Hierarchical Rough Fuzzy Self Organizing Map for Weblog Prediction

Arindam Chaudhuri[1(✉)] and Soumya K. Ghosh[2]

[1] Samsung R&D Institute Delhi, Noida 201304, India
arindam_chau@yahoo.co.in
[2] Department of Computer Science Engineering,
Indian Institute of Technology Kharagpur, Kharagpur 721302, India
skg@iitkgp.ac.in

Abstract. Web servers play a vital role in conveying knowledge and information to end users. With rapid growth of WWW over past decades discovering hidden information about the usage pattern is critical towards determining effective strategies as well as to optimize server usage. Most of the available server analysis tools provide statistical data only without much useful information. Mining useful information becomes challenging task when user traffic data is huge and keeps on growing. In this work we propose hierarchical rough fuzzy self–organizing map (HRFSOM) to analyze useful information from the statistical data through weblog analyzer. We use cluster information generated by HRFSOM for data analysis and a variant of takagi sugeno fuzzy inference system (TSFIS) to predict daily and hourly traffic jam volumes. The experiments are performed using web user access sample patterns available at Yandex Personalized Web Search Challenge where statistical weblog data is generated by AWStats web access log file analyzer. The proposed classifier has superior clustering accuracy compared to other classifiers. The experimental results demonstrate the efficiency of proposed approach.

Keywords: Weblog prediction · Weblog data · SOM · RFSOM · HRFSOM

1 Introduction

In the present competitive business scenario web has become a place where people of all ages, languages and cultures conduct their digital lives [1]. The web usage entails a wide array of devices, networks and applications [2, 3]. When searching, creating and disseminating information, users leave behind great deal of data revealing their information needs, attitudes and personal facts. The web designers collect these artifacts in weblogs for subsequent analysis. WWW is always expanding with rapid increase of information transaction from web users. For web administrators, discovering hidden information about users' access patterns improves web information service performance quality. From business point of view, knowledge obtained from access patterns manages e-business services. The statistical data obtained from weblog files provide information explicitly. The analysis relies on past usage patterns, shared content degrees and inter-memory associative links. This leads to content intelligence

R. Silhavy et al. (eds.), *Artificial Intelligence Trends in Intelligent Systems*,
Advances in Intelligent Systems and Computing 573, DOI 10.1007/978-3-319-57261-1_19

improving overall system quality. The pattern discovery of web usage mining consists of statistical analysis, clustering etc. Most of the existing research focuses on finding patterns with insignificant pattern analysis. Some of the important weblog analysis techniques are conceptual framework, phenomenology, content analysis, discourse analysis etc.

Considering different weblog analysis techniques access patterns available at Yandex Personalized Web Search Challenge [4] are used to predict daily and hourly traffic jams. The predicted results provide information for decision making activities. To achieve this hierarchical rough fuzzy self-organizing map (HRFSOM) is proposed to cluster and discover patterns from data that is used for statistical analysis. To make analysis more intelligent clustered data is used for prediction. A variant of takagi sugeno fuzzy inference system (TSFIS) [2] explores prediction of average daily and hourly traffic jams. The statistical data from sample patterns is provided by AWStats log file analyzer [5]. The generated data covers different aspects of users' access log records, weekly based reports, domain summary navigation summary etc. The challenge lies in finding relevant hidden information through extraction of patterns. The information analysis and prediction from huge datasets entails requirement of hybrid intelligent systems. The major contributions of this work includes: (a) input representation of SOM [6] as rough granules or nuggets (b) rough fuzzy sets [7] formulation to extract domain knowledge from data (c) training HRFSOM to cluster data and (d) using variant of TSFIS to predict daily and hourly traffic jam. This paper is organized as follows. In Sect. 2 computational of HRFSOM is highlighted. This is followed by experiments and results in Sect. 3. Finally in Sect. 4 conclusions are given.

2 Computational Method

In this section mathematical framework of proposed HRFSOM model [2] is presented. The schematic representation of prediction system is given in Fig. 1.

