S15-5

RECENT RESEARCH AND DEVELOPING TREND OF ENGINEERING MANAGEMENT IN CHINA BASED ON TEXT MINING

Shaohua Jiang[1], Wenling Zhang, Zhaohong Qiu, Shaojun Wang

[1] Lecturer, School of Civil & Hydraulic Engineering, Dalian University of Technology, China
Correspond to shjiang@dlut.edu.cn

ABSTRACT: With the rapid development of China economy, many engineering projects with large scale and investment were constructed in China and some were the biggest ones in the world. With the development of engineering practice, great progress in the research of engineering management of China was made and a large number of research findings were embodied in content of research papers and were represented by technical words. To know the state of arts in the research field of engineering management in China, three major parts, namely title, abstract and keywords of research papers in last five years from three representative Chinese journals about engineering management were chose as research materials. Unlike western languages, there are no delimiters between the words of Chinese, so the maximum matching and frequency statistics (MMFS) method, a text segmentation technique of text mining Chinese, was presented to extract the features consisting of technical words, phrases and words from the research materials. Recent research and developing trend of engineering management in China were found by comparing and analyzing the difference of technical words in the research materials of last five years.

Keywords: Text mining; Engineering management; Text segmentation; Recent research and development trend

1. INTRODUCTION

In the last decades China economy achievement has attracted worldwide attention and has become the third economic entity in the world. With the rapid economic development of China, plenty of infrastructure projects with large scale and investment were constructed, such as the bird's nest and water cube in Beijing Olympic Games, even quite a few projects with particular design are unique in the world and need complicated and innovative construction and management technique. With the development of infrastructure construction industry, great progress has achieved in the research field of engineering management and a great deal of research findings were published in papers, especially in Chinese.

To find the state of arts in the field of engineering management in China, text mining was adopted to discover knowledge implied in the published electronic papers. Three most important parts, i.e. title, abstract and keywords of research papers in last five years from three representative Chinese journals, namely Construction Management Modernization, Construction Technology, and Construction Economy, in the field of engineering management were chose as material of research.

Different from Indo-European languages, such as English, French, German, Chinese texts have no blank between words, so segmentation is a necessary step in processing Chinese texts, such as machine translation and information retrieval.

Words and phrases are basic elements of texts and form the features of text presentation, so feature extraction is a basic precondition of 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 texts, since 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 [1, 2].

The goal of Chinese text feature extraction is thus to transform a plain Chinese text to meaningful words and phrases.

The maximum matching and frequency statistics (MMFS) method [3, 4], which can extract special terms and proper nouns effectively, was introduced to segment Chinese text with no lexicon.

Based on the analysis and contrast about the features extracted by MMFS, recent research and developing trend of engineering management in China were summarized.

2. THE STATE OF THE ART OF CHINESE TEXT SEGMENTATION

Chinese text segmentation can be categorized as four approaches including segmentation based on lexicon [5], the method based on syntax and rules [6], the method based on statistics, for example the N-gram method [7], and the combination method of the above methods [8].

Segmentation based on lexicon is the most basic segmentation method for Chinese text. It performs segmentation process using string matching algorithms supported by a well pre-pared lexicon with sufficient amount of lexical entries which covers all of the Chinese words as possible. However 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 degraded.

Segmentation based on syntax and rules processes the segmentation and syntactic and semantic analysis synchronously. It utilizes syntactic and semantic information to carry out part of speech tagging and solves the segmentation ambiguity problem. The existing syntax knowledge and rules are too general and complex to avoid conflict between them with the increase of their quantity, so the precision of this method isn't satisfying and still in the phrase of test.

To conquer the disadvantage of the methods based on lexicon, syntax and rules, N-gram model was adopted. It's a statistical language model which was used very often. 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 [9]. Because the limit of computation cost in actual application, the N-gram model often takes into account only several historical information and forms models like bigram, trigram and so on.

