Artificial Neural Network Approach for Student's Behavior Analysis

Harish Kumar, Anil Kumar Solanki, Anuradha Taluja

Abstract: India is one of the leading countries in the word for technical education and management education. A study shows that in India e-education is growing with a massive growth. The students find relevant information from various e-education web sites. Numbers of pages are indexed on www and finding the desired information is not an easy task. For solving this problem we need an approach which helps in finding precise solution as per students need. In this paper we also use Web log analysis via neural network structure. We are using a neural approach with fuzzy clustering that shows the analysis of web logs, depends on the performance of the clustering the number of requests. Clustered data is used to analyze with neural approach for effective web site modification and behavior analysis. The proposed model use neural network approach with a firing rule to discover and analyze useful knowledge from the available Web log data.

I. INTRODUCTION

Since last twenty years, IT industry has observed an explosive growth on the World Wide Web (WWW). Impressive development in the past decades in education and research, e-material is increased with a massive growth rate. According to the survey, the details of which were released by the online search giant, over 45% Indian students use the Internet for research on education[1][3]. Recent studies estimate the educational web having over millions of pages [7], and pages are being added and deleted every day. Advancement in day by day puts a special effect on the education system. This excellent growth of the Web prompted the development of new domains of application, the Web Mining being one of them [4]. Besides a large amount of educational content information stored on web pages, web pages also contain a rich and dynamic collection of hyperlink information [6]. Students uses the Internet to find out the required information because the Internet is an infinite source of data that can come either from the Web content, represented by the billions of pages publicly available, or from the Web usage, represented by the log information collected daily by all the servers around the world [1] [2]. Extract useful knowledge from WWW data is considered as web mining. The users want to have the effective search tools to find relevant information easily and precisely Millions of students and millions of web pages are on Internet, numbers of pages are indexed and finding the desired information is not an easy task.

Manuscript received March, 2014.

Prof A.K.Solanki has obtained his PhD in Computer Science and Engineering from Bundelkhand University.he has published good no of papers in national and International Journals.His area of specialization is Data mining.

Harish Kumar has completed his M.Tech (IT) from Guru Gobind Singh Indraprastha University, Delhi. Now currently pursuing his PhD from Mewar University, Chittorgarh.

Anuradha Taluja has completed her M.Tech from Guru Gobind Singh Indraprastha University, Delhi. Her area of specializati on is Data mining.

The Internet browsing results generates a huge quantity of data. The Web service providers want to find the way to predict the users' behaviors and personalize information to reduce the traffic load and design the Web-site suited for the different group of users (students). Their motto is to collect more information for their students.

The existing web solutions provided are not sufficient to satisfy the needs of different web users. Online education concerns rely on web usage analysis to obtain students behavior for education promotion. This helps in upgrading higher education. Web navigation generates the massive data related to student's interactions with the educational web sites [5]. This massive data is in the form on web logs or server log files. Web Log analysis is used for discovering similar patterns in Web log data .These patterns are used for analyzing the navigational behavior of the students. Web usage mining also known as web log mining. Web log Mining has become very critical for effective Web site management, business, support services, personalization, network traffic flow analysis and so on.

II. RELATED WORK

Ajith Abraham et, al. provides an approach for effective Web site management, business and support services, personalization, network traffic flow analysis and so on[1]. Their work is based on neuro-fuzzy approach that has shown that the usage trend analysis which is very much depends on the performance of the clustering of the number of requests. They proposed an 'intelligent-miner' (i-Miner) to optimize the concurrent architecture of a fuzzy clustering algorithm (to discover data clusters) and a fuzzy inference system to analyze the trends[10]. In the concurrent neuro-fuzzy approach, self organizing maps were used to cluster the web user requests.

Wang X et.al proposed that self-organizing map (SOM) has been used to cluster the usage requests and also developed several soft computing paradigms to analyze the Web usage trends. Empirical results have clearly shown the importance of the clustering algorithm to analyze the user access trends [10]. A hierarchical evolutionary approach has been proposed in this paper to optimize the clustering algorithm and the fuzzy inference system to improve the performance.

