A Comparative Study on the Distribution of Chinese and English Multi-Words

Wen Zhang†　　Taketoshi Yoshida†　　Xijin Tang‡

† School of Knowledge Science, Japan Advanced Institute of Science and Technology, 1-1
Ashahidai, Tatsunokuchi, Ishikawa 923-1292, Japan
{zhangwen,yoshida}@jaist.ac.jp
‡ Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese
Academy of Sciences, Beijing 100080, P.R. China
xjtang@amss.ac.cn

Abstract

A study about the distribution of multi-words in both Chinese text and English text was carried out to explore a theoretical basis for probabilistic term-weighting scheme. Poisson distribution and G-distribution are comparatively studied to describe the relationship between words' frequency and number of occurrences, for both technical multi-words and non-technical multi-words. Also, a rule-based multi-word extraction algorithm was proposed to extract the multi-words from texts based on occurring structures and syntactical patterns in texts. Our experimental results demonstrated that G-distribution has a better capability than Poisson distribution in description of the relationship between multi-words' frequency and number of occurrences for technical multi-words and non-technical multi-words.

Keywords: multi-word, term-distribution, Poisson distribution, G-distribution

1 Introduction

Distribution of terms, which focused on frequencies of word occurrences along with other characteristics, has attracted great interest in the field of textual information processing such as information retrieval and speech recognition. Further, various term distribution models were proposed to capture the regularities of word occurrences in texts, and discover the underlying mechanisms of terms (words' behavior) in texts. A good understanding of distribution patterns is useful on occasions when we want to assess the likelihood of a certain number of occurrences of a specific term in a collection of texts.

In the area of text mining and information retrieval, term weighting is a necessary and crucial process if we want to transform the text into numerical vectors, so that statistical methods can be employed on these vectors for mining non-trivial patterns in texts. However, most term weighting methods, such as vector space model, are based on empirical observation and linguistic intuition, rather than theoretical analysis of the term distribution and properties in corpus or texts. Because of this reason, term distribution was studied to shed light on distinguishing between unimportant (function, non-content, semantically-unfocused) terms and important (content, topical, semantically-focused) terms in texts, according to explicit statistical characteristics. After the analysis of term distribution, important terms can be extracted from texts, and unimportant terms can be ignored accordingly, when numerical transformation is carried out in text representation.

Generally, features are associated with single words during the process from textual information to numerical vectors. However, there are some cases, such as technical papers and professional theses, where it helps to consider a group of words as a feature which is used to describe a specialized concept in that field. Multi-word features are not found too frequently in a document collection, but when they do occur they are often highly predictive and informative in explaining learning methods. While "multi-word" is the fundamental notion of this paper, this notion had no satisfactory formal definition until now. It can only be intuitively characterized: it occurs only in specialized types of discourse, and is often specific to subsets of domains, and when it occurs in general types of discourse or in a variety of domains it often has broader or more diverse meaning, for example, name entities, terminological noun phrases and so on. Although some work has been done in term distribution and many prominent proposals have been presented [1-2], little work has been done on the comparison of different term distribution models, which was carried out in this paper.

The rest of this paper is organized as follows. Two different term distribution models, Poisson model and G-model, are introduced in Section 2. Section 3 describes our data set and data preprocessing to examine these two models. Also, the multi-word extraction method is given based on structures and characteristics of text. The experiments on term distribution of Chinese and English multi-words are carried out, and the results are showed in Section 4. Finally, discussions and concluding remarks, and further research, are given in Section 5.

2 Term distribution models

The classical probabilistic term distribution models, such as Poisson model and G-model, are introduced in this section with their basic assumptions for the words occurring in a text.

2.1 Poisson distribution

The definition of the Poisson distribution is as follows.

$$\mathrm{p(k;}\lambda_{\mathrm{i}}) = e^{-\lambda_{\mathrm{i}}} \frac{\lambda_{\mathrm{i}}{}^{k}}{k!} \quad \text{for some } \lambda_{\mathrm{i}} > 0 \qquad (1)$$

