Research on the Vocabulary Relevancy Algorithm of the Improved Search Engine

Taoyu Wang¹,a, Xianjun Shi¹, Qijie Chen¹

¹Control Science and Engineering, Naval Aviation University, Yantai, 264001, China
a1000479524@smail.shnu.edu.cn
751997574@qq.com

Abstract: Aiming at the limitation of the existing search engine search algorithm and the low reliability of the derived vocabulary relevancy algorithm, a vocabulary relevancy algorithm based on search engine and vocabulary relevancy algorithm is proposed under the background of making full use of the knowledge base of Hownet. Firstly, the main defects of search engine algorithm are considered in two aspects, meaningless text and redundancy. Then, the two problems are solved by using "Hownet" and DSC algorithm. Finally, the search engine-based vocabulary relevancy algorithm is improved by integrating multiple factors such as weight reduction and noise reduction. Experimental results show that compared with the pure improvement method based on the search engine, the method of spearman coefficient and Pearson coefficient were improved, as well as reduce meaningless language segment on the calculation of correlation, verified the "Hownet", and DSC algorithm into the search engine algorithm can effectively improve the computing performance of relevant vocabulary.

1. Introduction

The quantitative calculation methods of lexical similarity can be generally divided into two categories [1], those with and without background. The method with background is divided into two methods, rule-based and statistics-based. The rule-based approach[2] is based on the fact that two terms have a certain semantic relevance if and only if they have a path in the structural hierarchical network diagram between concepts. This method often requires the help of a large semantic dictionary with relatively complete domain knowledge to calculate the word similarity according to the superordinate and isotopic relations between concepts. The statisms-based approach [3] requires two terms to have some degree of semantic similarity if and only if they appear in the same context on the basis of the assumption. This method [4] mostly relies on large-scale corpus training to judge whether two lexical contexts have similar related word sets. For those without background, the method based on word literal matching is generally adopted.

However, the existing single lexical association degree calculation methods have their own limitations. Guo Li et al. [5] showed that when the data was sparse, the correlation degree calculated by the statistical method was not ideal. Zhu Zhengyu et al. [6] improved the semantic similarity of words based on Hownet, but relied too much on the word system constructed by Hownet. Gao Guoqiang et al. [7] used Internet search engines to calculate the semantic similarity of Chinese word pairs, which did not need to build a huge dictionary base but could easily be misled by wrong information on the Internet.

Based on the relevancy algorithm of search engine and Knowledge network, with the improved
search engine algorithm as the core, the knowledge construction of Knowledge Network is screened and corrected. Considering that the updating speed of the word system constructed by the knowledge network is slow, but the basic relation is more correct. When a search engine is dealing with the relevancy of new words, its basic categories are determined, and some disturbing and meaningless corpus can be screened through the word system of Hownet.

2. Material and Methods

2.1 Calculation of Vocabulary Relevancy Based on the Number of Query Pages

The number of page pages [8] refers to the number of pages containing the searched words $p$. For example, there are 100 million results for "Arabia" on Baidu. In the rest of this article, symbols $N(p)$ will be used to represent the number of pages returned by Baidu queries $p$. Of course, the terms $p$ and $q$ the individual query pages are not enough to get semantic similarity between them; you should also add the number of query pages that appear together, i.e., the number of query pages for "p AND q". For example, when you search for "Arabic" and "number" on Baidu, you can find 23,500,000 results, i.e. $N(\text{Arab} \cap \text{number}) = 235000000$.

The heart of the calculation of relevancy based on the number of query pages is "You shall know a word by the company it keeps". In this paper, Jaccard method is used as the basic theory to calculate vocabulary relevancy. The specific formula (1) is as follows.

$$Jaccard(p,q) = \frac{N(p \cap q)}{N(p) + N(q) - N(p \cap q)}$$

(1)

Where, $N$ is the number of pages in the search engine. The baidu search term limit is $1 \times 10^9$, so in the article is $N = 1 \times 10^9$.

This algorithm has some disadvantages:

(1) The noise existing in the network data is ignored. Since most of the text on the network is unaudited and calibrated, there will be a large number of meaningless statements. Therefore, the illogical occurrence of two words should be reduced to improve the accuracy of semantic relevancy calculation.

(2) The redundancy in network data is ignored. As a result of the search engine itself algorithm loopholes, web pages will appear a lot of repetition. Therefore, the number of pages of queries returned by search engines is inaccurate, and a large number of duplicate pages should not be counted to improve the accuracy of vocabulary relevancy calculation.

2.2 Use HowNet to Remove Noise

Words are associated with each other to a relatively low degree in noise, which is as well as the meaningless paragraphs. By using the knowledge system of "Hownet", the meaningless fields in the return page of search engine can be screened out, so as to improve the accuracy of word correlation degree calculation.

