TELKOMNIKA, Vol. 13, No. 4, December 2015, pp. 1414~1421
ISSN: 1693-6930, accredited A by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v13i4.2389

Automatically Generation and Evaluation of Stop Words List for Chinese Patents

Deng Na[1], Chen Xu*[2]
[1]School of Computer, Hubei University of Technology, Wuhan 430068, P.R. China
[2]School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan 430073, P.R. China
*Corresponding author, email: iamdengna@163.com, xuchen2014@yeah.net

Abstract

As an important preprocessing step of information retrieval and information processing, the accuracy of stop words' elimination directly influences the ultimate result of retrieval and mining. In information retrieval, stop words' elimination can compress the storage space of index, and in text mining, it can reduce the dimension of vector space enormously, save the storage space of vector space and speed up the calculation. However, Chinese patents are a kind of legal documents containing technical information and the general Chinese stop words list is not applicable for them. This paper advances two methodologies for Chinese patents. One is based on word frequency and the other on statistics. Through experiments on real patents data, these two methodologies' accuracy are compared under several corpuses with different scale, and also compared with general stop list. The experiment result indicates that both of these two methodologies can extract the stop words suitable for Chinese patents and the accuracy of Methodology based on statistics is a little higher than the one based on word frequency.

Keywords*: stop word; patent; statistics; information retrieval; word frequency*

1. Introduction

Along with the development of Internet and Information Technology, massive amounts of data are accumulated in every domain. Reported by Survey report on the quantity of Chinese Internet Information Resources, up to the end of 2005, the number of Chinese web pages has reached 2.4 billion, and this number is increasing continually and explosively. Not only the web pages, some other character carriers, such as academic papers, patents and so on, their amounts are going up at an alarming rate too. How to find out the needed data from these big data quickly and precisely and how to mine out useful information from these data are the problems demanding prompt solution. In information retrieval, there are three basic steps, that is, word segmentation, stop words' elimination and indexing. In text mining, before several kinds of mining, word segmentation, stop words' elimination and key words' extraction are still the essential work [1]. Therefore, as an important preprocessing step of information retrieval and information processing, the accuracy of stop words' elimination directly influences the ultimate result of retrieval and mining.

Stop words are those words emerging frequently in corpus but with no important information [2-3]. Zipf's Law [4] shows that, in English language, only a few words are used regularly, most words are rarely used. The languages of the other countries, including Chinese, have the same character. Stop words are such words used frequently but can not differentiate documents. For example, in English, the most frequent words are the, of, and, to, a, in, that and is [5], all of them are included in English stop words list.

The elimination of stop words can bring in advantages from two aspects. In information retrieval, since stop words have no actual meanings, there is no need to index them. Stop words' elimination can compress the storage space of index. In text mining, it can reduce the dimension of vector space enormously, save the storage space of vector space and speed up the calculation.

At present, there are many stop words lists for English language, such as [6] and [7], and there have been some researches about English stop words. [8] [9] evaluate the influence on retrieval performance brought by stop words lists with different length and content. In [9], its

Received August 25, 2015; Revised October 9, 2015; Accepted October 24, 2015

own stop words list's generation method is given. This method is based on two existing lists and the most simple and general strategy, that is, computing the frequency of each word in the corpus, ranking them in descending order, choosing those words with high frequencies, and filtering them manually. If the words left are not in the two existing lists, add them into the union of the two lists. Finally, add some special words in, like times, file names, roman numerals, prefixes, adjectives, adverbs, dates, foreign words, scale unites and so on. [10] verifies that after using stop words list, meta-search engine can obtain better result. [11] adopts stop words list to detect paper plagiarism.

Except for English, about other non-Chinese language, [12] [13] studies the construction of stop words list for Arabic. [14] makes use of union entropy to study the stop words for Mongolian. [15] uses the methodology of [2] to build stop words list for Thai.