In the most common model of the Poisson distribution in information retrieval, the parameter $\lambda_i > 0$ is the average number of occurrences of $\mathrm{w_i}$ per document, that is, $\lambda_i = \frac{cf_i}{N}$ where cf_i is the collection frequency and N is the total number of documents in the collection. With Poisson distribution, we can estimate the probability of a word occurring a certain number of times in a document. That is, $\mathrm{p(k;}\lambda_{\mathrm{i}})$ is the probability of a document having exactly k occurrences of w_i, where λ_i is appropriately estimated for each word. The basic assumption for the Poisson distribution is that the occurrences of a term are independent of each other, i.e., there is no correlation between the different occurrences of a term in documents. However, in most cases, this assumption may not hold, because of different occurrence patterns of content words and non-content word in texts. Based on this idea, the two-Poisson model and further Poisson mixtures were developed to estimate the probability of occurrences of a term, but they all have a variety of problems in practical application [3]. All these models have the same basic assumptions regarding word occurrence with Poisson distribution. To simplify, we only use Poisson distribution for estimation in this paper, because we should develop more complex models using Poisson distribution only if the simple Poisson distribution is validated as promising in some cases with our multi-words.

2.2 G-distribution

The G-distribution (G means general), also known as a three-parameter probability distribution, is defined as follows:

$$\begin{aligned}\Pr(k;\alpha,\gamma,\beta) &= (1-\alpha)\cdot\delta_{k,0} + (1-\gamma)\cdot\delta_{k,1} \\ &+ \frac{\alpha\gamma}{B-1}\cdot\left(1-\frac{1}{B-1}\right)^{k-2}\cdot(1-\delta_{k,0}-\delta_{k,1})\end{aligned} \qquad (2)$$

In practical application, $\alpha = \sum_{r\geq 1} p_r = 1 - p_0$, is the sum of frequency of terms whose occurrence is more than zero, and $\gamma = \frac{\sum_{r\geq 2} p_r}{\sum_{r\geq 1} p_r} = 1 - \frac{p_1}{1-p_0}$, is the proportion of the frequency of more than or equal to two to the frequency as more than or equal to one. $p_r = \Pr(k = r)$ is the probability of having exactly r instances of a term in a document. $B = \frac{\sum_{r\geq 2} p_r r}{\sum_{r\geq 2} p_r}$ is a measure of topical burstiness. Also, a Kroneker symbol was employed here as $\delta_{i,j} = \begin{cases} 0, i \neq j \\ 1, i = j \end{cases}$.The basic assumption with the G-distribution is that there are mainly two kinds of words existing in a document: one is non-content words, and the other is content words. As for the multi-words, they are usually content words in a document, and they can be separated into topical and non-topical words. When a content word is present in a document, but the concept named by a content word is not topical (non-topical) for that document, then this word would typically occur only once in this document. But when a concept named or expressed by a content word is topical for the document, then the content word is characterized by multiple (times>=2) and then often bursty occurrence. The notion of burstiness is fundamental for obtaining G-distribution, which means the multiple occurrence of a content word or phrase in some documents but in other documents, they do not occur at all. For more details about G-distribution, readers can refer to [4].

3 Data preprocessing

In order to compare both distribution models described above, two types of texts in different languages were selected as our experimental object. First, the profiles of those text collections were specified, and then a multi-word extraction method was proposed to extract the multi-words from both document collections in this section.

3.1 XSSC texts in Chinese and Reuters texts in English

Based on our previous work, 184 texts concerning the details of each conference were collected from XSSC Website (http://www.xssc.ac.cn), in which many academic topics of a wide scope from basic research to advanced techniques are included. In this paper these documents were used to conduct multi-word extraction, and their distribution characterization was performed. By our calculation, there are 14 categories related to this document collection, and the average to number of sentences per-document is 41.46, i.e., the average length of XSSC document.

The Reuters-21578 data set (http://www.research.att.com/~lewis.) was selected as our experiment text in English. It appeared as Reuters-22173 in 1991, and was indexed with 135 categories by personnel from Reuters Ltd in 1996. In the area of text mining, it was usually adopted as a bench marking data set for text categorization. But in this paper, the data was used to extract the English multi-words, and to observe the distributions of these multi-words in newswire texts. By our statistics, this data set contains in total 19403 valid texts, with an average of 5.4 sentences in each text. For convenience, the texts from 4 categories, “grain”, “crude”, “trade” and “interest” were selected as our target data set, because obvious distinctions instead of overlapping can be drawn between them which may benefit the distribution of the multi-words. With this method, 574 texts from “grain”, 566 texts from “crude”, 424 texts from “interest” and 514 texts from “trade” were assigned as our target data set.

