Enhanced Integrated Approach to Predict Web User's Future Requests using K-Means and FP-Growth

Tanveer Kaur Dewgun
M.Tech Scholar
Bansal Institute of Science and Technology
Bhopal, India

Pushpraj Singh Chauhan
Information Technology Department
Bansal Institute of Science and Technology
Bhopal, India

ABSTRACT
The tremendous growth in the World Wide Web has led to the user perceived latency when requesting for resources from the web servers. Millions of users are connected to the web server for different needs. To improve the performance of the servers, caching is used where the frequently accessed pages are stored in proxy server caches. Pre-fetching of web pages is the new research area which when used with caching greatly increases the performance. In this paper, a better algorithm for predicting the web pages is proposed. Clustering of web users according to their location using K-Means clustering is done and then each cluster is mined using FP-Growth algorithm to find the association rules and predict the pages to be pre-fetched for storing in cache.

General Terms
Web Usage Mining, Web Caching, Web Pre-fetching

Keywords
Web Usage Mining, Apriori, FP-Growth algorithm, K-Means clustering

1. INTRODUCTION
Large amount of data is generated today due to the explosive development of information technology. Researchers have provided different approaches to store and manipulate this data for knowledge extraction. The process of analyzing and extracting new useful and previously unknown information from the huge data is known as data mining. Different areas where data mining is deployed are banking, e-commerce, biology, World Wide Web etc.

World Wide Web is one of the largest data source. It contains billions of data created by millions of internet users daily. This data can be analyzed to extract valuable information which leads to the new area of data mining application known as web mining. Web Mining can be defined as application of data mining techniques to extract knowledge from the web data including web documents, hyperlinks between documents, usage logs of websites, etc. [1]. In simple terms, web mining is just the use of data mining techniques on the data provided by www.

1.1 Web Usage Mining
The application of data mining techniques to discover interesting patterns from the web logs is known as web usage mining. Web usage mining provides better understanding for serving the needs of Web-based applications [1]. Web usage mining consists of three phases, preprocessing, pattern discovery, and pattern analysis. After the completion of these three phases the user can find the required usage patterns and use this information for the specific needs. Web usage mining has achieved great success in various fields, such as personalization of Web content and user recommendation, pre-fetching and caching pages for user navigation, improvement of Web design and e-commerce.

1.2 Web Caching
Owing to increasing network bandwidth and computing power, the usage of internet has grown at the breakneck rate. Over the years, there has been a significant increase in the number of users who access Internet through high speed DSL but still the access latencies perceived by them is high. Therefore, a significant reduction in web latency assumes importance for the users of the Internet and also for the Internet service providers who desire to increase the web surfing speed.

Web latency can be reduced either by pushing the bandwidth at the expense of incurring higher costs or by implementing better technological solutions such as introduction of cache(s) at the server, proxy or the client side. Web caching is an effective technique to alleviate the server bottleneck and reduce network traffic, thereby reducing network latency. It is the automatic creation of temporary copies of information residing on computers other than host servers in order to make this information readily available to people around the world.

1.3 Web Pre-fetching
Web caching is adopted to reduce the user perceived latency by storing the objects already visited in a place closer to the end user. However due to the tremendous growth in the web technology web caching alone is not sufficient enough to improve the system performance. World Wide Web is now not just a simple information sharing device, it has grown into a huge repository of dynamic and interactive data. System performance can be greatly enhanced by complementing the web caching technique with the effective technique known as web pre-fetching.

Web pre-fetching predicts the web objects that are likely to be accessed in the near future by studying the user behavior. These objects are those which are not yet requested by the users. The forecasted web objects are retrieved from the origin servers and saved on the caches to serve the end user requests. This helps in increasing the cache hits in turn reducing the latency.

In the proposed work, the two popular web mining techniques are used to predict the pages likely to be accessed in future. The data from a proxy server is collected and then preprocessed. Later the location of the users is used to cluster them using k-means algorithm. After clustering, the FP-growth algorithm is applied to generate the association rules from the clusters.

