http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 11 (1): 148-153, 2012
ISSN 1812-5638 / DOI: 10.3923/itj.2012.148.153

A Novel Feature Selection Framework in Chinese Term Definition Extraction

[1,2]Xu Pan, [1,2]Hong-Bin Gu and [1,2]Zhi-Qing Zhao
[1]College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing, China
[2]Engineering Research Center for Flight Simulation and Advanced Training, Nanjing, China

Abstract: In this study, a novel feature selection framework was proposed for extracting definition from aviation professional corpus. The framework combined between-class and within-class distribution difference to express contribution of small disjuncts in classification and increase precision and efficiency of definition extraction. In this study, definition of the framework was introduced and influence of feature selection made by the framework and traditional methods was compared. Results using traditional methods and the proposed framework in different classification strategy were contrasted in the end. The results indicated that the framework introduced in this paper is better than traditional methods in extraction definitions.

Key words: Feature selection, unbalanced corpus, definition extraction, text categorization, small disjunct

INTRODUCTION

In the past few years the automatic extraction of definitions from textual data became a common research area in Natural Language Processing (NLP). This technology was widely used in information extraction, text mining, ontology development and e-learning (Walter and Pinkal, 2006; Aguilar and Sierra, 2009; Alarcon *et al*., 2009). In Computer Based Training (CBT) of aircraft industry, definition extraction is a novel approach to collect professional knowledge for knowledge base and ontology base (Pan *et al*., 2010).

A number of feature selection metrics have been explored, such as Information Gain (IG), Chi-square (CHI) and Mutual Information (MI) (Sergios and Konstantinos, 2007; Tenenbaum *et al*., 2000). There are two disadvantages of these methods. (1) They tend to choose high-frequency words in corpus which will lead minority class to be submerged by majority class. (2) They cannot express the contribution of small disjunct within-class in classification.

Traditional methods select feature using inherent distribution of data that is independent of subsequent learning algorithm and has the best computation efficiency. Recent research has focused on selecting rich-information feature and tended to apply features more between-class differences or balance the importance of different class by adjusting the weight of different class in formula (Li and Zong, 2005; How and Narayanan, 2004). For holding rich classification information features, some other methods were proposed. The (categorical descriptor term) CTD (How and Narayanan, 2004) method combines inverse document frequency with inverse category frequency. The (strong class information words) SCIW (Li and Zong, 2005) method tends to retain strong class information words. Christy and Tharnbidurai (2006) introduced a method for automatically extracting key elements from a collection of text documents by extracting a set of features using Genetic algorithms.

In response to feature selection in imbalanced classification, Dunja Mladenic's work shown that among IG, cross entropy, MI, (Term Frequency) TF, (the Weight of Evidence of Text) WET. IG achieved the worst result because IG preferred to high-frequency feature (Mladenic and Grobelnik, 1999). Zheng *et al.* (2004, 2003) proposed a nearly optimal method combined positive and negative features based on existing algorithms. It performed better than traditional methods when the algorithms found two types of features accurately. Researchers in Hong Kong proposed a method called weighed frequency and odds (WFO) that combined the two measurements with trained weights. The experimental results on data sets from both topic-based and sentiment classification tasks shown the method was robust (Li *et al*., 2009). Zheng and Srihari (2003) proposed a feature selection strategy combined with one-sided metrics, two-sided metrics, positive and negative features. The experiment proved that the two-sided metrics performed well with balanced data and after mixed positive and negative features in appropriate proportion it could performed well with imbalanced data.

Some researchers also proposed new feature selection framework not only for balanced data classification but also for the case of imbalanced data. Hongfang *et al.* (2009) archived a Category Distribution-based Feature Selection (CDFS) framework by selecting

Corresponding Author: Xu Pan, College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing, China

features that have strong discriminative power using distribution information of features and assigning weights flexibly to categories. Xu *et al.* (2008) developed feature selection function KG using quantitative category discrimination ability. Cui *et al.* (2007) defined a feature selection algorithm by theory of Markov Blanket and Chi-Square test which obtain an approximate optimal feature subset. In this study, definition extraction is regarded as two-class imbalanced classification and a novel feature selection framework is proposed. The framework takes into account the characteristic of within-class and between-class distribution difference of features and reflects the contribution of different feature distribution between sub-concepts within definition in classification. A balanced-level parameter is also defined according to sample distribution in training set to adapt to various imbalanced data classification strategy.

