SciGraph: A Knowledge Graph Constructed by Function and Topic Annotation of Scientific Papers

Yuchen Yan
School of Government
Beijing Normal University
Beijing China
yanyuchen@mail.bnu.edu.cn

Chong Chen*
School of Government
Beijing Normal University
Beijing China
chenchong@bnu.edu.cn

ABSTRACT

In researchers' literature exploration, a knowledge structure of certain domain helps those without clear purposes gain a quick understanding of new topics. Besides that, a paper set classified by function save the exploration time of those with clear purposes. This paper proposes a solution to annotate the function and topics of scientific papers, and construct a knowledge graph, named SciGraph, which helps researchers obtain both purpose-oriented papers and the knowledge structure of a domain. A dataset of 0.9 million scientific papers from different disciplines are collected, and the proposed solutions are applied to form SciGraph, within which a total ca. 1.9 million nodes and ca. 16.4 million relations are generated. The contribution of this study includes, (1) organizing the functions and topics of scientific papers within a unified knowledge graph, which may support explorative retrieval; (2) proposing a DF-ITF method to identify the candidate feature words in function definition; and (3) applying a method based on term co-occurrence to extract and rank the hyponymy relation of topic keywords.

CCS CONCEPTS

Information systems→Information retrieval→Retrieval models and ranking→Combination, fusion and federated search

KEYWORDS

Scientific papers, Paper annotation, Knowledge graph, Natural language processing

1 Introduction

Scientific paper is a necessary medium with which researchers analyze the research trends and form the understanding on specific domains[1,2].But the domain knowledge keep changing under the rapid development of science and technology, thus researchers have to frequently face new topics and numerous papers while establishing their own studies[3].A fine-grain knowledge organization approach, which integrates both the functions and topics, is needed to save researchers' exploration efforts in a certain domain.

On one hand, different types of papers are needed in different stages of a research. For example, a newcomer of a topic probably needs review papers to form a general understanding, while a proficient researcher may prefer papers on specific problems, such algorithms, theory, application details, etc. Thus, scientific papers should be classified into functions to meet the various purposes of researchers in different cases.

On the other hand, the relation among related topics helps researchers building a complete understanding on unfamiliar topics. If hyponym concepts could be identified, researchers would not have to suffer so much difficulty in finding relevant papers at their beginning stage on unfamiliar topics.

To achieve these goals, this study proposed a scientific knowledge organization system, SciGraph, which integrates the function and topic annotation results of scientific papers from multiple scientific disciplines, trying to meet researchers' explorative academic information needs.

2 Methodology

2.1 Dataset

The dataset is composed of 0.9 million Chinese scientific papers sampled from 1,203 domains in 11 scientific disciplines with titles, abstracts, and keywords to keep a reasonable domain coverage. The domains and the disciplines are chosen based on Chinese Library Classification (CLC) system. The CLC label is letter-numeral format with the first letter for discipline and the following numbers for subordinate domains.

2.2 Function Annotation

The function of papers is defined according to the search purpose of researchers who are in certain study stage, such as to know the research progress, find methods, etc. This paper proposes a method to filter out feature words for the function definition and trains a model to annotate the function of papers.

1. Function Types Inspired by the well-known TF-IDF method, this study puts forward a DF-ITF method to select words, that are widely used in research papers based on consensus of paper-writing in multiple domains[4]. The words such as *theory*, *approach*,

[*] Corresponding Author

experiment, *progress*, etc., are identified as feature words according to DF-ITF, that appear as less as possible in one paper while widely appear in papers of the whole corpus, i.e. low term frequency and high document frequency.

Given dataset $D = \{d_1, d_2, d_3, \dots, d_m\}$, $|D| = m$, term set $T = \{t_1, t_2, t_3, \dots, t_n\}$, $|T| = n$, for each term t_i,

$$itf_i = \frac{1}{\left|\sum_{j=1}^{m} \sum_{t \in d_j} \{t : t = t_i\}\right|}, i \in \{1,2,3,\dots,n\} \quad (1)$$

$$df_i = \left|\sum_{j=1}^{m} \{j : t_i \in \mathrm{d_j}\}\right|, i \in \{1,2,3,\dots,n\} \quad (2)$$

$$df - itf_i = df_i \times itf_i, i \in \{1,2,3,\dots,n\} \quad (3)$$

The value of *df-itf*$_i$ is in range (0,1].The higher it is, the more widely-used the term t_i is.

