Journal of Theoretical and Applied Information Technology
10th March 2013. Vol. 49 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN APPROACH BASED ON ITERATIVE LEARNING ALGORITHM FOR CHINESE TEXT HIERARCHY FEATURE EXTRACTION WITHOUT LEXICON

[1]SHAOHUA JIANG
[1]Asstt Prof., Faculty of infrastructure engineering, DALIAN UNIVERSITY OF TECHNOLOGY
E-mail: [1]shjiang@dlut.edu.cn

ABSTRACT

A great deal of information included in Chinese text is invaluable asset for further text mining, but the difference between Chinese and the western languages imposes restrictions on further utilization of Chinese text. No distinction indication between words by using spaces is one of the major differences between Chinese, also some other Asian languages, such as Japanese, Thai, etc., and Western languages. Chinese segmentation and features extraction is essential in Chinese natural language processing because it is a precondition for further Chinese text information retrieval and knowledge discovery. Maximum matching and frequency statistics (MMFS) segmentation method based on length descending and string frequency statistics is an effective segmentation and extraction method for Chinese words and phrases, but there are still some shorter words and phrases included in the longer ones extracted by MMFS can't be obtained. In order to solve this problem, this paper presents a novel Chinese hierarchy feature extraction method combined MMFS with iterative learning algorithm. This method can extract hierarchy feature according to morphology with no need for lexicon support, no need for acquiring the probability between words in advance and no need for Chinese character index. Experimental results confirm the efficiency of this statistical method in extracting Chinese hierarchy feature. This method is also beneficial to feature extraction for other Asian languages similar to Chinese.

Keywords: *Hierarchy Feature Extraction, Chinese text, Iterative Learning Algorithm*

1. INTRODUCTION

A significant amount of information is located in text documents, so it's critical to handle unstructured text data by utilizing information technology (IT). Successful research efforts can be beneficial to text information management and knowledge discovery. There is a unique challenge to handle Chinese text compared to English text [1]-[2]. Chinese, also some other Asian languages, such as Japanese, Thai, etc., are quite different from Indo-European languages, such as English, French, German, etc. In Chinese language, the word is the smallest independent meaningful element and the character is the basic written unit. There is no delimiting space to separate words and obvious punctuations only exist between clauses in Chinese text [3]-[4]. So it's difficult for computer to process Chinese text unless it has been separated into individual words and phrases in advance, so segmentation is a prerequisite step in Chinese texts processing, such as machine translation, information retrieval and text mining [5]-[6]. Words and phrases are basic elements of text and form the feature of text presentation, so feature extraction is a basic precondition of further text mining.

The identifying of distinct words in English or other Indo-European languages texts is trivial task. However, it is a difficult task for Chinese as a non-space-delimited language [7]-[8]. Chinese texts consist of a string of ideographic characters without any delimiters to indicate word boundaries between words except for punctuation signs at the end of each sentence, and occasional commas within sentences [9]-[11]. The goal of Chinese text feature extraction is thus to transform a plain Chinese text to a series of meaningful words and phrases.

The maximum matching and frequency statistics (MMFS) method [12]-[13] can segment Chinese text and extract feature without the support of lexicon. Special terms and proper nouns can be extracted effectively by means of MMFS, unfortunately, shorter words and phrases included in the longer ones extracted by MMFS can't be extracted which hinder MMFS' effectiveness [14]. Therefore, this paper proposes a novel, statistics-based scheme of Chinese text hierarchy feature

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

extraction by iterative learning algorithm on the basis of MMFS. Experimental results revealed that the proposed scheme of feature extraction can effectively distill hierarchy feature for Chinese text without lexicon.

The rest of this paper is organized as follows. Section 2 gives a brief overview of previous work of Chinese text segmentation. Section 3 introduces the maximum matching and frequency statistics segmentation method. Section 4 describes the details of the approach for Chinese text hierarchy feature extraction without lexicon based on iterative learning algorithm. Section 5 reports the experimental results. Finally Section 6 concludes the research findings and discusses the directions in future.

