Research and Improvement of TextRank Algorithm Adding Degree Adverbs

Chengxia Liu*, Yuheng Che, Ruixue Duan
Computer School, Beijing Information &Technology University, Beijing,100101, China
*Corresponding author’s e-mail: liucx@bistu.edu.cn

Abstract. Textrank algorithm is an approach to extract keywords in articles which will be the premise to find useful information in massive network data. After analysing the characteristics of this algorithm, this paper improves the accuracy of keyword extraction by adding degree adverbs. Based on the function of the original algorithm, a degree adverb table is established, which contains all the degree adverbs that need to be marked. By adding the degree adverbs marker, the results of keyword extraction for the same text will be improved. According to experimental studies, for most articles, adding degree adverbs judgment function increases the importance of key words in these articles. At the end of this paper, the keyword extraction results for different types of articles is demonstrated and compared.

1. Introduction
Keyword extraction methods mainly include supervised keyword extraction and unsupervised keyword extraction. Among them, supervised keyword extraction algorithms mainly include Kea [1] (keyword extraction algorithm), GeneX [2], and so on. These algorithms all need manually annotated corpus. However, in most cases, the text information that users need to extract is not manually annotated in advance. For example, the media content, such as news on the Internet or microblog, is not manually annotated. The unsupervised text keyword extraction method does not need manual annotation. It is mainly divided into three categories. The first category of keyword extraction algorithm is based on statistical features, such as TF-IDF (term frequency – inverse document frequency) algorithm [3]. The second category is based on topic model, such as LDA (latent Dirichlet allocation) algorithm [4]. The third category is based on word graph model, such as TextRank algorithm [5]. This paper mainly studies the implementation and improvement of TextRank algorithm.

2. Review of TextRank algorithm
TextRank algorithm is improved from PageRank algorithm [6]. And many improvements have been made to TextRank algorithm[7-8]. This paper explores the influence and improvement of TextRank algorithm after adding degree adverbs on keyword extraction. Let's begin with the PageRank algorithm.

2.1. Principle of PageRank algorithm
PageRank algorithm is a web page sorting algorithm proposed by Lawrence page and Sergey Brin. Based on the analysis of web page link structure, it represents the whole network structure relationship as the style of Web graph. The algorithm regards the link of web page V to web page U as a vote of V to U, so Web page V needs to assign part of its PageRank value to web page U. Therefore, the higher...
the authority value of web page $V$, the higher the value of web page $U$ can get. At the same time, the more times web page $U$ is linked, the higher its PageRank value is. Finally, PageRank algorithm sorts the search results according to the PageRank value of each web page.

PageRank algorithm is based on the random web surfing model, assuming that users browse the web content randomly on the web. And all the pages on the web can be accessed, then users can have the following two choices:

a) Click the previous link on the current page.

b) By typing the web address, you can jump to a web page randomly for access.

Suppose that the probability of users clicking on a web link is $d$, then the probability of direct jump is $1-d$. The variable $d$ here is also called damping coefficient. According to this principle, the calculation formula of PageRank value of a web page is shown in Formula (1).

$$PR(U) = (1 - d) + d \left( \frac{PR(V_1)}{Out(V_1)} + \cdots + \frac{PR(V_n)}{Out(V_n)} \right)$$

(1)

Among them, $V_i$ is the web page linking to $u$. $\{V_1, V_2, \cdots, V_n\}$ is the collection of all web pages linking to $u$. $Out(V_i)$ is the number of links which link out from web page $V_i$.

2.2. Principle of TextRank algorithm

The textRank keyword extraction algorithm is very similar to PageRank algorithm in basic principle. PageRank algorithm calculates the importance of each network page according to the link relationship between network pages on the Internet, while TextRank algorithm regards words as "nodes on the World Wide Web". Combining with text analysis of this application scenario, the "web page" is changed to "words". And the importance of each word is calculated according to the co-occurrence relationship between words. The core formula of TextRank algorithm is shown in formula (2).

$$WS(V_j) = (1 - d) + d \times \frac{\sum_{V_i \in Out(V_j)} \omega_{ji} \cdot WS(V_i)}{\sum_{V_k \in Out(V_j)} \omega_{jk}}$$

(2)

Among them, $\omega_{ji}$ is used to indicate that importance of the edges connected between two nodes is different.

TextRank algorithm is an algorithm to determine the importance of nodes in a graph based on all the information extracted from the whole graph. Its basic idea is "vote" or "recommend". When a node in a graph wants to connect to another node, it can be regarded as voting for another node. The higher the number of votes cast for a node, the higher the importance of the node.

