Multi-feature automatic abstract based on LDA model and redundant control

Tang Dan\textsuperscript{1, a}, Sun Yu\textsuperscript{1*}

\textsuperscript{1}Yunnan Normal University, Kunming, Yunnan Province, 650500, China

\textsuperscript{*}Corresponding author's e-mail: sunyu_km@hotmail.com

Abstract: With the continuous popularization of computer application technology and the rapid development of Internet technology, there has been an explosion of information in every field, and more and more information is transmitted and stored on the Internet in the form of electronic text. Automatic summarization can solve the problem that the speed of traditional hand-written summaries cannot keep up with the speed of electronic text information generation. Automatic summarization can use a computer to generate a comprehensive and accurate coherent short essay on one or more documents, so people can quickly understand the main content of the text, which greatly improves the efficiency of people's learning and work. Nowadays, automatic summarization technology has achieved rich results, but Chinese document summarization is still in the shallow semantic analysis of text, and there are problems such as redundancy of summarization and poor quality of abstract content. In view of the above problems, this article is in the research and implementation of Chinese automatic summarization technology. On the one hand, LDA is used to mine the topic distribution of documents to obtain high-quality abstracts; on the other hand, the similarity between the sentences to be extracted is calculated to remove redundant information.

1. Introduction

Abstract generation methods are divided into two types: abstract based on extract and abstract based on comprehension [1]. The extraction-based automatic summarization system mainly extracts the sentences related to the topic from the sentences of the document as the summary sentence. The comprehension-based automatic summarization system obtains the main idea of the article by analyzing the syntax and grammar, and reorganizes the words and sentences to generate the summary. At present, there is no direct method for comprehensible summarization, so most of them use abstraction-based methods.

The current multi-document automatic summarization has achieved remarkable results, but the abstracts generated by the current automatic summarization technology have problems such as content redundancy, incoherent sentences, and inaccurate and comprehensive topics. Therefore, this article aims to obtain comprehensive, accurate, concise and coherent abstract sentences. So a multi-feature automatic abstracting method based on LDA and redundant control is proposed. First, we perform LDA modeling on multi-document collections. LDA contains a three-layer structure of words, topics, and documents. Words and documents can be associated through shallow topics to obtain a specific topic distribution and then calculate the KL divergence distribution. Secondly, the importance of the sentence is calculated by combining weights such as word frequency, length characteristics, location characteristics, sentence and title similarity. Finally, the sentences with the highest ranking of sentence importance are extracted, and the similarity of each extracted sentence is compared with the selected abstract sentence. If the
similarity exceeds a certain threshold, the sentence will not be included in the abstract sentence. This avoids extracting similar sentences, and can greatly reduce redundancy.

2. **Context-related recognition**

Foreign people such as Luhn, L.F.Rau, Haghighi, Je Nichols, Blei, Asli and others have made great contributions to the research of automatic abstracts[2-7]. Domestic scholars, such as Wang Yongcheng, Wang Kaizhu, Wu Lide, Zhong Yixin, Qin Bing, Liu Ting, Song Xuanchen, etc. have also achieved great success in the field of automatic abstracts[8-15]. Multi-feature automatic summarization based on LDA and redundant control is mainly divided into the following three parts:

2.1 **Multi-document LDA modeling**

LDA is a multi-layered production probability model, which contains a three-layer structure of words, topics, and documents, and associates words and documents through shallow topics. LDA makes the Bag of Words hypothesis, that is, the grammar or order of words is not considered in the model, only the number of times they appear.

2.2 **Sentence weight calculation**

The score of a sentence can be used as the weight of the sentence. The greater the weight, the more likely it is an abstract sentence. The importance of the sentence is evaluated by calculating the similarity between the topic probability distribution of the sentence and the document topic probability distribution, and then the KL divergence is used to calculate the similarity between the two distributions. The score of the sentence is calculated as the weighted sum of the feature values.

2.3 **Extraction of abstract sentences**

The score of each sentence in the document set is calculated according to the above features, and the sentences are uniformly sorted according to the size of the score. The system extracts sentences according to the sentence scores from large to small, and compares the extracted sentences with the existing abstract sentences. Calculate and filter out sentences that are too similar until the abstract reaches the specified number.

3. **Feature combination based on LDA model**

3.1 **Basic features**

3.1.1 **Word frequency features**

In this paper, the term frequency is calculated by the TF-IDF (Term Frequency/Inverse Document Frequency) method. TF-IDF is the current general keyword importance measurement. TF stands for text word frequency, IDF stands for inverse text frequency index. The TF-IDF method can retain important vocabulary and filter out frequently used words.

