A Comparative Study on Unsupervised Feature Selection Methods for Text Clustering

Luying LIU, Jianchu KANG, Jing YU, Zhongliang WANG
School of Computer Science, Beihang University
37 Xueyuan Road, Haidian District, Beijing, 100083
E-mail:{lly, kang, yujing, wzl}@nlsde.buaa.edu.cn

Abstract- **Text clustering is one of the central problems in text mining and information retrieval area. For the high dimensionality of feature space and the inherent data sparsity, performance of clustering algorithms will dramatically decline. Two techniques are used to deal with this problem: feature extraction and feature selection. Feature selection methods have been successfully applied to text categorization but seldom applied to text clustering due to the unavailability of class label information. In this paper, four unsupervised feature selection methods, DF, TC, TVQ, and a new proposed method TV are introduced. Experiments are taken to show that feature selection methods can improves efficiency as well as accuracy of text clustering. Three clustering validity criterions are studied and used to evaluate clustering results.**

I. INTRODUCTION

In recent years, a tremendous growth in the volume of text documents available on the Internet, digital libraries, news sources, and company-wide intranets has been witnessed. This has led to an increased interest in developing methods that can help users to effectively navigate, summarize, and organize this information with the ultimate goal of helping them to find what they are looking for. Fast and high-quality text clustering algorithms play an important role towards this goal as they have been shown to provide both an intuitive navigation/browsing mechanism by organizing large amounts of information into a small number of meaningful clusters as well as to greatly improve the retrieval performance either via cluster-driven dimensionality reduction, term-weighting, or query expansion [1].

Compared with structured data in database, because of lingual diversity, text data with unstructured form is more diversiform and complex. Researches have shown that under the current automatic understanding level of natural language, word is still the best unit for text representation and processing. Nowadays, Vector Space Model (VSM) [2] is the most popular method in representing documents. A key assumption of VSM is that the sequence of words appear in the document is not important. A text or document is represented as a series of words without order information, known as a bag of words. This representation raises two severe problems: the high dimensionality of feature space and the inherent data sparsity. A single document has a sparse vector over the set of all terms and may contain hundreds of thousands of terms. The performance of clustering algorithms will decline dramatically due to the problems of high dimensionality and data sparseness [3].

In general, the feature dimensionality can be cut down by removing stop-words and words with high frequency. Stop-words are usually given as a word list. Most of these words are conjunctions or adverbs which have no contribution to cluster process, and sometimes have negative influence. Words with high frequency which can be gotten in word frequency dictionary appear in most documents, so they are not helpful for cluster either. Words appear in no more than three documents and at least 33% of all documents can be removed [4]. But to evaluate the performance of feature selection methods, this technique is not used in this paper.

For the negative influence of high dimensionality and data sparsity, it is highly desirable to reduce the feature space dimensionality. There are two commonly used techniques to deal with this problem: feature extraction and feature selection. Feature extraction is a process that extracts a set of new features from the original features through some functional mapping [5], such as principal component analysis (PCA) [6] and word clustering [7]. The feature extraction methods have a drawback that the generated new features may not have a clear physical meaning so that the clustering results are difficult to interpret [8].

Feature selection is a process that chooses a subset from the original feature set according to some criterions. The selected feature retains original physical meaning and provides a better understanding for the data and learning process. Depending on if the class label information is required, feature selection can be either unsupervised or supervised. For supervised methods, the correlation of each feature with the class label is computed by distance, information dependence, or consistency measures [8]. Some supervised feature selection methods have been successfully used in text classification [9], such as Information Gain (IG) and $\chi 2$ Statistics (GHI). But due to the lack of class label, supervised feature selection methods can not be used in text clustering. In this paper, we investigate (a) what the strengths and weaknesses of existing feature selection methods are when applied to text clustering, (b) how much of the document vocabulary can be reduced without losing useful information in text clustering, and (c) the influence of validity criterions to clustering results.

