Original Research Paper

Volume-9 | Issue-7 | July - 2019 | PRINT ISSN No. 2249 - 555X

Computer Science

SIGNIFICANCE STUDY OF USER WEB ACCESS RECORDS MINING FOR BUSINESS INTELLIGENCE

Amish D. Vyas Ph. D. Scholar JJTU (Rajasthan)

Dr. Yogesh Kumar Sharma* Associate Professor, (HOD), Computer Department JJTU (Rajasthan) *Corresponding Author

ABSTRACT This paper discusses about how business intelligence on a website could be obtained from users' access records instead of web logs of "hits". In this paper, we present the important concepts of Web usage mining and its various practical applications. We further present a novel approach 'intelligentminer' (i-Miner) to optimize the concurrent architecture of a fuzzy clustering algorithm (to discover web data clusters) and a fuzzy inference system to analyze the Web site visitor trends. A hybrid evolutionary fuzzy clustering algorithm is proposed in this paper to optimally segregate similar user interests. The clustered data is then used to analyze the trends using a Takagi-Sugeno fuzzy inference system learned using a combination of evolutionary algorithm and neural network learning. Due to intense competition on one hand and the customer's option to choose from several alternatives business community has realized the necessity of intelligent marketing strategies and relationship management. Web usage mining attempts to discover useful knowledge from the secondary data obtained from the interactions of the users with the Web. Web usage mining has become very critical for effective Web site management, creating adaptive Web sites, business and support services, personalization, network traffic flow analysis and so on. In this paper, we present the important concepts of Web usage mining and its various practical applications. Data mining and business intelligence techniques can be integrated in order to develop more advanced decision support systems. They describe the Web mining process as a sequence of steps for the development of advanced decision support systems. By following such a sequence, the authors can develop advanced decision support systems, which integrate data mining with business intelligence, for websites.

KEYWORDS : Business intelligence, data mining, i-Miner, fuzzy clustering

1.0 INTRODUCTION:

Nowadays most companies have moved into e-business where they use Internet as a platform to interact with their customers. This has created more touch points than the traditional face-to-face visits, and therefore we have a new realm called Web Business Intelligence to deal with. In Web Business Intelligence, for targeting the business areas of decision support and personalized service, we have identified three main technologies namely Data-warehousing (that includes data-mining modules and decision support modules), Web log analyzer and Web spy. Each one of them provides certain insight or understanding of one of the many aspects of a business respectively. In particular, we state that having a web log analyzer alone is not sufficient to provide a complete picture of business intelligence. Nevertheless, web-log analysis gives very good detailed reports on web site traffic statistics. Other information such as competitors, market positions, and most importantly the customers and their behaviors on the website needed also to be known. In our case, e-Personalization that means personalizing web-content according to the customer's interest/needs will be guided by the knowledge that we have on him.

To realize the objectives of Web Business Intelligence, we propose an Access-control (AC) architecture that can progressively built upon a basic website. The objective is to distribute the functions of e-CRM and business intelligence using an agent-architecture and to monitor users' accesses on the website.

We have applied our proposed web mining process in the development of intelligent monitoring/management systems to guarantee the quality of web sites. Examples of monitoring activities include:

- **Usage:** Keep track of the paths users take during their accesses, the efficiency of pages/hyperlinks in guiding the users to accomplish their goals;
- **Users:** How users are grouped taking into account their browsing behavior, how groups change with time, how groups of users relate with the success of the site;
- **Data quality:** How adequate the content and meta-data of a web site are;
- **Automation:** The effect of personalization actions. For instance, if users are following the recommendations of products and pages or not.

2.0 LITERATURE REVIEW

Suresh Babu D, (2011) Web usage mining is a main research area in Web mining focused on learning about Web users and their interactions with Web sites. The motive of mining is to find users' access models automatically and quickly from the vast Web log data, such as frequent access paths, frequent access page groups and user clustering. Through web usage mining, the server log, registration information and other relative information left by user provide foundation for decision making of organizations. This article provides a survey and analysis of current Web usage mining systems and technologies. This paper also discusses an application of WUM, an online Recommender System that dynamically generates links to pages that have not yet been visited by a user and might be of his potential interest. Differently from the recommender systems proposed so far, ONLINE MINER does not make use of any off-line component, and is able to manage Web sites made up of pages dynamically generated.