Vaishali A.Zilpe et al. proposed that web servers are surely the richest and the most common source of data. They can collect large amounts of information in their log files and in the log files of the databases they use. These logs usually contain basic information e.g. name and IP of the remote host, date and time of the request, the request line exactly as it came from the client, etc. Accurate Web usage information could help to attract new customers, retain current customers, improve cross marketing/sales, effectiveness of promotional campaigns, tracking leaving customers and find the most effective

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

logical structure for their Web space. User profiles could be built by combining users' navigation paths with other data features, such as page viewing time, hyperlink structure, and page content [15]. ART can also be implemented with all previous techniques like semantic Web log, hybrid information filtering, fuzzy immunity clonal selection neural network, and fuzzy multi-set to build Multi-pass ART, and provide more efficient result.

Olfa et al .proposed that flow of information in a Web personalization system can be prone to significant amounts of error and uncertainty [11]. This uncertainty pervades all stages from the user's web navigation patterns to the final recommendations. Fuzzy approximate reasoning seems to be a natural framework for the recommendation process. They presented a simple, intuitive, and fast approach to provide dynamic predictions in the Web navigation space. Real noisy Web usage data was used as a simulation test bed for the fuzzy recommendation system. The proposed approach is efficient since only pre discovered profiles (offline) need to be compared. They took advantage of the sparsity of Web usage data to enable a direct storage and access to the relation matrix's columns by hashing. Their proposed approach is fit for real-time recommendations (on average less than 0.02 secs. per recommendation on a 2 GHz Pentium 4 Linux PC). This makes fuzzy recommendations suitable for real time recommendations in a live setting on today's most active and huge websites. We are currently performing more offline and online tests using various recommendation strategies.

Archana et,al. deals with the ambiguity and the uncertainty underlying Web interaction data, fuzzy reasoning appears to be an effective tool[13]. In her work, she use the fuzzy clustering to categorize user sessions in order to derive groups of users which exhibit similar access patterns from web log data. The obtained clusters which can be exploited to implement different personalization functions, such as dynamic suggestion of links to Web pages retained interesting for the user.

III. WEB MINING

Web mining may be classified into three categories

1. Web Content Mining
2. Web Structure Mining.
3. Weblog mining(Web Usage Mining)

Web content mining (WCM) is to find useful information from the content of web pages [3][12] like free semi-structured data such as HTML code, pictures, and various unloaded files. Web structure mining (WSM) is use to generating a structural summary about the web site and web pages. Web structure mining tries to discover the link structure of the hyperlinks at the inter document level [12]. Web usage mining (WUM) is related to the data generated by visitors of a web site. User leaves their navigation footprints in the form of web logs. Web usage mining (WUM) or web log mining is used for web site modification and web users' behavior analysis. Web usage mining (WUM) or web log mining, users' behavior or interests is revealed by applying data mining techniques on web. Web log files are of different types [3][12].

1. Access Log File.
2. Agent Log File
3. Referrer Log File
4. Error Log File

Access Log File: It records information about which files are being requested from web server. It is located in the directory www/logs/.

Agent Log File: It records information about the web clients that make requests on your server.

Referrer Log File: It records information about the URL that the web browser had been viewing immediately before making the request on your server. This is particularly useful when you want to determine where requests on your web server come from and what websites are referring web traffic to your server. It is located in the www/logs/ directory and called Referrer Log File.

Error Log File: It records information about failed requests of your server. If someone tries to access a file on your server that doesn't exist, your server automatically generates an error message. Each of these error messages is recorded in the referrer log. It is located in the www/logs/ directory and called Error Log File.

