Journal of Engineering and Applied Sciences 11 (8): 1859-1865, 2016
ISSN: 1816-949X

Features Selection from Data in Order to Improve Classification Methods Performance

Reyhaneh Khademi and Mahdi Afzali
Department of Computer, Islamic Azad University, Buinzahra Branch,Tehran, Iran

Abstract: Web pages classification is one of the main and challenging subjects in the field of data mining. Web page classification knowledge helps users to obtain useful information from massive data sets on the Internet automatically and efficiently. Many efforts have been made by researchers for web page classification, however, there is still opportunity to improve current approaches. Source of one of the main challenges in the educational categories is that the current data set is unbalanced. Because the size of pages in one subject is not the same with the other subject and its distribution is not uniform. Standard machine learning algorithms are influenced by main and big classes (groups) and secondary groups are ignored so accuracy standard for grouping is reduced. In this research, for solving this problem and for grouping web page a new approach based on collective grouping of support vector machine is proposed. To reduce and select features, principal components analysis and independent component analysis tools have been used respectively. Results show that proposed methods in better than other methods (which are widely used on web pages categories).

Key words: Web pages categories, support vector machine, collective categories, accuracy, approach

INTRODUCTION

With the increasing development of the world wide web, online text data size development has become rapid and significant in recent years. Due to information explosion and big data in the network, data recovery is facing with very serious challenges. So useful information extraction ways from raw data with high-size on the internet is now more important. Researchers have studied a lot on web mining with various data on world wide web. Researches on different fields like web mining data extraction, opinion mining, usage mining, data integration, social networks analysis have been done. Main focus is on searching engines and web directories. in fact every field classification including web page classification is one methods of subject organization. Usually classification between studying components is done according to some rules (such as: hide and obvious similarities). existing pattern extraction is a complex process, because these patterns are hide and cannot be seen obviously. So machine learning algorithms for classification is required. This need has led to extensive research on web pages classification technology. Web pages classification, may face unorganized data in web. The goal of web pages classification is classifying web pages on the internet in template of specific and predefined classes. Number of web pages that has become available for users by search engines has increased exponentially. Organizing web's big data needs an accurate and well-order way for using vital information sources. one of the approaches is web pages classification. Web pages classification using supervised learning tool, devotes predefined classes to these pages. this inductive learning way builds a model from previously viewed web pages and in next step uses this model for web pages classification. lots of machine learning classifications has been used for web pages classification.among them we can point to support vector machine,-K neighborhood algorithm and Bayesian classification. Classification is a supervised way for classifying data in a way that more similar elements perch in the same group. In contrast, clustering is an unsupervised learning method which discover hidden relationships between data. These relationships can be used even for better classification of one class elements. Finding proper pattern between data is complex because usually this pattern is hide and cannot be seen easily. That's why machine learning algorithms are needed for classification. web pages classification can also be done with unorganized data. Final goal of web pages classification is to classify them into some predefined classes. Web pages classification process consists of several stages: page retrieval, stemming, Stop-word Filtering, words weight calculation, reduction and feature selection and finally classifying documents with intended classifier in the proposed scheme for last phase, a

Corresponding Author: Reyhaneh Khademi, Department of Computer, Islamic Azad university, Buinzahra Branch, Tehran, Iran

development for support vector machine will be proposed in which a set of support vector machine classification will be used that are learnt on a normal distribution that its data are extracted from Iran's english-language journals database. Using ensemble classifier can reduce variance of data sets and introduce more apparent concept than a simple classifier.

Literature review: Bhimavaram and Govindarajulu (2015) says an optimal system has been proposed for process Improvement of ontology classification. In ontology, concepts are classified. to do this partial or structural named concepts are calculated for inductive test. Because inductive tests are costly, reduction of such test numbers is vital. For this purpose Top-down and bottom-up queries are optimized (Bhimavaram and Govindarajulu, 2015). Classification accuracy can be improved with combination of text classification based on feature selection and pre-processing with reduced dimensions. Comparison between-K neighborhood algorithm and support vector machine in (Gayathri and Marimuthu, 2013) shows that an efficient-K neighborhood has less accuracy compared with support vector machine. Except clear features of -K neighborhood algorithm, in high computing issues failure in classification occurs. To deal with this problem, the authors have proposed a support vector machine technique. Support vector machine can be used as document classifier and has showed its superiority to other ways in classification operations.

