International Journal of Soft Computing 8 (6): 435-439, 2013
ISSN: 1816-9503

Prefetching of Web Pages Based on Clustering Patterns for Enhancing the Performance of Web Retrieval Process

[1]H.K. Yogish and [2]G.T. Raju
[1]Department of Computer Science and Engineering, Bharathiar University,
641046 Coimbatore, Tamilnadu, India
[2]Department of Computer Science and Engineering, RNS Institute of Technology,
560061 Bangalore, Karnataka, India

Abstract: Web is a huge, explosive, diverse, dynamic and mostly unstructured data repository which supplies incredible amount of data and also raises the complexity of how to deal with the data because most of this data are not too much interest to most of the users, web service providers and business analysts. Web data is categorized into web content data, web usage data and web structure data. Based on the type of web data used in mining, web mining has three categories-web content mining, web usage mining and web structure mining. Usage data is related to web activity. This study presents a methodology which preprocesses the web data, extract and cluster using ART1 NN technique and preloads web pages which are likely to be accessed in near future.

Key words: Preprocessing, clustering, web usage data, ART1 NN, prefetching

INTRODUCTION

Today, the World Wide Web (WWW) has turned to be the largest information source available in the planet. It is huge, explosive, diverse, dynamic and mostly unstructured data repository which supplies incredible amount of data and also raises the complexity of how to deal with the data because most of this data are not too much interest to most of the users, web service providers and business analysts. The users want to have the effective search tools to find relevant information easily and precisely. The web service providers want to find the way to predict the user's behaviours and personalize information to reduce the traffic and design the web site suited for the different group of users. The business analysts want to have tools to learn the users/consumers' needs. Due to this abundance, it became essential for finding ways in extracting relevant data from this ocean of data. During the past decade, researchers have proposed a new unifying area called web mining (Cooley *et al*., 1999; Mobasher *et al*., 2011).

Web mining is a very hot research topic which combines two research areas: Data mining and World Wide Web which apply data mining techniques to web data with intelligent analysis. Based on the type of web data used in mining, web mining has three categories (Cooley *et al*., 1999; Mobasher *et al*., 2011), web content mining, web usage mining (Chen and Zhang, 2002; Palpanas and Mendelzon, 1999) and web structure mining. Web content mining focuses on the discovery/ retrieval of the useful information from the web contents/ data/documents while the web structure mining emphasizes to the discovery of how to model the underlying link structures of the web. Web usage mining tries to discovery the useful information from the secondary data derived from the interactions of the users while surfing on the web.

LITERATURE REVIEW

Prefetching means fetching the URL objects before the users request them. The web prefetching approaches can be characterized into following:

Short-term prefetching: Future requests are predicted to the cache's recent access history. Based on these predictions, clusters of web objects are pre-fetched (Palpanas and Mendelzon, 1999). In this context, the short-term prefetching schemes use Dependency Graph (DG) where the patterns of accesses are held by a graph and Prediction by Partial Matching (PPM) where a scheme is used adopted from the text compression domain (Chen and Zhang, 2002; Deshpande and Karypis, 2001). In addition, several short-term prefetching policies (Padmanabhan and Mogul, 1996; Cadez *et al*., 2003)

Corresponding Author: H.K. Yogish, Department of Computer Science and Engineering, Bharathiar University,
641046 Coimbatore, Tamilnadu, India

are based on Markov Models which are used for modelling and predicting user's browsing behaviour over the web.

Long-term prefetching policies: Global object access pattern statistics (e.g., objects' popularity, objects' consistency) are used to identify valuable (clusters of) objects for prefetching. In this type of scheme, the objects with higher access frequencies and no longer update time intervals are more likely to be prefetched (Chakrabarti, 2003).

The main objective of this study is to demonstrate that web prefetching is an effective solution to reduce web latency perceived by the users. The major research challenges includes pre-processing (Cooley *et al*., 1999) of the large raw web usage data, extractions of interesting and potentially useful patterns and reduce the significant increase in user perceived latency due to heavy traffic in the internet.

