Journal of Web Engineering, Vol. 16, No.1&2 (2017) 126-144

AN AUTOMATED WEB PAGE CLASSIFIER AND AN ALGORITHM FOR THE EXTRACTION OF NAVIGATIONAL PATTERN FROM THE WEB DATA

ABDUL RAHAMAN WAHAB SAIT T. MEYYAPPAN

Research Scholar *Professor*

Dept. of Computer Science and Engineering

Alagappa University, India

rahamaan@gmail.com *meyyappant@alagappauniversity.ac.in*

Received January 18, 2016
Revised June 27, 2016

There is a demand for web intelligence in e-business and internet oriented markets. Many data crunching tools are available for the vendors to predict the customer behaviour on their website; still, there is a vacuum exist, and they fail to grab visitor attention on their products. Internet crimes are increasing exponentially with the growth of popularity of the internet. Web page classification (WPC) is a technique to classify the web page into a particular category by using its content and attributes like URL, Meta, and Title tags. Classification of web pages provides an option for an organization/ University to either block or allow a web page to the employees / students. Weblog pattern (WLP) mining is a favourite tool to extract useful patterns and deduce knowledge for the development of the website. The proposed work found the solutions for the extraction of WLP and WPC. The work has executed neural fitted Q-Iteration (NFQ) [1] method to classify Tamil and English web pages and extract the types of visitor visits the web page using a weblog. The experiment results show that there are an economic time and memory usage of the proposed method and improved percentage of accuracy comparing to existing methods.

Key words: *Web page classification, Browsing pattern, Neural fitted Q – Iteration, Weblog, Web mining, Machine learning, Reinforcement learning*
Communicated by: M. Gaedke & O. Diaz

1 Introduction

Semi – structured nature of web leads to further research to build up a structured way to access the internet and obtain knowledge for the development of business depending on the web [2]. The introduction of web mining in the field of computer science becomes a milestone for the enterprise people depends on the Internet [3]. The nature of web helps business people to develop their network across world. The introduction of new technology and development in the structure of internet made data analysis on web more complex. The various developments in the web mining made an interactive web for the users. Web mining has a variety of applications like web page optimization, fraudulent activity detection, link analysis, search engine optimization, social network analysis, WPC and WLP generation [4][5][6].

Figure 1 illustrates the goal of the proposed research and its activities on weblog and web page. Initially, web pages and weblog are collected and preprocessed for machine learning and NFQ algorithms to generate results and applied for the desired purpose.

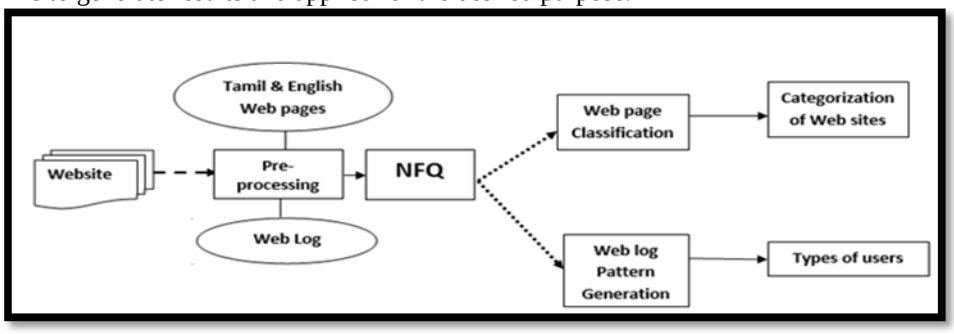

Fig. 1 NFQ as a WPC and WLP generator

WPC and web user pattern generation are the interesting application of web mining, useful for the web developer and owners to evaluate web pages and modify the way of presenting the information to grab users' attention towards them [7].

WPC used to identify the webpage using the content provided in it. The research uses n-gram method to generate keywords and compare the content of web page then classifies it. Web content filtering is the prime application of WPC and the research utilised NFQ with n-gram to classify web pages.

URL and content based WPC generates accurate results comparing to the other methods. The results can be used to recommend a particular site to web administrator to either block or allow in the network [8]. Applications of WPC are Search optimization, Answering system, Web content filtering, Web crawling and ontology annotation [9]. Email and Online Community websites can be classified using WPC but access control and privacy laws are restricted the boundaries of researchers.

NFQ used to cluster the type of users from the weblog to promote the site according to the users' interest. Wide varieties of techniques are available to study the log, and the performance of NFQ is more accurate than the existing methods.

The Web log mining is widely used for the development of a website and finds the traces of criminal activities on the website and finds the traces of criminal activities on the website. The server keeps the weblog; different formats are available, depends on the software service provider for the website [10].

The proposed research classify the Tamil and English websites into seven broad categories composed of Education, Entertainment, Sports, Blog, Porn, News, and Economy. WEBKB dataset used for the classification of web pages into Student, Faculty, Staff, Department, Course, and Project. The other part of research mine weblog and categorize the users into three different types: frequent, potential, and synthetic.

1.1 Frequent users

Users have the habit to visit the website in regular interval using the URL and interact with the site by posting queries and feedbacks. Usually, this kind of users spends time in the website for the purpose of transactions or fond of an attraction of information provided on the site [11].

1.2 Potential users

Users visit the page without any intention to do transaction / purpose. They visit the site without any referrals like google.com, yahoo.com, and facebook.com. There is a chance that this kind of users may become frequent users. Website owners will attract them to do more visit to the site by applying some changes on the site design / Information [11].

