Journal of Theoretical and Applied Information Technology
30th November 2013. Vol. 57 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

MULTI LAYER PERCEPTRON FOR WEB PAGE CLASSIFICATION BASED ON TDF/IDF ONTOLOGY BASED FEATURES AND GENETIC ALGORITHMS

N.VANJULAVALLI[1], DR.A.KOVALAN[2]

1. Research Scholar, Department of Computer Science and Applications, PMU, Vallam, Thanjavur.

2. Assistant Professor (S.S) Department of Computer Science and Applications, PMU, Vallam, Thanjavur

ABSTRACT

Now a day, Millions of web servers are available to give a huge amount of electronic content to the end users using the internet. Searching the content relevant to the need of a user is a challenging task because of the ambiguity in our natural language. Classification of web pages based on their contents is useful to the search engines to give appropriate and desired data to the user. In this paper, an optimized approach is used for classifications of web pages. Feature extraction and selection of best features play a key role in classification. In this work, features are extracted from the ontology representation of content and Inverse Document Frequency (IDF). Then the best features are selected by using genetic algorithm. Using the selected features by GA, an Artificial Neural Network (ANN) is trained to classify the web pages. Results were compared with the classification methods based on neural networks with IDF based feature extraction, neural networks with ontology based feature extraction, neural networks with combined IDF and ontology based feature selection and other three methods are based on applying genetic algorithm to select the best features to classify the web pages. The parameters such as a percentage of accuracy, precision, recall and root mean square error are considered for performance evaluation. Numerical results showed that hybrid classifier trained by multilayer neural network with GA for selecting IDF and ontology based features gave 93% of accuracy, high precision and recall and lowest RMSE when comparing to all other methods.

Keywords: *Web page classification, Inverse Document Feature, Ontology, Neural network classifier, Genetic Algorithms*

1. INTRODUCTION

Web applications are useful to share the general and specific information and to do many business activities globally, usingthe internet. The most important component of these applications is the web pages. A web page is created by using standard languages such as Hypertext Mark-up Language (HTML) formatting tags and Extensible Mark-up Language (XML) formatting tags. A web page may contain text, multimedia content, links to other web pages and graphical pages. There are two types of web pages. They are static and dynamic. Static web pages have contents that remain unchanged. But dynamic web pages display the contents based on some input values given by the user. To create a successful web site, the required information is gathered, contents are organized and published by several connected pages and published to the identified audience.

When retrieving information from the web sites, finding only the relevant content as per the need of the search is essential. Several search engines are available to search the content by giving queries or by giving keywords. There are two types of search engines. They are directory style and robot style. Yahoo, INSIZE and JAPAN are some of the examples of directory style search engines. AltaVista and Google are the search engines based on the robot-style. Directory style search engine uses lots of human power to categorize the web pages. Robot style search engines search the web pages based on the keywords without checking the contents, because most of the web pages are assigned with a set of keywords at the time of publishing in on line [1].

In natural language, many words are ambiguous giving different meaning based on the context and situation. Therefore, development of web directories, classification of web pages and analysis of topic-specific search are useful. Classification of contents makes an important part of most of the content management and retrieval activities [2]. Generally, if the number of classes is two, it is called as a binary classification; otherwise it is called as multiclass-classification.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Assignment of classes can be of two types: hard classification and soft classification. Based on the organization of categories, there are two types of web page classification. They are flat classification and hierarchical classification. Flat classification considers all the categories as equal and parallel, but hierarchical classification maintains the categories in a tree-type structure and one category may have many sub types.

The rapid growth of contents of the World Wide Web (WWW) needs an automated assistance for the classification of web pages to help the users of web [3]. These classifications help the web servers to create and maintain the contents in organized way and are also helpful to the web search engines which are based on the keywords search. Many researches, about web page search, concluded that users have a preference of navigating using the list of already classified contents. Therefore, automated classification of web pages is a labour-intensive task. Even though, the web page content is designed by standard HTML tags, pure text extraction and classification are major difficult tasks because of noisy data by some advertisement banners.

