Fusion of Domain Knowledge and Text Features for Query Expansion in Citation Recommendation

Yanli Hu[1], Chunhui He[1], and Bin Ge[1]

Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha Hunan 410073, P.R. China
{huyanli,gebin}@nudt.edu.cn, {xtuhch}@163.com

Abstract. Academic citation recommendation addresses the task of recommending citations for a scientific paper. Effective citation recommendation is greatly important for literature reviewing, literature-based discovery and a wide range of applications. In this paper, we propose a query expansion framework via fusing domain-specific knowledge and text features for academic citation recommendation. Starting from an original query, domain-specific and context-aware concepts are derived to expand the query to improve the performance of citation recommendation. From the perspective of enriching knowledge structure, domain-specific concepts are extracted from domain knowledge graphs such as ACM Computing Classification System and IEEE thesaurus. From the perspective of providing query scenarios, the query is extensively considered with context-aware concepts derived from text feature extraction. Then candidate concepts are filtered via distributed representations like BERT to expand the query. Experiments of citation recommendation for papers in public data sets show that our proposed model of query expansion improves the performance of academic citation recommendation.

Keywords: Query Expansion · Citation Recommendation · Domain Knowledge Graph · Feature Extraction · Distributed Representation.

1 Introduction

Literature reviews are in great demand arising from the rapidly-increasing publications of scientific articles as well as maintaining awareness of developments in scientific fields. Finding and evaluating relevant material to synthesize information from various sources is a crucial and challenging task.

One main method of literature searching is keyword search. However, keyword search can't meet the need of literature searching partly due to the *synonymy* and *polysemy* problems. On the other hand, literature reviews consist of correlated articles of variant words, resulting in worse performance of citation recommendation.

The situation improves little despite of the development of search techniques in recently years. Take the literature searching of surveys as examples. Three surveys are randomly selected from a bibliography sharing service, CiteULike, and the retrieval performance by the search engine Google Scholar is evaluated

Table 1: The number of references of surveys hit in the top 100 retrieval results.

Surveys	#references	#references hit in the top100 results
Title: A Survey of Mobility Models for Ad Hoc Network research **keywords**: ad hoc networks, entity mobility models, group mobility models	33	0
Title: A Survey of Web Information Extraction Systems **keywords**: Information Extraction, Web Mining, Wrapper, Wrapper Induction	47	14
Title: A Survey of Controlled Experiments in Software Engineering **keywords**: Controlled experiments, survey, research methodology, empirical software engineering	52	13

according to the amount of references located among the searching results, as illustrated in Table 1. The amount of references of the surveys hit in the top 100 retrieval results indicates poor performance for citation recommendation.

In this paper, we propose a novel query expansion framework for academic citation recommendation. By incorporating multiple domain knowledge graphs for scientific publications, we use domain-specific concepts to expand an original query with appropriate spectrum of knowledge structure. Meanwhile, text features are extracted to capture context-aware concepts of the query. Then candidate concepts, namely domain-specific and context-aware concepts, are further filtered via distributed representations to derive the expanded query for citation recommendation.

2 Related Work

Traditional methods of query expansion choose terms from relevant/irrelevant documents. Terms are usually weighted according to their frequency in single document and in the collection, and the top terms with the highest frequency are added into the initial query. To provide structured knowledge of a topic, Ref [1] addresses the problem of citation recommendation by expanding the semantic features of the abstract using DBpedia Spotlight [14], a general purpose knowledge graph.

A content-based method for recommending citations [2] is suggested, which embeds a query and documents into a vector space, then reranks the nearest neighbors as candidates using a discriminative model. A context-aware citation recommendation model [8] is proposed with BERT and graph convolutional networks. In the context of heterogenous bibliographic networks, ClusCite[12], a cluster-based citation recommendation framework, is proposed.

Text feature extraction is fundamental to citation recommendation. *TF-IDF* [9] (Term Frequency–Inverse Document Frequency) and their variants are essential to represent documents as vectors of terms weighted according to term frequency in one document as well as in the number of documents in the corpus.