By combining support vector machine and decision tree better results can be obtained in text mining. This combined method uses efficient computation property of tree structure and high classification accuracy of support vector machine (He and Liu, 2008) a new criterion has been proposed based on support vector area descriptor. Main idea of proposed algorithm is that one class of a problem can be divided into a set of two classes problems and these problems can be easily solved with support vector machine. SVDD 16 support vector data descriptor has been used for data description while in (Li and Chen, 2012) fuzzy features are used. parameters named number of occurrences of defined words and words that have a high value in this parameter are extracted and added to the list of selected properties. Features that show document subject are selected as weighted repeat in order to fix repeat words way flaws. Then support vector machine is used for classifier learning. Mr. khabaz and kianmehr in (Khabbaz *et al*., 2012) have used soft words clustering in combination with feature reduction in order to build information features vectors.these vectors show structural and textual XML documents aspects. For extracting information in order to identify richest structural information in XML document, tree mining algorithm and information gain filter are used. A heavy test has been done on standard data set including 20 newsgroup and XML document sets with web server logs. They use LOGML data set (Khabbaz *et al*., 2012) because the classifier is made only based on textual features, results shows that their proposed way works better than simple support vector machine. With applying support vector machine and decision tree using feature vector magnifier on XML document sets, classification accuracy has been obtained 85.79%. These results are better than Xrules accuracy of one structure base known classifier for XML documents classification.

A new fast learning algorithm for support machine vector has been proposed in (Arora *et al*., 2012) for conditions that samples are experiencing aliasing. These aliasing data which are not in the same class are separated and dependence vectors among them are calculated. According to proposed way in this research this vector can be used for identifying samples that may be vital but are apparently alias. Important point is that classifier accuracy in this method has the same quality as that of the entire data set is used directly.

An algorithm has been introduced for classifying news into different groups in order to identify most popular news groups in special time and place (country) in (Xu and Geng, 2012). Short massages are extracted from twitter and classify into 12 groups. These 12 groups are used for learning in machine learning. Each word in short massages is considered as a feature and a feature vector using bag of words strategy is built. Main reason for using support vector machine is its application in high dimensional data. To avoid data bias cross validation method used. A simple Bayesian classifier algorithm based on Independent Component Analysis (ICA) can be used for web pages classification. This method has been used (Dilrukshi *et al*., 2013; He and Liu, 2008; Xu *et al*., 2010; Araujo and Martinez, 2010; Chen *et al*., 2010; Ofuonye *et al*., 2010) and significant improvement on a set of data used (Dilrukshi *et al*., 2013) have been made using Bayesian classifier Enhancements. Web mining means analysis of identified patterns, in other words by analyzing users web pattern, his special pattern can be identified. Focus is on techniques that predict user interacts with web. To do this they use the web secondary data. Compare previous studies have shown that according to the vastness of the web, the need to classify these pages is inevitable. Many efforts have been made to classify web pages optimally. However, further improvements can be achieved. One of the main challenges in this area is the lack of balance between data of this data sets. Support vector machine performance is improved by having a larger data set. Learning process in this algorithm is rather easy. Over the past two decades, many methods including: NB classifier (Fan *et al*., 2001), self-organized neural networks (Zhang *et al*., 2001), SVM (Xue *et al*., 2006) and others have been proposed for web pages classification. Some combination of the above methods approach has been used to categorize web pages. For example, Vmin and Agzin (Xue *et al*., 2006) in SVM and NB classifier have used Body, title, heading and meta text (Explanatory texts) for features selection. Results show combination of these features with SVM classifier has been provided better efficiency for web pages classification. Jin *et al*. (2007) have used concepts such as relief network, information gain, Chi-square function as selected feature to improve efficiency of web page classification. Zhang-Ching and Chang-san (Chen and Hsieh, 2006) have proposed a method that two types of characteristics were used as SVM classifying input- for web page classification.