2. PREVIOUS WORK

In general, approaches to Chinese text segmentation proposed before can be divided into four categories: segmentation based on lexicon [15], the method based on syntax and rules [16], the method based on statistics, for example the N-gram method [17], and the hybrid method [18].

2.1 Lexicon-Based Method

Lexicon-based method is the most basic and intuitive segmentation method for Chinese text. It performs segmentation process using string matching algorithm supported by a predefined lexicon with sufficient amount of lexical entries which covers as more Chinese words as possible. Unfortunately, such a large lexicon is difficult to be constructed or maintained by manpower since the set of words is open-ended. Therefore, many new proper nouns and specialized terms which appear continuously as a result of the intersection and fusion of various subjects are often out-of-lexicon words due to insufficient amount of lexical entries so that the accuracy of Chinese segmentation is often limited.

2.2 Syntax And Rules-Based Method

In segmentation method based on syntax and rules, segmentation and syntactic and semantic analysis are processed synchronously. It utilizes syntactic and semantic information to carry out part of speech tagging and solves the segmentation ambiguity problem. Due to the increase of the quantity of the existing syntax knowledge and rules, it's difficult to avoid conflict between the knowledge and rules because they are too general and complex, so the precision of this method isn't satisfying and still need further development.

2.3 Statistics-Based Method

To overcome the shortcoming of the methods based on lexicon, syntax and rules, statistics-based method was presented which relies on statistical information, such as word and/or character occurrence frequencies. Most statistics-based methods use N-gram model, which in fact is an N-1 order Markov process. The N-gram model analyzes large amounts of test data statistically, and then provides transitional probabilities from the prior N-1 words to the next word [19]. In spite of its popularity, due to the limit of computation cost in actual application, the N-gram model often takes into account only several historical information and then forms models like bigram, trigram and so on.

For Chinese character string (CCS) composed of L words, it contains $L(L+1)/2$ information items when N's value is from 1 to L. So N-gram model's computation cost rises dramatically as N increases.

In sum, N-gram model suffers from three primary drawbacks. Firstly, the amount of training corpus can't include all language phenomena with the increase of application fields; the main problem of N-gram model is estimate the probability of these language phenomena accurately. Secondly, the existing hardware is difficult to meet the requirement of the computation cost of N-gram model with the increase of corpus and the number of N. Thirdly, Chinese character strings obtained by N-gram model lack semantic meaning. The above drawbacks limit N-gram model's applicability.

2.4 Hybrid Method

The hybrid method combines part of the above methods, but it still cannot avoid the shortcomings of its each part radically.

To deal with the above-mentioned weaknesses of the previous methods, a method named MMFS was proposed. This statistics-based method can extract CCS whose support degree is bigger than a predefined value.

3. THE MMFS METHOD

3.1. Basic Idea

It is still a controversial problem that whether the basic processing unit in Chinese is word or phrase owing to the characteristics of Chinese. In practice, the definition "word is the smallest language element, it can be used independently and has semantic meaning" lacks of operability. Phase has steady structure, so phase should also be regarded as the basic processing unit.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The main characteristics of word in Chinese text are as follows:

(1) If a CCS has a higher frequency, the possibility of it being a word is higher.

(2) Only the CCS with unambiguous semantic can be a word.

(3) The Chinese character' combination mode can be observed in statistical sense.

(4) The shorter word has higher frequency and it is function-oriented. On the contrary, the longer word has lower frequency and it is content-oriented.

As a result of the above characteristics of Chinese text, the corresponding processing technology is different from English. Text content is the basis of text processing, so a statistics-based Chinese text segmentation method was put forward [12]. On the basis of segmentation of CCS by segmentation tag in the preprocessing phase, the CCS's frequency is counted according to the principle of matching the longer string first. The content-oriented CCS including more Chinese character is processed first, then the length of CCS intended to be segmented is reduced and the corresponding CCS's frequency is analyzed. So the segmentation of Chinese text can be finished without the support of thesaurus and preliminary probability estimation.