The rest of this paper is organized as follows. In Section II, we give a brief introduction on several feature selection methods including a new method TV. In Section III, we introduce the validity criterions we use in the paper. We introduce the datasets and validity criterions in Section IV. In Section V, several experiments are conducted to compare the

0-7803-9361-9/05/$20.00 ©2005 IEEE

effectiveness of different feature selection methods. And we will show that the effects of feature selection methods are different when they are used on different dataset or under different validity criterions. Finally, we summarize our major contributions in Section VI.

II. Feature Selection Methods

As we discuss above, unsupervised feature selection methods select a subset of important feature for clustering over the whole data. In this section, we give an introduction on several feature selection methods, including DF, TC, TVQ and TV.

Notations used in the paper are as follows: D denotes the document set, M is the dimension of the feature, N is the number of documents in the dataset, D_j is j^{th} document in the dataset, t_{ij} is i^{th} feature of D_j, f_{ij} is the frequency of t_{ij}.

A. Document Frequency (DF)

Document frequency is the number of documents in which a term occurs in a dataset. It is the simplest criterion for term selection and easily scales to a large dataset with linear computation complexity. A basic assumption of this method is that terms appear in minority documents are not important or will not influence the clustering efficiency. It is a simple but effective feature selection method for text categorization [9].

B. Term Contribution (TC)

Because the simple method like DF assumes that each term is of same importance in different documents, it is easily biased by those common terms which have high document frequency but uniform distribution over different classes. TC is proposed to deal with this problem [10].

We will introduce TF.IDF (Term Frequency Inverse Document Frequency) first [11]. TF.IDF synthetically considers the frequency of a term in a document and the document frequency of the term. It believes that if a term appears in too many documents, it's too common and not important for clustering. So Inverse Document Frequency is considered. That is, if the frequency of a term in a document is high and it does not appear in many documents, the term is important. A common form of TF.IDF is:

$$f(t_i, D_j) = TF_{ij} * \log\left(N / DF_j\right) \tag{1}$$

The result of text clustering is highly dependent on the documents similarity. So the contribution of a term can be viewed as its contribution to the documents' similarity. The similarity between documents D_i and D_j is computed by dot product:

$$Sim(D_i, D_j) = \sum_t f(t, D_i) * f(t, D_j) \tag{2}$$

So the contribution of a term in a dataset is defined as its overall contribution to the documents' similarities. The equation is:

$$TC(t_k) = \sum_{i,j \cap i \neq j} f(t_k, D_i) * f(t_k, D_j) \tag{3}$$

C. Term variance quality (TVQ)

Term variance quality method is introduced by Inderjit Dhillon, Jacob Kogan and Charles Nicholas [12]. It follows the ideas of Salton and McGill [13]. The quality of the term t is measured as follows:

$$q(t_i) = \sum_{j=1}^{n} f_{ij}^2 - \frac{1}{n}\left[\sum_{j=1}^{n} f_{ij}\right]^2 \tag{4}$$

Where n is the number of documents in which t occurs at least once, and $f_{ij} >= 1$, j=1,...,n.

D. Term Variance (TV)

We introduce a new method called Term Variance to evaluate the quality of terms. That is to compute the variance of every term in all dataset. Methods like DF assume that each term is of same importance in different documents, it is easily biased by those common terms which have high document frequency but uniform distribution over different classes. TV follows the idea of DF that the terms with low document frequency is not important and can solve the problem above at the same time. A term appears in very few documents or has uniform distribution over documents will have a low TV value. The quality of the term is measured as follows:

$$v(t_i) = \sum_{j=1}^{N}\left[f_{ij} - \overline{f_i}\right]^2 \tag{5}$$

III. Cluster Validity Criterions

The application of a clustering algorithm to a dataset aims at, assuming that the dataset offers such a clustering tendency, discovering its inherent partitions. However, the clustering process is perceived as an unsupervised process, since there are no predefined classes and no models that would show what kind of desirable relations should be valid among the data. So some validity criterions used on text classification, such as precision and recall [14], are not suit for evaluating the clustering results.