1.3 Synthetic users

Search engines use robots to index the websites and list according to the user search keywords. Keywords are some part of information about the website. Users visit the site through the search engine. This type of users are not direct users of the site but referred by search engines. There are some referrals found on the web. Search engines, Internet marketing, social networks, and Youtube are the social referrals in the web [11].

The proposed research identifies the different type of users using their navigational behaviour. All web pages will fall under the categories mentioned in the previous paragraphs. The extraction of patterns from the weblog depends on the behaviours of the three types of users.

The paper organized as follows. In section 2, the literature survey on web classification and WLP generation are introduced. Section 3 discusses the pre-processing of web pages and a weblog. Section 4 presents the proposed work for web classification and WLP generation based on NFQ. In section 5, the experiments on a pre-processed data set are reported. Section 6 concludes the research.

2 Related works on WPC and WLP generation

Many kinds of literature are available on automatic WPC and WLP mining. Web pages are classified based on URL and bag of words methods. Some literature combined both methods and showed some improvements comparing to the other methods. WLP mining is an application oriented technique that generates results according to the criteria for the sake of promotion of the website. The machine learning methods implement the WLP generation. The following part of the section discusses the existing literature related to the proposed research.

In [8], A Ph.D. thesis, WPC based on F- Neighbor algorithm. The information of neighboring pages used for the classification of web pages through the proposed algorithm. The research categorizes the web pages as a multi – label classification, multiple classes are assigned to an instance. F – Neighboring algorithm utilizes text on web pages and balances the different fields and captures the topics of web pages. The experiment on dataset shows more accuracy than the existing methods.

In [12], Support vector machine (SVM) and maximum entropy (ME) classifier with URL based feature extraction followed by the WPC. 3 – Grams were used to extract the words from the URL and stored in the feature vector with a maximum feature of 17576. SVM can handle the high number of non – linear data. Term frequency used to add as a feature of n- gram to the ME classifier to classify

the URL with more efficiency. The resulting analysis shows that the SVM accuracy is better than the ME classifier.

In [13], A URL-based approach proposed for WPC based on machine learning methods. The research used n-grams word formation to generate words related to the URL. Structure-oriented weight technique was followed to assign weightages to the words representing a web page. Hamming loss, One-error ranking loss, coverage and micro – averaged precision are the metrics used to evaluate the multi-label classification methods. The performance of neural network method is better than other existing methods.

In [14], a Random classifier method was employed to classify the web pages. The bootstrap sample of the dataset used to construct the tree in the classifier. Gini index used to choose the relevant attributes for the best split of the tree. Random forest is the multitude of decision tree. It has more efficiency than decision tree. The proposed method experimented with Yahoo web corpus and World Wide Knowledge Base (WEBKB) data set and generated better result comparing to the decision tree method.

In [15], a decision tree method based classifier used to generate attributes with co-occurrence analysis to classify web pages. Yahoo! Japan used for the purpose of training and the morphological analysis applied to split the sentences into part of speech as nouns, adjective and adverb. Basket analysis used to generate frequent itemset and association rules. Arts & Humanities, Business & Economy, Education, Government and Health were the top categories implemented for WPC.

In [16], a WPC based on web summarization using Bayesian Classifier and Support vector machine proposed to classify web pages. Luhn's summarization technique and latent semantic analysis used for web page summarization. Significance factor used to generate a words pool from the web page. The research experimented 2 millions web pages from the looksmart web directory. Stop-word remover and porter stemming were applied to generate tokens from the text extracted from the webpage.

F – neighbor algorithm is the old algorithm and does not have the ability to produce results in limited time.WPC needs more web pages to generate better results. The algorithm has scalability issue and not suitable for WPC.

In [17], a Birch algorithm was implemented to cluster the users of the websites based on the generalized session extracted from the weblog. The attribute – oriented induction method used to generalize the session from the weblog. In this study, a session represents a set of page / time pair. A client-side applet was implemented to calculate the total time spent by the user on the web page. The approach tested with the large dataset and produced optimum clusters of the web pages.

In [18], an Ontology-based pattern generation was proposed to extract the information and integrated with semantic information of the web page uses for web recommendation system. The user session converted into a sequence of objects for the process of k-means clustering method. Needleman – Wunch algorithm utilized for the comparison of sequences of objects. The experiment results show that the research is effective than the existing methods.

In [19], an ant-based clustering was implemented to extract frequent pattern for the weblog, University websites URL were classified into 28 elements and stored as a corpus, and an array was

implemented to hold the clusters formed from the weblog. Dissimilarity function performed by the proposed research between the corpus and the array. The final result is a pattern of any length for the next stage of web mining.

3 Pre-processing of web data

3.1 WPC

The pre-processing stage gather URL address, meta tag, title tag, heading tag and keywords used in the web page; repeated words are considered as a keyword and limited to some four per website [20]. The research classifies the web pages into 7 broader topics as a flat categorization of 158 Tamil and 572 English websites and 6 broader topics of 4 Universities web pages of WEBKB[21] dataset. Multiple labels collected from the website transformed into a numerical pattern and accessed to classify the website by a classifier according to the parameters. The pre-processing of web pages is shown in figure 2. The N – gram and keyword collection extracted features then transformed into accessible form for the proposed method and other machine learning methods.