Extraction of features is a most important step in any classification model. For effective classification, the extracted features should give valuable information about the categories, and it should be inexpensive in terms of computation [4, 5]. Some of the features used for web page classification are textual, image and other information. In the personal web pages, the amount of text is very sparse. Other information includes audio, video and multimedia contents.

Tree banking Decisions Feature (TDF) is based on a set of candidate trees which is created by the grammar of the language and disambiguation is done by annotators [6, 7]. This reduces the man power to create the tree and a better annotation is built. TDF is used to improve the annotation by humans and to evaluate the difference between individual's analyses. After creating the n number of trees, sentences are applied to all the trees to make the decision about ambiguities. Usually n candidate trees produce approximately log(n) decisions for a given sentence. The decisions are similar to the accurate judgments by the human annotator who created the decision trees.

Ontology was introduced in the field of Artificial Intelligence (AI) during 1990's for the proper representation of the real things in a program code. Recently ontology is used in information retrieval, management of information, knowledge management and information integration systems [8]. Ontology has a structure to maintain the relationships and constrains among the different concepts. The approaches used for ontology permit the reuse and sharing in a computational form from the different knowledge bodies. The computation form is an agreement created cooperatively by different people in different places. Ontology approaches are used to extract the group of features and the group can be divided into sub categories. For example, the extracted features can be divided into qualitative and quantitative features.

IDF stands for inverse document frequency used in most of the IR systems. IDF is calculated by – $\log_2 dfw/D$, where D represents the number of documents in the collection and dfwis the frequency of document. Therefore, dfw represents the number of documents that have the word w [9].

Genetic algorithms are stochastic search algorithms used for solving NP-hard problems. GA uses a type of natural selection with operations inspired from the genetics operations [10]. The operations of GA are selection, cross over, mutation and reproduction. Selection is used to select the chromosomes available in the population for reproduction. Cross over selects two chromosomes to provide a new combination and mutation alters some parts within the chromosomes. Every iteration uses fitness function to evaluate the solution. The number of iterations are performed till the fitness function is satisfied, or the time for selection of features is expired.

In this work, features are extracted from the ontology representation of the web contents with inverse document frequency. Then best features are selected by using genetic algorithm. Using the selected features by GA, an Artificial Neural Network (ANN) is trained to classify the web pages.

2. RELATED WORK

Myo Myo Than Naing presented Ontology-Based Web Query Classification for Research Paper Searching [11]. To get the desired content from the web pages, search engines were used. The user could use the search engine without giving a request in a fixed format query.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Classification of user's request into a set of pre-determined categories was useful to improve the relevant data retrieval. A Query Classification algorithm (QCA) was developed to classify the user queries into categories, and if the request belonged to more than one category the categories were ranked. Ontology of the domain was used to use the vocabulary. Using the vocabulary of top history library systems ontology model was developed for the query classification system. This ontology based approach produced better results for subject specific search to the users.

S. Lovely Rose and K.R. Chandran presented Normalized Web Distance Based Web Query Classification [12]. This approach categorized the user's queries or search request into an intermediate set of categories. Then features were extracted using feature selection algorithms. Features were extracted from the available search engines for the given retrieval request. Normalized web distance was used for mapping the categories of features into the target categories. When a feature set was mapped to more than one target category, then the target categories were mapped. For ranking, the parameters such as name, frequency and combination of these features were used. Experiments were conducted with a result of 40 pages. Numerical results proved that the precision and recall was increased by 10 % than existing methods.

J. Alamelu Mangai et al presented A Novel Feature Selection Framework for Automatic Web Page Classification [13]. Most of the classification algorithm's performance depended on the elimination of the noisy and outlier data. For pre-processing, the content of the web page were converted into a text file, and textual features were extracted. Redundant or irrelevant attributes used for classification decreased the accuracy of classification. Therefore, best features should be used for the classification. Selection of features also reduced the dimensionality, number of resources and time required for the classification. An entropy measure called as ward entropy was used to calculate the most relevant features for web page classification. Artificial neural networks were used for training the classifier with selected features, and remaining features were eliminated. This method improved performance by reducing the learning rate.