Currently, there are few scholars researching on stop words for Chinese. [2] [16] calculate the rank of words from statistics and informatics these two viewpoints, and consider both of them to get the final list. In statistics model, the average probability and variance of each word is computed. Those words with high average probability and low variance are deemed as candidate stop words. In information model, the entropy of each word is calculated and they extract the words with low entropy out as candidates. Finally, Borda ranking method is adopted to determine the final list. [17] gives the definition of stop words from the angle of statistics. It thinks that a stop word should satisfy two conditions. One is high document frequency and the other is that it has little relationship with classification categories. On the basis of contingency table, calculate words' weighted Chi, and those with the lowest values are ranked on the top of list. [18] calculate the probabilities of word in sentence and sentence in corpus respectively, and extract stop words list according to their union entropy. [19] divides stop words into two groups, that are, absolute words and relative words. It makes use of left/right entropy and Ngram to filter stop words for users' request in Information Retrieval.

Our work in this paper contains: 1) give two methodologies to generate stop words lists for Chinese patents; 2) through experiments on the real patent data, compare the accuracies of these two methodologies under corpuses with different scales, and compare them with the list for general Chinese texts.

The main innovation of our paper is that:

1) Currently, there have no relative research on stop words list for Chinese patents. We fill in the blanks. We analyze the contents in the stop words list for general Chinese texts, classify them into some categories, and clarify why this list is not suitable for Chinese patents.

2) Through the experiments on real patent data, we find that the algorithm in [2] [15] is not applicable for Chinese patents too, and on the basis on which, we give some modification and adjustment.

3) Compare and evaluate the accuracies of these two methodologies under corpuses with different scales, and compare them with the list for general Chinese texts.

2. The Stop Word List for General Chinese Texts

In Internet, there are some popular stop words lists, such as Harbin Institute of Technology's and Baidu Corporation's.

These lists contain some words frequently occurring in general texts. We classify them into eight categories.

1) Modal particle. Eg: "啊(ah), 阿(ah), 哎(hey), 哎呀(oh,my!), 哎哟(ouch), 唉(gosh)" and so on.
2) Onomatopoeia. Eg: "嗡嗡(buzz), 吧哒(ba-da), 叮咚(ding-dong), 沙沙(rustle), 瑟瑟(se-se)" and so on.
3) Local Dialect And Folk Adage. Eg: "敞开儿(unrestrictedly), 打开天窗说亮话(Speak frankly and sincerely), 赶早不赶晚(The earlier the better), 那末(then)" and so on.
4) Conjunction. Eg: "不管(regardless of), 并非(really not), 换句话说(in other words)" and so on.
5) Adverb. Eg: "马上 (immediately), 略微 (a little), 默默地 (silently), 必定 (certainly), 果然 (as expected)" and so on.
6) Pronoun. Eg: "你(you), 他们(they)" and so on.
7) Preposition. Eg: "在(at), 从(from), 当(when)" and so on.
8) Emotional words. These words usually contain commendatory or derogatory sence in them. Eg: "不择手段 (play hard), 不亦乐乎 (pleasureably), 老老实实 (conscientiously), 故意 (intentionally), 成心(on purpose)" and so on.

However, these lists are not suitable for Chinese Patents. It can be explained from two aspects.

1) Many words in the list for general Chinese texts will never appear in Chinese Patents. Patens are a kind of legal documents containing technology information, and their wording is usually rigorous and serious. Thus, patents will not contain those words refereed above because they are not serious enough. Therefore, since these words will never appear in patents, the list for general texts is not fit for patents.

2) Chinese patents are usually written according to specific mode and sentence pattern, and include many conventional words. For example, "发明(invention), 实用新型(utility model), 技术(technology), 系统(system), 装置(equipment), 采用(adopt)" and so on. These words have actual meanings in general texts, but in patents, they are template words would be used by most patents, and could not differentiate patents.