In general terms, there are three approaches to investigate cluster validity [15]. The first is based on external criteria. This implies that we evaluate the results of a clustering algorithm based on a pre-specified structure, which is imposed on a dataset and reflects our intuition about the clustering structure of the dataset. The second approach is based on internal criteria. In this case the clustering results are evaluated using only quantities and features inherited from the dataset. The third approach of clustering is based on relative criteria. Here the basic idea is the evaluation of a clustering structure by comparing it to other clustering schemes, resulting by the same algorithm but with different input parameter values.

In our experiments in next section, we use datasets offering by some organizations and having been labeled manually. So the external criteria are used here. Consider $C = \{C_1...C_m\}$ is a clustering structure of a dataset D and $P = \{P_1...P_s\}$ is a

defined partition of the data. Every pair of points from the dataset using the following terms:

SS: if both points belong to the same cluster of the clustering structure C and to the same group of partition P.

SD: if points belong to the same cluster of C and to different groups of P.

DS: if points belong to different clusters of C and to the same group of P.

DD: if both points belong to different clusters of C and to different groups of P.

Assuming that a, b, c and d are the number of SS, SD, DS and DD pairs respectively, then a+b+c+d=M, which is the maximum numbers of all pairs in the dataset. The following indices are defined to measure the degree of similarity between C and P:

Averaged Accuracy:

$$AA = \frac{1}{2}\left[\frac{a}{a+c} \cdot \frac{d}{b+d}\right] \tag{6}$$

Rand Statistic:

$$RS = {(a+b)}/{M} \tag{7}$$

Folkes and Mallows index:

$$FM = \sqrt{\frac{a}{a+b} \cdot \frac{a}{a+c}} \tag{8}$$

IV. Datasets And Cluster Validity Criterions

In Section IV and Section V, we will use DF, TC, TVQ and TV methods to reduce the feature dimensionality of four datasets: FBIS, RE1, TR45 and TR41. Then cluster validity criterions will be used to evaluate the effect of these feature selection methods.

A. Datasets

Text classification performance varies greatly on different dataset. So we chose four different text datasets to evaluate the performance of the feature selection methods. The characteristics of the various document collections used in our experiments are summarized in Table 1.

Data set FBIS is from the Foreign Broadcast Information Service data of TREC-5 [16]. Data sets RE1 is from Reuters-21578 text categorization test collection Distribution 1.0 [17]. Data sets TR45 and TR41 are derived from TREC-6 collections. For all data sets, we used a stop-list to remove common words, and the words were stemmed using Porter's suffix-stripping algorithm [18].

B. Evaluation of Cluster Validity Criterions

As we talked in Section III, there are many cluster validity criterions can be used to evaluate the performance of clustering algorithms. But the performance of cluster validity criterions themselves is different. In this section, we will first evaluate these validity criterions by applying a single feature selection method DF on different datasets on which the performance has already reached a compatible view in this research field.

TABLE 1
Characteristics of Datasets

Data	# of doc	# of class	avg class size	# of words
FBIS	2463	17	144.9	2000
RE1	1657	25	66.3	3758
TR45	690	10	69.0	8261
TR41	878	10	87.8	7454

DF is a simple but effective feature selection method. When applying DF on text datasets, if minority of terms are removed, the clustering performance will be improved or no loss. When more terms removed, the clustering performance will drop quickly.

The values of different validity criterions when applying DF on different datasets are showed in Fig. 1.

The results of AA, RS, FM are respectively range from 0.5714 to 0.7201, from 0.7370 to 0.8928, from 0.1422 to 0.5157. As can be seen in Fig. 1, four curves of RS methods on different four datasets approximately follow the rule we mentioned above. But the curves are very gently, so the trends are not distinct. Four curves of AA are all follow the rule well except the curve of TR45. Curves of FM on datasets FBIS and RE1 follow the rule of DF very well, while curves of TR45 and TR41 surge randomly.