Proposed methodology: To address the challenges mentioned above this study propose a complete pre-processing methodology, a novel approach to extract cluster patterns and use of clusters in prefetching. Figure 1 shows the architecture of the proposed methodology.

Data pre-processing: This is the most time consuming and important task of the web usage mining process. It transforming the web usage data into a Relational Data Model and also removes the irrelevant entries such as image, graphics, audio and video files from the weblog data. The data used for this research is from the NASA server logs which is in the common log file format. The architecture of the proposed preprocessing methodology is shown in Fig. 2.

The objectives of the preprocessing are: to convert the raw log file into a set of transactions, to discharge the non-interesting or noisy requests and to reduce the quantity of data being analyzed and to enhance its quality. The tasks performing in the pre-processing step are (Cooley *et al*., 1999) given.

Data cleaning: Web log is examined to remove irrelevant information for example the log entries with figures (jpg, gif, etc.) can be removed.

User identification: The user identification plays a significant role to identify the distinct and unique users of website. Although, users alone play no role in web session clustering, they provide significant information about who the distinctive website users are.

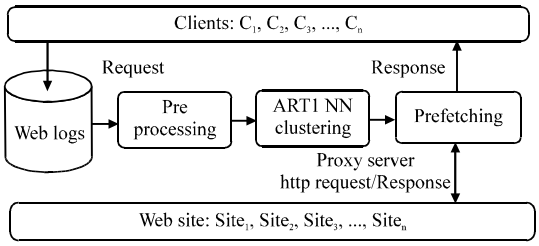

Fig. 1: Architecture of proposed methodology

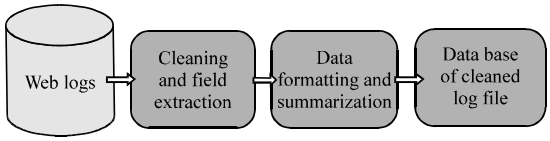

Fig. 2: Architecture of preprocessing methodology

Session identification: A session is a set of page references from one source site during one logical period. Historically a session would be identified by a user logging into a computer, performing work and then logging off. Identifying the user sessions from the log file is complex task due to proxy servers, dynamic addresses and cases where multiple users access the same computer (at a library, Internet cafe, etc.) or one user uses multiple browsers or computers (Cadez *et al*., 2003). In most of the session identification techniques, 30 min timeout was taken and transactions made by user with web site are in 30 min are grouped as session (Fu *et al*., 1999; Raju and Satyanarayana, 2007).

Discovery of cluster patterns: Given the input binary pattern vector P_H, the problem is to group N patterns into C clusters such that the entries within clusters are more similar than across clusters (Fu *et al*., 1999; Zhang *et al*., 1996; Pallis *et al*., 2005). A clustering algorithm takes a set of input vectors and gives a set of clusters as output thus mapping of each input vector to a cluster. The technique adopted in this study for clustering of web users and pages based on Adaptive Resonance Theory 1 (ART1) Neural Network (NN) that uses unsupervised learning, it partitions pre-processed log data into several high quality clusters without prior information about the number of clusters. Since, ART1 NN is adaptive in nature, it is well suited to understand the user interests and the changes in these interests over a period of time. Figure 3 shows the proposed clustering methodology. For each host H, the feature extractor forms an input binary pattern vector P that is derived from the base vector D. The procedure given in Fig. 4 generates the pattern vector which is the input vector for ART1 NN based clustering algorithm.

Prefetching: This technique is used to reduce the user perceived latency present in every web based application, the prefetching module as shown in the Fig. 1 prefetches the URLs that are most frequently accessed by all the members (hosts) of that cluster represented by a prototype vector. The proxy server responds to the client with prefetched URLs (Chen and Zhang, 2002; Palpanas and Mendelzon, 1999; Deshpande and Karypis, 2001). The prefetching accuracy is measured by predicting the URLs for each member of the cluster and then the prediction is verified with access logs recorded for the next t days (prediction period).