For Chinese characters 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. There are primary three drawbacks in N-gram model:

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 how to estimate the probability of these language phenomena.

The existing hardware is difficult to meet the requirement of the computation cost of N-gram model with the increase of corpus and N.

Chinese character strings obtained by N-gram model lack semantic meaning.

There are several typical statistical segmentation methods for Chinese text without lexicon. Fu and Yuan, etc [10] requires computing the frequency of a shorter CCS which is contained by a longer one repeatedly and processes segmentation by some word filtering criterions. Liu and Wu, etc [11] calculates word frequency in the local context and processes segmentation by constructing hashing index of Chinese characters. Xu and Su, etc [11] need word index. Jin and Sun, etc [13] adopts the method of length increasing and it is appropriate to distill Chinese words whose frequency is medium and high.

The method which integrates part of the above methods still can not avoid the shortcomings of its each part radically.

To conquer drawbacks of the above methods, a method named MMFS was proposed. This method can extract CCS whose support degree is bigger than a predefined value.

3. THE MAXIMUM MATCHING AND FREQUENCY STATISTICS SEGMENTATION METHOD STYLING

3.1 Basic Idea of MMFS

Whether the basic processing unit in Chinese is word or phrase, is still a controversial problem owing to the properties of Chinese. The definition "word is the smallest language element, it can be used independently and has semantic meaning" lacks maneuverability [16]. Phase has steady structure, so phase should be regarded as the basic processing unit.

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

- If a CCS in text has a higher frequency, the possibility of it being a word is higher.
- Only the unambiguous semantic CCS can be a word.
- The Chinese characters' combination mode can be observed in statistical sense.

In addition, 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 these properties of Chinese text, the processing technology is different from west languages. Text processing is on the basis of content, so a new Chinese text segmentation method is put forward in this paper. On the basis of segmentation of CCS by segmentation tag in the pretreatment phase, the CCS's frequency is analyzed according to the idea of matching the longer string first. The segmentation of content-oriented CCS having more Chinese characters is processed first, then the CCS's length is decreased and the CCS's frequency is analyzed. So the segmentation of Chinese text can be finished without thesaurus and probability estimation in advance.

3.2 Design of the MMFS Algorithm

Some definitions are given before discussion.

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

Definition 2: CCS is all the Chinese strings in text.

The basic idea is enriching the set of segmentation tag and turning longer text string into shorter CCS, so it is convenient for later

processing.

To explain the segmentation of text more explicitly, the inputted Chinese text is denoted by T and some definitions are given as follows.

Definition 3: Segmentation Denotation (*SD*) is th e set of denotations which can't appear in phras e and word of Chinese text. It comprises natural segmentation denotation and non-natural segmentation denotation. Natural segme ntation denotation includes punctuations and no n-
natural segmentation denotation includes numbe r and non-Chinese characters.

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

Definition 5: Pretreated String (*PS*) is the set of the CCS which is formed after the pretreatment through *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 statistic s on the basis of *PS*.

Definition 7: Support degree is the frequency of the CCS in the text. The predefined support de gree 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, *C*, the set of all Chinese characters, is defined. *NC*, the set of all non-Chinese characters, is defined. Λ deno t e s blank, $\Lambda \in NC$. So $T = C \cup NC$. The proposed algorithm i ncludes five main steps.

- Preliminary segmentation. *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, $T_1 = C \cup \{\Lambda\}$.
- Further segmentation. T_2 is the set of CCS which is formed by further preliminary segmentation by *SS* and comprises more blanks. The continuous Λ in T_2 is incorporated to one, so *PS*, the result of incorporation, is formed. $PS = c_1 c_2 \Lambda \Lambda c_n$, $c_i \in C \cup \{\Lambda\}, 1 \leq i \leq n$, and c_i, c_{i+1} can't be Λ at the same time.
- 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.
- Filtered by predefined support degree. Take the CCS whose support degree is more than or the same as Φ as the final segmentation result and produce *SR*. Φ should change with the different length of text. Φ is set to 2 because of the limited text length.
- Feedback. The new discovered segmentation denotation and Segmentation String based on *CS* is added to *SD* and *SS*, thereby the system's capability is more perfect. Feedback is an optional function.