3.2. Algorithm Design

In order to facilitate further discussion, some definitions are given in advance.

Definition 1: Text string is all the strings in text, including Chinese and non-Chinese character.

Definition 2: Chinese character string (CCS) is all the strings comprising of consecutive Chinese character in text.

It's straightforward that the more segmentation tags exist, the more possible for the longer text string to be segmented into shorter CCS, and it will be more convenient for subsequent processing. To explain the segmentation of text more explicitly, the input Chinese text is denoted by T and some definitions are given as follows.

Definition 3: Segmentation Denotation (SD) is the set of denotations which can't appear in phrase and word of Chinese text. It comprises natural segmentation denotation and non-natural segmentation denotation. Natural segmentation denotation includes punctuations, non-natural segmentation denotation includes number and non-Chinese character.

Definition 4: Segmentation String (SS) is the set of the CCS which has definite meaning and can be used independently and the separate Chinese character which can't form phase or word with other Chinese character. SS can form part of segmentation result directly. SS comprises the empty word which has high frequency and is consisted of one or two Chinese character and the substantive which has high frequency.

Definition 5: Preprocessed String (PS) is the set of the CCS which is formed after the preprocessing through the use of SD and SS.

Definition 6: Candidate String (CS) is the set of CCS which is formed after the segmentation by length descending and string frequency statistics based on PS.

Definition 7: Support degree is the frequency of the CCS in the text. The predefined support degree is denoted by Φ, where $\Phi \geq 2$.

Definition 8: Segmentation Result (SR) is the set of CCS which is filtered by Φ on the basis of CS.

Definition 9: Special indicatory semantic CCS is composed of phrases, words or the combination of phrases and words and has more special indicatory semantic property than phrase.

To explain the algorithm in detail, let C represents the set of all Chinese character, NC represents the set of all non-Chinese character. Λ denotes blank, $\Lambda \in NC$. So $T = C \bigcup NC$. The proposed algorithm includes five main steps, namely preliminary processing, further processing, automatic segmentation, filtered by predefined support degree, feedback. The above five steps are described as follows.

First step. Preliminary processing. The input Chinese text T is processed by SD and paragraph compartmentation. If there is a denotation belonging to SD in text, the denotation is replaced by Λ. So T_1, the set of short CCS, is formed, $T_1 = C \bigcup \{\Lambda\}$.

Second step. Further processing. T_2 is the set of CCS which is formed by further preliminary processing by SS and comprises more blanks. The continuous Λ in T2 is incorporated into one, so PS, the result of incorporation, is formed.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$PS = c_1 c_2 \cdots\cdots c_n$, $c_i \in C \bigcup \{\Lambda\}, 1 \le i \le n$, and c_i , c_{i+1} can't be Λ simultaneously.

Third step. Automatic segmentation. According the principle of processing longer CCS first and length descending, the frequency of the string belonging to PS in the context is computed. The CCS whose concurrent frequency is more than 1 is extracted. So automatic segmentation is finished and CS is formed.

Fourth step. Filtered by predefined support degree. The CCS whose support degree is more than or the same as Φ forms the final segmentation result and SR is produced. Φ can be changed with the different length of text.

Fifth step. Feedback. The new discovered segmentation denotation and Segmentation String based on CS are added to SD and SS, by doing so, the system's capability is improved. Feedback is an optional function.

Matching CCS from left to right and the longer CCS first, so it is a maximal matching method with left combination first.

With regard to the time spending of this algorithm, assume that the length of maximal CCS to be extracted is L , the number of total CCS of text after preprocessing is N . Only the CCS including L to 2 Chinese character is extracted. The time complexity in the worst case is $LN^3/2$, i.e. the time complexity is $O(N^3)$ in the worst case, but the actual time requirement is far less than this value.

