A Page-topic Relevance Algorithm Based on BM25 and Paragraph-Semantic Correlation

Shi Peng*, Xiaodan Xie, Jia Zhai, Yusheng Jia, Yuxuan Gong

Science and Technology on Optical Radiation Laboratory, Beijing 100854, China

*pengshi1828@163.com

Abstract. It is difficult for the traditional topic relevance algorithm based on word frequency and probability statistics model to deal with the ambiguous topic of search keywords, which leads to the retrieval results containing much information that users are not interested in. To solve this problem, a page-topic relevance algorithm based on BM25 and paragraph-semantic correlation is proposed in this paper. The semantic correlation of one page is calculated by the paragraph-semantic classification using a pre-trained deep neural network model, and is weighted with the BM25 retrieval score. Experimental results show that compared with the traditional BM25 algorithm, this algorithm can effectively improve the retrieval accuracy.

Keywords: Topic relevance, BM25, Semantic correlation, Information retrieval.

1. Introduction

With the rapid development of computers and communication technology, the Internet has become the largest carrier of information release, dissemination and circulation. Search engines, providing users with the function of fast information retrieval by searching keywords, have become an important tool for people to browse massive information on the Internet. However, search keywords are generally short and boast various meanings, which leads to many irrelevant contents in the search results that do not meet the users’ search intentions. Taking the keyword “Harry Potter” as an example, the content involved may be JK Rowling’s series of books, or a series of movies adapted, or the report of Daniel Radcliffe, who plays the role. This ambiguous topic can seriously affect the efficiency and accuracy of information retrieval.

The topic relevance algorithm of web pages or documents determines whether search engines can effectively find the content that users are interested in, thus it is a research hotspot in the field of information retrieval. One of the most typical algorithms for ranking the web topic relevance is BM25[1]. It mainly uses word frequency to establish the retrieval probability model between search keywords and page content and calculate the correlation between them. BM25 has been widely used in such fields as web page ranking and natural language processing. Different methods have been employed in many previous studies to improve the traditional BM25 algorithm. Robertson et al. proposed a method of weighing term frequencies before applying the frequency saturation function[2]. Svore et al. proposed a method to improve BM25 retrieval by using LambdaRank and machine learning[3]. Blanco et al. proposed a method to extending BM25 by defining the notion of virtual region and introducing multiple query operators[4]. These methods have improved the results of
traditional BM25 retrieval to a certain extent, but still failed to effectively solve the problem of ambiguous topics.

A page-topic relevance algorithm based on BM25 and paragraph-semantic correlation is proposed in this paper. Deep neural network is used to classify the topic of each paragraph in the web page and calculate the semantic relevance score, which will be integrated with the BM25 retrieval score to get the topic relevance score of the web page.

2. Page-topic relevance algorithm

The page-topic relevance algorithm proposed in this paper mainly includes three steps: first, calculate the relevance score between search keywords and page content (BMscore) by BM25; Then, extract the characteristics and classify the topics of the semantics of each paragraph in the page through the pre-trained BERT model, and calculate the semantic correlation score (SC Score) of the page content through the classification results of all paragraphs; Finally, linearly weight the retrieval score and semantic relevance score of BM25 by the iterative training weight function to obtain the Page-Topic Relevance Score (PTR Score).

2.1 BM25 algorithm

BM25(Best Match) is a very typical probabilistic and statistical retrieval model in the field of information retrieval. The algorithm firstly divides the retrieval sentence(Q) into several morphemes(qi), then calculates the correlation score between morphemes(qi) and documents(d), and finally calculates the topic relevance between Q and d by weighted summation of each morpheme score. The general formula is as follows:

\[ BMscore(Q,d) = \sum_{i} W_i \cdot R(q_i,d) \]  

Where, \( W_i \) is the weight of morpheme \( q_i \), \( R(q_i,d) \) is the correlation score between morpheme \( q_i \) and document \( d \), and the formula is as follows:

\[ W_i = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \]

\[ R(q_i,d) = \frac{f_i \cdot (k_1 + 1)}{f_i + K} \cdot \frac{qf_i \cdot (k_2 + 1)}{qf_i + k_2} \]  

Where, \( N \) is the number of all documents in the index, and \( n(q_i) \) is the number of all documents containing the keyword \( q_i \); \( k_1, k_2 \) and \( K \) are adjustment factors, which are usually set according to experience. \( f_i \) is the frequency of \( q_i \) in \( d \), and \( qf_i \) is the frequency of \( q_i \) in \( Q \). Considering that \( W_i \) in formula (2) may have negative values in practical application, a slight adjustment is made in this paper to ensure that it will not have negative values, and the specific formula is as follows:

\[ W_i = \log \frac{N + 1}{n(q_i) + 0.5} \]  

2.2 Semantic correlation algorithm based on paragraph-semantic classification

BERT is the most popular deep neural network model in the field of natural language processing.\(^5\)\(^6\)\(^7\) It mainly uses the encoder of Transformer as the basic model and uses multi-dimensional attention mechanism to build the model. In this paper, the semantics of each paragraph in the page are classified
by the pre-trained BERT model, and the semantic correlation score of the whole page and the topic of interest to users is calculated according to the classification results. The specific formula is as follows:

\[
SCScore = \frac{\sum_{i=1}^{n} m_i^2}{\sqrt{n}}, \quad p
\]  

Where, \( m_i \) is the probability that paragraph semantics match the correct topic, \( n \) represents the number of main paragraphs contained in the number \( d \), and \( p \) is the adjustment factor, which is used to adjust SCScore and BMScore to the same order of magnitude.

