Improve Diversity-oriented Biomedical Information Retrieval using Supervised Query Expansion

Bo Xu, Hongfei Lin, Liang Yang, Kan Xu, Yijia Zhang, Dongyu Zhang, Zhihao Yang,
Jian Wang, Yuan Lin, Fuliang Yin
Faculty of Electronic Information and Electrical Engineering
Dalian University of Technology Dalian, China
{xubo, hflin}@dlut.edu.cn

***Abstract*—In the paper, we propose a novel diversity-oriented biomedical information retrieval method based on supervised query expansion. Our method aims to obtain the most relevant and diversified terms to enrich user queries for better interpreting the information needs. We first propose a diversity-oriented labeling strategy to annotate the usefulness of candidate expansion terms. We then extract both the context-based and resource-based term features to represent terms as feature vectors. In model training, we propose a diversity-oriented group sampling method to modify the loss function of learning-to-rank for accurate biomedical term ranking. Experimental results on TREC Genomics datasets show that our method is effective in improving the performance of biomedical information retrieval in terms of both the relevance and the diversity.**

***Index Terms*—biomedical information retrieval, supervised query expansion, term ranking model, diversity**

I. Introduction

In recent years, the rapid development of biomedical research has led to an explosive increase in the number of related articles, which poses a great challenge for researchers to obtain the needed information. To meet the challenge, biomedical IR systems are designed to retrieve a list of articles, which is not only relevant to given queries, but also diversified to completely meet the information needs. The relevance of search results reflects the similarity between a given query and each candidate document, while the diversity of search results captures different aspects of the query.

Existing studies have focused on improving the relevance and the diversity of biomedical IR. However, few studies have addressed the diversity of search results using query expansion. Generally, query expansion methods can be categorized into two types: unsupervised query expansion (UQE) and supervised query expansion (SQE). To comprehensively measure the usefulness of expansion terms, SQE methods are developed recently, and have been demonstrated effective to improve the quality of the selected expansion terms [1]–[4]. SQE methods represent candidate expansion terms as feature vectors. Term feature vectors are treated as the inputs of a classifier or a ranker for further refinement. SQE methods have exhibited two advantages over UQE methods for improving retrieval performance. For one thing, SQE methods take multiple features of terms into consideration to choose high-quality expansion terms instead of using one certain scoring function. For another, SQE methods can achieve satisfactory performance particularly in retrieval with constraints due to the flexibility in the model optimization. Unlike general IR tasks, biomedical IR has the constraints on the diversity, and few studies have integrated SQE methods in biomedical IR.

TABLE I
Supervised Query Expansion for Biomedical information retrieval

Algorithm 1 Supervised Query Expansion Pipeline
Training the SQE model M
1: For each training query q, select k candidate terms via PRF
2: Label each term based on the diversity-oriented strategy
3: Represent each term as a feature vector using different term features
4: Train term ranking model M using the modified loss function
Testing the model M in query expansion retrieval
1: For each testing query q, select k candidate terms via PRF
2: Represent each term as a feature vector using the term features
3: Apply M to obtain the top m terms for query expansion

In this paper, we propose a novel supervised query expansion method for diversity-oriented biomedical information retrieval. Our method aims to address the relevance and the diversity of search results simultaneously. We propose a biomedical term labeling strategy to measure the relevance and the diversity of each candidate expansion terms, and extract both the context-based and the resource-based term features to comprehensively represent the terms for further refinement. In model training, we integrate the group sampling and diversity-oriented weighting function into the loss function of ranking support vector machines to improve the quality of expansion terms. We conduct extensive experiments on the datasets from TREC Genomics tracks. Experimental results demonstrate the effectiveness of our method in biomedical IR, outperforming other state-of-the-art baseline methods.

II. General Framework

We introduce our supervised query expansion framework for diversity-oriented biomedical information retrieval. We illustrate the pipeline of our framework in Table I.

