FUZZY C-MEANS CLUSTERING BASED PREFETCHING TO REDUCE WEB TRAFFIC

Neha Sharma and Sanjay Kumar Dubey
Amity University, Noida (U.P.), 201303, India

ABSTRACT

Web caching is used to reduce the network traffic but in the present scenario of heavy usage of web, caching itself is not enough. Hence, prefetching used with caching to improve the cache performance. Web caching and prefetching are the solution of many network problems like as access latency used by users and performance degradation due to heavy load on web. It helps in fetching the resources previously, which can be referenced by the user in the near future. It basically uses the concept of spatial locality of web objects. The data need to be stored, before the request comes for its access. For storage of such a prefetched data, web cache is used. For deciding the spatial locality of web objects, data is taken from proxy server. The paper proposes a framework for web traffic reduction. 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 pre-fetched.

KEYWORDS: *Prefetching, clustering, proxy server, cache, prediction.*

I. INTRODUCTION

From a simple information-sharing system, the web has now grown as a rich collection of dynamic and interactive services such as image, video conferencing etc. The tremendous growth of web has resulted into unavoidable demand of web servers and other network resources. While retrieving web pages from remote servers delay can be experienced. One solution is to increase the bandwidth, which will result in increased system cost. To reduce cost and to enhance the performance, cache-based approach is used.

A web cache is a technique used for the short-term caching of web objects, to reduce server load, bandwidth consumption and apparent lag. Cache only stores recent and frequently accessed objects. Web cache can be implemented using proxy server. Reverse Proxy acts as a bridge between web server and clients. Clients send their requests to proxy running between web server and client. The proxy connects the server and relays data between the user and the server. At each request, the proxy server is contacted first to find whether it has a valid copy of the requested object [1]. If the proxy has the requested object this is considered as a cache hit, otherwise a cache miss occurs and the proxy must forward the request on behalf of the user to web server. Upon receiving a new object, the proxy services provide a copy to the end-user and keep another copy to its local storage [2].

The web surfing pattern and web requests of a single user are same most of the time at a specific instant. Even different users also access same web document simultaneously. The web server load is also not distributed uniformly for all web objects access, few popular servers has the most of the load. Since, number of users requests same web object, so, if we store that common object in web cache, client will face lower latency. Web cache keeps the copy of frequently accessed objects closed to client side to reduce latency.

User web requests patterns are required to be pre-fetched to improve web performance, for which user's next request pattern need to be known previously. For this, techniques like as clustering, sequence analysis, markov models, association rule mining etc. are used. To increase the efficiency of web prediction techniques and to reduce the number of sequences to be selected (as cache size is limited), prediction rules are applied on clustered set of data. Fuzzy c-means clustering is performed on the preprocessed data and on those cluster sets Prediction by Partial Matching (PPM) technique of markov source applied to generate pre-fetching rules.

In this paper, a framework is proposed to support the prefetching criteria's on web servers. According to the framework, there are basically five steps to predict and prefetch the user's requests, viz. Data extraction from proxy web log, data preprocessing, data clustering, Prediction by partial matching and at last prefetching according to the PPM results obtained. The proposed framework has few merits over the previously used techniques, it gives better clustering results for overlapped data sets and also one data point may belong to more than one cluster, unlike k-means and other clustering algorithms.

The paper is divided into different sections for the ease. Section 2 provides an overview of primarily research done in this field. A lot of work has been done previously in this field and this part of paper explains the same. Section 3 concentrates on the proposed framework given in the paper. It's further divided into subsections preprocessing, Fuzzy c-means clustering, Partial Prediction Matching (PPM) and then after prefetching. In section 4, merits and demerits of the proposed technique are explained. Finally, the conclusion of the paper has been written, explaining the work done in brief with the merits and scope of the work in the present scenario.

II. LITERATURE REVIEW

Pallis *et. al* [1] addressed the short-term prefetching problem on a Web cache environment using an algorithm (clustWeb) 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.

Wan *et. al* [3] focuses on discovering latent factors of user browsing behaviours based on random indexing and detecting clusters of Web users according to their activity patterns acquired from access logs.