Three main sources of web log file are

1. Client Log File,
2. Proxy Log File
3. Server Log File.

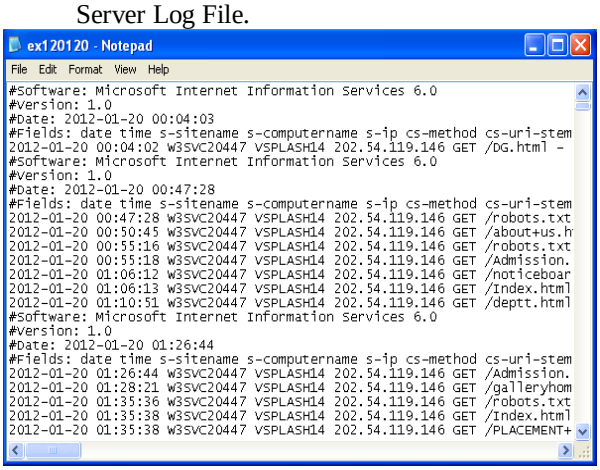

Figure 1: Web Log File

A log file contains the following field

- The client's host name or its IP address. example: 174.129.228.67
- The client id (generally empty and represented by a \-")
- The user login (if applicable),
- The date and time of the request. Example: 2012-01-20 00:47:28
- The operation type (GET, POST, HEAD, etc.),
- The requested resource name.
- The request status.
- The requested page size.
- The user agent (a string identifying the browser and the operating system used)
- The referrer of the request which is the URL of the Web page containing the link that the user followed to get to the current page.

For discovery and analysis of usage patterns from the available web log data, it is necessary to perform three steps: Preprocessing, Pattern Discovery, Pattern Analysis.

1. Preprocessing: Data preprocessing phase of web usage mining is completed in four steps.

A) Cleaning
B) User/Session Identification
C) Transaction Identification

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

D) Transformation

DUSTER ALGO [12] is use for cleaning the web log data and its complexity is much better than the other cleaning algorithms. Before Compiling size of web log file is 3153Kb. After 1 pass cleaning size is reduce up to 1,289 Kb. Second pass of cleaning size is approximate 1.1 Mb. In this phase web log data must be cleaned, filtered, integrated and transformed in such a way that the irrelevant and redundant data can be removed[16].User session and transaction are identified and stored in a 2 dimensional matrix data form for analysis. This data must be assembled into a reliable and integrated form in order to be used for pattern discovery.

2. Pattern discovery: Pattern discovery uses methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. Once the domain-dependent data transformation phase is completed, the resulting transaction data must be formatted to conform to the data model of the appropriate data mining task [16].
3. Pattern analysis: The discovery of user access patterns from the user access logs, referrer logs, user registration logs etc is the main purpose of the Web Usage Mining activity. Pattern discovery is performed only after cleaning the data and after the identification of user transactions and sessions from the access logs. The analysis of the pre-processed data is very beneficial to all the organizations performing different businesses over the web. The tools used for this process use techniques based on AI, data mining algorithms, psychology, and information theory.

IV. NEURAL NETWORK IN WEB USAGE MINING

Neural network is an extremely simplified model of the brain. Brain contains about 1010 basic units of neurons. Each neuron in turn is connected to about 104 other neurons. An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information[14]. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (Neurons) working in unison to solve specific problems. An ANN is configured for a specific application, such as similar pattern matching or data classification, through a learning process. Main function of brain and ANN system is to transforms inputs into outputs to the best of its ability

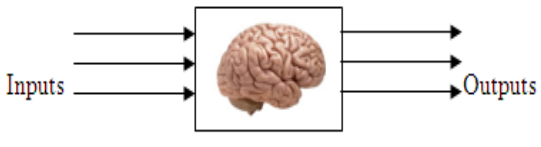

Figure2: Brain Neural system

Figure 3: ANN system

The neuron has two modes of operation: the training mode and the using mode. In the training mode, the neuron can be trained to fire (or not), for particular input patterns. In the using mode, when a taught input pattern is detected at the input, its associated output becomes the current output. If the input pattern does not belong in the taught list of input patterns, the firing rule is used to determine whether to fire or not.

