Web Page Classification using an Ensemble of Support Vector Machine Classifiers

Shaobo Zhong*
College of Elementary Education, Chongqing Normal University, Chongqing 400700, China
Email: zshaob@163.com

Dongsheng Zou
College of Computer Science, Chongqing University, Chongqing, 400044, China
Email: dszou@cqu.edu.cn

***Abstract*-Web Page Classification (WPC) is both an important and challenging topic in data mining. The knowledge of WPC can help users to obtain useable information from the huge internet dataset automatically and efficiently. Many efforts have been made to WPC. However, there is still room for improvement of current approaches. One particular challenge in training classifiers comes from the fact that the available dataset is usually unbalanced. Standard machine learning algorithms tend to be overwhelmed by the major class and ignore the minor one and thus lead to high false negative rate. In this paper, a novel approach for Web page classification was proposed to address this problem by using an ensemble of support vector machine classifiers to perform this work. Principal Component Analysis (PCA) is used for feature reduction and Independent Component Analysis (ICA) for feature selection. The experimental results indicate that the proposed approach outperforms other existing classifiers widely used in WPC.**

Index-Terms— **Web Page Classification, Support Vector Machine, Ensemble Classifier.**

I. Introduction

With the rapid development of the World Wide Web, the mass of online text data has grown at very fast speed in recent years. Information retrieval is facing great challenge due to the explosion of the network scales. How to obtain useable information from the huge internet raw data automatically and efficiently becomes more and more important than any time before. Researchers have been actively studying on web mining with various data in the World Wide Web. They study various fields such as focused crawler, information extraction, opining mining, usage mining, information integration, social network analysis and so on. Search engines and Web directories are the essential attempts. Actually in each field, classification is one of the methods that organize the subject. Classification is a supervised method of grouping data in a way, that more similar elements come together in the same group, but clustering is an unsupervised method that can find hidden relations among data, which can be used to divide members of a class to even more related clusters. Usually classification is done according to some rules such as latent or obvious analogies among things which are studied. Finding existent pattern is a complicated procedure because these patterns are usually hidden and can not be seen obviously. Therefore, machine learning algorithms are needed for classification. This makes many researchers focus on the issue of WPC technology. WPC can deal with the unorganized data on the web. The purpose of WPC is to classify the Internet web pages into a certain number of pre-defined categories.

During the past two decades, many methods have been proposed for WPC, such as Naive Bayes (NB) classifier [1], self-organization neural networks [2], Support Vector Machine [3], etc. Recently some methods attempt to use some hybrid approach for WPC. For example, Weimin and Aixin [4] used body, title, heading and meta text as feature by using SVM and Naive Bayesian classifier. The result shows that combination of these features with SVM classifier gives higher efficiency for web page classification system. Xin Jin et al. [5] used ReliefF, Information Gain, Gain ratio and Chi Square as feature selection technique for improving the web page classification performance. Rung-Ching and Chung-Hsun [6] proposed a web page classification method by using two types of features as inputs to SVM classification. The output of two SVM is used as inputs of voting schema to determine the category of the web page. The voting improves the performance when compares with the traditional methods. Fang et al. [7] proposed a web page classification by using five classification methods. The output of these SVMs is used as inputs of voting

Manuscript received July 1, 2010; revised January 1, 2011; accepted January 22, 2011.

This work was funded by the Key Project of Chinese Ministry of Science and Technology (No. 2008ZX07315-001), Major scientific and technological special project of Chongqing (No.2008AB5038).

*corresponding author: Shaobo Zhong.

doi:10.4304/jnw.6.11.1625-1630

method and picks the class with the most votes as the final classification result. This method improves the performance when compared with the individual classifiers. Zhang et al. [8] presented a web page categorization based on a least square support vector machine (LS-SVM) with latent semantic analysis (LSA). LSA uses Singular Value Decom-postion (SVD) to obtain latent semantic structure of original term-document matrix solving the polysemous and synonymous keywords problem. LS-SVM is an effective method for learning the classification knowledge from massive data, especially on condition of high cost in getting labeled classical examples. The F-value is 98.2% by using LS-SVM method. Moayed et al. [9] used a swarm intelligence algorithm in the filed of WPC by focusing on Persian web pages. Ant Miner II is the used algorithm. The highest accuracy for News site 1 is 89%. Hossaini et al. [10] used Genetic Algorithm (GA) for classification and clustering. The algorithm works on variable size vectors. At the GA part they combined standard crossover and mutation operators with K-means algorithm for improving diversity and correctness of results. By means of this method they achieved more accurate classes and defined subclasses as clusters. Their method shows more accurate results than fixed size methods. The accuracy rate is about 90.7% and also overload of unnecessary elements in vectors is bypassed.

