# A Feature Selection Method for Document Clustering Based on Part-of-Speech and Word Co-Occurrence

Zitao Liu
International School of Software
Wuhan University
Wuhan, China
zitao.whu@gmail.com

Wenchao Yu
International School of Software
Wuhan University
Wuhan, China
issyuwenchao@gmail.com

Yalan Deng
International School of Software
Wuhan University
Wuhan, China
dengyalan@gmail.com

Yongtao Wang
International School of Software
Wuhan University
Wuhan, China
williampaladin@gmail.com

Zhiqi Bian
International School of Software
Wuhan University
Wuhan, China
bianzhiqi@yahoo.cn

*Abstract*—**Feature selection is a process which chooses a subset from the original feature set according to some rules. The selected feature retains original physical meaning and provides a better understanding for the data and learning process. However, few modern feature selection approaches take the advantage of features' context information. Based on this analysis, we propose a novel feature selection method based on part-of-speech and word co-occurrence. According the components of Chinese document text, we utilize the words' part-of-speech attributes to filter lots of meaningless terms. Then we define and use co-occurrence words by their part-of-speech to select features. In the evaluating process, we use the text corpus from Sogou Lab to do some experiments and use Entropy and Precision as criteria to give an objective evaluation of document clustering performance. The results show that our method can select better features and get a more pleasant clustering performance.**

***Keywords-feature selection; document clustering; part-of-speech; word co-occurrence***

## I. INTRODUCTION

With the growth of the Internet, huge amounts of text data expand in a geometric way. How to improve the efficiency of utilizing the information resource through information fusion has become a hot researching spot [1]. Document clustering, which is an unsupervised information organization method, attracts lots of researchers and developers [2]. One basic problem of obtaining useful and helpful information from massive data sets is how to effectively represent and denote this text data. When we process a large number of document set, we encounter two problems. For one thing, the processing time increase dramatically with the growth of the number of documents and the length of each document. For another thing, some of the features may be redundant, some may be irrelevant, and some can even misguide clustering results。In such a case, selecting a subset of features, which containing more information, from the initial feature space always bring us a better performance [3][4].

## II. RELATED WORK

Feature selection is one of the most basic problems in machine learning area [5]. It is aiming at extract a small subset of feature from the problem domain while retaining a suitably high accuracy in representing the original features [6]. In recent years, a lot of feature selection methods have been proposed. There are some traditional feature selection methods based on Document Frequency (DF) and Term Strength (TS) [7]. Dash and Liu proposed a feature selection method based on entropy to weigh each feature's importance to each cluster [8]. Yun Zhang gave a hierarchical method based on co-occurrence in the search engine's results clustering [9]. YUAN-CHAO LIU considered that words occurred in a same document are co-occurrence words and he proposed a feature selection algorithm for document clustering based on word co-occurrence frequency [10]. Meanwhile, MING LIU thought that a word's co-occurrence words are those occurs before and after certain word and he gave relevant feature selection method to improve the clustering performance [11]. Peter utilized the relationship between sentences and sentences' frequency to select co-occurrence words [12]. However, these researches have some common limitation: First, to some text document after segmentation, different words' part-of-speech means different amount of clustering information, but above researches treat them equally. Second, the definition of word co-occurrence is obscure and do not utilize the word's context information. Based on above analysis, we give our feature selection method based on part-of-speech and word co-occurrence, which firstly filtering features by part-of-speech and secondly defining word co-occurrence by part-of-speech. The results show that our feature selection methods own a better clustering performance.

The rest of this paper is organized as follows. In section 3, we first briefly introduce our new feature extraction model. Then we use both the part of speech and word co-occurrence information to filter and select some well features. In Section 4, we conduct several experiments to compare the effectiveness of different feature selection methods in ideal and real cases. Finally, we summarize our major contributions in Section 5.

978-1-4244-5934-6/10/$26.00 ©2010 IEEE

## III. FEATURE SELECTION USING POS AND WORD CO-OCCURRENCE

### *A. Feature Extraction Model*

In the traditional process of feature selection in document clustering, researchers segment each document into a list of terms and use some criterions to score and sort these candidate features. In this way, we treat each segmented term equally. However, this approach ignores the information of each term's part of speech. A document or a sentence is made of a list of full words (noun, verb, adjective…) and functional words (preposition, conjunction, auxiliary…). These functional words do not contain the semantic meanings and only are used to be some syntactic elements, which are meaningless to the whole document in the semantic aspect.