An advantage of the proposed approach is that better network resource utilization by prefetching the web pages for each cluster of web users instead of individual user. The pseudo code of proposed approach is given in the study.

Pseudo code for prefetching:

```
prefetcher (Host_Id )
 { //Takes input:  Host_Id of the host that request a URL, Cluster the hosts
using ART1 NN Clustering algorithm;
   ART1_Clustering (P,  ρ, n );
 /* P is the array of pattern vectors and ρ is the Vigilance parameter. Let 'n'
is the number of clusters  and C1,  C2,…, Cn are the clusters represented by
the prototype vectors. The prototype vector for the kth cluster is of
the form Tk = ( tk1, tk2, …, tkm) where tkj for j = 1, 2,..., m are the top-down
weights corresponding  to node k in layer F2 of the ART neural network.
*/
        Initialize count = 0;
        Repeat for each cluster Ck of the n clusters
```

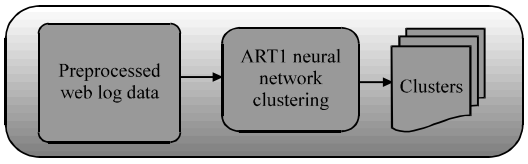

Fig. 3: Clustering Model

```
Procedure:  Pattern_Vector_Genarator( )
Begin
 for each pattern vector PH, where H = 1 to n
       for each element pi in pattern vector PH, i = 1 to m
            if URLi requested by the host more than twice
                        then pi = 1;
                 else
                        pi = 0;
  End
```

Fig. 4: Procedure for generating pattern vector

```
        If (Host_ID is a member of cluster Ck)
        {
                Repeat for j = 1,2,…m
                        If (tkj = 1)  {
                          Prefetched_URLs [Count] = URLi
                          Count++; }
            }
          Return Prefetched_URLs [];
} // End of CPF prefetching scheme
```

EXPERIMENTAL RESULTS

Experiments have been conducted on log files collected from NASA web site and academic site. The results are shown in Table 1. It shows that the proposed preprocessing methodology reduces the initial size of the log data to 72-82% by eliminating unnecessary requests and also increases their quality through better structuring of the web data.

Experiments were conducted for varying number of hosts. The value of the vigilance parameter ρ which controls the number of clusters to be formed is varied between 0.3 and 0.5 (for quality clusters). The following observations are made from the experimental results as shown in the Fig. 5.

The value of the vigilance parameter ρ which controls the degree of mismatch is varied between 0.3 and 0.5 (for quality clusters). Increasing the value of the vigilance parameter 'ρ' increases the number of clusters. Lower value of ρ causes more number of hosts to be in one cluster resulting in lesser number of clusters.

Higher value of ρ causes smaller number of hosts to be in one cluster causing more number of clusters. ART1 takes less time compared to SOM and K-Means and hence appears to be better.

Table 2 presents the results obtained by executing the prefetching model for 6 days (1/Aug/1995 to 6/Aug/1995). Figure 6 shows the web traffic (the number of URLs requested by each host/users). It is observed from the Fig. 7 that the prediction accuracy ranges from 83.33-98.38%. A deviation has occurred in one case (U12 in third row) in which the user does not request any of the prefetched URLs. Experimental results reveal that there is a significant reduction in user perceived latency with an average prediction accuracy of 93.16%.

