A Comparative Study of Commonly Used Stop-word Lists for Chinese Text Clustering (Post-print)

Authors: Guan Qin, Deng Sanhong, Wang Hao

Date: 2017-11-08T00:00:00+00:00

Abstract

【Purpose】 To conduct experimental comparative analysis on the effects of different stop word lists on various types of text data, and to provide reference recommendations for the construction and application of stop word lists. 【Method】 Three stop word lists were selected: Baidu stop word list, Harbin Institute of Technology stop word list, and Sichuan University Machine Intelligence Laboratory stop word list. Text processing was performed on three different corpora using Chinese word segmentation technology, TF-IDF feature evaluation function, and VSM model. Clustering experiments were conducted using a K-means algorithm implemented in Java, and the results were evaluated using three metrics: precision P, recall R, and F1 score. 【Results】 The effects of different stop word lists on various types of text data differ significantly; the length and content structure of the word lists are direct factors influencing the effectiveness, with two-character stop words demonstrating the most pronounced effect. 【Limitations】 The types and quantity of experimental texts are limited. Additionally, only simple analysis and comparison of word quantity and content were performed for different stop word lists, without conducting experimental analysis on stop words categorized by class. 【Conclusion】 Stop word lists significantly impact the accuracy of text clustering, making the construction or selection of appropriate Chinese stop word lists crucial. Moreover, excessively increasing the number of stop words does not consistently improve clustering results.

Full Text

A Comparative Study of Commonly Used Chinese Stopword Lists for Text Clustering

Guan Qin, Deng Sanhong, Wang Hao
(School of Information Management, Nanjing University, Nanjing 210023, China)
(Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing 210023, China)

Abstract

[**Objective**] Through experimental comparison and analysis, this study examines the effects of different stopword lists on various types of textual data, providing reference for the construction and application of stopword lists. [**Methods**] We selected three mainstream stopword lists: the Baidu stopword list, the Harbin Institute of Technology stopword list, and the Sichuan University Machine Intelligence Laboratory stopword list. Based on three different corpora, we processed texts using Chinese word segmentation technology, the TF-IDF feature evaluation function, and the VSM model. We then conducted clustering experiments using a K-means algorithm implemented in Java, evaluating the results through three metrics: precision (P), recall (R), and F1-score. [**Results**] Different stopword lists exhibited significantly varying effects on different text types. The length and content structure of the lists were direct influencing factors, with two-character stopwords showing the most pronounced impact. [**Limitations**] The experimental text types and quantities were limited, and our analysis of different stopword lists focused only on simple comparisons of word counts and content, without categorical classification of stopwords for experimental analysis. [**Conclusions**] Stopword lists substantially affect text clustering accuracy, making it crucial to construct or select appropriate Chinese stopword lists. However, excessively increasing the number of stopwords does not consistently improve clustering results.

Keywords: Text Clustering; Stopword List; K-means

1. Introduction

In the era of rapidly developing information on the Internet, text mining technology has attracted widespread attention for processing and utilizing massive amounts of information. In 1995, Feldman et al. proposed the concept of text mining [1]. Subsequently, Ahonen-Myka et al. applied data mining techniques directly to preprocessed textual information, noting that text preprocessing is crucial for mining efficiency [2]. Text preprocessing typically consumes most of the time in text mining. For Chinese texts, this process includes Chinese word segmentation, stopword removal, feature extraction, and vector space representation, making stopword research significant.

Stopwords originated in information retrieval. Luhn discovered that some words appeared frequently but contributed poorly to retrieval effectiveness [3], and he first proposed using "noise" to represent these words [4], forming the prototype of stopwords. Later research statistically showed that the top 10 most frequent terms in English documents account for 20%-30% of total term frequency [5]. Frakes et al. argued that eliminating high-frequency words during automatic indexing could improve retrieval speed, reduce storage space, and maintain accuracy [6]. Consequently, Lo et al. defined stopwords as frequently occurring words that do not help information retrieval and should be eliminated [7]. In word-based retrieval systems, stopwords are high-frequency words with little retrieval significance, such as "的" (de), "是" (shi), "太" (tai), and "of" [8]. In question-answering systems, stopwords change dynamically depending on the question [9]. In SVM-based automatic classification, they refer to function words and neutral words with weak category characteristics [10]. In text mining, stopword identification focuses more on whether they can represent text features.

