International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), Nov 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

An Integrated Approach to Predict User's Behavior from Web Transactions

Pushpraj Singh Chauhan[1], Dr. Asheesh Shah[2] ,Dr. Suresh Jain[3]

[12]CSE,MewarUniversity, (India)

[3]CSE, PIEMR, (India)

ABSTRACT

The Internet and the World Wide Web (WWW) have made dramatic impact on individuals and organizations in less than a decade. The tremendous growth in the World Wide Web has led to the user perceived latency when requesting for resources from the web servers. Pre-fetching of web pages and then applying caching to those pages greatly increases the performance of the servers. In this paper we have tested the results in terms of Run Time and Pattern Generated. When we compare our proposed algorithm with Apriori and Fp-Growth we find that though the proposed algorithm is taking more time to run than the other two, but it (proposed algorithm) increase the number of patterns generated. This in turn will help us in better prediction of web user's behavior.

Keywords: *FP-Growth algorithm, K-Means clustering, Pre fetching, Web caching*

I. INTRODUCTION

The promise of the Internet and e-commerce has led to the increasing use of the web for transaction processing. Many organizations have adopted web-enabled transaction processing for applications such as processing payments online, selling products online, making travel reservations to name a few. E-commerce has been on a steady rise. However, transaction processing on the web is not the dominant use of the Internet or e-commerce although it is an essential application. Some transactions are very simple, such as purchasing a book or transferring funds, and can be processed immediately. Other transactions are simply defined, transaction processing is the unambiguous and independent execution of a set of operations on data in a database, which treats the set of actions as a single event [1].

1.1 Overview of Web Transaction

Web Transaction can be defined as an activity, or group of activities, that is responsible for performing some application-specific work. Web Transaction is a sequence of URLs combined into one complete process. Typical web transactions are when a customer logs in a member website, makes a purchase on a shopping site, fills in and submits a web form and performs other interactions with a website and web application. The Web Site Pulse Transaction Monitoring allows Customers to measure the experience of online users navigating through multiple steps of their websites. Client-Server computing systems in which we use Web browsers as

client systems are called Web application systems. Web application systems dealing with database transactions are called Web-based transaction systems

1.2 Web Usage Mining

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site [2]

1.3 Web Caching

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.

Caching in simple words is a process to keep the frequently accessed documents in places near to the end user to decrease the user perceived latency. Based on the places where the caches are placed web caches can be categorized as below:

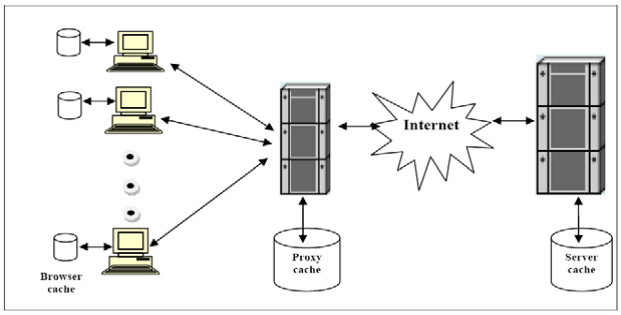

Figure 1 Types of Web Caches

1.3.1 Browser Cache

This type of cache is the one placed in the client's machine. Browser Caches are those created by the popular web browsers such as Google Chrome, Internet Explorer, Netscape, Mozilla etc. on the client machine. Internet browsers use caching to store HTML web pages by storing a copy of visited pages and then using that copy to render when you re-visit that page.

1.3.2 Proxy Server Cache

Proxy server is a computer system which is found in between the end device and the origin server. Proxy server cache provides the same functionality as the browser cache but on a larger scale. Browser cache is created only for one user but a cache created at the proxy server serves many different users in the same way. Whenever a client request any object, the request first goes to the proxy server where it checks if the object is available in its

own cache. If it is available the request is fulfilled by the proxy itself, if not the request is forwarded to the origin server.