3.2 Multi-word extraction

Basically, there are two types of methods to extract multi-words from documents: one is to utilize the mutual information between words, which is a statistical method [5-6], and another is to analyze the syntactical structure of multi-words, which is a rule-based method [7]. Usually, the multi-word extraction methods vary with different languages from a linguistic perspective; here a rule-based multi-word extraction method is proposed which is independent of language. From previous study on the structure and characteristics of multi-words, a conclusion was widely accepted that a multi-word has the properties of a noun phrase (NP) at its ending and repetition of occurrence. Accordingly, a simple hypothesis is that an NP having a frequency of two or more can be a candidate as a multi-word in a text [7]. With this hypothesis, we proposed a multi-word extraction method to extract the multi-words from both the Chinese and English texts. The basic idea for this method is to identify the repetitive patterns (a group of consecutive words) in sentences as candidates in a document first, and then determine the part of speech of these identified patterns. If a candidate’s part of speech is a noun (not a pronoun), it should be accepted as a multi-word. Otherwise, it should be refused as a multi-word. The following is our method to identify the repetition of any two sentences in the same document.

```
Input:
$s_1$, the first sentence
$s_2$, the second sentence
Output:
  Repetitive and consecutive words extracted from $s_1$ and $s_2$.

  Procedure:

  $s_1 = \{w_1, w_2, …, w_n\}$ $s_2 = \{w_1', w_2', …, w_m'\}$
k=0

  for each word $w_i$ in $s_1$
          for each word $w_j$ in $s_2$
           while($w_i = w_j$)
              k++
           end while
           if k>1
           combine the words from $w_i$ to $w_{i+k}$ as the output of this procedure
           End if
          End for
  End for
```

After the repetition is extracted from sentences in a document, the ICTCAS [8] and JWNL [9-10] were employed to determine the part of speech of the last word of the repetition, for Chinese and English respectively. Also, in the case that the last word is not a noun, such as “prime minister agreed”, the last noun of this repetition was determined, and “prime minister” was regarded as a multi-word. Moreover, the length and the alignment of each word also were considered to make the extraction more accurate, for example, multi-word usually has a length no more than 6 single words. For the XSSC documents, 5087 multi-words were extracted, and 4024 multi-words were extracted from Reuters texts. It is very interesting to notice that although the total number of documents of Reuters texts (2074) is far larger than the XSSC texts (184), the number of multi-words of both texts are approximately similar. We conjectured that this may be because the algorithm for multi-word extraction is applied to sentence directly. Reuters data sets have a total of 7628 sentences, while 11200 sentences are found in XSSC testing texts. Moreover, the types of text are different, XSSC text is about academic and scientific reports, while Reuters texts belong to brief news reports.

4 The distribution of Chinese and English multi-words in text

The comparison of Poisson distribution and G-distribution was conducted in XSSC and Reuters texts, respectively. Although traditionally the words in texts were separated into non-content words (function words, semantically-unfocused words) and content words (semantically-focused words, topical words), here, we only discriminate the technical multi-words and non-technical multi-words. Multi-words are semantically focused in their essence; otherwise there would not be accidents of combination of these words. Technical multi-word refers to a group of words which are highly related to the contents of the texts, such as terminological noun phrases, while non-technical multi-words are not so highly related to the content of the texts, for example the commonly used phrases in a field, and the names of places. Also two measures are introduced here to evaluate the performance of Poisson distribution and G-distribution. They are gross error as Ea and local error as El. Their definition is as follows.

$$E_a = \sum_{r \ge 0} |act - est| \tag{3}$$

$$E_l = \sum_{r \ge 2} |act - est| \tag{4}$$

Here, *act* is the actual frequency of multi-word occurrence in texts, and *est* is the estimated frequency of multi-word occurrence in texts by Poisson distribution or G-distribution. E_a was introduced to evaluate the overall estimation, and E_l was introduced to evaluate the local estimation, because there are never errors for G-distribution if $r \le 1$, according to its formula.

4.1 Overall distribution of the multi-words in XSSC and Reuters texts

Before the distribution of single multi-words was examined, the overall distributions of the multi-words in XSSC and Reuters texts are plotted, as shown in Figure 1 and Figure 2. The overall distributions of the multi-words in XSSC and Reuters text have a very similar curves. Their only difference is that the latter has a larger range of number of occurrences (count) and frequency. Moreover, they conform to Zipf's law [11], which states that the product of the frequency and the rank order (order of count) is approximately constant and is adopted by some practical information retrieval applications [12]. We verified this statement in our research for both XSSC and Reuters. For the multi-words of XSSC, this constant is about 0.02-0.1 for most cases but with some exceptions at the smaller order. For the multi-words of Reuters, also the constant is about 0.02-0.1 for most cases also with some exceptions at the smaller order.

