Research on Query Term Expansion based on RankSVM and LDA Model

Lin Peng, Guo-quan Lu

1College of Big Data Yunnan Agricultural University Kunming 650201
2College of Big Data Yunnan Agricultural University Kunming 650201
apenglin2286351@163.com, b769321019@qq.com

Abstract—The lack of semantic association will lead to poor performance of retrieval system. This paper proposed a query terms extension method based on RankSVM and LDA model. Firstly, with the aid of RankSVM’s excellent performance of sorting, high-quality initial retrieved results will be obtained. Then utilizing the LDA model to select the documents related to query term and to generate topic model for sequenced initial retrieval results. In the post-processing step, using threshold to select the highest probability among each topic and regards them as query terms set. Experiments show that the query terms expansion method proposed in this paper is superior to the current query expansion methods based on pseudo-relevance feedback in the five evaluation index, therefore this is a feasible method.

1. INTRODUCTION

With the development of information technology, information is exponentially increasing. The traditional retrieval systems have some problems, such as mismatching between query terms and retrieval results and lack of semantic function, which leads to low recall rate and accuracy. Query expansion technology is a technology that makes query terms have semantic association function and since it was proposed in the 1960s, the recall rate and accuracy have been remarkably improved[1]. One of query expansion technology exercises is as follows: concepts similar to or related to the initial query term will be added to the initial query set, and we will get a extended set of query terms which is longer than the initial query terms set. The methods widely used in the research of query expansion are as follows: 1. a method of query expansion based on Thesaurus[2-4]; 2. A method of query expansion based on Ontology[5-8]; 3. Global analytical methods, such as a method of query expansion method based on clustering analysis[9-10]; 4. Local analytical methods, such as a method of query expansion based on pseudo feedback model[11-12]. The performance of the method of query expansion based on Thesaurus depends on the quality of thesaurus, Therefore, dictionaries need to be constantly updated. At the same time, this method can only retrieve the results similar to the query terms but cannot retrieve the results relate to. The method of query expansion based on Ontology has excellent performance, but the construction of domain ontology needs the support of domain experts and the quality of domain ontology plays a decisive role. Therefore, this method gets poor marks for stability. Using global analysis requires calculate all the data, as a results, the system’s time and space overhead is too high to handle volumes of data. The quality of the first retrieve text is important to local analysis, if the quality of the first retrieve text is improved, we will see significant performance boosts. At present, the method of query expansion based on LDA model had been implemented and a good performance had been obtained[13]. However, only one factor was taken into
consideration by the ranking function, namely, the similarity between the topic terms and query terms. Similarly, the method of query expansion based on rank learning algorithm can provide excellent performance, however query terms were expanded by computing the statistical relationship between characteristic words and text, therefore the potential information of topics cannot be mined. In this paper, the initial results were ranked by introducing RankSVM algorithm then topics were extracted by using LDA model, and the filtered topic terms were added to the set of initial query terms.

2. Query Expansion Algorithm Based on RankSVM and LDA Model

RankSVM is an algorithm for page ranking based on SVM, and it can transform sorting problem into binary classification problem on the basis of the training set. Assuming that the query set is set \( Q = \{q_1, q_2, \cdots, q_n\} \), \((x_i^u, x_i^v)\) represents a partially ordered document pair for query \( Q \). For query \( Q \), if \( x_i^u \) is more relevant than \( x_i^v \), the document pair is marked as “1”, otherwise marked as “0”. By repeating the above procedure, we can get the tag set \( y_{u,v} \) for query \( Q \). At this time, the problem of document sorting has changed into the problem of classification document pair. Therefore, the sorting function can be defined as \( f(x) = \omega^T X \), and its loss function is as follows:

\[
\min \frac{1}{2}\|\omega\|^2 + \lambda \sum_{i=1}^{n} \sum_{u,v} y_{u,v} \xi_{u,v}^i \geq 1 - \xi_{u,v}^i, y_{u,v} = 1 \]

\( \xi_{u,v} \geq 0, i = 1, 2 \cdots n \)

In this formula, \( \omega^T \) represents the solution of the weight vector of the support vector machine, \( n \) represents the number of text, and \( \xi \) represents the relaxation variable.

LDA model was proposed by David M. Blei in 2003. It is an unsupervised learning algorithm, which can mine the hidden topic information in this document. In this model, the generation process of a text can be described as follows: a topic was selected according to the topic distribution of each previous article, and then words was selected according to the word distribution of the topic, repeating the above process. The model of LDA model is shown in Fig.1.

![Figure I Graph Model of LDA Topic Model](image)

In this figure, \( D \) represents all texts, \( w \) represents words in texts, \( z \) represents topics, LDA model considers that topics and words obey dirichlet distribution, so \( \alpha, \beta \) represents the prior parameters of dirichlet distribution, \( \theta \) represents the probability distribution of topics. In summary, the joint probability distribution of documents can be expressed as follows:

\[
p(D | \alpha, \beta) = \prod_{m} \prod_{\alpha} p(\theta_j | \alpha) \prod_{c_j} \sum_{\phi} p(z_{j,c} | \theta_j) p(w_{j,c} | z_{j,c}, \beta) d \theta_j
\]
3. EXPERIMENT

3.1. Data Sources
The corpus used in this experiment is derived from the professional literature of HowNet, including 112 articles about "acid bamboo", 39 articles about "Yunnan tiger hazel", 300 articles about "wild vetch Artemisia" and 549 noise documents. In this paper, we use an open source word segmented called HanLP, version 1.2.8 to preprocess the experimental text, such as word segmentation and deletion of stop words.