The DF-ITF method is applied to score the weight of each term in our dataset. The top 1000 terms are selected as candidate feature words for function definition. A total of six function types, namely Review & Progress, Demonstration & Comparison, Argumentation & Discussion, Theory & Computation, Technology & Method, and Design & Application, are defined according to observation on these words. And the feature words are selected for each type to represent the function.

2. Function Annotation Annotating the function of scientific papers is a text classification task. The training set is prepared by an easy-to-start way, i.e., if the title of a paper contains a feature word of a function type *f*, the paper will be annotated as *f*. This study applies a text classification model with BERT embedding on abstract, using a softmax function to map the model output to the possibility of class, and select the class with the highest possibility as the function annotation of output given paper. The overall accuracy achieves 0.72.

2.3 Topic Annotation

This study try to annotate quality topic keywords of papers, and then evaluate the hyponymy relation weights of each keyword pairs, extracting and ranking the hyponymy relation in the corpus.

1. Keyword Extraction Referring to previous studies, this study takes it as a sequence labeling task[5,6,7]. The paper abstracts are labeled with BIO labels according to the keywords assigned by the authors. A token classification model with BERT embedding is trained and applied to abstract texts in order to identify more quality keywords from papers. Finally, the candidate keywords are extracted from the output of the last layer according to the rule of BIO labels. The F1 value of this model reaches to 0.475.

2. Hyponymy Relation Identification Based on term co-occurrence, this study designs a method to identify the hyponymy relations among different topic keywords by ranking the relation weight C, as shown in Equation 4. A relation between keywords k_1 and k_2, denoted as $C(k_1, k_2)$, can be represented by their frequency $F(k_1), F(k_2)$ and their co-occurrence $F(k_1, k_2)$. ε is a threshold in relation identification. If $C(k_1, k_2) > 0$, k_1 is likely to be a hypernym keyword of k_2 since k_1 is more widely used in the corpus than k_2, and its meaning is taken as more "general", and vice versa. $C(k_1, k_2) = 0$ means there is not an obvious relation between k_1 and k_2, i.e. the relation is not considered in our study.

$$C(k_1, k_2) = \begin{cases} \dfrac{F(k_1, k_2)}{F(k_2)} & , F(k_1, k_2) \geq \varepsilon, F(k_1) > F(k_2) \\ 0 & , F(k_1, k_2) < \varepsilon \\ -\dfrac{F(k_1, k_2)}{F(k_1)} & , F(k_1, k_2) \geq \varepsilon, F(k_1) \leq F(k_2) \end{cases} \quad (4)$$

In order to reduce the calculation cost in relation identification, a threshold of $F(k_i)$, i.e. δ, is set. The candidate keyword k_i is kept only when $F(k_i)$ is not less than δ . The threshold δ measures the generality of a topic keyword. In this study, the threshold $\varepsilon = 50, \delta = 20$.

2.4 Knowledge Graph Construction

For the sake of interaction and flexibility[8,9], we construct a knowledge graph, SciGraph, to organize the annotated functions and topics of papers. There are four kinds of nodes and four kinds of relations. The total ca. 1.9 million nodes include 0.9 million scientific papers, 1,203 CLC labels, the six functions and ca. 1 million topic keywords. The ca. 16.4 million relations among these nodes include the CLC label of scientific papers (HAS_CLC), the functions of papers (HAS_FUNCTION), the topic keywords of papers (HAS_KEYWORD), and the hyponym relation among topic keywords (SUB-OF). To better show the structure of SciGraph, we choose a paper in machinery industry, showing the nodes and relations related to this paper. The episode of SciGraph is shown in Figure 1.