Output of two SVM was done to determine class of web page with a voting schema. Voting operation can improve efficiency in comparison with other methods. Fang and others have proposed a web page classifier that use 5 classification method. Outputs of these SVM is used as voting schema input and final classification is done based on highest rate. This method compared with the separate classifiers (single-single) has improved performance. Zhang *et al*., 2008) have shown a web page classification based on LS-SVM30. The LSA to obtain semantic structure, apply the original document terms matrix for solving SVD decomposition keywords problem. The LS-SVM is an effective method for learning mass data classification knowledge (especially valuable) by having labeled samples. Moayed *et al*. (2008) have used a swarm intelligent algorithm in context of web pages classification focused on Persian web pages. Algorithm name is ant miner that highest accuracy for Iran's Broadcasting News Network 1 is the 89%. Hossaini *et al*. (2008) have been used Genetic algorithms for classifying and clustering web pages. This algorithm is used on vectors with variable size. In genetic algorithm part they have focused on crossover standard operators and mutation combined with k-means algorithm for diversity improvement and results validity. Based on this method, with access to more detailed categories, sub-categories were defined as clusters. Their method leads to more accurate results with accuracy rates of approximately 90.7% as compared to fixed procedures and also unnecessary elements in vectors were ignored. He and Liu (2008) have used a method based on independent component analysis for web page classification using NB classifier. The source of one of the main challenges in the education categories is unbalanced data set. Numbers of one web page kind can be more or less than others. Standard machine learning algorithms ignoring this lack of balance, trends to main class and ignore small and subsidiary classes and this will leads to appearance of high False Negative Rate (FNR).

Unknown aspects and variables related to the research questions: Unknown aspect of the problem is the quality of web page classification using a method that ensures the accuracy. One of the main problem variables is the quality of generated categories in data before and after using the collective support vector machine. And this is quantifiable and measurable with algorithms efficiency criterion. The aim of this research is to improve web pages classification of Iranian English-language publications quality in a way that researchers visiting magiran website get their desired response from their conducted search the best time. Reaching this goal requires building a strong model for classifying this set contents. Comparison between previous researches shows that considering the extent of web pages, the need for classification is inevitable. Many efforts have been made to classify web pages optimally. However, further improvements can be achieved. One of the main challenges in this field is made by imbalance between these data of these data sets on the internet. Support Vector Machine performance is improved by having a larger data set. learning process in this Binary algorithm is somewhat easy.

General aims and hypothesis of research:

- Web pages classification of Iranian English-language publications quality can be improved using machine learning algorithm
- Solve the problem using machine learning algorithms has lower costs and greater efficiency

Class of Iranian English-language publications quality improvement using machine learning Introducing conceptual effective features of Iranian English-language publications web pages

MATERILAS AND METHODS

The data is Thrown in an Excel file Line by line.first line of this data set that is stored in excel file is like Table 1. This dataset finally includes 245 rows and each row has some variable keywords between 2-5 words. Lable of columns in this data set includes 9 class in alpha betical order: artificial intelligence, communication, control, electricity, image processing, machine learning, network, power system and security. Tagging is done from title and abstract of journal and a word has been selected that is suitable represent of this article's class.

Table 1: Samples number of each

Class label	Class no	Count	Percent
Artificial intelligence	1	27	0.05
Communication	2	121	0.24
Control	3	30	0.06
Electricity	4	119	0.24
Image processing	5	28	0.06
Machine learning	6	28	0.06
Network	7	47	0.09
Power system	8	47	0.09
Security	9	48	0.10
Sum	-	495	-

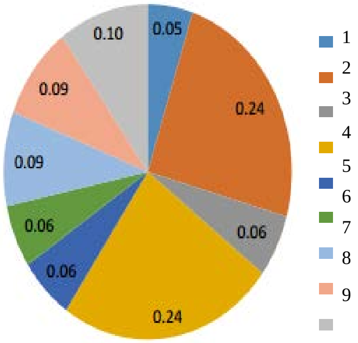

Fig. 1: Each class ration of the entire data set

for example despite for some articles fuzzy, mining, data, text mining, face detection, electronic has been considered but finally With reviewing and reasons like overlapping and small numbers of these tags sample, these classes have been merged with other tags and finally 9 tag classes appears. Table 1 shows each class ration in all data sets. class According to Table 1, this set includes 495 web page. Means 27 artificial intelligence documents, 121 communications documents, 30 control documents, 119 electricity documents, 28 image processing and machine learning documents 47 power and network documents and 48 security documents. According to chart number of two classes (electricity and communications) is equal to 7 other classes (half of data set). Figure 1 shows each class ratio. in Fig. 1, there is three 6%, two 9 and 24, one 5 and one 10% classes. As shown smallest ration is for artificial intelligence with 5%.