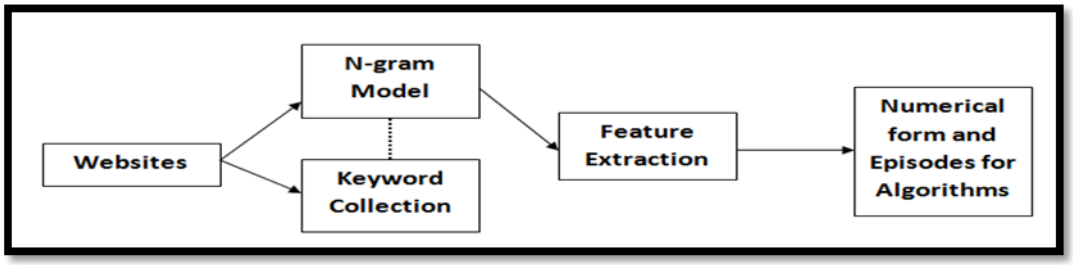

Fig. 2 Pre-processing for WPC

N-gram based approach is followed to gather the multiple labels from the web page. Bag of words does not yield optimum results. To classify the Tamil web page, frequently used words are collected in a separate file because the Tamil language has 247 letters. The research uses 2 –gram based approach to capturing the features from the web page. Tamil script with 2-grams generates more words resulting more time and space. Therefore, the frequently used keywords used to complete the 2 –gram into a full word and fulfill the process in a small amount of time. If 2 – gram from URL and title and meta tag generated "**அர**", "**ரஅ**" matches words like "**அரசு**" means for Government and "**அரங்கம்**" means for the theater, so the two words matched with the keyword collections for the website type "Entertainment" and "Government". The process iterated for sometimes to extract features from the website [22]. The same procedure followed in the English language and WEBKB Dataset. The keywords generated from the websites transformed into numerical form and episodes (Batch files) to train the Machine learning algorithms and NFQ to learn the classification of Tamil and English websites [23]. There is no relationship between training time and a number of web pages as the pre-processing work uses 2-gram technique to generate keywords from the web pages. 3-gram and 4-gram generate more number of combinations to generate keywords leads to the requirement of more memory and time, and there will be no improvement in the generation of keywords.

3.2 WLP Generation

Figure 3 describes the process of formatting the data for the methods discussed in the research. The identification of sessions and removing irrelevant data from the weblog are pre-processed and transformed into numerical form / episodes for NFQ and other methods.

$$F(x) = \int_{p1}^{p2} pdx + \int_{p1}^{p2} prdx$$

P1 and p2 are the specific periods and pdx is the past data and prdx is the present data, and F(x) is the future trend generated through the two functions. A Weblog is the log file stored on the server, keeps the users past and present data in formats like IIS (Internet information Services) and Apache. The research has followed the Apache web server format.

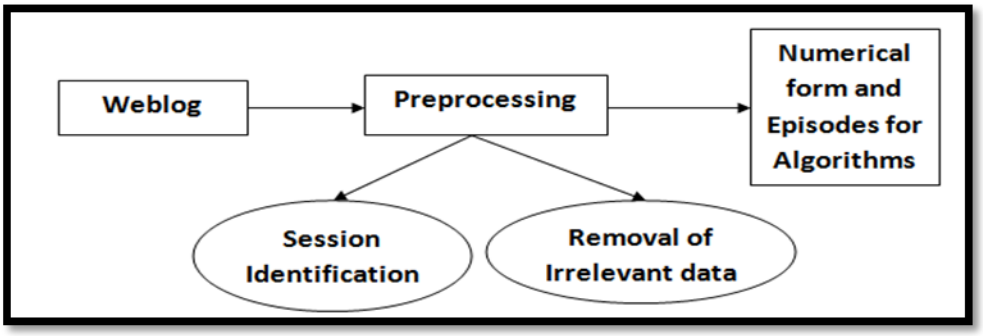

Fig. 3 Pre-processing for Web log pattern generation

Pre-processing of a weblog is the tedious process because the identification of session leads to the number of users. The weblogs collected for the period of seven days from Clarknet [24], ninety days from www.rahablog.com [25], ninety days from www.urakkapesu.com[26]and 150 days from www.ijcsit.org [27]. A 45 seconds set as a limit to identify a single user and unwanted files like images, favicon, and incomplete data removed from the dataset. The parameters like referrals, post, get, and head methods were added with the identified users. The pre-processed data transformed into numerical form and episodes for the generation of pattern.

4 NFQ

Reinforcement learning (RL) is the right option to attain a maximum precision. It is a slow learner comparing to the existing methods. It has the ability to take decision according to the environment and the process will be reinforced to attain the optimum level. The learning process of RL follows trail and error method and the rules are stochastic in nature. Q – Learning is the popular learning method in the RL techniques. [1] proposed a neural fitted Q – value for the adequate training for the model to reach the accuracy within the limited time. It is a neural network, multi – layer perceptron based method to implement reinforcement learning for a machine. It is possible to train the NFQ with a minimum amount of training samples [1] with the limited time to get the maximum accuracy. The following snippets show the NFQ for the WPC derived from [1]. The research used iRprop+,[28] a Resilient

back propagation with weight backtracking to train the method with the sample episodes (Batch files) and reduce the total time for the computation.

```
NFQ_WebClass_main()
{
Input: A preprocessed web pages W; output: Q-value function Q_N
K=0
Init_MLP()->Q_0;
Do{
Generate_category P = { (input^L,target^L), L=1, ….  …,#W} where:
Input^L = s^L,u^L
target= c(s^L,U^L,S'^L)+μmin_b Q_k(s',b)
iRprop^+_training(P) ->Q_k+1
k:=k+1
}While(K<N)
```