2. RELATED WORK

Due to rapid growth in the number of internet users, the user perceived latency has become a serious issue for the web service providers. Researches have been done which combines different techniques from multiple domains to overcome this issue.

To reduce perceivable network latency, researchers focused on pre-fetching popular documents. The integration of pre-fetching and caching techniques greatly improves the performance and also reduces the running time of the applications by 50% [2].

Garofalakis et al. basically provided a survey on data mining techniques and algorithms for discovering structures of web, hypertext and hyperlink [3]. In [4], a generalization based clustering approach has been presented, which also incorporates attribute oriented induction.

Pitkow et al. predicted the web surfer's path in pattern extraction mechanism [5]. [6] Worked on prediction of future requests and has built n-gram model for the same.

Cooley et al. categorized the web mining and then presented possible research areas. A scheme for fast allocation of web pages using data mining techniques and competitive neural network is being discussed in [7].

Zhang et al. proposed an efficient data clustering approach for very large databases, by generating hierarchical clustering of web users based on their access patterns [8]. In order to user's web page requests, clustering technique using first-order Markov models has been provided in [9]. Short-term pre-fetching uses Dependency Graph (DG), where graph consist of access patterns and Prediction by Partial Matching (PPM) is used. The merit of short-term pre-fetching is that it reduces the user-perceived latency. Other than this, it also has two demerits. Firstly, it may cause excessive network traffic, if pre-fetching policy is not designed cautiously. Secondly, optimization of cache space is not good in this pre-fetching scheme. The long-term pre-fetching uses global object access pattern statistics, where clusters of valuable objects are identified. This scheme may be used in places like as Content Distribution Network (CDN), mobile computing environments etc.

Different benefits of web pre-fetching are provided in [11, 12], whereas [10] motivates in research in web caching.

Vakali et al. described an extensive range of web data clustering schemes, in most of the cases clusters belongs to intra-site web pages [14]. In grouping inter-site web pages, web clustering performance reduces due to increase in complexity of web. If there is some change in web user's pattern, then it must to be updated in the resulted clusters.

Schloegel et al. used graph theory for working with web log files. The paper represented the web log files using web navigational graph and then using web partition techniques [13].

Nanhay Singh et al. used the two web mining techniques, K-Means clustering and Apriori algorithm together to predict and pre-fetch the web pages from the proxy server [15].

3. PROPOSED WORK

In the existing works the performance of the servers is improved by pre-fetching the likely pages and caching them in the web caches. Prediction of the pages can be performed using different algorithms such as Markov model, Apriori algorithm etc. The recent works also includes the integration of more than one of algorithms to overcome the limitations of each other.

In the proposed work two different data mining approaches to predict the web pages which are likely to be accessed in near future are used. The existing works try to cluster the data based on the user interests or the time taken by the server to respond back to the requests. In this work improvement of the performance is achieved by clustering the users in different group based on the location from which the request is sent.

Clustering the users based on the location improves the hit ratio. Suppose the user staying in Delhi searches for Cineplex in Delhi. It is most likely that he may also search for restaurants near him. By adding the location in the algorithm we can provide him better search results which he is looking for.

In this proposed work an improvement in the performance is also achieved by using the FP growth algorithm for finding the frequent itemset instead of the Apriori algorithm.

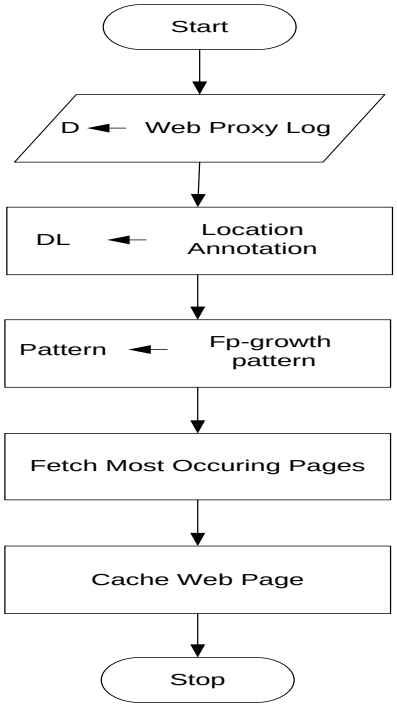

Figure 1. Proposed Work Flowchart

The following subsections explain the steps on the proposed work.