A. Term Labeling Strategies

Ground truth labels of terms are taken as the learning targets in training a term ranking model. During model training, term labels are used to measure the ranking loss. To assign term labels, an initial retrieval is conducted using an original

978-1-5386-5488-0/18/$31.00 ©2018 IEEE

query q. We evaluate its performance using an evaluation metric $Eval$, denoted as $Eval(q)$. $Eval$ can be any evaluation metrics for measuring retrieval performance, such as mean average precision. Then, we conduct another retrieval using an expanded query containing q and one candidate term t. We record its performance as $Eval(q,t)$. By comparing $Eval(q)$ with $Eval(t,q)$, we determine whether t is relevant to the query q or not. To generate the diversity-oriented term labels, an intuitive way is to determine the diversity of a term based on whether the term is contained in query-related aspects. Namely, if the term occurs in any query-related aspect, we assign the label 1 to the term, indicating its usefulness. Otherwise, we assign the label 0 to the term, indicating its uselessness. Although this labeling strategy seems simple and feasible, it may ignore important information of the term: a term contained in several aspects may be more diversified than the term contained in only one aspect; potential impact of terms on retrieval performance is still an important factor in term labeling. Based on the above consideration, we propose a labeling strategy based on both diversity and relevance of terms, and categorize term labels into four classes: definitely useful (labeled as 3), partly useful (labeled as 2), probably useful (labeled as 1) and not useful (labeled as 0). The advantage of the labeling strategy with multiple labels lies in that it may compute the ranking loss more accurately than that with binary labels, thus obtaining more effective models.

B. Term Features

We represent all terms as input feature vectors for term ranking models to refine obtained candidate terms. Each dimension of these vectors, as one type of term feature, can reflect the term usefulness to the given query from a certain perspective. In our method, two types of term features are extracted: one is based on term distribution in the context; the other is based on term importance in biomedical resources. Based on these two factors, we define two sets of features below for candidate expansion terms.

1) Context-based Features: Context-based features are extracted to measure the relative importance of terms with respect to a given query within the retrieval collection. Different types of textual statistics can be adopted to measure term importance. We extract two categories of context-based features: features based on term frequency and inverse document frequency ($tfidf$) and features based on co-occurrences ($cooc$), which have been proved effective as term features [4].

For $tfidf$ based features, we extract term frequency, inverse document frequency and their combination as different features, respectively. Term frequency counts the occurrences of a certain term within each document, and inverse document frequency counts the number of documents containing a certain term. These two factors, as classical textual statistics, can jointly reflect term importance in the entire collection in terms of $tfidf$. Moreover, since term distribution in the whole corpus and in the feedback documents (top-ranked document from initial retrieval) may characterize candidate expansion terms differently, we extract term features from these two document sets, respectively, as different term features.

To further consider query information, we extract co-occurrence based term features by accounting for co-occurrences of each candidate term and its corresponding query in documents. Intuition for the co-occurrence lies in that two terms tend to be more relevant when they co-occur more frequently in the context. Moreover, we split original documents as sliding windows, and consider the co-occurrences of each two terms within each smaller textual windows. Context-based features are extracted to address term importance and term relevance to given queries. Since queries in biomedical literature retrieval always contains domain-specific terms, we further characterize the terms using biomedical resources for comprehensively representing candidate expansion terms.

2) Resource-based Features: Biomedical semantic resources involve large amounts of semantic and syntactic information of biomedical terminologies for modeling domain-specific terms. These resources have been widely used in biomedical text mining tasks. To comprehensively represent terms, we incorporate resources into modeling term characteristics to encode domain-specific term information. We investigate two resources in our study: Medical Subject Headings (MeSH) and MetaMap.