Mangai et al presented A Novel Approach for Automatic Web Page Classification using Feature Intervals [14]. Automatic web page classification algorithm extracted the features from the web pages, and each feature was discredited using standard algorithms. Then the features were ranked by using the measure of information gain with respect to the label of the class. Based on the ranking, initial weights are assigned to the features. Then using the training data and the weighed features neural network was trained. For experiments the computer science and engineering department details were taken from top universities, the punctuation symbols and tabs were removed from the HTML tags. After extracting and ranking features, the classifiers were trained to categorize the instances into staff, students and research scholars. Performance achieved was better than existing classifiers.

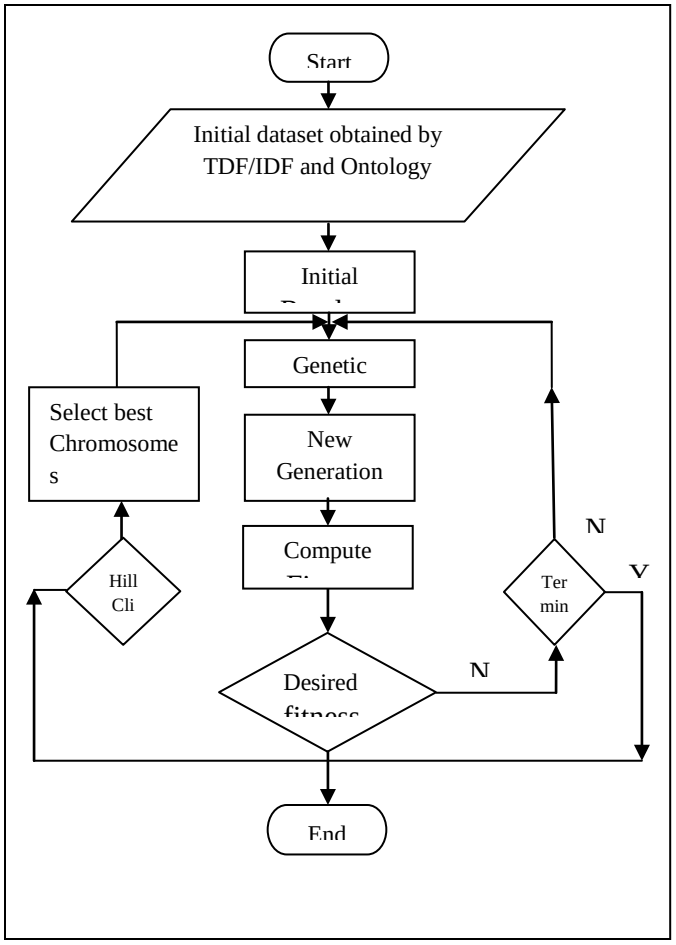
Hernández et al presented a statistical approach to URL-based web page clustering [15].Distance based measures were used in the traditional classification algorithms. Two URLs might have a small distance, but the content available in the two URLs was different concepts. Applying distance measures to the URL did not provide enough accuracy in the web page classification. In this approach probabilistic based approach was used instead of using the pattern set for the classification. Feature vectors were created for each token given in the URL and classifier algorithm was similar to the token bucket to classify the web pages. But results showed improvement in accuracy when compared to traditional distance based measures using URLs for classification of web pages.

3. MATERIALS AND METHODS

Some of the parameters used for web page URL based classification are baseline, URI components, length, precedence, orthographic and Sequential features [16]. This method does not use the available contents in the web pages. Based on the contents available some of features used are IDF, stemming and stopping words. Inverse document frequency is the most useful feature used for web page classification. IDF is high for infrequent words and IDF is low for frequent words. Ontology is used to represent the objects or concepts and their relationships using a shared vocabulary and taxonomy. In this work, ontology based representation is used for maintaining the semantics of the web content. IDF and TDF were used for feature extraction. After extracting large number of features, the dimensionality of the features is reduced by genetic algorithm for an

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

efficient classification. From the initial population, many numbers of iterations are used to select the best combination of features using hill climbing algorithm. At the end of iteration, fitness function was evaluated to stop the feature selection process. The architecture diagram of the proposed approach is given in the following figure 1.