So as to the result of our first experiment, AA is the best validity criterion. And we can see from the result that text classification performance varies greatly on different dataset. The performance of FBIS and RE1 are much better than the others. And if we only consider the result of FBIS and RE1, AA and FM validity criterions are both good, and FM may even better. So in the experiments below, we will mainly use FBIS and RE1 datasets, as well as AA and FM validity criterions.

V. Evaluation Of Feature Selection Methods

The following experiments we conducted are to compare the unsupervised feature selection methods DF, TC, TVQ and TV.

We chose K-means to be the clustering algorithm. Since K-

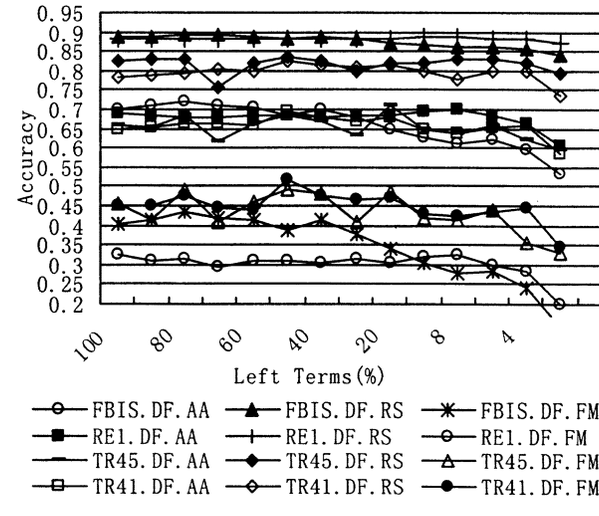

Fig.1. Precision comparison on datasets

means clustering algorithm is easily influenced by selection of initial centroids, we random produced 5 sets of initial centroids for each dataset and averaged 5 times performance as the final clustering performance.

The AA and FM results on FBIS and RE1 are shown in Fig. 2 to Fig. 5.

From these figures, first, we can see that the unsupervised feature selection methods can improve the clustering performance when a certain terms are removed. For all methods in our experiments, at least 70% terms can be removed with no loss in clustering performance on both

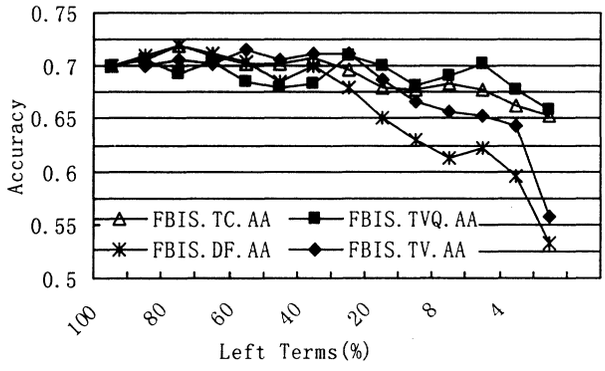

Fig.2. Precision comparison on FBIS (AA)

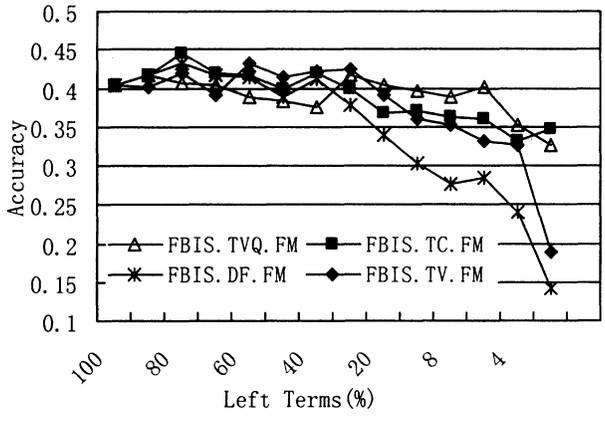

Fig.3. Precision comparison on FBIS (FM)

