An Automatic Text Summary Extraction Method Based on Improved TextRank and TF-IDF

Xinxin Guan¹,a, Yeli Li¹,b, Qingtao Zeng¹,c and Chufeng Zhou¹,d
¹ School of Information Engineering, Beijing Institute of Graphic Communication, Beijing 102600, China
a 1145512971@qq.com, b liyl@bigc.edu.cn, c zengqingtao@bigc.edu.cn,
d 576213376@qq.com

Abstract. In the era of information explosion, more and more people advocate "fast reading". Compared with some texts longer texts can express vivid images but it will waste time. Therefore, automatic text abstraction has become one of the research hotspots in natural language processing. In order to ensure the accuracy of automatic text abstraction by computer, a weighting method based on comment factor, position factor and weighting of part-of-speech factor is designed which mainly combines and improves TextRank and TF-IDF algorithm. This method focuses on introducing the reader's comments as a comment factor, supplementing the shortcomings of paying too much attention to the original text and ignoring the reader's comments. Secondly, the frequency factor and part-of-speech factor are introduced to ensure the readability of the generated abstracts. The experimental results show that the method has good improvement in accuracy, consistency and connectivity.

1. Introduction
With the rapid development of science and technology, more and more people access the Internet through computers, mobile phones and other terminals, which has also changed people's general reading and writing habits, making a large number of electronic forms of information appear on the Internet, even though the network has become an unprecedented treasure house of knowledge. In the era of data explosion, the number of network users has also increased. More and more people rely on the Internet to obtain information and resources. Network text is also growing exponentially. Now, fast reading is advocated. How to extract similar text abstracts from the original text without changing its meaning has become one of the current research hotspots.

Text Abstract extraction is to use computer to automatically extract and generate topic sentences of articles, so that users can grasp the central idea of articles in a short time, reduce the time spent by users to find information, and has a wide range of application value. In recent years, text summarization technology has been paid attention and studied by scholars at home and abroad, and scholars have put forward many excellent algorithms.

According to the number of documents covered by the summary, automatic text summary can be divided into single document summary and multi-document summary. Single document digest technology generates digest for a single document, while multi-document digest technology generates digest for documents with similar topics. [1]

In recent years, a large number of research results have been achieved on Chinese text automatic summarization technology based on Chinese characteristics. In 2005 Xiaodan Xu proposed that we should first analyze the information characteristics of the web page, and then obtain the text
information of the web page. Then the statistical method and heuristic rules are combined to extract keywords and key sentences from documents. Finally, according to the proportion of summary output qualified summary sentences [2]. In 2014 Yang Liu and Lei Cui proposed to study the citation context, to explore the consistency between the citation context, the abstract of the target document and the target document, and to match the subject words in segments [3]. In 2014 Rui Yang proposed that the abstract feature attributes should be summarized from the abstract point of view. The weight of each feature attributes can be obtained by genetic optimization algorithm, and then the optimal abstraction can be found by particle swarm optimization algorithm [1]. In 2015, Shibo Zhang and Boai Liu proposed that the method of generating multi-document summaries is mainly based on multiple semantic features and semantic similarity to calculate the importance score of sentences. The stack decoder algorithm is used to generate abstracts from multiple stacks. [4]. In 2016 Yanqing Guo, Rui Zhao, Xiangwei Kong and Haiyan Fu proposed a news summary extraction method based on weighted event elements, which improves the transfer probability matrix and increases the detail and readability of the output timeline summary by weighting the event elements [5]. In 2017 Xiaoping Chen proposed that extracts the short text of micro-blog automatically, calculates the weight of word vector by gradient descent algorithm, and obtains the weight of word-sentence relationship by iteration[6]. In 2017 Nainai Li, Peiyu Liu, Wenfeng Liu and Weitong Liu proposed, for Chinese single document, TextRank algorithm is used to generate text summary by combining the whole structure information of the text and the context information of the sentence [7]. In 2018 Fang Su, Xiaoyu Wang and Zhi Zhang proposed that the neural network model of sequence-to-sequence learning is analyzed, so that the words appearing at the beginning of the original text appear more in the abstract than in the abstract[8]. In 2019 Yun Wu, Changchun Yang, Jiajun Mei and Huan Gu proposed that an automatic summary extraction algorithm (F-CoRank) for word-sentence collaboration is proposed. Enhance the word frequency of feature words similar to the title, construct undirected network graph, and get the weight of each node to generate summary [9].