2.1 Problem Description

The research problem entails in predicting the daily and hourly traffic jams on ever growing traffic data. In order to achieve this prediction task, we propose a SOM based predictor viz RFSOM to analyze the web user access sample patterns at Yandex Personalized Web Search Challenge [4]. The statistical data from sample patterns is provided by AWStats web access log file analyzer [5]. The clusters generated by RFSOM are used by takagi sugeno fuzzy inference system (TSFIS) variant for prediction. The prediction task performed on daily and hourly traffic jams provide information for several decision making activities.

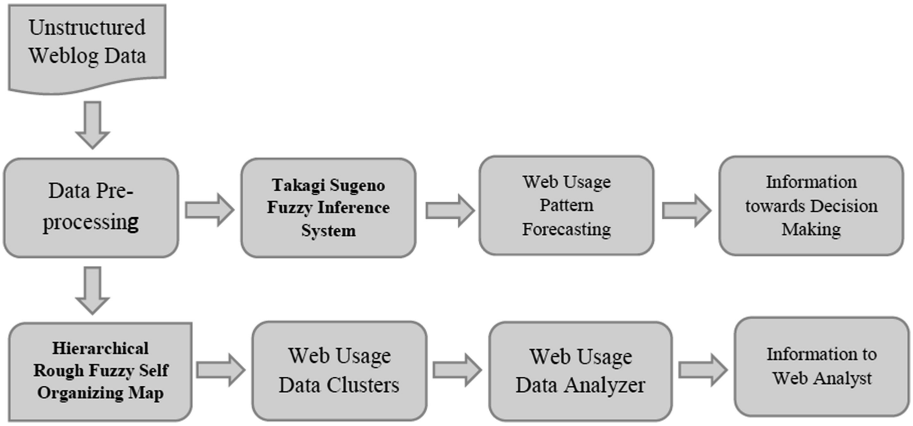

Unstructured Weblog Data
Data Pre-processing
Takagi Sugeno Fuzzy Inference System
Web Usage Pattern Forecasting
Information towards Decision Making
Hierarchical Rough Fuzzy Self Organizing Map
Web Usage Data Clusters
Web Usage Data Analyzer
Information to Web Analyst

Fig. 1. The schematic representation of prediction system for web pattern analysis

2.2 Datasets

The experimental data is taken from Yandex Personalized Web Search Challenge 2014 which re-ranked web documents using personal preferences [4]. Here an opportunity was provided to consolidate and scrutinize the work from industrial labs on personalizing web search using user-logged search behavior context. It provided fully anonymized dataset shared by Yandex which had anonymized user ids, queries, query terms, urls and clicks. Each user record is specified in terms of user identification, query and url. The content tag contains an individual record. The record is followed by query term and clicks. Each of these records in dataset has recorded entity and level of prediction involved. After performing experiments with dataset more than 70000 records are labelled manually. The total dataset log contained 167,413,039 labelled records. The data is pre-processed to fit record mentioned in specific buckets. The logs are about 2 years old. The queries and users are sampled from only one region. The sessions containing queries with unwanted intent detected with Yandex classifier are removed considering the top-k most popular queries where k is user defined.

2.3 Rough Fuzzy Self-organizing Map for Weblog Mining

In this section RFSOM [2, 8] is evaluated in terms of rough fuzzy sets and SOM. It performs clustering and discovers patterns from data which are used for statistical analysis. The clustered data is used to predict daily and hourly traffic using TSFIS variant. The experimental data is adapted from single site weblog data generated by AWStats web access log file analyzer [5]. The web user access data is taken from 16th May 2010 to 16th May 2011. This raw data is unstructured in nature and is converted to structured format. After initial analysis statistical data comprising of number of domain requests, daily requests and hourly page traffic volume is selected to develop cluster models for finding web users usage patterns. The data is cleaned by removing irrelevant noise. The datasets are scaled to two tone format [2]. Other inputs such as number of users and total processing time are considered to distinguish temporal sequence of data. The data having most recent access are given higher index. The preprocessed data is presented to RFSOM for clustering and discovering patterns. The major phases are:

Representation of input vectors of SOM in terms of rough granules or nuggets: SOM input vector is described in terms of rough granules lower, median and upper. The human mind performs tasks based on perceptions presented through rough granules or nuggets [2]. A rough granule is group of patterns defined by generalized constraint $Y\, rs\, R$; R is constrained relation, rs is random set constraint combining probabilistic and possibilistic constraints and Y is rough set valued random variable. A pattern $y \in U$ is assigned value with membership function $\mu_Y^A(y)$ as:

$$\mu_Y^A(y) = \frac{\left\|[y]_A \cap Y\right\|}{\left\|[y]_A\right\|} \quad \text{for}\, y \in U \tag{1}$$

In Eq. (1) membership values are defined by π-membership function as:

$$\pi(y, C, \lambda) = \begin{cases} \quad 2\left(1 - \frac{\|y-C\|_2}{\lambda}\right)^2 & \text{for}\, \frac{\lambda}{2} \le \|y - C\|_2 \le \lambda \\ \quad 1 - 2\left(\frac{\|y-C\|_2}{\lambda}\right)^2 & \text{for}\, 0 \le \|y - C\|_2 \le \frac{\lambda}{2} \\ & \qquad\qquad \text{otherwise} \\ 0 & \end{cases} \tag{2}$$

In Eq. (2) $\lambda > 0$ is scaling factor of π function with C as central point and $\|\cdot\|_2$ is Euclidian norm. The π functions parameters are governed by p patterns set with n features $\left\{P_{ij}; i = 1, \ldots\ldots., p, j = 1, \ldots\ldots., n\right\}$. Here P_{jmin_m} and P_{jmax_m} are minimum and maximum values respectively along j^{th} feature for p patterns. The outliers affect parameters, centre and scaling factor of π function. Their effect is reduced by taking average of feature values of p patterns along j^{th} feature P_j which is centre of linguistic term c_{m_j}. The average pattern values having labels in ranges $\left[P_{j_{min_m}}, c_{m_j}\right)$ and $\left(c_{m_j}, P_{j_{max_m}}\right]$ are defined as middle points of linguistic terms *lower* and *upper* by c_{low_j} and c_{up_j} respectively. Similarly patterns with values in above ranges can be considered along j^{th} axis $P_{j_{min_{low}}} = P_{j_{min_m}}$, $P_{j_{max_{low}}} = c_{m_j}$, $P_{j_{min_{up}}} = c_{m_j}$ and $P_{j_{max_{up}}} = P_{j_{max_m}}$ [2]. To incorporate granular concept a n-dimensional pattern is represented as 3n-dimensional linguistic vector. If $Ft_{i1}, \ldots\ldots., Ft_{in}$ represent n features of i^{th} pattern $\boldsymbol{Ft_i}$ with μ being π-membership function then rough granule of features is:

$$\begin{aligned} \boldsymbol{Ft_i} = [&\left\{\mu_{lower(Ft_{i1})}(\boldsymbol{Ft_i}), \mu_{median(Ft_{i1})}(\boldsymbol{Ft_i}), \mu_{upper(Ft_{i1})}(\boldsymbol{Ft_i})\right\}, \ldots\ldots., \\ &\left\{\mu_{lower(Ft_{in})}(\boldsymbol{Ft_i}), \mu_{median(Ft_{in})}(\boldsymbol{Ft_i}), \mu_{upper(Ft_{in})}(\boldsymbol{Ft_i})\right\}] \end{aligned} \tag{3}$$

Crystallization of linguistic input data based on α-cut: Once input vectors of SOM are represented as rough granules, linguistic input is crystallized. This is done in two phases through computation of similarity matrix and generation of crystallization structures. The algorithm for first phase defines pairwise similarity matrix along linguistic inputs through rough set connectives. The similarity matrix develops crystallization structures based on α-value $(0<\alpha<1)$. In second phase crystallization structures are determined in q groups. Both algorithms are available in [2]. The resultant structures (q groups) are partitions or clusters. These partitions are arranged in decreasing order according to group size. The top m user defined groups based on their size are selected for every α-value in all q groups. The compactness of first m groups for every α-value are calculated using rough fuzzy entropy [7] and crystallization structures. For any particular α-value lowest average rough entropy are accepted.