3.2. Method
Firstly, LDA model was used to train all the texts in plant field in corpus, and then Lucene, a full-text retrieval tool, was used to retrieve documents based on keywords. Then, according to the posterior distribution of the trained LDA model, the documents related to the query words were screened out. The RankSVM package, developed by Thorsten Joachims of the Department of Computer Science, Cornell University, was used to sort the filtered results. Finally, the ranked results are passed through the trained LDA model, and the top ten words were selected as the new query word set.

In order to compare with other query word expansion methods, we designed the following four experiments: A. Keyword-based retrieval; B. Query word expansion method using ranking algorithm only; C. Query word expansion method using LDA model only, D. The method in this paper.

3.3. Result and Analysis
Since the corpus used in this paper is about three kinds of plants, the names of plants were used as the initial query words. Taking "acid bamboo" and "soil" as examples, the following candidate sets can be obtained by using the algorithm in this paper.

In this paper, recall rate, accuracy rate, MAP and P@k were used as the evaluation indicators of the experimental results. The results are shown in the following fig.4.

![Figure II Recall Rate of Four Groups](image)

From the figure, we can see that the recall rate after query word expansion is higher than that without query word expansion, which shows that the method of query word expansion can retrieve more comprehensive results. At the same time, the performance of group C is better than that of group B, which shows that LDA model can mine potential information of documents. Group D combines the advantages of Group B and Group C, so it has the better performance.
From the graph, we can see that the retrieval accuracy of the experimental group with function of query word expansion is higher than that without query word expansion, which shows that using query word expansion can improve the retrieval accuracy. The performance of group B is better than that of group C, which shows that the RankSVM algorithm of group B has better performance than group C which considered the correlation between query words and topics. Group D can not only get high quality initial retrieval documents, but also have the advantage of mining potential topics, so it has the better performance.

From the graph, we can see that the performance of the method with function of query word expansion is higher than that of without query word expansion, and the performance of group D which combines the advantages of B and C is the best.

3.4. Conclusions and Prospects

Based on the above experimental results, the following conclusions can be drawn in this paper: 1. Because query expansion expands query words, the performance is better than that without query word expansion. 2. High-quality initial results are conducive to the acquisition of high-quality query word extension set. A set of high-quality initial results can be obtained by using RankSVM algorithm, which is conducive to the improvement of accuracy; LDA model can mine potential topic information, which is conducive to the improvement of recall rate. 3. Because the method in this paper combines the advantages of RankSVM algorithm and LDA model, it is superior to the other query word expansion methods based on pseudo-relevance feedback in each index. The experimental results show that this method is superior to other query word expansion methods, but the size of the corpus limits the performance. Therefore, in the future, we will focus on enriching corpus.
ACKNOWLEDGMENTS
The findings and the opinions are partially supported by projects of the National Natural Science Foundation of China (NSFC), No. 31960290. This work is supported by Engineering and Technology Research Center for Yunnan Agricultural Big Data and Yunnan Agricultural University, China.

REFERENCES
[1] LI Ai-ming. (2015). Research on Query Expansion Method Based on Ontology and User Query Intention. Information Science, 33(05):68-71.
[2] MA Hui-nan etal. (2007). An approach to fuzzy query expansion based on synonymy thesaurus. Journal of Dalian University of Technology, 47(3):439-443.
[3] Xu Jianxin etal.(2010). Improved Local Co-occurrence Query Expansion Based on Synonymous. JOURNAL OF INTELLIGENCE, 29(9):145-147.
[4] Wu Qin etal.(2010). a local Query Expansion Method Based On Semantic Dictionary.Journal of Nanjing University (Natural Science), 50(4): 526-533.
[5] OUYANG Liubo, TAN Ruizhe. (2015) . Query expansion method based on ontology and user log. Computer Engineering and Applications, 51(1):151-155.
[6] HAO Zhi-feng etal. (2014). Query expansion integration of correlation rules and ontology weight graph. Application Research of Computers,31(10):3028-3032.
[7] Wang Hongxia. (2016). Application Research of Semantic Query Expansion Based on Ontology. BULLETIN OF SCIENCE AND TECHNOLOGY, 32(1):118-122.
[8] Wan Jing etal. (2013). Search Extension Based on Ontology and Local Context Analysis.Control Engineering of China, 20(3): 558-561.
[9] Yang Jing etal. (2013). A Query expansion Method Based on Constrained Semi-supervised Clustering. Chinese Science Paper, 8(10): 994-997.
[10] YUAN Jinsheng, CHENG Chaoran. (2012). Query recommendation algorithm based on text clustering search engine. Computer Engineering and Applications, 48(3): 129-132.
[11] Feng Ping, Huang Mingxuan. (2011). Query Expansion of Pseudo Relevance Feedback Based on Feature Terms Extraction and Correlation Fusion. Data Analysis and Knowledge Discovery, 2011, 22(1): 52-56.
[12] Wang Meng etal. (2014). Short Text Expansion and Classification Based on Pseudo-relevance Feedback. Journal of Zhejiang University(Engineering Science), 48(10): 1835-1842.
[13] Yan Zequan. Research On Query Expansion Based On Topic Model. (2014). Harbin: Harbin Institute of Technology.
[14] Xu bo etal. (2015). A Query Expansion Method Based on Learning to Rand. Journal of Chinese Information Processing, 29 (3):155-160.
[15] Lin yuan. (2012). Research of Learning to Rank in Information Retrieval. Dalian:Dalian University of Technology.
[16] D.M.Blei, A.Y.Ng, M.I.Jordan. (2003). Latent Dirichlet Allocation.Journal of Machine Learning Research, 4(3): 993-1022.