He et al. [11] used an approach using Naive Bayes (NB) classifier based on Independent Component Analysis (ICA) for WPC. Some other researchers also addressed this problem [13-22]. However, there is significant room for improvement of current approaches.

One particular challenge in training classifiers comes from the fact that the dataset used for WPC is unbalanced [12] to some extent. The number of one kind of web pages can be much smaller or greater than another. Standard machine learning algorithms without considering class-imbalance tend to be overwhelmed by the major class and ignore the minor one and lead to high false negative rate by predicting the positive point as the negative one[23]. However, the accurate classification of web page from the minority class is equivalently important as others. In order to overcome this disadvantage, a common approach is to change the distribution of positive and negative sites during training by randomly selecting a subset of the training data for the majority class. But this approach fails to utilize all of the information available in the training data extracted from the original web pages.

In this paper, a novel approach for WPC is proposed. Our approach uses an ensemble classifier to deal with WPC. The novel approach implements an ensemble of SVM classifiers trained on the ‘‘natural’’ distribution of the data extracted from the original web pages. The ensemble classifier can reduce the variance caused by the peculiarities of a single training set and hence be able to learn a more expressive concept in classification than a single classifier. In addition, PCA algorithm is used for feature reduction and ICA algorithm for feature selection. The experimental results indicate that the proposed approach was indeed providing satisfactory accuracy in web page classification.

This paper is organized as follows. Section II focuses on the method. Section III describes the experiments. The conclusion and future work are discussed in Section IV.

II. Methods

The process of WPC consists of web page retrieval processing, stemming, stop-word filtering, the weight of regular words calculating, feature reduction and selection, and finally the document classification using ensemble classifier. In web page retrieval phase, we will also retrieval the latest news web pages category form the Yahoo.com, and store them in our local databases according to Ref. [4]. In this way out research work can be compared with previous efforts.

A. Web Page Representation

It is difficult to carry on the WPC directly because the words in web documents are huge and complex. In this paper, we extract character words constitutes eigenvector with Vector Space Model (VSM), which is considered as one of most popular model for representing the feature of text contents. In this model, each document is tokenized with a stop-word remover and Porter stemming [24] in order to get feature words used as Eigen values. Finally the documents are projected to an eigenvector, as follow:

$$V(d) = (t_1, w_1(d); t_2, w_2(d); \cdots, t_n, w_n(d)), \quad (1)$$

Where t_i denotes the i-th keyword and $w_i(d)$ is the weight of t_i in document d .

B. Weight calculation

One obvious feature that appears in HTML documents but not in plain text documents is HTML tags. The information derived from different tags bear different importance. For example, a word present in the TITLE element is generally more representative of the document’s content than a word present in the BODY element. So, according to the HTML tags in which the terms are included in, we defined a new method of weight calculation as follows:

$$W_j(d) = \frac{1}{2}\left[\left(W_j(t, \tilde{d})\right) + \left(\Psi(t_j, d_i)\right)\right] \quad (2)$$

where $W(t, \tilde{d})$ is the weight of t in document $\tilde{d}$ according to frequency of words appeared in the HTML documents.

$$W_j(t,\tilde{d}) = \frac{tf(t,\tilde{d}) \times \log(N/n_t + 0.01)}{\sqrt{\sum_{N\in d}\left[tf(t,\tilde{d}) \times \log(N/n_t + 0.01)\right]^2}} \quad (3)$$

where, $tf(t,\tilde{d})$ is the frequency of t in document $\tilde{d}$. N is the number of total documents. And n_t is the number as documents in which i -th keyword appears. $\Psi(t_j, d_i)$ is the location of the words appeared in the HMTL document as following functions

$$\Psi(t_j, d_i) = \sum_{e_k} \left(\partial(e_k) \cdot TF(t_j, e_k, d_i)\right) \quad (4)$$

Where e_k is an HTML element, $\partial(e_k)$ denotes the weight assigned to the element e_k and $TF(t_j, e_k, d_i)$ denotes the number of times term t_j is present in the element e_k of HTML page d_i . We define the function $\partial(e_k)$ as:

$$\partial(e) = \begin{cases} \alpha, & if\ e\ is\ META or\ TITLE \\ 1. & elsewhere \end{cases} \quad (5)$$

where, $\alpha = 2,3,4,5,6$ were tested and compared with standard $TF(t_j, d_i)$ .The experimental results showed that using ensemble classifier can obtain the best results while the value of α equals 6.