Meanwhile, to certain separated term in the whole document set, those words, which occur front or behind this term, hold valuable information to explain its meaning. We call these words Context Words. In other words, one term co-occurs with its context words in a document. Considering the analysis above, we use terms' context words to select the features and to measure the contribution of certain term to document clustering. For pairs of co-occurring words above certain threshold of frequency, the hold more clustering information than other words, so based on above opinions, we propose our feature selection model based on part of speech and word co-occurrence, see Figure 1.

### *B. Selection Based on Part-of-Speech*

No matter to the semantic meaning in a sentence or to relevance of document topic, noun and verbs hold much more information than preposition and other functional words. However, according to the syntactic need, a huge number of functional words used in one document. There large numbers of functional words which exist in the feature vectors not only cause longer time in clustering, but also influence the precision of clustering. We do some statistical research in the public corpus of China Daily published in January, 1998. There are 1140931 Chinese terms after segmentation in this corpus. It contains 614451 functional words, which is 53.9% of the total terms. See Figure 2.

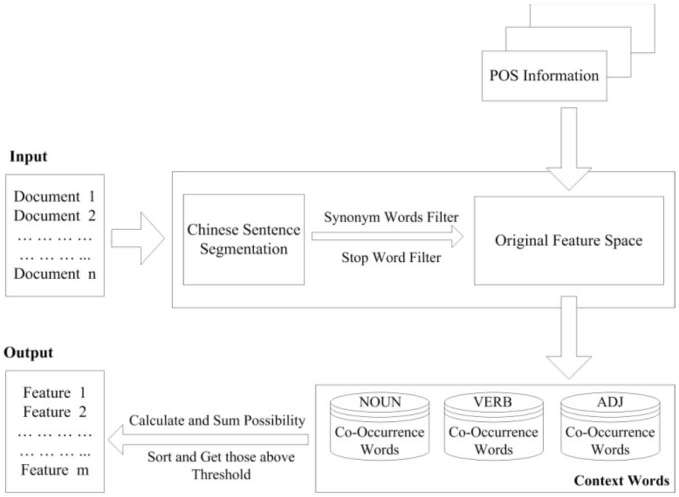

Figure 1. Feature selection model based on pos and word co-occurrence

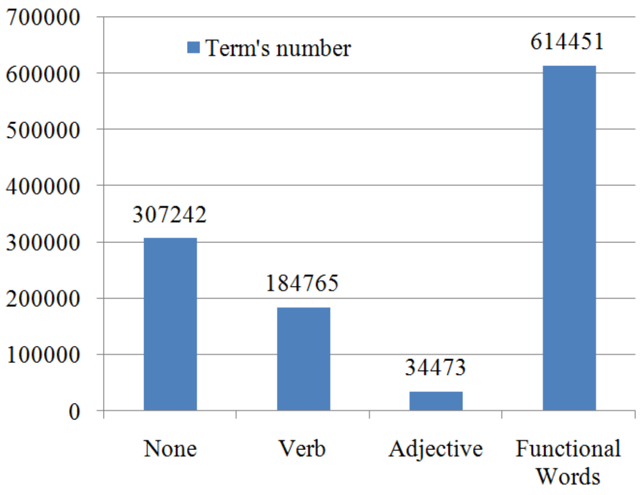

Figure 2. Statistical term's number in China Daily in January 1998

To those segmented initial feature selection set, we use the following tags of part-of-speech in Table I to filter those initial features. We only reserve noun, verbs and adjectives.

$$\mathrm{Term}_i \equiv (\mathrm{Term}^i_{\mathrm{LiteralVal}}, \mathrm{Term}^i_{\mathrm{POS}})\ \mathrm{Term}_{\mathrm{POS}} \in \{\mathrm{NOUN}, \mathrm{VERB}, \mathrm{ADJ}\}$$

### *C. Selection Based on Correlated Word Pairs*

#### *1) What word co-occurrence is*

Before the concrete description of word co-occurrence, let us give the definition of word co-occurrence. In the traditional feature selection based on co-occurrence, researchers always consider that two words are correlated word pairs if they co-occur in the same document or same sentence. However, in almost circumstance, those pairs are non-related in their own context, which is meaningless to feature selection. What's worse, they may reduce the precision of final document clustering. So based on the difference of each term's part of speech, we use the following rules to define the word co-occurrence.

