Submitted on: 03.10.2019

Title of the Satellite Meeting: Knowledge Management section Satellite Conference
Date: Thursday, 22, August, 2019
Location: Ionian University，Corfu, Greece

Describe Library Resources with Knowledge Graph

Lu Zhang
Department of Resource Collection, National Science Library, Chinese Academy of Sciences, Beijing, China.
E-mail address: zhanglu@mail.las.ac.cn

Abstract:
Libraries have large amount of credible knowledge. But unfortunately, advanced Internet search tools and knowledge graphs cannot fully cover the valuable library collections. Using knowledge graph to describe collections can optimize knowledge services in several aspects. Specifically, the application of knowledge graph can help library enhance retrieval efficiency, integrate various resources and improve reference services. In order achieve these improvements, National Science Library of Chinese Academy of Sciences tries to build a knowledge graph connecting all the collections and open resources. The work of academic knowledge graph has been completed. However, in order to reveal the semantic connection of resources, a concept knowledge graph still needs to be working on, which is the focus of this research. This study proposes a process of constructing concept knowledge graph and discusses possible applications of knowledge graph in library.

Keywords: Academic knowledge graph, Concept knowledge graph, Library collections.

1. Knowledge Graph and Library

Knowledge graph was launched by Google in 2012 to improve search results. The essence of knowledge graph is a semantic network. By connecting kinds of entities together in a network, the knowledge graph describes entities and concepts in the real world structurally, and provides the ability to analyze information using relationships. Therefore, knowledge graph is applied in semantic search, question-and-answer system and intelligent knowledge service. Many industries, especially information service industry, place great effort on building domain-specific knowledge graph.

Many search engine companies and research institutions have built various knowledge graphs for improving searching. For example, in the academic search engine of Microsoft Academic Service(MAS), the graph has been used in providing interactive search experience by harvesting the syntactic and semantic cues for parsing and predicting user queries, and taking advantage of the relationships across different types of entities to offer heterogeneous suggestions.[1]

Libraries have large amount of credible knowledge. But unfortunately, advanced Internet search tools and knowledge graphs cannot fully cover the valuable library collections. In the future, library should be the best "knowledge acquisition tool" rather than the biggest "knowledge warehouse". Therefore, libraries need to build knowledge graph to describe their collections.

2. Construction of Knowledge Graph

The knowledge graph of library collections should consists of two parts as shown in the Fig. 1. The first part is academic knowledge graph, which describes academic entities and their relations. The entities includes researcher, article, database, project, conference and so on. These data will be extracted from structured meta data. The second part is concept knowledge graph, which describes entities at semantic level. In a concept knowledge graph, we got entities like concept, drug, plant and so on. We use them describe academic entities by build relations between them just like the red lines in the picture below. Concept entity will be used to describe researcher, article and so on. Other entities in the concept knowledge graph will do the same thing. The data in the concept knowledge graph are extracted from semi-structured corpus like title, abstract.

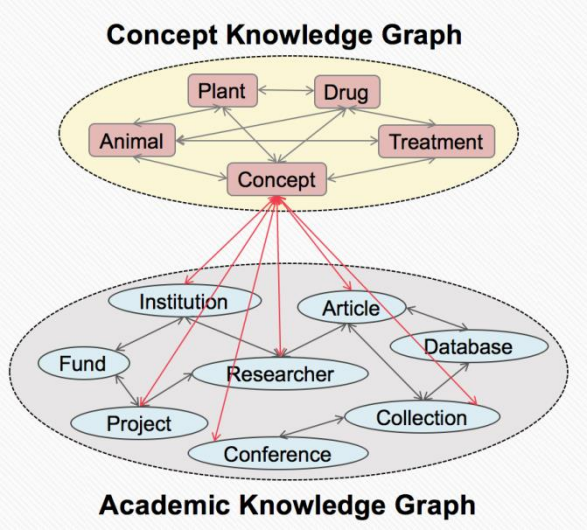

Fig. 1 Two Parts of Knowledge Graph

3.1 Construction of Academic Knowledge Graph

With the development of artificial intelligence, the technology of knowledge graph building based on academic research data have become increasingly mature. The construction of academic knowledge graph simply includes two parts: entity recognition and relation building. We could get instances of academic entities in the structured meta data directly, and

just build the relations between them according to the schema predefined as shown in the table.