Sum up the TF-IDF of each non-stop word in the sentence to know the importance of the sentence in the text. The calculation formula is as follows:

$$F(s_i) = \sum_{w_l \in s} f_{TF-IDF}(w_l)$$

Among them, $f_{TF-IDF}(w_l)$ represents the value of TF*IDF of vocabulary $w_l$ in $s$ in the sentence.

3.1.2 **Location feature**

In Chinese documents, the author generally has a rigorous structure for the content of the article, instead of stacking together the information to be expressed. The beginning of the article generally assumes the function of leading the full text and highlighting the main theme; the end generally sublimates the theme and summarizes the full text; the body of the article generally elaborates the information to be conveyed. From this we can know that the beginning and end of the article usually contains important information
about the document. Based on this feature of the text structure, we can regard the text as a linear sequence of several sentences. The weight formula of the sentence is as follows:

\[ P(s_i) = \begin{cases} 0.8, & s_i \text{ is the first sentence of the paragraph} \\ 0.2, & s_i \text{ is the end sentence of the paragraph} \end{cases} \]

### 3.1.3. Sentence length feature

When selecting abstracts for automatic abstracts, the length of sentences will also affect the quality of abstracts. Too short sentences may convey too little important information, and too long sentences may have some redundant information. Therefore, extract sentences avoiding extracting sentences that are too long or too short, and try to extract key sentences that fit the topic. Use regular expressions to remove the interference of non-critical sentences on the abstract, the formula is as follows:

\[ L(s_i) = 1 - \frac{|l(s_i) - l_{avg}(D_i)|}{l_{avg}(D_i)} \]

Among them, \( l(s_i) \) represents the number of words contained in sentence \( s \), that is, the length of sentence \( s \), and \( l_{avg}(D_i) \) represents the average length of sentences in the document collection.

### 3.1.4. Sentence and title similarity

If the sentence is more similar to the title of the document to which it belongs, then it can be considered that the sentence expresses the central theme of the document and contains more important information. Then the sentence is most likely to be called an abstract sentence. The similarity between the sentence and the document title can be converted into a comparison between the sentence vector and the document title vector, and the closeness between the sentence and the title can be judged by the angle between the vectors. Because we want to calculate the angle between vectors, this article uses cosine similarity to calculate the similarity between the sentence and the title. The formula is as follows:

\[ \text{Sim}(s, t) = \frac{\sum_{i=1}^{n} w_{si} \times w_{ti}}{\sqrt{\sum_{i=1}^{n} (w_{si})^2 \times \sum_{i=1}^{n} (w_{ti})^2}} \]

Among them, \( w_{si} \) represents the weight of the word frequency feature of the i-th word in the sentence \( s \), and \( w_{ti} \) represents the weight of the word frequency feature of the i-th word in the document title \( t \).

### 3.2 Feature combination based on LDA model

This article chooses to use the commonly used LDA topic model to model the document collection. The most important part of the model is the calculation of the similarity of the text unit. We can regard the text unit as a mixture of topics, using a K-dimensional vector Represents the mixing ratio of this topic, namely the document-topic mixed distribution probability, and the sentence-topic mixed distribution probability. If the document-topic distribution probability is similar to the sentence-topic distribution probability, we think that the topic of the sentence is similar to the topic of the document, and the more likely it is that the sentence is a summary sentence.

Based on the above ideas, we use KL (Kullback-Leibler) divergence to calculate the dispersion between the topic distributions of two text units.

\[ D_{KL}(P||Q) = \sum_{i} P(i) \times \log \frac{P(i)}{Q(i)} \]

Among them, \( P \) and \( Q \) represent two probability distributions, and \( P(i) \) and \( Q(i) \) represent the i-th component of the P and Q distributions.

Because the distance represented by KL divergence is not symmetric, when calculating the correlation between two probability distributions \( D_{KL}(P, Q) \), \( D_{KL}(P||Q) \) and \( D_{KL}(Q||P) \) are used. In order to make the distance between the two probability distributions can be used to express the relevance of the distribution, we reverse \( D_{KL}(P, Q) \), the formula is as follows:

\[ T_{sim}(P, Q) = -\frac{D_{KL}(P||Q) + D_{KL}(Q||P)}{2} \]
Combining the feature values of the above sentences and summing them by weight, the final weight calculation formula of the sentence can be obtained as:

$$\text{Score} = \lambda_1 F(s_i) + \lambda_2 P(s_i) + \lambda_3 L(s_i) + \lambda_4 \text{Sim}(s, t) + \lambda_5 \text{Tsim}(P, Q)$$

In this experiment, when the parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ are set to \{0.1, 0.1, 0.2, 0.2, 0.4\}, the experimental effect is the best.