Podlipnig *et. al* [4] provides an exhaustive survey of cache replacement strategies proposed for Web caches. The paper concentrated on proposals for proxy caches that manage the cache replacement process at one specific proxy. A simple classification scheme for these replacement strategies was given and used for the description and general critique of the described replacement strategies. Although cache replacement is considered as a solved problem, we showed that there are still numerous areas for interesting research.

Greeshma *et. al* [5], web prefetching techniques and other directions of web prefetching are analyzed and discussed. These techniques are applied to reduce the network traffic and improve the user satisfaction. Web prefetching and caching can also be integrated to get better performance.

Kasthuri *et. al* [6] shows that ddeduction 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.

Santra *et al* [7] presented an efficient Cluster based Web Object Filters from Web Pre-fetching and web caching scheme to evaluate the web user navigation patterns and user references of product search. The proposed Web Pre-fetching and Web caching scheme efficiently integrates Web caching and pre-fetching contents. Clustering was done with the requested web page objects obtained from pre-fetched and web cached contents.

Lou *et. al* [8] investigated the problem of user transaction identification in proxy logs. In a proxy logs, a single user transaction may include pages references from one site as well as from multiple sites. Moreover, different types of transactions are not clearly bounded and are sometimes interleaved with each other as well as with noise. Thus an effective transaction identifier has to identify interleaved transactions and transactions with noise, and capture both the intra-site transactions and the inter-site

transactions. It presented a cut-and pick method for extracting all these transactions, by cutting on more reasonable transaction boundaries and by picking the right page sequences in each transaction.
Sathiyamoorthi *et. al* [9], authors discusses various data preprocessing techniques that are carried out at proxy server access log which generate Web access pattern and can also be used for further applications.
Venketesh *et. al* [10] presented a prediction model that built a Precedence Graph (PG) 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.
Patil *et. al* [11] proposed an Intelligent Predictive Web caching algorithm, IPGDSF#, capable of adapting its behavior based on access statistics.
Sharma and Dubey [12] provided the literature survey in the area of web mining. 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.

III. THE PROPOSED FRAMEWORK

Web caching is used to reduce network traffic and congestion by caching web at proxy. The paper presents a framework for prefetching scheme to improve the web performance. The prefetching scheme interprets the user's request depending upon their previous access behaviour. The previous access records of the user are extracted from reverse proxy log data and this extracted data is then preprocessed. Then after, the preprocessed data is clustered depending upon the user access patterns. Clustering is done using Fuzzy c-means based clustering approach. Till now, most of the researchers have used K-means clustering algorithm, but due to some limitations of k-means clustering algorithm, this paper presents a replacement algorithm for it. Clustered data is used for pattern finding for fulfilment of future requests of users. A markov model is used as web prediction technique here.
A Markov model is a finite-state machine where the next state depends only on the current state [13]. Associated with each arc of the finite-state machine network (a directed cyclic graph) is the probability of making the given transition [14]. The Prediction by Partial Matching (PPM) algorithm is used under it [15, 16, 17, and 18]. PPM algorithm uses markov model of m orders to store previous contexts.
Now, after applying PPM, i.e. analysis of user access sequences, prediction for the next request of the user will be obtained. Hence, depending upon this prediction technique web documents will be pre-fetched in the proxy server cache, so to improve the cache hit ratio. Improvement of cache hit ratio will lead to decrease in the latency and would also decrease web traffic. The architecture of the proposed framework has been shown in figure 1 and figure 2 shows the pseudocode for the framework steps.

3.1. Data Extraction

Firstly, data from reverse proxy web log extracted, as it keeps the user records for their previous requests. Data extraction is the process of retrieving data out from data sources for preprocessing. Depending upon these access patterns of users, prediction for next requests of users will be done. Hence, data is extracted very carefully from the proxy web log so that accurate information about the user access behaviour taken.

3.2. Data Preprocessing

Preprocessing means removal of noise and irrelevant information from the data set present. Preprocessing is important before applying clustering on the extracted data. Pre-processed data improves the efficiency and ease of mining process [19]. Data preprocessing is not a simple task, it consists of basically 4 important steps as explained below:

- First step, data cleaning, is the process of cleaning the data entries by filling the missing values, removing inconsistencies and smoothing the noisy data.
- Second step is data integration. Data might be taken from multiple resources for analysis or prediction, as used in the framework scenario. So, data taken from multiple resources need to

be clumped together into a single file, this process of clumping data from all the sources is known as data integration.

