Journal of Zhejiang University SCIENCE
ISSN 1009-3095
http://www.zju.edu.cn/jzus
E-mail: jzus@zju.edu.cn

An improved TF-IDF approach for text classification*

ZHANG Yun-tao (张云涛)[1,2], GONG Ling (龚 玲)[2], WANG Yong-cheng (王永成)[2]

([1]*Network & Information Center,* [2]*School of Electronic & Information Technology, Shanghai Jiaotong University, Shanghai 200030, China*)

E-mail: ytzhang@mail.sjtu.edu.cn; lgong@mail.sjtu.edu.cn; ycwang@mail.sjtu.edu.cn

Received Dec. 5, 2003; revision accepted June 26, 2004

Abstract: This paper presents a new improved term frequency/inverse document frequency (TF-IDF) approach which uses confidence, support and characteristic words to enhance the recall and precision of text classification. Synonyms defined by a lexicon are processed in the improved TF-IDF approach. We detailedly discuss and analyze the relationship among confidence, recall and precision. The experiments based on science and technology gave promising results that the new TF-IDF approach improves the precision and recall of text classification compared with the conventional TF-IDF approach.

Key words: Term frequency/inverse document frequency (TF-IDF), Text classification, Confidence, Support, Characteristic words

doi:10.1631/jzus.2005.A0049 **Document code:** A **CLC number:** TP31

INTRODUCTION

The widespread and increasing availability of text documents in electronic form increases the importance of using automatic methods to analyze the content of text documents, because the method using domain experts to identify new text documents and allocate them to well-defined categories is time-consuming and expensive, has limits, and does not provide continuous measure of the degree of confidence with which the allocation was made (Olivier, 2000). As a result, the identification and classification of text documents based on their contents are becoming imperative.

The classification can be done automatically by separate classifiers learning from training samples of text documents. The main aim of the classifier is to obtain a set of characteristics that remain relatively constant for separate categories of text documents and to classify the huge number of text documents into some particular categories (or folders) containing multiple related text documents.

In text classification, a text document may partially match many categories. We need to find the best matching category for the text document. The term (word) frequency/inverse document frequency (TF-IDF) approach is commonly used to weigh each word in the text document according to how unique it is. In other words, the TF-IDF approach captures the relevancy among words, text documents and particular categories.

We put forward the novel improved TF-IDF approach for text classification, and will focus on this approach in the remainder of this paper, and will describe in detail the motivation, methodology, and implementation of the improved TF-IDF approach. The paper discusses and analyzes the relationship among confidence, support, recall and precision, and then presents the experimental results.

IMPROVED TF-IDF APPROACH

Text classification can be effected by various learning approaches of classifier, such as *k*-nearest neighbor (Sun *et al*., 2001), decision tree induction,

* Project (No. 60082003) supported by the National Natural Science Foundation of China

naïve Bayesian (Fan *et al*., 2001), support vector machine (Huang and Wu, 1998; Larry and Malik, 2001) and latent semantic index (Lin *et al*., 2000). Some of these techniques are based on, or correlated with, the TF-IDF approach representing text with vector space in which each feature in the text corresponds to a single word.

VSM assumes that a text document d_i is represented by a set of words (t_1, t_2, …, t_n) wherein each t_i is a word that appears in the text document d_i, and n denotes the total number of various words used to identify the meaning of the text document. Word t_i has a corresponding weight w_i calculated as a combination of the statistics $TF(w_i, d_i)$ and $IDF(w_i)$. Therefore, d_i can be represented as a specific n-dimensional vector $\boldsymbol{d}_i$ as

$$\boldsymbol{d}_i=(w_1, w_2, \ ..., w_n) \tag{1}$$

Weight is the measure that indicates the statistical importance of corresponding words. The weight w_i of word t_i can be determined by the value of $TF(w_i, d_i)$*$IDF(w_i)$. The *TF* value is proportional to the frequency of the word in the document and the *IDF* value is inversely proportional to its frequency in the document corpus. The function encodes the intuitions that: (1) The more often a word occurs in a document, the more it is representative of the content of the text; (2) The more text the word occurs in, the less discriminating it is (Fabrizio, 2002).