To capture the latent semantic associations of terms, Latent Semantic Analysis (LSA) [5] is proposed to analyze the relationships of documents based on common patterns of terms. To uncover semantic structures of documents, Latent Dirichlet Allocation (LDA) [3] is a generative probabilistic model to represent one document as a mixture of topics, and the words in each document imply a probabilistic weighting of a set of topics that the document embodies.

In the above representations, each term in documents is represented by a one-hot vector, in which only the corresponding component of the term is 1 and all others are zeros. To provide a better estimate for term semantics, distributed representations of terms are proposed to boost the semantics of documents. Instead of sparse vectors, embedding techniques, such as Word2Vec[10] and BERT[6], learn a dense low-dimensional representation of a term by means of its neighbors.

3 The Proposed Framework

The proposed query expansion framework for citation recommendation is composed of three major steps: (1) From the perspective of enriching knowledge structure, expanding an original query with domain-specific concepts based on multiple domain knowledge graphs. (2) From the perspective of providing query scenarios, extending the query with context-aware concepts derived from text feature extraction. (3) Filtering out the above candidate concepts via distributed representations to derive the expanded query for citation recommendation.

3.1 Model Overview

To capture diverse information needs underlying the query, we propose a query expansion framework combining domain knowledge with text features, as illustrated in Fig. 1.

Starting from an initial query q_0 in Fig. 1, domain knowledge and text features are used to expand the query, and then candidate concepts are filtered to derive the enriched query for citation recommendation.

More specifically, domain knowledge are used to provide knowledge structure in the form of domain-specific concepts, with the aid of multiple domain knowledge graphs such as ACM Computing Classification System [1] (CCS for short) and IEEE thesaurus [2]. Meanwhile, text feature extraction is utilized to derive context-aware concepts from document collections to provide query scenarios.

Among the above domain-specific and context-aware candidate concepts, filtering techniques are applied to choose a set of closely-related concepts to formulate a new query q, and citations are recommended with respect to the expanded query q.

3.2 Domain-specific Expansion Using Multiple Knowledge Graphs

With the characteristics of citation recommendations in mind, we propose query expansion based on knowledge graphs. However, general-purpose knowledge graphs are not applicable in academic disciplines.

[1] https://dl.acm.org/ccs

[2] https://www.ieee.org/content/dam/ieee-org/ieee/web/org/pubs/ieee-thesaurus.pdf

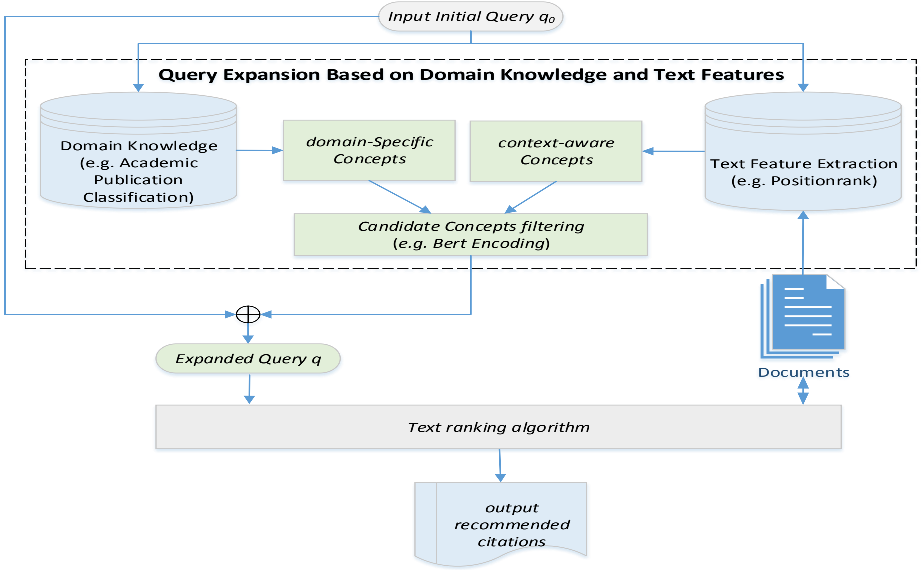

Input Initial Query q_0
Query Expansion Based on Domain Knowledge and Text Features
Domain Knowledge (e.g. Academic Publication Classification)
domain-Specific Concepts
context-aware Concepts
Text Feature Extraction (e.g. Positionrank)
Candidate Concepts filtering (e.g. Bert Encoding)
Documents
Expanded Query q
Text ranking algorithm
output recommended citations

Fig. 1: The query expansion framework based on domain knowledge and text features.