- Third step is data transformation of pre-processed data. Data transformation is the process of transforming the data into the forms relevant for the mining process.
- Fourth Step, data reduction is done. In data reduction, data set is reduced in a manner that after and before reduction data set produces the same analytical results. Basically, data reduced by smaller representations of the data, statistical reduction of data and data aggregation.

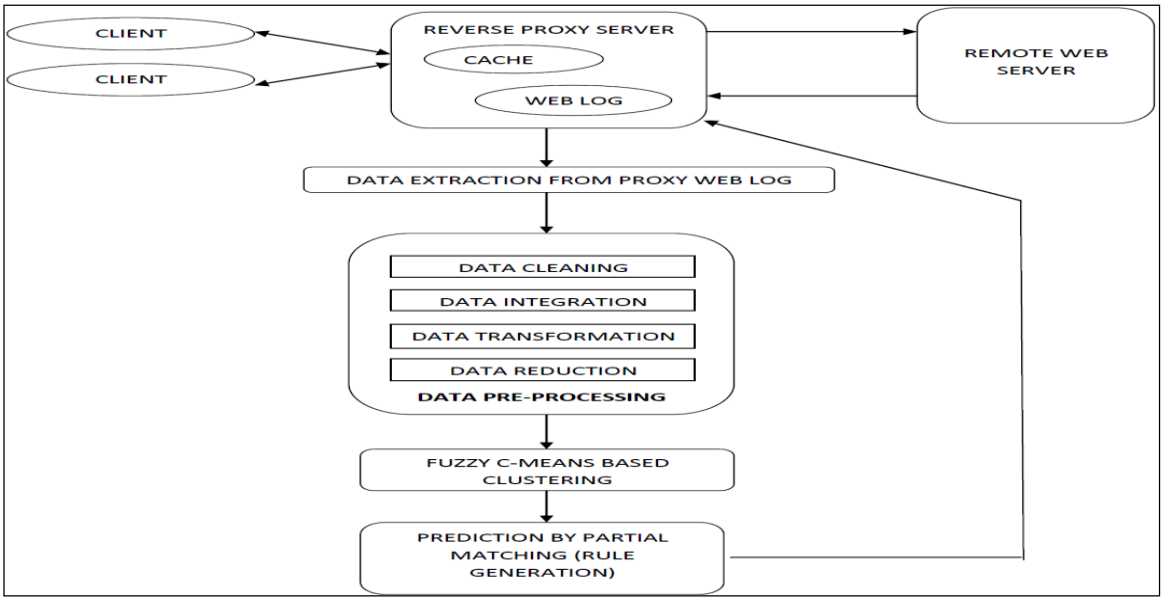


Figure 1. The Proposed Framework

framework_algo

Objective: To improve cache hit ratio.
Input: Proxy web log data
Output: Prefetching sequence pattern
1: Extract web log data from reverse proxy server.
2: Preprocess the extracted data.
3: Data cleaning
4: Data Integration
5: Data transformation
6: Data reduction
7: Begin
8: **Call** fuzzy_clust
9: **Call** build_ppm
10: **Call** predct_ppm
11: **end**

Figure 2. Pseudo code of Fuzzy c-means based clustering Algorithm

3.3. Fuzzy c-means based Clustering

The framework uses fuzzy c-means based clustering as the next step. It's a data clustering algorithm used for clustering the data sets. It works by assigning the membership to each data point corresponding to each cluster centre on the basis of distance between the cluster centre and data point. The pseudocode for the algorithm is also being provided if figure 5.

fuzzy_clust

Objective: Obtain the clusters of the given data sets.