FRAMEWORK FOR FEATURE SELECTION FUNCTION

MI considers information of low frequency words but sometimes it also aggrandizes the effect of low frequency words and ignores the impact of different category distribution. IG counts in the information provided by features for classification but it does not distinguish positive correlation or negative correlation between features and category. Generally speaking, in imbalanced data classification, MI results in poor precision because of its sensitivity to low frequency words, IG induces poor recall because it tends to retain positive features and there are a mass of negative features in data set. Chi-Square test has the same matter with IG.

When extracting definitions, small disjuncts is another problem besides imbalanced data set. Small disjuncts, also known as within-class imbalance, is a relative conception of between-class imbalance. It represents the imbalanced distribution of training samples among sub concepts in one category (Lin *et al.*, 2008). Feng (1997) divides term definition into six types and four modes. Each one of these types and modes has particularly syntactic and lexical patterns and could be regarded as the small disjuncts in definitions. In this study, these problems are solved by integrating within-class distribution difference and between-class distribution difference of features into a unitary framework.

Definition of feature selection framework: The framework proposed in this study is composed of three segments.

The between-class distribution difference of feature t is used to measure the difference of occurrence frequency of t between category A and category B. It is defined as follow:

Definition 1: Between-class distribution difference of feature t within category C_i could be denoted as:

$$BC_{(i,t)} = \log \frac{E\left\{\left[TF_{(i,t)} - E\left(TF_{(i,t)}\right)\right]^2\right\}}{\bar{X}} = \log \frac{\sum_{i=1}^{n} TF_{(i,t)}^2 - n \times \bar{X}^2}{n \times \bar{X}} \quad (1)$$

where, $TF_{(i,t)}$ is occurrence frequency of t in category C_i and:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} TF_{(i,t)}$$

is mean value of occurrence frequency of t in sample space.

The within-class distribution difference of feature t is used to measure the difference of occurrence frequency of t among sub-concept within different categories. It is defined as follow:

Definition 2: Within-class distribution difference of feature t within category C_i could be denoted as:

$$WC_{(i,t)} = \log \frac{stdF_{i(d,t)}}{meanF_{i(d,t)}} = \log \frac{\sqrt{\frac{\sum_{i=1}^{n}\left(F_{(d,t)} - meanF_{i(d,t)}\right)^2}{n-1}}}{\frac{1}{n}\sum_{dei,d=1}^{n} F_{(d,t)}} = \log \frac{\sqrt{(n-1)\sum_{i=1}^{n}\left(F_{(d,t)} - meanF_{i(d,t)}\right)^2}}{(n-1)\sum_{dei,d=1}^{n} F_{(d,t)}} \quad (2)$$

where, $TF_{(d,t)}$ is occurrence number of t in text d of category C_i:

$$F_{(d,t)} = \frac{TF_{(d,t)} + 1}{|d|}$$

is occurrence frequency of t in text d and the mean value of occurrence frequency of t in category C_i is:

$$meanF_{i(d,t)} = \frac{1}{n}\sum_{dei,d=1}^{n} F_{(d,t)}$$

$$stdF_{i(d,t)} = \sqrt{\frac{\sum_{i=1}^{n}(F_{(d,t)} - meanF_{i(d,t)})^2}{n-1}}$$

is the standard deviation of occurrence frequency of t in category C_i.

For the purpose of adapting to various policies in extracting definitions, we define a weighting coefficient for the feature selection framework.

Definition 3: The weighting coefficient of balance degree could be denoted as:

$$Weight_d = (W_p - W_i)^2 \quad (3)$$

where, W_p is the ratio of occurrence number of feature t in category C_i to the occurrence number of t in other categories of training data and W_i is the ratio of instance number of category C_i to instance number of other categories in training data.