This method doesn't segment single Chinese character without reference to its frequency, because single Chinese character has no information itself and is useless for text classification and retrieval in practice.

4. CHINESE HIERARCHY FEATURE EXTRACTION BY ITERATIVE LEARNING ALGORITHM

4.1 Basic Idea Of Chinese Hierarchy Feature Extraction Method

Since most single Chinese character has no ideographic information and can't express the meaning clearly, so only the feature including more than single Chinese character is taken into account in this paper [20].

To be convenient for discussion, some definitions are given here.

Definition 10: Original Feature (OF) is the set of segmentation result of MMFS. $OF = \{T_1, T_2, \cdots T_n\}$, T_i is the element of Segmentation Result (SR), $|T_i| \ge 2$, $|T_i|$ indicates the length of T_i , namely the number of Chinese character included in T_i , $T_i \ne T_j$, when $i \ne j$, $1 \le i \le n$, $1 \le j \le n$.

Definition 11: Hierarchy Feature (HF) is the set of segmentation result of MMFS in OF . The words and phrases in HF are comprised jointly by some words and phrases in OF , so the feature in HF has semantic hierarchy relation with the feature in OF which contains it, that is the origin of the word HF . The frequency of feature in HF is larger than 3 because support degree $\Phi \ge 2$ in MMFS. $HF = \{F_1, F_2, \cdots F_m\}$, $|F_i| \ge 2$, $F_i \ne F_j$, when $i \ne j$, $1 \le i \le m$, $1 \le j \le m$. There are two situations about F_i : $F_i \in OF$ or $F_i \notin OF$. Here is an example of hierarchy feature, in case of $OF = \{C_1C_2C_3C_4, C_3C_4C_5C_6\}$, C_i indicates Chinese character, then $HF = \{C_3C_4\}$, C_3C_4 is the hierarchy feature comprised by $C_1C_2C_3C_4, C_3C_4C_5C_6$ jointly.

Definition 12: Iterative Feature (IF) is the union of OF and HF . $IF = OF \bigcup HF = \{W_1, W_2, \cdots W_l\}$, $l \le n+m$, $|W_i| \ge 2$, $W_i \ne W_j$, when $i \ne j$, $1 \le i \le l$, $1 \le j \le l$.

Definition 13: Iterative Feature Information (IFI) is the set of the feature included in IF and its frequency and length. $IFI = \{(W_i, Freq_i, Len_i)\}$, $1 \le i \le l$.

There are special indicatory semantic CCS, phrase and word in OF . In order to extract more comprehensive features of Chinese text, especially those hierarchy features which are not included in OF , iterative learning algorithm is presented in this paper. The basic idea of Chinese hierarchy feature extraction method based on iterative learning algorithm is as follows: segment Chinese text by MMFS on the basis of OF and incorporate the extracted hierarchy feature which is not belonged to OF into OF to match continuously, then get IF by combination of HF and OF . For example, supposed OF ={知识发现(knowledge discovery), 知识管理(knowledge management)}, then a hierarchy feature comprised by the two features in OF is "知识(knowledge)", so IF ={知

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

识发现(knowledge discovery), 知识管理(knowledge management),知识(knowledge)}.

Known from above basic idea, $OF \subseteq IF$, $HF \subseteq IF$. The hierarchy feature extracted by iterative learning algorithm has higher frequency, which is larger than 3, so it can figure Chinese text more comprehensively and provides a foundation for further Chinese text modeling.

4.2 Design Of The Hierarchy Feature Extraction Algorithm

The proposed iterative learning algorithm includes four steps.

First step. Segment Chinese text by MMFS and take the result as OF , $OF = \{T_1, T_2, \cdots T_n\}$, $T_i, 1 \le i \le n$ is the element of segmentation result set by MMFS. Let $HF = \varphi$.