Extraction of domain knowledge through rough fuzzy sets: The lowest average rough fuzzy entropy values [7] are presented to decision system DT to extract domain knowledge. Let p be the number of patterns in all m groups obtained using selected α-value. These p patterns from m groups $\{y_1, \ldots., y_p\}$ are presented to decision system $DT = (U, B \cup \{v\})$ where U and B represent universe and attributes $\{b_1, \ldots., b_{3n}\}$ respectively. Here each attribute is constructed by considering corresponding decision from $3n$-dimensional linguistic vectors from Eq. (3). The decision attribute d defined as Y_j; $j = 1, \ldots., m$ corresponding to pattern is assigned according to its group. Each Y_j is treated as decision class. Each pattern $y_i \in U$ is classified by its decision class. The rough fuzzy reflexive relation R_b between any two patterns p and q in U with respect to quantitative attribute $b \in B$ is:

$$R_b = \begin{cases} max\left(min\left(\frac{b(q)-b(p)+2\sigma_{b_{j_1}}}{2\sigma_{b_{j_1}}}, \frac{b(p)-b(q)+2\sigma_{b_{j_1}}}{2\sigma_{b_{j_1}}}\right), 0\right) & \text{if } b(p),\, b(q) \in R_d\left(Y_{j_1}\right) \\ max\left(min\left(\frac{b(q)-b(p)+2\sigma_{b_{j_2}}}{2\sigma_{b_{j_2}}}, \frac{b(p)-b(q)+2\sigma_{b_{j_2}}}{2\sigma_{b_{j_2}}}\right), 0\right) & \text{if } b(p) \in R_d\left(Y_{j_1}\right),\, b(q) \in R_d\left(Y_{j_2}\right) \wedge j_1 \neq j_2 \end{cases} \tag{4}$$

In Eq. (4), $j_1 = 1, \ldots., m$, $j_2 = 1, \ldots., m$ and $\sigma_{b_{j_1}}$, $\sigma_{b_{j_2}}$ represent standard deviation of decision classes Y_{j_1}, Y_{j_2} respectively; $b(p), b(q) \in R_d\left(Y_{j_1}\right)$ which implies patterns p, $q \in Y_{j_1}$ with respect to decision attribute $\{d\}$ where $b \in \{d\}$. Again $b(p) \in R_d\left(Y_{j_1}\right)$ and $b(q) \in R_d\left(Y_{j_2}\right)$ imply that patterns p, q belong to two different decision classes Y_{j_1}, Y_{j_2} respectively. The decision system DT contains p decision classes. The $3n$-dimensional vectors M_{kj} and S_{kj}; $j = 1, \ldots., 3n$ are mean and standard deviation respectively of patterns belonging to k^{th} decision class. The weighted distance of pattern $\boldsymbol{Ft_i}$; $i = 1, \ldots., p$ from k^{th} decision class is:

$$Z_{ik} = \sqrt{\sum\nolimits_{j=1}^{n} \left[\frac{Ft_{ij} - M_{kj}}{S_{kj}}\right]^2} \,\forall\, k = 1, 2, 3, \ldots\ldots., p \tag{5}$$

In Eq. (5), Ft_{ij} is j^{th} part of i^{th} pattern and the pattern membership is:

$$\mu_k(\boldsymbol{Ft_i}) = \frac{1}{1 + \left(\frac{Z_{ik}}{fz_d}\right)^{fz_e}} \tag{6}$$

In Eq. (6), fz_d and fz_e are rough entities. When a pattern has different membership values then its decision attribute becomes quantitative. This is shown in two ways. The membership values of all patterns in k^{th} class to its own class is $E_{kk} = \mu_k(\boldsymbol{Ft_i})$ if $k = v$ and membership values of all patterns in k^{th} class to other classes is $E_{kv} = 1$ if $k \neq v$; k, $v = 1, \ldots., p$. For any $C \subseteq B$ rough positive region is defined based on C-indiscernibility relation R_C for $p \in U$ as:

$$POS_c(q) = \left(\cup_{p \in U} R_C \downarrow R_d p \right)(q) \, \forall q \in U \tag{7}$$

Incorporating domain knowledge in SOM: The decision table DT explains granulation concept by partition and rough fuzzy set approximations based on rough reflexive relation. By this knowledge data is extracted and incorporated into SOM which is used for competitive learning. The knowledge about encoding procedure considers decision table DT with its set of conditional attributes, decision attributes, set of patterns and labeled values of patterns corresponding to $3n$-dimensional conditional attributes. The decision table DT extracts domain knowledge about data using following steps: (a) Generate rough reflexive relational matrix on all possible patterns pairs and obtain additional granulation structures (b) Using rough reflexive relational matrix compute memberships belonging to lower approximation of every pattern for each conditional attribute (c) Calculate rough positive region of every pattern for each conditional attribute (d) Calculate degree of dependency of each conditional attribute with respect to each decision class and assign resulting dependency factors as initial weights between input layer and (user defined clusters) output layer nodes.