Input:

X: set of data points
V: set of centres
m: any value from1 to infinity
c: number of cluster centres
u: fuzzy membership

Output: Clustered data set

1: **for**(i=1;i<=n;i++) **do**
2: **for**(j=1;j<=n;j++) **do**
3: **for**(k=1;k<=c;k++) **do**
4: $u_{ij}=1/(d_{ij}/dik)^{(2/m-1)}$
5: **end for**
6: **end for**
7: **end for**
8: **for**(i=1;i<=n;i++) **do**
9: **for**(j=1;j<=n;j++) **do**
10: $v_j=v_j + [(u_{ij})^m x_i/(u_{ij})^m]$
11: **end for**
12: **end for**

Figure 3. Pseudo code of Fuzzy c-means based clustering Algorithm

3.4. Web Prefetching - Prediction by Partial Matching (PPM)

Predictions are obtained from the comparison of the current context with each Markov model**.** Predictions are useful to improve cache performance as using prediction, the web documents which are required in near future will be kept in cache. Optimal page replacement algorithm will be used in cache. Pages in the cache will be replaced according to the optimal algorithm. Web prediction will let cache know, what might be the requests of the users in near future, accordingly data will be maintained in the cache and as a result cache hit will be improved. Figure 4 and figure 5 provides algorithm given by Domenech *et. al* for PPM evaluation [13].
To understand the concept of PPM lets we take a hypothetical example. Let we assume that the proxy log is having two user sessions at present for subject choice. First access sequence is DCS, CG, ADA, OS and second access sequence is like as DCS, CG, CS, OS. So, now the corresponding state graph for PPM is shown in figure 6.

IV. MERITS OF THE FRAMEWORK

The base of the framework is web cache and fuzzy c-means based clustering algorithm. There are number of advantages of using these in this framework, which actually are the merits of the proposed approach:

- It reduces bandwidth consumption resulting in reduction in network traffic and congestion.

build_ppm

Objective: Build a PPM prediction model
Input:
S: set of users' sessions
m: order of the Markov model
Output:
T: prediction model

```
1: for each session s € S do
2:      current-context[0]  = root node of T
3:      for j = 1 to m do
4:      current context[j]  = NULL
5:      end for
6: for each object i € s do
7:      for j = m down to 0 do
8:          if current context[j] has child node C representing i then
9:          increment node C occurrence count
10:         current context[j + 1] ← node C
11:         else
12:         construct child node C representing i
13:         node C occurrence count ← 1
14:         current context[j + 1] ←  node C
15:         end if
16:         current context[0] ← root node of T
17:         end for
18:     end for
19: end for
20: return T
```

Figure 4. Algorithm for making the prediction model in PPM [13]

predct_ppm

Objective: Obtain predictions of the next requests of the user
Input:
T : prediction model
Ri: last k user's accesses; $0 \leq i \leq k \leq$ order of the prediction model
th: threshold
Output:
P: Set of predictions

```
1: for j = 1 to k do
2:    current context[j]   node of depth j, representing the access sequence Rk−j+1, · · · ,Rk
3: end for
4: for j = k down to 1 do
5:      for each child node C of current context[j] do
6:          if (occurrence count of C) / (occurrence count of parent) ≥th then
7:          P←P Ů{a}
8:          end if
9:      end for
10: end for
11: remove duplicates from P
12: return P
```

Figure 5. Algorithm for making predictions in PPM [13]

- Access latency is also reduced.
- It disseminates the load of web server among proxies and leads to load reduction of remote server.
- If the remote server is not available due to the remote server's crash or network partitioning, it can access a cached copy at proxy. Thus, the robustness of the Web service is enhanced.
- It helps to analyse the access patterns of the organisation.
- Fuzzy c-means based clustering algorithm gives best results for overlapped data sets comparatively better than that of k-means algorithm (used by some researchers).
- One data point may belong to more than one cluster centre.

