Available online at www.sciencedirect.com

SCIENCE DIRECT

Journal of Network and
Computer Applications 28 (2005) 147–165

Journal of NETWORK and COMPUTER APPLICATIONS

www.elsevier.com/locate/jnca

Intelligent web traffic mining and analysis

Xiaozhe Wang[a,*], Ajith Abraham[b], Kate A. Smith[a]

[a] *School of Business Systems, Faculty of Information Technology, Monash University, Clayton, Victoria 3800, Australia*

[b] *Department of Computer Science, Oklahoma State University, 700 N Greenwood Avenue, Tulsa, OK 74106-0700, USA*

Received 7 January 2004; accepted 7 January 2004

Abstract

With the rapid increasing popularity of the WWW, Websites are playing a crucial role to convey knowledge and information to the end users. Discovering hidden and meaningful information about Web users usage patterns is critical to determine effective marketing strategies to optimize the Web server usage for accommodating future growth. Most of the currently available Web server analysis tools provide only explicitly and statistical information without real useful knowledge for Web managers. The task of mining useful information becomes more challenging when the Web traffic volume is enormous and keeps on growing. In this paper, we propose a concurrent neuro-fuzzy model to discover and analyze useful knowledge from the available Web log data. We made use of the cluster information generated by a self organizing map for pattern analysis and a fuzzy inference system to capture the chaotic trend to provide short-term (hourly) and long-term (daily) Web traffic trend predictions. Empirical results clearly demonstrate that the proposed hybrid approach is efficient for mining and predicting Web server traffic and could be extended to other Web environments as well.

Keywords: Web traffic; Self organizing map; Fuzzy inference system

1. Introduction and Motivation for Research

The World Wide Web (WWW) is continuously growing with the information transaction volume from Web servers and the number of requests from Web users. Providing Web administrators with meaningful information about users access behavior (Levene and Loizou, 1999) and usage patterns (Brin, 1998; Buchner and Mulvenna, 1998)

[*] Corresponding author.

1084-8045/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jnca.2004.01.006

has become a necessity to improve the quality of Web information service performances. As such, the hidden knowledge obtained (Buchner et al., 1999) from mining Web server traffic and user access patterns could be applied directly for marketing and management of E-business, E-services, E-searching, E-education and so on.

However, the statistical data (Hastie et al., 2001) available from the normal Web log files (Masseglia et al., 1999; Pohle and Spihopoulou, 2002) or even the information provided by most conventional Web server analysis tools including commercial Web trackers could only provide explicit information due to the natural limitation of statistic methodology used. Computational Web Intelligence (CWI), a recently coined paradigm, is aimed to improve the quality of intelligence in the Web technology (Pal et al., 2002; Zhang and Lin, 2002) and includes Web mining as one main stream. Generally, the Web analysis relies on three general sets of information given: (i) past usage patterns (ii) degree of shared content (Boley et al., 1999) and (iii) inter-memory associative link structures (Pirolli et al., 1996) corresponding to the three subsets in Web mining namely: (i) Web usage mining (Masseglia et al., 1999) (ii) Web content mining and (iii) Web structure mining. In Web usage mining, the pattern discovery consists of several steps including statistical analysis, clustering (Kitsuregawa et al., 2002; Lingras, 2002), classification and so on (Ng and Smith, 2000; Srivastava et al., 2000; Wang et al., 2002). For instance, in E-commerce, analyzing the Web usage data can help organizations to understand the customer Web browsing patterns (Cheung et al., 1997) which inturn might help to facilitate E-commerce specific processing such as: Web structure management for designing a better Website, and promotional campaigns for building customized advertisements and for making better strategic marketing decisions (Chang et al., 2001). Most of the current research are focusing on finding patterns but with little effort on the detailed pattern/trend analysis that varies with the Web environments and the intelligent paradigms considered (Wang et al., 2002; Cheung et al., 1997, 2001; Han et al., 2000; Jespersen et al., 2002).

The Web personalization system proposed by Mobasher et al. (1999) is based on the direct clustering of URLs instead of clustering the user sessions. This work comprises of two components: (i) offline tasks related to the mining of usage data and (ii) online process of automatic Web page customization based on the knowledge discovered. In the offline part, association rule hypergraph partitioning technique is used to provide automatic filtering capabilities to capture the relationship among items based on their patterns of co-occurrence across transactions. Such groups of items are referred to as frequent item sets that are used as hyerredges to form a hypergraph, and then partitioned into set of clusters.

In ‘LumberJack’ (Chi et al., 2002) proposed by Chi et al., user profiles were built up by combining both user session clustering and traditional statistical traffic analysis. K-means algorithm was used to cluster the data and the retrieved content (data features) including users’ navigation paths, page viewing time, hyperlink structure, and page content. Then they were used for building up a vector space model in order to find the Web users’ usage patterns represented as online user profiles (Martín-Bautista et al., 2002; Pazzani and Billsus, 1997).

Joshi et al. (1999) used a relational OLAP approach for creating a Web log warehouse using access logs (raw log data) and mined logs (association rules and clusters discovered from logs). Data pre-processing was used to clean and format the log data into formats that

Oracle SQL Loader could accept. The warehousing used CGI and SQLPLUS and the Web interface used Perl CGI scripts. SGI's 'Mineset' (Agrawal and Srikant, 1994) implements a variation of the a priori algorithm to discover association rules and fuzzy C-medoids algorithm developed by Krishnapuram et al. (1999) to cluster and capture a graded notion of the similarity between sessions. It was possible to analyze both the Web logs and traversal patterns using an online query. Zaïane et al. (1998) combined OLAP and data mining techniques for finding usage patterns.

To discover the access patterns and sequential navigational patterns, association-rule mining and clustering had been used in many research projects (Berkan and Trubatch, 2002; Chen and Kuo, 2000). For general access pattern tracking, some of the popular projects are 'WebLogMining' (Zaïane, 2001), 'WUM' (Spiliopoulou and Faulstich, 1998) and 'WebSIFT' (Cooley et al., 1999). Customized usage tracking research includes 'Adaptive Web sites' (Perkowitz and Etzioni, 1998) and WebWatcher (Joachims et al., 1997) and so on. More detailed overview of Web usage mining research could be obtained from (Srivastava et al., 2000; Wang et al., 2002; Cheung et al., 1997, 2001).

In our research, we used a clustering algorithm to discover hidden relationships among the Web server data and access patterns. The unsupervised learning algorithm Self Organising Map (SOM) (Kohonen, 1990) was used for the clustering task to discover usage patterns from Web server logs. The clustered data and clustering information (Fu et al., 1999) were further used for learning the trend patterns by using statistical analysis methods. In order to make the analysis more intelligent we also used the clustered data to predict the daily and hourly traffic including request volume and page volume. Using a Takagi Sugeno Fuzzy Inference System (TSFIS) (Sugeno, 1985), we explored the prediction of the daily request volume (1–5 days ahead) in a week and the hourly page volume (1, 12 and 24 h ahead) in a day.

We analyzed the Web user access and server usage patterns of Monash University's main Web server located at http://www.monash.edu.au. We made use of the statistical/text log file data provided by Web log analyzer 'Analog' (Analog, 2002) which is a popular Web server analysis tool. It can generate numerical and text information based on original server log files covering different aspects of the users access log records. The weekly based reports include traffic volume, types of files accessed, domain summary, operating system used, navigation and soon. The typical Web traffic patterns of Monash University in Fig. 1(a) and (b) are showing the daily and hourly Web traffic (request volume and page volume) on the main server site for the week starting from 14-Jul-2002, 00:13 A.M to 20-Jul-2002, 12:22 A.M.