Training RFSOM and clustering the data: After RFSOM is developed its training is done through following steps: (a) Transform input data into 3-dimensional granular space (b) Choose initial RFSOM connection weights using rough fuzzy sets (c) Train RFSOM through competitive learning (d) Partition data into clusters of granulation structures (e) Update weights connecting input layer to winning node and neighboring nodes (f) Repeat steps (c)–(e) by adjusting connection weights and neighborhood size (g) Map each term to node and label winning nodes to make output space ordered. To select RFSOM parameters trial and error approach is adopted. This reduces normalized distortion and quantization error. The clustering results accuracy obtained from RFSOM are better than SOM and fuzzy SOM (FSOM) algorithms [2, 6, 8].

2.4 Hierarchical Bidirectional Recurrent Neural Network for Semantic Analysis

The hierarchical version of RFSOM viz HRFSOM [2] is proposed here. The computational benefits [2] serve the major motivation. HRFSOM is different from RFSOM in terms of efficient classification accuracy based on similarities and running time when volume of data grows [2]. The model architecture is shown in Fig. 2. HRFSOM architecture correlates data behavior across multiple features of relevance. This facilitates distribution of computational overheads in predictor construction. At 1st and 2nd layers relatively small RFSOMs are utilized (6 × 6). The number of RFSOMs are increased at layers 3, 4 and 5. RFSOMs in last layer are constructed over subset of examples for which neuron in 5th layer is Best Matching Unit (BMU). RFSOMs at last layer can be larger (40 × 40) than used in 1st to 5th layers. It improves resolution and discriminatory capacity of RFSOM with less training overhead. Building HRFSOM requires several data normalization operations. This provides for initial temporal pre-processing and inter-layer quantization between 1st to 5th layers. The pre-processing provides suitable representation for data and supports time based representation. The 1st SOM layer treats each feature independently with each data instance mapped to sequential values. In case of temporal representation standard RFSOM has no capacity to recall histories of patterns directly. A shift register of length l is employed in which tap is taken at predetermined repeating interval k such that $l \% k = 0$ where $\%$ is modulus operator. The 1st level RFSOMs only receive values from shift register. Thus, as each new connection is encountered (at left), content of each shift register location is transferred one location (to right) with previous item in l^{th} location being lost. In case of n-feature architecture it is necessary to quantize number of neurons between 1st to 5th level RFSOMs. The purpose of 2nd to 5th level RFSOM is to provide an integrated view of input feature specific RFSOMs developed in 1st layer. There is potential for each neuron in 2nd to 5th layer RFSOM to have an input dimension defined by total neuron count across all 1st layer RFSOMs. This is brute force solution that does not scale computationally. Given topological ordering provided by RFSOM, neighboring neurons respond to similar stimuli. The topology of each 1st layer SOM is quantized in terms of fixed number of neurons using potential function clustering algorithm [2, 6, 8]. This reduces number of inputs in 2nd to 5th layers of RFSOM. The neurons in 4th layer acts as BMU for examples with same class label thus maximizing detection rate and minimizing false positives. However, there is no guarantee for this. In order to resolve this 4th layer SOM neurons that act as BMU for examples from more than one class are used to partition data. The 5th layer RFSOMs are trained on subsets of original training data. This enables size of 5th layer RFSOMs to increase which improves class specificity while presenting reasonable computational cost. Once training is complete 4th layer BMUs acts to identify which examples are forwarded to corresponding 5th layer RFSOMs on test dataset. A decision rule is required to determine under what conditions classification performance of BMU at 4th layer RFSOM is judged sufficiently poor for association with 5th layer RFSOM. There are several aspects that require attention such as minimum acceptable misclassification rate of 4th layer BMU relative to number of examples labeled at 4th layer BMU and number of examples 4th layer BMU represent. The basic implication is that there must be optimal number of connections associated with 4th layer BMU for training of corresponding 5th layer RFSOM and misclassification rate over examples associated with 4th layer BMU exceeds threshold. HRFSOM is characterized in terms of success probability in recovering true hierarchy H^* and runtime complexity. Some restrictions are placed to similarity function S [2] such that similarities scale with hierarchy upto some random noise: (a) For each $y_i \in Cs_j \in Cs^*$ and $j' \neq j$: $\min_{y_p \in Cs_j} \mathbb{Exp}[S(y_i, y_p)] - \max_{y_p \in Cs'_j} \mathbb{Exp}[S(y_i, y_p)] \geq \gamma > 0$. Here expectations are taken with respect to noise on S. (b) S2 For each $y_i \in Ct_j$, a set of V_j words of size v_j drawn uniformly from Cs_j satisfies:

