The frequently used web pages by the users are identified and integrated with the prefetching scheme. The web caching is used to achieve performance improvement for the cache proxy server. The most important problem of predicting a user's behavior on web sites is gained importance due to the rapid growth of the World Wide Web and the basic to personalize user's browsing history.

KEYWORDS- *Data mining, Web caching, K-means, Web logs, B-CART.*

I. INTRODUCTION

Data mining is an availability of huge amount of data and the need for resolving such data into useful information and knowledge. Web is a magnificent source of data and a large number of internet users access the web to find data. The web is a way of accessing information over the medium of the Internet.

It is an information-sharing model that is built on top of the Internet. Web mining can be classified into three areas: content mining, structure mining and usage mining. Web content mining extracts useful information or knowledge from the contents of Web pages. Web structure mining discovers useful knowledge from hyperlinks that portrays the structure of the Web.

Web usage mining refers to the discovery of user access patterns from Web server's logs, which keeps record of every click made by each user. In general, Web usage mining consists of three processes: Data pre-processing, patterns discovery and patterns analysis.

Web usage mining is used to make efficient data access for the internet users in various web caching and pre-fetching techniques are applied. These techniques are applied on data, which is needed to distribute.

Web servers provides two diverse types of data,

1. Static data are kept in files at a server side.
2. Dynamic data which are constructed by programs at runtime.

The program is executed when a user makes request for a particular content.

The dynamic data makes the web sites very slow, because of data is organized on request basis. A web cache is a temporary storage of web documents. A web cache stores copies of web documents transient through it. The web architecture is categorized by

i) Web clients: The software employed by users to access the web
ii) Web servers: It contains the information what the user requests.

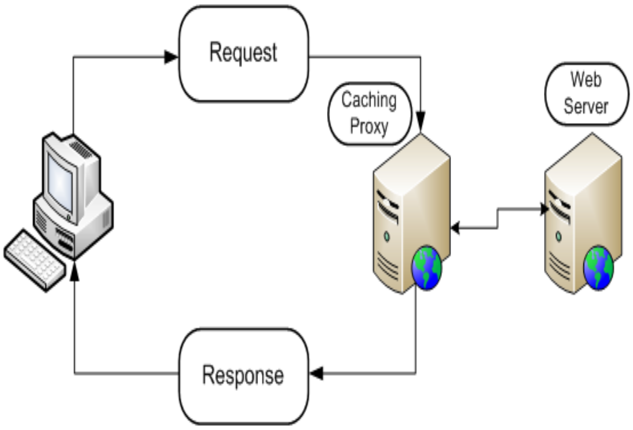

Figure 1.1. Web architecture

The basic web architecture in Figure 1.1 shows that the client requests the web server for web pages. Through caching proxy the client gets the response. The web log files are implemented using cluster and classification techniques to find out the association between the web data so that the most frequently used web logs are identified and stored in the web cache.

II. LITERATURE SURVEY

The web prefetching strategy proposed Association Rule Mining (ARM) algorithm to discover the pre-fetched documents. It discovers dependencies between pairs of documents. A web server can get the most probable next request reducing the time taken to respond to a request significantly that helps us to make up the web latency that are faced on the Internet today. The vigilant implementation of these methods can reduce access time and latency making optimal usage of the servers computing power and the network bandwidth.

Association rule mining is one of the strategies that find out association among the items sold together by relating market basket data analysis. The clustering method is also used for recognizing class of most traded products, categorizing customers based on their buying behavior and their influence of purchase. The various researches have implemented ARM and Clustering algorithms in different tools to find out the efficiency among them. This paper compares the results of Apriori and K-Means algorithms against their implementation in Weka and XL Miner. In this comparison the authors have used the sales transaction data set. The outcomes are very promising and also formed valuable information for sales and business improvements and also analyzed the data for hidden knowledge and the results showed some very interesting patterns in user buying behavior and buying timings [2].

The designing and implementing of a pre -fetching model in the form of a Markov model and association rule mining. The Markov property states that, given the present state, upcoming states are independent of the past states. The Web prefetching strategy also proposed in association rule mining algorithm to discover the pre-fetched documents. It discovers dependencies between pairs of documents [3].

A novel Frequent-Pattern tree (FP-tree) structure, which is an extended prefix-tree structure for storing compressed, crucial information about frequent patterns, and develop an efficient FP-tree based mining method, FP-growth, for mining the complete set of frequent patterns by pattern fragment growth [7].

This paper emphasizes on clustering among the different mining processes. It defines various clustering algorithm for similar kind of web access pattern. These algorithms serve as foundation for the web usage clustering that were described and found that web mining methods and clustering technique are used for self-adaptive websites and intelligent websites to provide personalized service and performance optimization[8].

