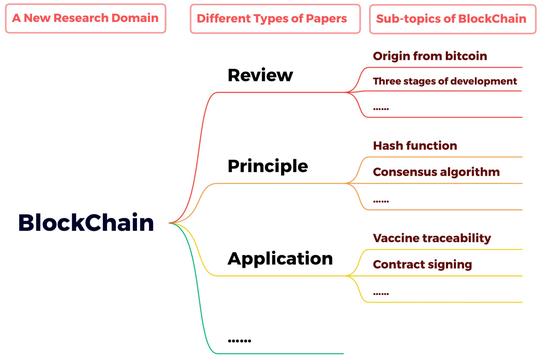
SciGraph: A Knowledge Graph Constructed by Function and Topic Annotation of Scientific Papers

Yuchen Yan, Chong Chen

Beijing Normal University, Beijing, China

INTRODUCTION

◆**Background:** The knowledge of a domain keeps changing under the rapid development of science and technology, researchers have to frequently face new topics and numerous papers while establishing their own studies. On one hand, different types of papers are needed in different stages of a research. On the other hand, the relations among related topics helps researchers building a complete understanding on unfamiliar topics.



◆**Target:** (1) Annotate scientific papers in two dimension, i.e. function and topic, (2) Construct a knowledge graph named SciGraph to organize the annotation results.

◆**Significance:** (1) Use SciGraph to form a literature retrieval system; (2) Help researchers obtain both purpose-oriented papers and the knowledge structure of a domain.

MATERIALS & METHODS

◆**Dataset:** The dataset is composed of ca. 0.9 million Chinese scientific papers with titles, abstracts and keywords, and sampled from 1,203 domains in 11 scientific disciplines in order to keep a reasonable domain coverage.

◆**Function Annotation:** (1) Propose a DF-ITF method to select the function words with top 1000 DF-ITF values; (2) Cluster the identified function words into six function types: *Review & Progress, Demonstration & Comparison, Argumentation & Discussion, Theory & Computation, Technology & Method, and Design & Application*; (3) Apply a BERT-based text classification method to annotate a suitable function on scientific papers, the overall accuracy of the model achieves 0.72 .

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

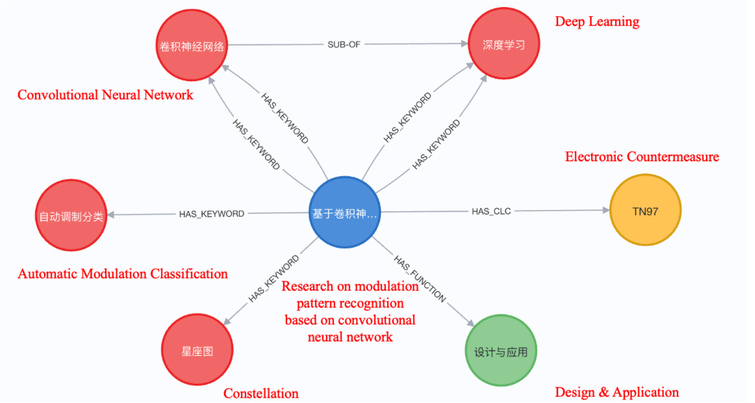
$$df_i = \left|\sum_{j=1}^{m}\{j: t_i \in \mathrm{d_j}\}\right|, i \in \{1,2,3,\dots,n\}$$

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

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

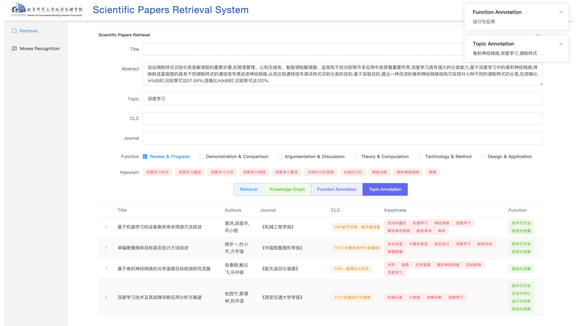
◆**Topic Annotation:** (1) Using a BERT-based token classification to extract topic keywords from abstracts of scientific papers, the F1 value reaches 0.475; (2) Apply a term-occurrence-based method to calculate the hyponym relation weight $C(k_1,k_2)$ of topic keyword pairs k_1 and k_2; (3) Identify and rank the hyponym relations based on the weights above.

◆**Knowledge Graph Construction:** This study construct a knowledge graph, SciGraph, to organize the annotated functions and topics of papers. There are four kinds of nodes (total ca. 1.9 million) and four kinds of relations (total ca. 16.4 million) among above nodes.

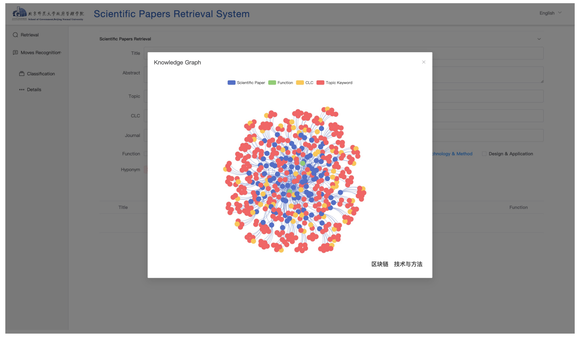


APPLICATION DEMO

In order to show the significance of the proposed solution on explorative retrieval, this study design and construct a scientific paper retrieval system.



This system contains a distinguishing function, i.e. displaying knowledge graph, which meet the need of explorative retrieval particularly.



CONTACT

Yuchen Yan (yanyuchen@mail.bnu.edu.cn)

Chong Chen (chenchong@bnu.edu.cn)

北京師範大學
BEIJING NORMAL UNIVERSITY