3. Word Segmentation of Chinese

Before eliminating stop words, word segmentation is required. At present, there have been many popular tools, such as IKAnalyzer [20] and JE-analysis [21]. JE-analysis is not open source software with its own stop words list. Before its segmentation result comes out, stop words have been deleted. By comparison, IKAnalyzer is open source, and users can customize their own stop words list. To guarantee keep all the words, we choose IK Analyzer as the word segmentation tool in this paper.

4. Two Methodologies of Generating Stop Words Lists for Chinese Patents

[22] mentions that, the most simple and general strategy is computing the frequency of each word in the corpus, ranking them in descending order, choosing those words with high frequencies, and filtering them manually. [2] [15] use a statistics method to calculate average probability, variance and SAT of word and rank them. However, our experiment shows that this method is not suitable for Chinese patents. We give it some modification and adjustment. Inspired by the research above, in this paper, we give two methodologies to generate stop words list. One is based on the most simple and general strategy and the other is based on modified SAT. The details are as follows.

4.1. Methodology One: Based on the Most Simple and General Strategy

The procedure of generating stop words list using Methodology One is as follows:

1. segment word for patents
2. get the frequency of each word in the corpus
3. rank the words according to their frequencies in descending order
4. extract the words with high frequency as candidates, and filter them manually

Here, the frequency in step 2 refers to the occurring count of word in the whole corpus, not document frequency.

4.2. Methodology Two: Based on Statistics

Suppose the corpus D={d_i}, 1=<i<=N. N refers to the count of patents. The set of words in corpus is denoted as W={w_j}.

Definition 1: average probability MP

The average probability of word w_j in D is:

$$MP\left(w_j\right)=\frac{\sum_{1\le i\le N} p_{ij}}{N}$$

p_{ij} is the frequency probability of w_j in d_i. In other words, p_{ij} equals to w_j 's frequency in d_i divided by the number of words in d_i. If a word has a high MP value, it implies that this word occurs frequently in the whole corpus.

Definition 2: variance VP

The variance of w_j in D is:

$$VP(\mathrm{w}_j)=\frac{\sum_{1\le i\le N}(\mathrm{p}_{ij}-\mathrm{MP}(\mathrm{w}_j))^2}{N}$$

If a word has a low VP value, it implies that this word occurs uniformly in the whole corpus.

Definition 3: SAT

The SAT of w_j in D is:

$$SAT(\mathrm{w}_j)=\frac{MP(\mathrm{w}_j)}{\sqrt{VP(\mathrm{w}_j)}}$$

If a word has a high SAT value, it implies that this word occurs frequently and uniformly in the whole corpus. The word like this is very likely to be a stop word.

Modification and adjustment of SAT

On the basis of [2] [15], we give some modification and adjustment on SAT. Our experiment shows that the old definition of SAT is not suitable for Chinese patents. In the experiment, if we use the old definition, after ranking the words according to SAT in descending order, those words on the top are not correct stop words. Many words with low MP and lower VP are ranked on the top improperly. The reason is that MP is in magnitude of frequency, but VP is in magnitude of the square of frequency. Therefore, we modify the definition of SAT, and adjust the square root of VP as SAT's denominator. In this way, VP and MP are in the same magnitude.

The procedure of generating stop words list using Methodology Two is as follows:

1. Segment word for patents
2. Calculate each word's SAT value in the corpus
3. Rank the words according to SAT in descending order
4. Extract the words with high SAT as candidates, and filter them manually