Generally, in a week, Monash University's main Web server (Server Usage Statistics) receives over 7 million hits. Since the data is not only large but also cover different aspects (domains, files accessed, daily and hourly access volume, page requests, etc.), it becomes a real challenge to discover hidden information or to extract usage patterns from such data sets. It becomes more difficult when the traffic volume keeps on growing due to the growth of the organization itself. The mere complexity of the data volume paves way for the requirement of hybrid intelligent systems for intelligent information analysis and trend prediction. In the subsequent Section 2, we show the structure of the proposed hybrid neuro-fuzzy model for mining Web usage patterns. In Section 3, we present the analysis of the clustered Web data using SOM followed by modeling the TSFIS to learn and predict

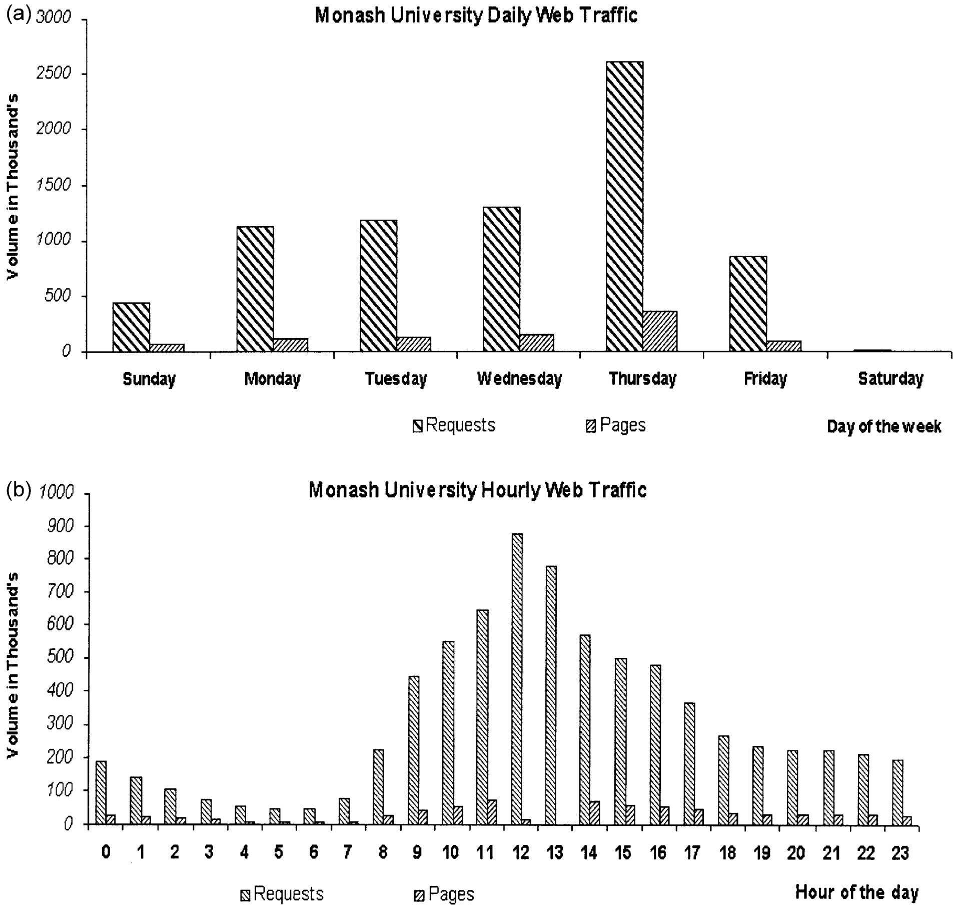


Fig. 1. Monash University's daily and hourly web traffic patterns, (a) Daily website traffic (request volume and page volume) in a week, (b) Hourly website traffic (request volume and page volume) in a day.

the short-term and long-term usage patterns in Section 4. Finally, some conclusions and future works are given in Section 5.

2. Hybrid neuro-fuzzy approach for web traffic mining and prediction

The hybrid framework combines SOM and Fuzzy Inference System (FIS) operating in a concurrent environment as shown in Fig. 2. In this concurrent model, neural network assists the fuzzy system continuously to determine the required parameters especially when certain input variables cannot be measured directly. Such combinations do not optimise the fuzzy system but only aids to improve the performance of the overall system (Abraham, 2001). Learning takes place only in the neural network and the fuzzy system remains unchanged during this phase. The pre-processed data (after cleaning and scaling)

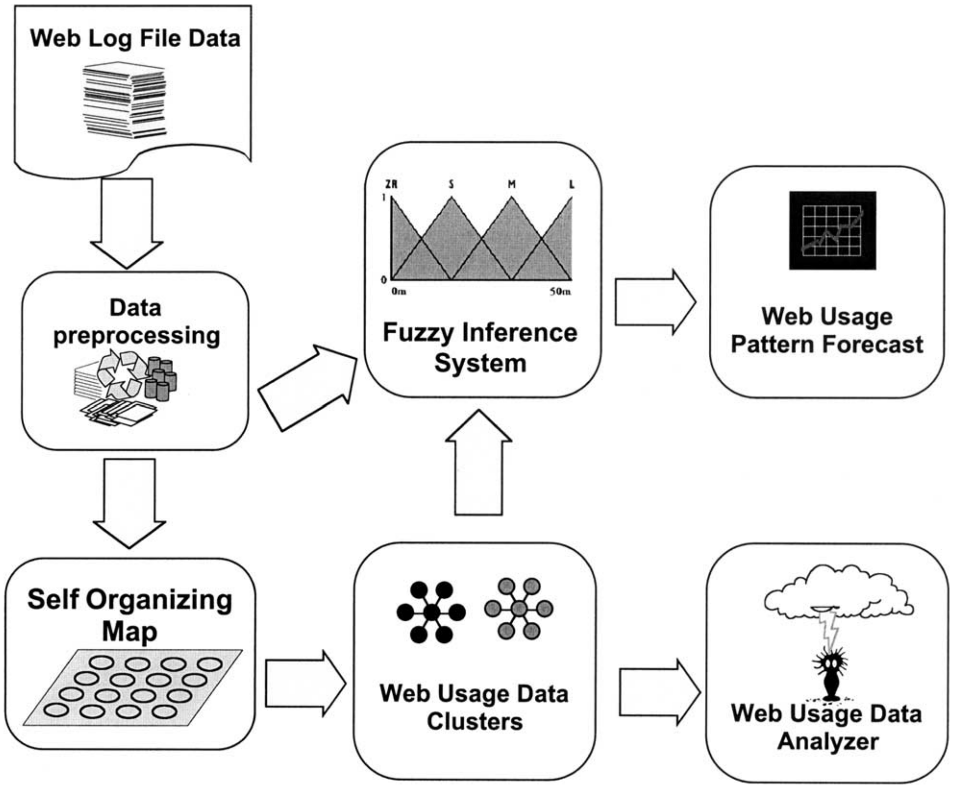


Fig. 2. Architecture of the concurrent neuro-fuzzy model for web traffic mining.

is fed to the SOM to identify the data clusters. The clustering phase is based on SOM—an unsupervised learning algorithm (Kohonen, 1990), which can accept input objects described by their features and place them on a two-dimensional (2D) map in such a way that similar objects are placed close together.