Govindasamy V, (2014) Web Usage mining helps in finding the user needs by analyzing the web server log files to make the administrators of the web sites to modify their web site to attract more number of users. This is very vital in commercial sites where the administrator has to know what his customer wants or for what the users prefer their site to other such web sites. Web usage mining is the process of getting that information about the users through the log files. This mining creates a access pattern which shows how a user uses a particular site. i.e the sequence of web pages the user has used.

Vijiyarani S, (2015) Web is a collection of inter-related files on one or more web servers while web mining means extracting valuable information from web databases. Web mining is one of the data mining domains where data mining techniques are used for extracting information from the web servers. The web data includes web pages, web links, objects on the web and web logs. Web mining is used to understand the customer behaviour, evaluate a particular website based on the information which is stored in web log files. Web mining is evaluated by using data mining techniques, namely classification, clustering, and association rules. It has some beneficial areas or applications such as Electronic commerce, E-learning, Egovernment, E-policies, E-democracy, Electronic business, security, crime investigation and digital library. Retrieving the required web page from the web efficiently and effectively becomes a challenging task because web is made up of unstructured data, which delivers the large amount of information and increase the complexity of dealing information from different web service providers.

3.0 METHODOLOGY:

Mining Framework Using Intelligent Miner (i-Miner):

The i-Miner hybrid framework optimizes a fuzzy clustering algorithm using an evolutionary algorithm and a Takagi–Sugeno fuzzy inference

system using a combination of evolutionary algorithm and neural network learning. The raw data from the log files are cleaned and pre-processed and a fuzzy C means algorithm is used to identify the number of clusters. The architecture is illustrated in Fig. 1. The developed clusters of data are fed to a Takagi-Sugeno fuzzy inference system to analyze the trend patterns. The if-then rule structures are learned using an iterative learning procedure by an evolutionary algorithm and the rule parameters are fine-tuned using a back propagation algorithm.

The hierarchical distribution of the *i*-Miner is depicted in Fig. 2. The arrow direction depicts the speed of the evolutionary search. The optimization of

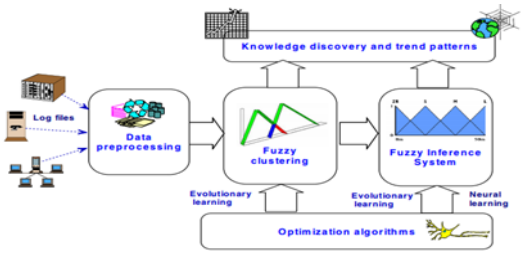


Figure 1: *i*-Miner framework.

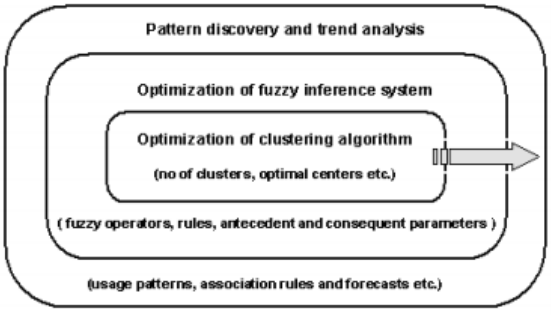


Figure 2: Hierarchical architecture of *i*-Miner.

clustering algorithm progresses at a faster time scale in an environment decided by the inference method and the problem environment. More technical details are provided in the subsequent sections.

Clustering algorithms

Fuzzy clustering algorithm

One of the widely used clustering methods is the fuzzy c-means (FCM) algorithm developed by Bezdek (1981). FCM partitions a collection of n vectors x_i, i = 1, 2 . . . , n into c fuzzy groups and finds a cluster center in each group such that a cost function of dissimilarity measure is minimized. To accommodate the introduction of fuzzy partitioning, the membership matrix U is allowed to have elements with values between 0 and 1. The FCM objective function takes the form

$$J(U, c_1, \ldots, c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{ij}^m d_{ij}^2 \quad (1)$$

where u_{ij} is a numerical value between [0, 1]; c_i is the cluster center of fuzzy group *i*; $d_{ij} = k_{ci} - x_{jk}$ is the Euclidian distance between *i*th cluster center and *j*th data point; and m is called the exponential weight which influences the degree of fuzziness of the membership (partition) matrix.

