PROXY DRIVEN FP GROWTH BASED PREFETCHING

Devender Banga1 and Sunitha Cheepurisetti2
1,2Department of Computer Science Engineering,
SGT Institute of Engineering and Technology, Gurgaon, India

ABSTRACT
Now days, number of people are connected to the web to get results for their various queries from the servers. Servers are also heavily loaded to respond to large number of user’s. At this moment of time when millions of users are connected to the web, results for their queries should be as fast as possible. There are number of schemes by which we can reduce the response time of a query to a server, such as caching scheme, prefetching scheme. This paper proposes a framework for prefetching in the proxy server. For the reduction in the web latency and improving user satisfaction, prefetching is important. The proposed framework uses FP growth algorithm for determining the user’s frequent sets of patterns. User data will be collected from the web log history. Now, depending upon the threshold and patterns generated by the FP Growth, list of patterns to be prefetched will be generated and passed to the predictor module. Predictor module will prefetch the web objects by creating session with the main server. Using FP Growth algorithm of association rule mining frequent patterns can be determined without candidate key generation. Working of FP Growth is shown using Data-applied. The proposed framework improves the efficiency of the existing network.

KEYWORDS: Prefetching, Proxy server, cache, prediction.

I. INTRODUCTION

Ones dependency on network is increasing day by day, in developing countries where mostly every person is addicted to web, as we can get result to any query of any subject we heard even once. The moment we start searching with a statement or a keyword, that keyword is stored by browser cache for further providing fast results. In today’s network where speed is of major concern we need schemes and algorithms to reduce the response time from server to a client, where using only a cache is not sufficient to reduce the latency. As today’s network is full of congestion and getting a response faster is a major issue of concern. Many researchers worked in this area and also there are schemes and frameworks available which reduces the response time. One of the schemes available is prefetching with web caching. Web mining is used to mine the data received from the proxy server in three tier web architecture and then results are fed back to proxy cache which then responds to user queries on demand. One of the example of mining is, if we visited a shop to buy bread slices, then the milk can, butter bar and cheese slices which are placed nearby bread can be mined by using association rule based mining, in this example we have only one query named bread but other results such as slices, milk can are not required but these are the expected results of query for persons which may looked for milk can and cheese slices along with bread at some earlier point of time. So by applying mining rules we can get results to expected queries from the cache rather than asking the results for various pre-questioned queries again and again from the server. This greatly reduces the latency of a network. There are three types of caches; browser cache, proxy cache and server cache. Applying a cache with a browser gives faster results but while using three tier architecture where proxy server is the mediocre; much of the burden is bared by proxy server instead of main server. So, by applying web caching rules on proxy server instead of main server will help in providing the results much faster. Proxy servers are used not only to reduce web latency but for achieving various tasks like: NAT (Network Address Translation), Firewall and for security of web servers so that unauthorized access could be stopped [1]. Proxy servers also provide a way where we can grant
access to any of the link on the web by hiding the actual details of the users like IP (Internet Protocol) address.

Paper is divided into different sections for the ease of understanding. Section II focuses on the primary work done in this field. Section III depicts the proposed framework and section IV provides the experimental work done to prove the proposed framework. At the end, conclusion of the work has been provided.

II. LITERATURE REVIEW

Pallis et. al[2] addressed the short-term prefetching problem on a web cache environment using an algorithm for clustering inter-site web pages. The proposed scheme efficiently integrates web caching and prefetching. According to this scheme, each time a user requests an object, the proxy fetches all the objects which are in the same cluster with the requested object.

Sharma and Dubey proposed a framework for web traffic reduction [3]. The approach first extracts data from proxy server web log, and then the extracted data is preprocessed. The preprocessed data is then mined using clustering, and sequence analysis is being done to know the patterns to be prefetched.

Divya and Kumar presented a survey on three different association rule mining algorithms AILS, Apriori and FP-tree algorithm and their drawbacks which would be helpful to find new solution for the problems found in these algorithms [4].

Kosala et.al surveyed the research in the area of web mining [5]. The paper explores the connection between web mining categories and the related agent paradigm. This paper focuses on representation issues, on the process, and on the learning algorithm, and the application of the recent works as the criteria.

Chang et.al[6] describes that Information extraction (IE) from semi-structured web documents is a critical issue for information integration systems on the Internet. The discovery of repeated patterns is realized through a data structure call PAT tree. The paper also focuses that incomplete patterns are further revised by pattern alignment to comprehend all pattern instances.