Referring to Fig. 3, data *U*, *V* and *X* may be segregated into three different clusters according to the SOM algorithm and they have the different degree association with each cluster. The clustered data by SOM are then used by ‘Web Usage Data Analyzer (WUDA)’ for discovering different Web traffic patterns and providing useful information to the Web analysts.

Fig. 2. Architecture of the Concurrent Neuro-Fuzzy Model for Web Traffic Mining FIS is used to learn the chaotic (Coenen et al., 2000) short-term and long-term Web traffic patterns (example depicted in Fig. l(a) and (b)). FIS is a popular computing framework based on the concepts of fuzzy set theory, fuzzy *if-then* rules, and fuzzy reasoning. The basic structure of the FIS consists of three conceptual components: (i) a rule base, which contains a selection of fuzzy rules; (ii) a database, which defines the membership functions used in the fuzzy rule and (iii) a reasoning mechanism, which performs the inference procedure upon the rules and given facts to derive a reasonable output or conclusion.

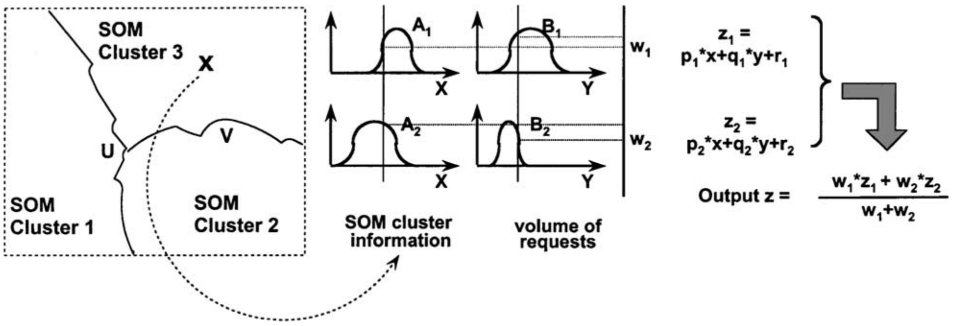


Fig. 3. Fuzzy association of data with SOM generated clusters.

As shown in Fig. 3, data X is associated with Cluster 3 strongly but data U and V only have weak associations with the associated clusters (Cluster 1 and Cluster 2).

Example: Data U is associated with Cluster 1 but can be also considered to have some weak association with Clusters 2 and 3. The degree of association of the data with a particular cluster is modeled as an additional fuzzy variable. We used the TSFIS in which the rule consequence is constituted by a weighted linear combination of the crisp inputs rather than a fuzzy set and has the following structure (Sugeno, 1985):

$$\text{If } x \text{ is } A_1 \text{ and is } B_1, \text{ then } f_1 = p_1 x + q_{1y} + r_1 \tag{1}$$

The inference method uses two input variables (cluster information from SOM and volume of requests) as shown in Fig. 3. Based on the clustered data of Web server log files, our target is to let the FIS learn and capture the chaotic web traffic flow and provide a short-term (hourly) and long-term (daily) server access trend prediction few time steps ahead. Compared to a neural network an important advantage of the FIS is the interpretability of the developed model in the form of simple if-then rules.

3. Data clustering and experimental analysis using SOM

Web usage mining normally contains four processing stages including data collection, data preprocessing, pattern discovery and pattern analysis (Chang et al., 2001) The data source selected for our approach is from the Web traffic data generated by the ‘Analog’ Web access log file analyzer (Analog, 2002). It is a usual practice to embed Web trackers or Web log analysis tools to analyze the Web server log files for providing useful information to Web administrators. After browsing through some of the features of the best trackers available on the market (Wang et al., 2002), it is easy to conclude that rather than generating basic statistical data they really cannot provide much meaningful information. In order to overcome the drawbacks of available Web log analyzers, the hybrid approach is proposed to discover hidden information and usage pattern trends, which could aid the Web managers for improving the management, performance and controlling of the Web servers. In our approach, after selecting the required data from the large data source, all the logs were cleaned, formatted and scaled to feed the SOM clustering algorithm. The SOM data clusters could be presented as 2D maps for each Web traffic feature, and WUDA was used for detailed Web user access and server usage patterns analysis.

3.1. Data pre-processing

We used the data from 01 January 2002 to 07 July 2002 for the cluster analysis process. Selecting useful data is an important task in the data pre-processing stage. After some preliminary analysis, we selected the statistical data comprising the traffic data on the domain, hourly and daily basis including request volume and page volume in each data type to generate the cluster models for finding Web user access and server usage patterns. To build up a precise model and to obtain more accurate analysis, it is also important to remove irrelevant and noisy data as an initial step in pre-processing task. Since SOM cannot process text data, any data in text format also have to be encoded according to a specific coding scheme into numerical format. Further, the datasets were scaled to 0–1. Besides the two inputs of ‘volume of requests’ and ‘volume of pages’, which directly from the original data, we also included an additional input ‘time index’ to distinguish the access time sequence of the data entries. The most recently accessed data were indexed with higher value of ‘time index’ while the least recently accessed data were placed at the bottom with lowest value (Aggarwal et al., 1999). This is a very critical step to obtain more precise analysis result due to time dependence characteristic of Web usage data itself.

3.2. Data clustering using SOM

With the increasing popularity of Internet, millions of requests (with different interests from different countries) are received by Web servers of large organizations Monash University is a truly international university with its main campus located in Australia and campuses in South Africa and Malaysia. The university has plans to extend its educational services around the globe. Therefore, the huge traffic volume and the dynamic nature of the data require the necessity to implement efficient and intelligent Web mining framework.

In Web usage mining research, the method of clustering is broadly used in different projects by researchers for finding the usage patterns or user profiles (Wang et al., 2002). Among all the popular clustering algorithms, SOM has been successfully used in Web mining projects (Honkela et al., 1997; Kohonen et al., 2000). In our approach, with the Web usage data from high dimensional input data, finally a 2D map of Web usage patterns with different clusters could be formed from the SOM training process. The related transaction entries are grouped into the same cluster and the relationship between different clusters is explicitly shown on the map. We used the Viscovery SOMine (Eudaptics, 2002) to simulate the SOM. All records after pre-processing stage were used by the SOM algorithm and the clustering results were obtained after the unsupervised learning. We adopted a trial and error approach by comparing the ‘normalized distortion error’ and ‘quantization error’ to decide the various parameter settings of the SOM algorithm. We finally decided the parameter setting, which could minimize both ‘normalized distortion’ and ‘quantization’ errors. From all the experiments with different parameter settings, the best solution was selected when minimum errors were obtained.

3.3. WUDA to find domain patterns

From the SOM clustering process, five clusters were mapped according to the user access data on the country/domain basis. To analyze the difference among the clusters, we have illustrated the comparison on the number of the unique country/domain and averaged request volume for each cluster in Table 1.