$$Precision = \frac{|\{relevant\ documents\} \cup \{retrieved\ documents\}|}{|retrieved\ documents|}$$

Recall is the fraction of relevant documents retrieved successfully. Recall is calculated by the following formula.

$$Recall = \frac{|\{relevant\ documents\} \cup \{retrieved\ documents\}|}{|relevant\ documents|}$$

To evaluate the classification results, the measures such as true positive, true negative, false positive and false negative are used with external's judgments [17]. Based on these measures the precision, recall and average are calculated by the following formulas.

$$Precision = \frac{True\ positive}{True\ positive + False\ positive}$$

$$Precision = \frac{True\ positive}{True\ positive + False\ negative}$$

$$Classification\ Accuracy = \frac{True\ positive + True\ negative}{True\ positive + True\ negative + False\ positive + false\ negative}$$

Experiments and Results

For conducting the experiments, the details from the department of computer science in a US University are collected. From the collected data, the stopping characters and the punctuation symbols are removed. Courses are classified into under graduation and graduation courses. People are classified into students, staffs and faculties. Figure 2 shows the sample data.

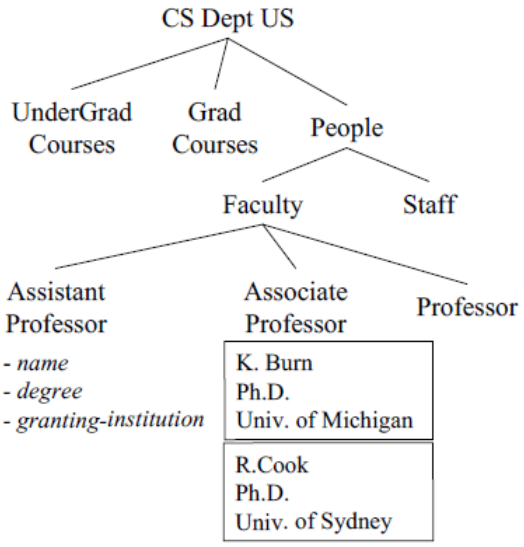


Figure 2: Sample Data

The parameters such as a percentage of accuracy, precision, recall and root mean square error are considered for performance evaluation. Table 1 gives the value of these parameters for the various classification methods using the same data set. Figure 3 to 6 shows the results graphically.

Journal of Theoretical and Applied Information Technology
30th November 2013. Vol. 57 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1: Results Of Classification Methods

Method	Classification accuracy	Precision	Recall	RMSE
MLP NN with IDF based feature extraction	76	0.783	0.76	0.3129
MLP NN with IDF and ontology based feature extraction	86	0.863	0.86	0.25
MLP NN with IDF based feature extraction and GA based feature selection	87	0.882	0.87	0.2286
MLP NN with IDF and ontology based feature extraction and GA based feature selection	88	0.894	0.88	0.2111
MLP NN with IDF based feature extraction and Hybrid GA based feature selection	86	0.863	0.86	0.25
MLP NN with IDF and ontology based feature extraction and Hybrid GA based feature selection	93	0.938	0.93	0.1758

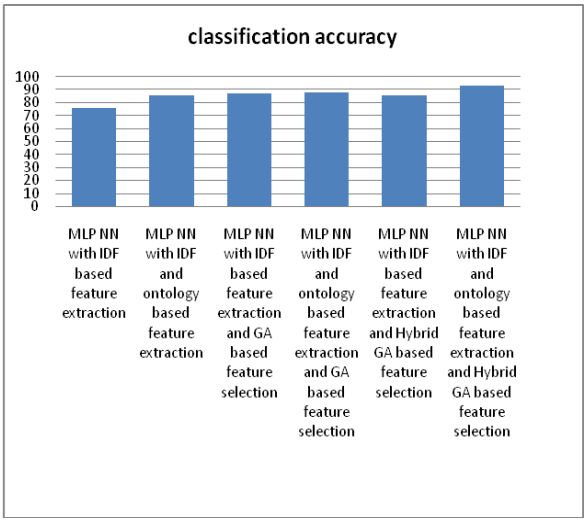


Figure 3: Classification accuracy

Figure 3 shows the classification accuracy of methods for neural network classifiers based on IDF, ontology based feature extraction and genetic algorithm based feature selection. Bar chart in figure 3 shows that IDF and ontology based feature extraction and hybrid GA based feature selection gives highest accuracy 93 % when comparing to all other methods.