3. Improvement of TextRank algorithm

For a Chinese article, if you want to judge which words are key words, in addition to "vote", you can judge whether there are some special words in a sentence, such as "extra(倍加)", "very(非常)", "too(过于)" and other adverbs representing degree. Here let’s make a hypothesis: if there are degree-adverbs in a sentence of an article, the higher the degree, the other words in the sentence will more likely become the keywords.

3.1. Acquisition of degree adverbs

In modern Chinese, such words as "extra" and "absolute" are defined as adverbs of degree. Compared with other types of Chinese adverbs, degree adverbs have stronger positioning. Degree adverbs can be divided into relative degree adverbs and absolute degree adverbs [9], and the specific categories are shown in Table 1.
Table 1. Classification of degree adverbs.

| Classification | Degree adverbs                                      |
|----------------|-----------------------------------------------------|
| **Very high**  |                                                    |
| **High**       | more (更、更加) even more (更为、更其、越发) the more(越) more and more(越加) Extra(倍加) further(愈、愈加、愈发) increasingly(愈为、愈益) especially(格外) all the more(益发) |
| **Medium**     |                                                   |
| **Low**        | Slightly(稍、稍微) a little(稍微、稍微) a bit(稍许) briefly(略、略微) somewhat(略微、略微) something of (一些、多少) |
| **Very high**  | Extremely(极、极至为) exceedingly(极其、极度 to the utmost(极端) utmost(至、至为) top(顶) excessively(过、过分) too(过于) particularly(分外) very much(万分) absolute(绝对) |
| **High**       | very(很、挺、怪、老、甚、甚为、大、大为) extraordinarily(非常、何等、何其) especially(特别、相当、尤为) well(好、好不) quite(颇、颇为、十分、满、蛮) rather(异常、深为、不胜) enough(够) more(多、多么) specially(特、尤其) Incomparable(无比) |
| **Medium**     | not very(不大、不太、不很、不甚) |
| **Low**        | Seldom(有点、有些) |

These degree adverbs are stored in the dictionary. And the degree adverb dictionary is used to judge whether there are degree adverbs in the sentence.

3.2. Improvement of TextRank algorithm after adding degree adverbs
Before the text segmentation, read the storage dictionary to obtain all the Chinese degree adverbs. Special marks are made for the words in the sentences with degree adverbs. After segmentation and word filter, judge whether there are degree adverbs. If there is, all the other words belonging to the same sentence are added with special marks. When the connecting edge of graph is constructed with these words with special marks, the weight of corresponding edge will be increased. Because the statements with high degree adverbs are often the statements emphasized by the author, the weight should be heavier. For example, the weight of the connected edge with special markers is set to 5.0, while the weight of the other connecting edge is still 1.0. The keywords will be more accurate in this case.

Therefore there are three main changes in the program of TextRank algorithm. First, establish a special dictionary of adverbs of degree. Second, the method of reading words in special dictionaries and judging whether there are words in special dictionaries should be established. Third, special word markers are added in keyword extraction and ranking, and high weight values are set for the word nodes of special markers.

4. The Implementation of Textrank algorithm and its improved
TextRank algorithm is to select words with high importance from a large number of filtered words. The basis of judging the importance of words is to set TextRank score for each word. After several
rounds of iteration of TextRank algorithm, the higher the score, the higher the importance of words in the text. At the same time, it is more qualified to be the keywords of the article.

In the initial stage of calculating the TextRank value of each node, the initial textrank value of each node in the graph is considered. Here the initial TextRank value of each node can be set to \(1/n\), where \(n\) is the number of nodes in the whole graph. The TextRank value is calculated according to formula (2).

4.1 Implementation of textrank core algorithm
After realizing the TextRank value calculation formula, the graph will be build and TextRank values of all words need be sorted.

4.1.1 Constructing the edge of undirected graph – Combine()
By combining words, the edge of undirected graph is constructed. The process is shown in Figure 1.

![Flow chart of Combine() function](image1)

**Figure 1.** Flow chart of Combine() function

In the TextRank algorithm, if every word in the same window is regarded as a node in the undirected graph, these words in the window can connect an edge in the construction process of the graph.

4.1.2 Sort the TextRank value of words - SortWords()
The damping factor is an indispensable variable in the subsequent calculation of TextRank value of each node, which is usually assigned to 0.85. The words sort process is shown in Figure 2. Firstly, the first part realizes the construction of undirected graph. The second part iteratively calculates the final score of TextRank of each node through formula 2. And the third part receives the calculated score of TextRank of each node and sorts their values in reverse order.