As evident from Table 1, Clusters 4 and 5 are distinguished from the rest of the clusters. Cluster 4 has almost 6 times of the volume of requests compared with the average requests volume of the other 3 clusters (Clusters 1, 2 and 3), and Cluster 5 has the maximum number of requests which is nearly 10 times that of Cluster 4. However, by comparing the number of domain countries, Clusters 1, 2 and 3 all have around 150 different domain sources, Cluster 4 contains only Australian domains and Cluster 5 accounts for only *.com and *.net users. The majority of the requests originated from Australian domains followed by *.com and *.net users. This shows that even though Monash University’s main server is accessed by all users around the globe, the majority of the traffic still originated within Australia.

To identify the interesting patterns from the data clusters of Clusters 1, 2 and 3, we have plotted the ‘time index’ value which we used as an additional input in Fig. 4 for further analysis. For Clusters 1, 2 and 3 the number of requests of each cluster are very similar and also shared by similar number of users from different countries which made it difficult to identify their difference depending on the volume of requests and pages. However, as shown in Fig. 4, Cluster 1 (marked with ‘◇’), 2 (marked with ‘_’) and 3 (marked with ‘Δ’), can be distinguished with reference to the time of accessing the server. So, Clusters 1, 2 and 3 have very similar patterns for the requests, but their time of access is separated very clearly. Cluster 2 accounts for the most recent visitors and Cluster 3 represents the least recent visitors. Cluster 1 accounts for the users that were not covered by Clusters 2 and 3. Therefore, the different users were clustered based on the time of accessing the server and the volume of requests.

Table 1
Analysis of requests volume for domain clusters

Cluster number	Unique country or domain	Request volume (averaged)
1	160	2009.91
2	157	2325.18
3	162	3355.73
4	1 (.au)	199258.93
5	2 (.com and.net)	1995725.00

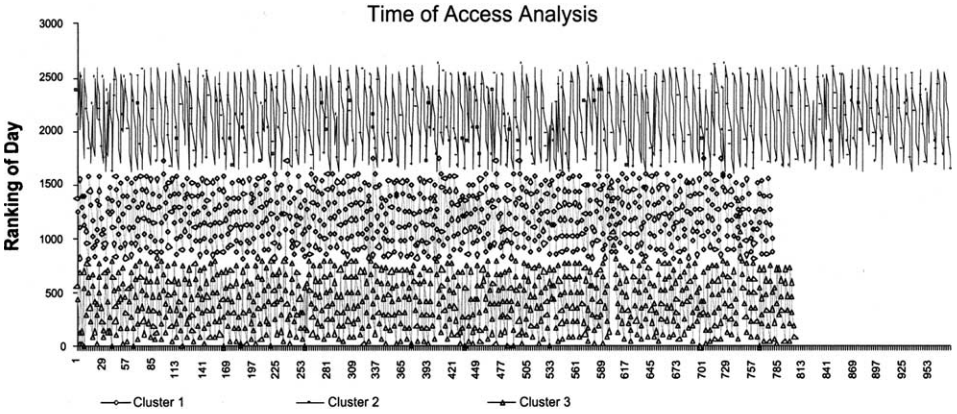


Fig. 4. WUDA for time of access of cluster 1, 2 and 3 in domain cluster map.

3.4. WUDA to analyze hourly web traffic request patterns

The training process generated four clusters for the Web traffic requests on the hourly basis. The developed cluster map indicating the hour of the day when the request was made is shown in Fig. 5.

From the developed cluster map depicted in Fig. 6, it is very difficult to tell the difference between each cluster, as the requests according to the different hours in a day are scattered. But from Fig. 6, it may be concluded that Cluster 2 (marked with '◇') and Cluster 3 (marked with '_') have much higher requests and pages (nearly double) than Cluster 1 (marked with 'Δ') and Cluster 4 (marked with 'x'). This shows that two groups of clusters are separated based on the volume of requests for different hours of the day.

By analyzing the feature inputs of the SOM clusters (Fig. 6), it is difficult to find more useful information. However, by looking at the each hour as shown in Fig. 7 more

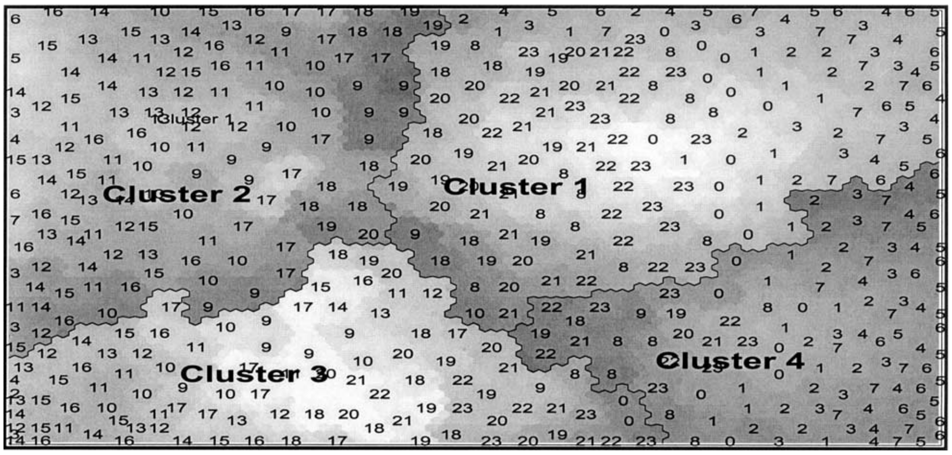


Fig. 5. Hourly web traffic requests cluster map.

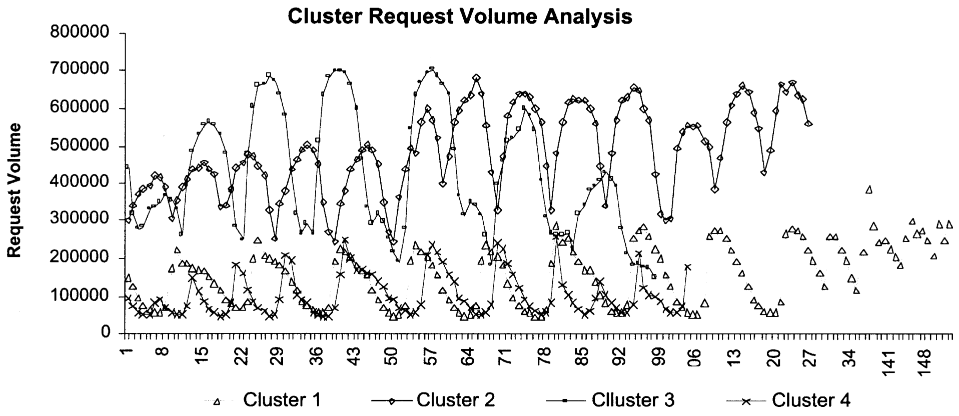


Fig. 6. WUDA for request volume in hourly cluster map.

meaningful information could be obtained. Clusters 2 and 3 are mainly responsible for the traffic during the office hours (09:00–18:00), and Clusters 1 and 4 account for the traffic during the university off peak hours. It is interesting to note that the access patterns for each hour could be analyzed from the cluster results with reasonable classification features. By combining the information from Figs. 6 and 7, the hourly access patterns could be understood very clearly.

3.5. WUDA to discover daily requests clusters

Due to the dynamic nature of the WWW, it is difficult to understand the daily traffic pattern using conventional Web log analyzers. We attempted to cluster the data depending on the total activity for each day of the week using 'request volume', 'page volume' and 'time index' as input features. The training process using SOM produced seven clusters and the developed 2D cluster map is shown in Fig. 8.

