International Journal of Automation and Computing 9(4), August 2012, 442-448
DOI: 10.1007/s11633-012-0665-x

A Novel Feature Selection Framework for Automatic Web Page Classification

J. Alamelu Mangai[1] V. Santhosh Kumar[1] S. Appavu alias Balamurugan[2]

[1]Department of Computer Science, BITS Pilani, Dubai Campus, DIAC, Dubai 345055, UAE

[2]Department of Information Technology, Thiagarajar College of Engineering, Madurai 625015, India

Abstract: The number of Internet users and the number of web pages being added to WWW increase dramatically every day. It is therefore required to automatically and efficiently classify web pages into web directories. This helps the search engines to provide users with relevant and quick retrieval results. As web pages are represented by thousands of features, feature selection helps the web page classifiers to resolve this large scale dimensionality problem. This paper proposes a new feature selection method using Ward′s minimum variance measure. This measure is first used to identify clusters of redundant features in a web page. In each cluster, the best representative features are retained and the others are eliminated. Removing such redundant features helps in minimizing the resource utilization during classification. The proposed method of feature selection is compared with other common feature selection methods. Experiments done on a benchmark data set, namely WebKB show that the proposed method performs better than most of the other feature selection methods in terms of reducing the number of features and the classifier modeling time.

Keywords: Feature selection, web page classification, Ward′s minimum variance, information gain, WebKB.

1 Introduction

With web being used to its full potential, retrieving more relevant results to a user query is a challenge to the search engines. With the introduction of the WWW, data accumulates in a speed unmatchable by the humans capacity of data processing. This huge repository needs to be meticulously organized to enable efficient and quick information retrieval. Manually classifying web pages into predefined categories needs time, human expertise and utmost care. All these factors have paved the way for automatic web page classification (WPC). WPC also helps many information management and retrieval tasks such as constructing and expanding web directories, improving the quality of web search results, building efficient focused crawlers, helping question answering systems and building domain specific search engines[1]. WPC has a strong connection with natural language processing, data mining, text mining, machine learning, information retrieval and knowledge management.

Since web pages have text and multimedia data, they can be viewed as structured, semi-structured or unstructured. This has imposed additional challenges to WPC than traditional text classification. Many algorithms and approaches for selecting the best features for WPC and modeling web page classifier itself have been stated in literature. The performance of all such algorithms depends highly on the preprocessing done on the web pages and the initial set of extracted features. Furthermore, the presence of redundant and irrelevant attributes could mislead the analysis. Including all the attributes in the data mining procedures not only increases the complexity of the analysis, but also degrades the accuracy of the result. So, the best representative features of a web page have to be extracted and this has a significant role in reducing the dimensionality, time and resources needed to classify a web page. The objective of this paper is to propose a feature selection method which helps to improve the performance of web page classifiers. This method uses Ward′s minimum variance measure to identify the redundant features in a web page. Two features are said to be redundant, if the variance between them is less, otherwise they are non-redundant. The best representative features in a cluster are retained and the others are eliminated. This method of reducing the dimensionality helps in improving the performance of the web page classifiers by reducing their learning time.

The rest of the paper is organized as follows. Section 2 highlights the related work, proposed work is described in Section 3, details of the experiments are summarized in Section 4, and Section 5 highlights the results and findings.

2 Related work

Many approaches for automatic web page classification have been witnessed over years in literature. Without preprocessed data, there is no good mining results. The performance of the web page classifiers are improved from different perspectives, e.g., dimensionality reduction (feature selection), using the word occurrence statistics in a web page (content based), using the relationship between different web pages (link based), using the association between queries and web pages (query log based), and using the structure of page, images, links contained in the page and their placement (structure based). This paper focuses on improving WPC by reducing the feature space.

Since web pages are of high-dimensions and have noisy information, they need to be properly preprocessed to reduce the learning time and the complexity of the classifiers. Feature selection is one way of solving the curse of dimensionality for content based web page classifiers. It has been an active research area in pattern recognition, statistics, and data mining communities. The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information.

Manuscript received January 2, 2012; revised February 8, 2012

Feature selection can significantly improve the comprehensibility of the resulting classifier models and often build a model that generalizes better to unseen points. Further, it is often the case that finding the correct subset of predictive features is an important problem in its own right.