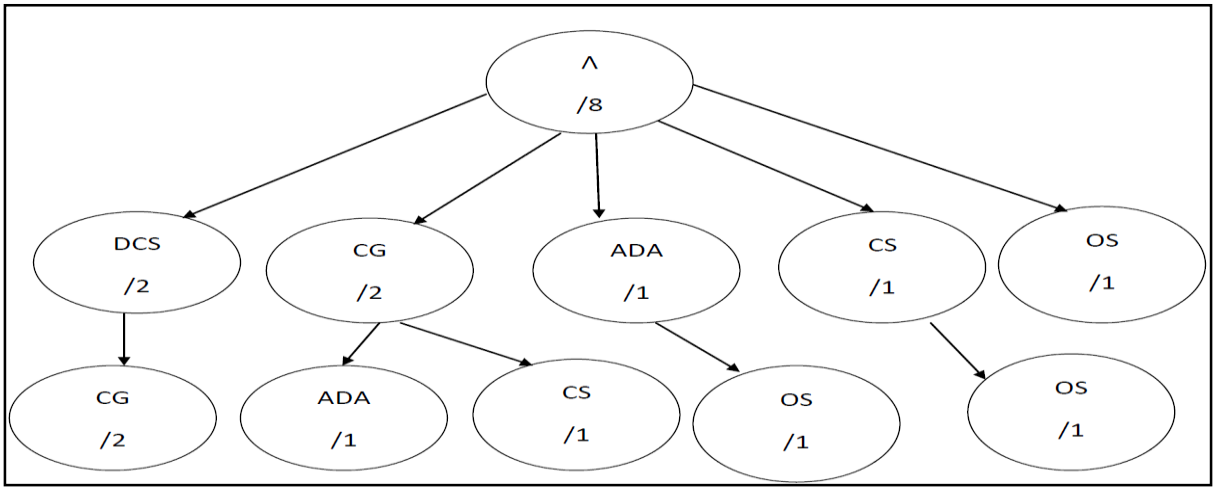
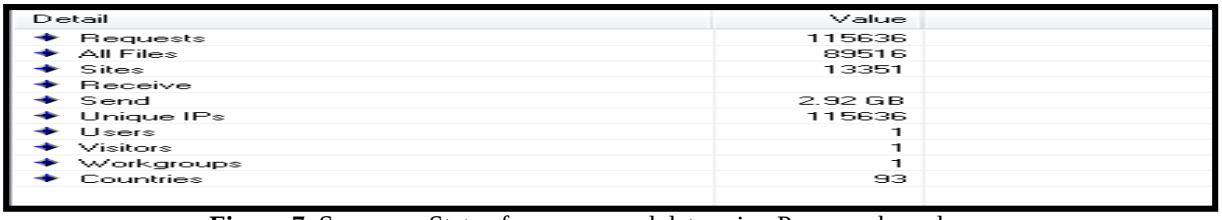


Figure 6. State graph of PPM Algorithm

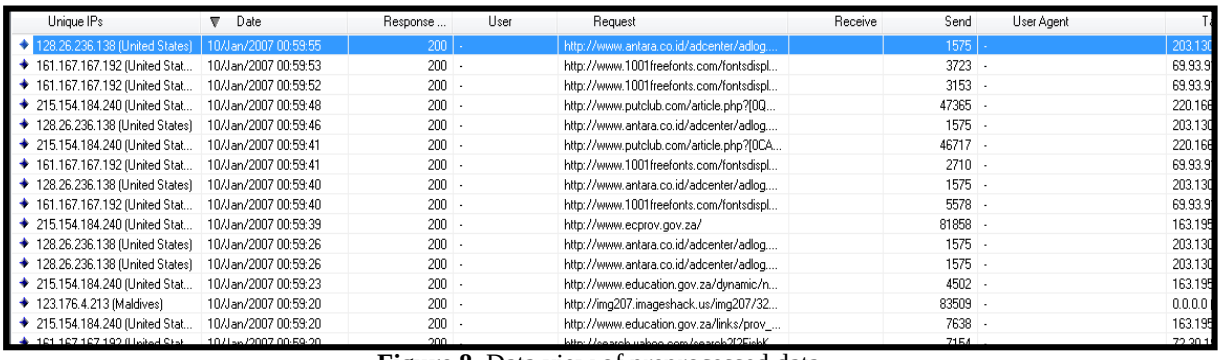
V. SIMULATION

In this paper, simulation is being performed using two tools. One is for preprocessing the data and another one for cluster formation. For illustration, dataset "*pa.sanitized-access.20070110.gz"* is used which was collected from *ftp://ircache.net*. This data set need to be preprocessed before clustering and for this purpose proxy log explorer was used. Then after, the preprocessed data set was clustered using GA Fuzzy Clustering tool.