RESULTS AND DISCUSSION

Linear SVM method: Different performances has been done on initial data set and different results have been checked. Data set has 1976 base features (non-repetitive keyword) and 495 fix columns. The 9 column label with the weight of keywords 5 and 4, 3 and 2 is intended. Weighted action and feature vector creation of this data set is done with C# program language to:

Weighting data set algorithm:

```
Input: raw data set with dimensions of 1977-1976 (495 features and one column label).
First put all keywords of all 495 articles in first line
Remove duplicate key words.
Do the following 495 times for each row from the second row to later.
    In the first phase of weighting, in each row for keywords depending on the number of them, initialize the corresponding column from 5-2.
    For other non keywords with C#functions by connecting to Wikipedia compute cosine similarities of both words pages (each keyword with non keyword) and replace its average with word's weight. In this stage with assumption of 4 keywords for every 1972 other non keywords(totally 1976) in this line this word's similarity with keywords is computed and averaged and is considered as that word weigh.
Output: data set includes 495 rows and 1977 columns (1 tag column).
```

Combined SVM classification algorithm:

```
Input: weighting data set with dimensions of 495-1977.
The 5th Fold method: for j from 1-5 do the following 5 times:
Divide data set into two classes of train with 400 members and test with 95 members.
Apart first column of test and train and store it in testy and trainy. put the rest in testx and trainx From trainy.
Apart train members classes and determine each class percentage (with dividing numbers by 400)
Build svmstruct number 1-9 from 9 classes with svmtrain function by linear parameter.
Run svmclassify from svmstructs 9 times
Build confusion matrix elements with 9-9 members on 95 members of testx set.
Do for the number of classes.
Compute accuracy and Repeatability 10 member arrays and F1 from confusion matrix. 10th places is reserved to other class
keep accuracy and repeatability and F1 in an excel file
Output: excel file includes 10 rows (class) and 3 column (accuracy-repeatability-F1).
```

SVM implementation stage (combined and train and test) is done in MATLAB. In order to compare with other methods, classification on same data set is done using SVM-linear and SVM-RBF and Decision Tree and results have been shown in Table 2-5 respectively. Table 5 shows the comparison between linear SVM , radial SVM and decision tree using F criterion. Figure 2 shows that linear SVM is more efficient than two other methods. RBF method also is not good for this special data set. Radial SVM efficiency is between two other methods. Reason for using weighted average instead of ordinary averaging is that number of 9 class members are not equal. It's clear that accuracy in a class with more members should have more effect than class with less members. So, for computing weighted average, each class weigh has been considered as a factor.

Table 6 and Fig. 3 shows running time of combined SVM algorithm with linear kernel in a core i7 processor, ram 6 and 4 GB graphics software.

Table 2: SVM inear paarameter results

Class No.	1	2	3	4	5	6	7	8	9	Other
FOL.D1										
Precision	0.499998	0.375	0.599999	0.24	0.499998	0	0.111111	0.111111	0.333333	0
Recall	0.166666	0.5	0.499999	0.25	0.2	0	0.125	0.1	0.3	0
F1	0.249999	0.428571	0.545454	0.244898	0.285713	0	0.117647	0.105263	0.315789	0
FOL.D2										
Precision	0	0.447368	0.749998	0.419355	0.399999	0	0	0.285714	0	0
Recall	0	0.68	0.428571	0.541666	0.399999	0	0	0.222222	0	0
F1	0	0.539682	0.545454	0.472727	0.399999	0	0	0.25	0	0
FOL.D3										
Precision	0.499999	0.296296	0.499998	0.290322	0.333333	0.499998	0.111111	0	0.333332	0
Recall	0.333333	0.333333	0.249999	0.375	0.333333	0.2	0.1	0	0.1	0
F1	0	0.313725	0.333332	0.327273	0.333333	0.285713	0.105263	0	0.153846	0
FOL.D4										
Precision	0	0.361111	0.166666	0.444444	0	0.99999	0.5	0	0.384615	0
Recall	0	0.541666	0.222222	0.333333	0	0.2	0.5	0	0.555555	0
F1	0	0.433333	0.499998	0.380952	0	0.333332	0.5	0	0.454545	0
FOL.D5										
Precision	0	0.384615	0.166666	0.388889	0	0.499999	0.214286	0	0.266666	0
Recall	0	0.625	0.249999	0.291667	0	0.399999	0.3	0	0.444444	0
F1	0	0.47619	0.536665	0.333333	0	0.444443	0.25	0	0.333333	0
Average										
Precision	0.199999	0.372878	0.536665	0.356602	0.246666	0.399997	0.187301	0.079375	0.263589	0
Recall	0.1	0.536	0.285614	0.358333	0.186666	0.16	0.205	0.064444	0.289	0
F1	0.3	0.438301	0.362626	0.351837	0.203809	0.212698	0.194582	0.071053	0.251503	0