![Flow chart of SortWords() function](image2)

**Figure 2.** Flow chart of SortWords() function

4.2 The Implementation of TextRank improvement algorithm after adding degree adverbs
After adding degree adverbs, the SpecialWords variable is created, which reads and stores degree
adverbs. In addition, a new function, JudgeSpecialWords is added which is responsible for judging whether there are degree adverbs in a sentence. If there are, all strings in the sentence are marked with special marks. After text segmentation and before stop words text filtering, you need to call judge_SpecialWords function to get the result of degree adverb annotation. If a word is a degree adverb, all words in the same sentence are marked as true, while other unmarked words are marked as false. The keyword sort algorithm, SortWords, which constructs the connection edge of the node in a circular way, need to add the judgment of the word tag. If the word tag is true, the weight of the connecting edge constructed by them is 5.0, otherwise it is 1.0.

5. Comparison test before and after algorithm improvement

Before and after adding the degree adverbs, the test program tested 75 articles, including different types of articles, such as news and journal papers, with the article number of N01-N75.

5.1. News’ keyword extraction test

During the test, a brief news from a regular press conference held by the Ministry of Foreign Affairs on May 14, 2021 was intercepted. The comparison results of keyword extraction are shown in Figure 3.

It can be seen that the degree adverbs make the results of extraction keywords changed. Before adding degree adverbs, abstract verbs like "express(表示)" rank in the top ten of the keywords. After adding the function of degree adverbs, the word "international(国际)" has become a new word in the top ten keywords. The scores of the other nine words have some changed too.

5.2. Journal paper’s keyword extraction test

In the test, the journal article "Discussion of the hierarchical deployment and development of edge computing" [10] was used to test. It mainly introduces the relationship between edge computing and cloud computing, the intelligent hierarchical deployment model of edge computing and the development of edge computing deployment. The keyword extraction results are shown in Figure 4.

It can be seen that after adding the degree adverbs, the key words degree of "application(应用)" and "intelligence(智能)" has been significantly improved, while the key word degree of "data(数据)" has slightly decreased. The degree of other key words has no obvious change.
5.3. Test summary
Among the 75 articles, 10 of them had no difference in keyword extraction results, accounting for 13% of the total. There is a little difference in the scores of the top keywords in 35 articles, which the top keywords is same but the degree is different, accounting for 47% of the total. However, there are obvious differences in the keyword extraction results of the other 30 articles, which the new keywords can better reflect the theme of the articles, accounting for 40% of the total.

6. Summary and prospect.
In order to find useful information in the complex and huge network information, the key information in the text need to be extracted first. TextRank, the keyword extraction algorithm, is researched in this paper. After adding the degree adverbs in TextRank algorithm, the keyword extraction is more in line with the author's intention and the reader's understanding, because the degree adverbs strengthen the meaning to be expressed in the text. In the future, further study can be made by dividing degree adverbs according to the degree of severity. In addition, other adverbs can be added and other factors that affect the semantic can be used to extract keywords more accurately.

References
[1] Eibe Frank, Gordon W. Paynter, Ian H. Witten, Carl Gutwin, and Craig G. Nevill-Manning. (1999) Domain-specific key phrase extraction. In Proceedings of the 16th International Joint Conference on Artificial Intelligence. pp. 668–673.
[2] Peter D Turney. (2000) Learning algorithms for keyphrase extraction. Information Retrieval 2(4):303–336.
[3] Ken Barker and Nadia Cornacchia. (2000) Using noun phrase heads to extract document key phrases. In Advances in Artificial Intelligence. pp. 40–52.
[4] Blei D M, Ng A Y, Jordan M I. (2003) Latent dirichlet allocation[J]. Machine Learning Research Archive, 3: 993-1022.
[5] Rada Mihalcea and Paul Tarau. (2004) Textrank: Bringing order into text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. pp. 404–411.
[6] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. (1998) The pagerank citation ranking: bringing order to the web. Technical report, Standford Digital Library Technologies Project.
[7] Lin Huang, Guo Shuhui. (2003) The characteristics, scope and classification of degree adverbs. Journal of Shanxi University (Philosophy & Social Science), 2:71-74
[8] Florescu C, Caragea C. (2017) PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents[C] //Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. pp. 1105-1115.
[9] Zhou Qingyun, Fang Yuansheng, Shang Zhenlei, Zhong Wanli. (2020). Keyword Extraction Method for Complex Nodes Based on TextRank Algorithm. 2020 International Conference on Computer Engineering and Application (ICCEA).
[10] Xiao Xia, Ruan Yiming. (2018) Discussion of the hierarchical deployment and development of edge computing. China Science and Technology Achievements. 2:59-60,64