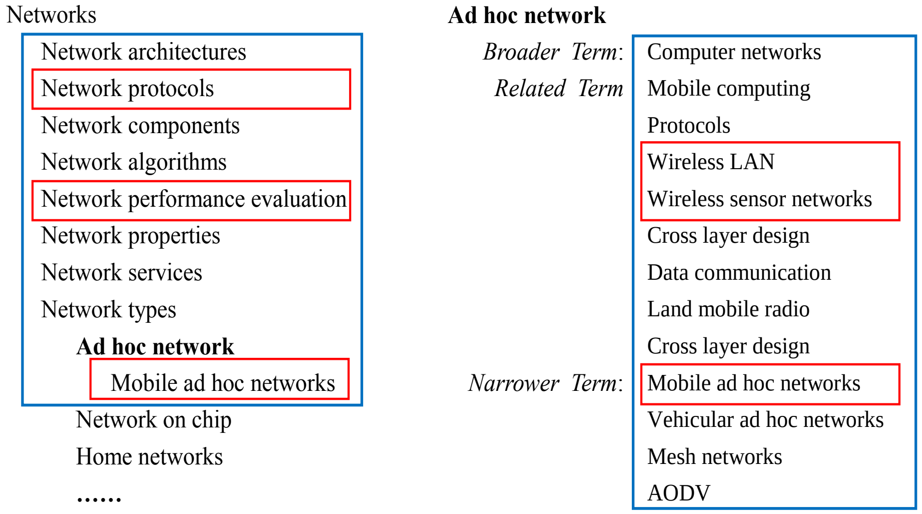
Take the citation recommendation for 'A Survey of Mobility Models for Ad Hoc Network research' in Table 1 as an example (hereinafter called the survey). Using Microsoft Concept Graph [3], the related concepts for *ad hoc network* consists of *research domain*, *wireless network*, *network*, *system* and so on. As to the concept *mobility model*, the suggested concepts include *parameter*, *component*, *user criterion* and so on. Analyzing the references cited in the survey, the suggested concepts derived from general-purpose knowledge graph contribute little to recommend appropriate citations. On the other hand, knowledge graphs such as WordNet [11] have also limited contribution, as no concepts for *ad hoc network* and *mobility model* .

We utilize domain-specific knowledge graphs, such as ACM CCS and IEEE thesaurus, to expand academic concepts. Take the concept *ad hoc networks* as an example. The concept hierarchies of ACM CCS and IEEE thesaurus are illustrated in Fig. 2, which suggest structures of concepts and their relations with *ad hoc networks*. Exploring the references of the survey, highly relevant concepts appearing in the references are circled in red rectangles in Fig. 2.

For ACM CCS, we notice that the concepts of the siblings of *network types*, a hypernym of *ad hoc networks*, are also highly related to the topics of the references, rather than the direct siblings of *ad hoc networks* like *networks on chip* or *home networks*. Here a concept's siblings refer to those concepts that share the same parents with the given concept.

Therefore, to expand domain-specific concepts of a term to the original query, we design adaptive policies to multiple domain knowledge graphs. For IEEE thesaurus, the concepts of broader terms, related terms and narrower terms are

[3] https://concept.research.microsoft.com/



(a) The concept hierarchies of ACM CCS for *ad hoc network*

(b) The concept hierarchies of IEEE thesaurus for *ad hoc network*

Fig. 2: The diverse concept hierarchies for *ad hoc networks* in multiple domain knowledge graphs.

added into the expansion set of candidate concepts. For ACM CCS, a two-level policy is suggested to add the concepts of the hyponyms and the siblings of the hypernyms of the term to the original query.

Using the policy, domain-specific knowledge for *ad hoc networks* with 22 concepts are derived from the ACM CCS and IEEE thesaurus, as circled in the blue rectangles in Fig. 2. The concept *mobile ad hoc networks* is counted once.