In [2], a study of the appropriate feature selection techniques for WPC is explored to obtain a minimum number of highly qualitative features. The authors have combined CfsSubset evaluator with term frequency to achieve good classification accuracy[2]. Three distinct features of a web page namely, the URL, title and meta data which are believed to have more predictive information about a web page are used[3] with machine learning methods to classify a web page. The output of principal component analysis (PCA) is combined with a manual weighting scheme to classify web pages using neural networks in [4]. A fuzzy ranking analysis with discriminating power measure[5], rough set theory[6] and an integrated use of ant colony optimization with fuzzy-rough sets[7], is used to reduce the dimensions of web pages. A study of rough sets, their extension and applications in data mining is explored in [8]. A new feature selection method that incorporates hierarchical information about the categories is used in [9]. This prevents the classifying process from going through every node in the hierarchy. On page features (content based) and features of neighboring pages (context based) are used[10] to classify a web page.

A genetic algorithm that determines the best features for a given set of web pages is proposed in [11], to decrease the feature space. The feature space can also be reduced by identifying and eliminating redundant (relevant) features in a web page. A feature selection method using Bayes theorem is proposed in [12]. The dependence between two attributes is determined based on the probabilities of their joint values that contribute to the positive and negative classification decisions. This paper proposes Ward′s minimum variance measure to identify clusters of similar or redundant features in a web page. In each such cluster, the best representative features are retained and the others are eliminated. This helps in reducing the dimensions of the data, which is used to model the web page classifiers.

3 Proposed work

It is widely accepted by the machine learning community that, there is no good mining results without good data. So, in this paper, web page classification is implemented in three steps namely, preprocessing, feature selection using the proposed method and classifier induction.

3.1 Preprocessing

For high-dimensional data sets, like web pages, if features are not properly selected, the time taken to model the classifier will increase. This work attempts to minimize the time taken to build the classifier by using a series of preprocessing steps. The preprocessing steps for improving the performance of web page classifiers are described as follows.

1) Convert each web page to a text file.

2) Extract the best features from each web page and construct a web page-feature matrix. Find the term frequency and inverse document frequency of each new feature in the web pages. These features will be the best representative features of a web page category, F_{initial}. Each web page is represented as a vector with the weight of a feature f in a web page i calculated as

$$w_{ij} = tf_{ij} \times idf_i = tf_{ij} \log_2 \left(\frac{N}{df_i} \right) \tag{1}$$

$$tf_{ij} = \frac{f_{ij}}{\max_i f_{ij}} \tag{2}$$

where N is the number of web pages in the collection, tf_{ij} is the frequency of term i in web page j, df_i is the web page collection frequency of a feature, $\max_i \{f_{ij}\}$ is the frequency of the most common term in the web page.

3) Remove web pages which have all feature weights as zero.

4) Identify and eliminate the duplicate and conflicting web pages.

3.2 Proposed feature selection framework

The proposed method has two steps: Identify clusters of redundant features (using Ward′s minimum variance measure) and eliminate redundant features.

3.2.1 Identify clusters of redundant features

Ward′s algorithm[13] is a commonly used procedure for forming hierarchical groups (clusters) of mutually exclusive subsets. Other than computing distances between clusters, it forms clusters by maximizing within-clusters homogeneity. The within-group (i.e., within-cluster) sum of squares is used as the measure of homogeneity. That is, the Ward′s method tries to minimize the total within-group or within-cluster sum of squares. Clusters are formed at each step such that the resulting cluster solution has the fewest within-cluster sums of squares. The within-cluster sums of squares is also known as the error sums of squares (ESS) or variance E. This method is considered to be the very efficient[14]. Although it tends to create clusters of small size, the intra-cluster similarity will be more. Motivated by this key property of this method, this paper proposes to use the same idea for identifying redundant features in a web page.

Input: Feature matrix F, of order $n \times m$, where n is the number of instances and m is the number of features.

Output: Clusters with redundant features $C_{\text{reduntant}}$ where $C_{\text{reduntant}} = \{C_1, C_2, C_3, \cdots, C_{n_k}\}$. Each C_i in $C_{\text{reduntant}}$ is a cluster of redundant features, p_i is the size of cluster C_i and n_k is the number of clusters.