1.4 Web Pre-fetching Techniques

However due to the tremendous growth in the web technology web caching alone is not sufficient enough to improve the system performance. System performance can be greatly enhanced by complementing the web caching technique with the effective technique known as web pre-fetching. Pre-fetching is done by predicting the future user requests either by studying the content of the web pages or by analyzing the history of user's past behavior. Many researchers are interested in predicting the future request based on their past activities. Following are some of the techniques working on this principle:

1.4.1 Markov Model

Markov models are very commonly used in the identification of patterns based on the sequence of previously accessed pages [3] [4]. They are the natural candidates for sequential pattern discovery for link prediction due to their suitability to modeling sequential processes. The Markov model process calculates the probability of the page the user will visit next after visiting a sequence of Web pages in the same session. Markov model implementations have been hindered due to the fact that low order Markov models do not use enough history and therefore, lack accuracy, whereas, high order Markov models incur high state space complexity.

1.4.2 Association Rules

Association rule mining is a major pattern discovery technique [5]. Every user leaves traces of information behind his every action which reveals his behavior. In web usage mining the web server logs collects all the information about the user activities. Association rule mining helps in predicting the future requests by studying the history. The original goal of association rule mining is to solve market basket problem. For a data set containing shopping transactions, association rules summarize relationships illustrated by the following example. Customers who buy bread and milk will most likely buy eggs, or, bread and milk -> eggs. The use of association rules is not limited to the traditional market-basket problem. It is widely used as an approach to predict future accesses in web usage mining. The main limitation of association rule mining is that many rules are generated, which result in contradictory predictions for a user session.

II. RELATED WORK

To account for some of the shortcomings in TLA, Cochrane & Markey [6] suggest combining TLA with another type of analysis (either questionnaire or protocol) to provide a more complete picture which can draw on the strengths of both types of studies.

Much literature was reviewed to identify the factors that may influence the adoption of web-enabled transaction processing. Studies in the field of innovation, which span many disciplines and focus on both organizations and individuals, have defined an innovation as an idea, practice, or object that is perceived as new by an individual or another unit of adoption. Young [7] illustrated the use of Web Transaction Logs as a collection management tool.

Wallace [8] demonstrated how analysis of transaction can identify bibliographic instruction needs and point out weaknesses in information system design.

Web Transaction logs have been used by librarians for over a quarter of a century to unobtrusively monitor user behavior with information systems [9].A transaction log is the output product of transaction monitoring. The transaction monitoring of an information system is defined as "the automatic logging of the type, content, or time of transactions made by a person from a terminal with that system" [10].

Kurth [11] states: "Transaction log data effectively describe what searches patrons enter and when they enter them, but they don't reflect, except through inference, who enters the searches, why they enter them, and how satisfied they are with their results" (p. 98). Kurth further goes on to explain that errors in TLA can arise through limitations of the online system, the inability to isolate and characterize individual users, and decisions and biases of the researcher analyzing the logs.

Polly & Cisler [12] point out two weakness of the use of the Web as an information system: slowness and "chaotic disorganization". While the issue of speed will have to be taken up by computer scientists and engineers, the disorganization of the Web is a prime target for librarians to tackle.

III. PROPOSED WORK

In the existing works the performance of the servers is improved by pre-fetching the likely pages and then caching them in the server. 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 proposed work improvement of the performance is achieved by clustering the users in different group based on their location from which the request is sent. Clustering the users based on the location improves the hit ratio. The web log file provides all the data about the user such as user name, IP address, Time Stamp, Access Request, number of Bytes

3.1 Web Proxy Log Data and Preprocessing

The transaction performed by the users are stored as web log in servers. 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 created by the web agents.

3.2 Location Annotation

The proposed work K-Means algorithm is used for clustering. Clustering is done as per the area/location of the users. Since 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. By adding the location in the algorithm we can provide him better search results which a person is looking for.

3.3 Clustering using K-Means

The K-Means algorithm is yet the simplest but the most efficient algorithm which solves the objective of the cluster analysis.In K-Means Algorithm 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 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 item set. 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. 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.

International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), Nov 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

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

Most browsers cache text and images, so when a user returns to a previously viewed page, the server is often not accessed; thus, the server access log contains no record of the return to the cached document and this action cannot be registered in the transaction log. This means that the transaction logs will, by necessity, be incomplete.