This paper mainly focuses on the automatic abstraction of news, blog articles, forum posts, technical papers and other types of text in the online network.

2. Related Work

2.1. Data Source Acquisition

The data source of this paper is to crawl news, blog articles, forum posts and technical papers in the online network through Python language programming. According to the data needs of this paper, the main content is to obtain news, blog articles, forum posts, technical papers and comments.

2.2. Data Preprocessing

De-stop words refer to those words that occur very frequently in natural environments but have no substantial impact on the meaning of text or pages. Removing stop words can prevent noise interference and effectively improve keyword density.

Chinese word segmentation refers to dividing a sequence of Chinese characters into individual words. Word segmentation is the process of reorganizing continuous word sequences into word sequences according to certain norms. There are many Chinese word segmentation tools, such as Jieba word segmentation, LTP, Boson NLP, IK Analyr. The Chinese word segmentation tool used by researchers is HanLP Chinese word segmentation, which calls the HanLP Java toolkit through Python. HanLP provides a rich function, stable and clear framework, and the corpus in HanLP is constantly updated to ensure the effect of word segmentation.

2.3. Denoising

In this paper, the chaotic text and pictures on the web page are removed, and only the text that has been well-arranged is retained. As well as grabbing the effective content of Internet comments for effective information acquisition, while deleting comments that are not related to the subject.
2.4. TextRank

TextRank arithmetic is derived from PageRank, which is a text processing model based on graph sorting. The weight of a word in TextRank depends on the weight of the edge that is composed of the previous points and the sum of the weights of this point to the other edges. If a word appears after many words, then it is more important to say that the word is important.

If a word with a high TextRank value is followed by a word, the TextRank value of the word will be increased accordingly so that the formula of TextRank can be changed from PageRank formula to:

\[
S(v_i) = (1-d) + \sum_{(j,k) \in d} w_{jk} \sum_{(i,j) \in d} w_{ij} S(v_j)
\]

2.5. TF-IDF

TF-IDF (term frequency-inverse document frequency) is a commonly used weighting technology for information retrieval and text mining. The main idea is that if a word or last sentence is relatively rare, but it appears many times in a text, then it probably reflects the characteristics of this article.

In preprocessing, stop words are filtered out, and the remaining words are meaningful words. For a text review, the words that appear many times are more important words, namely word frequency. The specific expression is 2 and the formula is 3.

\[
TF(t) = \frac{\text{Number of times term } t \text{ appears in a document}}{\text{Total number of terms in the document}}
\]

\[
tf_{ij} = \frac{n_{t,j}}{\sum_i n_{t,i}}
\]

In the above formula, \(n_{t,j}\) is the number of occurrences of the word in the file \(d_j\), and the denominator is the sum of the occurrences of all the words in the file \(d_j\).

\[
IDF(t) = \frac{\text{Total number of files in corpus } d}{\text{Number of words } t \text{ appears in the document}}
\]

The word frequency in TF is limited to only one text, but when a word is also a high frequency word in a collection or database, then the word is not so important. IDF, an important weight adjustment parameter, is introduced to measure the importance of a word. Its size is inversely proportional to the common degree of a word. The concrete expression is 4 and the formula is 5.

\[
idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}
\]

among them:

\(|D|\) indicates the total number of files in the corpus;

\(\{j : t_i \in d_j\}\) Indicates the number of words \(t\) appearing in the file.