4.2 Distributions of technical multi-words

For the technical multi-words of XSSC, "纳米材料"(nanophase materials) and "生态环境"(ecological environment) were assigned as the testees, because they are the hot topics in new technology areas, only some professional articles may have these words. And for the technical multi-words of Reuters, we selected "crude oil" and "interest rates" as our samples, because they are the topics of the categories we picked out from Reuters texts. Tables 1-4 show the results of the Poisson distribution and G-distribution on these examined multi-words.

From Tables 1-4, it can be seen that the G-distribution has a better capability in estimating the probability of exactly $r(\geq 0)$ occurrences of a multi-word in texts. In addition, it is obvious that the estimation error of the Poisson distribution of 0 and 1 occurrences has a significant proportion of the overall estimation error for $E_a >> E_l$ in most cases of Poisson distribution. Furthermore, the estimation based on Reuters texts are better than the estimation based on XSSC texts, as the estimation based on Reuters texts always has less error using any Poisson distribution or G-distribution. We will discuss this phenomenon in Section 5.

4.3 Distributions of non-technical multi-words

For the non-technical multi-words of XSSC, "基础研究"(basic research) and "科学问题"(scientific problem) were assigned as the testees, because they are popular words in XSSC academic discussion and have a very extensive meaning other than a concrete professional concept. For the non-technical multi-words of Reuters, we selected "United States" and "Soviet Union" as our samples, as they are the names of countries and can be used anywhere related to these two countries in newswire reports. Tables 5-8 show the results of the Poisson distribution and G-distribution on these examined multi-words.

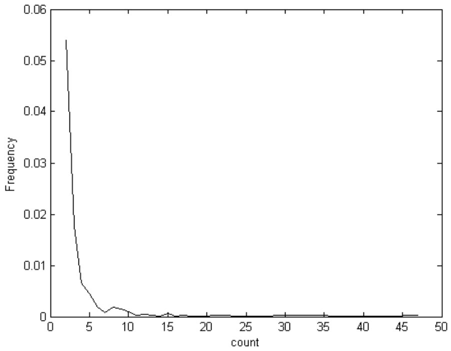

Figure 1.The overall distribution of the numbers of multi-words and their frequency in XSSC

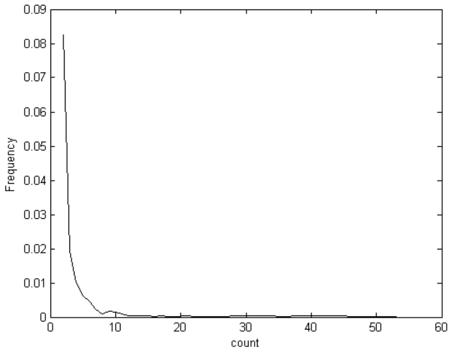

Figure 2.The overall distribution of the numbers of multi-words and their frequency in Reuters

It has been shown in Tables 5-8 that the G-distribution has a better capability in estimating the probability of exactly $r(\geq 0)$ occurrences of non-technical multi-words in texts. As with the technical multi-words, it is clear that the estimation error for the Poisson distribution at 0 and 1 occurrences has a significant proportion of the overall estimation error, and the estimation based on Reuters texts are better than the estimation based on XSSC texts. However, when the results were compared with the former results on technical multi-words, it can be seen that in XSSC, the technical multi-words have less estimation error than non-technical multi-words, but when it comes to Reuters texts, it is the opposite case, i.e., non-technical multi-words have less error than technical multi-words. We will discuss this phenomenon in Section 5.

5 Discussion and Concluding Remarks

In this paper a study on the distribution of multi-words in texts was carried out. Two kinds of models, Poisson distribution and G-distribution, were examined in Chinese and English texts, using as XSSC text collection and Reuters data set. Moreover, a simplified multi-word extraction method was proposed to extract the multi-words from texts independent of language, based on the structures and syntactical rules of multi-words in texts.