Method

1) Initially, each feature itself is in an individual cluster, with variance, $E = 0$.

2) Repeat the following steps for each adjacent pairs of rows F_iF_{i+1}, where $i = 0$ to $n - 1$. (Note: Clusters formed in an earlier stage are never unmerged).

2.1) Find all possible mergers of features in the two rows.

2.2) Calculate the mean and variance E of each possible merger.

2.3) Choose the merger with minimum variance E, where E of a possible merger is defined by

$$E = \sum_{i=1}^{n_k} \sum_{j=1}^{p_i} [F_{ij} - m_i]^2$$

where n_k represents the number of clusters in the current merger, m_i represents the mean of cluster i in the current merger, F_{ij} represents the value of j-th feature in cluster i, and p_i represents the number of features in cluster i.

2.4) Calculate the vote of each merger with minimum E, identified for these two rows.

3) By majority voting, choose the merger with highest vote. Each merger will have clusters with different number of features. Clusters with more than one feature are said to have redundant feature groups

3.2.2 Eliminate redundant features

Input: Clusters with redundant features $C_{\text{reduntant}}$ where $C_{\text{reduntant}} = \{C_1, C_2, C_3, \cdots, C_{n_k}\}$. Each C_i in $C_{\text{reduntant}}$ is a cluster of reduntant features, p_i is the size of cluster C_i, n_k is the number of cluster, and p_i is the size of a cluster.

Output: Reduced feature set $R = \{C_{1\text{best}}, C_{2\text{best}}, \cdots, C_{n_k\text{best}}\}$

Method

for $i = 1$ to n_k do
  for each C_i in $C_{\text{reduntant}}$
    for $j = 1$ to p_i
      1) Rank each feature F_i, using its information gain in predicting the class label
      2) Select the feature F_i with the highest rank
    end
  end
end

3.3 Classification

To evaluate the proposed feature selection method for web page classification, various machine learning classifiers like neural network based classifier (NN), k nearest neighbor (kNN), support vector machine based classifiers (SVM) and ensembled classifiers (boosting), naive Bayesian classifier (NB), decision tree based classifier (J48) are trained using WEKA. Previous studies have shown that NN, kNN, SVM and boosting methods exhibit good performance with classifying high dimensional data. NB and J48 classifiers are known for their simplicity and easy interpretation of the results respectively. Features are selected by other common feature selection methods namely, PCA, information gain, relief, gain ratio, oneR Attribute eval, and symmetric uncertainty attribute eval. The performances of the classifiers are evaluated with the full set of features and with the features selected by various other methods. The classifiers are tested using 10 fold cross-validation for various sizes of web page collections.

4 Experiments and results

Experiments were done on a benchmark data set called WebKB[15]. This data set contains WWW-pages collected from computer science departments of various universities. The pages are manually classified into the following categories: student, faculty, staff, department, course, project and others. For the analysis of the proposed work, course web pages are considered as positive examples and student web pages as negative examples. Table 1 shows the details of the various web page collections taken for experiments.

Table 1 The initial data set

Data set	No. of instances	No. of features
70–30	100	2774
100–100	200	4185
200–200	400	6654
300–200	500	7874
300–300	600	8963
350–150	500	7651
400–200	600	8508
400–300	700	9563
400–400	800	10363

To reduce the high dimensions of the web pages, they undergo a series of preprocessing steps. The HTML tags, stop words, punctuations, digits and hyphens are removed from the web pages in our preprocessing phase. Then, the web pages are subsequently stemmed. Table 2 shows the results after preprocessing over the various number of positive–negative web page collections.

Table 2 The data set after preprocessing

Data set	No. of instances	No. of features
70–30	56	5
100–100	92	6
200–200	291	13
300–200	298	9
300–300	414	14
350–150	391	13
400–200	432	15
400–300	557	17
400–400	585	17

The results of Table 2 indicate that preprocessing reduces the dimensions of the web pages considerably. This is required for improving the performance of the web page classifiers. Table 3 shows the results of the proposed feature selection method using Ward′s minimum variance measure. The clusters with more than one members are features with minimum variance E, found by the Ward′s method. So, features in these clusters are identified as redundant features. Modeling a classifier using all these redundant features will require maximum utilization of computer resources. Therefore, only the features in a cluster that have more predictive information of the category of a web page are retained and the others are eliminated. Such best representative features in each cluster are selected by ranking feathers using information gain.