3.1 Web Proxy Log Data and Preprocessing

Web users traverse different web pages in search of information. They may visit multiple web sites and spend some time on different web sites. All the information about the web user's visit to the web pages is stored in a file called the web log file. The web server writes all the information about the user to this web log file whenever the user requests for a web page to the web server.

The web log file provides all the data about the user such as user name, IP address, Time Stamp, Access Request, number of Bytes Transferred, Result Status, URL that Referred and User Agent.

The log files are stored on the proxy servers in the popular log format known as the Common Log Format (CLF). This file keeps track of all the user requests that come to the web sites. This web log data is then preprocessed to make it appropriate for the mining algorithms. The Preprocessing phase involves the removal of all irrelevant and noise data from the web log file. In our proposed work we have performed data cleaning by removing all the unwanted entries such as entries with code 200 or all the entries created by the web agents.

3.2 Location Annotation

The proposed work uses the K-Means clustering algorithm which clusters the users based on their location. But the location information is not available in the log file. To obtain the user location the IP address is used. In this step, the location of the user is obtained from its IP address using a web service and this information is added to the processed logs for further mining the data.

3.3 Clustering using K-Means

The algorithm includes the clustering of web data using one of the simplest and efficient web mining clustering algorithms, the K-Means algorithm. Clustering can be defined as given the desired number of cluster k and a dataset of n points, and a distance-based measurement function, we are asked to find a partition of the dataset that minimizes the value of the measurement function. In simple words, clustering is a process of separating the data entities in different groups based on some similarities and dissimilarities.

The K-Means algorithm is yet the simplest but the most efficient algorithm which solves the objective of the cluster analysis. The procedure followed by K-Means is simple and straight forward in which we classify a data set of say n items into different clusters(say k clusters). The main idea is to find k centroids, one for every cluster. The next step is to find the entities from the data set belonging to the same centroid i.e. those which are nearest to the centroid. When all the items of the data set are assigned to one of the centroid the first stage is completed and an early set of clusters is obtained. After the first stage we recalculate to find the new centroids and then again find the distances between the data set entities and the centroids. The same process is iterated till the centroids become stable and there are no more changes in it. The K-Means algorithm is fast robust and easier to understand compared to the other clustering algorithms. Also it provides better results when the data items are well separated or distinct from each other.

In this study, the K-Means algorithm is used to group the web data into different clusters based on the location of the web users which is obtained from the IP addresses. The work assumes to separate the users based on the location from where the request is being generated. After obtaining the clusters, the algorithm to generate the association rules is applied.

3.4 Pattern Discovery using FP- Growth Algorithm

The frequently occurring patterns (list of itemsets or subsequences) in the data set are known as the frequent patterns. For instance, a subset of items from the data set such as bread and butter appearing frequently in the transactions can be called as a frequent itemset. These types of frequent patterns play an important role in mining associations and finding relationships between the items in the transaction. Finding associations between items means to discover which items appear together in the transaction and are frequently found in the transactions.

Association rule mining is the most popularly used method for discovering interesting relationships between the items in the databases. For example, the rule {bread, butter} => {milk} discovered from the transactional data of any supermarket indicates that the customer buying bread and butter is likely to buy milk also. Finding such rules can greatly help the business analysts to provide better solutions and enhance the performance.

A web log file also provides a lot of information about the web users and their behavior. Association rule is the widely used data mining technique which can be applied to the web data as well to discover frequent patterns. When applied to Web Usage Mining, association rules are used to find associations among web pages that frequently appear together in users' sessions. The typical result has the form {X.html, Y.html} => {Z.html} which states that if a user has visited page X.html and page Y.html, it is very likely that in the same session, the same user has also visited page Z.html.