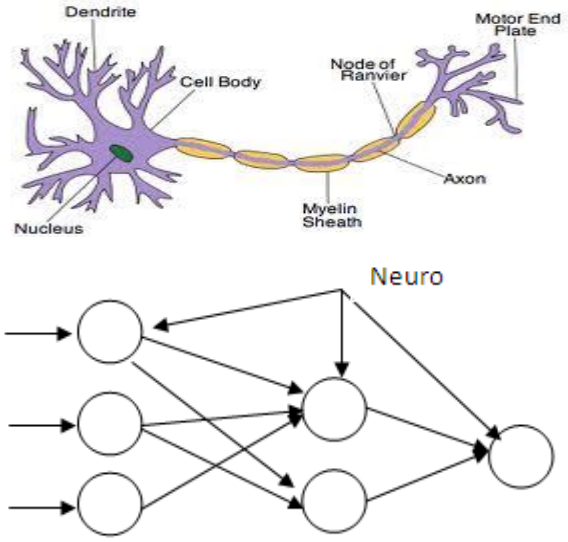

Figure 4: Brain Neuron System and Artificial neuron system

Web graph is similar to the neuron system and we can use ANN approach for predicting the next user movement and web site modification. The main idea is move around the link graph of web logs. Each IP is associated with a user and having a threshold time variant with it. The output of a neuron is a function of the weighted sum of the inputs plus a bias. An artificial neuron is a device with many inputs and one output.

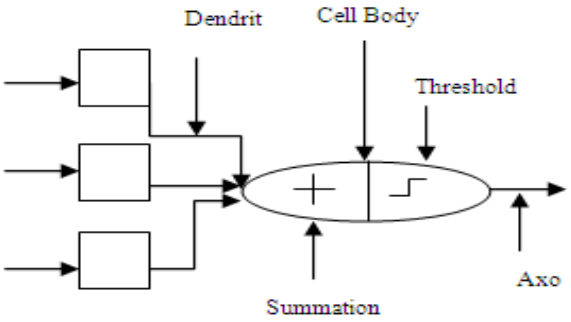

Figure 5: The Neuron Model

V. EXPERIMENTAL SETUP

In web usage mining P1, P2, P3, ……., Pn. are the input to the artificial neurons and t1,t2,t3,…….tn are the time attach to the input links. It is .this acceleration or retardation of the input signal that is modelled by the weights (Time). Means a page having a larger time and if extra bias value also supports the page, it indicates that the particular page having higher precedence with others. Consider the neural network with two inputs, one hidden neuron, and one output. It has page visit and their associated elapsed time. We apply firing rule on this with the firing algorithm. Then, we will train learning network to approximate them. Besides these changes, we will sample all parameter and one expected value on output.

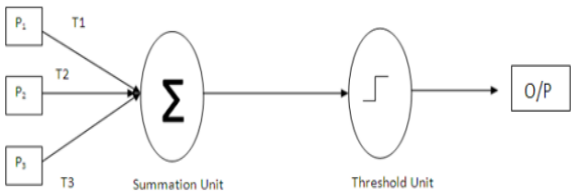

Figure 6: Shows the Neuron model for a web session

The output is used for clustering designing. The developed clusters of data are fed to a web ASTRO fuzzy system to analyze the patterns. The optimization of clustering algorithm progresses at a faster time scale in an environment decided by the inference method and the problem environment. A simple firing rule can be implemented by using Hamming distance technique. The rule goes as follows: Take a collection of training patterns for a node, some of which cause it to fire (the 1-taught set of patterns) and others which prevent it from doing so (the 0-taught set). Then the patterns not in the collection cause the node to fire if, on comparison, they have more input elements in common with the 'nearest' pattern in the 1-taught set than with the 'nearest' pattern in the 0-taught set. If there is a tie, then the pattern remains in the undefined state. For the purpose of testing the efficiency of algorithm in, our college data is used.

Access Page	Actual Corresponding web Page
A	Department
B	Intake
C	Assignment
D	Placement
E	HOD

Table 1: Actual Corresponding Page Table

For example, a 3-input neuron is taught to output 1 when the input (U1, U2 and U3) is 111 or 110 and to output 0 when the input is 000 or 001. Then, before applying the firing rule, the table is

	A	B	C	D	E	F
U1	0	1	1	0	1	1
U2	0	0	1	0	1	1
U3	0	0	0	1	0	1
O/P	0	0	1	0	0	1

Table 2: Table generated after applying firing rule

After checking the algorithm for a particular session, we apply this for bulk of session file. Suppose that Navigation pattern sequences are as follows

Navigation Pattern	Frequency of visit
S A B C D E F T	3
S A C F T	2
S A C E T	3
S B C D T	2
Total No of web site navigate	10

Table 3: Page Frequency Table

Suppose we have state space say S= {S1, S2…, Sn) at the time t sate sequence is represented by St and transition probability is represented by Pi j. In first order Markov chain model state probability is depend on the previous state for example probability of state j depends on the previous state i. So transition probabilities are represented by following expressions.