Table 4 compares the performance of the proposed method with PCA, information gain, relief, gain ratio, oneR Attribute eval, symmetric uncertainty attribute eval. Compared to the other feature selection methods, it can be inferred from Table 4 and Fig. 1, that the proposed method using Ward′s minimum variance achieves the highest level of dimensionality reduction. The methods info gain, relief, gain ratio, oneR and symmetric uncertainty attribute eval suggest that all attributes are significant for classification. PCA reduces the number of features for seven out of ten input data sets. The significance of the features selected is tested by running various machine learning classifiers namely, NN, kNN, SVM, ensembled boosting method, NB, and J48. The performances of the classifiers are

Table 3 Clusters of redundant features formed and the final selected features

Data set	Clusters	Selected features
70–30	(1) (2,3,4,5)	1,5
100–100	(2,4,5,6) (3) (1)	1,3,5
200–200	(2,3,12) (11,13) (1) (4) (5) (6) (7) (8) (9) (10)	1,4,5,6,7,8,9,10, 12,13
300–200	(4,5,7,9) (8) (1) (2) (3) (6)	1,2,3,4,6,8
300–300	(12) (3,4,5,14) (1) (2) (5) (6) (7) (8) (9) (10) (11) (12) (13)	1,2,6,7,8,9,10,11, 12,13,14
350–150	(2,3,4,13) (1) (5) (6) (7) (8) (9) (10) (11) (12)	4,1,5,6,7,8,9,10,11,12
400–200	(2,3,5,15) (1) (4) (6) (7) (8) (9) (10) (11) (12) (13) (14)	2,1,4,6,7,8,9,10,11,12,13,14
400–300	(1,4,5,17) (2) (3) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16)	2,3,6,7,8,9,10,11,12,13,14,15,16,17
400–400	(1,4,5,17) (2) (3) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16)	2,3,6,7,8,9,10,11,12,13,14,15,16,17

Table 4 Number of features selected by each feature selection algorithm

Data set	Original No. of attributes	Ward′s method	PCA	Info gain	Relief	Gain ratio	OneR	Symetric
70–30	5	2	5	5	5	5	5	5
100–100	6	3	6	6	6	6	6	6
200–200	13	10	12	13	13	13	13	13
300–200	9	6	8	9	9	9	9	9
300–300	14	11	13	14	14	14	14	14
350–150	13	10	12	13	13	13	13	13
400–200	15	12	14	15	15	15	15	15
400–300	17	14	16	17	17	17	17	16
400–400	17	14	16	17	17	17	17	17