Web pre-fetching techniques are categorized into three- probability based, clustering based and using weight-functions. In the probability based pre-fetching, probabilities are calculated using the history access data. This method assumes that the request sequence follows a pattern and the probabilities are trying to follow this pattern. Clustering based pre-fetching methods make decisions using the information about the clusters containing pages that have been fetched previously, assumes that pages that are close to those previously fetched pages are more likely to be requested in the near future.

The data mining techniques are an efficient tool for extracting hidden and useful information from large amount of data. Data mining is the young and fastest growing technology to deal with large data [14]. Data visualization supports the collaboration among users/data-analysts and data sets involved in data mining process. The data mining techniques are used for visualize and understanding of data.[15].The dynamic and self-reinforcing interactions between the macro- and micro-level, which increases dependence of shopping [16]. An effective way for finding multiple targets is to set a threshold for the candidates' scores[17].

Clustering approach is used to assume the structure of objects in their leading feature space and all the objects in training data sets are divided into several clusters. Each cluster is correspond to a preference degree and an ordinal regression function is then learned [18].An attribute cluster-based method is proposed to deal with the various dimensionality. The wide ranges of experiments over real datasets endorse the effectiveness and efficiency of our algorithm [19]. Some of semantically related words probably have a low or even zero co-occurrence probability [21]. The most important part in mining the web log files from the web server is to mine the session time, based on that the web user are identified.[22].The user navigational behavior plays a major role in mining the user access patterns.[23]. In this approach web access patterns of user are analyzed by using information present in the proxy server log files using clustering and classification techniques

III. RESULTS AND REVIEW

Web mining is the use of data mining techniques to discover outlines from the Web. According to analysis goals, web mining can be divided into three different types, which are Web usage mining, content mining and Structure mining.

Web usage mining is a procedure of picking up information from user how to custom web sites. Web content mining is a process of selecting up some information from the text, images and contents of a web page. Web structure mining is a process of picking up information from associations of web pages. The Web Usage Mining contains of three main steps,(1)Data Preprocessing, (2)Pattern Discovery and (3)Pattern analysis.

3.1 Data Preprocessing: In this phase, a series of processing tasks are functional on web log file such as data cleaning, user identification number, session identification number, path completion and transaction identification numbers.

3.2 Pattern Discovery: In this phase, the techniques from several research areas, such as data mining, Image processing, machine learning, statistics, and pattern recognition are inspected and applied on data to obtain a frequent patterns after preprocessing

3.2.1 Client Side: Association rules are defined in prefetching engine. The Prefetching engine is at client side. Prefetching engine stores web objects that are pre-fetched by the server side prediction engine by preprocessing web log.

For a client's request, pages predicted by the prediction engine are stored in prefetching engine at the client side.

3.2.2 Server Side: when a client sends a request for a web page/web object, the prediction engine will predict extra pages stored in the prefetching engine,

Web prefetching is to provide low recovery latency for users, that can be described as high hit ratio. Prefetching also upturns system resource requirements in order to increase hit ratio.

Resources disbursed by prefetching include server CPU cycles, server disk I/O's, and network bandwidth. Among them, bandwidth is possible to be the primary limiting factor.

Proxy caches have become a central mechanism for sinking the latency of Web documents recovery. While caching alone reduces latency for previously requested documents. The web document prefetching could mask latency for previously unseen, but correctly predicted requests.

K-Means Clustering Algorithm

Input: Let X = $\{x_1,x_2,x_3,\ldots\ldots..,x_n\}$ be the set of data points and V = $\{v_1,v_2,\ldots\ldots.,v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..
4) Recalculate the new cluster center using:

$$vi = \left(\left(\frac{1}{ci}\right) \sum_{j=1}^{ci} xi \right) \text{……………………..(1)}$$

where, *'c_i'* represents the number of data points in i^{th} cluster.
5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3).
The k-means clustering algorithm will be applied on web log data to obtain the k number of clusters to identify the association among them, and then it will be classified using Classification algorithm.