As a dictionary base, every word has its corresponding thesaurus. The calculation formula (2) of thesaurus similarity [9] is as follows:

$$\text{simS}(a,b) = \frac{1}{2} \times \frac{\alpha}{\alpha + \sum_{i=1}^{n} \text{weight}(level(i))} + \frac{1}{2} \times \frac{2 \times \log f(LCN)}{\log f(a) + \log f(b)}$$

(2)

Where, $\alpha$ represents the semantic distance value when the similarity is 0.5; weight($k$) is the weight function of each edge; level($i$) is the level of each edge on the reachable path in the semantic hierarchy tree; LCN represents the minimum common parent node in the tree; $f(a)$ is the ratio of the number of nodes connected to the total number of nodes in the tree, which is used to reflect the density
information in the tree where the semantic source $a$ is located.

Lexical similarity is based on thesaurus similarity, and the similarity of each thesaurus will have an impact on the overall lexical similarity. The weighted bipartite graph can be used to obtain the calculation formula (3) of vocabulary similarity as follows:

$$\text{Same}(p, q) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \text{SimS}(p_i, q_j)}{mn}$$

(3)

When removing the noise, the four words before and after the keywords in the paragraph returned by the search engine are selected to calculate the lexical similarity, and the noise threshold is set. The noise removal formula (4, 5) is as follows:

$$D_{no}(p) = \frac{\sum_{i=1}^{L} \varepsilon \left(\sum_{j=1}^{4} \text{Same}(p, p_j) / 4 - no\right)}{L}$$

(4)

$$D_{no}(p \cap q) = \frac{\sum_{i=1}^{L} \varepsilon \left(\sum_{j=1}^{4} \text{Same}(p, q_j) / 4 - no + \sum_{j=4}^{8} \text{Same}(q, p_j) / 4 - no\right)}{L}$$

(5)

Where, $L$ is the noise sample number, in this paper $L = 10000$.

2.3 DSC algorithm and Bronc filter

There are four reasons for duplicate pages, mirror pages, duplicate urls, different formats, plagiarism and reprinting of article content. Among them, duplicate URL and duplicate content are the two most important aspects of duplicate content. Considering that duplicate content contains duplicate urls, in order to reduce the workload of deduplication, the page returned by the search engine should be URL deduplication processed first.

This paper uses counting bloom filter [10] to process URL. If the accuracy of the filter is $\gamma$ and the number of duplicate URL pages removed is $URL(p)$, then the actual number of URL pages removed is $\gamma URL(p)$.

The improved DSC algorithm [11] is adopted to remove duplicate content, and the formula (6) is as follows:

$$DSC(p) = \sum_{i=1}^{m} \sum_{j=1}^{m} \varepsilon \left(\frac{\sum_{k=1}^{n} d_i(k)d_j(k)}{\sqrt{\sum_{k=1}^{n} d_i(k)^2 + \sum_{k=1}^{n} d_j(k)^2}} - dsc\right)$$

(6)

Where, $dsc$ is the threshold of repetition rate. If the threshold exceeds the threshold, the repeated content shall be deemed to be deleted.

In order to improve the efficiency of redundancy removal, only the first 10,000 search pages are searched and filtered. Considering the page recommendation mechanism of Baidu, the possibility of page redundancy in the top ranking is high, so the redundancy coefficient of the design $Rd$ is adjusted for the overall redundancy. The calculation formula (7) of the total redundancy is as follows:

$$RY(p) = \frac{N(p)Rd(DSC(p) + \gamma URL(p))}{M}$$

(7)

3. Results

This section improves the search engine algorithm by integrating noise, redundancy, overall search page volume and other factors. The main improvements are as follows:
(1) Although the total number of search pages is fixed in this paper, the impact of the total number of search pages on the overall relevancy still needs to be taken into account by the algorithm. The impact of total search pages is designed with the following considerations in mind:

1. The greater the total number of searches, the higher the accuracy of relevance. In general, the higher the correlation.
2. The change of total search volume has a relatively small influence on relevancy compared with other factors.

To sum up, \( e^{-\frac{1}{N}} \) is selected as the influence factor of the total amount of search. \( e^{-\frac{1}{N}} \) is a singly increasing function with respect to \( N \), and its range is \([0,1]\), when \( N \) tends to positive infinity, the range of change is very small, which accords well with the influencing factors on the total amount of search.