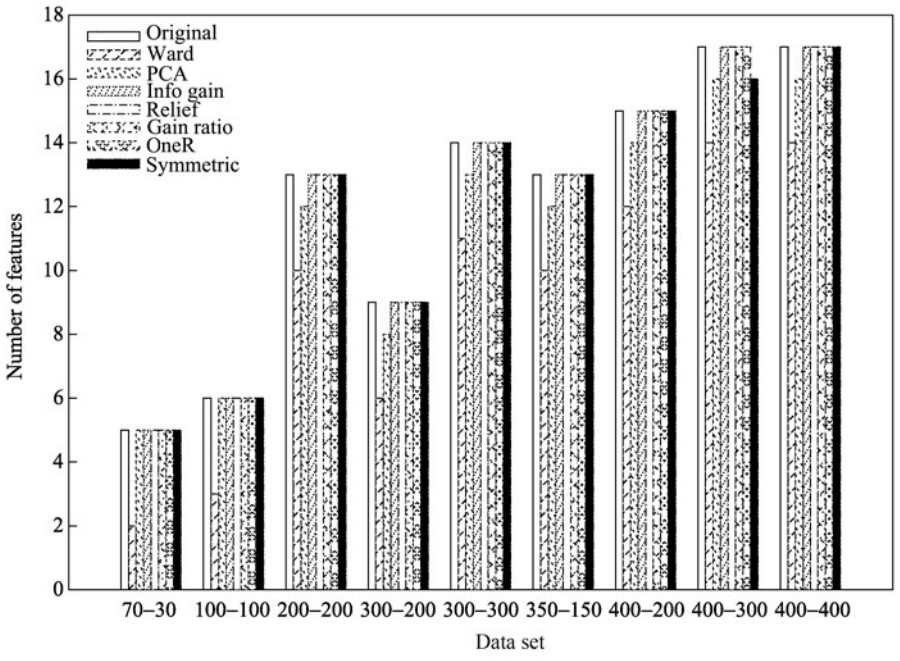

Fig. 1 Number of features selected by various feature selection methods

evaluated with three sets of features namely full set, feature subset chosen by Ward′s, and feature subset chosen by PCA. The classification accuracy of the kNN classifiers with features selected by various methods are shown in Table 5.

The results of Table 5 show that the classification accuracy is considerably the same in all three cases. With the reduced number of features, the proposed method is able to maintain the classification accuracy. The features selected are themselves more predictive of the category of the web page. The classification accuracy of the NN classifiers with features selected by various methods are shown in Table 6.

The results in Table 6 show that the performance of the NN classifiers with features selected by the proposed method is better than those selected by PCA. Although the accuracy with full features is slightly better than that with features selected by the proposed method, there is a significant difference in the time taken to model the classifier in all three cases as shown in Fig 2.

Table 5 Accuracy of kNN classifier on the features selected by each feature selection algorithm

Data set	Full features	Ward′s method	PCA
70–30	96.42	98.21	96.42
100–100	94.56	96.74	94.56
200–200	96.21	90.03	94.15
300–200	94.29	91.95	93.95
300–300	95.65	95.17	96.13
350–150	97.18	96.68	96.93
400–200	96.71	96.06	96.99
400–300	95.33	95.51	95.51
400–400	96.06	94.70	96.06
Average	95.82	95.00	95.63

Table 6 Accuracy of NN classifier on the features selected by each selection algorithm

Data set	Full features	Ward′s method	PCA
70–30	100.00	100.00	100.00
100–100	92.39	96.74	92.39
200–200	94.84	89.69	93.15
300–200	93.28	91.28	93.23
300–300	96.61	95.65	95.63
350–150	95.90	95.95	95.39
400–200	96.75	96.79	96.52
400–300	96.65	96.25	96.12
400–400	95.21	95.56	95.12
Average	95.74	95.33	95.28

It can be inferred from Fig. 2 that the time taken to model the NN classifier with features selected by the proposed method is the least of all three of feature selection methods. Running NN classifier with features selected by the proposed method is significantly faster than NN classifier with full set of features and features selected by PCA. The classification accuracy of the ensemble classifier namely AdaBoost, SVM, NB with features selected by various methods are shown in Tables 7 and 8, respectively. Tables 7 and 8 show that the classification accuracy is the same in all three cases. However, the classification accuracy is maintained with the reduced number of features selected by the proposed method rather than using the full set of features.

Table 7 Accuracy of boosting classifier on the features selected by each feature selection algorithm

Data set	Full features	Ward′s method	PCA
70–30	100.00	100.00	100.00
100–100	96.74	96.74	96.74
200–200	91.75	90.03	90.72
300–200	96.64	91.61	93.62
300–300	94.44	94.44	94.68
350–150	95.65	94.37	93.86
400–200	95.13	93.75	93.75
400–300	94.07	94.97	94.43
400–400	92.80	92.14	92.47
Average	95.24	94.23	94.47