MeSH provides hierarchically-organized terminologies for indexing and cataloging of biomedical information on PubMed search engines, which have been widely used in biomedical information retrieval tasks. The frequency of term occurrences in MeSH can reflect the importance of terms in biomedicine. We define two indicators to measure the term importance in MeSH: MeSH-based term frequency and MeSH-based concept frequency. MeSH-based term frequency counts the number of occurrences of each term in the entire vocabulary. Intuitively, if a term appears more frequently in MeSH, it will be more important as a terminology. We formalize this indicator as follows.

$$tf_{MeSH}(t_j) = \frac{log(freq(t_j, MeSH) + 1.0)}{log|T|} \quad (1)$$

where $|T|$ represents the number of terms in MeSH. $freq(t_j, MeSH)$ represents the raw frequency of the term t_j in MeSH. Inspired by document frequency used in IR, we propose MeSH-based concept frequency as another indicator for measuring term importance. MeSH-based concept frequency counts the number of unique concepts containing a certain term in MeSH. If a term is contained within more concepts, it will be more important to reflect domain-specific characteristics of terms. We formalize this method as follows.

$$idf_{MeSH}(t_j) = \frac{M - m(t_j) + 1.0}{m(t_j) + 1.0} \quad (2)$$

where M represents the number of concepts in MeSH. $m(t_j)$ represents the number of unique concepts containing the term t_j. $idf_{MeSH}(t_j)$ measures the importance of the term t in MeSH in analogy with inverse document frequency used in

IR. Moreover, we also combine $tf_{MeSH}(t_j)$ and $idf_{MeSH}(t_j)$ as a new term feature as follows.

$$tfidf_{MeSH}(t_j) = idf_{MeSH}(t_j)log(tf_{MeSH}(t_j) + 1.0) \quad (3)$$

MetaMap as a powerful natural language processing tool, has been widely used in biomedical text mining tasks [5], which is developed by the National Library of Medicine (NLM). MetaMap seeks to discover domain-specific concepts from biomedical text in the Unified Medical Language System (UMLS) metathesaurus. We adopt MetaMep to map expanded queries to a concept query, and extract term features using the concept query. Specifically, we combine one candidate expansion term with the original query to form an expanded query, and then convert the expanded query from a text query to a concept query. The canonical forms of Concept Unique Identifiers (CUIs) are contained in the concept query. If the concept query contains more biomedical concepts, it is more likely to convey useful information about the term, and the term may be more useful for expansion. The number of recognized concepts is treated as a term feature. We formalize this feature as follows.

$$concept(t) = count(t, Q_{expand}(t)) \quad (4)$$

where $Q_{expand}(t)$ is the expanded query with the term t. $count(t, Q_expand(t))$ measures the number of times term t appearing in the concept representations of the expanded query. Since MetaMap returns several candidates for an expanded query with several concepts, the number of returned candidates may also reflect the importance of the term. We define two term feature based on this consideration as follows.

$$conceptnum(t) = count_{CUI}(t, Q_{expand}(t)) \quad (5)$$

$$candidate(t) = \frac{\sum_{q \in Q_{expand}(t)} |R(c)|}{count_{CUI}(t, Q_{expand}(t))} \quad (6)$$

where $conceptnum(t)$ counts the total number of concepts in the concept query Q_{expand}, which measures term importance at the query level. $|R(c)|$ is the number of returned candidates for the concept c with respect to Q_{expand}. We normalize the feature values by the number of concepts contained in the concept query to make the feature values comparable to each other.

We represent each candidate expansion term as a feature vector using the context-based and resource-based features. In model training, we treat the term feature vectors as inputs and the term labels as targets for optimizing the intermediate models by pre-defined ranking loss functions.

C. Group Enhanced Loss Function for Term Ranking

In this section, we introduce the ranking loss function in our method. We adopt a group sampling method based on group-wise learning to rank methods [6]. To apply group-wise learning to rank for biomedical term selection, we divide the set of terms for each query into small groups based on the divide-and-conquer strategy, and each group of terms comprises one term with higher label and several terms with lower labels. We then adopt ranking support vector machines (RankSVM) [7] to examine the performance of our model. Formally, the objective function of RankSVM is defined as follows.

$$\begin{aligned} &min\frac{1}{2}\omega^T\omega + C\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{u,v,y^i_{u,v}} \xi^{i,j}_{u,v} \\ &s.t.\ \ \omega^T(t^{i,j}_u - t^{i,j}_v) \geq 1 - \xi^{i,j}_{u,v}, t^{i,j}_u \succ t^{i,j}_v, \xi^{i,j}_{u,v} \geq 0 \end{aligned} \quad (7)$$

where $t^{i,j}_u \succ t^{i,j}_v$ implies the term u should be ranked ahead of term v with respect to the j^{th} group of i^{th} query. C is the trade-off coefficient between the ranking loss and the model complexity.