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

With regard to the time spending of this algorith m, supposing the maximal CCS to be extracted whose length is *L*, the number of total CCS of te xt is *N* after pretreatment. The CCS having *L* to 2 Chinese characters is extracted. The time co mplexity 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 valu e.

This method doesn't segment single Chinese ch aracters without reference to its frequency, bec ause it has no information itself and is useless f or classification and retrieval of text in practice. Since phrase is the basic processing unit and s pecial indicatory semantic CCS has more special semantic meaning than phrase, so that the basi c processing unit of Chinese text should be the combination of special indicatory semantic CCS, phrase and word.

4. RESULTS OF EXPERIMENT AND ANALYSIS

4.1 Design of Experiment

Three important Chinese journals in the field of engineering management were chose as representatives,

namely Construction Management Modernization, Construction Technology, and Construction Economy. Three most important parts, i.e. title, abstract and keywords of research papers in last five years from above three journals were chose as material of research. The experimental corpus of the three journals is summarized in table 1-3 respectively.

Table 1. Corpus of Construction Management Modernization

Year	2004	2005	2006	2007	2008
Pieces of paper	124	122	122	117	96

Table 2. Corpus of Construction Technology

Year	2004	2005	2006	2007	2008
Pieces of paper	343	527	650	606	804

Table 3. Corpus of Construction Economy

Year	2004	2005	2006	2007	2008
Pieces of paper	309	295	465	507	503

4.2 Result of Segmentation

Text segmentation result of the three journals is shown in table 4-6 respectively.

Table 4. Segmentation result of Construction Management Modernization

Year	2004	2005	2006	2007	2008
Amount of segmentation result	1124	748	964	892	754

Table 5. Segmentation result of Construction Technology

Year	2004	2005	2006	2007	2008
Amount of segmentation result	1702	3340	3315	3386	5918

Table 6. Segmentation result of Construction Economy

Year	2004	2005	2006	2007	2008
Amount of segmentation result	1999	1810	2766	3494	3635

4.3 Analysis of Research Highlight

Many universal words in the segmentation result, most of which are words and expressions composing two Chinese characters, can't act as representative of research status and trend, so a procedure of eliminating universal terms was implemented to form technical words set which include technical words merely.

In addition, there are still too many technical words to express recent research highlight clearly, so only the top 5 technical words in descending order of frequency are considered, which are summarized in Table 7-9 in Chinese and English respectively.

Table 7. The top 5 technical words in the journal of Construction Management Modernization

Year	2004		2005		2006		2007		2008	
No.	Chinese	English	Chinese	English	Chinese	English	Chinese	English	Chinese	English
1	项目管理	project management	项目管理	project management	招投标	bidding and bid	项目管理	project management	项目管理	project management
2	招投标	bidding and bid	房地产	real estate	房地产	real estate	代建制	system of building as an agent	合同管理	contract management
3	风险管理	risk management	风险管理	risk management	管理模式	management model	风险管理	risk management	风险管理	risk management
4	合同管理	contract management	质量控制	quality control	新农村	new country	房地产市场	real estate market	代建制	system of building as an agent
5	房地产市场	real estate market	后浇带	casting strip	房地产评价	real estate evaluation	承包商	contractor	承包商	contractor

From Table 7 the conclusions can be drawn as follows: research related to real estate is one highlight in the journal of Construction Management Modernization because the appearance number of technical words including real estate is 5 in the top 5 technical words from 2004 to 2008, project management and risk management are also hot research fields because their appearance numbers are all 4.