Table. 1 Example of Relations Built in the Academic Knowledge Graph

Relation	Subject	Object
publish	institution	collection
source_is	article	collection
affiliation	researcher	institution
fund_by	project	institution

National Science Library of Chinese Academy of Sciences has successfully extracted information from Sci-Tech big data and built an academic knowledge graph[2] connecting all the collections and open resources, which includes more than 300 million entities and 1.1 billion relations. It is used to support knowledge discovery platform and smart personal research assistant apps "Scholar-In" for scientific big data. The schema of the graph is designed by Wang et al.(2019)[2], as shown in the Fig. 2. The article in the schema could be a journal paper, conference paper, book chapter or thesis. And the collection could be a journal, book or proceedings. Libraries can take SKOS as the basic data model. Besides, the entity graph of Microsoft Academic Service (MAS)[1] and VIVO[3] ontology have built efficient schema for academic knowledge graph, which could be used as core model of library's knowledge graph. For example, the entity graph of MAS is comprised of six types of entities that model the real-life academic communication activities: field of study, author, institution, paper, venue, and event.

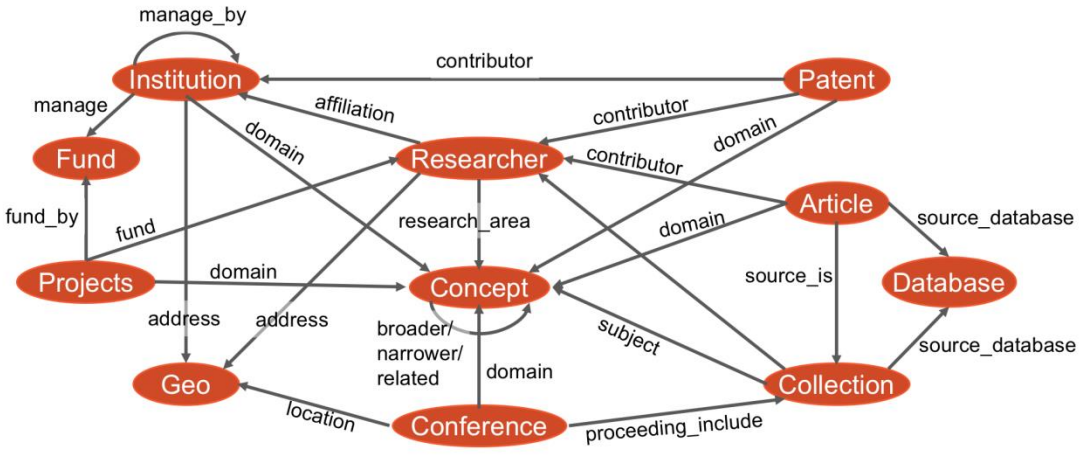

Fig. 2 Schema of Academic Knowledge Graph

The difficulties of building academic knowledge graph includes data cleaning and entity alignment. Data cleaning aims at recognize duplicate meta data from different sources, which could be solved with Authority files and some rules. For example, using ISBN to identify the

same book. Entity alignment means to recognize same entities, like same author, same institutions and so on. This can also be solved with authority files and some rules. For example, if two researchers with different name have the same ORCID or e-mail address, they could be recognized as the same person.

It should be noted that, academic knowledge graphs focus on establishing external associations between different types of entities and is relatively easy to build. There are still many difficulties in concept knowledge graph construction. However, concept knowledge graph is the key to reveal semantic associations between scattered resources. Some researches transformed traditional knowledge organization system data into a SKOS knowledge ontology structure[4], which make sense but are inefficient. Only by establishing the concept knowledge graph can we effectively describe and manage a large number of rapidly growing information resources.