C. Feature reduction

The method presenting feature words will generally create multidimensional datasets. PCA is certainly the most widely used method for multivariate statistical analysis. It reduces data dimensionality by performing a covariance analysis between factors. As such, it is suitable for datasets in multiple dimensions. The efficiency of the filter approach of PCA is relatively high. According to the different processing manners, PCA can be divided into data method and matrix method. We choose matrix method, and represent the training sample in the form of document-lemma matrix $\Re = (w_{ij})_{m\times n}$, where covariance is the weight of terms existing in the set of documents. All data which calculated the variance and covariance are represented in matrix. Then, get the eigenvectors of the covariance matrix, which are corresponding to the main component of the original data. We selected the first-used eigenvectors $\xi \le m$, the ξ herein, as eigenvectors is 100, 200, 400, etc. The principal components set is $n \times \xi$ matrix $M = (\ell_{ij})_{n\times\xi}$, where ℓ_{ij} is the eigenvectors being extracted out of the reduced state from original data size $m \times n$ to data size $n \times \xi$. The complete analysis of the PCA method used in this paper is given in Ref. 25 and Ref. 26.

D. Feature selection

Independent Component Analysis (ICA) [27] is a novel statistical signal and data analysis method. The purpose of ICA is to linearly transform the original data into components which are as much as statistically independent [28]. The task of ICA is to find Separation matrix W to make $y = Wx$ where $y = (y_1, y_2, \cdots, y_N)^T$ is called output variable, and $x = (x_1, x_2, \cdots, x_N)^T$ is an observed random variable. If y_i is mutually independent, then y_i is the estimated value of an independent random variable $s = (s_1, s_2, \cdots, s_N)^T$. It can be seen as an extension of PCA towards higher order dependencies.

E. An ensemble of SVM classifiers

Support vector machine

Support vector machine (SVM)classifier, motivated by results of statistical learning theory[29][30], is one of the most effective machine learning algorithms for many complex binary classification problems .Given the training set $T=\{(x_1,y_1),(x_2,y_2),\cdot\cdot;(x_l,y_l)\in(X\times Y)^l\}$ when the penalty factor C and kernel function $K(.,.)$ are selected properly, we can construct a function

$$g(x) = \sum_{i:x\in X_+} \alpha_i K(x,x_i) - \sum_{i:x\in X_-} \alpha_i K(x,x_i) + b, \quad (6)$$

where the non-negative weights α_i and b are computed during training by solving a convex quadratic programming. In order to estimate the probability of an unlabeled input x belonging to the positive class, $P(y = 1 | x)$, we map the value $g(x)$ to the probability by (Platt, 1999)

$$\Pr(y=1|x) = P_{AB}(g(x)) = 1/[1+\exp(A*g(x)+B)] \quad (7)$$

Where A and B are then obtained by solving the optimization problem

$$\underbrace{\min}_{z=(A,B)} F(z) = -\sum_{i=1}^{l} (t_i \log(p_i) + (1-t_i)\log(1-p_i))$$
$$s.t.\ \ ti = \begin{cases} (N_+ + 1)/(N_+ + 2) \ if\, y_i = +1, \\ 1/(N_- + 2) \ \ if\, y_i = -1, \end{cases} \qquad (8)$$
$$pi = P_{A,B}(g(xi)),\ i = 1,2,\cdots l$$

Where N_+ and N_-, respectively, represent the number of positive and negative points in training set. Then the label of the new input x is assigned to be positive if the posterior probability is greater than a threshold, otherwise negative, i.e.

$$f(x) = \begin{cases} 1, & if\ \Pr(y=1|x) > threshold \\ -1. & otherwise \end{cases} \qquad (9)$$

where 1 corresponds to positive class, whereas -1 corresponds to negative class.