Detail	Value
Requests	115636
All Files	89516
Sites	13351
Receive	
Send	2.92 GB
Unique IPs	115636
Users	1
Visitors	1
Workgroups	1
Countries	93

Figure 7. Summary Stats of preprocessed data using Proxy web explorer



Unique IPs	Date	Response ...	User	Request	Receive	Send	User Agent	T
128.26.236.138 (United States)	10/Jan/2007 00:59:55	200	-	http://www.antara.co.id/adcenter/adlog....		1575	-	203.130
161.167.167.192 (United Stat...	10/Jan/2007 00:59:53	200	-	http://www.1001freefonts.com/fontsdispl...		3723	-	69.93.9
161.167.167.192 (United Stat...	10/Jan/2007 00:59:52	200	-	http://www.1001freefonts.com/fontsdispl...		3153	-	69.93.9
215.154.184.240 (United Stat...	10/Jan/2007 00:59:48	200	-	http://www.putclub.com/article.php?[0Q...		47365	-	220.168
128.26.236.138 (United States)	10/Jan/2007 00:59:46	200	-	http://www.antara.co.id/adcenter/adlog....		1575	-	203.130
215.154.184.240 (United Stat...	10/Jan/2007 00:59:41	200	-	http://www.putclub.com/article.php?[0CA...		46717	-	220.168
161.167.167.192 (United Stat...	10/Jan/2007 00:59:41	200	-	http://www.1001freefonts.com/fontsdispl...		2710	-	69.93.9
128.26.236.138 (United States)	10/Jan/2007 00:59:40	200	-	http://www.antara.co.id/adcenter/adlog....		1575	-	203.130
161.167.167.192 (United Stat...	10/Jan/2007 00:59:40	200	-	http://www.1001freefonts.com/fontsdispl...		5578	-	69.93.9
215.154.184.240 (United Stat...	10/Jan/2007 00:59:39	200	-	http://www.ecprov.gov.za/		81858	-	163.195
128.26.236.138 (United States)	10/Jan/2007 00:59:26	200	-	http://www.antara.co.id/adcenter/adlog....		1575	-	203.130
128.26.236.138 (United States)	10/Jan/2007 00:59:26	200	-	http://www.antara.co.id/adcenter/adlog....		1575	-	203.130
215.154.184.240 (United Stat...	10/Jan/2007 00:59:23	200	-	http://www.education.gov.za/dynamic/n...		4502	-	163.195
123.176.4.213 (Maldives)	10/Jan/2007 00:59:20	200	-	http://img207.imageshack.us/img207/32...		83509	-	0.0.0.0
215.154.184.240 (United Stat...	10/Jan/2007 00:59:20	200	-	http://www.education.gov.za/links/prov_...		7638	-	163.195

Figure 8. Data view of preprocessed data

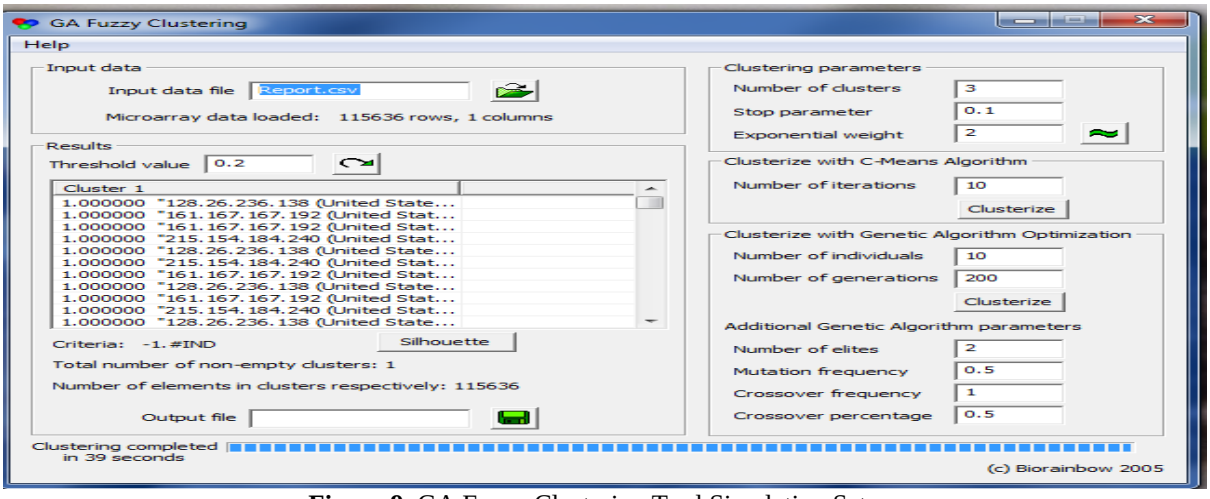


Figure 9. GA Fuzzy Clustering Tool Simulation Setup

Figure 7 depicts the Summary Stats of preprocessed data using Proxy web explorer and figure 8 represents the view of preprocessed data. Figure 9 shows GA Fuzzy Clustering Tool Simulation Setup and figure 10 shows the Cluster Set view. All this done using open source tools GA Fuzzy Clustering tool and Proxy Web explorer.