Table 8. The top 5 technical words in the journal of Construction Technology

Year	2004		2005		2006		2007		2008	
No.	Chinese	English	Chinese	English	Chinese	English	Chinese	English	Chinese	English
1	混凝土	concrete	混凝土	concrete	混凝土	concrete	混凝土	concrete	钢结构	steel structure
2	预应力	prestressing force	高性能混凝土	high performance concrete	钢结构	steel structure	大体积混凝土	massive concrete	预应力	prestressing force
3	钢结构	steel structure	预应力	prestressing force	预应力	prestressing force	预应力	prestressing force	混凝土	concrete
4	钢筋焊接网	welded steel fabric	粉煤灰	coal fly ash	基坑工程	footing groove engineering	质量控制	quality control	施工方案	construction scheme
5	脚手架	falsework	地下连续墙	underground diaphragm wall	混凝土浇筑	concrete construction	钢结构	steel structure	仿真计算	simulation calculation

Table 9. The top 5 technical words in the journal of Construction Economy

Year	2004		2005		2006		2007		2008	
No.	Chinese	English	Chinese	English	Chinese	English	Chinese	English	Chinese	English
1	项目管理	project management	项目管理	project management	项目管理	project management	代建制	system of building as an agent	项目管理	project management
2	工程总承包	general contract	代建制	system of building as an agent	代建制	system of building as an agent	项目管理	project management	房地产	real estate
3	拖欠工程款	default payment problem	建筑市场	construction market	房地产	real estate	房地产	real estate	基础设施	infrastructure
4	建筑市场	construction market	风险管理	risk management	风险管理	risk management	风险管理	risk management	房地产市场	real estate market
5	核心竞争力	core competitive power	招投标	bidding and bid	承包商	contractor	建筑节能	energy conservation in building	成本控制	cost control

From Table 8 the conclusions can be drawn as follows: researches of project management and real estate-related are highlight in the journal of Construction Economy because the appearance numbers of technical words of project management and including real estate are 5 respectively in the top 5 technical words from 2004 to 2008, system of building as an agent and risk management are also hot research fields because their appearance numbers are all 3.

From Table 9 the conclusions can be drawn as follows: researches related to concrete are highlight in the journal of Construction Technology because the appearance numbers of technical words including concrete are 8 in the top 5 technical words from 2004 to 2008, prestressing force and steel structure are also hot research fields because their appearance numbers are 5 and 4 respectively.

4.4 Analysis of Research Trend

To obtain the trend of engineering management research in China, each technical words set of the above three journals from 2004 to 2007 was gathered to form a history glossary and technical words of 2008 were compared to the history glossary one by one, then the new technical words representing research trend of the three journals were gained. Owing to the convenience for expression and the limit of the space of the whole page, only some new technical words of the three journals in 2008 with higher frequency are listed in Table 10-12 respectively according the rank of frequency in descending order.

Table 10. The new technical words with higher frequency in the journal of Construction Management Modernization in 2008

No.	Frequency	Chinese	English
1	7	安全监理	safety supervision
2	7	组织模式	organization pattern
3	6	动态联盟	dynamic alliance
4	6	项目清单	project bill
5	6	同业担保	craft's guarantee
6	6	工程造价咨询企业	engineering cost consulting enterprises
7	5	司法鉴定	judicial appraisal
8	5	串通投标	collusion in bidding

From Table 10 the conclusions can be drawn as follows: researches of safety supervision and organization pattern are the most important emerging hotspots in the journal of Construction Management Modernization due to their highest frequency; beside this, research topics about dynamic alliance, project bill, craft's guarantee, engineering cost consulting enterprises, judicial appraisal and collusion in bidding are also new arisen highlights.