Apriori is one of the frequently used algorithms to mine and discover the association rules. It uses the breadth first search technique to calculate the support value for the items and also a candidate generation function is used to exploit the downward closure property of support. This candidate generation step of Apriori provides good results but it also suffers from two nontrivial costs:

a) It may need to generate a huge number of candidate sets.

b) It may need to repeatedly scan the database and check a large set of candidates by pattern matching.

Another influential algorithm which does not use the candidate generation technique to mine the complete frequent itemset is the frequent-pattern growth or simply the FP-Growth algorithm.

FP-Growth algorithm uses the FP tree data structure which represents all the database transactions. It then applies the divide and conquer technique to solve the mining problem by first breaking it into multiple small problems.

3.5 Fetch Most Occurring Pages

After the application of the FP-Growth algorithm a number of rules which will help in predicting the pages which are likely to be requested in future by the web users are obtained. In this step, those web pages are found by studying the association rules discovered.

3.6 Caching the Web Pages

The predicted pages are then fetched from the server and stored on to the proxy server cache so that when requested they can be provided to the users and in turn reduce the latency.

4. EXPERIMENTAL WORK AND RESULTS

The dataset from the ircache.net website is collected. The file "pa.sanitizedaccess. 20070109.gz" is used for experiment which is available at the proxy server installation ftp://ircache.net.

In the proposed framework, the raw data from the log file is collected initially which is processed later.

Following fig. 2 shows the sample log file collected from the proxy server.

```
1168387181.350 258 17.219.121.198 TCP_REFRESH_HIT/200 3236 GET http://www.1001freefonts.com/fontsdisplay/eldarado.gif - DIRECT/69.93.91.218 image/gif
1168387182.185 190 17.219.121.198 TCP_REFRESH_HIT/200 2575 GET http://www.1001freefonts.com/fontsdisplay/electrofied.gif - DIRECT/69.93.91.218 image/gif
1168387186.300 7 235.121.3.62 TCP_MISS/400 212 GET http://pa.us.ircache.net:3128/squid-internal-dynamic/netdb - NONE/- -
1168387186.943 173 27.109.138.213 TCP_CLIENT_REFRESH_MISS/304 165 GET http://www.qvc.com/qvc/gif/hp/tab_blue_center.gif - DIRECT/167.140.19.2 -
1168387187.220 100 27.109.138.213 TCP_CLIENT_REFRESH_MISS/200 477 GET http://www.qvc.com/qvc/gif/hp/lft_grey_corner.gif - DIRECT/167.140.19.2 image/gif
1168387187.570 92 27.109.138.213 TCP  CLIENT  REFRESH  MISS/200 994 GET http://www.qvc.com/qvc/gif/bttn  watchqvc.gif - DIRECT/167.140.19.2 image/gif
```

Figure 2. Sample Proxy Log Data File before Preprocessing

After preprocessing, the number of lines reduced from 915 to 672. A snapshot of data after preprocessing is shown in fig. 3.