5. Analysis and Evaluation of Experiment

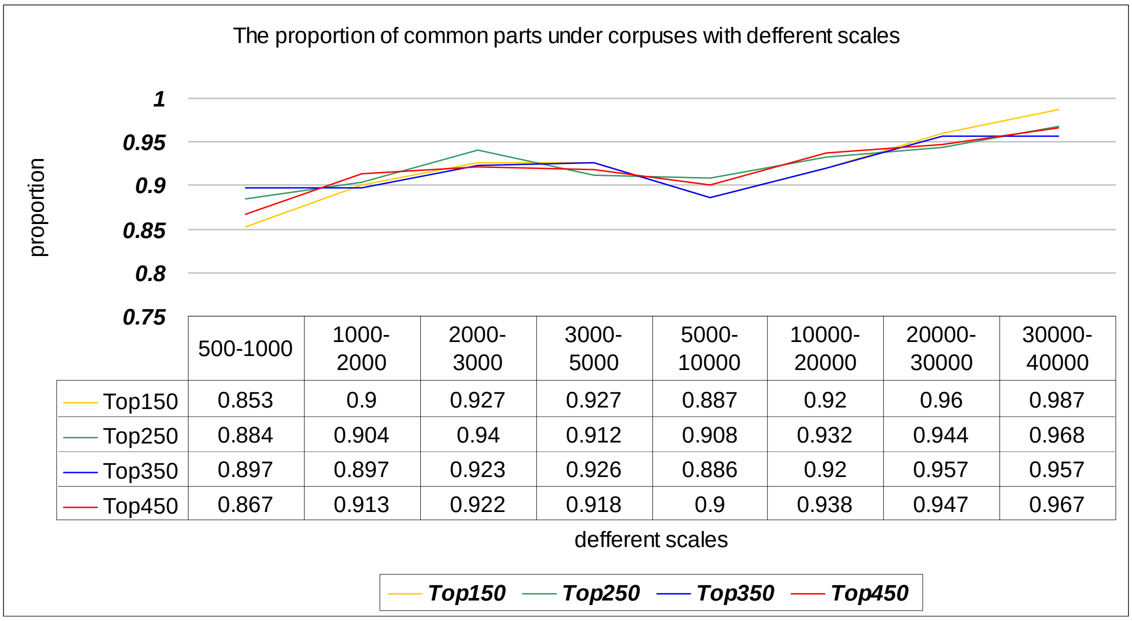
5.1. Dataset

The data source in our experiment is more than 40000 Chinese patents applied by Chinese universities and scientific research institution from 1985-9-10 to 2010-10-6. These data include application number, application date, IPC (International Patent Classification), applicant, patentee, title, abstract, deputy and so on. In this paper, we are concerned only with application number and abstract. The application numbers are used as keys, and the abstracts compose the corpus.

The experiment is conducted under 9 corpuses with different scales. They represent the number of patents are 500, 1000, 2000, 3000, 5000, 10000, 20000, 30000 and 40000 respectively.

5.2. Using Methodology One to Generate Stop Words List under Corpuses with Different Scales

Different scale of corpuses generates 9 lists in all. Figure 1 shows the proportion of common parts of these 9 lists. We can see that along with the ever-increasing of the corpus's scale, though there are a few wave hollows, the proportion of common parts is on the rise totally. In other words, along with the ever-increasing of the corpus's scale, the words on the top of the lists become stable gradually. In addition, Figure 1 indicates that when the compared scales are 30000 and 40000 patents, the proportion of common parts of top 150 words reaches 0.987. Those wave hollows may be caused by the data's unbalanced distribution in the corpus. The phenomenon of broken lines' gliding disappears going with the scale's increase.



	500-1000	1000-2000	2000-3000	3000-5000	5000-10000	10000-20000	20000-30000	30000-40000
Top150	0.853	0.9	0.927	0.927	0.887	0.92	0.96	0.987
Top250	0.884	0.904	0.94	0.912	0.908	0.932	0.944	0.968
Top350	0.897	0.897	0.923	0.926	0.886	0.92	0.957	0.957
Top450	0.867	0.913	0.922	0.918	0.9	0.938	0.947	0.967

Figure 1. The proportion of common parts of these 9 lists under corpuses with different scales

5.3. The Accuracies of Stop Words List under Corpuses with Different Scales Using Methodology One

Figure 2 shows the accuracies of Top 150, 250, 350, 450 words in stop words lists under corpuses with different scales. It is clear that all of the nine broken lines in the figure reveal a downtrend. This indicates that the more the Top words' number, the lower is the accuracy of the stop words list. In other words, in each list, the accuracy of Top 150 words is highest. In addition, Figure 2 also shows that it is not to say the bigger the corpus, the more accurate is the list.