3.3 Redundancy control

At present, the commonly used redundancy control methods include cluster-based methods and candidate-based methods. The cluster-based method is to calculate the similarity between all sentences, and then use the clustering method to identify the topic of public information. The candidate-based method is to calculate the similarity between the candidate abstract and the selected abstract. Only when the candidate abstract has enough different information can it be selected as an abstract sentence.

The text reduces redundant sentences with the same subject in the abstract by calculating the similarity between the candidate abstract and the selected abstract. The following method is proposed to determine whether the candidate abstract is redundant:

$$\text{Sim}(s_1, s_2) = \frac{1}{2} \times \left( \frac{\text{Samewords}(s_1, s_2)}{l(s_1)} + \frac{\text{Samewords}(s_1, s_2)}{l(s_2)} \right)$$

Among them, $s_1, s_2$ represent the selected abstract and candidate abstract respectively, Samewords($s_1, s_2$) represents the number of the same words in the sentence $s_1, s_2$, and $l(s_1), l(s_2)$ represent the different words in the $s_1, s_2$.

Therefore, the summary data selection process of redundancy control in this paper is as follows:

1. Input the extracted candidate abstract set;
2. Through the calculation of feature values, the sentence with the highest score, that is, the highest weight, is stored in the abstract as the first selected abstract sentence, and this sentence is removed from the abstract candidate set;
3. Use the above formula to calculate the similarity between each sentence in the candidate abstract and the selected abstract sentence. If the similarity is less than a given threshold, then the sentence can be used as an abstract sentence, and this sentence is removed from the abstract candidate set; sentence;
4. Repeat step 3, if the selected summary length reaches the specified length, end this process;
5. Output a collection of summary sentences.

4. Experimental results and analysis

4.1 Evaluation method

The evaluation method of this paper uses fuzzy labeling method for each topic. In the labeling process, in addition to marking the standard abstract sentence in the source document collection, it also indicates that the standard abstract sentence can be replaced in the source document and cannot be combined with the standard abstract sentence in the abstract. Sentences that co-occur in the middle are called candidate abstract sentences. Each candidate abstract sentence is assigned a weight between (0, 1] according to the degree of substitution. The evaluation corpus obtained in this way can use the accuracy, redundancy and overall quality to evaluate the quality of the abstract system. In order to solve the problem that there are multiple replaceable abstract sentences in the text collection that cannot be considered in the traditional multi-document automatic abstract evaluation. On this basis, the accuracy, redundancy and comprehensive quality are used to evaluate the system under test:

$$\text{Precision} = \frac{\left( \sum_{i=1}^{K} \varphi_i \right)}{K}$$

$$\text{Redundancy} = \frac{\left( \sum_{i=1}^{K} \left( \sum_{j=1}^{K} \varphi(S_i, S_j) \right) \right)}{K}$$

$$\text{Total(summary)} = \text{Precision} - \text{Redundancy}$$
Among them, K is the total number of sentences in the abstract to be evaluated. K1 is the total number of sentences in the sentence to be evaluated in the sentence of the standard abstract, (w1,w2,….wk) is the weight of each sentence, which is obtained by the above manual annotation method; \( \varphi(S_i, S_j) \) is a binary discriminant function, when \( S_i, S_j \) are similar sentences, \( \varphi(S_i, S_j)=1 \), otherwise it is 0.

4.2 Result analysis

The three indicators of the experiment in this article are tested according to the length of 10 sentence abstracts. Model-1 is a traditional abstract algorithm, that is, the weighted average of four basic features is used as the weight of the sentence. Model-2 is the topic distribution algorithm of LDA modeling, namely The similarity of the topic probability distribution between the sentence and the document is used as the weight of the sentence (Table 1 is the experimental result).