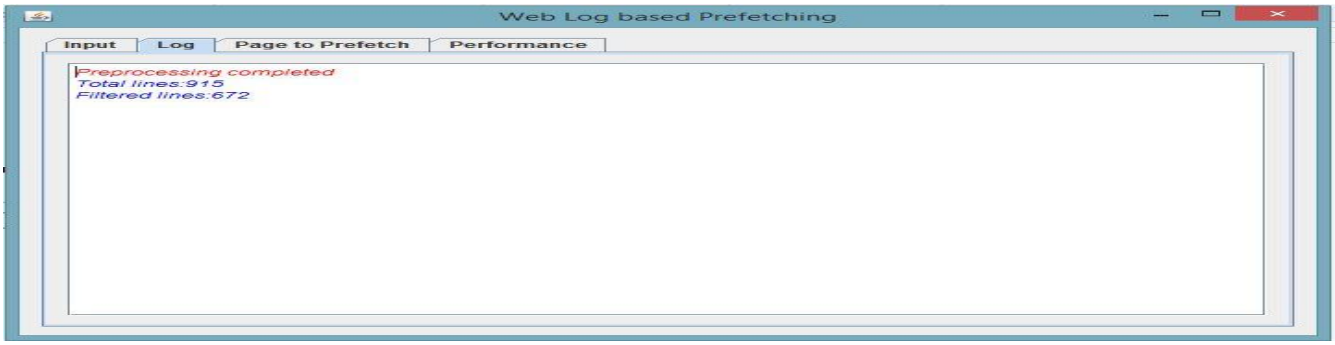

Figure 3. Number of filtered lines after preprocessing

After this the data is clustered based on location using k-means algorithm. Later the work finds the association rules for predicting the pages using the FP-Growth algorithm. The predicted pages are fetched from the server and then can be stored onto the cache of the proxy servers.

The fig. 4 shows the time taken by different algorithms when clustered with location and without location.

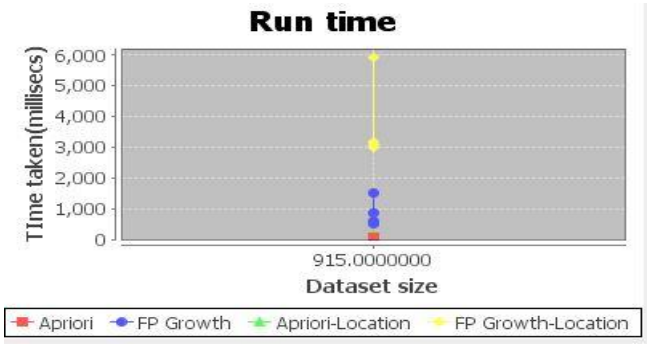

Figure 4. Time take by apriori and FP- Growth

From the graph above it can seen that the time taken by Apriori algorithm is the least but the advantages over the Apriori algorithm as compared to FP growth can be seen in the next two figures.

Fig. 5 shows the number of patterns generated which help the pages to be predicted and pre-fetched. More the number of rules better are the chances of having the required page found in the cache.

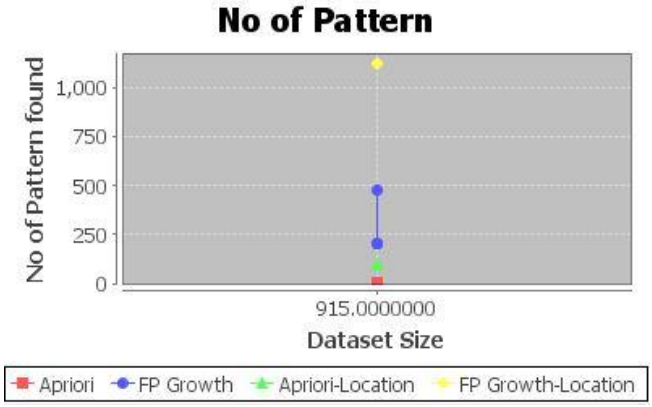

Figure 5. No of patterns generated

The next figure fig. 6 shows the hit ratio of the algorithms. The results show that when clustering the data based on location there is an increase of 5% to 8% in the hit ratio. Also the FP- Growth algorithm is better than the Apriori algorithm can be seen from the results.

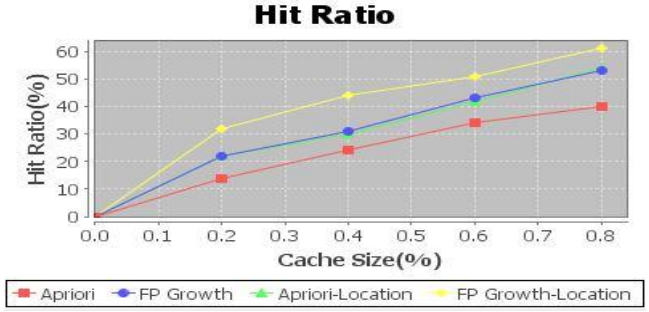

Figure 6. Hit Ratio of the algorithm

5. CONCLUSION AND FUTURE WORK

In the proposed work, the two mining algorithms are used to predict the pages likely to be requested by the users and then pre-fetch them to store them in the cache. Results have shown that the algorithms used and the clustering of data using web user's location greatly improves the performance of the proxy server.