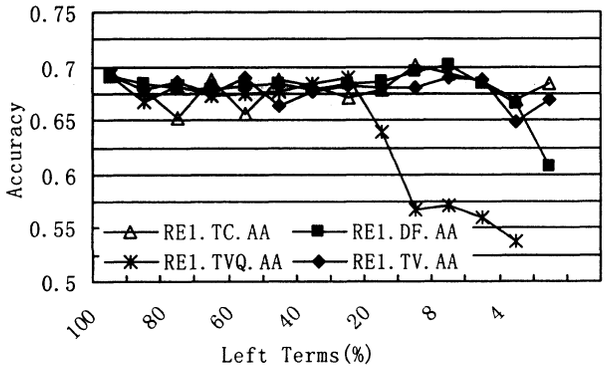

Fig.4. Precision comparison on RE1 (AA)

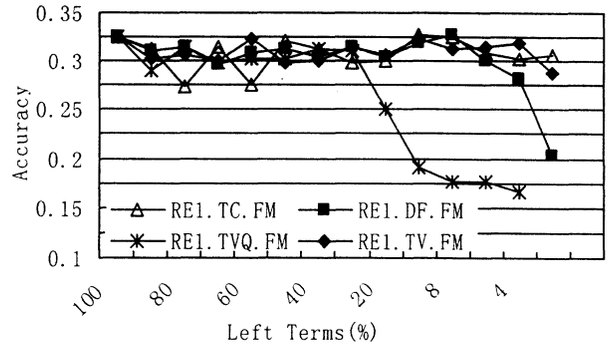

Fig.5. Precision comparison on RE1 (FM)

datasets. And for most feature selection methods, when certain features are removed, the clustering performances can be improved. For instance, when 20% terms of FBIS are removed by TC method, it can achieve 9.4% FM value improvement.

Second, TC is the steadiest method in all. The performance of clustering will not descend distinctly when terms are removed. The results of TC method are shown in Fig. 6.

Third, TV method is a little worse than TC, but much better than DF and TVQ. DF method drop quickly when more than 60% terms are removed, and the performance of TVQ is very bad when more than 70% terms are removed from RE1 dataset. The results of TV method are shown in Fig. 7. When no more than 80% terms are removed from datasets by TV method, there will be no loss in clustering performance.

VI. CONCLUSION

Clustering is one of the most important tasks in the data mining process for discovering groups and identifying interesting distributions and patterns in the underlying data. In order to solve the high dimensionality and inherent data sparsity problems of feature space, feature selection methods are used. In real case, the class information is unknown, so only unsupervised feature selection methods can be exploited.

In this paper, we evaluate several unsupervised feature selection methods, including DF, TC, TVQ and a new proposed method TV. TC and TV are better than DF and