Fig. 4 NFQ Loop for WPC

iRprop$^+$ is an improved version of RProp [29][30][31] can heal itself from the worst situation by adjusting the weight using backtrack method. In figure 4 and 5, Q value – function for the multi – layer perceptron for the purpose of WPC and WLP generation is provided with the use of NFQ [8]. L is the training experience for the initial state s and target state s' and action a. l is incremented up to W episodes. c(sL,UL,S'L)is the condition for the target state s' and μminbQk(s',b)is the discounting factor adjusted by the Q-function during iteration of the method. Q0 is the initial state of Q – function and iterated through the loop.

```
NFQ_WebLog_main()
{
   Input: A preprocessed web log W; output: Q-value function QN
   K=0
Init_MLP()->Q_0;
Do{
Generate_category P = { (input^L,target^L), L=1, ….  …,#W} where:
Input^L = s^L,u^L
target= c(s^L,U^L,S'^L)+μmin_b Q_k(s',b)
iRprop^+_training(P) ->Q_k+1
k:=k+1
}While(K<N)
```

Fig. 5 NFQ Loop for Web Log Pattern Generation

NFQ for WPC and WLP mining trained with sample dataset for few time to attain the maximum accuracy. The cost setting of NFQ does not need previous knowledge of the environment and useful for the limitation of computation time and memory usage.

5 Experiment and Results

The experiments conducted in Intel© Core ™ i7- 5500u processor, 2.40 GHz, 6 GB RAM with Windows 10 operating system. The algorithm implemented in Java for machine learning methods, and NFQ in R-L Glue, Matlab, and Weka [32] applied to generate the results.

5.1 WPC

158 Tamil and 572 English web pages downloaded and classified manually into the following categories: Education, Entertainment, Sports, News, Porn, Blog, and Economy. A large dataset contains 8,282 classified University web pages from WEBKB also used to verify the efficiency of the proposed method with other machine learning methods. Table 1 (a),(b) and (c) shows the categories with the number of web pages classified manually for the evaluation of accuracy of the methods. Naïve Bayes(NB), J48, Support Vector Machine (SVM), and Random Forest (RF) are the algorithms compared with NFQ and trained with sample dataset derived from the pre-processed dataset discussed in the previous section.

NFQ training differs from other methods. The training of NFQ for downloaded web pages' corpus required a Q-value function of a multi – layer perceptron with 9 inputs having 2 state and 7 action variables and 3 hidden layers with 7 neurons each and 1 output neuron with a control interval 0.03 seconds. Actions are restricted between -7 and 7. The controller function modified the weights with the allocated interval. Initially 5 episodes or 172 cycles conducted for Education category and extended to 10 episodes or 253 cycles.

Category	Web pages
Education	24
Entertainment	27
sports	14
news	12
porn	19
blog	15
economy	29
other	18

Table 1 a. Categories of Tamil web pages

Category	Web pages
Education	72
Entertainment	78
sports	65
news	73
porn	61
blog	77
economy	72
other	74

Table 1 b. Categories of English Web pages

Category	Web pages
Student	1641
Faculty	1124
Staff	137
Department	182
Course	930
Project	504
Other	3764

Table 1 c. Categories of WEBKB Dataset

The process repeated for different categories and the computation time calculated using episodes and cycles and arranged with the computation time of other methods. The Table 2, 3, and 4 describe the computation time of Training phase of the methods. Figure 6 (a),(b), and (c) are represented as trend over time of methods during training phase. NFQ performance is better than the other methods; the reason is the implementation of the iRprop+ algorithm to accelerate the computation time of NFQ.

The NFQ training for WEBKB dataset required a Q – value function with 8 input (2 state and 6 action) variables and 2 hidden layer with 6 neurons each and 1 output neuron with a control interval 0.02 seconds. Actions are restricted between -6 and 6. The best policy found for the student category found with 14 episodes or 351 cycles.

Category/ Algorithm	*Edu*	*Enter.*	*Sports*	*News*	*Porn*	*Blog*	*Economy*
NB	0.078	0.094	0.130	0.272	0.279	0.739	0.439
J48	0.089	0.123	0.110	0.213	0.341	0.639	0.279
SVM	0.439	0.479	0.178	0.302	0.412	0.472	0.211
RF	0.094	0.148	0.147	0.248	0.361	0.578	0.310
NFQ	0.063	0.153	0.092	0.167	0.379	0.576	0.372

Table 2 Training time (in Seconds) of Tamil web pages

Category/ Algorithm	**Edu**	**Enter.**	**Sports**	**News**	**Porn**	**Blog**	**Economy**
NB	0.098	0.254	0.101	0.145	0.215	0.198	0.156
J48	0.135	0.289	0.356	0.156	0.325	0.318	0.356
SVM	0.415	0.456	0.118	0.185	0.457	0.278	0.189
RF	0.084	0.178	0.197	0.298	0.345	0.325	0.345
NFQ	0.078	0.179	0.114	0.189	0.112	0.312	0.418

Table 3 Traininig time (in Seconds) of English web pages

Category/ Algorithm	***Student***	***Faculty***	***Staff***	***Dept***	***Course***	***Project***
NB	0.325	0.341	0.195	0.135	0.129	0.164
J48	0.445	0.368	0.478	0.097	0.134	0.127
SVM	0.586	0.568	0.198	0.115	0.095	0.102
RF	0.205	0.254	0.235	0.201	0.125	0.095
NFQ	0.201	0.235	0.174	0.109	0.112	0.091