Table 3: SVM inear paarameter results

Class No.	1	2	3	4	5	6	7	8	9	Other
FOL.D1										
Precision	0.99999	0	0.99999	0.99999	0	0	0	0	0	0
Recall	0.16667	0	0.16667	0.16667	0	0	0	0	0	0
F1	0.28571	0	0.28571	0.08	0	0	0	0	0	0
FOL.D2										
Precision	0	0	0.99999	0.99999	0	0	0	0	0	0
Recall	0	0	0.14286	0.16667	0	0	0	0	0	0
F1	0	0	0.25	0.08	0	0	0	0	0	0
FOL.D3										
Precision	0.99999	0	0	0	0	0	0	0	0	0
Recall	0.16667	0	0	0	0	0	0	0	0	0
F1	0.28571	0	0	0	0	0	0	0	0	0
FOL.D4										
Precision	0	0	0	0	0	0	0	0	0	0
Recall	0	0	0	0	0	0	0	0	0	0
F1	0	0	0	0	0	0	0	0	0	0
FOL.D5										
Precision	0	0	0	0	0	0	0	0	0	0
Recall	0	0	0	0	0	0	0	0	0	0
F1	0	0	0	0	0	0	0	0	0	0
Average										
Precision	0.4	0	0.4	0.4	0	0	0	0	0	0
Recall	0.06667	0	0.0619	0.01667	0	0	0	0	0	0
F1	0.11429	0	0.10714	0.032	0	0	0	0	0	0

Table 4: Decesion tree algorithm results

Class No.	1	2	3	4	5	6	7	8	9	Other
Fold1										
Precision	0.50	0.40	0.75	0.23	0.33	0.00	0.00	0.43	0.33	0.00
Recall	0.17	0.42	0.50	0.25	0.20	0.00	0.00	0.30	0.10	0.00
F1	0.25	0.41	0.60	0.24	0.25	0.00	0.00	0.35	0.15	0.00
Fold2										
Precision	0.00	0.53	1.00	0.29	1.00	0.50	0.25	0.67	0.50	0.00
Recall	0.00	0.40	0.29	0.33	0.20	0.17	0.20	0.22	0.10	0.00
F1	0.00	0.45	0.44	0.31	0.33	0.25	0.22	0.33	0.17	0.00
Fold3										
Precision	0.50	0.39	0.00	0.29	1.00	0.33	0.33	0.40	0.00	0.00
Recall	0.33	0.50	0.00	0.25	0.17	0.20	0.10	0.22	0.00	0.00
F1	0.40	0.44	0.00	0.27	0.29	0.25	0.15	0.29	0.00	0.00
Fold4										
Precision	0.00	0.19	0.50	0.22	1.00	0.00	0.20	0.50	0.25	0.00
Recall	0.00	0.25	0.17	0.17	0.17	0.00	0.10	0.40	0.11	0.00
F1	0.00	0.22	0.25	0.19	0.29	0.00	0.13	0.44	0.15	0.00