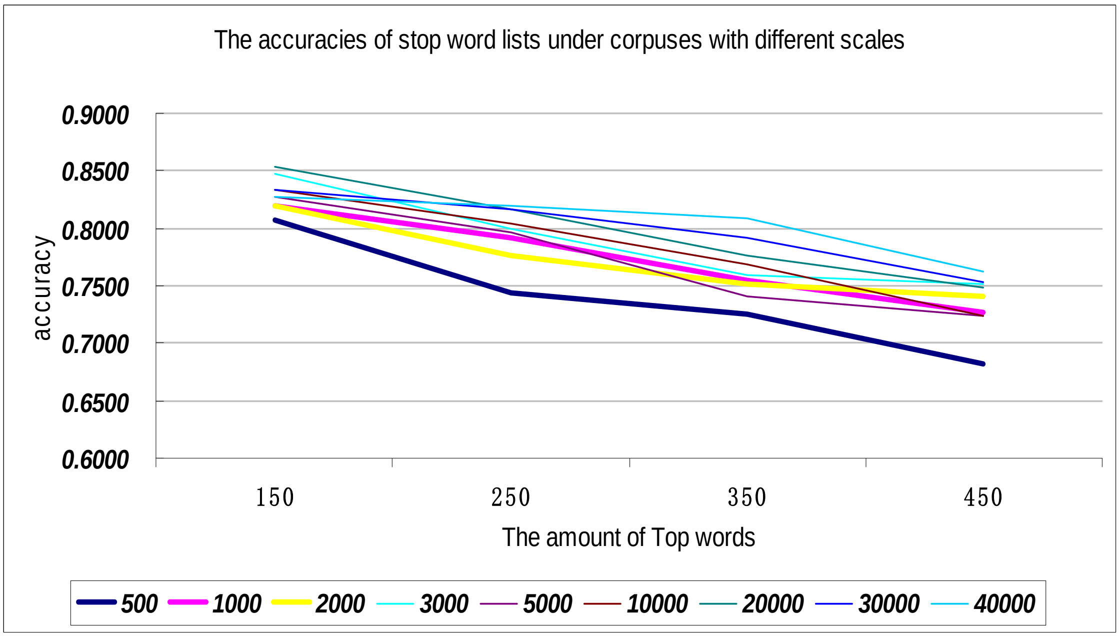


Figure 2. the accuracies of Top 150, 250, 350, 450 words in stop words lists under corpuses with different scales

The concrete accuracies data in Figure 2 is shown as follows.

Table 1. The accuracies data of Top 150, 250, 350, 450 words in stop words lists under corpuses with different scales

	150	250	350	450
500	0.8067	0.7440	0.7260	0.6820
1000	0.8200	0.7920	0.7540	0.7270
2000	0.8200	0.7760	0.7510	0.7400
3000	0.8467	0.8000	0.7600	0.7510
5000	0.8267	0.7960	0.7400	0.7240
10000	0.8330	0.8040	0.7690	0.7240
20000	0.8530	0.8160	0.7770	0.7490
30000	0.8330	0.8160	0.7910	0.7530
40000	0.8267	0.8200	0.8090	0.7620

5.4. Methodology One’s List Compared with the List for General Texts

It has been referred above that, some words have actual meanings in general texts, but are common templates and sentence pattern in patents with no actual meanings. We compare a sublist from Methodology One (i.e. Top 150 words when the scale is 20000) with the list of Harbin Institute of Technology (767 words). In our sublist, there are 110 new words. 30 words are chosen, shown in Table.2.