$\mathbb{Prob}\left(\min_{y_p \in Cs_j} \mathbb{Exp}\left[S(y_i, y_p)\right] - \sum_{y_p \in V_j} \frac{S(y_i, y_p)}{v \cdot} > \epsilon\right) \leq 2e^{\left\{\frac{-2v_j\epsilon^2}{\sigma^2}\right\}}$. Here $\sigma^2 \geq 0$ parameterizes noise on similarity function S. From viewpoint of feature learning stacked RFSOMs extracts temporal features of sequences in weblog data. Various trade-offs are done towards improving representation ability and avoiding data over fitting. It is easy to overfit the network with limited data training sequences. This algorithm can be fine-tuned with heuristics.

3 Experiments and Results

In this section experimental results are presented for weblog prediction. When data is clustered by HRFSOM it is taken up by WUDA to discover request patterns and daily requests clusters. The fuzzy inference system uses patterns to infer meaningful information from data. During training process HRFSOM generated 5 clusters for hourly number of requests as shown in Table 1. The clusters generated in Table 1 are scattered in nature with respect to hourly requests and are almost identical. This classification ambiguity is resolved through WUDA by hourly page traffic volume and page requests. The clusters 4 and 5 have higher request and pages than clusters 1, 2 and 3. Using conventional web log analyzers it is difficult to understand daily traffic pattern. As a result of this data is clustered based on total activity for each day of week using volume of daily requests, pages and index value as input features. The training through HRFSOM generated 7 clusters as shown in Table 2. The clusters are separated according to access time as revealed by WUDA. After patterns are discovered through

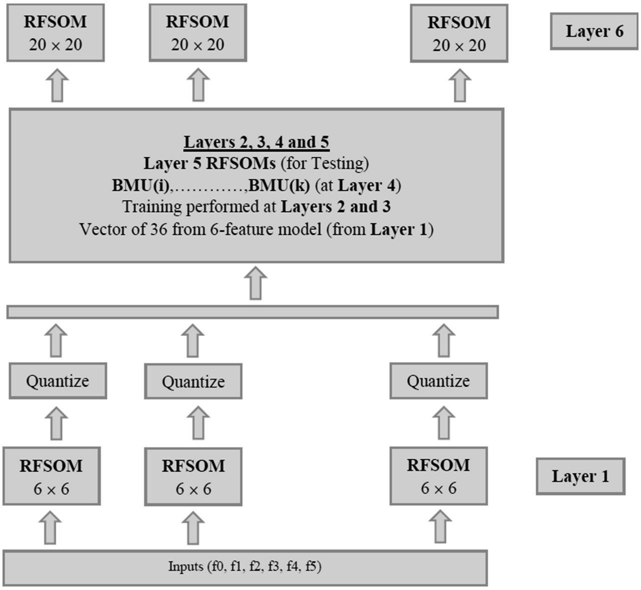

RFSOM
20 × 20
RFSOM
20 × 20
RFSOM
20 × 20
Layer 6
Layers 2, 3, 4 and 5
Layer 5 RFSOMs (for Testing)
BMU(i),…………,BMU(k) (at Layer 4)
Training performed at Layers 2 and 3
Vector of 36 from 6-feature model (from Layer 1)
Quantize
Quantize
Quantize
RFSOM
6 × 6
RFSOM
6 × 6
RFSOM
6 × 6
Layer 1
Inputs (f0, f1, f2, f3, f4, f5)