Definition 4: The feature selection framework function could be denoted as:

$$\begin{aligned} FS_{(t)} &= Weight_d \times \sum_{i=1}^{n}(BC_{(i,t)} \times WC_{(i,t)}) \\ &= (W_p - W_i)^2 \times \\ &\sum_{i=1}^{n}(\log\frac{\sum_{i=1}^{n} TF_{i,t}^2 - n\times\bar{X}^2}{n\times\bar{X}} \times \log\frac{\sqrt{(n-1)\sum_{i=1}^{n}(F_{(d,t)} - meanF_{i(d,t)})^2}}{(n-1)\sum_{dei,d=1}^{n} F_{(d,t)}}) \end{aligned} \quad (4)$$

Specially, in 2-category classification, the $BC_{(i,t)}$ of different category is same because the same variance of two category and $FS_{(i,t)}$ retrogress to a function taking into account within-class distribution difference only. Under the circumstances, another $BC_{(i,t)}$ could be denoted as follow.

Definition 5: In 2-category classification, Between-class distribution difference of feature t within category C_i could be denoted as:

$$BC_{(i,t)} = \frac{TF_{(i,t)}}{\bar{X}} \quad (5)$$

where, $TF_{(i,t)}$ is occurrence frequency of t in class C_i and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} TF_{(i,t)}$ is mean value of occurrence frequency of t in sample space.

Characteristic of feature selection framework: Definition 4 integrates characteristic of two distribution difference into unique formula. Relative to the formula only reckon between-class distribution difference in, features with higher within-class distribution difference obtain higher value in Eq. 4 and is separated from other features. Framework value is summated after calculated in different category and the contribution of both positive features and negative features is reflected in this way.

The distribution of feature values in training data of Balanced Random Forests (BRF) method is shown in Fig. 1. The feature values in Fig. 1a are calculated using framework $FS_{(t)}$, in Fig. 1b are calculated using MI. The distribution using other traditional feature selection methods is similar in drawing to Fig. 1b.

In Fig. 1b, the majority of features in training data are distributed densely over the bottom of parabola. As a result, occurrence number of these features in positive instances and negative instances is very close and between-class distribution difference of these features is not obvious. Specially, it happens frequently in Fig. 1b that one function value corresponding to more than one feature and it is difficult to filter features in the situation.

In Fig. 1a, because of introducing (2) into feature selection framework, the features with outstanding within-class distribution difference will obtain higher framework values than in Fig. 1b and are separated from the bottom of parabola. The probability of the phenomenon that one framework value is corresponding to more than one

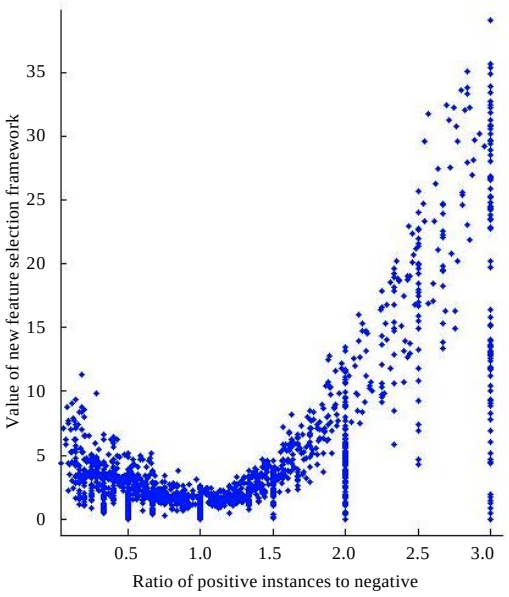

Fig. 1(a): Features distribution in training set of BRF (the new framewok)

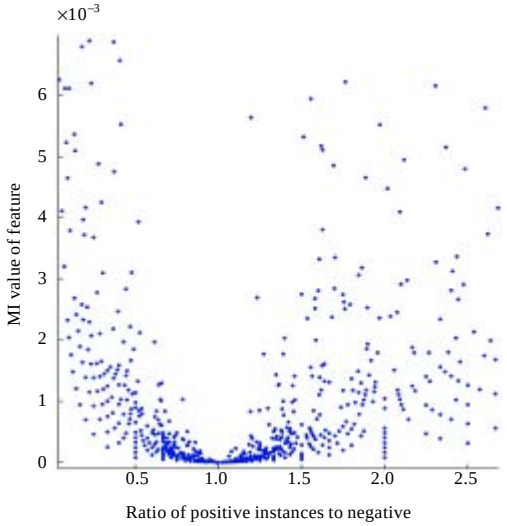

Fig. 1(b): Features distribution in training set of BRF (MI)