Combine (3) and (5) to get (6):

\[
\text{tf-idf}_{ij} = tf_{ij} \times idf_i
\]

The traditional TF-IDF algorithm only considers the importance of word frequency in the text. It does not take into account the semantic information of words and the comments of some readers on the text below. If only the method is used alone, the abstracts extracted are not particularly accurate. Therefore, this paper first extracts the key words of a comment with TF-IDF, then according to the topic sentences extracted with TextRank, weights them again and finally gets the text summary.
3. Experimental Process

3.1. Experimental Procedure
The main flow charts of the experiment are as Figure 1:

The specific steps are as follows:
A. Enter text, segment the document by line breaks, and clause each paragraph by line breakers.
B. Mark the position of the article. Through the mark list, mark whether it is the first, last, first and last sentence, that is, the position factor of the sentence.
C. TextRank is used to extract the key words in the text; the results are merged into a sentence for word segmentation; at the same time, part of speech is extracted, nouns and verbs are extracted, and the part of speech factors of nouns and verbs are marked respectively.
D. Input comments and denoise them, removing comments and symbols that are not related to the topic.
E. uses TF-IDF to extract the key words of comments and retain the first 20 keywords, and extract parts of speech.
F. Calculates the overall sentence weight according to the above-mentioned part of speech factor, comment factor and position factor.
G. Sort the weights of sentences.
H. According to the ranking results, the top 10% sentences are taken as abstracts.

3.2. Improvement of Algorithms
Frequency: Although the number of occurrences of a word can not be generalized, to some extent, it reflects the importance of the word.

Part of speech factor: According to the part of speech in Chinese, it can be divided into substantive words and functional words. The notional words include n, v, adj, numerals, quantifiers and pronouns. Functional words include adv, preposition, conjunction, auxiliary, interjection and onomatopoeia. The distribution of keyword parts of speech is mainly noun or noun phrase. The second is verb, numeral,
adverb and other modifiers, which can effectively avoid the shortcomings of traditional linguistic methods.

User comments: There are corresponding user comments below each article. User comments are also the content corresponding to the article keywords. This paper mainly counts the word frequency Term Frequency in the comments, that is, TF value.

4. Experimental Results
Taking the news of https://new.qqq.com/omn/20190331/20190331A02SPQ.html as an example, the article of crawling the web page is Figure 2, The results of this method is Figure 3.