Second step. Let $S = \sum_{i=1}^{n-1}(T_i + \Lambda) + T_n$, S is the set of CCS, $T_i \in OF$, Λ indicates blank. The symbol "+" indicates connection of strings. *MaxLen* indicates the length of the longest features included in OF , $MaxLen = Max|T_i|$, $1 \le i \le n$.

Third step. Match CCS on S according the length which is shorter than a CCS's length (the length of one Chinese character is 2) of *MaxLen* by MMFS. If a hierarchy feature F_i is extracted in the matching process, $HF = HF + \{F_i\}$. If $F_i \notin OF$, $OF = OF + \{F_i\}$. Let $S = \sum_{i=1}^{n-1}((T_i - F_i) + \Lambda) + (T_n - F_i) + \Lambda + F_i$, $(T_i - F_i)$ indicates that if T_i includes F_i , then delete F_i from T_i . $(T_n - F_i)$ represents that if T_n includes F_i , then delete F_i from T_n . Then matching process is continued on the basis of S until the end.

Fourth step. Let $IF = OF = \{W_1, W_2, \cdots W_i\}$, compute each W_i 's frequency and length in Chinese text then form the final *IFI* .

The pseudocode of the third step is shown in Figure 1.

```
while k > sl
        do fp
                bp = the position of the first blank after fp
                do tk = the CCS under the circumstance of continuous blanks between
fp and bp are deleted
                if ( tk's length < k)
                        start from the next CCS
                else
                        do tk = the CCS whose length is k started from fp
                        if can't match tk started from fp
      extract CCS whose length is k from the next Chinese character
 else
                extract the matched CCS and amalgamate it to HF
                delete all matched CCS in S and connect the matched CCS to S by
Λ
                if the matched CCS is not included in OF
                     then amalgamate the matched CCS to OF
        fp = fp + 1
         bp = the position of the first blank after fp
k = k – 2
```

Figure 1. Pseudocode Of The Third Step

In Figure 1, k is the length of the CCS intended to be extracted, owing to the length of Chinese character is 2, $k = MaxLen - 2$, fp is the beginning position of S , bp is the first blank position after fp, tk is the continuous CCS whose length is k from S , sl is the predefined shortest length of CCS need to be extracted.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

HF and *IF* can be formed after the above process, and *IFI* can be formed accordingly at last.

5. EXPERIMENTS AND RESULTS

5.1 Effect Of Chinese Feature Extraction By Iterative Learning Algorithm

Take a Chinese scientific research paper about knowledge management and information system project management for example. Choose the text's title, abstract and keywords as experimental corpus, *OF* and *HF* are as follows respectively:

OF={信息系统项目管理(information system project management), 知识管理(knowledge management),信息系统(information system),模型(model), 项目(project)}. *HF* is formed after iterative learning, *HF* ={信息系统(information system), 项目(project), 管理(management), 知识(knowledge)}. "管理(management)" is a new hierarchy feature not included in *OF* .

The final *IF* is the union of *OF* and *HF* , so *IF* ={信息系统项目管理(information system project management), 知识管理(knowledge management),信息系统(information system),模型(model),项目(project),管理(management)}.

So the method of extraction Chinese text hierarchy features by iterative learning algorithm not only can extract the feature included in *OF* , the more important thing is that it can extract the feature not included in *OF* . So it lays a foundation for further Chinese text mining.

5.2 Experiment And Analysis Of Chinese Hierarchy Feature Extraction

To verify the Chinese hierarchy feature extraction method presented by this paper, 74 Chinese scientific research papers about management and information science are taken as experiment corpus. The length of these papers ranges from 0.8K to 40.7K. In this paper support degree Φ is set to 2 because of the text length is limited.

5.2.1. The number relation of hierarchy feature and feature in *of*

The relation between the number of hierarchy feature and the feature in *OF* is shown as Figure 2.