Optimization of fuzzy clustering algorithm Usually a number of cluster centers are randomly initialized and the FCM algorithm provides an iterative approach to approximate the minimum of the objective function starting from a given position and leads to any of its local minima. No guarantee ensures that FCM converges to an optimum solution (can be trapped by local extrema in the process of optimizing the clustering criterion). The performance is very sensitive to initialization of the cluster centers. An evolutionary algorithm is used to decide the optimal number of clusters and their cluster centers. The algorithm is initialized by constraining the initial values to be within the space defined by the vectors to be clustered. A very similar approach is given in. In the i-Miner approach, the fuzzy clustering algorithm is optimized jointly with the trend analysis algorithm (fuzzy inference system) in a single global search.

Self organizing map (SOM)

A self organizing map was used to cluster the user access records. The SOM is an algorithm used to visualize and interpret large high-dimensional data sets. The map consists of a regular grid of processing units, "neurons". A model of some multidimensional observation, eventually a vector consisting of features, is associated with each unit. The map attempts to represent all the available observations with optimal accuracy using a restricted set of models. At the same time the models become ordered on the grid so that similar models are close to each other and dissimilar models far from each other. Fitting of the model vectors is usually carried out by a sequential regression process, where t = 1, 2, . . . is the step index: For each sample x(t), first the winner index c (best match) is identified by the condition

$$\forall_i, \|x(t) - m_c(t)\| \leq \|x(t) - m_i(t)\| \quad (2)$$

After that, all model vectors or a subset of them that belong to nodes centered around node c = c(x) are updated

$$m_i(t+1) = m_i(t) + h_{c(x),i}(x(t) - m_i(t)) \quad (3)$$

Here $h_{c(x),\ i}$ is the neighborhood function, a decreasing function of the distance between the ith and cth nodes on the map grid. This regression is usually reiterated over the available samples.

Artificial neural network (ANN)

ANNs were designed to mimic the characteristics of the biological neurons in the human brain and nervous system. Learning typically occurs by example through training, where the training algorithm iteratively adjusts the connection weights (synapses). Back propagation (BP) is one of the most famous training algorithms for multilayer perceptrons. BP is a gradient descent technique to minimize the error E for a particular training pattern. For adjusting the weight (w_{ij}) from the i^{th} input unit to the jth output, in the batched mode variant the descent is based on the gradient $\nabla E \frac{\delta E}{\delta E\, \delta w_{ij}}$ for the total training set

$$\Delta w_{ij}(n) = -\varepsilon^* \frac{\delta E}{\delta w_{ij}} + \alpha^* \Delta w_{ij}(n-1) \quad (4)$$

The gradient gives the direction of error E. The parameters ε and α are the learning rate and momentum respectively.

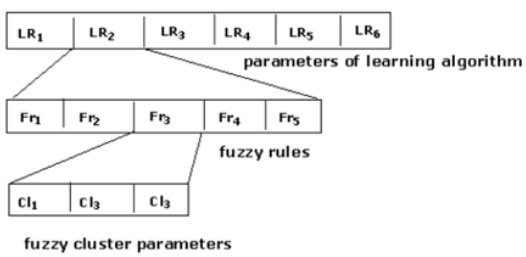
Linear genetic programming (LGP)

Linear genetic programming is a variant of the GP technique that acts on linear genomes. Its main characteristics in comparison to tree-based GP lies in that the evolvable units are not the expressions of a functional programming language (like LISP), but the programs of an imperative language (like c/c ++). An alternate approach is to evolve a computer program at the machine code level, using lower level representations for the individuals. This can tremendously hasten up the evolution process as, no matter how an individual is initially represented, finally it always has to be represented as a piece of machine code, as fitness evaluation requires physical execution of the individuals.

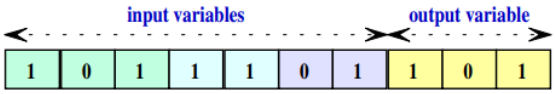
Fuzzy inference systems (FIS)

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 fuzzy inference systems, which are the most important modelling tools based on fuzzy set theory. A Takagi Sugeno fuzzy inference scheme was made used of in which the conclusion of a fuzzy rule is constituted by a weighted linear combination of the crisp inputs rather than a fuzzy set. The Adaptive Network Based Fuzzy Inference System (ANFIS) was used in the simulation, which implements a Takagi Sugeno fuzzy inference system learned using neural network learning.