TABLE I. PART-OF-SPEECH TAGS IN FEATURE SELECTION

| General PoS Category | Specified PoS Tag | Explanation |
|---|---|---|
| ADJ | a | Adjective |
| | ag | Adjective Morpheme |
| | ad | Adverb |
| | an | Adnoun |
| NOUN | n | Noun |
| | nr | People's Name |
| | ns | Place's Name |
| | nt | Organization' Name |
| | nz | Other Proper Nouns |
| VERB | v | Verb |
| | vd | Avendo |
| | vn | Gerund |

We use Doc to represent each document and use $S_{Doc}$ to represent the whole corpus ($Doc \in S_{Doc}$). $Term_i$, $Term_j$ ($i < j$) denote the two distinct terms in the corpus.

**Rule 1:**
If ( $Term^i_{POS}$ .Equlas(NOUN) && | $Term_i - Term_j$ | <= 3) $Term_j \in S^i_{CoOccurWord}$
**Rule 2:**
If ( $Term^i_{POS}$ .Equlas(VERB) && | $Term_i - Term_j$ | == 1) $Term_j \in S^i_{CoOccurWord}$
**Rule 3:**
If ( $Term^i_{POS}$ .Equlas(ADJ) && $Term_i - Term_j$ == 1) $Term_j \in S^i_{CoOccurWord}$

In the above rules, $S^i_{CoOccurWord}$ denotes $Term_i$ 's co-occurred words set. | $Term_i - Term_j$ | denotes the distance between two terms which is filtered by PoS in one document. For example, in the sentence "He is a cute baby," |He - baby| = |1 - 5| = 4. To those nouns, according to the research conducted by linguists, two nouns in certain near distance always have a high relevance in their meaning. To a verb, we always concern with this action's subject and object, so we choose the word in front of or behind certain verb as our co-occurred words. To an adjective, we put an emphasis on noun modified by this adjective instead of some words physically nearby.

*2) How to use correlated word pairs*

Each feature $Term_i$ has its own co-occurred word set $S^i_{CoOccurWord}$ and its co-occurred words' corresponding possibility set $S^i_{P_{CoOccurWord}}$.

To each element $S^i_{CoOccurWord}$ in $S^i_{CoOccurWord}$, we use (1) to calculate the co-occurrence possibility between $CoOccurWord_j$ and $Term_i$ . Then we add the outcome into $S^i_{P_{CoOccurWord}}$.

$$P^i_{CoOccurWord_j} = \frac{Freq(Term_i ,CoOccurWord_j)}{Freq( CoOccurWord_j) * Freq(Term_i )} \quad (1)$$

In the above equation, the co-occurrence possibility is decided by the frequency of $Term_i$ and its$CoOccurWord_j$. Freq ($Term_i$,$CoOccurWord_j$) denotes the number of $Term_i$ and $CoOccurWord_j$'s co-occurrence. Freq ($Term_i$) and Freq ($CoOccurWord_j$) denotes the frequence of $Term_i$ and frequence of $CoOccurWord_j$ in the whole document set respectively.

We use the following (2) to calculate the weighting possibility of $Term_i$.

$$P_{Term_i} = \frac{1}{n}\sum_{i=0}^{n} S^{NOUN}_{P_{CoOccurWord}}[i] * \alpha + \frac{1}{n}\sum_{i=0}^{n} S^{VERB}_{P_{CoOccurWord}}[i] * \beta + \frac{1}{n}\sum_{i=0}^{n} S^{ADJ}_{P_{CoOccurWord}}[i] * \gamma$$

where $\alpha + \beta + \gamma = 1$. (2)

Each element $P^i_{CoOccurWord_j}$ in the co-occurrence possibility set $S^i_{P_{CoOccurWord}}$ of $Term_i$. If $Term_i$'s $CoOccurWord_j$ is noun, we multiply $\alpha$ and its$P^i_{CoOccurWord_j}$; if $Term_i$'s $CoOccurWord_j$ is verb, we multiply $\beta$ and its $P^i_{CoOccurWord_j}$; if $Term_i$'s $CoOccurWord_j$ is adjective, we multiply $\gamma$ and its $P^i_{CoOccurWord_j}$. In our experiment, we use $\alpha = 0.7$, $\beta = 0.2$, $\gamma = 0.1$. At last, we sort each $Term_i$'s co-occurrence possibility $P_{Term_i}$ and set a threshold $\theta_t$ to select those above $\theta_t$ as our final features.