Fig. 2. The architecture of proposed HRFSOM model

Table 1. Hourly request clusters using HRFSOM

Clusters	Contribution percentage	Cluster contents (Percentage)
Cluster 1	18	14, 15, 16, 17, 18, 19, 20
Cluster 2	16	1, 2, 3, 4, 5, 6, 7, 21, 22, 23
Cluster 3	14	7, 8, 9, 10, 11, 13, 14, 20, 21, 22, 23
Cluster 4	25	9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21
Cluster 5	27	0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 18, 19, 20, 21, 22, 23

Table 2. Clustering of days using HRFSOM

Clusters	Contribution percentage	Cluster contents (Percentage)
Cluster 1	18	Sunday, Monday, Tuesday, Wednesday, Thursday
Cluster 2	14	Thursday, Friday
Cluster 3	30	Sunday, Monday, Tuesday, Wednesday, Thursday
Cluster 4	5	Monday, Tuesday, Wednesday
Cluster 5	24	Thursday, Friday
Cluster 6	5	Saturday, Sunday
Cluster 7	4	Saturday, Sunday

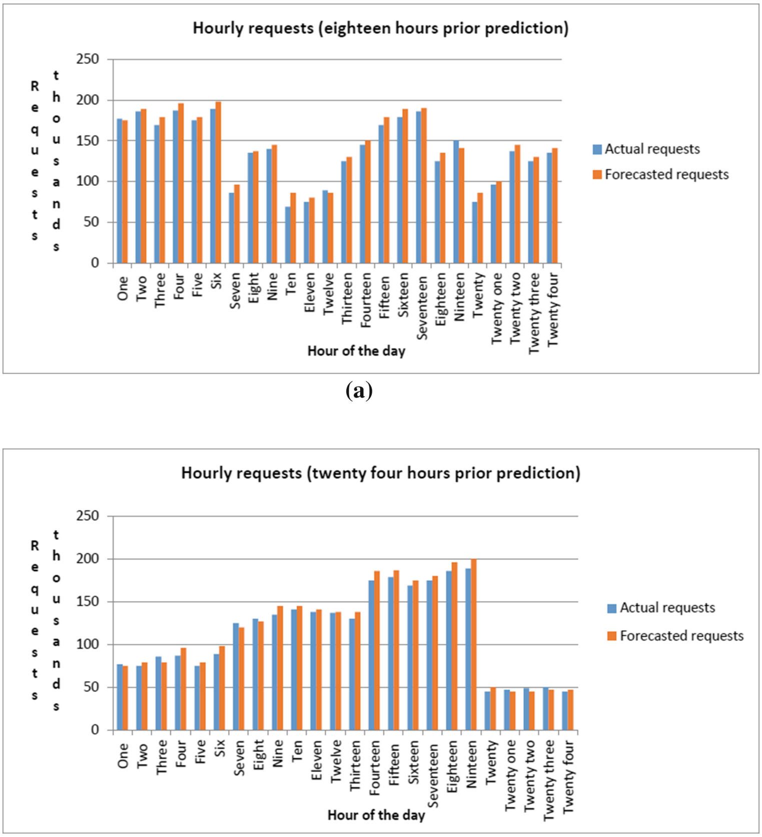
WUDA prediction is performed using variant of TSFIS [2]. The computing framework is based on rough fuzzy sets, fuzzy if then rules and fuzzy reasoning. The fuzzy rule is constituted by weighted linear combination of crisp inputs [2]. TSFIS performs interpretation in form of simple if then rules. Adaptive neuro fuzzy inference system (ANFIS) [8] and least mean squares (LMS) estimation [8] are used to fine tune antecedent and consequent rule parameters of TSFIS respectively. The data from 16th May 2010 to 31st December 2010 and data from 1st January 2011 to 16th May 2011 are used for training and testing purposes respectively. Grid partitioning is used to generate