For every fuzzy inference system, there exist a global search of neural network learning algorithm parameters, parameters of the fuzzy operators, if-then rules and



Chromosome structure of the *i*-Miner.



Chromosome representing an individual fuzzy rule (3 input variables and 1 output variable).

4.0 RESULTS

Experimentation Setup-Training and Performance Evaluation

In this research, we used the statistical/ text data generated by the log file analyzer from 01 January 2002 to 07 July 2002. Selecting useful data is an important task in the data pre-processing block. After some preliminary analysis, we selected the statistical data comprising of domain byte requests, hourly page requests and daily page requests as focus of the cluster models for finding Web users' usage patterns. It is also important to remove irrelevant and noisy data in order to build a precise model. We also included an additional input 'index number' to distinguish the time sequence of the data. The most recently accessed data were indexed higher while the least recently accessed data were placed at the bottom. Besides the inputs 'volume of requests' and 'volume of pages (bytes)' and 'index number', we also used the 'cluster information' provided by the clustering algorithm as an additional input variable. The data was re-indexed based on the cluster information. Our task is to predict (few time steps ahead) the Web traffic volume on a hourly and 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.

Table 1. Parameter settings of *i*-Miner

Population size	50
Maximum no of generations	32
Fuzzy inference system	Takagi Sugeno
Rule antecedent membership functions Rule consequent parameters	3 membership functions per input variable (parameterized Gaussian) linear parameters
Gradient descent learning	10 epochs
Ranked based selection	0.50
Elitism	5 %
Starting mutation rate	0.50

The initial populations were randomly created based on the parameters shown in Table 1. We used a special mutation operator, which decreases the mutation rate as the algorithm greedily proceeds in the search space defines the mutation step, where γ is the random number from the interval [0,1] and t is the current generation and tmax is the maximum number of generations. The function Δ computes a value in the range [0,*x*] such that the probability of returning a number close to zero increases as the algorithm proceeds with the search. The parameter b determines the impact of time on the probability distribution Δ over [0,*x*]. Large values of b decrease the likelihood of large mutations in a small number of generations. The parameters mentioned in Table 1 were decided after a few trial and error approaches. Experiments were repeated 3 times and the average performance measures are reported. Figures 3 and 4 illustrates the meta-learning approach combining evolutionary learning and gradient descent technique during the 35 generations

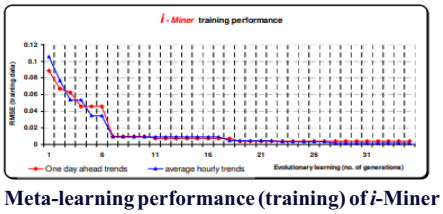


Figure 3. Meta-learning performance (training) of *i*-Miner

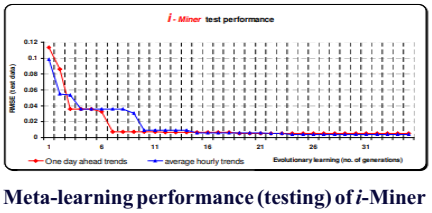


Figure 4. Meta-learning performance (testing) of *i*-Miner

Table 2 summarizes the performance of the developed i-Miner for training and test data. Performance is compared with the previous results wherein the trends were analyzed using a Takagi-Sugeno Fuzzy Inference System (ANFIS), Artificial Neural Network (ANN) and Linear Genetic Programming (LGP). The Correlation Coefficient (CC) for the test data set is also given in Table 2. The 35 generations of meta-learning approach created 62 if-then Takagi-Sugeno type fuzzy rules (daily traffic trends) and 64 rules (hourly traffic trends) compared to the 81 rules reported in.

Table 2. Performance of the different paradigms

	Period					
	Daily (1 day ahead) RMSE		CC	Hourly (1 hour ahead) RMSE		CC
Method	Train	Test		Train	Test	
i-Miner	0.0042	0.0053	0.9967	0.0012	0.0041	0.9981
TKFIS	0.0164	0.0402	0.9953	0.0432	0.0433	0.9841
ANN	0.0345	0.0481	0.9292	0.0546	0.0639	0.9493
LGP	0.0543	0.0749	0.9315	0.0654	0.0516	0.9446