Table 4: Continue

Class No.	1	2	3	4	5	6	7	8	9	Other
Fold5										
Precision	0.00	0.29	0.00	0.38	0.50	0.29	0.00	0.20	0.37	0.00
Recall	0.00	0.21	0.00	0.33	0.17	0.40	0.00	0.10	0.33	0.00
F1	0.00	0.24	0.00	0.36	0.25	0.33	0.00	0.13	0.35	0.00
Average										
Precision	0.20	0.36	0.45	0.28	0.77	0.22	0.16	0.44	0.29	0.00
Recall	0.10	0.35	0.19	0.27	0.18	0.15	0.08	0.25	0.13	0.00
F1	0.13	0.35	0.26	0.27	0.28	0.17	0.10	0.31	0.17	0.00

Table 5: Evaluation of classification method

		FI		
Class label	Percentage	SVM-linear	SVM-RBF	DT
1	0.056	0.129999715	0.114285388	0.129999715
2	0.242	0.438300505	0	0.352231177
3	0.062	0.362625517	0.107142569	0.258888326
4	0.240	0.351836589	0.031999974	0.272078028
5	0.056	0.339681816	0	0.280951582
6	0.056	0.354496383	0	0.166666306
7	0.096	0.177631367	0	0.101880210
8	0.096	0.177631367	0	0.309952905
9	0.096	0.314378289	0	0.165459797

Average (Weighted); 0.329728998, 0.020722815, 0.254356453

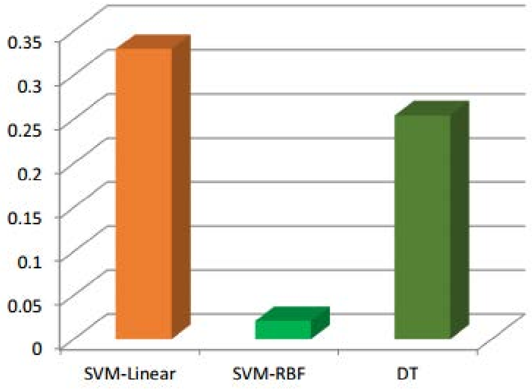

Fig. 2: The weighted average of F_1 criteria in three classification methods

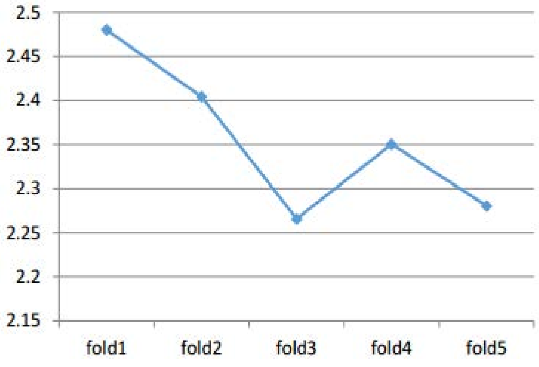

Fig. 3: Running time of SVM with linear kernal function

Table 6: Folds running time in linear SVM algorithm

Folds	Time (sec)
1	2.480214
2	2.404239
3	2.265610
4	2.350260
5	2.280246

CONCLUSION

Automatic classification of web pages is one of challenging research in text mining that plays an important role in building the semantic web. Many efforts have been made to categorize web pages. However, more opportunities is available to improve current methods. One of the main challenges in educational classifiers is the fact that the existing dataset is unbalanced. (one sample is very small while another is great). So F_1 value is unstable in most existing methods. In this research, unbalanced data set has been studied in web page classification and to solve this problem, SVM classifiers approach is proposed. accuracy comparison between radial SVM and decision tree with proposed combined classifier method is shown. Results shows that proposed approach almost solves the problem of lack of web pages balance.

REFERENCES

Araujo, L. and R.J. Martinez, 2010. Web spam detection: New classification features based on qualified link analysis and language models. IEEE Trans. Inf. Forensics Secur., 5: 581-590.

Arora, M., U. Kanjilal and D. Varshney, 2012. Efficient and intelligent information retrieval using Support Vector Machine (SVM). Int. J. Software Comput. Eng., 1: 39-43.

Bhimavaram, S. and P. Govindarajulu, 2015. An enhanced approach for ontology based classification in semantic web technology. Int. J. Adv. Res. Comput. Commun. Eng., Vol. 4,

Chen, R.C. and C.H. Hsieh, 2006. Web page classification based on a support vector machine using a weighted vote schema. Expert Syst. Applic., 31: 427-435.