Per word vector $\boldsymbol{d}_i$ commonly contains a lot of vector elements. In order to reduce the vector dimension, we shall select elements by calculating the value of *TF***IDF* for each element in the vector representation. The words selected as vector elements are called feature words by us. In documents, the higher-frequency words are more important for representing the content than lower-frequency words. However, some high-frequency words such as "the", "for", "at" having low content discriminating power are listed at the stop-list. The Chinese stop words were partially discussed by Wang (1992). It is clear those words appearing in the stop-list will be deleted from the set of feature words in order to reduce the amount of dimensions and enhance the relevancy between words and documents or categories.

Another preprocessing procedure is stemming. In stemming, each word is regarded as word-stem form in documents. For example, "development", "developed" and "developing" will be all treated as "develop". Similarly to stop-list, stemming reduces the amount of dimensions and enhances the relevancy between word and document or categories.

We observed that the author probably uses various words to express the same/similar concept in a text document. Moreover, a particular category consists of a set of documents produced by different authors, each of whom has different personal traits and stylistic features. Therefore, it is common that similar contents in the same category are expressed with various words by authors.

To reduce the amount of dimensions and further enhance the relevancy between word and document and the relevancy between word and category, synonyms are defined by an experimental lexicon constructed by us and all synonyms will be processed and considered as the same word.

The calculation of *TF*, *IDF* and word weight was discussed by Salton and Buckley (1988), and Salton (1991). Thorsten (1996) analyzed the relationship between text classification using the vector space model with TF-IDF weight and probabilistic classifiers. The analysis offered theoretical explanation for the TF-IDF word weight and gave insight into the underlying assumptions. These papers consider identically all the words for the text classification. However, we observe that some words play a dominant role for some particular categories. These words are called feature word in this paper. In this case, the likelihood that the document belongs to a particular category is very high when the document contains feature words. The improved TF-IDF approach uses the feature words to improve the accuracy of text classification.

We initiated two new terms, confidence and support, into the TF-IDF approach. The terms confidence and support were first used in data mining discipline. However, they have some new and concrete meaning in our approach.

Above all, we need to define several variables. We represent document d_i as a feature word vector $\boldsymbol{d}_i$, in which each component $w_j(d_i)$ represents the frequency that the word w_j appears in document d_i. In addition:

$N(all, c_m)$ represents the total number of documents among a particular category c_m.

$N(w_j, all)$ represents the total number of documents containing feature word w_j among the entire training documents corpus composed of all categories.

$N(w_j, c_m)$ represents the total number of documents containing feature word w_j among the particular category c_m.

$N(w_j, \neg c_m)$ represents the total number of documents containing feature word w_j among the entire training documents corpus except the particular category c_m.

$N(all)$ represents the total number of documents among the entire training documents corpus.

We define $conf(w_j, c_m)$ as the confidence degree of feature word w_j on the particular category c_m. The value of $conf(w_j, c_m)$ equals the quotient of the total number of documents containing feature word w_j among the particular category c_m divided by the total number of documents containing feature word w_j among the entire training documents corpus composed by all categories, as follows

$$conf(w_j, c_m) = \frac{N(w_j, c_m)}{N(w_j, all)} \tag{2}$$

We define $sup(w_j)$ as the support of feature word w_j. The value $sup(w_j)$ equals the quotient of the total number of documents containing feature word w_j among the entire training documents corpus divided by the total number of documents among the entire training documents corpus, as follows

$$sup(w_j) = \frac{N(w_j, all)}{N(all)} \tag{3}$$

where

$$0<conf(w_j, c_m)\leq 1 \tag{4}$$

$$0<sup(w_j)\leq 1 \tag{5}$$

Confidence is the measure of certainty to determine a particular category by a particular feature word. The potential usefulness of a particular feature word is represented by support.

The confidence and support of feature words are the dominant measures for text classification. Each measure can be associated with a threshold that can be adjusted by user aiming different types of documents corpus. The dominant measure is defined as

$$dom(w_j, c_m)=f(conf(w_j, c_m), sup(w_j)) = \begin{cases} 1 & ((conf(w_j, c_m) \geq threshold) \wedge (sup(w_j) \geq threshold)) \\ 0 & ((conf(w_j, c_m) < threshold) \vee (sup(w_j) < threshold)) \end{cases} \tag{6}$$

Feature words that meet the threshold ($dom(w_j, c_m)=1$) are considered as characteristic words $cw(w_j, c_m)$. When the set of feature words of a document d_j contains characteristic word $cw(w_j, c_m)$, the text document d_j will be classified into the category c_m. In other words, we can use characteristic word to determine whether a document is classified into a particular category or not.