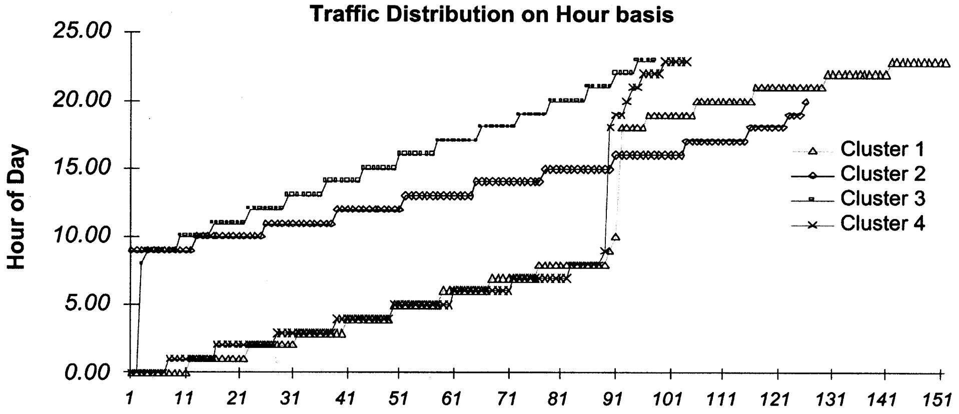


Fig. 7. WUDA for hour of the day access in hourly cluster map.

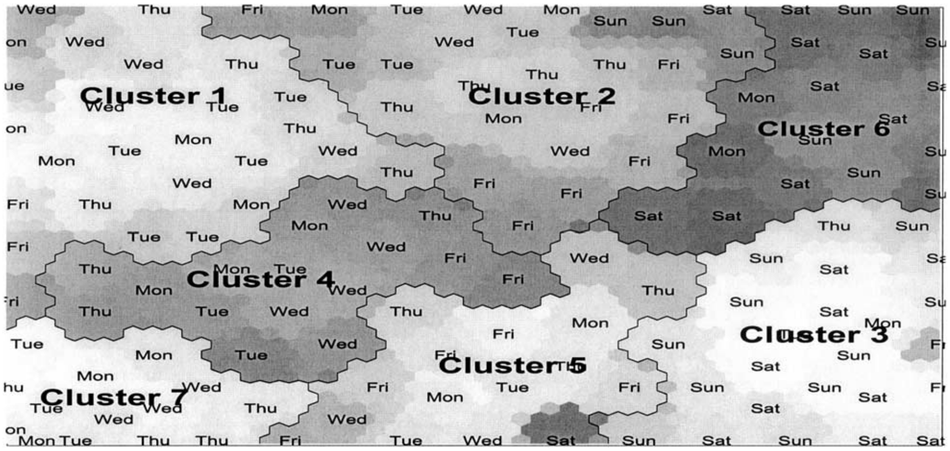


Fig. 8. Daily web traffic requests cluster map.

First, each cluster represents the traffic for only a certain access period by checking the time index value in each cluster records, and the ranking of the clusters are ordered as 2, 6, 1, 4, 3, 7 and 5 according to the descending order of the access time. In Table 2, WUDA reveals that the clusters are further separated according to the date of access in a week. Clusters 3 and 6 account for access records, which happened during weekend (Saturday and Sunday). The big group consists of Clusters 1, 2, 4, 5 and 7, which account for the transactions occurred with heavy traffic volume during normal working weekdays (Monday to Friday). With further detailed checking, Clusters 2 and 7 are different from other clusters because Cluster 2 covered heavier traffic on Friday but Cluster 7 missed Friday.

4. Fuzzy inference systems for web traffic trend prediction

The world of information is surrounded by uncertainty and imprecision. The human reasoning process can handle inexact, uncertain and vague concepts in an appropriate

Table 2
WUDA for time of access and day of the week in daily traffic cluster map

	Cluster 1 (%)	Cluster 2 (%)	Cluster 3 (%)	Cluster 4 (%)	Cluster 5 (%)	Cluster 6 (%)	Cluster 7 (%)
Monday	19.23	11.54	4.17	18.75	12.50	11.11	28.57
Tuesday	26.92	15.38	4.17	18.75	12.50	0.00	21.43
Wednesday	23.08	11.54	0.00	31.25	18.75	0.00	21.43
Thursday	19.23	15.38	4.17	18.75	18.75	0.00	28.57
Friday	11.54	30.77	8.33	12.50	31.25	0.00	0.00
Saturday	0.00	3.85	37.50	0.00	6.25	50.00	0.00
Sunday	0.00	11.54	41.67	0.00	0.00	38.89	0.00

manner. Usually, the human thinking, reasoning and perception process cannot be expressed precisely. These types of experiences can rarely be expressed or measured using statistical or probability theory. Fuzzy logic provides a framework to model uncertainty, human way of thinking, reasoning and the perception process. Fuzzy if-then rules and fuzzy reasoning are the backbone of FIS, which are the most important modeling tools based on fuzzy set theory. We made use of the TSFIS in which the conclusion of a fuzzy rule is constituted by a weighted linear combination of the crisp inputs rather than a fuzzy set (Sugeno, 1985). A conventional FIS makes use of a model of the expert who is in a position to specify the most important properties of the process. Expert knowledge is often the main source to design FIS.

The derivation of if-then rules and corresponding membership functions depend heavily on the a priori knowledge about the system under consideration. However there is no systematic way to transform experiences of knowledge of human experts to the knowledge base of a FIS. On the other hand, artificial neural network (ANN) learning mechanism does not rely on human expertise. To a large extent, the drawbacks pertaining to these two approaches seem complementary. Therefore, it is natural to consider building an integrated system combining the concepts of FIS and ANN modeling. A common way to apply a learning algorithm to a FIS is to represent it in a special ANN like architecture. However, the conventional ANN learning algorithms (gradient descent) cannot be applied directly to such a system as the functions used in the inference process are usually non differentiable. This problem can be tackled by using differentiable functions in the inference system or by not using the standard neural learning algorithm (Abraham, 2001). According to the performance measure of the problem environment, the membership functions, rule bases and the inference mechanism are to be adapted. Evolutionary computation and neural network learning techniques are used to adapt the various fuzzy parameters.

In this research, we used the Adaptive Neuro-Fuzzy Inference System (ANFIS) (Jang, 1992) framework based on neural network learning to fine tune the rule antecedent parameters and a least mean squares estimation to adapt the rule consequent parameters of the TSFIS. A step in the learning procedure has two parts: In the first part the input patterns are propagated, and the optimal conclusion parameters are estimated by an iterative least mean square procedure, while the antecedent parameters (membership functions) are assumed to be fixed for the current cycle through the training set. In the second part the patterns are propagated again, and in this epoch, back propagation is used to modify the antecedent parameters, while the conclusion parameters remain fixed. More details of the learning algorithm could be obtained from (Jang, 1992).