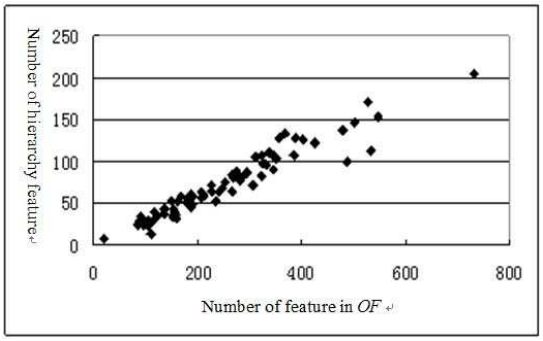

Figure 2. Number Of Hierarchy Feature And Number Of Feature In OF

It can be drawn from figure 2 that, on the whole, the number of Chinese text hierarchy feature increases monotonously with the increase of the number of feature in *OF* .

So in general, the longer of the text is, the number of feature in *OF* will increase, and the number of hierarchy feature will increase correspondingly.

5.2.2. Extraction chinese hierarchy feature not included in *of*

By statistics of experimental results of the above 74 Chinese scientific research papers, the percentage of Chinese hierarchy feature not included by *OF* in the total number Chinese hierarchy feature is 28.6%. The percentage of Chinese hierarchy feature not included by *OF* in *OF* is 3.1%. The accuracy of Chinese hierarchy feature not included by *OF* is about 80.2%.

The percentage of improper CCS constituted by 3 Chinese characters in all improper CCS is 75.8%. So, the most Chinese hierarchy feature not included in *OF* is accurate and a majority of improper CCS is the CCS constituted by 3 Chinese characters.

The percentage of Chinese hierarchy feature not included in *OF* by iterative learning in *OF* is not very high, but due to the high frequency nature (larger than 3), it's important for the hierarchy feature to be extracted to represent the Chinese text more comprehensive and lay a solid foundation for further Chinese text modeling and text mining.

6. CONCLUSION AND FUTURE WORK

With the growing importance of information retrieval applications based on text, there is a need to extract feature of text in support of further text processing. Due to the difference between Chinese

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

and western languages, feature extraction method need to be developed based on Chinese characteristics. In this paper, a novel Chinese hierarchy features extraction method based MMFS and iterative learning algorithm is presented to improve the effect of Chinese text expression. The presented work aims to extract Chinese hierarchy feature which has semantic relation with original feature according to morphology under the circumstance of no support of lexicon and no text corpus studying beforehand. The experimental results demonstrate that the proposed method can effectively obtain the Chinese hierarchy feature, such as Special indicatory semantic CCS, phrase and word in Chinese text, including new universal words, special terms and proper nouns. This paper lays a sound foundation for developing comprehensive model text in Chinese or other similar Asian languages.

The semantic relation between the extracted Chinese hierarchy feature and the original feature is important to develop ontology automatically, so methods which can improve the precision of Chinese hierarchy feature extraction and catch relation of concepts in domain ontology automatically need to be further developed in the future.

ACKNOWLEDGEMENT

Financial support (Grant No. 51178084) from National Natural Science Foundation of China is gratefully acknowledged. Thank the anonymous reviewers for their valuable comments.

REFRENCES:

[1 W. Li, T. Zhang, W. Lin, S. Deng, J. Shi, and C. Wang, “A new method for Chinese text content identification”, *Journal of Theoretical and Applied Information Technology*, Vol. 44, No. 1, 2012, pp. 131-136.

[2] F. Li, J. Li, “Studying of classifying Chinese SMS messages based on Bayesian classification”, *Journal of Theoretical and Applied Information Technology*, Vol. 44, No. 1, 2012, pp. 141-146.

[3] C. Fu, C. Zhuang, X. Gao, H. Wang, and Y. Wei, "Research on and construction of network question-answering system", *International Journal of Digital Content Technology and its Applications*, Vol. 6, No. 17, 2012, pp. 279-286.

[4] X. Zhu, Q. Cao, and F. Su, “A Chinese intelligent question answering system based on domain ontology and sentence templates”, *Journal of Theoretical and Applied Information Technology*, Vol. 5, No. 11, 2011, pp. 158-165.