Because we only use a characteristic word to determine the category of documents, it is briefly called "one-word-location".

Similarly, we may define $N(w_j \cup w_k, all)$, $N(w_j \cup w_k, c_m)$, $conf(w_j \cup w_k, c_m)$ and $sup(w_j \cup w_k)$. For example, we define $sup(w_j \cup w_k)$ as the support of feature words w_j and w_k. The value $sup(w_j \cup w_k)$ equals the quotient of the total number of documents containing feature words w_j and w_k among the entire training documents corpus divided by the total number among the entire training documents corpus, as follows

$$sup(w_j \cup w_k) = \frac{N(w_j \cup w_k, all)}{N(all)} \tag{7}$$

In the above situation, we use the feature words w_j and w_k to determine the category of documents containing feature words w_j and w_k. Here, the set of feature words w_j and w_k is considered as characteristic word. The case is briefly called "two-word-location".

We also call the characteristic word as location-word. To find the location-word is not an arduous task. In order to avoid combinatorial explosion of feature words, we adopt heuristic method to choose the location-word. Our experimental knowledge revealed that, it is adequate to choose higher-frequency feature words as the alternate location-word.

DISCUSSION AND ANALYSIS

Precision of the particular category c_m is the

percentage the number of correctly classified documents among category c_m divided by the total number of documents among category c_m, and is written as

$$precision(c_m) = \frac{TP(c_m)}{N(all,\ c_m)} \tag{8}$$

where, $TP(c_m)$ represents the number of correctly classified documents among the category c_m. *TP* is the abbreviation for "true positive". In other words, $TP(c_m)$ is the number of documents in category c_m classified correctly.

Recall on particular category c_m is the percentage $TP(c_m)$ divided by $S(c_m)$ the total number of documents that should be among the category c_m, and is written as

$$recall(c_m) = \frac{TP(c_m)}{S(c_m)} \tag{9}$$

where $S(c_m)$ represents the total number of documents that should be among the category c_m.

The total number of documents that should belong to the category c_m and be incorrectly classified into other categories is represented as $FN(c_m)$. *FN* means "false negative". Similarly, the total number of documents that should not belong to the category c_m and be incorrectly classified into c_m category is represented as $FP(c_m)$. *FP* means "false positive". $FN(c_m, c_j)$ represents the total number of documents that should belong to the category c_m and be incorrectly classified into c_j category.

Suppose N is the total number of all categories. According to above definitions, Eqs.(10)~(14) can be inferred

$$S(c_m) = TP(c_m)+FN(c_m) \tag{10}$$

$$N(all, c_m) = TP(c_m)+FP(c_m) \tag{11}$$

$$N(all) = \sum_{m=1}^{N} N(all,\ c_m) = \sum_{m=1}^{N} S(c_m) = \sum_{m=1}^{N} TP(c_m)+FN(c_m) \tag{12}$$

$$FN(c_m) = \sum_{j=1, j\neq m}^{N} FN(c_m,\ c_j) \tag{13}$$

$$FP(c_m) = \sum_{j=1, j\neq m}^{N} FN(c_j,\ c_m) \tag{14}$$

Substituting Eq.(11) into Eq.(8) yields expression

$$\begin{aligned} TP(c_m) &= precision(c_m)N(all, c_m) \\ &= precision(c_m)(TP(c_m)+FP(c_m)) \\ &= \frac{precision(c_m)FP(c_m)}{1-precision(c_m)} \end{aligned} \tag{15}$$

Substituting Eq.(10) into Eq.(9) yields expression

$$\begin{aligned} TP(c_m) &= recall(c_m)TP(c_m)+recall(c_m)FN(c_m) \\ &= \frac{recall(c_m)FN(c_m)}{1-recall(c_m)} \end{aligned} \tag{16}$$

By Eqs.(15) and (16), we obtain :

$$FP(c_m) = \frac{(1-precision(c_m))recall(c_m)}{(1-recall(c_m))precision(c_m)} FN(c_m) \tag{17}$$

In addition, we can compute the accuracy of text classification on the entire document corpus Ω:

$$accuracy(\Omega) = \frac{\sum_{m=1}^{N} TP(c_m)}{|\Omega|} \tag{18}$$

$|\Omega|$ denotes the number of documents in corpus Ω.