4.1. Design and experimentation results

We used the popular grid partitioning method (clustering) to generate the initial rule base. This partition strategy requires only a small number of membership functions for each input. Besides the inputs ‘volume of requests’“ and ‘volume of pages’ and ‘time index’”, we also used the ‘cluster location information’ provided by the SOM output. The data was re-indexed based on the clustering information inputs. We attempted to develop fuzzy inference models to predict (few time steps ahead) the Web traffic on the hourly and

Table 3
Training and test performance for daily web traffic prediction

Forecast period	Root mean squared error (RMSE)			
	Fuzzy inference system (with cluster information)		Fuzzy inference system (without cluster information)	
	Training	Test	Training	Test
1 day	0.01766	0.04021	0.06548	0.09565
2 days	0.05374	0.07082	0.10465	0.13745
3 days	0.05264	0.06100	0.12941	0.14352
4 days	0.05740	0.06980	0.11768	0.13978
5 days	0.06950	0.07988	0.13453	0.14658

daily basis. We used the data from 17 February 2002 to 30 June 2002 for training and the data from 01 July 2002 to 06 July 2002 for testing and validation purposes.

4.2. Daily traffic prediction

We used the MATLAB environment to simulate the various experiments. Given the daily traffic volume of a particular day the developed model could predict the traffic volume up to five days ahead. Three membership functions were assigned to each input variable. Eighty-one fuzzy if-then rules were generated using the grid based partitioning method and the rule antecedent/consequent parameters were learned after fifty epochs. We also investigated the daily web traffic prediction performance without the cluster information input variable.

Table 3 summarizes the performance of the FIS for training and test data. Fig. 9(a)–(e) depict the test results for one day, two days, three days, four days and five days ahead prediction of daily Web traffic volume.

4.3. Hourly page request forecast

Three membership functions were assigned to each input variable Eighty-one fuzzy if-then rules were generated using the grid based partitioning method and the rule antecedent/consequent parameters were learned after 40 epochs. We also investigated the volume of hourly page requests prediction performance without the cluster information input variable. Table 4 summarizes the performance of the FIS for training and test data. Fig. 10(a)–(c) show the test results for 1, 12 and 24 h ahead prediction of the volume of hourly page requested.

5. Conclusions and future work

The discovery of useful knowledge, user information and server access patterns allows Web based organizations to predict user access patterns and helps in future developments, maintenance planning and also to target more rigorous advertising campaigns aimed at groups of users (Cooley et al., 1997). Previous studies have indicated that the size of the Website and its traffic often imposes a serious constraint on the scalability of the methods (Paliouras et al., 2000). Our study on Monash University's Web access patterns reveals the necessity to incorporate CWI techniques for mining useful information and predicting trend patterns. From the SOM clusters, WUDA provided useful information related to the user access patterns, which could not be possible by using conventional statistical approaches. The developed FIS could predict the Web traffic volume on the daily and hourly basis within reasonable error

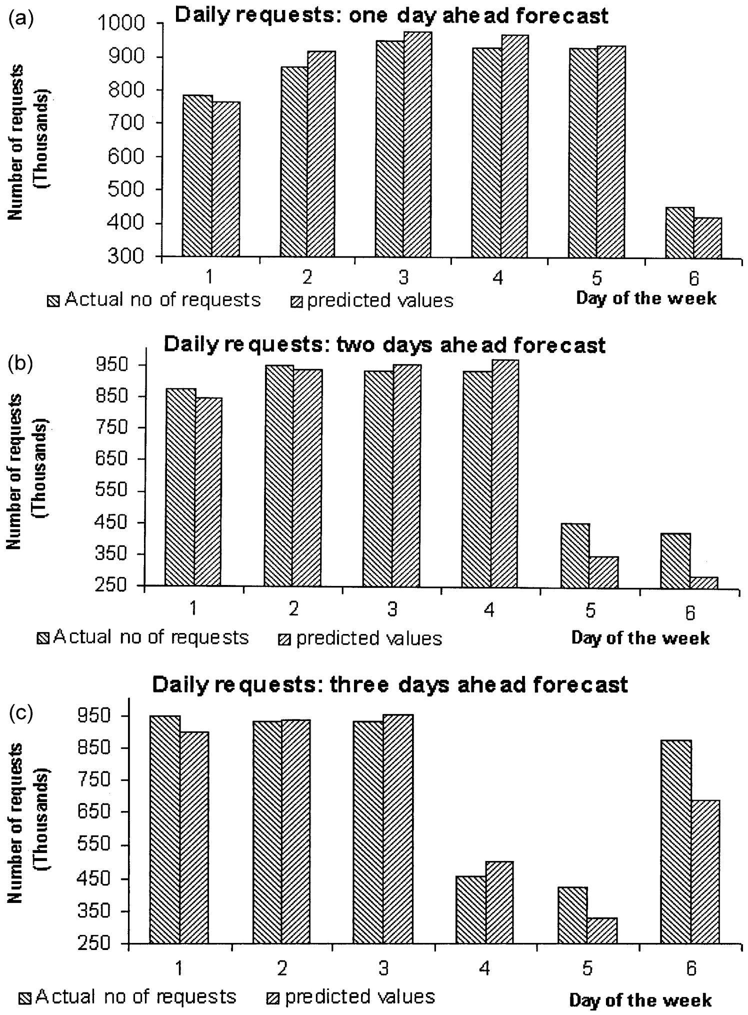
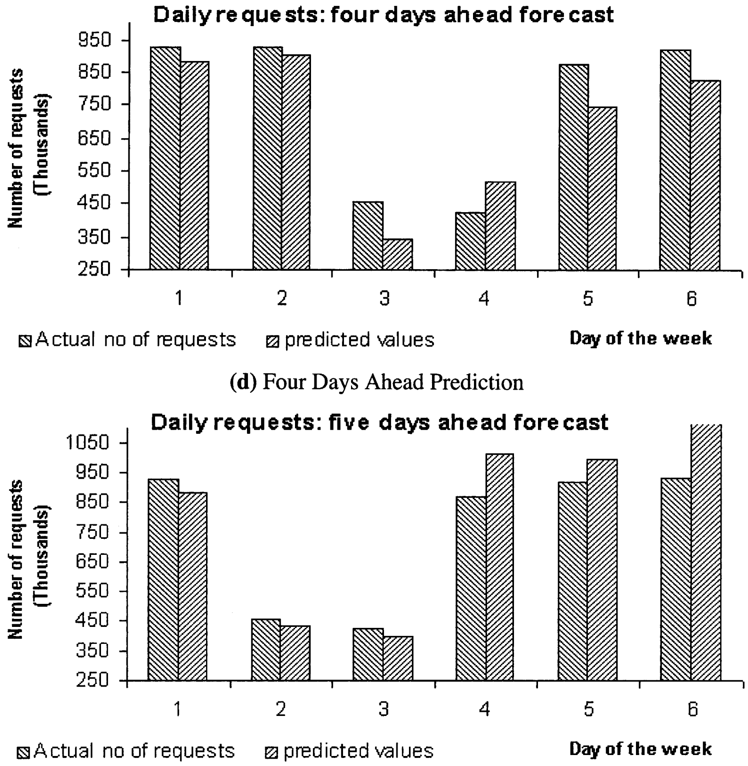


Fig. 9. Test results of daily prediction of website traffic on request volume, (a) One day ahead prediction, (b) Two days ahead prediction, (c) Three days ahead prediction, (d) Four days ahead prediction, (e) Five days ahead prediction.



(**d**) Four Days Ahead Prediction

(**e**) Five Days Ahead Prediction

Fig. 9 (*continued*)

limits. Our experiment results also reveal the importance of the cluster information to improve the prediction accuracy of the FIS. These techniques might be useful for the Website tracker software vendors to provide more useful information to the Web administrators.