3.3 Context-aware Expansion based on Text Feature Extraction

To expand a query with context awareness, the basic idea is to derive context-aware terms from document collections based on text feature extraction, as context-aware terms enriched the original query with characteristics of scenarios for better matching of relevant documents.

Key phrases are first extracted using PositionRank algorithm [7] as features of documents. Then document collections are processed to derive semantically related key phrases with terms in the original query, such as highly co-occurred key phrases. Providing the scenarios of co-occurrence with respect to the original query, semantically related key phrases are appended to the candidate set of query expansion.

Fig. 3 lists ten context-aware concepts of *ad hoc networks* and *mobility model* based on text feature extraction, respectively. Here, the context-aware concepts are computed from the CiteULike corpus which will be described in detail in Section 4.

3.4 Candidate Concept Filtering via Distributed Representation

Rather than directly expanding an original query with candidate concepts in Section 3.2 and Section 3.3, the aim is to filter the set of candidate concepts to derive a subset of closely-related concepts for query expansion to reduce noise.

#	Context-aware concepts to *ad hoc networks*
1	networks simulator
2	multihop wireless networks
3	mobile devices communicate
4	bandwidth channel
5	networks protocol design
6	mobility patterns
7	simulation environments
8	mobility models
9	wireless networks
10	Network protocol

#	Context-aware concepts to *mobility model*
1	mobile users
2	wireless networks
3	wireless links
4	performance differences
5	mobility pattern
6	random movement
7	mobile station
8	cellular environment
9	traffic parameters
10	cellular systems

Fig. 3: The context-ware concepts of *ad hoc networks* and *mobility model.*

In order to solve this problem, we propose a candidate concept filtering method via distributed representations. The input consists of two sets, a candidate concept for expansion and the original query. Then, the vectorized representations of the inputs are concatenated via BERT distributed representations, and each input will be converted into a 1*1024 dimensional vector. For detailed calculation principles, please see the BERT-as-service tool [4]. The cosine similarity is used to calculate the distance between the vector representations of the candidate concept and the original query, and output the normalized result (between 0 and 1) as a matching score. With the matching scores between candidate concepts and query, candidate concepts are sorted, and the top-k closely-related concepts are chosen to expand the original query.

4 Experiments

4.1 Data Sets and Evaluation Metrics

In this section, we test our model for citation recommendation tasks on two public data sets. The DBLP data set contains citation information extracted by Tang et al.[13]. The CiteULike data set consists of scientific articles from CiteULike database [5]. Statistics of the two data sets are summarized in Table2.

Table 2: Data Sets Overview.

Data sets	*CiteULike*	*DBLP*
#papers	24,167	149,363
#terms	6,488	30,849
#average citations per paper	8.25	5.52
#papers in training set	14501	89619
#papers in validation set	4833	29872
#papers in testing set	4833	29872

The performance is mainly evaluated in terms of precision and recall. The precision at top N results (P@N for short), and the recall at top N results

[4] https://github.com/hanxiao/bert-as-service

[5] http://static.citeulike.org/data/current.bz2

(R@N for short) are reported respectively. Additionally, mean average precision (MAP)[4] is evaluated to measure the performance averaged over all queries,

$$MAP(Q) = \frac{1}{\|Q\|} \sum_{j=1}^{\|Q\|} \frac{1}{m_j} \tag{1}$$

where Q is the set of queries. For the query q_j, $\{d_1, ..., d_{m_j}\}$ is the set of cited articles and R_{jk} is the set of ranked results from the top result to the article d_k.

4.2 Experimental Results

The experimental results on the CiteULike data set are presented in Table 3. We vary our model with text analysis and filtering methods. For example, the *domain KG + TF-IDF* method expands query without filtering, namely implementing domain-specific expansion based on multiple domain knowledge graphs and context-aware expansion with TF-IDF. All candidate concepts are appended to the original query without filtering.

Similarly, the *domain KG + LSA* and *domain KG + LDA* methods use domain knowledge plus LSA and LDA, respectively. In contrast to the first three methods, BERT embeddings of candidate concepts are used for filtering in the *domain KG + TF-IDF + BERT filtering* method.