Here, 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.

IV. EXPERIMENTAL WORK AND RESULTS

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

After preprocessing, the number of lines reduced from 3131 to 2431. A snapshot of data after preprocessing is shown in fig. 3.

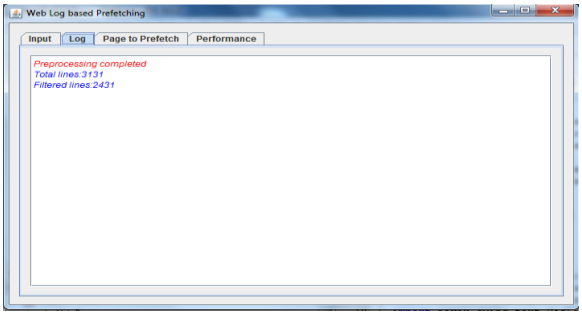

Figure 3. Number of filtered lines after preprocessing

```
GET http://www.quiethits.com/hitsurfer.php?[387.vlwHJccIgu:aIqlvda] - DIRECT/204.92.87.134 text/html1168300919.377 280 17.219.121.198 TCP_MISS/200 544
GET http://nuclearhits.com/external/target2.php?[Ogpmv6tDjKe6xMWQ.PuaEU] - DIRECT/66.98.166.88 text/html1168300931.828 142 194.124.123.75 TCP_MISS/504 1599
GET http://www.mister-wong.de/tags/seo-tools/ - NONE/- text/html1168300935.244 501 194.124.123.75 TCP_MISS/504 1617
GET http://offerz.directtrack.com/42/1/718&dp=4902 - NONE/- text/html1168300935.959 1612 254.118.93.171 TCP_MISS/304 248
GET http://stb.msn.com/i/9F/62C18E2491DC914FEC5A90147E6D39.gif - TIMEOUT_DIRECT/69.108.159.61 -1168300936.604 1 194.124.123.75 TCP_MISS/504 1617
GET http://www.mister-wong.de/tags/backlings_bekommen/ - NONE/- text/html1168300938.524 853 99.143.68.131 TCP_CLIENT_REFRESH_MISS/304 277
GET http://us.i1.yimg.com/us.yimg.com/i/ww/thm/1/search_1.1.png - DIRECT/64.215.172.70 image/png1168300941.345 2 194.124.123.75 TCP_MISS/504 1593
GET http://www.mister-wong.de/user/msuess/ - NONE/- text/html1168300942.648 80 99.143.68.131 TCP_CLIENT_REFRESH_MISS/304 278
GET http://us.i1.yimg.com/us.yimg.com/i/ww/news/2007/01/08/jobbig.jpg - DIRECT/64.215.172.70 image/jpeg1168300943.033 372 194.124.123.75 TCP_MISS/504 1621
GET http://www.bolanews.com/edisi-cetak/hl1.jpg - NONE/- text/html1168300950.163 50 99.143.68.131 TCP_CLIENT_REFRESH_MISS/304 278
GET http://us.i1.yimg.com/us.yimg.com/i/mntl/crr/07q1/img_0108.jpg - DIRECT/64.215.172.70 image/jpeg1168300953.645 648 99.143.68.131 TCP_MISS/304 260
GET http://us.js2.yimg.com/us.js.yimg.com/lib/bc/bc_2.0.3.js - DIRECT/64.215.172.97 application/x-javascript1168300957.585 186 194.124.123.75 TCP_MISS/504 1585
```

Figure 2. Sample Proxy Log Data File before Preprocessing

International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), Nov 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

After this the data is clustered based on location using k-means algorithm. Later on most likely visited pages are predicted using Apriori and FP-Growth algorithm. The predicted pages are then fetched from the server stored onto the cache of the proxy servers.