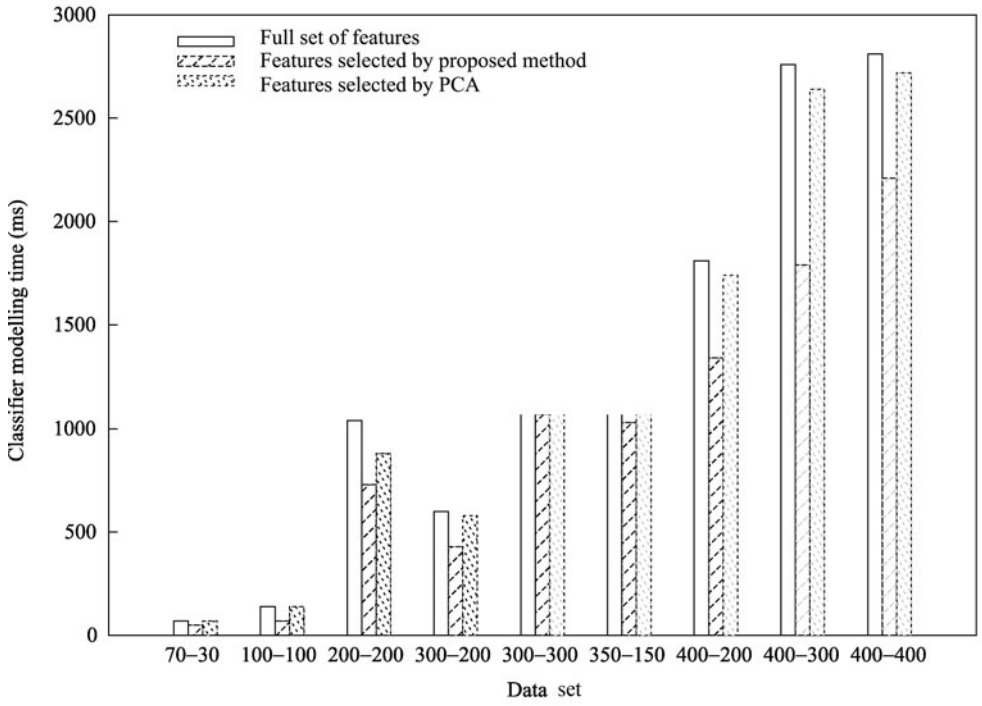

Fig. 2 Comparison of classifier modelling time of NN classifier

Table 8 Accuracy of SVM classifier on the features selected by each feature selection algorithm

Data set	Full features	Ward′s method	PCA
70–30	100.00	100.00	100.00
100–100	96.74	96.74	96.74
200–200	95.87	92.09	94.84
300–200	96.64	91.95	93.28
300–300	96.13	96.13	96.13
350–150	96.67	96.16	96.67
400–200	96.54	95.83	96.06
400–300	97.66	95.87	97.48
400–400	96.41	96.07	95.89
Average	96.96	95.65	96.34

Tables 9 and 10 show the results of classification accuracy with full set of features and reduced features using NB and J48 classifiers. In case of NB, the classification accuracy with features selected by PCA is the same as that obtained with the full set of features. In case of J48 classifier, the results in both cases are not as accurate as the results with the full set of features. However, the features selected by the proposed method result in better classification accuracy than those selected by PCA.

Table 9 Accuracy of NB classifier on the features selected by each feature selection algorithm

Data set	Full features	Ward′s method	PCA
70–30	100.00	100.00	100.00
100–100	93.47	88.04	93.47
200–200	94.15	91.75	93.81
300–200	96.64	90.27	93.95
300–300	95.89	95.65	95.89
350–150	96.41	94.63	95.65
400–200	93.98	93.29	93.75
400–300	95.15	95.15	94.97
400–400	94.70	94.70	94.18
Average	95.58	93.71	95.07

Table 10 Accuracy of J48 classifier on the features selected by each feature selection algorithm

Data set	Full features	Ward′s method	PCA
70–30	100.00	100.00	100.00
100–100	96.73	96.74	96.12
200–200	92.09	89.04	91.40
300–200	95.64	91.95	93.25
300–300	94.44	91.55	91.00
350–150	93.35	94.25	93.13
400–200	94.44	93.50	92.50
400–300	92.99	93.72	93.17
400–400	93.50	93.25	93.00
Average	94.78	93.78	93.73

To conclude, in terms of the factors such as running time and the number of features selected, the proposed method shows superior results over all the six feature selection methods tested. In terms of classification results, the average classification accuracy with features selected by the proposed method either increases or remains the same in 4/6 cases as compared to features selected by PCA.

5 Conclusions

Since web pages are high-dimensional data, they need to be preprocessed well before modeling a web page classifier. In this paper, the number of dimensions used to model a web page classifier is reduced after using a novel feature selection algorithm. This algorithm involves two steps: 1) identifying clusters of redundant features using Ward′s minimum variance measure; 2) the best representative feature of each cluster is then selected and the others are eliminated using information gain with ranked features. The proposed algorithm is used for classifying web pages from WebKB repository. The experimental results show that the proposed method selects significantly less number of features than many of the other existing feature selection methods. Various classification algorithms have shown good performance with these reduced features in terms of classification accuracy.