B-CART Algorithm

Derivation: D : Set of tuples
Each Tuple is an 'n' dimensional attribute vector
Input: X : {x1,x2,x3,…. xn}
Output: 'm' Classes : c1,c2,c3…cm
Naïve Bayes classifier predicts X belongs to Class Ci

$$P\left(\frac{Ci}{X}\right) > P(Ci)/P(X)$$

$$P\left(\frac{Ci}{X}\right) = P\left(\frac{X}{Ci}\right) P(Ci)/P(X) \text{ …………. (2)}$$

Naïve Assumption of "class conditional independence"

$$P(X/Ci) \; = \prod_{k=1}^{n} P\left(\frac{Xk}{Ci}\right) \text{ ……………. (3) } P\left(\frac{X}{Ci}\right) = P\left(\frac{X1}{Ci}\right) * P\left(\frac{X2}{Ci}\right) * \ldots * P\left(\frac{Xn}{Ci}\right) \text{ …. (4)}$$

The input data, also called the training set, consists of multiple records each having multiple attributes or features. Each record is tagged with a class label.
The objective of classification is to analyze the input data and to develop an accurate description or model for each class using the features present in the data.
This model is used to classify test data for which the class descriptions are not known. The clustered data is classified using Naive Bayes Classifier with the number of clusters selected by cross validation: 5

Table 1: The number of clusters with their attribute values

Clusters	0	1	2	3	4
Attributes	0.36	0.18	0.19	0.21	0.06

Table 2: Number of passes over the dataset

Cluster 0	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
7.3033	1.0699	1.0699	3.4025	1.1544
3.0171	1.0702	1.0702	10.71	1.1255
6.7251	6.9387	1.0858	1.0818	1.1686
2.1707	2.1868	1.2288	2.1487	5.2648
1.2249	1.1678	13.2787	1.1599	1.1686
7.0663	3.3092	1.2414	2.9265	1.4566
1.7994	8.9635	1.0672	1.064	1.1059
3.073	1.065	1.065	4.6721	1.1249
11.4516	1.0904	1.0904	2.2137	1.1538

Table 3: Number of most frequent patterns

Cluster0	Cluster1	Cluster2	Cluster3	Cluster4
1.4667	16.8939	1.1455	1.2183	1.2755
26.7017	1.7139	1.2467	1.7939	2.5439
5.7729	1.2676	1.2056	19.4345	1.3194
3.2602	1.4219	1.1638	1.402	2.7521
2.3536	1.2466	1.0822	1.235	1.0826
1.0514	1.1498	1.0749	1.1428	2.5811
1.2983	1.0833	4.4554	1.0793	1.0837

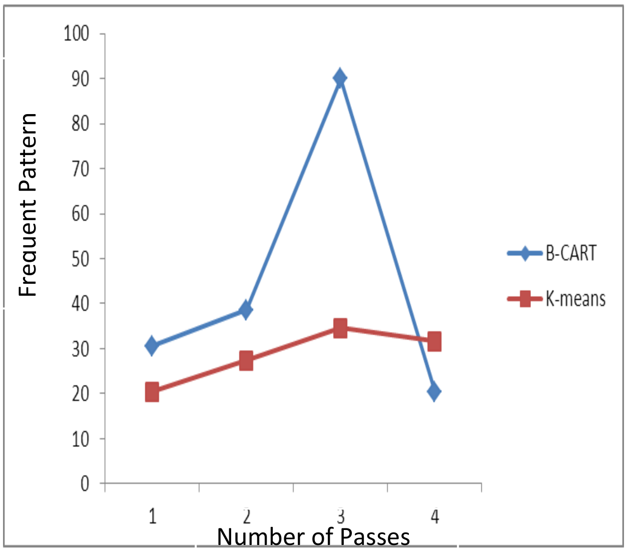

Figure 3.3 Comparisons on B-CART and ARM

The x axis denotes the number of passes over the dataset. y axis denotes the number of most frequent patterns. B-CART algorithm is more efficient because the most frequent item sets are mined compared to K-means algorithm.

IV. Conclusion & Future Work

The web log files are implemented using cluster and classification techniques to find out the association between the web data so that the most frequently used web logs are identified and stored in the web cache. The web log files in the proxy server are identified and it has been tested with k-means clustering, B-CART algorithms to find the user behavior over the internet.

Our future research work will be focused on classification techniques with the large data sets from web logs to identify the most frequently used patterns by the users on the web, applying different parameters to identify the frequent patterns and also increase the speed of server logs by removing the irrelevant log files.

References

[1] Vijayan and Jayasudha j. Greeshmas, A survey on web pre-fetching and web caching techniques in a mobile environment, cs & it-cscp 2012

[2] Ashok Kumar Loraine Charlet D,Annie M.C., "web log mining using K-Apriori Algorithm", volume 41, March -2012.

[3] Indla kasthuri , Ranjit KumarM.A, K. Sudheer Babu K ,Dr..Sai Satyanarayana Reddy, An Advance Testimony for Weblog Prefetching Data Mining, IJARCSSE, 2012.