3.2 Construction of Concept Knowledge Graph

The construction of concept knowledge graph mainly includes two parts: Knowledge Extractor and Concept Knowledge Graph Builder, as shown in Fig. 3. In the first part, corpus including title and abstract will be used to extract terms and relations. In the second part, the terms and relations will be further processed in order to build the concept knowledge graph.In the application of concept knowledge graph, librarians and users may give some feedbacks. The system administrator will adjust rules, patterns, parameters or only certain data according to feedbacks. The concept knowledge graph will be optimized during iterative adjustments.

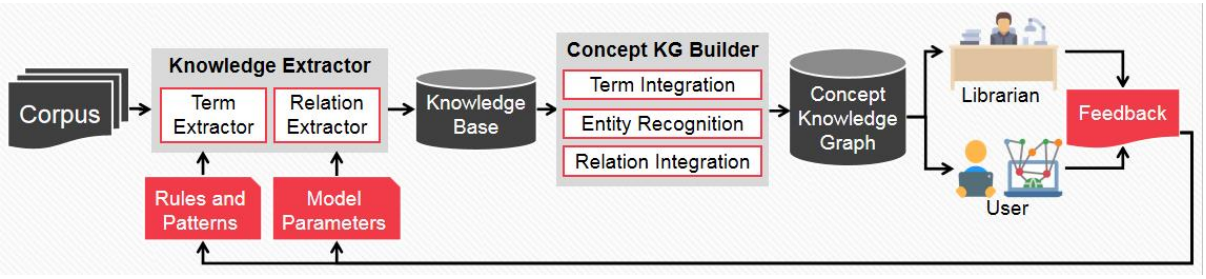

Fig. 3 Process of Building Concept Knowledge Graph

3.2.1 Knowledge Extractor

(1) Term Extractor

The term extractor aims to get terms with high quality. That is they are expected to be instances of entities. So, firstly terms in the thesaurus, entity base and etc. will be used as sample data for data training. In order to select qualified terms, system could evaluate the quality of extracted terms in 5 aspects below.

①Concordance[5]: PMI, PKL. This index calculates the solidification degree of a phrase, that is, the degree of closeness between words in the current phrase. For a phrase v, it can be split into most-likely sub-units <ul, ur> such that PMI is minimized. Pointwise Mutual Information (PMI) and Pointwise Kullback-Leibler Divergence(PKL) could be used as the concordance features.

$$PMI(u_l, u_r) = \log \frac{p(v)}{p(u_l)\,p(u_r)}$$

$$PKL(v \| < u_l, u_r >) = p(v) \log \frac{p(v)}{p(u_l)\,p(u_r)}$$

②Informativeness[5]: IDF. Function words or stop words are not informative. Using stop word dictionary, the average IDF of words in the concept will be calculated.

③Popularity[5]: Frequency. This traditional indicator reflects the importance of phrases, and high-quality phrases should appear frequently enough in corpus.

④Degree of freedom: Information entropy. The value of the entropy indicates the instability of the relationship between current word and adjacent words. The information entropy of high-quality phrases should be a small value.

⑤Composition of the phrase: The part of speech structure, and the proportion and position of stopping words. For example, the beginning word and the end word of a concept are generally not stop words, and the proportion of stop words in the whole phrase should be lower than a certain threshold. Rules and threshold varies in different studies, which could be obtained through machine learning.

Then, we could extract more terms from corpus. And these terms will be used as mature terms for data training later.

(2) Relation Extractor

The relation extractor aims to get relation with high credibility. At first, classic syntactic patterns will be used to extracted pairs of instances connected by relations. Nine isA syntactic patterns could be used as the target patterns, including five classical syntactic patterns suggested by Hearst (1992)[6] and four syntactic patterns obtained through machine learning by Snow et al. (2005)[7]. In the nine syntactic patterns below, BT means broader term, while NT is narrower term.