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

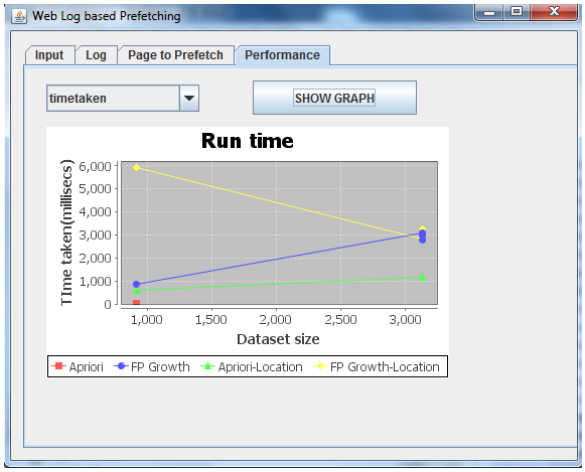

Figure 4. Run Time by Proposed Algorithm as compared to Apriori and FP- Growth

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. More the number of rules better are the chances of having the required page found in the cache.

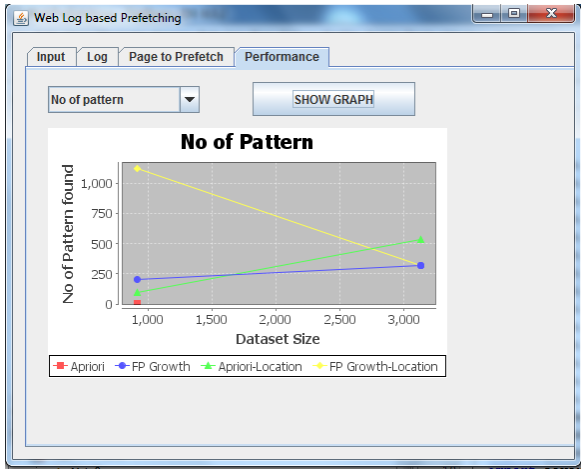

Figure 5. No of patterns generated

V. CONCLUSION AND FUTURE WORK

In this proposed work, we have integrated Apriori and Fp Growth Algorithm with location, which provides better result in terms of pattern generation. We have also used Apriori and Fp-Growth algorithms alone. The FP-Growth algorithm generates more rules (patterns) as compared to Apriori Algorithm which in turn increases the hit ratio but the Fp- Growth algorithm takes more time as compared to Apriori. The reduction of time taken can be improved in future by using an incremental approach.

REFERENCES

[1] Pete, L., "Transaction Processing," *Computerworld, Vol.35, No.40, 2001.*

[2] https://en.wikipedia.org/wiki/Webmining

[3] Bouras, C. & Konidaris, A. (2004), "Predictive prefetching on the web and its potential impact in the wide area", *WWW: Internet and Web Information Systems* (7), 143–179.

[4] Chen, M., LaPaugh, A. S. & Singh, J. P. (2002), "Predicting category accesses for a user in structured information space", *SIGIR'02, Finland* pp. 65–72.

[5] Mobasher, B., Dai, H., Luo, T. & Nakagawa, M. (2000), "Discovery of aggregate usage profiles for web personalization", *WebKDD'00, USA* pp. 61–82.

[6] Cochrane, P. A. & Markey, K. (1983). Catalog use studies since the introduction of online interactive catalogs: impact on design for subject access. Library & Information Science Research, 5(4), 337-363.

[7] Young, I. R. (1992). The Use of a general periodicals bibliographic database transaction log as a serials collection management tool. Serials Review, 18(4), 49-60.

[8] Wallace, P. M. (1993). How do patrons search the online catalog when no one's looking? Transaction log analysis and implications for bibliographic instruction and system design. RQ, 33(2), 239-252.

[9] Kaske, N. K. (1993). Research methodologies and transaction log analysis: issues, questions, and a proposed model. Library Hi Tech, 11(2), 79-85.

[10] Kurth, M. (1993). The Limits and limitations of transaction log analysis. Library Hi Tech, 11(2), 98-104.

[11] Peters, T. A., Kaske, N. K., & Kurth, M. (1993). Transaction log analysis. Library Hi Tech Bibliography, 8, 151-183.

[12] Polly, J. A. & Cisler, S. (1994). What's wrong with Mosaic? Library Journal, 119(7), 32-34.

[13] 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.

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