Stopwords significantly interfere with text processing. They not only carry minimal textual information but also inhibit other words, substantially affecting processing efficiency and accuracy. Yang and Pedersen found that using the top 10 stopwords to reduce feature vectors caused no negative impact, while using the top 100 had minimal negative effects [11]. Silva et al. experimentally verified that stopword removal can greatly reduce feature vector dimensionality and improve text classification accuracy [12]. Therefore, stopword removal is essential in text preprocessing.

Currently, stopwords can be removed by constructing stopword lists. These lists are categorized as general or domain-specific, with some scholars dividing them into True Full-stop Words and Semi-stop Words [13]. Their sources include manual construction and statistical automatic learning [14]. Luhn proposed the concept of "term discrimination ability," which became a common criterion for manual construction [3]. Van Rijsbergen used statistical methods to create a 250-word stopword list [15], and Fox statistically analyzed a stopword list for general English texts based on the Brown Corpus [16].

Statistical automatic learning involves continuously tagging and screening to extract high-frequency words from corpora, followed by manual verification. Mature stopword identification algorithms include document frequency, term frequency statistics, entropy calculation, and CHI statistics [17]. Reference [18] mentioned a method based on joint entropy for stopword selection. Reference [19] proposed a combined statistical and linguistic approach. Lo et al. designed a random sampling method based on terms, noting that the most effective stopword list combines classic lists with automatically extracted ones [7]. Zou et al. proposed a statistical and information-theoretic model for stopword selection [20]. In Chinese sentiment classification, five stopword lists containing different parts of speech were constructed [21]. Statistical automatic learning has become the primary method for stopword list construction, supplemented

by manual verification and achieving good results.

Domain-specific stopword lists have also gained attention in fields like medicine, chemistry, and computer science, mainly through probabilistic and content analysis of large domain text collections [22]. However, this method has limitations, with low accuracy when text distribution is uneven. Makrehchi et al. proposed using parameter- and input-sensitive classifiers to determine stopword content changes' impact on classification results, thereby identifying stopword list content [23].

While English stopword list research has achieved considerable results, Chinese stopword list research started later, with fewer in-depth studies and no widely recognized standard list. Reference [24] compared stopword lists removing different parts of speech for Chinese sentiment classification, finding that removing adjectives, adverbs, and verbs worked best, while traditional topic classification stopword lists offered little help for sentiment classification. This demonstrates that constructing or selecting precise stopword lists can yield 事半功倍 (twice the result with half the effort) effects.

Currently, mainstream general Chinese stopword lists include those from Baidu, Harbin Institute of Technology, and Sichuan University Machine Intelligence Laboratory. Given stopword removal' s significant impact on feature vector sets and classification effectiveness, this study aims to experimentally analyze commonly used stopword lists across multiple Chinese text corpora using clustering algorithms. We seek to compare their applicability and efficiency, identifying criteria for constructing, selecting, and using stopword lists for different domain text processing.

2. Experimental Design

2.1 Experimental Process The experiment consisted of four components: text collection, text processing, clustering, and effectiveness evaluation. The workflow is shown in Figure 1 [Figure 1: see original paper].

We employed three Chinese corpora: Sohu news data from Sogou Labs [25-27], the Fudan text corpus provided by the Natural Language Processing Group at the International Database Center of Fudan University' s Computer Science Department [22,28-29], and a Chinese text classification corpus mentioned in reference [30] [31]. These corpora feature high text quality, comprehensive types, and broad coverage. For text processing, we used ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) for word segmentation and selected three widely used stopword lists: Baidu, Harbin Institute of Technology, and Sichuan University [32-33]. We adopted TF-IDF for feature selection and the VSM model for text vector representation. For clustering, we used a simple and efficient K-means algorithm implemented in Java, evaluating results using precision (P), recall (R), and F1-score based on human-annotated standards [16].

Experimental texts were randomly sampled from each corpus. From five categories (economy, IT, military, sports, and arts), we selected texts with IDs divisible by 5, relabeled them, and extracted 640 texts from each corpus, totaling 1,920 test texts (Table 1). Each text belonged to at most one category, including news, literature, and abstracts. We used controlled variable methods for testing, accurately recording results for analysis.