An ensemble of SVM classifiers

An ensemble of SVM classifiers is a collection of SVM classifiers, each trained on a subset of the training set (obtained by sampling from the entire training points) in order to get better results [31]. The prediction of the ensemble of SVMs is computed from the prediction of the individual SVM classifier, that is, during classification, for a new unlabeled input x_{test} ,the j -th SVM classifier in the collection returns a probability $P_j(y=1|x_{test})$ of x_{test} belonging to the positive class, where $j = 1,2,\cdots m$ and m is the number of SVM classifiers in the collection. The ensemble estimated probability, $P_{Ens}(y=1|x_{test})$, is obtained by

$$P_{Ens}(y=1|x_{test}) = (1/m) \times \sum_{j=1}^{j=m} P_j(y=1|x_{test}) \quad (10)$$

Fig.1 shows the architecture of the ensemble of SVM classifiers.

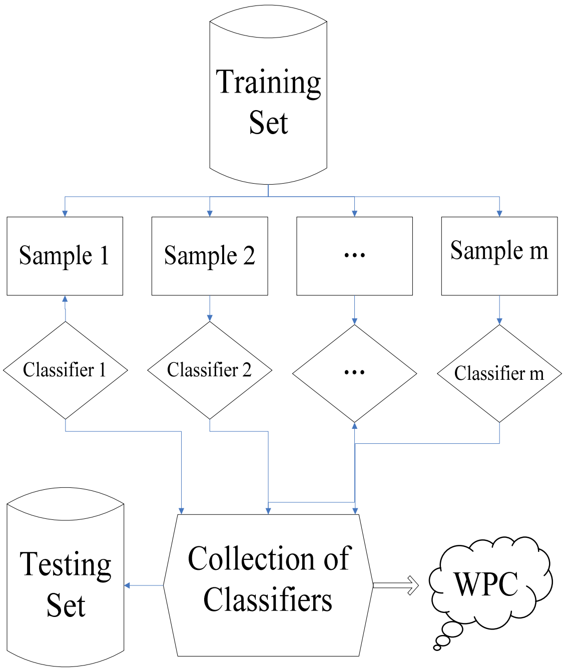

Figure 1. Architecture of the ensemble classifier fusing m SVM classifiers. Each one is trained on a balanced subsample of the training data.

III. Experimental Results and Discussion

For experimental purpose, we build the dataset in the similar way as He et al. [11]. We choose the web page dataset from the Yahoo sports news. The dataset includes six categories of web pages. They are Soccer, NBA, Golf, Tennis, Boxing and NFL. The whole set include 3,160 web pages, i.e.880 documents of Soccer, 560 documents of NBA, 320 documents of Golf, 640 documents of Tennis,280 documents of Boxing, 480 documents of NFL. Among the dataset, 2500 documents (about 80%) selected randomly from different classes were used for training data, and the remaining other document for test data.

As for performance measure, the standard information retrieval measures, such as recall (r), precision (p), and F1 ($F1 = 2rp/(r+p)$) are used to estimate the performance of our method. To compare with other approaches, we have done the classification on the same dataset by using TFIDF, NB classifier and He's improved NB (denoted as NBICA)[11] .

The experimental results of WPC on our dataset are shown in Table 1. For the category of Soccer, NBA, Golf, Tennis, Boxing and NFL, the value of F1 are 91.55%, 92.97%, 94.40%, 92.50%, 94.55% and 93.87%, respectively. Meanwhile, the overall average of F1 measure is 93.31%. Comparing with NBICA, the overall F1 value is increased modestly from 92.13 to 93.31% by using our approach. In addition, the F1 value for each category is relatively stable with our approach. However, the lowest F1 value is 75.85% for Soccer category while the highest one is 98.81% for rugby category with NBICA. The F1 value varies evidently because the sizes for each category of web pages are unbalanced with NBICA. As observed from Table1, we can summarize that this problem is solved with our approach by using an ensemble of SVM classifiers.

For comparison we used some other methods, such as TFIDF [32], NB and NBICA for WPC on the same dataset. The experimental results of WPC are shown in Table 2. By using TFIDF, NB and NBICA methods, the overall average F1 value are 81.78, 84.04 and 89.63%, respectively. Our method of ensemble classifier improves F1 by 3-11%. These results indicate the superior performance of our approach over that of some existing methods for WPC.

TABLE 1.

EXPERIMENTAL RESULT USING ENSEMBLE CLASSIFIER.