- NT{, NT}*{, }（and|or）otherBT：temples, treasuries and other civic buildings
- BT such as {NT, }*（or|and）NT：red algae such as Gelidium
- Such BT as {NT, }*（or|and）NT：works by such authors as A, B and C
- BT{, }including{NT, }*（or|and）NT：all common-law countries, including Canada and England
- BT{, }especially{NT, }*（or|and）NT：most European countries, especially France, England and Spain
- BT like NT；
- BT called NT；
- NT is a BT；

- NT, a BT.

The pairs with hierarchical relation may appears in other classes, which could help machine to find more hierarchical syntactic patterns. And those patterns will also be used as credible pattern to extract more possible instance pairs. In this way, more and more terms and relations could be found from the corpus in the iterative loop.

According to the syntactic patterns, candidate concept pairs with isA relations are extracted from the sentences. Referring to the Probase's algorithm[8] of detecting super-concept and sub-concept, system could select the most appropriate BT and one or more possible NTs from the sentences. Credibility of the relations refers to the possibility that there is an isA relationship between the candidate concept pairs. The credibility index refers to the Plausibility index used in Probase[8], judging the credibility of evidential relations extracted from the corpus. The algorithm of Probase takes the PageRank of web pages into account, which reflects different quality of different web pages. When dealing with library collections, system also needs to take into account the different authority of different resources. The system could design a priority weighting algorithm based on indicators such as impact factor of author, number of citations, and etc.

3.2.2 Concept Knowledge Graph Builder

This part establishes knowledge graph using terms and their relationships obtained from corpus though three steps.

① Term Integration. Synonymy terms will be connected through normalizing terms or referring to existed KOS like thesaurus.

② Entity Recognition. The entity type of term will be marked by matching with entity bases or using word2vec to infer entity type according to its contexts.

③ Relation Integration. As synonymy relations are established, contradiction and duplication of relations will appear, which need to dealt with in this step.

After all of these work, the concept knowledge graph could be put into practice.

4. Application of Knowledge Graph in Library

More and more practices in library have proved the value and potential of knowledge graph. In this section, possible application will be discussed.

4.1 Semantic Search

Search engine is like a mirror of the library collections, while knowledge graph is the key to describe the collections. Knowledge graph drives the evolution of search engines, and semantic search is one of the core applications of knowledge graph.

With the support of knowledge graph, searching based on keyword matching can convert to the search based on entity and relationship. And library's search engine can deconstruct and analyze the sentences or key words entered by the searcher accurately. Then the system will find the corresponding entity and related entities through relations and then get richer knowledge. For example, the Talk to Books system proposed by Google is a typical product of knowledge graph application. It let the search engine understand the user's question and

the content of each book, and then match the information accurately. User simply type in a statement or a question like “What is the best programming language?”. The system will find a book “C Programming for Arduino” and the sentences in book related to what you typed.

This particular application can help users find interesting books that may not be available by keyword search. What the system searches is not the "string" of query, but the semantic "thing" of it, which can hardly do without the help of semantic relations.

4.2 Information Integration

Libraries have various types of collections, including printed book, e-book, research paper, database, dissertation, multimedia material and so on. Resources of different type seem like separated by "walls", but in fact, they are related at semantic level. So, using knowledge graph to describe library collections can break the walls between different type of resources by connecting them in rich dimensions.

The integration of information will not only improve the efficiency of information acquisition, but also enable users to learn knowledge in various angles. For example, in the knowledge discovery platform of NSL as shown in the Fig. 4, different aspects about a researcher will be presented to the users, including the topic he/she is following now, the researchers he/she has cooperated with, and so on.

Fig. 4 Description of A Researcher in the Discovery Platform of NSL

In addition, knowledge graph could help libraries connect resources from different institutions or sources. For example, the school library can integrate school-level resources such as textbooks, slides, invited lectures, and course-selecting system. For example, when a student searches "artificial intelligence" in the library system, relevant books, papers, software, slides, lecture videos and courses he or she could take will be presented. In this way, students can learn things from multiple dimensions and get more clues for further learning.