The proposed method of feature selection can be used in domains, like medical domains, where obtaining features is really expensive. Our future work would be to explore the performance of this method on multiple categories of web pages.

References

[1] J. Han, M. Kamber, J. Pei. *Data Mining: Concepts and Techniques*, 2nd ed., San Francisco, USA: Morgan Kaufmann, 2005.

[2] M. I. Devi, R. Rajaram, K. Selvakuberan. Generating best features for web page classification. *Webology*, vol. 5, no. 1, Article 52, 2008.

[3] L. W. Han, S. M. Alhashmi. Joint web-feature (JFEAT): A novel web page classification framework. *Communications of the IBIMA*, vol. 2010, Artical ID 73408, 2010.

[4] A. Salamat, S. Omata. Web page feature selection and classification using neural networks. *Information Sciences*, vol. 158, no. 1, pp. 69–88, 2004.

[5] C. M. Chen, H. M. Lee, Y. J. Chang. Two novel feature selection approaches for web page classification. *Expert Systems with Applications*, vol. 36, no. 1, pp. 260–272, 2009.

[6] T. Wakaki, H. Itakura, M. Tamura. Rough set-aided feature selection for automatic web-page classification. In *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*, IEEE, Beijing, China, pp. 70–76, 2004.

[7] R. Jensen, Q. Shen. Web page classification with ACO-enhanced fuzzy-rough feature selection. In *Proceedings of the 5th International Conference on Rough Sets and Current Trends in Computing*, ACM, Berlin, Germany, vol. 459, pp. 147–156, 2006.

[8] Q. Shen, R. Jensen. Rough sets, their extensions and applications. *International Journal of Automation and Computing*, vol. 4, no. 3, pp. 217–228, 2007.

[9] X. Peng, Z. Ming, H. Wang. Text learning and hierarchial feature selection in web page classification. In *Proceedings of the 4th International Conference on Advanced Data Mining and Applications*, ACM, Berlin, Germany, vol. 5139, pp. 452–459, 2008.

[10] M. Farhoodi, A. Yari, M. Mahmoudi. A persian web page classifier applying a combination of content-based and context-based features. *International Journal of Information Studies*, vol. 1, no. 4, pp. 263–271, 2009.

[11] S. A. Ozel. A genetic algorithm based optimal feature selection for web page classification. In *Proceedings of International Symposium on Innovations in Intelligent Systems and Applications*, IEEE, pp. 282–286, 2011.

[12] S. Appavu alias Balamurugan, R. Rajaram. Effective and efficient feature selection for large-scale data using Baye′s theorem. *International Journal of Automation and Computing*, vol. 6, no. 1, pp. 62–71, 2009.

[13] J. H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244. 1963.

[14] K. P. Soman, S. Diwakar, V. Ajay. *Insight Into Data Mining*, India: Prentice Hall, 2006.

[15] The 4 Universities data set. [Online], Available: http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/, May 7, 2012.

J. Alamelu Mangai graduated from Annamalai University, India in 2005. She is a Ph. D. candidate of BITS Pilani, Dubai Campus, UAE, and she has been working as a senior lecturer in the Department of Computer Science in BITS Pilani, Dubai Campus.

Her research interests include data mining algorithms, text and web mining.

E-mail: mangai@bits-dubai.ac.ae (Corresponding author)

V. Santhosh Kumar received his Ph. D. degree from Indian Institute of Science, Bangalore, India. He is currently working as assistant professor in BITS Pilani, Dubai Campus, UAE.

His research interests include data mining and performance evaluation of computer systems

E-mail: santhoshkumar@bits-dubai.ac.ae

S. Appavu alias Balamurugan received his Ph. D. degree from Anna University Chennai, Chennai, India. He is currently working as assistant professor, Department of Information Technology at Thiagarajar College of Engineering, Madurai, India.

His research interests include pattern recognition, data mining and informatics.

E-mail: app_s@yahoo.com