Table 4 Training time (in Seconds) of WEBKB Dataset

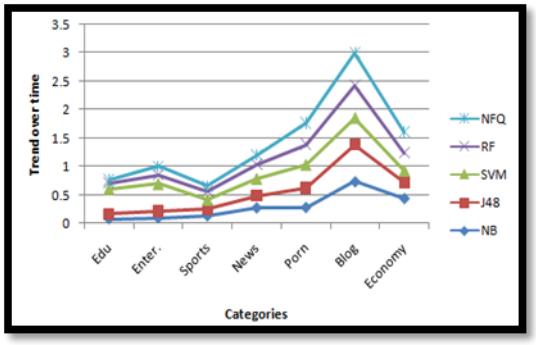

Fig. 6 a. Tamil Web page

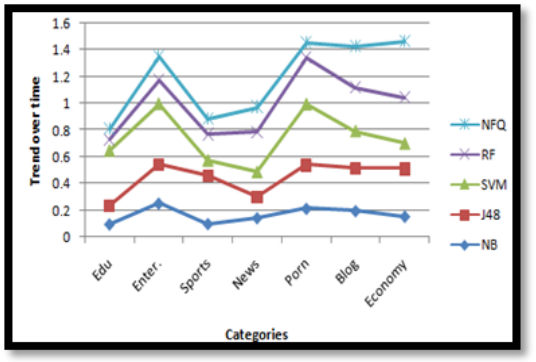

Fig. 6 b. English Web page

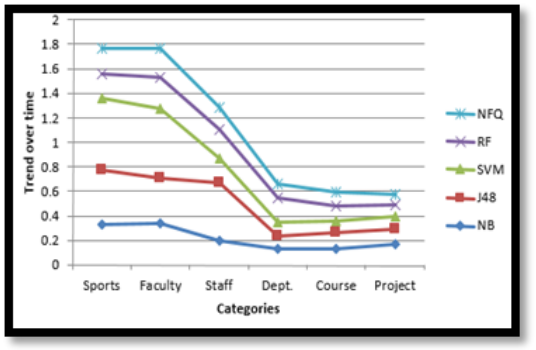

Fig. 6 c. WEBKB Dataset

The table 5, 6 and 7 summarizes the time required for testing of classification of web pages. The figure 7 (a), (b), and (c) show the trend over time of testing for each classification. NFQ took the short period to classify each web page, and overall performance is superior to other methods ad NB performance was very low comparing to the other methods.

Category/ Algorithm	***Edu***	***Enter.***	***Sports***	***News***	***Porn***	***Blog***	***Economy***
NB	0.018	0.093	0.012	0.129	0.094	0.149	0.142
J48	0.121	0.212	0.019	0.142	0.089	0.072	0.174
SVM	0.174	0.191	0.098	0.170	0.112	0.082	0.194
RF	0.079	0.094	0.101	0.109	0.170	0.092	0.214
NFQ	0.019	0.078	0.009	0.017	0.019	0.078	0.009

Table 5 Testing time (in Seconds) of Tamil web pages

Category/ Algorithm	**Edu**	**Enter.**	**Sports**	**News**	**Porn**	**Blog**	**Economy**
NB	0.008	0.095	0.108	0.112	0.101	0.198	0.098
J48	0.024	0.102	0.125	0.089	0.087	0.107	0.013
SVM	0.121	0.132	0.079	0.058	0.045	0.179	0.078
RF	0.045	0.102	0.018	0.023	0.078	0.163	0.015
NFQ	0.006	0.084	0.005	0.019	0.012	0.154	0.014

Table 6 Testing time (in Seconds) of English web pages

Category/ Algorithm	*Student*	*Faculty*	*Staff*	*Dept*	*Course*	*Project*
NB	0.012	0.018	0.027	0.006	0.027	0.025
J48	0.028	0.019	0.021	0.011	0.036	0.016
SVM	0.097	0.079	0.056	0.024	0.045	0.054
RF	0.025	0.019	0.024	0.009	0.014	0.016
NFQ	0.009	0.012	0.018	0.004	0.014	0.008

Table 7 Testing time (in Seconds) of WEBKB Dataset

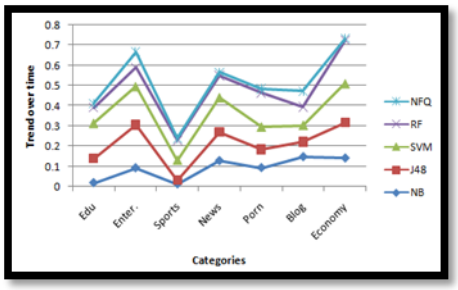

Fig. 7 a. Tamil Web pages

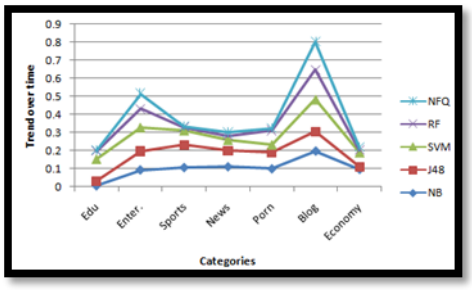

Fig. 7 b. English Web pages

Fig. 7 c. WEBKB Dataset

The table 8 describes the prediction accuracy for the Tamil web pages. Tamil language identification requires more training for all the methods. NFQ accuracy is better than the other methods, and the reason is that the method reinforced during the Training phase and the performance improved in the testing phase. The accuracy can be enhanced by increasing the number of episodes but the computation time will be increased, and the total cost will be more than the other methods. The figure 8 (a), (b) and (c) related to the prediction accuracy of the Tamil, English web pages and WEBKB Dataset.