## IV. Experiment Result and Analysis

Based on the feature selection model mentioned above, we use the Chinese web portal news data provided by Sogou Lab publicly to do our experiment [13]. We extract 5 categories (Information Technology, Military, Finance, Auto, Sports) from 18 categories of news. There are 22668 documents in total (4102 documents from Information Technology, 2979 documents from Military; 5728 documents from Finance; 4934 documents from Auto; 4222 documents from Sports.) See Table II. We use the ICTCLAS Chinese Word Segmentation and PoS system to preprocess our dataset [14].

Considering that the initial centroids have a deep influence on the K-means clustering, we random produced 5 sets of initial centroids for our data set and averaged 10 times performances as the final clustering performance. Before performing clustering, TF-IDF was used to calculate the weight of each term. We use feature selection methods based on entropy, document frequency, $\chi^2$ statistic (CHI) to reflect comparison.

We use two criterions: Entropy and Precision to evaluate the K-means clustering. Entropy measures the uniformity or purity of a cluster and Precision directly reflect the performance of clustering [15]. See the following (3), (4).

$$Entropy = -\sum_{j=1}^{\acute{K}} \frac{C_j}{N} \sum_{i=1}^{K} p_{ij} * \log_2 p_{ij}$$
$$\text{where} \quad p_{ij} = \frac{1}{C_j} \left| \{ d_i \mid label(d_i = C_j)\}\right| \quad (3)$$

$$Precision = \sum_{j=1}^{\acute{K}} \frac{C_j}{N} \, Precision(C_j)$$
$$\text{where} \quad Precision(C) = \frac{1}{C} \max(|\{d_i \mid label(d_i = C_j)\}|) \quad (4)$$

In the above formula, K denotes the number of original classes and $\acute{K}$ denotes the number of obtained classes. N is the total documents' number. $C_j$ represents number of documents in the j cluster.

TABLE II. Chinese Web Portal News Dataset Properties

| Data Category | Docs Num | Terms Num | Avg. Terms Per Doc | Distinct Term Num | Avg. DF Per Term |
|---|---|---|---|---|---|
| IT | 4102 | 895256 | 218 | 37017 | 24 |
| Military | 2979 | 569318 | 191 | 32091 | 17 |
| Finance | 5660 | 1295278 | 229 | 38288 | 33 |
| Auto | 4934 | 882478 | 179 | 30447 | 28 |
| Sports | 4222 | 526254 | 125 | 33763 | 15 |

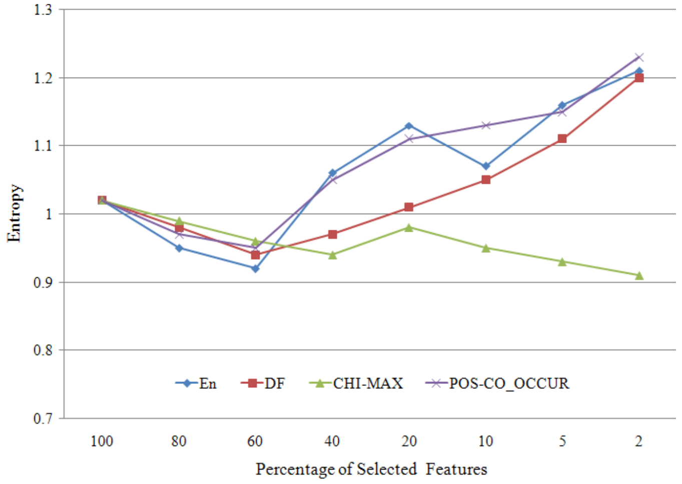

Figure 3. Entropy comparision by four feature selection methods

The entropy of the clusters based on several famous feature selection methods is depicted in Figure 3. From this figure, we can find that when we select 50% or more features, the four methods do not have an obvious effect on the cluster entropy. In their performance about entropy, they are equally effective. However, with the reduction of the number of features, different feature selection methods have different performance. When we use 10% of selected features, the worst method is CHI-MAX and its performance drops evidently. Among these four methods, our feature selection method based on PoS and Word Co-Occurrence has a pleasant and robust clustering performance even through the reduction of number of features.

In the performance of precision, feature selection based on DF has a good clustering precision when the features' amount is huge. From 60% to 90%, DF method always owns a better clustering performance than other three methods. However, when we select 20% or fewer features to conduct document clustering experiment, our method's precision ascends gradually. Especially from 10% to 2%, our method always has the best performance. See Figure 4.