We modify the objective function of RankSVM following our diversity-oriented loss function by incorporating the diversity-oriented weighting function for computing the total ranking loss. The final form of the objective function is defined as follows.

$$\begin{aligned} &min\frac{1}{2}\omega^T\omega + C\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{u,v,y^i_{u,v}} \xi^{i,j}_{u,v} \\ &s.t.\ \ \gamma(t^{i,j}_u)\omega^T t^{i,j}_u \geq \gamma(t^{i,j}_v)\omega^T t^{i,j}_v + 1 \\ &\qquad - \xi^{i,j}_{u,v}, t^{i,j}_u \succ t^{i,j}_v, \xi^{i,j}_{u,v} \geq 0 \end{aligned} \quad (8)$$

where γ is the diversity-oriented weighting function based on group sampling. We believe the model learned using this function can select more relevant and diversified terms for biomedical query expansion to enhance the retrieval performance.

III. Experiments

A. Experimental Settings

We conduct our experiments on the datasets from TREC Genomics tracks in 2006 and 2007 [8], [9]. We adopt four evaluation metrics designed for the tracks: Document MAP, Passage MAP, Passage2 MAP and Aspect MAP. We build our information retrieval system based on Indri search engine [10], which is widely used in existing works. We index articles from the experimental datasets with stemmed words and stopword removed in advance. We tune the parameters of our method for 2006 dataset with 2007 queries, and tune the parameters for 2007 dataset with 2006 queries. We perform five-fold cross validations, and report the average performance of all the folds. The division of training set, testing set and validation sets is based on query number at the ratio of 3:1:1, which is used for model training, prediction, and parameter selection, respectively. The division follows the standard learning-to-rank datasets LETOR [11].

TABLE II
OVERALL RETRIEVAL PERFORMANCE OF DIFFERENT MODELS FOR 2006 QUERIES

2006 queries	Document	Passage	Passage2	Aspect
Language model [12]	0.3178	0.0205	0.0239	0.1983
Relevance model [13]	0.3194	0.0207	0.0240	0.2023
Term dependency [14]	0.3198	0.0208	0.0254	0.1785
SVM-based SQE [1]	0.3050	0.0237	0.0292	0.2447
ListNet [15]	0.3216	0.0234	0.0290	0.2256
RankSVM [7]	0.3065	0.0235	0.0335	0.2632
Our model	0.3282*†	0.0249*†	0.0345*†	0.2828*†
2007 queries	Document	Passage	Passage2	Aspect
Language model [12]	0.2587	0.0646	0.0876	0.2000
Relevance model [13]	0.2678	0.0720	0.0963	0.2302
Term dependency [14]	0.2804	0.0683	0.0939	0.1974
SVM-based SQE [1]	0.2833	0.0729	0.0999	0.2298
ListNet [15]	0.2819	0.0739	0.1012	0.2255
RankSVM [7]	0.3226	0.0844	0.1160	0.2467
Our model	0.3337*†	0.0847*†	0.1155†	0.2713*†

B. Overall Retrieval Performance

For the models compared, the query-likelihood language model [12] is one of the classic retrieval models in IR field, which is also taken as the basic retrieval model in our experiments. Relevance model [13] and term dependency model [14] are two unsupervised query expansion models widely used in different tasks. Support Vector Machine (SVM), RankSVM and ListNet [15] are three learning to rank methods belonging to the pointwise approach, the pairwise approach and the listwise approach, respectively, in which SVM-based SQE method has been proved effective in [1]. We report the results of these models in Table II on the two datasets. We conduct two-tailed paired Student t-tests ($p < 0.05$) to examine whether the improvements are significant relative to the baseline models, where an asterisk indicates significant improvements over the RankSVM-based model and a dagger indicates significant improvements over the ListNet-based model. The table shows that compared to classic retrieval models, unsupervised query expansion methods can improve the retrieval performance of biomedical retrieval task, and supervised query expansion method can further enhance the performance for both query sets. Among the supervised query expansion methods, our method significantly outperforms other methods in terms of most evaluation metrics, which shows the effectiveness of our model.