Table 2. Some new words not in the stop words list of Harbin Institute of Technology

本发明 (the invention)	一种(a kind of)	方法(method)
中(in)	装置(device)	制备(equipment)
具有(have)	进行(conducted)	上(on)
包括(containing)	涉及(involved)	系统(system)
在于(lie in)	采用(adopt)	属于(belong to)
材料(material)	后(after)	实用新型(utility model)
实现(achieve)	用于(used for)	所述(according to)
组成(constitute)	处理(handle)	提供(provide)
技术领域(technology domain)	使(make)	特征(characteristic)
形成(form)	利用(utilize)	得到(get)

5.5. The Accuracies of Stop Words List under Corpuses with Different Scales using Methodology Two

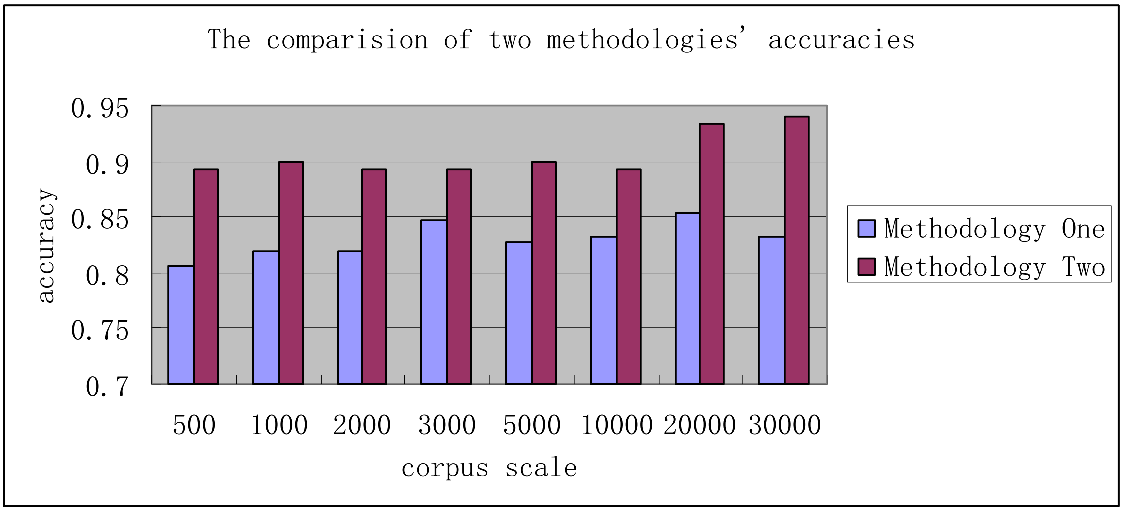


Figure 3. Methodology One and Two’s accuracies of Top 150 words under corpuses with different scales

From Figure 3, we can see that, under different scales of corpuses, Methodology Two's accuracies are higher than Methodology One's. Moreover, when the scale lays between 500 and 10000, the accuracies of Top 150 words of methodology have little differences, and when the scale promotes to 20000 and 30000, the accuracies amplifies obviously.

6. Conclusion

Aiming at the problem that the general stop words list is not suitable for Chinese patents, this paper proposes two methodologies to automatically generate stop words list. It classifies the words of general list into several categories and clarifies why the general list is not applicable for Chinese patents. Through the experiment on real patents data, we compare our two methodologies' accuracies under corpuses with different scales, and compare with the general list too. The experiment result indicates that both of our two methodologies are suitable for Chinese patents, and the accuracy of the methodology based on statistics is a little higher than the one base on word frequency.

Acknowledgment

This work was supported by Research Foundation for Advanced Talents of Hubei University of Technology (No. BSQD12131), the Fundamental Research Funds for the Young Teachers' Innovation project of Zhongnan University of Economics and Law (No. 2014147), the National Natural Science Foundation of China (No. 61201250), the Natural Science Foundation of Anhui Province (No. 1308085QF103), the Guangxi Natural Science Foundation (No. 2012GXNSFBA053174), and the Guangxi University Key Lab of Cloud Computing and Complex System Found. The authors sincerely thank the anonymous reviewers for their constructive comments and helpful suggestions.