| Document number | Model-1 P | Model-1 R | Model-1 T | Model-2 P | Model-2 R | Model-2 T |
|-----------------|-----------|-----------|-----------|-----------|-----------|-----------|
| 1               | 0.10      | 0.10      | 0.74      | 0.92      | 0.10      | 0.73      |
| 2               | 0.58      | 0.10      | 0.82      | 0.68      | 0.10      | 0.76      |
| 3               | 0.38      | 0.10      | 0.28      | 0.95      | 0.10      | 0.86      |
| 4               | 0.48      | 0.10      | 0.38      | 0.85      | 0.10      | 0.56      |
| 5               | 0.92      | 0.10      | 0.44      | 0.90      | 0.10      | 0.83      |
| 6               | 0.84      | 0.10      | 0.46      | 0.84      | 0.10      | 0.70      |
| 7               | 0.30      | 0.10      | 0.70      | 0.86      | 0.10      | 0.82      |
| 8               | 0.44      | 0.10      | 0.34      | 0.97      | 0.10      | 0.68      |
| 9               | 0.20      | 0.10      | 0.28      | 0.77      | 0.10      | 0.48      |
| 10              | 0.17      | 0.10      | 0.07      | 0.68      | 0.10      | 0.46      |
| average value   | 0.44      | 0.10      | 0.45      | 0.84      | 0.10      | 0.68      |

The data in Table 1 shows that the accuracy of the abstract results obtained by the method of calculating sentence importance based on the topic distribution is better than the abstract results obtained by the traditional abstract algorithm. It can be seen from Table 1 that the accuracy rates of 10 document collections in the following two methods have increased in the same trend, but the redundancy has not changed much, which shows that this method of distinguishing sentence redundancy needs improvement.

5.Conclusions

The method in this paper uses LDA to model the document collection, obtains the topic probability distribution of the sentence and the topic probability distribution of the document, calculates the similarity between the probability distributions as the importance of the sentence, and extracts abstract sentences according to the importance of the sentence. The experimental results show that the performance of the abstract obtained by this method is better than that of the traditional abstract method. However, the method in this paper still tends to extract long sentences, and long sentences affect the performance of the system. Future work will consider compressing the extracted abstracts on the basis of performance improvement; in addition, because the bag of words assumption is used in the LDA model, it cannot express the structural relationship among sentences, documents, and document collections well. In future work, the LDA model will be expanded to make it more suitable for automatic summarization.

References

[1] Naaman M, Rattenbury T, Good N. Automatic extraction of semantics from text information: US, US 8060491 B2[P]. 2011.
[2] Luhn H P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information[J]. IBM Journal, 1957, 309-317.
[3] Lisa F. Rau, Paul S. Jacobs, Uri Zernik. Information Extracting and Text Summarization Using Linguistic Knowledge Acquisition [J]. Information Processing & Management, 1989, 25(4):
[4] A Nenkova. LVanderwende. The impact of frequency on summarization. Technical report, Microsoft Research, 2005.

[5] Nichols J, Mahmud J. Summarizing Sporting Events Using Twitter [C]. Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces. ACM 2012: 189-198.

[6] Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation [J]. The Journal of Machine Learning Research, 2003; 993-1033.

[7] Asli Celikyilmaz. Dilek Hakkani-Tur. A hybrid hierarchical model for multi-document summarization [J]. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 815-824, Uppsala, Sweden, 11-16 July 2012.

[8] Su Haiju, Wang Yongcheng. Automatic compilation of Chinese scientific and technological literature abstracts [J]. Journal of Information, 1989, 8(6): 433-439.

[9] Wang Yongcheng, Xu Huimin. OA Chinese Document Automatic Summarization System [J]. Journal of Information, 1997, 16(2): 128-132.

[10] Wang Jianbo, Wang Kaizhu. Natural language text comprehension and automatic summarization based on comprehension [J]. Chinese Journal of Information, 1992, 6(2): 1-7.

[11] Wu Lide, Wei Xiong guan. Fudan Abstract System of Chinese Text [J]. Communication of COLIPS 1996(1): 35-39.

[12] Yang Xiaolan, Zhong Yixin. Research and implementation of automatic summarization system based on full-information dictionary [J]. Chinese Journal of Electronics 1997.12. Vol.16 Issue 6.

[13] Yang Xiaolan, Wang Minghui. Research and Implementation of Text Understanding, Knowledge Acquisition and Automatic Summarization System [C]. Electronic Devices, Vol. 21, No. 1, 1997: 394-399.

[14] Qin Bing, Liu Ting, Li Sheng. A review of multi-document automatic summarization [J]. Journal of Chinese Information Processing, 2005, 19(6): 13-20.

[15] Song Xuanchen. Research on automatic multi-document summarization based on statistics and semantic analysis [D]. University of Science and Technology of China. 2009.5