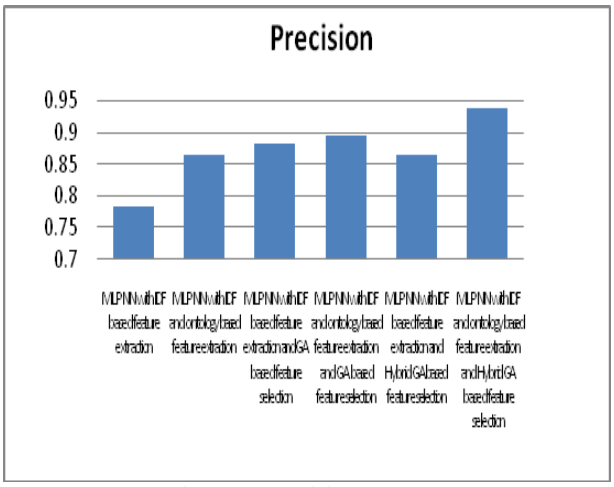


Figure 4: Precision

Figure 4 shows the precision value of methods for neural network classifiers based on IDF, ontology based feature extraction and genetic algorithm based feature selection. Bar chart in figure 4 shows that IDF and ontology based feature extraction and hybrid GA based feature selection gives highest precision value when comparing to all other methods.

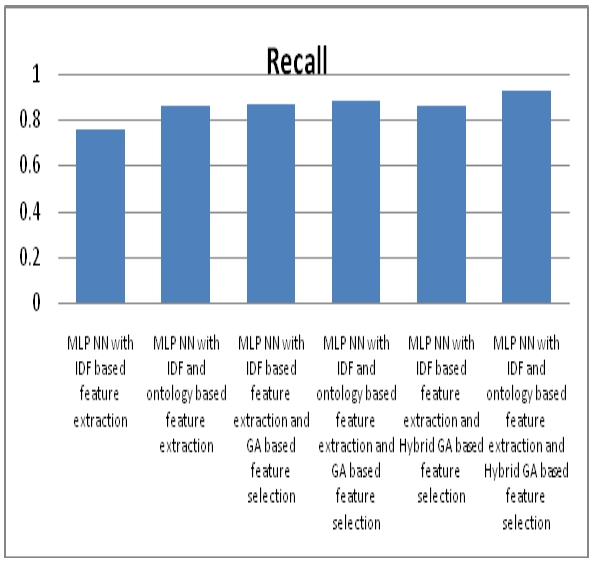


Figure 5: Recall

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 5 shows the recall value of methods for neural network classifiers based on IDF, ontology based feature extraction and genetic algorithm based feature selection. Bar chart in figure 5 shows that IDF and ontology based feature extraction and hybrid GA based feature selection gives highest recall value when comparing to all other methods.

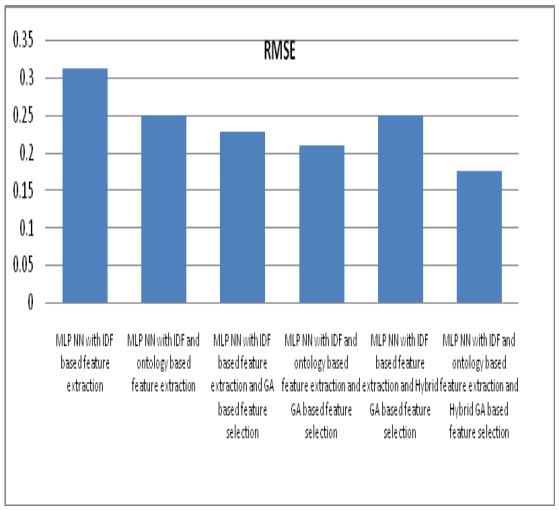


Figure 6: Root Mean Square Error (RMSE)

Figure 6 shows the RMSE value of methods for neural network classifiers based on IDF, ontology based feature extraction and genetic algorithm based feature selection. Mean squared error is used to compute squared error between the actual and classified vectors. Bar chart in figure 6 shows that IDF and ontology based feature extraction and hybrid GA based feature selection gives lowest root mean square error value when comparing to all other methods.