Table 1. The distribution of Chinese multi-word “纳米材料” and its probability estimation from Poisson and G-distribution.

r	0	1	3	4	5	8	11	15	17	26	E_a	E_l
act.($\times 10^{-2}$)	89.67	4.89	1.09	0.54	0.54	1.09	0.54	0.54	0.54	0.54		
Poisson est.($\times 10^{-2}$)	55.30	32.76	1.92	0.28	0.03	0.00	0.00	0.00	0.00	0.00	67.09	4.85
G-model est.($\times 10^{-2}$)	89.67	4.89	0.60	0.54	0.42	0.30	0.21	0.13	0.10	0.04	3.08	3.08

Table 2. The distribution of Chinese multi-word “生态环境” and its probability estimation from Poisson and G-distribution.

r	0	1	2	3	4	5	6	9	12	13	27	E_a	E_l
act.($\times 10^{-2}$)	80.98	7.07	4.35	1.63	1.09	1.09	1.63	0.54	0.54	0.54	0.54		
Poisson est. ($\times 10^{-2}$)	48.01	35.23	12.92	3.16	0.58	0.09	0.01	0.00	0.00	0.00	0.00	76.45	15.32
G-model est. ($\times 10^{-2}$)	80.98	7.07	2.63	2.05	1.60	1.25	0.97	0.46	0.22	0.17	0.01	4.77	4.77

Table 3. The distribution of “crude oil” and its probability estimation from Poisson and G-distribution

r	0	1	2	3	4	6	8	E_a	E_l
act.($\times 10^{-2}$)	90.51	6.21	2.45	0.59	0.10	0.10	0.05		
Poisson est. ($\times 10^{-2}$)	86.73	12.35	0.88	0.04	0.00	0.00	0.00	12.29	2.37
G-model est. ($\times 10^{-2}$)	90.51	6.21	2.26	0.70	0.22	0.02	0.00	0.55	0.55

Table 4. The distribution of “interest rates” and its probability estimation from Poisson and G-distribution

r	0	1	2	3	4	5	6	8	E_a	E_l
act.($\times 10^{-2}$)	92.22	4.26	2.01	0.98	2.04	0.10	0.15	0.05		
Poisson est. ($\times 10^{-2}$)	86.99	12.13	0.85	0.04	0.00	0.00	0.00	0.00	17.54	4.44
G-model est.($\times 10^{-2}$)	92.22	4.26	2.01	0.86	0.37	0.16	0.07	0.01	1.97	1.97

Our experimental results have shown that G-distribution has a better capability in describing the relationship between multi-word occurrence and frequency than the Poisson distribution. This validates the basic assumptions in G-distribution about the existence of word burstiness in texts, regarding content words. Especially the inability of Poisson distribution to estimate the probability of exactly 0 and 1 occurrence enhanced the assumption that the occurrences of multi-words in text may not be feasibly regarded as independent from each other. Although some researchers argued that the two-Poisson model or negative binomial may be another way to solve this kind of problem [13], and that the problem with Poisson is that documents differ widely in size [14], the basic assumption of Poisson model should be reconsidered.

Table 5. The distribution of Chinese multi-word “基础研究” and its probability estimation from Poisson and G-distribution.

R	0	1	2	3	4	5	6	7	8	12	18	E_a	E_l
act.($\times 10^{-2}$)	49.46	20.11	11.41	6.52	4.35	2.17	3.26	0.54	1.09	0.54	0.54		
Poisson est. ($\times 10^{-2}$)	24.88	34.61	24.08	11.17	3.88	1.08	0.25	0.05	0.01	0.00	0.00	63.62	24.54
G-model est. ($\times 10^{-2}$)	49.46	20.11	10.46	6.86	4.51	2.96	1.94	1.27	0.84	0.16	0.01	5.45	5.45

Table 6. The distribution of Chinese multi-word “科学问题” and its probability estimation from Poisson and G-distribution.

r	0	1	2	3	4	5	6	7	8	9	10	11	13	E_a	E_l
act.($\times 10^{-2}$)	49.46	20.65	12.50	5.98	3.26	1.63	2.17	1.09	0.54	0.54	1.09	0.54	0.54		
Poisson est. ($\times 10^{-2}$)	25.01	34.66	24.02	11.10	3.84	1.07	0.25	0.05	0.01	0.00	0.00	0.00	0.00	67.3	28.84
G-model est. ($\times 10^{-2}$)	49.46	20.65	10.15	6.70	4.43	2.92	1.93	1.28	0.84	0.56	0.37	0.24	0.11	12.59	12.59