Class No.	Recall (%)	Precision (%)	F1 (%)
1.Soccer	90.36	92.78	91.55
2.NBA	95.66	90.42	92.97
3.Golf	96.27	92.6	94.40
4.Tennis	94.50	90.58	92.50
5.Boxing	95.68	93.45	94.55
6.NFL	95.45	92.35	93.87
Average	94.65	92.0	93.31

TABLE 2.

F1 VALUE BY USING DIFFERENT APPROACHES

Class No.	TFIDF (%)	NB (%)	NBICA (%)	Ensemble classifier (%)
1.Soccer	84.32	85.85	90.25	91.55
2.NBA	83.44	93.56	93.68	92.97
3.Golf	74.37	76.30	84.56	94.40
4.Tennis	85.60	85.81	93.56	92.50
5.Boxing	80.16	78.69	83.40	94.55
6.NFL	82.76	84.05	92.30	93.87
Average	81.78	84.04	89.63	93.31

IV. CONCLUSION

Automated web pages classification, which is a challenging research direction in text mining, plays an important role to establish the semantic web. Many efforts have been made for WPC. However, there is significant room for improvement of current approaches. One particular challenge in training classifiers comes from the fact that the dataset used for WPC is unbalanced to some extent. Consequently, the F1 value of most existing methods is unstable. In this article, we have studied the problem of unbalanced dataset in WPC. We proposed a novel approach using an ensemble of SVM classifiers to address this problem. The comparison of performance among four methods, namely TFIDF, NB, NBICA and our ensemble classifier, has been presented in this paper. The experimental results indicate that the proposed approach could solve the problem well. Moreover, the F1 value is increased modestly with our approach.

In future research, we should address to increase the number of categories to a large extent to observe the F1 value with our approach. Moreover, combined with some existing algorithms, such as Genetic algorithm, our method of ensemble classifier can be further improved.

ACKNOWLEDGMENTS

The authors thank the editor and referees for their careful review and valuable critical comments. We also thank Prof. He for valuable suggestions and comments. This work is supported by the Key Project of Chinese Ministry of Science and Technology (No. 2008ZX07315-001), Major scientific and technological special project of Chongqing (No.2008AB5038)， Education project of Chongqing Normal University(080201), The Chongqing Key Research Base of Humanities and Social Sciences: the Financial Support from Chongqing Research Center of Elementary Teacher Education. The authors are grateful for reviewers who made constructive comments.

REFERENCES

[1] Fan Y., Zheng C., Wang Q. Y., Cai Q. S., Liu J. Web Page Classification Based on Naive Bayes Method (In Chinese), Journal of Software, 2001, pp. 1386-1392.

[2] Zhang Y. Z. The Automatic Classification of Web Pages Based on Neural Networks. Neural information processing, ICONIP2001 Proceedings, Shanghai, China, 14-18 November 2001, Vol.2, pp. 570- 575.

[3] Xue W. M., Bao H., Huang W. T., Lu Y. C. Web Page Classification Based on SVM. Intelligent Control and Automation, 21-23 June 2006, vol.2, pp. 6111- 6114.

[4] W. Xue, H. Bao, W. Huan, and Y. Lu, “Web Page Classification Based on SVM,” 6th World Congress on Intelligent Control and Automation, Dalian, China, 2006, pp. 6111-6114,.

[5] J. Xin, L. Rongyan, S. Xian, and B. Rongfang, “Automatic Web Pages Categorization with ReliefF and Hidden Naive Bayes,” Proceedings of the 2007 ACM symposium on Applied computing, Seoul, Korea, 2007, pp. 617-621.

[6]Chen R., Hsieh C., and Chen H. Web Page Classification

Based On A Support Vector Machine Using A Weighted Vote Schema. Expert Systems with Applications, 2006, vol. 31, pp. 427-435.

[7] Rui F., Alexander M., and Babis T. A Voting Method for the classification of Web Pages. IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, 2006, pp. 610-613.

[8]Zhang Y., Fan B., Xiao L. B. Web Page Classification Based-on A Least Square Support Vector Machine with Latent Semantic Analysis. Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008, pp. 528-532.

[9] Moayed M. J.; Sabery, A. H.; Khanteymoory, A. Ant Colony algorithm for Web Page Classification. 2008 International Symposium on information technology Kuala Lumpur, Malaysia, 26-29 August 2008, pp. 8-13.

[10]. Hossaini, Z Rahmani, A. M. Setayeshi, S. Web pages classification and clustering by means of genetic algorithm: a variable size page representing approach. 2008 International conference on Computational Intelligence for Modeling Control & Automation (CIMCA 2008), 10-12 December 2008, pp. 436-440.