(2) The influence of noise in the search return page needs to be removed from the calculation formula (8).

\[
N'(p) = N(p)(1 - Dno(p)) \\
N'(q) = N(q)(1 - Dno(q)) \\
N'(p \cap q) = N(p \cap q)(1 - Dno(p \cap q))
\]  

(8)

(3) The influence of redundancy in the search return page needs to be removed from the calculation formula (9).

\[
N'(p) = N'(p)(1 - RY(p)) \\
N'(q) = N'(q)(1 - RY(q)) \\
N'(p \cap q) = N'(p \cap q)(1 - RY(p \cap q))
\]  

(9)

(4) Add confidence to improve the formula (10).

\[
con(p \cap q) = \frac{N'(p \cap q)}{N'(q)}
\]  

(10)

As an index of the credibility of the correlation algorithm, the correlation algorithm can only play a role in setting the confidence threshold when the confidence is greater than the threshold. Set confidence thresholds \( \theta \).

To sum up, the improved formula (11) of term relevancy based on search engine is as follows:

\[
WC(p,q) = \begin{cases} 
0 & \text{if } con(p \cap q) < \theta \\
\frac{N'(p \cap q)}{N'(p) + N'(q) - N'(p \cap q)} e^{-\frac{1}{N}} & \text{if } con(p \cap q) \geq \theta
\end{cases}
\]  

(11)

4. Discussion

In this experiment, Chinese vocabulary similarity algorithm [8] and network search engine algorithm based on search engine were adopted, namely, reference 7 and Reference 8. To verify the validity of the method, 540 pieces of data provided by NLPCC meeting were selected for the experiment. The standard semantic similarity between the sample data and the experimental data was manually annotated and averaged by 20 participants, and the reliability was relatively high. 270 pieces were randomly selected as the weight training data, and the remaining 270 pieces were taken as the test data. Where selected three kinds of similarity algorithm, an improved algorithm of search engine is literature seven, eight of use of the algorithm and the algorithm of search engines, respectively 270 test data for lexical semantic similarity computation, and then use the Spearman coefficient and the coefficient of Pearson evaluate its accuracy, the results such as table 1, the results can be seen that the method is better than other algorithms.
Table 1. Correlation coefficient

| Algorithm          | Spearman | Pearson |
|--------------------|----------|---------|
| Reference[7]       | 0.327    | 0.361   |
| Reference[8]       | 0.349    | 0.385   |
| My algorithm       | 0.392    | 0.427   |

5. Conclusion
This paper fully considers the advantages and disadvantages of the return page algorithm of the search engine in order to propose an algorithm to modify the return page of search engine by using knowledge network and DSC algorithm. Besides, this paper puts forward a method to eliminate the noise of search engine by using Hownet and reduce the weight of the search engine return page by DSC algorithm, so the existing search engine return algorithm is improved. From the experimental data, the improved algorithm proposed in this paper is more accurate and more accurate than the algorithm of Hownet and search engine. Next, the algorithm of noise reduction and weight reduction will be further studied, so as to further improve the calculation rationality of lexical semantic correlation.

References
[1] Zhang Junjie, Gu Chongyang, Xu Haoyu. (2018) Text Similarity Calculation based on Lexical Semantic Information. J. Computer Application Research, 35(02): 391-395.
[2] Zhong Yiping, Li Shaoqing, Wu Chengrong. (2015) Calculation of Lexical String Similarity of Bad Text Variant Keyword Recognition. J. Computer Applications and Software, 32(03): 151-157.
[3] Wang Yongsheng. (2015) Algorithm of Similar Words for Automatic Acquisition of Different Meanings. J. Computer Application and Software, 32(03): 258-260.
[4] Chen Huaixin, Huang Hao. (2017) Lexical Semantic Similarity Calculation based on Knowledge Network in the Field of Transportation. J. Application of Computer Systems, 26(03): 169-174.
[5] Liu Lei, Guoli. (2013) Lexical Semantic Similarity Algorithm Research and Application. J. Software Guide, 12(07): 57-58.
[6] Sun Junhua, Zhu Zhengyu. (2013) Improved Lexical Semantic Similarity Calculation based on Zhizhi.com. J. Computer Applications, 33(08): 2276-2279.
[7] Chen Fengyu, Gao Guoqiang, Huang Lvwei. (2014) Web search engines are used to calculate semantic similarity of Chinese words. J. Computer Technology and Development, 24(07): 84-87.
[8] Chen H Y. (2015) Semantic Similarity computing based on Search Engine. J. Computer Science, 42(01): 261-267.
[9] Tang Qiguan, Ge Bin, Li Fangfang. (2010) Research on Lexical Semantic Similarity Calculation Method based on CnKI. J. Computer Application Research, 27(09): 3329-3333.
[10] Huang Tao. (2013) Research and Application of Bloom Filter in Web Page de-weighting. D. Dalian Maritime University.
[11] Yan Fulliang. (2012) Research on duplicate Page Detection Algorithm in Search Engines. D. Henan University of Technology.