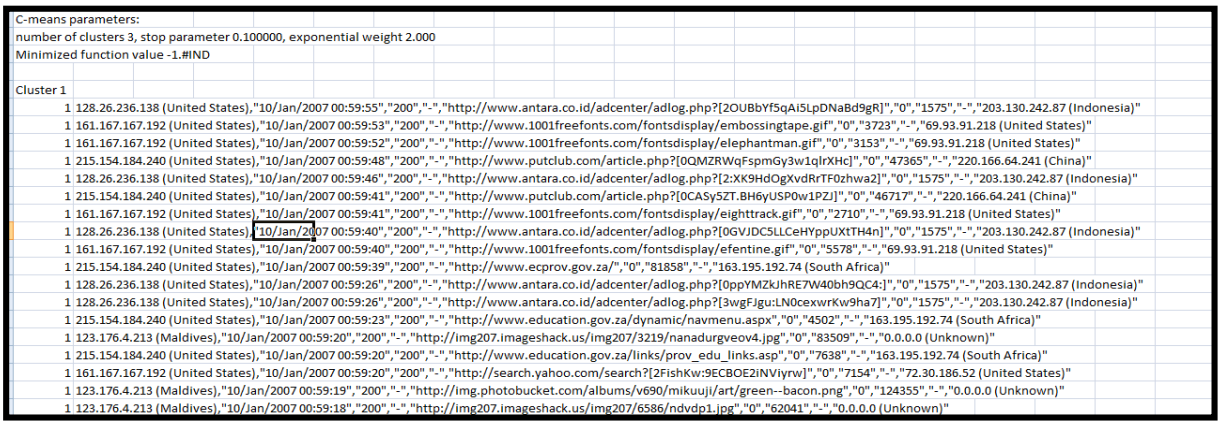


Figure 10. Cluster Set view

VI. CONCLUSION

In this paper, the problem of network congestion and web latency has been addressed. The paper proposes an efficient approach using prefetching for the same. Cache stores the copies of documents passing through it so that, if next time request comes for the same document, it can be fulfilled directly from the cache. The paper proposes a new approach to improve cache performance. After analysing the user access behaviour, proxy server can prefetch the next requests of the users.The paper basically changes the conventional technique by applying fuzzy c-means based clustering algorithm. The framework proposes that caching the data and pre-fetching the upcoming requests of the user's will improve the web server performance. The technique explains that proxy web log data must to be pre-processed and then clustered according to the user access behaviour. At last, PPM applied on the clustered set of data to generate the pre-fetching rules. For illustration, dataset "pa.sanitized-access.20070110.gz" is used which was collected from ftp://ircache.net. Clusters are made on the dataset using fuzzy c-means based clustering algorithm. Next step of the framework is to apply PPM so for that a dummy example taken and the technique and steps have been shown accordingly. It will result in performance improvement of the web via web latency reduction; cache hit ratio improvement; faster response of user requests and web traffic reduction. Application of this framework in the real world will actually improve the cache hit ratio and congestion problems in network.

VII. FUTURE WORK

For the future, our plan is to investigate the performance of web after applying the proposed framework and its implementation on proxy servers. Further, comparing the proposed approach with the present scenario and trying out other clustering algorithms for further improvement of the web are in the future scope. Finally, extending the use of the clustering and prefetching in other applications such as in mobile environment, content distribution network as well is another aim.