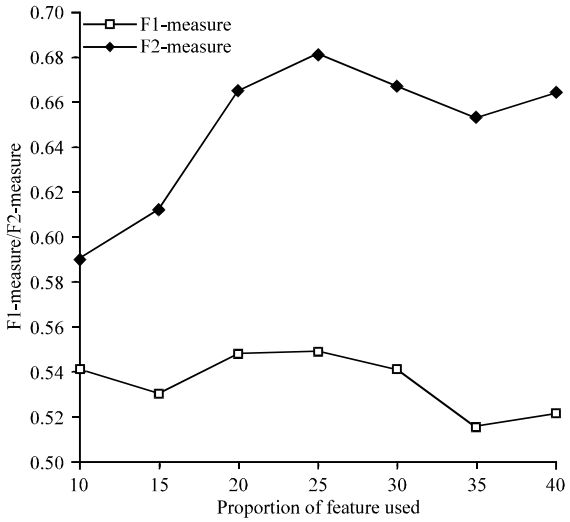

Fig. 2: Results of single C4.5 tree in BRF

feature will be reduced in this situation and this framework is more effective than traditional feature selection function in filtering features.

EXPERIMENTS

Experiment setting: Corpus used in this paper is collected from the Chinese aerospace professional literatures, a total of 16627 sentences which contains 1359 definition sentences. Two different classification strategies are used in this paper, one is improved BRF purposed by Pan *et al*. (2010), the other is Naïve Bayes with 10-cross validation. The two classifier are both implemented using WEKA (Waikato Environment for Knowledge Analysis).

Evaluation method used in this paper include recall, precision and F-measure. In the improved BRF experiment, because the ratio of positive instances to negative instances is 1 to 1, the weight W_i in Eq. 3 is set to 1. In the SVM experiment, positive instance accounted for 8.17% of total instance, so the weight W_i is set to:

$$W_i = \frac{8.17\%}{1-8.17\%} = 0.89$$

Experiment results: Figure 2 shows the average results of single C 4.5 trees used in BRF. The F1 and F2 results of C4.5 vary with different number of features used and reached highest value when the features used of total number of 20 to 30%. After reaching the highest value of F1 = 0.549 and F2 = 0.680, performance of C4.5 tree begin to decline.

As a result, the proportion of features selected using framework $FS_{(t)}$ in BRF experiment is 20 to 40%. Figure 3 is

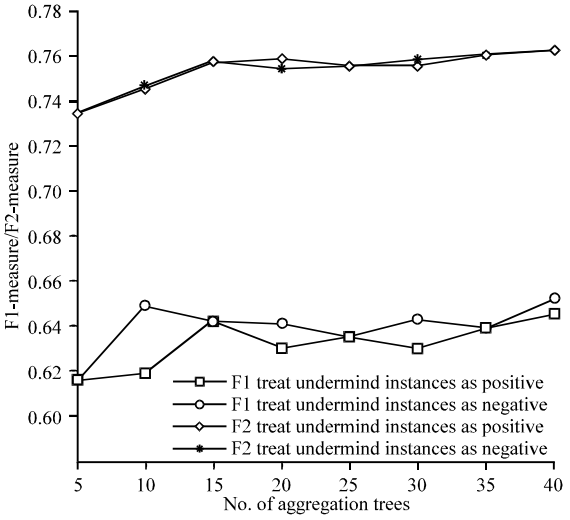

Fig. 3: Results of BRF using 25% features

the best results. As the number of aggregation tree increased the F1 and F2 results improve steadily. After the number of aggregation tree reached 15, F2 rise moderately up to 0.761, F1 remain fluctuant and reach the best score at 0.652.

Figure 4 shows the results of Naïve Bayes with 10-cross validation on original imbalanced data set using different feature selection methods. The classifier using $FS_{(t)}$ gets the best results in both F1-measure and F2-measure are when proportion of features used in experiments is more than 10%.