Next, we discuss the precision and recall of text classification among the set in which each document contains the location-word. In the situation, the following equation can be inferred according to the definitions of Eqs.(2) and (9)

$$recall(c_m) = conf(w_j, c_m) \tag{19}$$

However, precision of classification on particular category is relevant with the distribution of documents and the relationship among categories. It is reasonable to suppose that categories are independent and that the documents are the average distribution. In the situation, the following equation can be inferred according to the definitions of Eqs.(2), (3) and (8)

$$precision(c_m) \approx conf(w_j, c_m) \tag{20}$$

For testing documents corpus, the $precision(c_m)$

represents the precision and *recall*(c_m) represents the recall on the category c_m when the conventional TF-IDF approach is adopted. When the improved TF-IDF approach is adopted, the testing documents corpus can be divided into two sets of documents. One set is the set of documents that contain the location-word. Another is the set of documents that do not contain the location-word.

Applying Eq.(8), we obtain the precision on the particular category c_m for testing documents corpus. When the improved TF-IDF approach is adopted, the precision equals the quotient of the total number of documents classified correctly into the category c_m among the entire testing documents corpus is divided by the total number of documents among the entire testing documents corpus:

$$\begin{aligned} precision &\approx \frac{\sum_{w_j\in L} sup(w_j)N(test)conf(w_j,c_m)+\left(1-\sum_{w_j\in L} sup(w_j)\right)TP(c_m)}{\sum_{w_j\in L} sup(w_j)N(test)+\left(1-\sum_{w_j\in L} sup(w_j)\right)\left(TP(c_m)+FP(c_m)\right)} \\ &= \frac{\sum_{w_j\in L} sup(w_j)N(test)conf(w_j,c_m)+\left(1-\sum_{w_j\in L} sup(w_j)\right)TP(c_m)}{\sum_{w_j\in L} sup(w_j)N(test)+\left(1-\sum_{w_j\in L} sup(w_j)\right)\left(TP(c_m)+\frac{1-precision(c_m)}{precision(c_m)}TP(c_m)\right)} \\ &= precision(c_m)+\frac{\sum_{w_j\in L} sup(w_j)N(test)(conf(w_j,c_m)-precision(c_m))precision(c_m)}{\sum_{w_j\in L} sup(w_j)N(test)precision(c_m)+\left(1-\sum_{w_j\in L} sup(w_j)\right)TP(c_m)} \end{aligned} \tag{21}$$

where $N(test)$ represents the total number of testing documents corpus and L represents the set of all characteristic words in the category c_m. It is common, however that many a category only contains a particular location-word.

According to Eqs.(3) and (6), $\sum_{w_j\in L} sup(w_j)*N(test)$ is the number of documents that contain the location-word in the testing documents corpus and $\sum_{w_j\in L} sup(w_j)N(test)conf(w_j,c_m)$ is the number of documents that are correctly classified into category c_m by the location-word.

$TP(c_m)$ is the number of documents that will be correctly classified in the category c_m among the entire documents corpus by conventional TF-IDF. The value of $\left(1-\sum_{w_j\in L} sup(w_j)\right)N(test)$ is the number of documents that do not contain the location-word in the testing documents corpus. Therefore, the value of $\left(1-\sum_{w_j\in L} sup(w_j)\right)TP(c_m)$ is the number of documents that are correctly classified by the conventional TF-IDF approach among the set of documents that do not contain the location-word.

According to Eq.(20), the number of documents that are classified into the category c_m by the location-word is estimated as $\sum_{w_j\in L} sup(w_j)N(test)$. For documents that do not contain location-word and are classified into the category c_m by the conventional TF-IDF, the number is $\frac{\left(1-\sum_{w_j\in L} sup(w_j)\right)TP(c_m)}{precision(c_m)}$ according to Eqs.(8), (11) and (15).