[5] Y. Zhang, F.S. Tsai, and A.T. Kwee, "Multilingual sentence categorization and novelty mining", *Information Processing & Management*, Vol. 47, No. 5, 2011, pp. 667-675.

[6] R. T. Tsai, "Chinese text segmentation: A hybrid approach using transductive learning and statistical association measures", *Expert Systems with Applications*, Vol. 37, No. 5, 2010, pp. 3553-3560.

[7] M. Chau, J. Qin, Y. Zhou, C. Tseng, and H. Chen, "SpidersRUs: Creating specialized search engines in multiple languages", *Decision Support Systems*, Vol. 45, No. 3, 2008, pp. 621-640.

[8] S. Lo, "Web service quality control based on text mining using support vector machine", *Expert Systems with Applications*, Vol. 34, No. 1, 2008, pp. 603-610.

[9] S. Foo, and H. Li, “Chinese word segmentation and its effect on information retrieval”, *Information Processing and Management*, Vol. 40, No. 1, 2004, pp. 161-190.

[10] M. Zhang, Z. Lu, and C. Zou, “A Chinese word segmentation based on language situation in processing ambiguous words”, Information Sciences, Vol. 162, No. 3-4, 2004, pp. 275-285.

[11] C. Hong, C. Chen, and C. Chiu, “Automatic extraction of new words based on Google News corpora for supporting lexicon-based Chinese word segmentation systems”, *Expert Systems with Applications*, Vol. 36, No. 2, 2008, pp. 3641-3651.

[12] S. Jiang, and Y. Dang, “An Automatic Segmentation Method Combined with Length Descending and String Frequency Statistics for Chinese Text”, *the 6th International Symposium on Knowledge and Systems Sciences (KSS2005)*, International Institute For Applied Systems Analysis (Austria), August 29-31, 2005, pp. 81-86.

[13] S. Jiang, Y. Dang, and Z. Xuan, “Comparative Study on RMMFS and BMMFS of Chinese Word Extraction”, *Journal of the China society for Scientific and Technical Information*, Vol. 25, No. 4, 2006, pp. 499-503.

[14] S. Jiang, and Y. Dang, “Automatic Segmentation of Hierarchy Feature without Lexicon for Chinese Text Based on Iterative Learning”, *International Conference on*

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Computer Science and Software Engineering, IEEE Computer Society (China), December 12-14, 2008, pp. 657–661.

[15] M. Sun, Z. Zuo, and C. Huang, "An Experimental Study on Thesaurus Mechanism for Chinese Word Segmentation", *Journal of Chinese Information Processing*, Vol. 14, 1999, pp. 1-6.

[16] X. Zhang, and L. Wang, "Identification and Analysis of Chinese Organization and Institution Names", *Journal of Chinese Information Processing*, Vol. 11, 1997, pp. 21-31.

[17] W. Zhang, L. Yang, X. Sun, H. Yang, and Y. Liu, "An effective method of arbitrary length N-gram statistics for Chinese text", *International Journal of Digital Content Technology and its Applications*, Vol. 5, No. 3, 2011, pp. 143-155.

[18] H. Zhu, T. Ruan, and Q. Yu, "Studies on Text Segment Algorithms' Influence on Chinese-based Information Filtering", *Computer Engineering and Application*, Vo. 13, 2002, pp. 62-65.

[19] W. Naptali, M. Tsuchiya, and S. Nakagawa, "Studies on Text Segment Algorithms' Influence on Chinese-based Information Filtering", *IEICE Transactions on Information and Systems*, Vol. E95-D, No. 9, 2012, pp. 62-65.

[20] W. Zhang, H. Xu, and W. Wan, "Weakness Finder: Find Product Weakness from Chinese Reviews by using Aspects based Sentiment Analysis", *Expert Systems with Applications*, Vol. 39, No. 1, 2012, pp. 10283 – 10291.