Category /Algorithm	Edu	*Enter.*	*Sports*	*News*	*Porn*	*Blog*	*Economy*
NB	81.7	86	85	91.4	90.4	89.1	90.1
J48	83.1	87	86.2	90.8	89.4	91.3	94.4
SVM	79	83.2	79	87	75.3	89.4	85.3
RF	80	85	84	88	87	90.1	87.7
NFQ	*86.1*	*91.4*	*89.3*	*92*	*91.3*	*94.1*	*92.7*

Table 8 Prediction Accuracy of Tamil Web pages

Category /Algorithm	Edu	Enter.	Sports	News	Porn	Blog	Economy
NB	87.3	87.2	83.4	84.1	87.3	84.3	87.5
J48	90.1	89.2	87.4	84.3	86.4	88	90.1
SVM	78.9	87.3	79.2	80.4	84.6	79.5	82.3
RF	89.2	90.1	84.6	81.2	85.6	81.3	91.2
NFQ	*91.2*	*93.2*	*89.4*	*87*	*91*	*89*	*93.7*

Table 9 Prediction Accuracy of English Web pages

Category/ Algorithm	*Student*	*Faculty*	*Staff*	*Dept*	*Course*	*Project*
NB	91.4	89.2	90.5	87.6	89.6	89.3
J48	90.6	92.1	91.3	90.7	91.3	92.3
SVM	82.6	84.3	87.6	89.6	87.9	86.3
RF	91.5	92.3	91.3	92.3	91.9	94.3
NFQ	93.5	94.6	95.3	94.6	96.3	95.1

Table 10 Prediction Accuracy of WEBKB Dataset

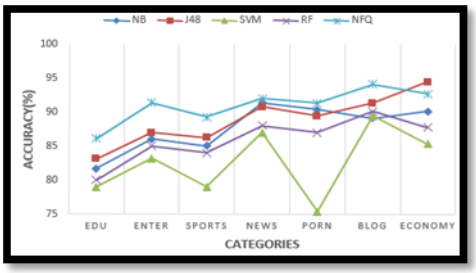

Fig. 8 a. Tamil Web pages

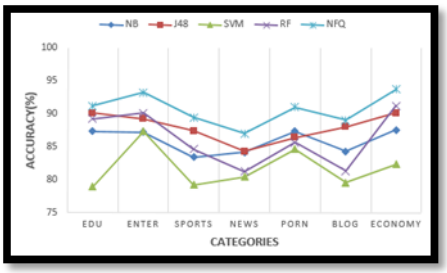

Fig. 8 b. English Web pages

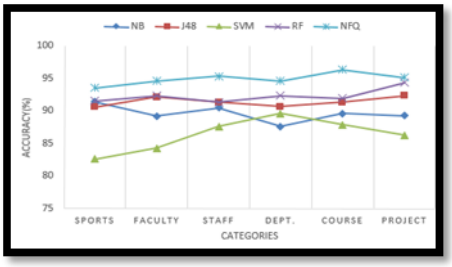

Fig. 8 c. WEBKB Dataset

The table 9 and 10 shows the prediction accuracy of English web pages and WEBKB Dataset. The accuracy of NFQ is superior to the other methods. The NFQ outperformed all other methods.

Algorithm /Web pages	**Memory Usage (MB)**				
	Hashmap				**NFQ**
	NB	**J48**	**SVM**	**RF**	
Tamil	0.64	0.534	0.72	0.67	0.52
English	12.1	15.21	14.5	15.3	12.2
WebKB	24.1	25.21	27.87	27.35	24.2

Table 11 Memory usage of Methods

Table 11 summarizes the memory usage of the methods during the Testing phase and NFQ took less data to classify the web page than the other methods. All other methods suffered from either memory, time and accuracy but NFQ has maintained the best level during the classification of web pages.

5.2 WLP Generation

Web traffic is the critical factor in the process of extraction of users from the weblog. If the page visited by some users, then there is a chance of generation of interesting pattern. The pre-processing

stage utilized web traffic to the number of users visited the page and filtered the user details and forwarded the data to the next stage of the research. The research used the dataset of web pages having enough web traffic for the comparison of existing and proposed methods time and memory complexity.

K- Means, BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), Ant Based Clustering are the methods compared with NFQ for the web log pattern generation. Table 12 summarizes the cluster of users extracted manually from the web logs for the evaluation of accuracy of the methods.

Users / Dataset	Frequent Users	Potential Users	Synthetic Users
Clarknet	1591	982	142
rahablog.com	90	43	38
urakkapesu.com	24	27	19
Ijcsit.org	175	96	47

Table 12 Total number of users in the Dataset

The training of NFQ for WLP generation required sample dataset made as the episode to cluster the users. Two actions are provided for controller -3 and 3. The interval of a controller is 0.02 second. The Q-value function iterated for each episode with sigmoidal activation function. The training phase has 14 episodes or 412 cycles for frequent users and the table 13 shows the overall time required for the training phase of the methods. NFQ computation time is calculated in the same format followed for the WPC. Figure 9 is the graphical representation of time taken by each method during the training phase for each clusters.