On March 31, I looked back on President's trip to Europe: a visit to "special friends". It's another spring when everything comes back to life. On both sides of Chang's Olympia Avenue, flags of China and France were raised in the wind. Their colors were distinctly different: one is yellow-green, the other is blue-white-red. It is the symbolic color of French and Chinese friendship. It reminds us of the 70th anniversary of the founding of New China and the 55th anniversary of the establishment of diplomatic relations between China and France. President Xi Jinping widely described Sino-French relations more than once: "China and France are special friends and win-win partners. Indeed, the closer a friend gets, the closer a good word gets." I came with a special female companion, who is the first female special envoy of the French government. She is a group of French students singing "Let's Go Two Roses Together" at the Assemblée Nationale in China. The blurred singing drew China and France ten thousands miles apart. "In 1994, after five years of preparation, the socialist state of Vietnam and the Democratic Republic of the East took the big step of radical reform and opening up to the outside world and signed the Paris Agreement, the restoration of diplomatic relations between China and France in an all-round way with extraordinary strategic vision. France became the first Western power to formally establish diplomatic relations with New China. Daring to be the first and keeping pace with the times" has always been an important magic weapon for Sino-French relations to be in the forefront of China relations with Western countries. Xi Jinping summed up this way: Opening up 55 years of real rich historical data, France has created many first contacts with New China: the first big western country to establish a comprehensive strategic partnership and strategic dialogue with China, the first western country to carry out civil nuclear energy cooperation with China, the first western country to sign an intergovernmental agreement on scientific and technological cooperation with China, and the first one to open a straight line train with China. Western countries on airlines... Over the past 55 years, no matter how changeable the international situation is, bilateral relations have been steadily moving forward along the road pioneered by our forefathers.Five years ago, in the presence of President Xi Jinping and French President Hollande, the Chinese people were proud and inspired to see the China-Gallic vaults of navigation in China. This is a more serious milestone, a new chapter in Chinese and French history. Last year, President Macron visited China for the first time after taking office. He presented Xi Jinping with a horse named "Mount Vernon Castle" by the French Government of the Republic, which was called "unprecedented" diplomatic gift by the media, implying that China and France are "positive and upward". On March 25, French President Xi Jinping held talks with French President Macron at the Elysee Palace in Paris. Next, what is the deployment of China-one road pragmatic cooperation? At the March 25 talks, Xi Jinping proposed that the two sides should deepen cooperative relations in nuclear, aerospace, automotive, and medical and health, green development, financial services and other fields has been steadily promoted, and cooperation in emerging areas such as agricultural integration, high-end manufacturing and artificial intelligence has shown encouraging momentum. Xi Jinping put forward the classic slogan "together for China" steady and far-reaching cultural novelty between China and France."Little by little, the birds build their nests." The blossoms are in full bloom at the end of the millennium: the silk-road platform is built at the end of the era." Xi Jinping quoted the above French proverbs and Chinese ancient sayings to describe the friendship between China and France. These two proverbs convey the grand vision of the sages of China and France. Although the historical and cultural backgrounds are different, their connotations are similar. Both China and France have a long history and unique civilization. Xi came with a beautiful feeling for France," Xi said in his signed article. This sentiment comes from the appreciation of the Chinese people and their individual for the brilliant history, culture and charm, from the mutual attraction and reflection of the two civilizations of China and France. A hundred years ago, China set off an upsurge of westward-work program in China. Nearly 2000 progressive young people in China dreamed of saving their country and studying in France. Zhu Enlai, Deng Xiaoping, Cai Heyun, Chen Yi and Nie Rongzhen are among the best. More than a half century later, Bai Xuefeng became one of the first French students to study in China. He was also the first French Geronimo of Chinese. In retrospect of decades of learning Chinese, Bai Xuefeng sighed: "Chinese culture is broad and profound, deeply attractive, and characterized by "the ten most influential Chinese books in France", "The Ancestors of Confucius", "Tao Te Jing", "Wang Mang", "Journey to the West" and "The Song of the Ocean". On March 24, 2019, Mark Long presented Xi Jinping with the French version of the first "Introduction to the Ancestors of Confucius" published in France in 1890. His book, Mark Long presented Xi Jinping with his first French version of the introduction of Confucianism, published in France in 1890, is a national gift, and describes the influence of Confucian Culture: "Confucianism leads deep influenced Voltaire and others, and provided valuable ideological enlightenment for the French Enlightenment Movement." This national gift is very precious. There are only two copies left. The other is kept in the National Museum of Asian Art, Jings, Paris. Open the brown cover, there are light pencil handwriting on the title page. Xi Jinping took the book gently, studied it carefully and solemnly said: "This gift is very precious. I want to return it to the National Library of China for collection." On March 27, 2014, Chinese President Xi Jinping attended the 50th anniversary of the establishment of diplomatic relations between China and France and delivered an important speech. The two great civilized and prosperous countries of the West, China and France are famous in the world. Chinese civilization is colorful because of exchanges and rich and rich because of mutual learning. Xi Jinping said: "Understanding the French culture will enable me to better understand the Chinese culture. The French culture, profound and interesting, has also provided us with a human civilization." "Understanding two civilizations will clear our hearts and minds. France has already set off a wave of learning China. There are 16 Confucius Institutes and 2 Confucius classrooms in the country. In 2018, more than 310,000 people are learning Chinese. The number of Chinese students studying in Japan is expected to reach 15,000 by 2023, which will become the largest group of foreign students in France. In 2021, the two countries will also hold the first "China-French Cultural Tourism Year". I came with ardent expectations of Sino-French relations. Compared to the one-sentence plan for the future development of Sino-French relations recent years, Xi Jinping and Mark Long have met, talked, and communicated many times. China and France have deep political mutual trust and frequent high-level exchanges. Together, they have established a deep co-development mechanism, such as economic and financial relations, they have carried out high-level economic talks and cultural tourism cooperation repeatedly in the French media entitled "Continuing to Walk Together on the Road of Common Development", in which he wrote: "I came with ardent expectations of Sino-French relations." This expectation comes from the deep potential and bright prospects of bilateral relations. Today's world is experiencing unprecedented changes in a century. The international situation is changing. Protectionism and unilateralism are rampant. China and France are both important countries on the international stage. In response to the common challenges facing mankind, Xi Jinping put forward China's plan. He has repeatedly elaborated on the idea of building a community of human destiny, which has many similarities or similarities with the idea of "the same planet" put forward by Mark Long. Last year, Mark Long's visit to China was Xian, the starting point of ancient Silk Road. He has repeatedly expressed his willingness to carry out pragmatic cooperation with China in "one belt and one road" initiative, an important path and grand practice for China to promote the construction of the community of human destiny, has become a broad consensus of the international community. Together, China and France can change the world, which has been proven many times over the past 55 years.On March 25, 2019, Chinese President Xi Jinping held talks with French President Macron at the Elysee Palace in Paris. Next, what is the deployment of China-one road pragmatic cooperation? At the March 25 talks, Xi Jinping proposed that the two sides should deepen cooperative relations in traditional fields of nuclear energy, aerospace and nuclear power generation in emerging areas such as scientific and technological innovation, agriculture, finance and pension services. More support and facilities should be provided for bilateral trade and investment. China will continue to substantially relax market access, optimize business environment, strengthen intellectual property protection, and create a new pattern of high-level opening to the outside world.Just the day before, in the beautiful coastal city of Nice, Xi Jinping and Mark Long held a business dinner. The two sides signed a number of important cooperation agreements, which are a valuable asset. We should inherit and carry forward them well so as to keep Sino-French relations at the forefront of the times." (CCTV Central Radio and Television Network)
4.1. Comparison of Three Algorithms