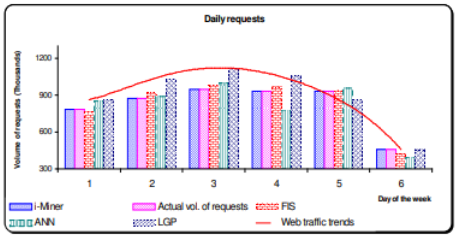


Figure 5. Test results of the daily trends for 6 days

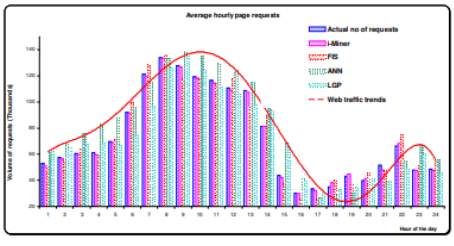
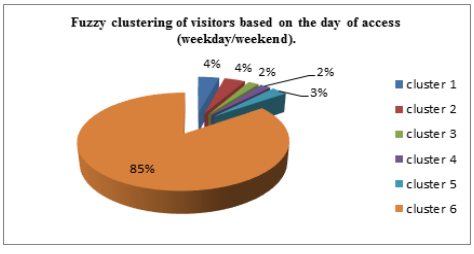


Figure 6. Test results of the average hourly trends for 6 days

Figures 5 and 6 illustrate the actual and predicted trends for the test data set. A trend line is also plotted using a least squares fit (6 th order polynomial). FCM approach created 7 data clusters for hourly traffic according to the input features compared to 9 data clusters for the daily requests. The previous study using Self-organizing Map (SOM) created 7 data clusters (daily traffic volume) and 4 data clusters (hourly traffic volume) respectively. As evident, FCM approach resulted in the formation of additional data clusters. Several meaningful information could be obtained from the clustered data. Depending on the volume of requests and transfer of bytes, data clusters were formulated. Clusters based on hourly data show the visitor information at certain hour of the day.



Graph 1: Fuzzy clustering of visitors based on the day of access (weekday/weekend).

5.0 CONCLUSIONS

Recently Web usage mining has been gaining a lot of attention because of its potential commercial benefits. The proposed i-Miner framework seems to work very well for the problem considered. The empirical results also reveal the importance of using soft computing paradigms for mining useful information. In this chapter, our focus was to develop accurate trend prediction models to analyze the hourly and daily web traffic volume. Several useful information could be discovered from

the clustered data. FCM clustering resulted in more clusters compared to SOM approach. Perhaps more clusters were required to improve the accuracy of the trend analysis. The main advantage of SOMs comes from the easy visualization and interpretation of clusters formed by the map. The knowledge discovered from the developed FCM clusters and SOM could be a good comparison study and is left as a future research topic. As illustrated in Table 2, i-Miner framework gave the overall best results with the lowest RMSE on test error and the highest correlation coefficient.

REFERENCES:

1. Abraham A. (2001), Neuro-Fuzzy Systems: State-of-the-Art Modeling Techniques, Connectionist Models of Neurons, Learning Processes, and Artificial Intelligence, Jose Mira and Alberto Prieto (Eds.), Lecture Notes in Computer Science 2084, Springer-Verlag Germany, Spain, pp. 269-276.
2. Aggarwal, C., Wolf J.L., Yu, P.S. (1999): Caching on the World Wide Web. IEEE Transaction on Knowledge and Data Engineering, vol. 11, no. 1, pp. 94-107.
3. Banzhaf. W., Nordin. P., Keller. E. R., Francone F. D. (1998), Genetic Programming : An Introduction on The Automatic Evolution of Computer Programs and its Applications, Morgan Kaufmann Publishers, Inc
4. Bezdek, J. C. (1981), Pattern Recognition with Fuzzy Objective Function Algorithms, New York: Plenum Press.
5. Chakrabarti S. (2003), Mining the Web: Discovering Knowledge from Hypertext Data, Morgan Kaufmann Publishers.
6. Chang, G., Healey, M.J., McHugh, J.A.M., Wang, J.T.L. (2001): Web Mining, Mining the World Wide Web. Kluwer Academic Publishers, Chapter 7, pp. 93-104.
7. Heer, J. and Chi E.H. (2001), Identification of Web User Traffic Composition using Multi- Modal Clustering and Information Scent, In Proc. of the Workshop on Web Mining, SIAM Conference on Data Mining, pp. 51-58.
8. Jang R. (1992), Neuro-Fuzzy Modeling: Architectures, Analyses and Applications, PhD Thesis, University of California, Berkeley.