Table 11. The new technical words with higher frequency in the journal of Construction Technology in 2008

No.	Frequency	Chinese	English
1	31	仿真计算	simulation calculation
2	12	曲率模态	curvature mode
3	10	沥青胶浆	asphalt mortar
4	10	张拉方案	tension scheme
5	9	施工技术措施	technical measures of construction
6	9	质量控制措施	control measures of quality
7	7	高性能水泥复合砂浆钢筋网薄层	high performance ferrocement laminate

From Table 11 the conclusions can be drawn as follows: a new research topic, namely simulation calculation, is drawing most attention in the journal of Construction Technology; researches related to curvature mode, asphalt mortar, tension scheme, technical measures of construction, control measures of quality and high performance ferrocement laminate are emerging topics too.

Table 12. The new technical words with higher frequency in the journal of Construction Economy in 2008

No.	Frequency	Chinese	English
1	14	限价房	priced housing
2	11	公共项目	public project
3	11	包工头	labor contractor
4	10	银行信贷	bank credit
5	8	钓鱼工程	angling engineering projects
6	8	契约理论	contract theory
7	8	健康诊断	health diagnosis
8	8	房地产产品软创新	soft innovation of real estate product
9	8	建筑市场交易	construction market trade

From Table 12 the conclusions can be drawn as follows: researches about priced housing, a new research topic, are drawing most attention in the journal of Construction Economy because priced housing, as new phenomena of China housing market, has caught researchers' eye; researches related to public project and labor contractor are also new research highlights; researches related to bank credit, angling engineering projects, contract theory, health diagnosis, soft innovation of real estate product, construction market trade are emerging topics too.

5. CONCLUSIONS

The maximum matching and frequency statistics (MMFS) method, a Chinese text segmentation technique which can distinguish technical words effectively, was introduced to distill technical words of three representative Chinese journals in the field of engineering management. Three major parts, namely title, abstract and keywords of research papers in last five years from the three journals were chose as research materials. By comparing and analyzing the technical words distilled from the research materials, recent research and developing trend of engineering management in China were summarized which can throw light on the research of engineering management in China for researchers and practitioners all over the world so as to promote their future work.

REFERENCES

[1] Schubert Foo, Hui Li, "Chinese word segmentation and its effect on information retrieval", *Information Processing and Management*, Vol. 40(1), pp. 161-190, 2004.

[2] Chin-Ming Hong, Chih-Ming Chen, Chao-Yang 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(2), pp. 3641-3651, 2009.

[3] Jiang Shaohua, Dang Yanzhong, "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)*, IIASA Laxenburg, Austria. pp. 81-86, 29-31 August, 2005,

[4] Shaohua Jiang, Yanzhong Dang. "Automatic Segmentation of Hierarchy Feature without Lexicon for Chinese Text Based on Iterative Learning", *International Conference on Computer Science and Software Engineering*, Wuhan, China, pp. 657-661, 12-14 December, 2008.

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

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

[7] Joon H. L., Hyun Y. C., Hyouk R. P., "n-Gram-based indexing for Korean text retrieval", *Information Processing and Management*, 35, pp. 427-441, 1999.

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

[9] Hirofumi Y., Yoshinori S., "Multiclass composite N-gram language model based on connection direction", *Systems and Computers in Japan*, 34, pp. 108-114, 2003.

[10] Fu S. X., Yuan D. R., Huang B. X., Zhong Z., "Word Extraction without Thesaurus Based on Statistics", *Journal of Guangxi Academy of Sciences*, 18, pp. 252-264, 2002.

[11] Liu T., Wu Y., Wang K. Z., "A Chinese Word Automatic Segmentation System Based on String Frequency Statistics Combined with Word Matching", *Journal of Chinese Information Processing*, 12, pp. 17-25, 1998.

[12] Xu G. X., Su X. W., Chen S. Y., "Arithmetic and Application of No Thesaurus Cutting Word in Chinese Text Mining", *Journal of JILIN Institute of Technology*, 23, pp. 16-18, 2002.

[13] Jin X. Y., Sun Z. X., Zhang F. Y., "A Domain-independent Lexical-acquisition Model to Chinese Document", *Journal of Chinese Information Processing*, 15, pp. 33-39, 2001.