IV. CONCLUSIONS AND FUTURE WORK

We propose a novel supervised query expansion method for diversity-oriented biomedical information retrieval. In the proposed method, we propose a term labeling strategy in consideration of diversity degree of terms, extract both context-based and resource-based term features for term representations, and modify the loss function with group sampling and diversity-oriented weighting function to learn more effective ranking model for term selection. Experimental results on TREC datasets show that our method outperforms baseline models, and effectively improves the performance of biomedical information retrieval in terms of both relevance-based and diversity-based evaluation measures. We will carry out our future work by extracting more powerful term features based on other useful biomedical resources to improve the performance, and also investigating other effective supervised learning methods for further optimizing our method.

V. ACKNOWLEDGEMENTS

This work is partially supported by grant from the National Natural Science Foundation of China (No. 61572102, 61632011, 61772103, 61602078, 61702080), Project funded by China Postdoctoral Science Foundation, the Fundamental Research Funds for the Central Universities (DUT18ZD102).

REFERENCES

[1] Guihong Cao, Jian Yun Nie, Jianfeng Gao, and Stephen Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 243–250, 2008.

[2] Yuanhua Lv, Cheng Xiang Zhai, and Wan Chen. A boosting approach to improving pseudo-relevance feedback. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 165–174, 2011.

[3] Zhiwei Zhang, Qifan Wang, Luo Si, and Jianfeng Gao. Learning for efficient supervised query expansion via two-stage feature selection. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 265–274, 2016.

[4] Bo Xu, Hongfei Lin, and Yuan Lin. Assessment of learning to rank methods for query expansion. *Journal of the Association for Information Science & Technology*, 67(6):1345–1357, 2016.

[5] A. R. Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, 2001(1):17, 2001.

[6] Yuan Lin, Hongfei Lin, Zheng Ye, Song Jin, and Xiaoling Sun. Learning to rank with groups. In *ACM International Conference on Information and Knowledge Management*, pages 1589–1592, 2010.

[7] Yunbo Cao, Jun Xu, Tie Yan Liu, Hang Li, Yalou Huang, and Hsiao Wuen Hon. Adapting ranking svm to document retrieval. In *International Acm Sigir Conference on Research & Development in Information Retrieval*, pages 186–193, 2006.

[8] William R. Hersh, Aaron M. Cohen, Phoebe M. Roberts, and Hari Krishna Rekapalli. Trec 2006 genomics track overview. In *Fifteenth Text Retrieval Conference, Trec 2006, Gaithersburg, Maryland, November*, pages 14–23, 2006.

[9] William Hersh and Ellen Voorhees. Trec genomics special issue overview. *Information Retrieval*, 12(1):1–15, 2009.

[10] Trevor Strohman, Donald Metzler, Howard Turtle, and W. Bruce Croft. Indri: A language-model based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*, 2005.

[11] Tieyan Liu, Jun Xu, Tao Qin, Wenying Xiong, and Hang Li. Letor: Benchmark dataset for research on learning to rank for information retrieval. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007.

[12] Chengxiang Zhai and John Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *Tenth International Conference on Information and Knowledge Management*, pages 403–410, 2001.

[13] Lavrenko, Victor, Croft, and W. Bruce. Relevance based language models. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 120–127, 2001.

[14] Yuan Lin, Hongfei Lin, Song Jin, and Zheng Ye. Social annotation in query expansion: a machine learning approach. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 405–414, 2011.

[15] Zhe Cao, Tao Qin, Tieyan Liu, Mingfeng Tsai, and Hang Li. Learning to rank: from pairwise approach to listwise approach. In *International Conference on Machine Learning*, pages 129–136, 2007.