Table 1: The results after preprocessing

Web sites	Duration	Original size	Size after preprocessing	Reduction in size (%)	No. of sessions	No. of users
NASA	1-10th Aug, 1995	75361 bytes (7.6 MB)	20362 bytes	72.98	6821	5421
NASA	20-24th July, 1995	205532 bytes (20.6 MB)	57092 bytes	72.22	16810	12525
Academic web site	12-28th May, 2001	28972 bytes (2.9 MB)	5043 bytes	82.50	1645	936

Hosts	Time (sec)		
	ART	K-means	SOM
100	0.156	0.188	0.422
250	0.875	0.931	1.797
500	5.156	10.219	23.359
1000	10.797	28.891	45.625

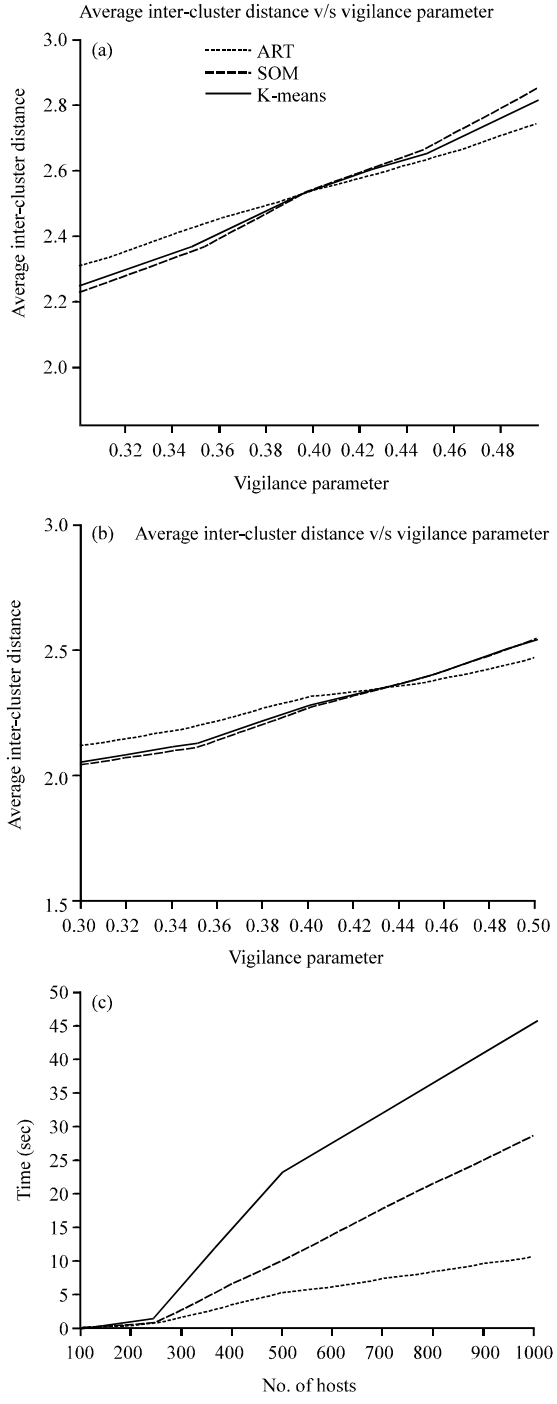

Fig. 5: a) No. of hosts 1000; b) No. of hosts 500 and c) Time complexity of clustering algorithms

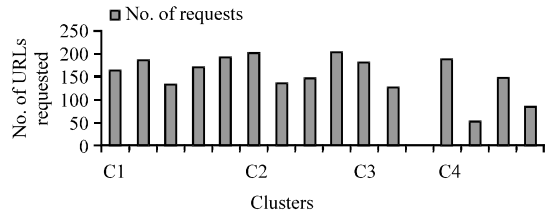

Fig. 6: Web requests traffic (NASA log data, Aug. 1995, 6 days)

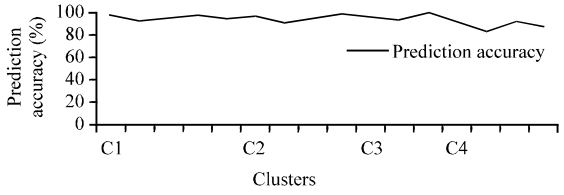

Fig. 7: Prediction Accuracy (NASA log data, Aug. 1995, 6 days)

Table 2: Results of Prefetching Model (NASA log data, Aug. 1995, 6 days)