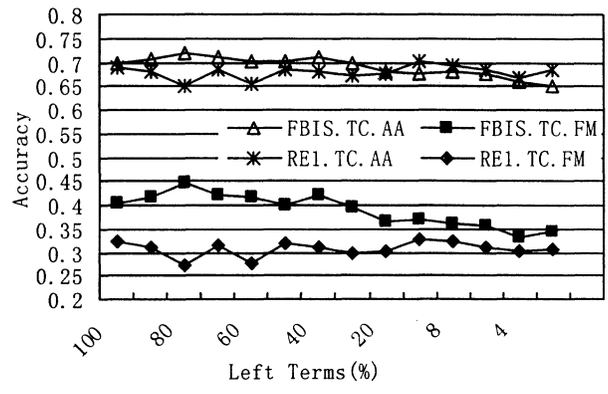

Fig.6. Performance of TC

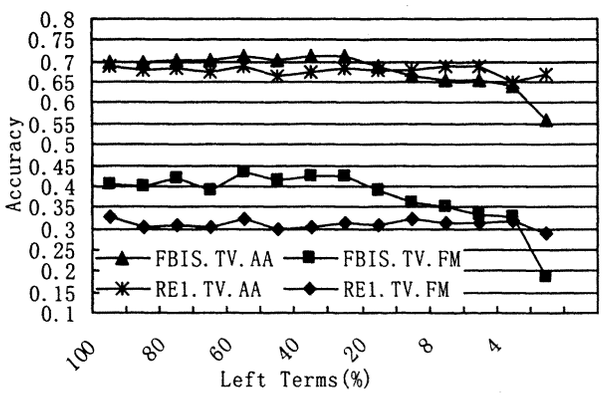

Fig.7. Performance of TV

TVQ. We also indicate in this paper that the performances of different cluster validity criterions are not same, and AA and FM criterions are better for evaluating the clustering results.

REFERENCES

[1] Y. Zhao & G. Karypis. "Criterion Functions for Document Clustering: Experiments and Analysis," *Technical Report #01-40*, Department of Computer Science, University of Minnesota, November 2001.

[2] Salton, G., Wang. A., & Yang, C.S., "A Vector Space Model for Information Retrieval," *Communications of the ACM*, 18(11): 613-620, November 1975

[3] Aggrawal, C.C., & Yu, P.S., "Finding Generalized Projected Clusters in High Dimensional Spaces," *Proc. Of SIGMOD'00* (pp. 70-81). 2000

[4] S. Ruger & S. Gauch. "Feature Reduction for Document Clustering and Classification," *Technical report*, Computing Department, Imperial College London, UK, 2000.

[5] Wyse, N., Dubes, R., & Jain, A.K., "A Critical Evaluation of Intrinsic Dimensionality Algorithms," *Pattern Recognition in Practice*, pp. 415-425, North-Holland, 1980

[6] Jolliffe, I.T., "Principal Component Analysis," *Springer Series in Statistics*, 1986

[7] Slonim, N., & Tishby, N., "Document Clustering using Word Clusters via the Information Bottleneck Method," *Proc. of SIGIR'00*, pp. 208-215, 2000

[8] Dash, M., & Liu, H., "Feature Selection for Classification," *International Journal of Intelligent Data Analysis, 1(3)*, pp. 131-156, 1997

[9] Yang, Y., & Pedersen, J. O., "A Comparative Study on Feature Selection in Text Categorization," *Proc. of ICML-97*, pp. 412-420, 1997

[10] Tao Liu, Shenping Liu, Zheng Chen, & WeiYing Ma, "An Evaluation on Feature Selection for Text Clustering," *Proc. of ICML-2003*, Washington DC, 2003.

[11] Gerard Salton, "Developments in Automatic Text Retrieval," *Science*, 253:974-980, August 1991

[12] Inderjit Dhillon, Jacob Kogan, &Charles Nicholas, "Feature Selection and Document Clustering," http://www,csee.umbc.edu/cadip/2002Symposium/kogan.pdf

[13] G.Salton, & M.J.McGill, "Introduction to Modern Information Retrieval," McGrawHill, New York, 1983

[14] Douthat A, "The Message Understanding Conference Scoring Software User's Manual," *In Proceedings of the Seventh Message Understanding Conference*, 1998

[15] Theodoridis, S., Koutroubas, K. "Pattern recognition," *Academic Press*, 1999

[16] TREC. *Text Retrieval Conference*. http://trec.nist.gov, 1999.

[17] D. D. Lewis, "Reuters-21578 Text Categorization Test Collection Distribution 1.0," http://www.research.att.com/~lewis, 1999.

[18] M. F. Porter. "An Algorithm for Suffix Stripping," *Program*, 14(3):130–137, 1980