Sharma and Dubey provided the literature survey in the area of web mining [7]. The paper basically focuses on the methodologies, techniques and tools of the web mining. The basic emphasis is given on the three categories of the web mining and different techniques incorporated in web mining.

Dominic and Abdullah dealt with FP-growth’s Variation algorithms [8]. The paper considers to the “Classic” frequent itemsets problem, which is the mining of all frequent itemsets that exist in market basket-like data with respect to support thresholds. The execution time and the memory usage were recorded to see which algorithm is the best. For the time consumption AFOPT(A Frequent Pattern Tree) algorithm took advantage for most of the data set even though it suffers from segmentation fault in the low support values on connect4 data set.

Venketesh et. al[9] presented a prediction model that built a Precedence Graph by considering the characteristics of current websites in order to predict the future user requests. The algorithm differentiated the relationship between the primary objects (HTML) and the secondary objects (e.g., images) when creating the prediction model.

Kasthuri et. al[10] shows that deduction of future references on the basis of predictive Prefetching, can be implemented by based on past references. The prediction engine can be residing either in the client/ server side. But in our context, prediction engine resides at client side. It uses the set of past references to find correlation and initiates Prefetching that is driving user’s future requests for web documents based on previous requests.

Sharma and Dubey proposed a framework for web traffic reduction [11]. The paper presented a framework for the prefetching and prediction in web. According to the framework, previous web requests of the user will be extracted from the proxy web log. From this web log, strong rules will be generated using FP Growth algorithm. These rules will be used to prefetch the upcoming requests of the current user.

Singh et. al[12] proposed a framework for prediction of web requests of users and accordingly, prefetching the content from the server. The proposed framework improved performance of web proxy server using web usage mining and prefetching scheme. They have clustered the users according to their access pattern and usage behaviour with the help of K-Means algorithm and then
Apriori algorithm is applied to generate rules for prefetching pages. This cluster based approach was applied on proxy server web log data to test the results using LRU and LFU prefetching schemes. Ford Lumban Gaol proposed web log sequential pattern mining using Apriori-all algorithm. The experiment will be conducted base on the idea of Apriori-all algorithm, which first stores the original web access sequence database for storing non-sequential data. An important application of sequential mining techniques is web usage mining, for mining web log accesses, where the sequences of web page accesses made by different web users over a period of time, through a server, are recorded [13].

III. THE PROPOSED FRAMEWORK

Prefetching and caching scheme is used to improve the performance of web even when more number of user are connected and demanding access to the content over the web. In our proposed framework a browser sends a request to the listener (forwarding function) to setup a connection between client and proxy server. Listener then forwards the request for connection establishment via TCP/IP Handshaking Protocol to the proxy server. Once the connection is established then proxy server sends a hint list to the predictor. Predictor maintains a web log history which tells us about the pages and content demanded with the number of hits of a particular web page with the IP address of the user. Hints from the predictor are then applied with FP (Frequent Pattern) growth algorithm. FP Growth algorithm gives the compressed data objects as an output, applying association rule mining approach. The proposed framework is as shown in Figure 1.

![Proposed Framework for web prefetching](image)

**Figure 1: Proposed Framework for web prefetching**

2.1. Predictor

Predictor collect hint list from the proxy server with the help of keywords from the user query. Hint list is very much essential to trace the type of data for which user is demanding, which is again forwarded to web log history and to the block where FP growth algorithm is applied.

2.2. Web Log History

Web log history is very important aspect in prefetching. Web logs redirect the prefetch engine to collect and cache particular type of data according to the demand of the user. Web log maintains the session of the user, IP address of the user, for how long a user was connected, type of information for which a user demanded, most frequently visited pages etc.