2.2 Stopword List Content Analysis Table 2 shows the basic characteristics of the three stopword lists. The lists differed considerably. The Baidu list contained single characters, English stopwords, and Chinese stopwords (e.g., "able," "一," "不是"), with a high proportion of two-character words. The Sichuan University list included many common idioms and three/four-character phrases (e.g., "打开天窗说亮话," "何乐而不为," "换言之"), with relatively few single-character words. The Harbin Institute of Technology list contained numerous Chinese and English characters (e.g., "*," "Δ," ".....").

The Baidu list had 1,395 stopwords, primarily due to 547 English stopwords. Two-character words dominated all three lists, with the Sichuan University list containing 663 such words, evidently to maximize matching and removal since most Chinese segmentation results are two-character strings [34].

Table 3 shows the overlap among the three lists. Single-character, three-character, and four-character words showed high overlap (over 80%), with differences mainly in two-character words. The Baidu and Sichuan University lists had similar two-character word counts with approximately 50% overlap, while the Baidu and Harbin Institute of Technology lists showed high overall overlap but differed in two-character word inclusion. The three lists shared 337 stopwords. We merged them into a new "full stopword list" to test whether list length affected clustering effectiveness.

3. Experimental Results

We fixed each stopword list for comparison and performed clustering on the Sogou, Fudan, and Chinese text corpora. For each cluster, we counted text types by category, assigned the majority type as correct, and calculated P, R, and F1 values. Tables 4 through 6 show clustering results for the three lists. Table 7 presents results after removing English words from the Baidu list, while Table 8 shows results using the full stopword list.

4. Experimental Analysis

The experimental data were directly calculated. To better compare the three stopword lists, we integrated the data, examining effects on the same text type (economy, military, etc.) and on the same corpus.

4.1 Text Domain Analysis Our data covered five domains: economy, IT, military, sports, and arts. With fixed stopword lists, we used three corpora to

obtain different clustering results. We extracted F1 values for each text type across corpora and averaged them, shown in Tables 9 and 10 .

Table 10 reveals: (1) Among the five domains, arts achieved the best clustering performance, while military performed poorly, with nearly 50% difference in F1 scores. This relates to the smaller military text quantity, indicating that clustering requires sufficient texts to extract accurate features and build precise vectors, avoiding interference in subsequent analysis. (2) The Baidu list performed best overall, with F1 scores 0.049 and 0.069 higher than the other two lists, particularly excelling in economy and military domains. Its strength lies in high-quality two-character stopwords. Despite similar counts to the Sichuan University list, its effectiveness was significantly higher, confirming that two-character stopwords most critically impact clustering. New stopword lists should prioritize two-character words. (3) The Harbin Institute of Technology list excelled in IT and arts clustering, while the Sichuan University list worked best for sports. The former contains unique Chinese/English characters, while the latter includes more three/four-character stopwords—key differences driving varying effectiveness. (4) This domain-specific analysis provides reference for constructing specialized stopword lists.

4.2 Different Corpora Analysis We summed and averaged F1 values across text types for each corpus to evaluate each stopword list' s clustering effects, shown in Table 11 and Figure 2 [Figure 2: see original paper].

The Baidu list worked best for the Sogou corpus, the Sichuan University list suited the Chinese text corpus, and the Harbin Institute of Technology list fit the Fudan corpus. Performance variations mainly depend on corpus text types and formats. The Fudan corpus comprises primarily academic literature with few news comments. The Sogou corpus contains only news from major portals. The Chinese text corpus mixes literature, news, and emails. These stopword lists perform better on academic literature, with Harbin Institute of Technology being superior. For news, Baidu shows clear advantages, while Sichuan University performs poorly but suits emails and literature better.

Combining Sections 4.1 and 4.2: (1) As an intermediate text processing step, stopword removal is crucial. Its predecessor, Chinese word segmentation, affects stopword matching. For example, "近年来" (in recent years) can be segmented as "近年/来" or as a whole. The first approach allows all three lists to remove stopwords, while only the Sichuan University list can handle the second segmentation because it contains this three-character stopword. Stopword list construction should consider multiple segmentation scenarios for maximum compatibility. (2) Stopword removal precedes feature vector extraction, aiming to eliminate useless words while preserving thematic representation and reducing dimensionality. Different lists' varying effectiveness stems from their source corpora. Selecting stopword lists similar to the target corpus yields better clustering results, suggesting that domain-specific lists should be built from large domain text collections.