[11] He Z. L., Liu Z. J. A Novel Approach to Naïve Bayes Web Page Automatic Classification. Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 2008, pp. 361-365.

[12] Japkowicz N. The class imbalance problem: significance and strategies. In: IC-AI'2000, Special Track on Inductive Learning Las Vegas, Nevada, 2000.

[13] Xu S.M., Wu B.,Ma C.. Efflcient SVM Chinese Web page classifier based on pre-classification. Computer Engineering and Applications, 2010, pp. 125-128.

[14] Araujo L.,Martinez R.J. Web Spam Detection: New Classification Features Based on Qualified Link Analysis and Language Models. IEEE Transaction on information forensics and security, 2010, Vol. 5 (3), pp. 581-590.

[15] Chen T.C., Dick, S., Miller, J. Detecting Visually Similar Web Pages: Application to Phasing Detection. ACM transaction on Internet techonology. 2010, Vol.10 (2), pp. 5

[16] Ofuonye E., Beatty P., Dick S.. Prevalence and classification of web page defects. Online Information Review, 2010, Vol. 34 (1), pp.160-174.

[17] Golub K., Lykke M. Automated classification of web pages in hierarchical browsing. Journal of documentation, 2009, Vol. 65 (6), pp. 901-925.

[18] Hou C.Q., Jiao L.C. Graph based Co-training algorithm for web page classification. Acta Electronica Sinica, 2009, pp.2173-80.

[19] Farhoodi, M., Yari A., Mahmoudi M. A Persian Web Page Classifier Applying a Combination of Content-Based and Context-Based Features.International Journal of Information Studies, 2009,pp.263-71.

[20] Selamat A., Subroto I.M.I., Choon C. Arabic script web page language identification using hybrid-KNN method.International Journal of Computational Intelligence and Applications,2009,pp.315-43.

[21] Zhu Z.G., Deng C.S., Kong L.P. Algorithm research on classifying Web users navigation patterns based on N-gram .Journal of the China Society for Scientific and Technical Information,2009,pp.389-394.

[22] Peng X.G., Ming Z., Wang H.T. WordNet based Web page classification system with category expansion. Journal of Shenzhen University Science & Engineering, 2009, pp.118-122.

[23] Liu, X. Y., Zhou, Z. H.. The influence of class imbalance on cost-sensitive learning: an empirical study. In: Sixth IEEE International Conference on Data Mining (ICDM'06), Hong Kong, 2006.

[24] The Porter Stemming algorithm, http://www. tartarus.org/~martin/PorterStemmer.

[25] Calvo R. A., Partridge M., Jabri M.. A comparative study of principal components analysis techniques. In Proceedings 9th Australian Conference on Neural Networks, Brisbane, QLD1998, pp. 276-281.

[26] Selamat, A., Omatu, S. Neural Networks for Web News Classification Based on PCA. Proceedings of the International Joint Conference, 20-24 July 2003, vol. 3, pp. 1792 - 1797.

[27] Hyvarinen A., Karhunen J., and Oja E., 2001. Independent Component Analysis, Wiley-Interscience, New York.

[28] Nacim F. C., Bernard R., Nathalie A.G. A Comparison of Dimensionality Reduction Techniques for Web Structure Mining. IEEE/WIC/ACM International Conference on, 2-5 Nov. 2007, pp. 116 – 119.

[29] Vapnik V., 1995. The Nature of Statistical Learning Theory. Springer, New York.

[30] Vapnik V., 1998. Statistical Learning Theory. Wiley, New York.

[31] Dietterich, T.G., 2000. Ensemble methods in machine learning. In: Lecture Notes in Computer Science, vol. 1857, pp. 1–15.

[32] Yang J. P., Honavar V., Miler L. Mobile intelligent agents for document classification and retrieval: a machine learning approach. Proceeding of the Eurpoean Symposium on cybemetics and Systems Research, Vienna, Austria, 1998, pp.707-712.

Shaobo Zhong was born in Sichuan, P.R. China, in January 24, 1973. He obtained the bachelor's the master's degree in Mathematics and Computer Science of the Chongqing Normal University, China in 1998, and the doctor's degree in College of Computer Science of the Chongqing University, China in 2008.His research interest includes machine learning, data mining and web page classification.