2.3. FP Growth

FP Growth approach is used for producing frequent item data sets based on divide and conquer method. FP Growth is mainly used for mining frequent item sets without candidate generation. Major step in FP-growth is to compress the database showing frequent item set in to FP-tree. To build a FP
tree further we need to apply 2 steps on the dataset. Further FP tree is divided into a set of more number of conditional databases in which we applied association rule mining for mining of frequent item datasets from FP tree [8]. FP growth algorithm pseudo code is as shown in Figure 2 [14].

```
FP_GROWTH

Input: FP-tree
Method: Call FP-growth (FP-tree, null).

procedure FP-growth (Tree, α)
{
1) if (single path P) do
2) for each combination: generate pattern β ⊢ α with support = minimum support of nodes in β.
3) else for each header ai do
   4) Generate pattern β = ai ⊢ α with support = ai.support;
5) Construct β.s conditional pattern base and then β.s conditional FP-tree Tree β
6) if Tree β = null
7) Then call FP-growth (Tree β, β)

Output: complete set of frequent patterns
```

Figure 2: FP Tree algorithm pseudo code [14]

2.4. Proxy Cache

Proxy cache is used with proxy server to save certain type of web data. The data after prefetching and mining is fed to proxy cache. A proxy cache works very similar to that of cache in operating system. When certain type information is demanded by a user / client through web browser passes through a proxy server than proxy server look for the response of that query in its cache memory. If the result of query for which the user is asking for found in cache memory of proxy server than it is called a hit else it is a miss. If no result is found in server cache then the query is forwarded to main server to get the result of the query.

2.4. Handshaking Process

As we know that TCP is a connection oriented protocol, but TCP does not establish connection on its own TCP uses handshaking protocol for establishment of connection. Connection establishment is very essential part in TCP so that further session of query and response could be maintained. Handshaking protocol is of two types:

2.4.1. Two way handshaking

In this type of handshaking protocol there are 2 types of messages passed between client and server for connection establishment and they are: Request message from host to server, accept/reject message from server to host.

2.4.2. Three way handshaking

In this type of handshaking protocol there are 3 types of messages passed between client and server for connection establishment and they are: Syc message from host to server, Syc_Ack from server to host, Ack from host to server.

IV. EXPERIMENTAL WORK

The result analysis is done using online tool data-applied. Demo data from the tool is being used and FP Growth is applied on that pre-processed data set. The results found shows that pattern generation is fast and more accurate as compared to existing methodologies. Since web log data is huge and
existing works like Apriori is slow because it generates candidate set of keys, whereas FP Growth doesn’t use this intermediate step. Hence, results in fast pattern generation process. Figure 3: shows the online web data which is pre-processed using FP Growth algorithm on the data set named Web Clickstream. It shows a log of data having various entities like; date, city, visits, page views, visit duration, origin etc. Figure 4 shows the Data set when association rule is applied on Clickstream data; it shows the association between various data elements. Figure 5 shows the descriptive information about Web Clickstream data in text format. It also shows association rule that how data is gathered and the percentage of data which is a hit.

Figure 3: Online pre-processed web using FP Growth algorithm.

Figure 4: Data set after applying association
CONCLUSION

By the use of FP Growth algorithm in association rule mining we reduced the response time of most of the user queries thereby increasing proxy server hit ratio, it improves the overall performance of three tier web architecture. The user query is forwarded to proxy server through the listener and then a predictor maintains web log history and passes the hints to FP Growth block where association is applied on the loaded data set using association rule mining. This proposed framework helps in reducing network congestion and it also helps in providing security to the main server as by increasing the hit ratio in proxy server somehow we kept the user away from main server. By using this framework we can also reduce DOS (Denial of service) attacks up to some extent. This framework will also reduces the time for which a network channel is allocated to a user in TCP, as soon as a user gets results to its queries, the allocated channel will become free and network traffic can be reduced. The merits of this framework includes: increase in proxy server cache hit ratio. The simulation of this framework is performed using Data-Applied tool on data set named Web Clickstream collected from data-applied.com. Actual realization of this framework results in improved network architecture.

VI. FUTURE WORK

In this framework, change in hit ratio will result thereby changing the size of proxy server cache and threshold value. We are now focusing on realization of this framework under various circumstances and varying network conditions mostly when network channels are heavily loaded. We are also planning to investigate the performance of this framework in different networks like mobile network where mobility can affect the performance of a network.
REFERENCES


AUTHORS

Devender Banga is Pursuing M.Tech. from SGT Institute of Engineering and Technology, Gurgaon, India. His current research area is Web Usage Mining.

Sunitha Cheepurisetti is Assistant Professor in Department of Computer Science Engineering, SGT Institute of Engineering and Technology, Gurgaon, India. Her research area is Operating System, Networks and Security. She has more than 6 publications in various National/International Journals.