The F1-measure of $FS_{(t)}$ starts with F1-measure at 0.363, then fluctuates from 0.348 to 0.376 gently. MI starts

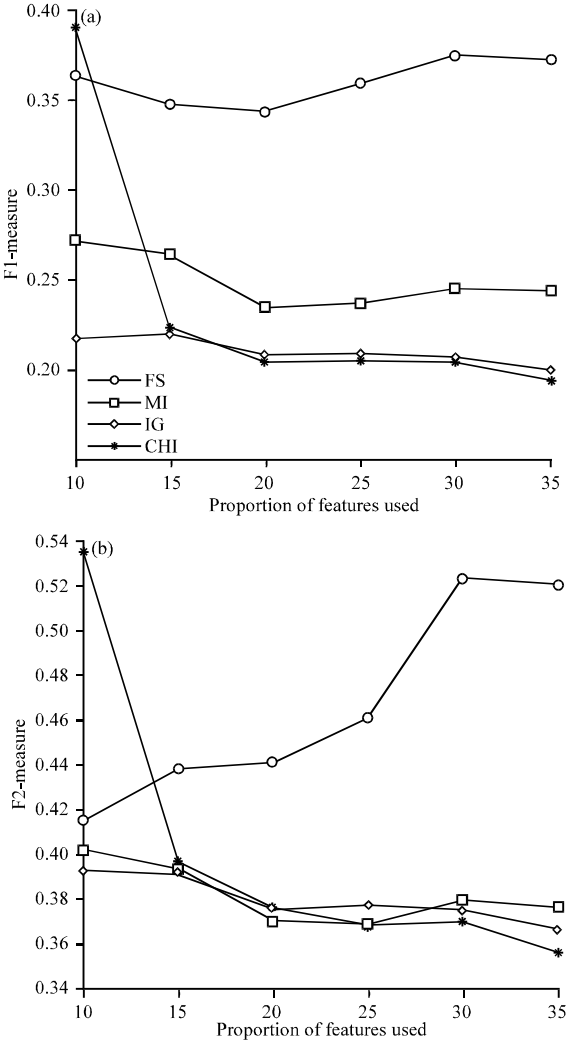

Fig. 4(a-b): (a) F1-measure and (b) F2-measure of Naïve Bayes

with F1-measure at 0.272 and declines to 0.235 when proportion of features used is 20%, then roses to 0.245. IG starts with F1-measure at 0.217 and reaches 0.223 when proportion of features used is 15%, then declines to 0.2. CHI starts with a high score at 0.39 but declines quickly to 0.223 when proportion of features used is 15%, then slowly down to 0.194.

The F2-measure of $FS_{(t)}$ improves from 0.418 to 0.52 substantially when the ratio of features used less than 30% and then remains stable. MI starts at 0.402 and declines to 0.368 when proportion of features used is 25%, then roses to 0.379. IG starts at 0.392, then declines slowly to 0.366. CHI starts with a high score at 0.535 but declines quickly when proportion of feature used increases and reaches 0.356 when proportion of features used is 35%.

It can be seen from the results, in experiments using balanced training data set for single C4.5 tree, contrast to traditional feature selection methods by Pan *et al*. (2010), F1-measure increase from 48 to 55% and F2-measure increase from 64.7 to 68.1% using framework $FS_{(t)}$. In experiment using BRF strategy, the best result of framework $FS_{(t)}$ is almost equal with that of traditional feature selection methods. However, the features required to archive the best results decrease from between 30 and 40% to between 20 and 30% and training time of single C4.5 tree decrease from an average of 520 to 330 sec. In experiments using Naïve Bayes classifier, with the increasing of number of selected features, the results of classifier using framework $FS_{(t)}$ surpass other classifiers.

For training data of single classifier, 75 to 80% of the features are distributed in the bottom area of the parabola in Fig. 1 and features on both sides of parabola with higher between-class distribution difference contribute less to classification. Therefore, in Fig. 1b, feature selection methods have to go into the midst of the bottom area of the parabola to obtain more effective features. In Fig. 1a, the bottom of the central region of parabola is uplifted to improve the value of features and compared to Fig. 1a, the framework obtains enough features without go into the midst of the bottom area of the parabola. As a result, compared to traditional feature selection methods, the framework proposed in this paper could achieve better results with less cost.