Table 3: Experimental Results of our model on the CiteUlike data set

CiteULike	MAP	P@10	P@20	R@20	R@50
domain KG + TF-IDF	0.211	0.448	0.364	0.277	0.313
domain KG + LSA	0.150	0.340	0.271	0.145	0.209
domain KG + LDA	0.113	0.278	0.222	0.122	0.182
domain KG + TF-IDF + BERT filtering	**0.287**	**0.694**	**0.605**	**0.329**	**0.361**

The results in Table 3 show that among the first three methods without filtering, the *domain KG + TF-IDF* method has better performance than the ones using LSA and LDA. Thus, the impacts of filtering technique on query expansion are further evaluated based on TF-IDF.

The last two methods with filtering outperformed the ones without filtering. And the *domain KG + TF-IDF + BERT filtering* method contributes the best performance among the five methods on the CiteULike data set. It indicates that the expansion of domain-specific and context-ware concepts plus BERT filtering improves the performance of citation recommendation.

The experimental results on the DBLP data set also proves the effectiveness of query exapnsion using domain-specific and context-ware concepts plus BERT filtering, as shown in Table 4.

Table 4: Experimental Results of our model on the DBLP data set

DBLP	MAP	P@10	P@20	R@20	R@50
domain KG + TF-IDF	0.095	0.160	0.151	0.203	0.365
domain KG + LSA	0.087	0.143	0.127	0.182	0.360
domain KG + LDA	0.073	0.137	0.122	0.172	0.282
domain KG + TF-IDF + BERT filtering	**0.168**	**0.295**	**0.228**	**0.331**	**0.439**

The above results suggest that combining domain knowledge with text features indicates remarkable advantages for citation recommendation, and filtering of candidate concepts is key to improve the performance.

5 Conclusions

In this paper, we address the problem of citation recommendation for literature review. Fusing domain knowledge and text feature to expand query is verified to improve the performance of locating citations scientific articles. Domain-specific concepts are extracted from multiple domain knowledge graphs to enrich knowledge structure for query expansion. Context-aware concepts are derived from text feature extraction to provide query scenarios. Then candidate concepts are filtered via distributed representations like BERT to expand the query with closely-related concepts. Experiments of citation recommendation on bibliographic databases show that our proposed model effectively improves the performance of citation recommendation.

Future research considers using large-scale scientific literature corpora to fine-tune the BERT pre-training vectors. In addition, the combination of named entity recognition technology to achieve the extraction of term features is also our focus.

References

1. Ayala-Gómez, et al.: Global citation recommendation using knowledge graphs. J. Intell. Fuzzy Syst. **34**(5), 3089–3100 (2018)
2. Bhagavatula, C., et al.: Content-based citation recommendation. arXiv preprint arXiv:1802.08301 (2018)
3. Blei, D.M., et al.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
4. Christopher, D.M., et al.: Introduction to information retrieval (2008)
5. Deerwester, S., et al.: Indexing by latent semantic analysis. J. Inf. Sci. **41**(6), 391–407 (1990)
6. Devlin, J., et al.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR **abs/1810.04805** (2018)
7. Florescu, C., et al.: PositionRank: An unsupervised approach to keyphrase extraction from scholarly documents. In: ACL. pp. 1105–1115. ACL (Jul 2017)
8. Jeong, C., et al.: A context-aware citation recommendation model with bert and graph convolutional networks. arXiv preprint arXiv:1903.06464 (2019)
9. Jing, L.P., et al.: Improved feature selection approach tfidf in text mining. In: ICMLC. vol. 2, pp. 944–946. IEEE (2002)
10. Mikolov, T., et al.: Distributed representations of words and phrases and their compositionality. In: NIPS. pp. 3111–3119 (2013)
11. Miller, G.A.: WordNet: An electronic lexical database. MIT press (1998)
12. Ren, X., et al.: Cluscite: Effective citation recommendation by information network-based clustering. In: SIGKDD. pp. 821–830. ACM (2014)
13. Tang, J., et al.: Arnetminer: Extraction and mining of academic social networks. In: SIGKDD. p. 990–998. ACM (2008)
14. Xu, B., , et al.: Cn-dbpedia: A never-ending chinese knowledge extraction system. In: IEA/AIE. pp. 428–438. Springer (2017)