4.3 Analysis of Baidu List Without English Stopwords Since the Baidu list contains many English words but our corpora are Chinese, we removed English stopwords and compared results in Table 12 .

Table 12 shows that some corpora (italicized) were unaffected due to minimal English content. Others (bolded) improved slightly after removal, while some declined. IT texts often contain domain-specific English terms that serve as valuable features; removing them reduces accuracy. Some military texts contain useless English words whose removal improves clustering. Overall, English words appear infrequently and often have specific meanings, suggesting English stopwords need not be considered for Chinese text clustering.

4.4 Full Stopword List Analysis We merged the three lists into a full stopword list for clustering and compared it against each list' s best performance (Table 13).

The full list improved over individual lists generally, but compared to each list' s optimal performance for specific text types, its advantage was minimal—only improving military texts by 0.082 while decreasing economy and sports texts by 0.108 and 0.133. This demonstrates that more stopwords are not always better; specificity matters. Effectiveness depends on optimizing for text source and type.

Comparative analysis reveals that differences among the three lists are significant, mainly in two-character words regarding content and quantity, stemming from different source corpora and application scopes. The Baidu list performed best on average across the three corpora, with slight improvement after removing English words. All three lists worked better on arts texts than other categories but poorly on military texts. The full list did not achieve optimal results and sometimes reduced precision. Therefore, selecting appropriate stopword lists is critical for clustering tasks. Building domain-specific lists or more comprehensive general lists would be beneficial and represents a future research direction. This study has limitations: small text quantities may cause 偶然性 (randomness), the single K-means algorithm may influence results, and we did not categorize stopwords by type for analysis. Future research will address these issues.

References

[1] Feldman R, Dagan I. Knowledge Discovery in Textual Databases (KDT)[C]//Proceedings of International Conference on Knowledge Discovery and Data Mining. 1995: 112-117.
[2] Ahonen-Myka H, Heinonen O, Klemettinen M, et al. Applying Data Mining Techniques in Text Analysis[R]. Technical Report C-1997-23, Department of Computer Science, University of Helsinki, 1997.
[3] Luhn H P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information[J]. IBM Journal of Research and Development, 1957, 1(4): 309-317.

[4] Luhn H P. The Automatic Creation of Literature Abstracts[J]. IBM Journal of Research Development, 1958, 2(2): 159-165.
[5] Francis W N, Kučera H, Mackie A W. Frequency Analysis of English Usage[J]. Frequency Analysis of English Usage Lexicon & Grammar, 1982, 18: 64-70.
[6] Frakes W B, Baeza-Yates R. Information Retrieval: Data Structures and Algorithms[M]. Prentice-Hall, Inc., 1992.
[7] Lo R W, He B, Ounis I. Automatically Building a Stopword List for an Information Retrieval System[J]. Journal of Digital Information Management, 2005, 3(1): 3-8.
[8] Jiang Zhaozhong. Chinese Words Segmentation Based on Context and Stopwords[D]. Hefei: Hefei University of Technology, 2010.
[9] Xiong Wenxin, Song Rou. Removal of Stop Word in Users' Request for Information Retrieval[J]. Computer Engineering, 2007, 33(6): 195-197.
[10] Zhou Qinqiang, Sun Bingda, Wang Yi. Study on New Pretreatment Method for Chinese Text Classification System[J]. Application Research of Computers, 2005(2): 85-86.
[11] Yang B Y, Pedersen J O. A Comparative Study on Feature Selection in Text Categorization[C]//Proceedings of International Conference on Machine Learning. 2010.
[12] Silva C, Ribeiro B. The Importance of Stop Word Removal on Recall Values in Text Categorization[C]//Proceedings of the International Joint Conference on Neural Networks. 2003, 3: 1661-1666.
[13] Tomov D T. Some Critical Remarks on the Stop Word Lists of ISI Publications[J]. Journal of Documentation, 2001, 57(6): 767-773.
[14] Hua Bolin. Stop-Word Processing Technique in Knowledge Extraction[J]. New Technology of Library and Information Service, 2007(8): 48-51.
[15] Van Rijsbergen C J. Information Retrieval[M]. London: Butterworths, 1975.
[16] Fox C. A Stop List for General Text[J]. ACM SIGIR Forum, 1990, 24(1-2): 19-21.
[17] Chen Xin, Zhang Jing, Li Xiaoguang, et al. A Text Classification Method for Chinese Pornographic Web Recognition[J]. Measurement & Control Technology, 2011, 30(5): 27-31.
[18] Gu Yijun, Fan Xiaozhong, Wang Jihua, et al. Automatic Selection of Chinese Stoplist[J]. Transactions of Beijing Institute of Technology, 2005, 25(4): 337-340.
[19] Cui Caixia. Research on the Effect of Stop Words Selection on Text Categorization[J]. Journal of Taiyuan Normal University: Natural Science Edition, 2008, 7(4): 91-93.
[20] Zou F, Wang F L, Deng X, et al. Automatic Construction of Chinese Stop Word List[C]//Proceedings of the International Conference on Applied Computer Science. 2006: 16-18.
[21] Wang Suge, Wei Yingjie. The Influence of Stoplist on the Chinese Text Sentiment Categorization[J]. Journal of the China Society for Scientific and Technical Information, 2008, 27(2): 175-179.