The hit ratio as seen in the results is increased by 5-10% when the web user's access data is clustered based on their location. Also the FP-Growth algorithm generates more number of rules as compared to Apriori which in turn helps in increasing the hit ratio.

The use of FP-Growth algorithm increases the hit ratio but the algorithm takes more time as compared to Apriori. The reduction of time taken can be improved in future. The results of this proposed work are tested on a small set of data. The implementation of the work on the real web server is under the future scope.

6. REFERENCES

[1] Web Usage Mining: Discovery and Applicaions of Usage Patterns from Web Data, Jaideep Srivastava, Robert Cooleyz , Mukund Deshpande, Pang-Ning Tan, ACM SIGKDD, Jan 2000.

[2] K. Chinen and S. Yamaguchi. An Interactive Pre-fetching Proxy Server for Improvement of WWW Latency. In Proceedings of the Seventh Annual Conference of the Internet Society (INEt'97), Kuala Lumpur, June 1997.

[3] Garofalakis M. N., Rastogi R., Sheshadri S., and Shim K.,"Data mining and the Web: past, present and future." InProceedings of the second international workshopon Web information and data management, ACM, 1999.

[4] Fu Y., Sandhu K., and Shih M., "Clustering of Web Users Based on Access Patterns." International Workshop on Web Usage Analysis and User Profiling (WEBKDD'99), San Diego, CA, 1999.

[5] Pitkow J. and Pirolli P. Mining longest repeating subsequences to predict www surfing. In Proceedings of the 1999 USENIX Annual Technical Conference, 1999.

[6] Z. Su, Q. Yang, Y. Lu, and H. Zhang. Whatnext: A prediction system for web requests using n- gram sequence models. In Proceedings of the First International Conference on Web Information System and Engineering Conference, pages 200-207, Hong Kong, June 2000.

[7] Phoha V. V., Iyengar S.S., and Kannan R., "Faster Web Page Allocation with Neural Networks," IEEE Internet Computing, Vol. 6, No. 6, pp. 18-26, December 2002.

[8] Zhang T., Ramakrishnan R., and Livny M., "Birch: AnEfficient Data Clustering Method for Very Large Databases." In Proceedings of the ACM SIGMODConference on Management of Data, pages 103-114, Montreal, Canada, June 1996.

[9] Cadez I., Heckerman D., Meek C., Smyth P., and Whire S., "Visualization of Navigation Patterns on a Website Using Model Based Clustering." Technical Report MSR-TR-00-18, Microsoft Research, March 2002.

[10] Podlipnig S, Boszormenyi L. A survey of Web cache replacement strategies. ACM Comput Surveys 2003;35(4):374–98.

[11] Rabinovich M, Spatsheck O. Web caching and replication. Addison Wesley; 2002.

[12] Teng WG, Chang CY, Chen MS. Integrating Web caching and Web prefetching in client-side proxies. IEEE Trans Parallel Distributed Syst 2005;16(5):444–55.

[13] Schloegel K, Karypis G, Kumar V. Parallel multilevel algorithms for multi-constraint graph partitioning. In: Proceedings of 6th international Euro-Par conference. September 2000. p. 296– 310.

[14] Vakali A, Pokorny J, Dalamagas T. An overview of Web data clustering practices. In: Proceedings of the EDBT Workshops 2004. Heraklion, Crete; 2004. p. 597–606.

[15] Nanhay Singh, Arvind Panwar and Ram Shringar Raw. Enhancing the performance of Web Proxy Server using Cluster Based Pre-fetching technique. IEEE 2013.