4. CONCLUSION

Classification of web pages based on their contents is useful to the search engines to give appropriate data to the user. In this work, features are extracted from the ontology representation of the content available in web pages and Inverse Document Frequency (IDF). Then the best features are selected by using genetic algorithm. Using the selected features by GA, an Artificial Neural Network (ANN) is trained to classify the web pages. For conducting the experiments, the details from the department of computer science in an US University are collected. From the collected data, the stopping characters and the punctuation symbols are removed. The data instance is classified into courses and people. Courses are classified into under graduation and graduation courses. People are classified into students, staffs and faculties. The parameters such as a percentage of accuracy, precision, recall and root mean square error are considered for performance evaluation. Numerical results showed that hybrid classifier trained by multilayer neural network with GA for selecting IDF and ontology based features gave 93% of accuracy, high precision and recall and lowest RMSE when comparing to all other methods.

REFERENCES:

[1]Sharon Vennix, "An introduction to website development for course web pages", MICHIGAN STATE UNIVERSITY, 1998.

[2]. Makoto Tsukada, Takashi Washio, Hiroshi Motoda, "Automatic Web-Page Classification by Using Machine Learning Methods", Institute of Scientific and Industrial Research.

[3]Qi, X., & Davison, B. D, "Web page classification: Features and algorithms", Department of Computer Science & Engineering, Lehigh University, 2007.

[4]Shen, D., Chen, Z., Yang, Q., Zeng, H. J., Zhang, B., Lu, Y., & Ma, W. Y, " Web-page classification through summarization" In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 242-249). ACM, July 2004.

[5] Asirvatham, A. P., & Ravi, K. K," Web page classification based on document structure", In IEEE National Convention, (2001, December).

[6] Chowdhury, M. F. M., Zhang, Y., & Kordoni. V"Using tree banking discriminates as parse disambiguation features", In Proceedings of the 11th International Conference on Parsing Technologies (pp. 226-229) Association for Computational Linguistics, (2009, October).

[7] Chowdhury, M. F. M , "Exploiting Treebanking Decisions for Parse Disambiguation " 2009.

[8] "Ontology The specification of a shared conceptualization", 2002,

[9]Kenneth W. Church and William Gale, "Inverse Document Frequency (IDF):A Measure of Deviations from Poisson ", 1995.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[10] Melanie, M "An introduction to genetic algorithms", Cambridge, Massachusetts London, England, Fifth printing,1999.

[11] Myo Myo than Naing, "Ontology-Based Web Query Classification for Research Paper Searching ", International Journal of Innovations in Engineering and Technology (IJIET), 2013.

[12] S. Lovely Rose, K.R. Chandran," Normalized Web Distance Based Web Query Classification", Journal of Computer Science 8 (5): 804-808, 2012

[13] J. Alamelu Mangai V, Santhosh Kumar S, Appavu alias Balamurugan, "A Novel Feature Selection Framework for Automatic Web Page Classification", International Journal of Automation and Computing, August, 2012.

[14] Mangai, J. A., Kothari, D. D., & Kumar, V. S. " A Novel Approach for Automatic Web Page Classification using Feature Intervals ", 2012.

[15]Hernández, I., Rivero, C. R., Ruiz, D., & Corchuelo, R, " A statistical approach to the URL-based web page clustering ",In the Proceedings of the 21st international conference companion on World Wide Web (pp. 525-526), ACM, April, 2012.

[16] Mahadevan, I., Karuppasamy, S., & Ramasamy, R,"Resource Optimization in Automatic web page classification using integrated feature selection and machine learning", International Arab Journal of e-technology, 2008.

[17] Kan, M. Y., & Thi, H. O. N, " Fast webpage classification using URL features" In the Proceedings of the 14th ACM international conference on Information and knowledge management (pp. 325-326). ACM, October, 2005.

[18] www.wikipedia.org.