Chen, T.C., S. Dick and J. Miller, 2010. Detecting visually similar web pages: Application to phishing detection. ACM Trans. Internet Technol., Vol. 10, 10.1145/1754393.1754394

Dilrukshi, I., D.K. Zoysa and A. Caldera, 2013. Twitter news classification using SVM. Proceeding of the 2013 8th International Conference on Computer Science & Education, April 26-28, 2013, IEEE, Kotte, Sri Lanka, ISBN:978-1-4673-4463-0, pp: 287-291.

Fan, Y., C. Zheng, Q.Y. Wang, Q.S. Cai and J. Liu, 2001. Web Page classification based on naive bayes method. J. Software, 12: 1386-1392.

Gayathri, K. and A. Marimuthu, 2013. Text document pre-processing with the KNN for classification using the SVM. Proceeding of the 2013 7th International Conference on Intelligent Systems and Control, January 4-5, 2013, IEEE, Coimbatore, India, ISBN:978-1-4673-4603-0, pp: 453-457.

He, Z. and Z. Liu, 2008. A novel approach to naive bayes web page automatic classification. Proceeding of the FSKD'08 5th International Conference on Fuzzy Systems and Knowledge Discovery, October 18-20, 2008, IEEE, Xi'an, China, ISBN:978-0-7695-3305-6, pp: 361-365.

Hossaini, Z., A.M. Rahmani and S. Setayeshi, 2008. Web pages classification and clustering by means of genetic algorithms: A variable size page representing approach. Proceeding of the 2008 International Conference on Computational Intelligence for Modelling Control and Automation, December 10-12, 2008, IEEE, Tehran, Iran, ISBN:978-0-7695-3514-2, pp: 436-440.

Jin, X., R. Li, X. Shen and R. Bie, 2007. Automatic web pages categorization with relieff and hidden naive bayes. Proceedings of the ACM Symposium on Applied Computing, March 11-15, 2007, Korea, pp: 617-621.

Khabbaz, M., K. Kianmehr and R. Alhajj, 2012. Employing structural and textual feature extraction for semistructured document classification. IEEE. Trans. Syst. Man Cybern. Appl. Rev., 42: 1566-1578.

Li, Y. and C. Chen, 2012. Research on the feature selection techniques used in text classification. Proceeding of the 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery, May 29-31, 2012, IEEE, Xi'an, China, ISBN: 978-1-4673-0024-7, pp: 725-729.

Moayed, M.J., A.H. Sabery and A. Khanteymoory, 2008. Ant colony algorithm for web page classification. Proceeding of the 2008 International Symposium on Information Technology, August 26-28, 2008, IEEE, Kuala Lumpur, Malaysia, ISBN:978-1-4244-2327-9, pp: 1-8.

Ofuonye, E., P. Beatty, S. Dick and J. Miller, 2010. Prevalence and classification of web page defects. Online Inf. Rev., 34: 160-174.

Xu, Q. and S. Geng, 2012. A fast SVM classification learning algorithm used to large training set. Proceeding of the 2012 2nd International Conference on Intelligent System Design and Engineering Application, January 6-7, 2012, IEEE, Lianyungang, China, ISBN:978-1-4577-2120-5, pp: 15-19.

Xu, S.M., B. Wu and C. Ma, 2010. Efflcient SVM chinese web page classifier based on pre-classification. Comput. Eng. Appl., 2010: 125-128.

Xue, W., H. Bao, W. Xue, W. Huang and Y. Lu, 2006. Web page classification based on SVM. Proceedings of the 6th World Congress on Intelligent Control and Automation, Volume 2, June 21-23, 2006, Dalian, China, pp: 6111-6114.

Zhang, Y., B. Fan and L.B. Xiao, 2008. Web page classification based on a least square support vector machine with latent semantic analysis. Proceeding of the FSKD'08 5th International Conference on Fuzzy Systems and Knowledge Discovery, October 18-20, 2008, IEEE, Lanzhou, China, ISBN:978-0-7695-3305-6, pp: 528-532.

Zhang, Y.Z., M. Zhao and Y. Wu, 2001. The automatic classification of web pages based on neural networks. Neural Inf. Process., Proc., 2001: 570-575.