Table 4
Training and test performance for hourly web traffic prediction

	Root mean squared error (RMSE)			
Forecast period (h)	Fuzzy inference system (with cluster information)		Fuzzy inference system (without cluster information)	
	Training	Test	Training	Test
1	0.04334	0.04433	0.09678	0.10011
12	0.06615	0.07662	0.11051	0.12212
24	0.05743	0.06761	0.10891	0.11352

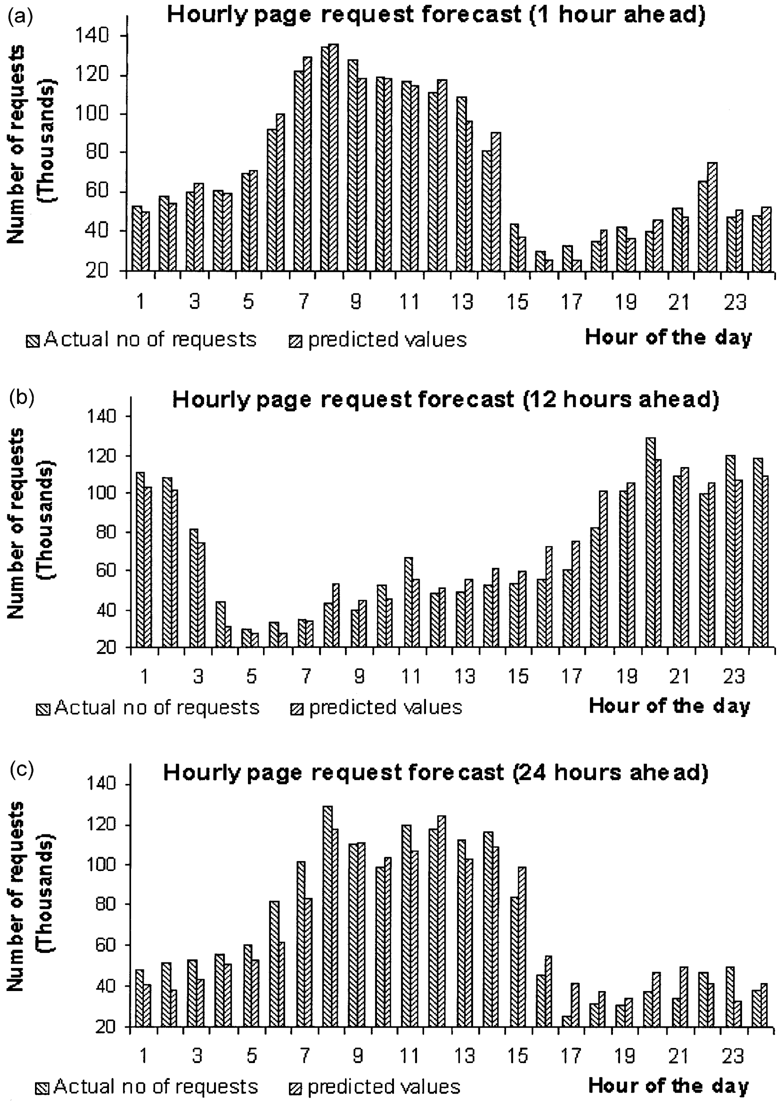


Fig. 10. Test results of hourly prediction of web traffic on page volume, (a) One hour ahead prediction, (b) Twelve hours ahead prediction, (c) Twenty-four hours ahead prediction.

We relied on the statistical/text data provided by the ‘Analog’ software embedded in the university’s Web server for experimentations. Due to incomplete details, we had to analyze the usage patterns for different aspects of log files separately, which virtually prevented us to link some common information between the different aspects such as trends, patterns etc. For example, the domain requests and the daily or hourly requests are all stand-alone information and are not interconnected. Therefore, a direct analysis from the raw Web access logs might be more helpful. We believe that if the detailed access information could cover different interlinked features, then the usage patterns analysis would be more comprehensive and useful.

In this research, we considered only the Web traffic data during the university's peak working time. Our future research will also incorporate off-peak months (summer semesters) and other special situations such as unexpected events and server log technical failures. We also plan to incorporate more data mining techniques to improve the functional aspects of the concurrent neuro-fuzzy approach.

References

Levene M, Loizou G. Computing the entropy of user navigation in the web. Department of Computer Science, University College London, Report No. RN/99/42 1999;.

Brin S. Extracting patterns and relations from the world wide web. In Proceedings of WebDB Workshop at the 6th International Conference on Extending Database Technology 1998;(EDBT(98):1998.

Buchner AG, Mulvenna MD. Discovering behavioural patterns in internet log files: playing the Devil's advocate. In Proceedings of the 12th Biennial International Telecommunications Society Conference (ITS(98)), Stockholm, Sweden; 1998.

Buchner AG, Mulvenna MD, Anand SS, Hughes JG. An internet-enabled knowledge discovery process. In Proceedings of the 9th International Database Conference, Hong Kong 1999;1999:13–27.

Hastie T, Tibshirani R, Friedman J. The elements of statistical learning-data mining. In: Hastie T, Tibshirani R, Friedman J, editors. Inference and prediction. New York: Springer-Verlag; 2001.

Masseglia F, Poncelet M, Teisseire M. Using data mining techniques on web access logs to dynamically improve hypertext structure. ACM SigWeb Lett 1999;8(3):1–19.

Pohle C, Spihopoulou M, In Proceedings of the 4th International Conference (DaWaK(02) on Data Warehousing and Knowledge Discovery, Aix-en-Provence, France, September 4–6, vol. 2454.; 2002. pp. 83–93.

Pal SK, Talwar V, Mitra P. Web Mining in Soft Computing Framework: Relevance. State of the Art and Future Directions. IEEE Transaction on Neural Networks 2002;13(5):1163–77.

Zhang YQ, Lin TY. Computational web intelligence (CWI): synergy of computational intelligence and web technology. In Proceedings of 2002 World Congress on Computational Intelligence, IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'02) Special Session on Computational Web Interlligence (CWI), Honolulu, Hawaii, USA, May 12–17; 2002.

Boley D, Gini ML, Gross R, Han EH, Hastings K, Karypis G, Kumar V, Mobasher B, Moore J. Document categorization and query generation on the world wide web using WebACE. J Artif Intell Rev 1999;13(5-6): 365–91.

Pirolli P, Pitkow J, Rao R. Silk from a sow's ear: extracting usable structures from the web. In Proceedings of Conference on Human Factors in Computing Systems (CHI(96), Vancouver, British Columbia, Canada 1996; 1996:118–25.

Masseglia F, Poncelet P, Cicchetti R. An efficient algorithm for web usage mining. J Networking Inf Syst (NIS) 1999;2(5-6):571–603.

Kitsuregawa M, Toyoda M, Pramudiono I. Web community mining and web log mining: commodity cluster based execution. In Proceedings of the 13th Australasian Database Conference (ADC(02), Melbourne, Australia 2002;5:3–10.

Lingras P. Rough set clustering for web mining. In Proceedings of 2002 World Congress on Computational Intelligence, IEEE International Conference on Fuzzy Systems (FUZZ-IEEE(02) Special Session on Computational Web Intelligence (CWI), Honolulu, Hawaii, USA 2002;2002:1039–44.