Pi,j = Probability of (St= j| St-1=i)

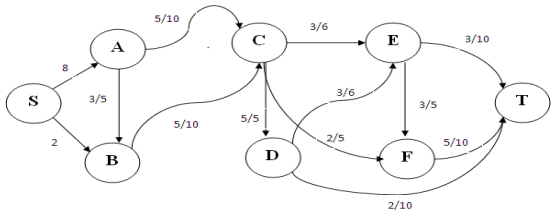

Figure7: Probability web Graph

Probability of hyperlink is based on the content of page being viewed. Matrix indicates navigation control can reach at total 10 times at T.

Session	Pages Visit
KECSESSION1(KS1)	C, B, A
KECSESSION2(KS2)	C, E, B, A, D
KECSESSION3(KS3)	D, E, B, A, E, D
KECSESSION4(KS4)	C, D, E, B, A
KECSESSION5(KS5)	A, D, B, E, D

Table 4: Session Table

	A	B	C	D	E	F	T
A	0	3/5	1/2	0	0	0	0
B	0	0	1/2	0	0	0	0
C	0	0	0	1	1/2	2/5	0
D	0	0	0	0	1/2	0	1/5
E	0	0	0	0	0	3/5	3/10
F	0	0	0	0	0	0	1/2
T	0	0	0	0	0	0	1

Table 5: Probability Table

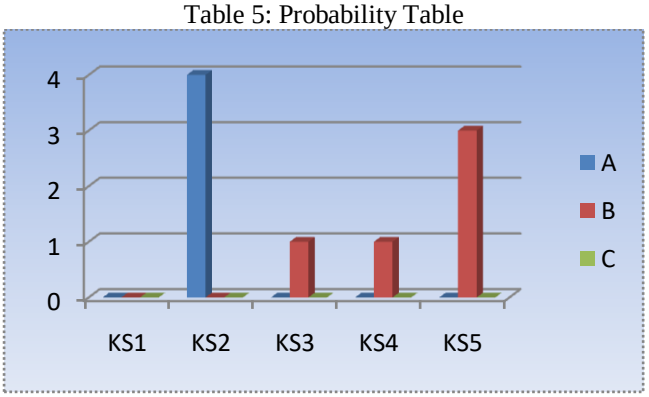

Figure 8: Page visit graph in a session

Table 6 indicates the accuracy and coverage of each page in a session and how much this is effective and help full in web user behavior analysis and web site modification.