Similarly, the recall on the particular category c_m for testing the documents corpus is obtained by Eqs.(9), (19) and (20) when the improved TF-IDF approach is adopted.

$$recall \approx \frac{\sum_{w_j \in L} sup(w_j)N(test)conf(w_j, c_m) + \left(1 - \sum_{w_j \in L} sup(w_j)\right) TP(c_m)}{\sum_{w_j \in L} sup(w_j)N(test) + \frac{\left(1 - \sum_{w_j \in L} sup(w_j)\right) TP(c_m)}{recall(c_m)}}$$

$$= recall(c_m) + \frac{\sum_{w_j \in L} sup(w_j)N(test)(conf(w_j, c_m) - recall(c_m))}{\sum_{w_j \in L} sup(w_j)N(test) + \frac{\left(1 - \sum_{w_j \in L} sup(w_j)\right) TP(c_m)}{recall(c_m)}} \quad (22)$$

The value of $\sum_{w_j \in L} sup(w_j)N(test)$ approximately represents the number of documents that contain the location-word and should be classified into the category c_m among the corpus tested.

The value of $\frac{\left(1 - \sum_{w_j \in L} sup(w_j)\right) TP(c_m)}{recall(c_m)}$ represents the number of documents that do not contain the location-word and should be classified into the category c_m among corpus tested.

According to Eqs.(21) and (22), the improved TF-IDF approach will improve the precision on category c_m if the value of $conf(w_j, c_m)$ is greater than the precision of the conventional TF-IDF approach. Similarly, the improved TF-IDF approach will improve the recall on the category c_m if the value of $conf(w_j, c_m)$ is greater than the recall of the conventional TF-IDF approach.

EXPERIMENTS

We performed two sets of experiments. The first experiment set was designed to evaluate the distribution of location-word and the relationship between confidence and support. Another was designed to assess the validity of the improved TF-IDF approach by comparison with the conventional TF-IDF approach. The two sets of experiments were based on same text corpus containing 2138 pieces of science and technology literature. The corpus was partitioned by 7 categories.

We randomly sampled two-thirds of the corpus (1425 pieces of documents) for training and used the remaining one-third for testing. We repeated the experiment 10 times and averaged the result.

Fig.1 shows the relationship between the value of confidence specified and ratio of documents that contain the location-word among the training corpus. When we raised the value of confidence, the number of documents containing the location-word decreased. It means that the usefulness of location-word specifying a particular category will be lowered if the confidence is increased. Meanwhile, the certainty of classification applying the location-word will be improved. In other words, when the threshold of confidence is increased:

(1) The number of texts containing the location-word will decrease. As a result, the number of texts that cannot be processed by the improved TF-IDF approach will increase.

(2) The precision and recall will be further increased for texts when the improved TF-IDF approach applied because the confidence is increased.

In brief, the threshold of confidence is a trade-off. Our experiment revealed that the best value of confidence was 96%, which however, varied in different sample space.

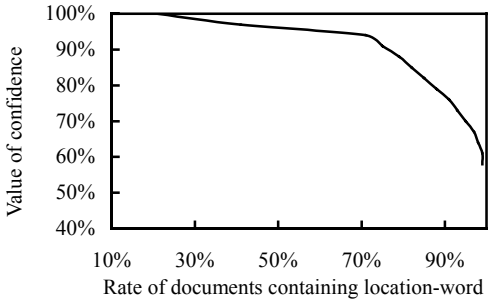

Fig.1 Confidence and location-word

Fig.2 shows the precision and recall of all seven categories when the improved TF-IDF approach and the conventional TF-IDF approach were respectively adopted and the value of confidence was 96%.