[4] Harish Kumar and Anil Kumar," Clustering Algorithm Employ in Web Usage Mining: An Overview", INDIA Com publication, Edition 2011.

[5] Santhosh Kumar B, RukmaniK.V," Implementation of Web Usage Mining Using Apriori and FP-Growth Algorithms", volume: 01, Issue: 06, Pages: 400-404(2010).

[6] Khattak M, KhanA. M, sungyoung lee*, andyoung-koo lee, Analyzing Association Rule Mining and Clustering on sales day Data with XLMiner and Weka

[7] Rajan Chattamvelli, "Data Mining Methods", Narosa publications, Edition 2009.

[8] Jiawei Han, Ian Pei, Yiwen Tin, Runying Mao, "Mining Frequent Pattern without Candidate Generation: A Frequent Pattern Tree Approach", Volume-8.

[9] Jian Pei, Jiawei Han, Behzad Mortazavi-asl, Hua Zhu, "Mining Access Pattern Efficient from Web Logs"

[10]Han J and Kamber,"Data Mining: Concepts and Techniques", Morgan Kaufman Publishers, 2000.

[11]Gomathi B ,sakthivel"Implementing Fusion to Improve the Efficiency of Information Retrieval Using Clustering and Map Reduction"springer ,2016.

[12]WenfeiFan,Xin Wang, YinghuiWu, "Answering Pattern Queries Using Views"IEEE Feb-2016.

[13]Zhun (Jerry) Yu, Fariborz Haghighat, Benjamin C.M. Fung "Advances and challenges in building engineering and data mining applications for energy-efficient communities"Elsevier-2016.

[14]Wilson Castillo Rojasa, Fernando Medina Quispea, Claudio Meneses Villegasb "Augmented visualization for data-mining models"Elsevier-2015.

[15]Giulio Mattioli, Jillian Anable, Katerina Vrotsou,"Car dependent practices: Findings from a sequence pattern mining study of UK time use data"Elsevier-2016.

[16]Xinjie Zhou, Xiaojun Wan, and Jianguo Xiao"CMiner: Opinion Extraction and Summarization for Chinese Microblogs",IEEE-July2016.

[17]Ezgi Can Ozan, Serkan Kiranyaz, Senior Member, IEEE, and Moncef Gabbouj, Fellow, IEEE," K-Subspaces Quantization for Approximate Nearest Neighbor Search"IEEE –July-2016.

[18] Ou Wu, Qiang You, Xue Mao, Fen Xia,Fei Yuan, and Weiming Hu," Listwise Learning to Rank by ExploringStructure of Objects",IEEE –July-2016.

[19] Weiguo Zheng, Xiang Lian, Lei Zou, Liang Hong, and Dongyan Zhao," Online Subgraph Skyline Analysis over Knowledge Graphs",IEEE,July-2016.

[20]Swee Chuan Tana,Kai Ming Tingb, Shyh Wei Tengb," Simplifying and improving ant-based clustering" , Elsevier-2016.

[21]Yuan Wang, Jie Liu, Yalou Huang, and Xia Feng," Using Hashtag Graph-Based Topic Model to Connect Semantically-Related Words Without Co-Occurrence in Microblogs" , IEEE July-2016.

[22]Kousalya. R. & Saravanan. V. (2012,May) Time based Web User Personalization and search. International Journal of Computer Applications 46 (23),11-17

[23]Kousalya. R. & Saravanan. V., (2014,Sep) Personalizing User Directories through Navigational behavior of Interesting Groups and Achieving Mining Tasks Journal of Theoretical and Applied Information Technology. 67(2), 328-333.

BIOGRAPHY

K. Nandhini received her B.Sc., from Bharathiar University, Coimbatore in 1996 and received M.C.A., from Bharathidasan University, Trichy in 2001. She obtained her M.Phil., in the area of Data Mining from Bharathidasan University, Trichy in 2004.She obtained her Ph.D., degree in the area of data mining from Bharathiar University, Coimbatore in 2012. At present she is working as an Assistant Professor at Department of Computer Science in Chikkanna Government Arts College, Tirupur. She has published more than 25 research papers in various National and International Journals in the area of Data Mining. Her research interest lies in the area of Data Mining and Artificial Intelligence.

K.Suguna received her BCA., from Bharathiar University, Coimbatore in 2009 and received MCA., from Anna University ,Chennai in 2012. She obtained her M.Phil., in the area of Data Mining from Bharathiar University, in 2014. At present she is working as an Assistant Professor in Department of Computer Applications, Dr.N.G.P Arts and Science College, Coimbatore. She has published more than 7 research papers in National, International journals and conferences. Her research interest in the area of Data Mining.