Table 7. The distribution of “United States” and its probability estimation from Poisson and G-distribution

r	0	1	2	3	4	5	6	9	10	E_a	E_l
act.($\times 10^{-2}$)	87.38	7.97	2.50	1.27	0.59	0.15	0.05	0.05	0.05		
Poisson est.($\times 10^{-2}$)	80.99	17.08	1.80	0.13	0.01	0.00	0.00	0.00	0.00	18.22	2.72
G-model est.($\times 10^{-2}$)	87.38	7.97	2.55	1.15	0.52	0.23	0.11	0.01	0.00	0.46	0.46

Table 8. The distribution of “Soviet Union” and its probability estimation from Poisson and G-distribution

r	0	1	2	3	4	5	E_a	E_l
act.($\times 10^{-2}$)	94.96	3.13	1.32	0.39	0.10	0.10		
Poisson est. ($\times 10^{-2}$)	92.47	7.24	0.28	0.01	0.00	0.00	8.22	1.62
G-model est. ($\times 10^{-2}$)	94.96	3.13	1.31	0.41	0.13	0.04	0.12	0.12

However, some questions have also come up with our experimental results. The first one is that the estimations on Reuters texts are better than the estimations based on XSSC texts. Perhaps it is because the XSSC texts are academic papers, and they have more terminological noun phrases but fewer texts than Reuters text, so that the multi-word behavior is not fully expressed in XSSC texts. The second possible reason is that the distribution of technical multi-words is in agreement with G-distribution than non-technical words, but this is not the case in Reuters texts. The reason for this point is also possibly because the XSSC texts are academic texts, so that the burstiness can more easily induced in their texts but the Reuters texts are newswire texts focused on more information included in short passages, so that the bustiness of content words can not happen in them naturally. Whatever the final reasons for these differences, further investigations are required to disclose these phenomena.

As for our further research, the term weighting methods based on term distribution theory are a potential direction in which to advance, especially for the multi-word features. For example, the importance of a multi-word in a text can be objectively measured and used for text representation, if the distribution of multi-word and the relative frequency of the multi-word in this text can be acquired from the text collection. Nevertheless, term distribution also provids a theoretical support for advanced applications such as speech recognition and machine translation, if a probability model is elaborated for the distribution of terms in text collections.

Acknowledgment

This work is supported by Ministry of Education, Culture, Sports, Science and Technology of Japan under the “Kanazawa Region, Ishikawa High-Tech Sensing Cluster of Knowledge-Based Cluster Creation Project” and partially supported by the National Natural Science Foundation of China under Grant No.70571078 and 70221001.

References

[1]. Anne De Roeck, Avik Sarkar, Paul Garthwaite. Frequent Term Distribution Measures for Dataset Profiling. Online: http://mcs.open.ac.uk/as5297/papers/LREC%20frequent% 20der oeck%2004.pdf

[2]. Bookstein, A. Swanson, D.R. Probabilistic models for automatic indexing. Journal of the American Society for Information Science, 25 (5): 312-318, 1974.

[3]. Kenneth W. Church and William A. Gale. Poisson mixtures. Natural Language Engineering, 1(2), 163--190, (1995).

[4]. S.Katz. Distribution of content words and phrases in texts and language modeling. Natural Language Engineering, 2(1): 15-59. 1996.

[5]. Mikio Yamamoto, Kenneth W. Church. Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. Proceedings of the 6th Workshop on very large corpora, Montreal, Canada, 1998, pp. 285-313.

[6]. Boxing Chen, Limin Du. Preparatory Work on Automatic Extraction of Bilingual Multi-Word Units from Parallel Corpora. Computational Linguistics and Chinese Language Processing. Vol.8, No. 2, August 2003, pp.77-92.

[7]. Justeson,F., Katz, S.M. Technical terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering, 1(1): 9-27. 1995.

[8]. Huaping, Zhang., Qun, Liu. ICTCLAST2.0 online: http:// www.ict.ac.cn/freeware/.

[9]. Java WordNet Library. Online: http:// sourceforge.net/projects/jwordnet.

[10]. WordNet: a lexical database for the English language. http://wordnet.princeton.edu/.

[11]. Zipf, George Kingsley. Human Behaviour and the Principle of Least Effort, Addison-Wesley, Cambridge, Massachusetts (1949)

[12]. Luhn,H.P. The automatic creation of literature abstracts, IBM Journal of Research and Development. 2, 159-165 (1958).

[13]. Harter, S.P. A probabilistic approach to automatic keyword indexing. Journal of the American Society for Information Science, 26,197-206, (1975).

[14]. Hinrich Scheutze, Christopher D. Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press. May, 1999.