CONCLUSIONS AND FURTHER WORK

The improved TF-IDF approach raises the pre-

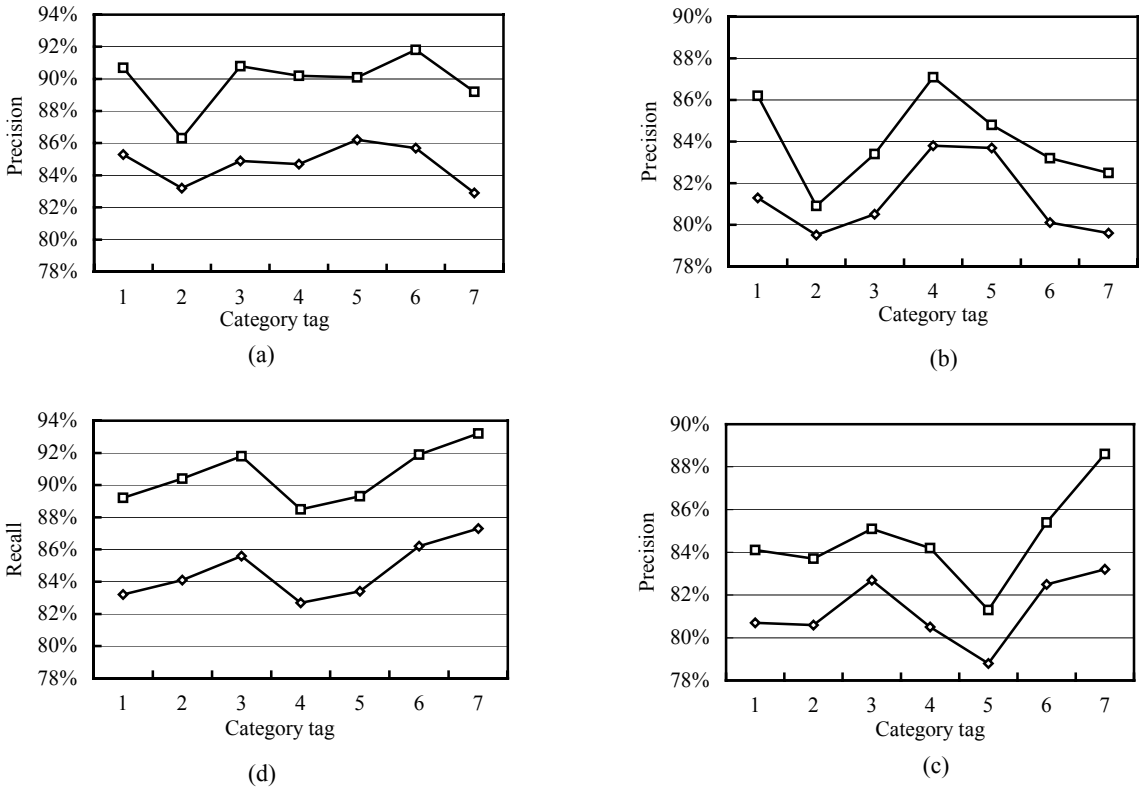

Fig.2 Contrasted precision and recall

(a) Closed test precision; (b) Open test precision; (c) Closed test recall; (d) Open test recall

—◇— Conventional TF-IDF —□— Improved TF-IDF

cision and recall of text classification when the value of confidence is evaluated properly. Moreover, the improved TF-IDF is a language-independent text classification approach.

Further experimental work is needed to test the generality of these results. Although science and technology articles can be considered as representative of various types of documents, we must see how the findings extend to broader types of documents such as news, web pages, email, etc. Another research issue is about how to choose a suitable value of confidence for different corpus.

References

Fabrizio, S., 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, **34**(1):1-47.

Fan, Y., Zheng, C., Wang, Q.Y., Cai, Q.S., Liu, J., 2001. Using naïve bayes to coordinate the classification of web pages. *Journal of Software*, **12**(9):1386-1392.

Huang, X.J., Wu, L.D., 1998. SVM based classification system. *Pattern Recognition and Artificial Intelligence*, **11**(2): 147-153 (in Chinese).

Larry, M.M., Malik, Y., 2001. One-class SVMs for document classification. *Journal of Machine Learning Research*, **2**:139-154.

Lin, H.F., Gao, T., Yao, T.S., 2000. Chinese text visualization. *Journal of Northeastern University*, **21**(5):501-504 (in Chinese).

Olivier, D.V., 2000. Mining E-mail Authorship. Proceedings of Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Boston, USA.

Salton, G., 1991. Developments in automatic text retrieval. *Science*, **253**:974-979.

Salton, G., Buckley, C., 1988. Term weighting approaches in automatic text retrieval. *Information Processing and Management*, **24**(5):513-523.

Sun, J., Wang, W., Zhong, Y.X., 2001. Automatic text categorization based on *k*-nearest neighbor. *Journal of Beijing University of Posts & Telecomms*, **24**(1):42-46 (in Chinese).

Thorsten, J., 1996. Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. Proceedings of 14th International Conference on Machine Learning. McLean. Virginia, USA, p.78-85.

Wang, Y.C., 1992. The Processing Technology and Basis for Chinese Information. Shanghai Jiaotong University Press, Shanghai, p.10-30 (in Chinese).