2.3 Page-Topic Relevance algorithm

The retrieval score of BM25 represents the correlation between search keywords and page content in probability and statistics. Semantic correlation score represents the semantic correlation between page content and topics of interest. The Page-Topic Relevance algorithm proposed in this paper comprehensively evaluates the relevance between web pages and the topics that users are interested in by linearly weighting them. The specific formula is as follows:

\[
PTRScore = \alpha \cdot BMScore + \beta \cdot SC \]  

The main parameters \( \alpha, \beta \) and \( p \) are optimized by iterative correction.

3. Experimental results and analysis

3.1 Experimental data

The automobile brand “Ferrari” is taken as the experimental object in this paper. The web crawler based on the Webcollector crawler framework crawls 2257 web pages related to the search keyword “Ferrari” from the search results of Yahoo. According to the data analysis results, all the data are divided into five topics: company, motorcade, vehicle type, dealer and other information, among which the information about company is designated as the topic of interest to users in this experiment. All the data have undergone unified preprocessing steps, including data cleaning, paragraph division, manual labeling etc.

3.2 Experimental environment

The main hardware environment of the test is as follows: CPU: Intel Core I7, GPU: GeForce RTX 2070; the main software environment is as follows: System: Ubuntu 16, Deep learning framework: Tensorflow1.12.

3.3 Models and parameters

In this experiment, the main parameters \( \alpha, \beta \) and \( p \) of Page-topic relevance algorithm are set as \([0.3, 0.7, 0.01]\) respectively. Paragraph-semantic classification model based on BERT is trained by a very small number of samples, and the main parameters are set as follows: the longest sequence dimension of sentences =128, Batch size=8, Epoch=20 and learning rate =0.001. Since the actual data volume of each topic is different, the number of training samples in each category is slightly different, with an average of 400 samples. By contrast, the traditional BM25 algorithm also calculates the topic relevance under the same experimental conditions. According to the subject specified in the experiment, the search keywords are set as “Ferrari” and “company”.

3.4 Evaluation method

In practical application, the ranking of topic relevance scores represents the web page retrieval results of corresponding algorithms. Therefore, whether the content of web pages with higher scores accords with the topics that users are interested in becomes an important basis for measuring the topic
relevance of the algorithm. Therefore, in this paper, the proportion of pages judged as topic-related by users in the top $K$ web pages is calculated, and the specific formula is as follows:

$$\text{Accuracy}(K) = \frac{K_c}{K}$$

Where, $K$ is the Top $K$ results in the selected ranking, and $K_c$ is the number of web pages related to the topic.

3.5 Experimental results
The first 50 pages selected by traditional BM25 and the method of this paper were manually evaluated by 5 users. The ranking accuracy of topic relevance of the two methods was counted from top50, top20 and top10 levels. The specific results are shown in Tab.1.

|            | BM25     | Method of this paper |
|------------|----------|----------------------|
| Accuracy(10) | 40%     | 50%                  |
| Accuracy(20) | 45%     | 60%                  |
| Accuracy(50) | 30%     | 62%                  |

Tab.1 Comparison table before and after geometric vertex optimization.

It can be seen from Tab.1 that the accuracy of the method of this paper is improved by 10%, 15% and 32% compared with the traditional BM25 algorithm in the top10, top20 and top50 levels respectively, which indicates that the accuracy of topic relevance algorithm can be effectively improved by fusing the semantic feature information of paragraphs and the probability statistics information of word frequency.

4. Conclusion
To solve the problem of topic ambiguity in traditional topic relevance algorithm based on probability and statistics model, this paper proposes a page-topic relevance algorithm based on BM25 and paragraph-semantic correlation. Through the pre-trained BERT model, the feature extraction and topic classification of paragraph paragraphs are carried out, and the calculated semantic correlation is combined with BM25 retrieval score. Experimental results show that this method can improve the ranking accuracy of the top 50 page topics by 32%. At the same time, the model training only needs a small number of samples of about 400 on average, which shows that this method has a great potential for practical application. Next, the research will mainly focus on the methods for the fast optimization of algorithm parameters.

References
[1] Robertson S and Zaragoza H 2009 The Probabilistic Relevance Framework: BM25 and Beyond J. Foundations & Trends in Information Retrieval 3(4) pp333-389
[2] Robertson S E, Zaragoza H and Taylor M J 2004 C Simple BM25 extension to multiple weighted fields Proc. of the 2004 ACM CIKM Int. Conf. on Information and Knowledge Management November 2004 Washington D C USA
[3] Svore K M, Burges C J C 2009 C A machine learning approach for improved BM25 retrieval Proc. of the 18th ACM CIKM Int. Conf. on Information and Knowledge Management
[4] Blanco R, Boldi P 2012 C Extending BM25 with multiple query operators Proc. of the 35th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval pp921-930
[5] Devlin J, Chang M W, Lee K and Toutanova K 2018 J Bert: Pre-training of deep bidirectional transformers for language understanding Preprint arXiv 1810.04805
[6] Bahdanau D, Cho K and Bengio Y 2014 J Neural machine translation by jointly learning to align and translate Preprint arXiv 1409.0473