Ng A, Smith KA. Web usage mining by a self-organizing map. In Proceedings of International Conference on Artificial Neural Networks In Engineering (ANNIE(00) Smart Engineering System Design: Neural Networks, Fuzzy Logic, Evolutionary Programming, Data Mining and Complex Systems 2000;10: 495–500.

Srivastava J, Cooley R, Deshpande M, Tan PN. Web usage mining: discovery and applications of usage patterns from web data. SIGKDD Explorations 2000;1(2):12–23.

Wang X, Abraham A, Smith KA. Web traffic mining using a concurrent neuro-fuzzy approach. In Proceedings of the 2nd International Conference on Hybrid Intelligent Systems, Computing Systems: Design, Management and Applications, Santiago, Chile 2002;2002:853–62.

Cheung D, Kao B, Lee J. Discovering user access patterns on the world wide web. In Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD(97), vol. 10.; 1997. pp. 463–70.

Chang G, Healey MJ, McHugh JAM, Wang JTL. Web minig. In Mining the World Wide Web—An Information Search Approach, Dordetch: Kluwer; 2001.

Han J, Pei J, Yin Y. Mining frequent patterns without candidate generation. In Proceedings of the 2000 ACM-SIGMOD International Conference on Management of Data (SIGMOD(00), Dallas, TX, USA 2000;2000: 1–12.

Jespersen SE, Thorhauge J, Pedersen TB. A hybrid approach to web usage mining. In Proceedings of the 4th International Conference (DaWaK(02) on Data Warehousing and Knowledge Discovery, Aix-en-Provence, France 2002;2002:73–82.

Mobasher B, Cooley R, Srivastava J. Creating adaptive web sites through usage-based clustering of URLs. In Proceedings of 1999 Workshop on Knowledge and Data Engineering Exchange, Chicago, Illinois, USA 1999; 1999:19–25.

Chi EH, Rosien A, Heer J. Lumberjack: intelligent discovery and analysis of web user traffic composition. In Proceedings of ACM-SIGKDD Workshop on Web Mining for Usage Patterns and User Profiles (WebKDD(02), Edmonton, Alberta, Canada 2002;2002:1–16.

Martín-Bautista MJ, Kraft DH, Vila MA, Chen J, Cruz J. User profiles and fuzzy logic in web retrieval. Springer J Soft Comput 2002;65(5):365–72.

Pazzani MJ, Billsus D. Learning and revising user profiles: the identification of interesting web sites. J Machine Learning 1997;27(3):313–31.

Joshi KP, Joshi A, Yesha Y, Krishnapuram R. Warehousing and mining web logs. In Proceedings of the 2nd ACM CIKM Workshop on Web Information and Data Management, Kansas City, Missouri, USA 1999; 1999:63–8.

Agrawal R, Srikant R. Fast algorithms for mining association rules. In Proceedings of the 20th International Conference on Very Large Databases, Santiago, Chile 1994;1994:487–99.

Krishnapuram R, Joshi A, Yi L. A fuzzy relative of the k-medoids algorithm with application to document and snippet clustering. In Proceedings of IEEE International Conference on Fuzzy Systems (FUZZIEEE(99), Seoul, Korea 1999;1999:1281–6.

Zaïane OR, Xin M, Han J. Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. In Proceedings of Advances in Digital Libraries Conference, Santa, Barbara, California, USA 1998;1998:19–29.

Berkan RC, Trubatch SL. Fuzzy logic and hybrid approaches to web intelligence gathering and information management. In Proceedings of 2002 World Congress on Computational Intelligence, IEEE International Conference on Fuzzy Systems (FUZZ-IEEE(02) Special Session on Computational Web Intelligence (CWI), Honolulu, Hawaii, USA 2002;2002:1033–8.

Chen PM, Kuo FC. An information retrieval system based on an user profile. J Syst Software 2000;54:3–8.

Zaïane OR. Building virtual web views. J Data Knowledge Engng 2001;39(2):143–63.

Spiliopoulou M, Faulstich LC. WUM: a web utilization miner. In Proceedings of Workshop on the Web and Data Bases (WebDB(98), Valencia, Spain 1998;1998:109–15.

Cooley R, Tan PN, Srivastava J. WebSIFT: the web site information filter system. In Proceedings of the Web Usage Analysis and User Profiling (WebKDD'99) Workshop on Web Mining, an Diego, CA, USA; 1999. pp. 163–82.

Perkowitz M, Etzioni O. Adaptive web sites: automatically synthesizing web pages. In Proceedings of the 15th National Conference on Artificial Intelligence and 20th Innovative Applications of Artificial Intelligence Conference (AAAI(98, IAAI(98), Madison, Wisconsin, USA 1998;1998:727–32.

Joachims T, Freitag D, Mitchell T. WebWatcher: a tour guide for the world wide web. In Proceedings of the 15th International Joint Conference on Artificial Intelligence (IJCAI(97), Nagoya, Japan 1997;1997: 770–5.

Kohonen T. The self-organizing maps. Proceedings of the IEEE, vol. 78.; 1990. pp. 1464–480.

Fu Y, Sandhu K, Shih MY. Clustering of web users based on access patterns. In Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD(99) Workshop on Web Mining, San Diego, CA, USA, vol. 5.; 1999. pp. 560–67.

Sugeno M. Industrial applications of fuzzy control. Amsterdam: Elsevier science; 1985.

Analog 2002. Website Log Analyser. At URL: http://www.analog.cx.

Server Usage Statistics, Monash University, Australia. At URL:http://www.monash.edu.au.

Abraham A. Neuro-fuzzy systems: state-of-the-art modeling techniques. In Proceedings of the 6th International Work-Conference on Artificial and Natural Neural Networks (IWAIVN 2001), Granada, Spain 2001;2001: 269–76.

Coenen F, Swinnen G, Vanhoof K, Wets G. A framework for self adaptive websites: tactical versus strategic changes. In Proceedings of Workshop on Webmining for E-commerce: challenges and opportunities (KDD(00), Boston, USA; 2000. pp. 75–81.

Aggarwal C, Wolf JL, Yu PS. Caching on the world wide web. IEEE Trans Knowledge Data Engng 1999;11(1): 94–107.

Honkela T, Kaski S, Lagus K, Kohonen T. WEBSOM-Self Organizing Maps of Documents Collections. In Proceedings of Workshop on Self-Organizing Maps (WSOM(97), Espoo, Finland 1997;1997:310–5.

Kohonen T, Kaski S, Lagus K, Salojrvi J, Honkela J, Paatero V, Saarela A. Self organization of a massive documents collection. IEEE Trans Neural Networks 2000;11(3):574–85.

Eudaptics Software 2002. Viscovery SOMine. At URL URL.com.

R. Jang 1992. Neuro-Fuzzy Modeling: Architectures, Analyses and Applications. PhD Thesis, University of California, Berkeley, USA.

Cooley R, Mobasher B, Srivastava J. Web mining: information and pattern discovery on the world wide web. In Proceedings of the 9th IEEE International Conference on Tools with Artifical Intelligence (ICTAI(97), Newport Beach, CA 1997;1997:558–67.

Paliouras G, Papatheodorou C, Karkaletsis V, Spyropoulos CD. Clustering the users of large web sites into communities. In Proceedings of the 17th International Conference on Machine Learning (ICML(00), Stanford University, Standord, CA, USA 2000;2000:719–26.