[22] Zhou Yao. Cloud Computing-based Research on Text Mining Techniques[D]. Changsha: National University of Defense Technology, 2011.
[23] Makrehchi M, Kamel M S. Automatic Extraction of Domain-Specific Stopwords from Labeled Documents. Proceedings of European Conference on IR Research (ECIR 2008), Glasgow, UK. 2008: 222-233.
[24] Hua Linsen. Study on Chinese Text Sentiment Classification[D]. Chongqing: Chongqing University, 2014.
[25] Sogou Labs. Sohu News Data[DB/OL]. [2016-07-05]. http://www.sogou.com/labs/resource/cs.php.
[26] Li Mei. Study of Chinese Text Clustering on Improved K-means Algorithm[D]. Hefei: Anhui University, 2010.
[27] Huang Lei, Wu Yanpeng, Zhu Qunfeng. Research and Improvement of TFIDF Text Feature Weighting Method[J]. Computer Science, 2014, 41(6): 204-207.
[28] Data Hall. Text Classification Corpus (Fudan) Test Corpus[DB/OL]. [2016-07-05]. http://www.datatang.com/datares/go.aspx?dataid=615059.
[29] Hu Xiaohui. The Research on Text Classification Based on Clique Model[D]. Nanchang: Jiangxi Normal University, 2008.
[30] Sun Guoju, Zhang Jie. An Evaluation of Feature Selection Methods for Text Categorization[J]. Journal of Harbin University of Science and Technology, 2005, 10(1): 76-78.
[31] Data Hall. Chinese Text Categorization Corpus[DB/OL]. [2016-07-05]. http://www.datatang.com/data/11971/.
[32] Data Hall. Stop Words Set[DB/OL]. [2016-07-05]. http://www.datatang.com/data/19300/.
[33] Yu Juan, Yin Jidong, Fei Shu. Identifying Synonyms Based on Sentence Structure Analysis[J]. New Technology of Library and Information Service, 2013(9): 35-40.
[34] Fei Hongxiao, Kang Songlin, Zhu Xiaojuan, et al. Chinese Word Segmentation Research Based on Statistic the Frequency of the Word[J]. Computer Engineering and Applications, 2005, 41(7): 67-68.

Author Contributions

Guan Qin: Data selection, experimental implementation, draft writing.
Deng Sanhong, Wang Hao: Research design, final manuscript revision.

Conflict of Interest Statement

All authors declare no conflict of interest.

Supporting Data

Supporting data is self-archived by the authors, available at: mf1614036@smail.nju.edu.cn.
[1] Guan Qin, Deng Sanhong, Wang Hao. SohuData.zip. Sohu News Data.
[2] Guan Qin, Deng Sanhong, Wang Hao. FudanData.zip. Text Classification Corpus (Fudan) Test Corpus.
[3] Guan Qin, Deng Sanhong, Wang Hao. ChineseTextData.zip. Chinese Text

Classification Corpus.
[4] Guan Qin, Deng Sanhong, Wang Hao. Stopwords.zip. Stop Words Set.

Received: December 5, 2016
Revised: December 25, 2016

Note: Figure translations are in progress. See original paper for figures.