References

[1] Erlin E, Rahmiati R, Rio U. Two Text Classifiers in Online Discussion: Support Vector Machine vs Back-Propagation Neural Network. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*. 2014; 12(1): 189-200.

[2] Zou F, Wang FL, Deng X, et.al. *Automatic construction of Chinese stop word list*. Proceedings of the 5th WSEAS international conference on Applied computer science. Hangzhou, China. 2006: 1010-1015.

[3] Yuang CT, Banchs RE, Siong CE. *An empirical evaluation of stop word removal in statistical machine translation*. Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra). Association for Computational Linguistics, 2012: 30-37.

[4] K Zipf. Selective Studies and the Principle of Relative Frequency in Language. MIT Press. 1932.

[5] Timothy C Bell, John G Cleary and Ian H Witten. Text Compression. Prentice Hall. 1990.

[6] DTIC-DROLS English Stop Word List , http://dvl.dtic.mil/stop_list.html

[7] English Stop Word List in Word Net, http://www.d.umn.edu/~tpederse/Group01/WordNet/words.txt

[8] Dolamic L, Savoy J. When stopword lists make the difference. *Journal of the American Society for Information Science and Technology*. 2010; 61(1): 200-203.

[9] Zaman ANK, Matsakis P, Brown C. *Evaluation of stop word lists in text retrieval using Latent Semantic Indexing*. Proceedings of 2011 the Sixth International Conference on Digital Information Management (ICDIM). Melborune, Australia. 2011: 133-136.

[10] Patel B, Shah D. *Significance of stop word elimination in meta search engine*. Proceedings of 2013 International Conference on Intelligent Systems and Signal Processing (ISSP). India. 2013: 52-55.

[11] Stamatatos E. Plagiarism detection using stopword n-grams. *Journal of the American Society for Information Science and Technology*. 2011; 62(12): 2512-2527.

[12] El-Khair IA. Effects of stop words elimination for Arabic information retrieval: a comparative study. *International Journal of Computing & Information Sciences*. 2006; 4(3): 119-133.

[13] Medhat W, Yousef AH, Korashy H. Corpora Preparation and Stopword List Generation for Arabic data in Social Network. *arXiv preprint arXiv*:1410.1135, 2014.

[14] Gong Zheng, Guan Gaowaa. Comparative Study on Between Mongolian Stop Words and English Stop Words. *Journal of chinese information processing*. 2011; 25(4): 35-38. (in Chinese)

[15] Daowadung P, Chen YH. *Stop Word in Readability Assessment of Thai Text*. Proceedings of 2012 12th International Conference on Advanced Learning Technologies (ICALT). Rome, Italy. 2012: 497-499.
[16] Zou F, Wang FL, Deng X, et.al. Stop word list construction and application in Chinese language processing. *WSEAS Transactions on Information Science and Applications*. 2006; 3(6): 1036-1044.
[17] Hao L, *Hao L. Automatic identification of stop words in Chinese text classification*. Proceedings of 2008 International Conference on Computer Science and Software Engineering. Wuhan, China. 2008; 1: 718-722.
[18] Gu Yijun, Fan Xiaozhong, Wang Jianhua, Wang Tao, Huang Weijin. Automatic Selection of Chinese Stoplist. *Transactions of Beijing Institute of Technology*. 2005; 25(4): 337-340. (in Chinese)
[19] Xiong Wenxin, Song rou. Removal of Stop Word in Users' Request for Information Retrieval. *Computer Engineering*. 2007; 33(6): 195-197.
[20] ik-analyzer. https://code.google.com/p/ik-analyzer/
[21] je-analysis.https://code.google.com/p/jinhe-tss/downloads/detail?name=je-analysis1.5.1.jar
[22] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze. An Introduction to Information Retrieval. Cambridge University Press, 2008.