Users / Methods	Frequent Users	Potential Users	Synthetic Users
K – Means	0.342	0.412	0.287
BIRCH	0.186	0.192	0.215
Ant Based Clustering	0.192	0.196	0.198
NFQ	0.148	0.174	0.194

Table 13 Training Time(in seconds) of Methods

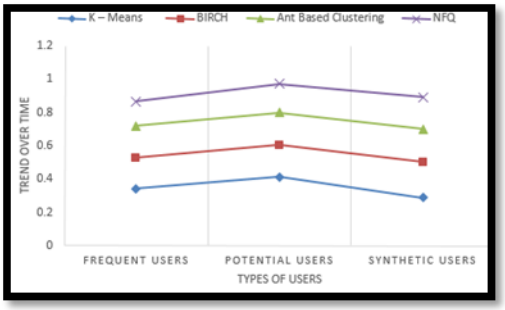

Fig. 9 Training Time

Table 14 and figure 10 shows the time required during the testing phase of the methods to cluster users and NFQ has taken limited amount of time comparing to other methods as it is reinforced during the training phase.

Users / Methods	Frequent Users	Potential Users	Synthetic Users
K – Means	0.114	0.178	0.143
BIRCH	0.097	0.104	0.118
Ant Based Clustering	0.084	0.094	0.104
NFQ	0.081	0.091	0.096

Table 14 Testing Time(in seconds) of Methods

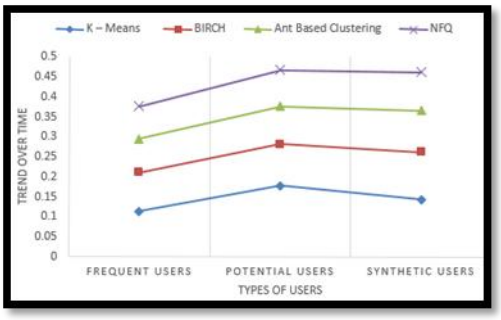

Fig. 10 Testing Time of Methods

Table 15 shows the memory usage of each method to cluster users from weblogs of websites used in the evaluation. The NFQ reinforced during the training phase to cluster the users so does not need more memory to generate clusters from a weblog.

Methods /Web pages	Memory Usage (MB)			
	Hashmap			NFQ
	K – Means	BIRCH	ANT BASED CLUSTERING	
CLARKNET	36.3	31.4	34.3	30.6
rahablog.com	17.2	14.8	16.2	17.4
urakkapesu.com	12.3	19.4	18.6	10.3
Ijcsit.org	11.1	12.5	14.8	8.9

Table 15 Memory usage of methods

Table 16 and Figure 11 shows the accuracy of each method and letters F, P, and S are frequent, potential and synthetic users of the website. NFQ accuracy is best among the methods used in the research of WLP generation. The accuracy can be improved to a maximum by increasing the number of episodes.

RL methods are slow learners and take more time to generate the accurate result. The proposed methods uses NFQ, an RL method with neural technique, implemented with iRprop+; a popular back-propagation method requires fewer samples and limited time to generate results.The structure of NFQ and ability to recover itself from the worst scenario and the pre-processing of a dataset are the factors of the proposed method to produce better results with limited time and space complexity comparing to the other methods.

Dataset / Methods	Clarknet			Rahablog.com			Urakkapesu.com			Ijcist.org		
	F	P	S	F	P	S	F	P	S	F	P	S
K – Means	89.4	91.1	85.2	84.1	90.7	91.6	86.1	91.5	87	92	84.2	90.6
BIRCH	94.6	93	94.2	81.3	92.8	90.5	91.3	90.8	89.5	91.2	94.2	91.5
Ant Based Clusterin	93.5	86.2	95.3	90.6	91.6	94.6	92.8	91.2	91.4	88.3	87.4	86.2
NFQ	97	96.5	97.4	92.4	93.7	95.1	93.5	95.6	92.6	93.4	96	95.4

Table 16 Prediction Accuracy of Methods

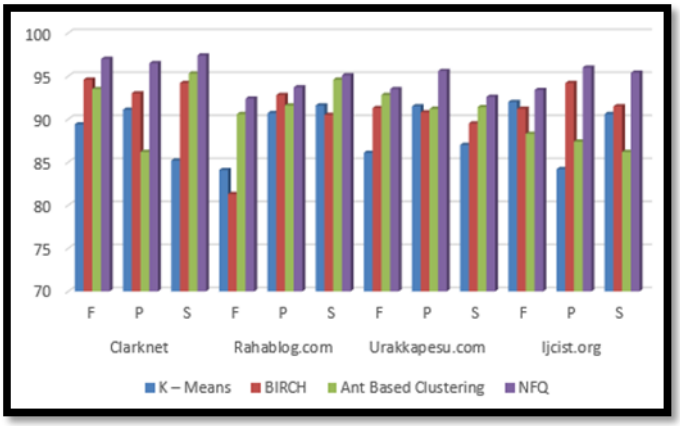

Fig. 11 Prediction Accuracy of Methods

6 Conclusions and Future Work

Web classification and the web log mining are the useful technologies to improve the standard of web-oriented services and businesses. NFQ is the RL method based on the neural technique to achieve optimum results in the limited amount of time. The pre-processing of web pages for WPC carried out by implementing 2 – grams method and pre-processing of weblog implemented by criteria based session identification methods. The proposed method based on NFQ performed multi –type flat categorization of web pages and clustered three types of users of a weblog. The performance of proposed research compared with existing methods and results shows better improvement in time and

space complexity. NFQ accuracy far better than other methods and further improved by increasing the number of episodes and cycles of Q-value function. The future scope of the research is to implement a multi – type, multi – label categorization of web pages and to predict the trend of capital market.