Cluster Id	Users in clusters	User Id	No. of requests	No. of URLs prefetched	Hits	Prediction accuracy
C1	U1, U2, U3, U4, U5	1	162	36	35	97.22
		2	184	-	33	91.66
		3	132	-	34	94.44
		4	168	-	35	97.22
		5	190	-	34	94.44
C2	U6, U7, U8, U9	6	200	62	60	96.77
		7	135	-	56	90.32
		8	146	-	58	93.54
		9	202	-	61	98.38
C3	U10, U11, U12	10	181	28	27	96.42
		11	126	-	26	92.85
		12	0	-	-	100.00
C4	U13, U14, U15, U16	13	186	24	22	91.67
		14	54	-	20	83.30
		15	147	-	22	91.67
		16	85	-	21	87.50

CONCLUSION

Web prefetching has been researched for years with the aim of reducing the user perceived latency, however studies mainly focus on prediction performance rather than on the user's point of view. The prefetching mechanisms is to be able to accurately predict which pages will be needed next, to minimize mistakes that result in wasted bandwidth and increased server loads. Hence, a fast and accurate prediction is crucial for prefetching performance so that proposed research shows its usefulness in reasonable utilization of network resources through prefetching of web pages for a community of users instead of a single user that is eliminates all the demerits of existing systems.

REFERENCES

Cadez, I., D. Heckerman, C. Meek, P. Smyth and S. White, 2003. Model-based clustering and visualization of navigation patterns on a web site. Data Min., Knowl. Discov., 7: 399-424.

Chakrabarti, S., 2003. Mining the Web: Discovering Knowledge from Hypertext Data. Part 2. Morgan Kaufmann Publishers, California, ISBN: 978155860 7545, Pages: 345.

Chen, X. and X. Zhang, 2002. Popularity-based PPM: An effective web prefetching technique for high accuracy and low storage. Proceedings of the International Conference on Parallel Processing, August 20-23, 2002, Canada, Vancouver, pp: 296-304.

Cooley, R., B. Mobasher and J. Srivastava, 1999. Data preparation for mining world wide web browsing patterns. Knowledge Infor. Syst., 1: 5-32.

Deshpande, M. and G. Karypis, 2001. Selective Markov models for predicting web-page accesses. Proceedings of the 1st SIAM International Conference on Data Mining, April 5-7, 2001, Chicago, USA.

Fu, Y., K. Sandhu and M.Y. Shih, 1999. Clustering of web users based on access patterns. Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining, August 15-18, 1999, San Diego, CA., USA.

Mobasher, B., N. Jain, E.H. Han and J. Srivastava, 2011. Web mining-pattern discovery from world wide web transactions: Technical report TR96-050. Department of Computer Science, University of Minnesota.

Padmanabhan, V. and J. Mogul, 1996. Using predictive prefetching to improve World Wide Web latency. ACM SIGCOMM Comput. Commun. Rev., 26: 22-36.

Pallis, G., L. Angelis and A. Vakali, 2005. Model-based cluster analysis for web users sessions. Proceeding of the 15th International Symposium on Methodologies for Intelligent Systems, May 25-28, 2005, Saratoga, Springs, pp: 219-227.

Palpanas, T. and A. Mendelzon, 1999. Web prefetching using partial match prediction. Proceedings of the 4th International Web Caching Workshop, March 31-April 2, 1999, Diego, California.

Raju, G.T. and P.S. Satyanarayana, 2007. Knowledge discovery from web usage data: Extraction of sequential patterns through ART1 neural network based clustering algorithm. Proceedings of the International Conference on Computational Intelligence and Multimedia Applications, Volume 2, December 13-15, 2007, Sivakasi, Tamil Nadu, pp: 88-92.

Zhang, T., R. Ramakrishnan and M. Livny, 1996. BIRCH: An efficient data clustering method for very large data bases. Proceedings of the ACM SIGMOD International Conference on Management of Data, June 4-6, 1996, Cananda, pp: 103-114.