Table 3(a). Prediction of daily traffic jam

Prediction period (days)	Root Mean Squared Error (RMSE)			
	Takagi Sugeno Fuzzy Inference System			
	With cluster location information		Without cluster location information	
	Training	Testing	Training	Testing
One	0.01569	0.03096	0.06475	0.09550
Two	0.05365	0.06998	0.10269	0.12747
Three	0.05067	0.05995	0.12845	0.14259
Four	0.05541	0.06886	0.11669	0.12998
Five	0.06847	0.07886	0.12486	0.14447
Six	0.06745	0.07465	0.12325	0.14396

Table 3(b). Prediction of hourly traffic jam

Prediction period (hours)	Root Mean Squared Error (RMSE)			
	Takagi Sugeno Fuzzy Inference System			
	With cluster location information		Without cluster location information	
	Training	Testing	Training	Testing
One	0.03434	0.03537	0.09579	0.09986
Six	0.04559	0.05477	0.09789	0.98989
Twelve	0.06519	0.06769	0.10095	0.11517
Eighteen	0.05596	0.06777	0.10477	0.10986
Twenty Four	0.05736	0.06696	0.10595	0.10769



(a)

(b)

Fig. 3. **(a)** Prediction of traffic jam eighteen hours ahead **(b)** Prediction of traffic jam twenty four hours ahead

Table 4. The comparison of clustering accuracy of SOM, FSOM, RFSOM and HRFSOM

Analysis algorithms	Accuracy percentage (with respect to web user access patterns)
SOM	70
FSOM	79
RFSOM	86
HRFSOM	96

initial rule base. Only small number of membership functions are required for each input and 2D spaces are partitioned using trapezoidal membership functions [2]. Alongwith regular inputs, cluster location information from HRFSOM output is also used. Based on these inputs fuzzy inference models are developed to predict web traffic volume on hourly and daily basis. Once traffic volume of a particular day is available the model predicts daily traffic upto 7 days ahead. The Tables 3(a) and (b) summarizes performance of fuzzy inference system for training and testing data in days and hours respectively. The prediction of hourly traffic is done upto 24 h ahead. The Fig. 3(a) and (b) represent test results for 18 and 24 h ahead prediction of hourly web traffic volume. The accuracy of clustering results of web user access patterns obtained from HRFSOM are highly significant are highly significant as shown in Table 4 when compared to other clustering algorithms such as SOM, fuzzy self-organizing map (FSOM) [2] and RFSOM.

4 Conclusion

The knowledge discovery from data in terms of relevant user information and access patterns allows organizations to predict user's future access patterns. This helps in further development, planning and maintenance towards advertising campaigns aimed at target user groups. The web user access patterns calls for incorporation of machine intelligence techniques for mining meaningful information. The statistical web log data is generated by AWStats web access log file analyzer. HRFSOM clusters and analyzes information related to user access patterns from statistical data. HRFSOM is developed hierarchically by incorporating rough fuzzy set with SOM and has superior clustering accuracy. The cluster information is used by TSFIS variant to predict daily and hourly traffic. The results indicate cluster information significance to improve prediction accuracy of inference system. The experimental results demonstrate superiority of proposed method.

References

1. Jansen, B.J.: Understanding User-Web Interactions via Web analytics. 1st edn. Synthesis Lectures on Information Concepts, Retrieval and S. Morgan and Claypool Publishers (2009)
2. Chaudhuri, A.: Weblog Prediction with Machine Leaning Methods. Technical report, Samsung R&D Institute Delhi India (2016)

3. Clifton, B.: Advanced Web Metrics with Google Analytics. 3rd edn., Sybex (2012)
4. Yandex Personalized Web Search Challenge 2014. https://www.kaggle.com/c/yandex-personalized-web-search-challenge/data
5. AWStats web log file analyzer. http://www.awstats.org/
6. Kohonen, T.: Self-Organizing Map, 3rd Extended edn. Springer Series in Information Sciences, vol. 30. Springer, Heidelberg (2001)
7. Lingras, P.: Fuzzy rough and rough fuzzy serial combinations in neurocomputing. Neurocomputing. **36**(1), 29–44 (2001)
8. Pratihar, D.K.: Soft Computing: Fundamentals and Applications, 1st edn. Alpha Science International Ltd. (2013)