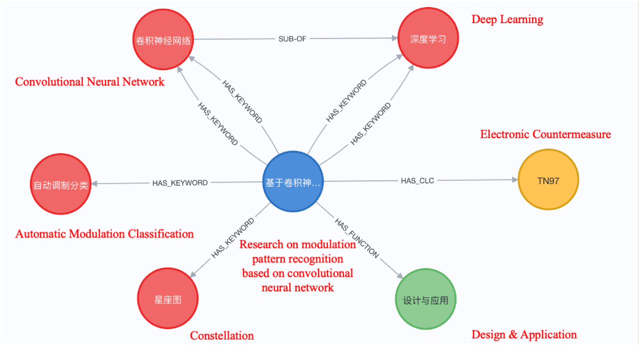

Figure 1. An Episode of the SciGraph

3 Conclusion and Discussion

Through the proposed DF-ITF method and term-co-occurrence-based hyponym extraction, this paper designs annotation models from two folds, i.e. function and topic, constructs SciGraph based on the annotation results, and organizes scientific knowledge on a huge dataset with ca. 0.9 million papers from 1,203 different domains of science and technology. The contribution of this study is to provide flexibility to both fine-grained knowledge organization of concerned domains and literature search to researchers of different purposes.

There are still many open problems to be solved, and this work can go further, for example, the usage of state-of-the-art models, the structure of function types, the relation identification methods and the value of threshold ε and δ, etc.

ACKNOWLEDGMENTS

This study is supported by National Social Science Foundation of China (grant number 21BTQ065).

REFERENCES

[1] Pavel Savov, Adam Jatowt, and Radoslaw Nielek. 2020. Identifying breakthrough scientific papers. Information Processing & Management 57, 2 (March 2020), 102168. DOI:https://doi.org/10.1016/j.ipm.2019.102168

[2] Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction.

[3] Zhenyu Gou, Fan Meng, Zaida Chinchilla-Rodríguez, and Yi Bu. 2021. Revisiting the Obsolescence Process of Individual Scientific Publications: Operationalisation and a Preliminary Cross-discipline Exploration.

[4] Hongseon Yeom, Youngjoong Ko, and Jungyun Seo. 2019. Unsupervised-learning-based keyphrase extraction from a single document by the effective combination of the graph-based model and the modified C-value method. Computer Speech & Language 58, (November 2019), 304–318. DOI:https://doi.org/10.1016/j.csl.2019.04.008

[5] Wasi Ahmad, Xiao Bai, Soomin Lee, and Kai-Wei Chang. 2020. Select, Extract and Generate: Neural Keyphrase Generation with Syntactic Guidance.

[6] Huanqin Wu, Wei Liu, Lei Li, Dan Nie, Tao Chen, Feng Zhang, and Di Wang. 2021. UniKeyphrase: A Unified Extraction and Generation Framework for Keyphrase Prediction. arXiv:2106.04847 [cs] (August 2021).

[7] Liangping Ding, Zhixiong Zhang, Huan Liu, and Yang Zhao. 2021. Design and Implementation of Keyphrase Extraction Engine for Chinese Scientific Literature. In EEKE '21, September 27-30, 2021, Illinois, USA.

[8] Xinyu Li, Chun-Hsien Chen, Pai Zheng, Zuoxu Wang, Zuhua Jiang, and Zhixing Jiang. 2020. A Knowledge Graph-Aided Concept–Knowledge Approach for Evolutionary Smart Product–Service System Development. Journal of Mechanical Design 142, 10 (May 2020). DOI:https://doi.org/10.1115/1.4046807

[9] Xiaoli Geng and Tianwen Deng. 2021. Research on Intelligent Recommendation Model Based on Knowledge Map. J. Phys.: Conf. Ser. 1915, 3 (2021), 032006. DOI:https://doi.org/10.1088/1742-6596/1915/3/032006

[10] Thomas Wolf, Lysandre Debut, Victor Sanh and et.al. 2020. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 38–45. DOI:https://doi.org/10.18653/v1/2020.emnlp-demos.6