References

1. Martin Riedmiller, "Neural Fitted Q Iteration – First experiences with a data efficient neural reinforcement learning method", ECML 2005, Volume 3720 of the series, Lecture note in computer science, pp. 317 – 328.
2. Ganesan S, Sivaneri A.I.U, and Selvaraju S.," Evolving interest-based user groups using PSO algorithm", International conference on recent trends in information Technology, 2014, pp. 1 – 6.
3. SuhasiniParvatikar and Bharti Joshi, " Analysis of user behavior through the web usage mining", IJCA Proceedings on International conference on advances in science and technology, IJCAST 2014 (3), Feb 2015, pp. 27 – 31.
4. E.Baykon, M.Henzinger, L.Marian, and I.Weber, " Purely URL – based topic classification. In proceedings of the 18th International conference on World Wide Web, pp. 1109 – 1110, New York, USA, 2009, ACM.
5. KobraEtminani, Mohammad – R. Akbarzadeh – T, and NooraliRaeejiyahehsari, " Web usage mining: users' navigational patterns extraction from weblogs using Ant – based clusters method", IFSA – EUSFLAT 2009, pp. 396 – 401.
6. C.Castillo and B.D.Davison, "Adversial web search, foundations and trends in Information Retrieval", 4(5), pp. 377 – 486, 2010.
7. P.N.Bennett and N.Nguyen, " Refined experts: improving classification in large taxonomies", In proceedings of the 32nd International ACM SIGR conference on research and development in information retrieval, pp. 11 – 18, ACM, 2009.
8. XG.Qi, "Web page classification and hierarchy adaptation", Ph.D. Thesis, Lehigh University, January 2012.
9. E.Baykon, M.Henzinger, L.Marian, and I.Weber, "A comprehensive study of features and algorithms for URL – based topic classification", ACM transactions on the web, pp. 5:15:1 – 15:29, July 2011.
10. SaelN.,MarkA.,andBehza H., "The Web usage mining data preprocessing and multi-level analysis on Moodle", International conference on computer systems and Applications (AICCSA), IEEE, ACS, 27 – 30 May 2013, pp 1 – 7.
11. Abdul rahaman and Dr.T.Meyappan,"Data processing and transformation technique to generate pattern from the web log", International conference on computer science and Information Systems, Oct 17 – 18,2014 Dubai(UAE), pp. 6 – 9
12. R.Rajalakshmi and Chandrabose Aravindan, "Web page classification using n – gram based URL features", 2013, 5th International Conference on Advance Computing, IEEE, pp. 15 – 21.
13. ChakerJebari, "A pure URL – based Genre classification of web pages", 25th International workshop on database and expert systems applications, 2014, pp.233 – 237.
14. Win Thanda Aung and Khin hay mar saw hla, " Random forest classifier for multi – category classification of web pages," IEEE Asia – pacific services computing conference, 7 – 11 Dec 2009, pp. 372 – 376.

15. Makoto Tsukada, Takashi Washio, Hiroshi Motoda Automatic web-page classification by using Machine Learning Methods, Web Intelligence: Research and Development, Volume 2198 of the series, Lecture Notes in Computer Science pp 303 – 313, 2001.
16. Dou Shen, Zheng Chen, Qiang Yang, Hua – Jun, Zeng Benyu Zhang, Yuchang Lu, Wei – Ying Ma ,Web – page classification through summarization, , Copyright 2004, ACM.
17. Fu, Y., Sadhu.K, and Shin M.Y.,"Clustering of web users based on access patterns", In. Proceedings of the 5th ACM SIGKDD, 1999, International conference on knowledge discovery and data mining, Springer, San Diego.
18. S.Haken Yilmaz and Pinar senkul, " Using ontology and sequence information for extracting behavior patterns from web navigation log", IEEE International conference on Data mining workshops – 2010, pp. 549 – 556.
19. Sameendra samarawickrama and Lakshmanjayaratne," Effect of named entities in web page classification", Fourth International Conference on Computational intelligence, modeling and simulation, 2012, pp. 38 – 42.
20. Sudheer Reddy, Kantha Reddy M, and SitaramuluV.,"An effective data preprocessing method for web usage mining", International conference on information communication and embedded systems, 21 – 22 Feb 2013, pp. 7 – 10.
21. http://www.cs.cmu.edu/~webkb/
22. Lin Kewen,"Analysis of preprocessing methods for web usage data", International Conference on Measurement, Information, and control (MIC), 18 – 20 May 2012, pp. 383 – 386.
23. GoongWei,"A new path filling method on data preprocessing in web mining", International conference on control engineering and communication technology (ICCECT), 7 – 9 Dec 2012, pp. 1033 – 1035.
24. ftp://ita.ee.lbl.gov/html/contrib/clarknet-http.html.
25. http://www.rahablog.com

26. http://www.urakkapesu.com
27. http://www.ijcsit.org
28. Igel C. and M.H Sken," Empirical evaluation of the improved RPROP learning algorithms", NeuroComputing, No. 50, pp. 105 – 123, 2003.
29. Riedmiller M., "RPROP – Description and Implementation details", Technical Report, Jan – 1994, University of Karlsruhe.
30. Riedmiller M and Braun H," A direct adaptive method for faster back – propagation learning: The RPROP algorithm", Proceedings of International conference on neural networks, pp. 586 – 591.
31. Anastasiadis A.D., Magoulas G.D., and VrahatisM.N.,"An efficient improvement of the RPROP algorithm", Proceedings of the first International Workshop on Artificial neural networks in pattern recognition, 2003.
32. http://www.cs.waikato.ac.nz/~ml/weka/