CONCLUSIONS

In this study, a Novel feature selection framework combined two types of distribution difference of features was proposed. Experiments show that the framework reflects the contribution of both between-class distribution difference and within-class distribution difference and solve the small disjunct problem in definition extraction to a certain extent. For different classification strategy, experiments using the framework achieve better results than experiments using traditional feature selection methods; meanwhile, the framework can be applied to different classification strategy by simply setting the weight.

Further research about this issue can focus on how to calculate the two types of distribution difference more effectively. The recent emergence of some new feature selection method is also very worthy of study.

REFERENCES

Aguilar, C. and G. Sierra, 2009. A formal scope on the relations between definitions and verbal predications. Proceedings of the 1st Workshop on Definition Extraction, September 18, 2009, Borovets, Bulgaria, pp: 1-6.

Alarcon, R., G. Sierra and C. Bach, 2009. Description and evaluation of a definition extraction system for Spanish language. Proceedings of the 1st Workshop on Definition Extraction, September 18, 2009, Borovets, Bulgaria, pp: 7-13.

Christy, A. and P. Thambidurai, 2006. Efficient information extraction using machine learning and classification using genetic and C4.8 algorithms. Inform. Technol. J., 5: 1023-1027.

Cui, Z.F., B.W. Xu, W.F. Zhang and J.L. Xu, 2007. An approximate markov blanket feature selection algorithm. Chin. J. Comput., 30: 2074-2081.

Feng, Z.W., 1997. An Introduction to Modern Terminology. Language and Culture Press, China.

Hongfang, J., W. Bin, Y. Yahui and X. Yan, 2009. Category distribution-based feature selection framework. J. Comput. Res. Dev., 46: 1586-1593.

How, B.C. and K. Narayanan, 2004. An empirical study of feature selection for text categorization based on term weightage. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, September 20-24, 2004, IEEE, Washington, DC., USA., pp: 599-602.

Li, S. and C. Zong, 2005. A new approach to feature selection for text categorization. Proceedings of the IEEE International Conference on Natural Language Processing and Knowledge Engineering, October 30-November 1, 2005, Beijing University of Posts and Telecommunications Press, Beijing, China, pp: 626-630.

Li, S., R. Xia, C. Zong and C.R. Huang, 2009. A framework of feature selection methods for text categorization. Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, August 2-7, 2009, Suntec, Singapore, pp: 692-700.

Lin, Z.Y., Z.F. Hao and X.W. Yang, 2008. Current state of research on imbalanced data sets classification learning. Appl. Res. Comput., 25: 332-336.

Mladenic, D. and M. Grobelnik, 1999. Feature selection for unbalanced class distribution and naive bayes. Proceedings of the 16th International Conference on Machine Learning, June 27-30, 1999, Morgan Kaufmann, San Francisco, CA., USA., pp: 258-267.

Pan, X., H.B. Gu and Z.Q. Zhao, 2010. Research on definition extraction based on over-sampling using distance distribution information of instances. Proceeding of the Chinese Conference on Pattern Recognition, October 21-23, 2010, Chongqing, China, pp: 168-173.

Sergios, T. and K. Konstantinos, 2007. Pattern Recognition. 3rd Edn., Academic Press, London, UK.

Tenenbaum, J.B., vin de Silva and J.C. Langford, 2000. A global geometric framework for nonlinear dimensionality reduction. Science, 290: 2319-2323.

Walter, S. and M. Pinkal, 2006. Automatic extraction of definitions from german court decisions. Proceedings of the Workshop on Information Extraction Beyond the Document, July 22, 2006, Sydney, pp: 20-28.

Xu, Y., J.T. Li, B. Wang and C.M. Sun, 2008. A category resolve power-based feature selection method. J. Software, 19: 82-89.

Zheng, Z. and R. Srihari, 2003. Optimally combining positive and negative features for text categorization. Proceedings of the ICML Workshop on Learning from Imbalanced Data Sets, (LIDS'03), Eashington, DC., USA., pp: 1-8.

Zheng, Z., R. Srihari and S. Srihari, 2003. A feature selection framework for text filtering. Proceedings of the 3rd IEEE International Conference on Data Mining, November 19-22, 2003, IEEE Computer Society, Washington, DC., USA., pp: 705-708.

Zheng, Z., X. Wu and R. Srihari, 2004. Feature selection for text categorization on imbalanced data. Proc. ACM SIGKDD Explorations Newslett., 6: 80-89.