In this paper, TF-IDF, TextRank and the method are compared to verify the effectiveness of this method. The experimental results are shown in Table 1.

| Evaluating indicator | TF-IDF | TextRank | Article method |
|----------------------|--------|----------|----------------|
| MacAvg_P             | 0.26   | 0.35     | 0.41           |
| MacAvg_R             | 0.526  | 0.56     | 0.62           |
| MacAvg_F1            | 0.34   | 0.44     | 0.55           |

According to Table 1, we can see that this method has improved in MacAvg_P, MacAvg_R and MacAvg_F1, so this method can effectively improve the accuracy of keyword extraction to a certain extent.

5. Conclusion

This paper summarizes a text summary extraction method based on reader's comments. Experiments show that this method is better than traditional TF-IDF and TextRank in MacAvg_P, MacAvg_R and MacAvg_F1. At the same time, the time difference between the improved algorithm and the traditional algorithm is small, so the algorithm is feasible.

Acknowledgement

Thanks to Beijing science and technology innovation service capability construction project (PXM2017_014223_000055) and Beijing Institute of Graphic Communication 2018 R&D Project (Ec201802).

References

[1] Rui Yang. Research and Design of Text Summarization System in the Network[D]. Nanjing University of Aeronautics and Astronautics, 2014.
[2] Xiaodan Xu. Research on Automatic Abstract System of Chinese Web Page[D]. National University of Defense Technology, 2005.
[3] Yang Liu, Lei Cui. Green network the summarization extraction system algorithm [J]. Library and Information Work, 2014, 58(06): 101-104.
[4] Shibo Zhang, Boai Liu. Multi-document Summarization Based on Sentence Score[D]. Journal of Shanxi University (Nat. Sci. Ed.), 2015, 38(03): 406-412.
[5] Yanqing Guo, Rui Zhao, Xiangwei Kong, Haiyan Fu, Jinping Jiang. News-summarization Extraction Method Based on Weighted Event Elements Strategy [J]. Computer Science, 2016, 43(01): 237-241.
[6] Xiaoping Chen. Research and Application of Topic-based Automatic Summarization of Short Text [D]. School of Computer Science & Engineering, 2017.
[7] Nainai Li, Peiyu Liu, Wenfeng Liu, Weitong Liu. Automatic digest optimization algorithm based on TextRank [J]. Application Research of Computers.
[8] Fang Su, Xiaoyu Wang, Zhi Zhang. Review Summarization Generation Based on Attention Mechanism [J]. Journal of Beijing University of Posts and Telecommunications, 2018, 41(03): 7-13.
[9] Yun Wu, Changchun Yang, Jiajun Mei, Huan Gu. Method of automatic summarization algorithm based on word-sentence co-ranking [J]. Computer Engineer and Design, 2018, 39(09): 2776-2779+2810.