REFERENCES

[1]. Pallis G., A. Vakali and J. Pokorny, (2008) "A clustering-based prefetching scheme on a Web cache environment", Computers and Electrical Engineering 34, Elsevier, pg 309–323.
[2]. Chauhan K. A., V. Chauhan and R. Gupta, (2011)"Exploring the Web Caching Method to Improve the Web Efficiency", International Journal of Computer Applications in Engineering Sciences, Vol I, issue IV.
[3]. Wan M., A. Jonsson, C. Wang, L. Li, and Y. Yang, (2012) "A Random Indexing Approach for Web User Clustering and Web Prefetching", pp. 40–52, Springer-Verlag Berlin Heidelberg 2012.
[4]. Podlipnig S. and L. Boszormenyi, "A Survey of Web Cache Replacement Strategies", ACM Computing Surveys, Vol. 35, No. 4, December 2003, pp. 374–398.
[5]. Greeshma G. Vijayan and J. S. Jayasudha, (2012) "A survey on web pre-fetching and web caching techniques in a mobile environment" Natarajan Meghanathan, et al. (Eds): ITCS, SIP, JSE-2012, CS & IT 04, pp. 119–136.
[6]. Kasthuri I., M.A. Ranjit Kumar , K. Sudheer Babu, and Dr. S. S. S. Reddy, (2012) " An Advance Testimony for Weblog Prefetching Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4.
[7]. Santra A.K. and S. Jayasudha, (2012) "An Efficient Cluster Based Web Object Filters From Web Pre-Fetching And Web Caching On Web User Navigation", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 2, ISSN (Online): 1694-0814.
[8]. Lou W., G. Liu, H. Lu, and Q. Yang , (2002)"Cut-and-Pick Transactions for Proxy Log Mining", C.S. Jensen et al. (Eds.): EDBT 2002, LNCS 2287, pp. 88–105,Springer-Verlag Berlin Heidelberg.
[9]. Sathiyamoorthi V. and Dr. M. Bhaskaran, (2011)" Data Preprocessing Techniques for Pre-Fetching and Caching of Web Data through Proxy Server", IJCSNS International Journal of Computer Science and Network Security, Vol.11,No.11.
[10]. Venketesh P. and R.Venkatesan, (2011)"Graph based Prediction Model to Improve Web Prefetching", *International Journal of Computer Applications (0975 – 8887) Volume 36– No.10*.
[11]. Patil B. J. and B.V. Pawar, (2011)"Improving Performance on WWW using Intelligent Predictive Caching for Web Proxy Servers", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 1.
[12]. Sharma N. and S. K. Dubey, (2012) "A Hand to Hand Taxonomical Survey on Web Mining", International Journal of Computer Applications (0975 – 8887), Volume 60– No.3.
[13]. Domenech J., A. J. Gil, J. Sahuquillo, and A. Pont, (2007) "Evaluation, Analysis and Adaptation of Web Prefetching Techniques in Current Web", Universitat Politecnica de Valencia.
[14]. Mitchell D. C., R. A. Helzerman, L. H. Jamieson and M. P. Harper, (1993)" A parallel implementation of a hidden Markov model with duration modeling for speech recognition", In Proceedings of the Fifth IEEE Symposium on Parallel and Distributed Processing, Dallas, USA.
[15]. Chen X. and X. Zhang, (2002)"Popularity-Based PPM: An Effective Web Prefetching Technique for High Accuracy and Low Storage", In Proceedings of the International Conference on Parallel Processing, Vancouver, Canada.
[16]. Chen X. and X. Zhang, (2003)"A Popularity-Based Prediction Model for Web Prefetching. IEEE Computer", vol. 36, no. 3, pages 63–70.
[17]. Fan L., P. Cao, W. Lin and Q. Jacobson, (1999)"Web Prefetching Between Low-Bandwidth Clients and Proxies: Potential and Performance", In Proceedings of the ACM SIGMETRICS Conference on Measurement and Modeling of Computer Systems, pages 178–187, Atlanta, USA.
[18]. Palpanas T. and A. Mendelzon, (1999)"Web Prefetching Using Partial Match Prediction", In Proceedings of the 4th International Web Caching Workshop, San Diego, USA.
[19]. Han J. and Kamber M., (2006)"Data Mining: Concepts and Techniques", Second Edition, Morgan Kaufmann Publishers,Elsevier.

AUTHORS

Neha Sharma is Pursuing M.Tech. from Amity University, Noida, India.Her current research interest is in the field of web mining.

Sanjay Kumar Dubey is Assistant Professor in Department of Computer Science and Engineering, Amity University Noida, India. He has more than 45 publications in National/International Journals. His Research area Soft Computing, Data Mining and usability engineering.