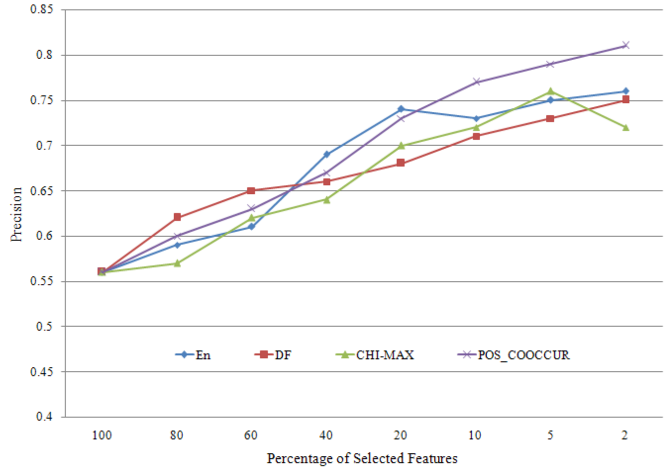

Figure 4. Precision comparision by four feature selection methods

## V. CONCLUSION

The main contribution of this paper is: First, considering the characteristics of Chinese documents, we propose our feature selection model in Chinese document clustering. Second, based on the syntactic attributes of Chinese language, we use part of speech to filter our candidate features. Meanwhile, we use our definition of distance between two terms to improve the traditional feature selection based on word co-occurrence. This fully utilizes the words' context information. According to different word's part of speech, we use different weights which highly improve the clustering precision. In the future, experiments will explore the optimal combination of various thresholds in the algorithm to further enhance the clustering quality and apply this approach into some web application.

## REFERENCES

[1] Yang, Y. Expert network: effective and efficient learning from human decisions in text categorization and retrieval. *In Proceedings of 17th ACM International Conference on Research and Development in Information Retrieval (SIGIR 1994)*, pp. 13–22.

[2] Luying Liu. A comparative study on unsupervised feature selection methods for text clustering. *In Proceedings of the 12th IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE 2005)*, pages 597-601.

[3] Meng, Xianjun. Semantic feature reduction in Chinese document clustering. *In Proceedings of International Conference on Systems, Man and Cybernetics(SMC 2008)*,pages 3721-3726.

[4] Yang, Y. and Liu, X. An-examination of text categorization methods. *In Proceedings of 22nd ACM International Conference on Research and Development in Information Retrieval (SIGIR 1999)*, pp. 42–49.

[5] Yang, Y. and Pedersen, J. O. A comparative study on feature selection in text categorization. *In Proceedings of 14th International Conference on Machine Learning (ICML 1997)*, pp. 412–420.

[6] M. Dash and H. Liu. Feature selection for classification. *Intelligent Data Analysis: An International Journal*, 1(3):131–156, 1997.

[7] Y. Yang. Noise reduction in a statistical approach to text categorization. *In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1995)* pp. 256-263, 1995.

[8] M. Dash and H. Liu. Feature selection for clustering. *In Proceedings of the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2000)*, pages 110–121,2000.

[9] Zhang, Yun and Feng, Boqin. A Co-occurrence based Hierarchical Method for Clustering Web Search Results. *In Proceedings of International Conference on Web Intelligence (WI 2008)*, pages 407-410.

[10] Liu, Yuan-Chao. A feature selection algorithm for document clustering based on word co-occurrence frequency. *In Proceedings of International Conference on Machine Learning and Cybernetics (ICMLC 2004)*, pages 2963-2968.

[11] Liu, Ming. FSSOM: One Novel SOM Clustering Algorithm based on feature selection. *In Proceedings of International Conference on Machine Learning and Cybernetics (ICMLC 2008)*, pages 429-435

[12] Schönhofen, Péter and Benczúr, András A. Feature Selection based on word-sentence relation. *In Proceedings of International Conference on Machine Learning and Applications (ICMLA 2005)*, pages 37-42.

[13] Sogou Lab Resource: http://www.sogou.com/labs/resources.html

[14] ICTCLAS Open Source: http://www.ictclas.org/Down_OpenSrc.asp

[15] Liu, Tao. An Evaluation on Feature Selection for Text Clustering. *In Proceedings of International Conference on Machine Learning (ICML 2003)*, pages 488-495.