Pages	Accuracy	Coverage
A	0.1	1
B	0.23	0.6
C	0.3	0.41
D	0.36	0.2
E	0.5	0.1

Table 6: Accuracy and Coverage

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

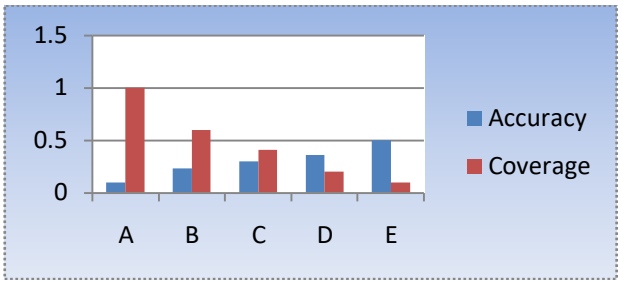

Figure 9: Accuracy and Coverage Graph

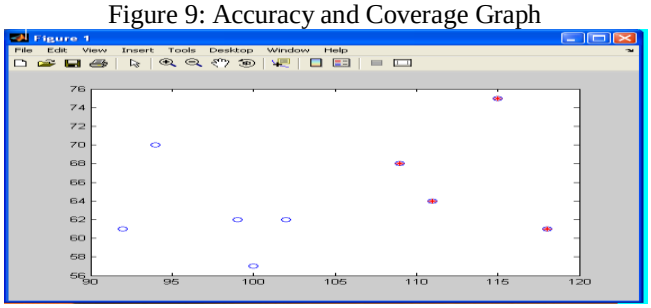

Figure 8: Students navigation cluster using FCM Clustering

VI. CONCLUSION

Web site modification and user behavior analysis can be done using various methods of neural network those are better in performance than other approaches due to efficient algorithm usage. Firing rules with hamming distance play a good role in identifying web usage patterns. FCM approach created many data clusters according to the input features. This paper emphasizes on dynamic model among the different processes of web usage mining. Thus clustering approach resulted in the formation of additional data clusters.

REFERENCES

[1] Ajith Abraham, "Business Intelligence from Web Usage Mining" Journal of Information & Knowledge Management, Vol. 2, No. 4 (2003) 375-390.
[2] Jos´e Borges, Mark Levene "An Average Linear Time Algorithm for Web Usage Mining" Sept 2003.
[3] Kumar Harish, Solanki A.K "Adaptive Markov Chain For Next Page Access Prediction" Vol. 9 No. 7 JUL 2011 Aug 25, 2011 by IJCSIS .
[4] Renáta Iváncsy, and Sándor Juhász, "Analysis of Web User Identification Methods" International Journal of Electrical and Computer Engineering 2:3 2007.
[5] . Suresh ,Madana ,A.RamaMohan Reddy, "Improved FCM algorithm for Clustering on Web Usage Mining" K IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 1, January 2011 ISSN (Online): 1694-0814
[6] Soumi Ghosh ,Sanjay Kumar Dubey "Comparative Analysis of K-Means and Fuzzy C-Means Algorithms" , International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013
[7] V Chitra, Dr. Antony "An Enhanced Clustering Technique for Web Usage Mining" ,International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 4, June - 2012
[8] Michal Munka, Martin `Drlíka Procedia "Impact of Different Pre-Processing Tasks on Effective Identification of Users' Behavioral Patterns in Web-based Educational System" Computer Science 4 (2011) 1640–1649 International Conference on Computational Science, ICCS 2011.
[9] Marquardt, Becker and D. Ruiz "A Pre-Processing Tool forWeb Usage Mining in the Distance Education Domain" . In Proceedings of the International Database Engineering and Applications, IDEAS 2004, pp. 78-87.
[10] Wang X., Abraham A. and Smith K.A.,, "Web Traffic Mining Using a Concurrent Neuro-Fuzzy Approach" In Proceedings Second International Conference on Hybrid Intelligent Systems, Chile, IOS Press Amsterdam, The Netherlands, 2002.
[11] Olfa Nasraoui and Christopher Petene. "Combining Web Usage Mining and Fuzzy Inference for Website Personalization"
[12] "Harish Kumar, Dr. Anil Kumar Solanki "Effective Cleaning of Educational website usage patterns and predicting their next visit,International journal of computer applications Volume 53 Nov 12.
[13] Archana N Boob " Fuzzy Clustering: An Approach for Mining Usage Profiles from Web" International Journal of Computer Technology & Applications, Vol 3 (1),329-331.
[14] Jaykumar, Kamlesh Patel, "A survey on web usage mining with neural network and proposed solutions on several issues" journal of information, knowledge and research in computer engineering, ISSN: 0975 – 6760 nov 12 to oct 13 volume – 02, issue – 02 page 330.
[15] Valishali A. Zilpe, Dr. Mohammad Atique "Neural network apprach for web usage mining", ,ETCSIT, published in IJCA [2011]

AUTHOR PROFILE

Prof A.K.Solanki has obtained his PhD in Computer Science and Engineering from Bundelkhand University.he has published good no of papers in national and International Journals.His area of specialization is Data mining.

Harish Kumar has completed his M.Tech (IT) from Guru Gobind Singh Indraprastha University, Delhi. Now currently pursuing his PhD from Mewar University, Chittorgarh .

Anuradha Taluja has completed her M.Tech from Guru Gobind Singh Indraprastha University, Delhi. Her area of specializati on is Data mining.