Libraries can also collaborate with social network sites to increase interest in reading and knowledge acquisition. For example, in the library system, user can see the latest popular books on the social networking sites and then get interested in reading them. Researchers can see discussions about the study they are interested in, which might help inspire thinking and innovation.

4.3 Reference Service

Intelligent libraries apply knowledge graph to upgrade reference services. Because knowledge graph is like the brain of reference services, helping to promote services in multiple applications.

QA system is popular nowadays and it needs to answer the user's questions explicitly. This requires the system to conduct semantic analysis of questions and provide answers based on the reasoning ability of knowledge graph. For example, if a user asks "Whether pregnant women can eat seafood", the reasoning process of the system includes three steps. Firstly, it need to find out what are the important nutrition in the amniotic fluid, and then find what are the materials may destroy those nutrition components, and then to check whether seafood contains those materials. Then the system could answer the question explicitly.

Traditional reference services could be upgraded with knowledge graph too. For example, when users need to learn the latest research results in a field, the traditional service is to provide them with relevant literature. But the system with knowledge graph could provide more clues, such as the relations of the researches, what are the hot concepts or terms in those papers and users of which area are following these researches too.

Intelligent reference services should also be able to provide personalized consultation results based on the user's identity, search history and so on. For example, for the same question "research progress of deep learning", a freshman who has never searched this field may need to know the background and development history of deep learning, while a professor of deep learning would like to know the latest breakthroughs. Knowledge graph can help librarians distinguish resources in greater detail and find resources better meet users' needs.

5. Conclusion

Compared with the messy Internet, libraries preserve highly ordered knowledge. There is a great deal of human wisdom in the work of purchasing, indexing, and shelving every book. Libraries with a high degree of organized knowledge, should be able to answer users' questions, and provide corresponding information resources. However, the current situation is that, most users prefer to search for news and papers from unknown sources on Internet search engine, to meet their needs quickly.

Each library is a part of Popper's "world three", containing a part of the human spirit products of culture, civilization, science and technology and other theoretical systems, recorded and stored by various carriers. It would be a pity if the library failed to arouse users' interest in knowledge exploration or failed to provide effective information for them due to inadequate services. Library knowledge service is not only be able to "find resources for people" and "find people for resources", but also make efforts to spread knowledge effectively. Knowledge graph will be the cornerstone of all of these work. Libraries should strengthen the use of knowledge graph to create a better environment for users to obtain information.

References

[1] Sinha, A. , Shen, Z. , Song, Y. , Ma, H. , Eide, D., Hsu, B.J., & Wang, K. (2015). An Overview of Microsoft Academic Service (MAS) and Applications. In Proceedings of the 24th International Conference on World Wide Web (WWW ’ 15 Companion). ACM, New York, NY, USA, 243-246. DOI=http://dx.doi.org/10.1145/2740908.2742839

[2] Wang,Y., Qian,L., Xie, J., Chang, Z., Kong, B.(2019). Building Knowledge Graph with Sci-Tech Big Data. Data Analysis and Knowledge Discovery, DOI: 10.11925/infotech.2096-3467.2018.1354.

[3] McCue, J., Chiang, K., Lowe, B., Caruso, B., Corson-Rikert, J., Devare, M.(2007). VIVO: Connecting people, creating a virtual life sciences community. D-Lib Magazine, ISSN 1082-9873, Vol. 13, Nº. 7-8, 2007. 13. 10.1045/july2007-devare.

[4] Chen, Q., Cao, J., Chen, R.(2019). Research and Practices from the Thesaurus to Knowledge Graph[J]. Agricultural Library and Information, 31(1): 44-53.

[5] Liu, J. , Shang, J. , Wang, C. , Ren, X. , & Han, J. (2015). Mining Quality Phrases from Massive Text Corpora. Acm Sigmod International Conference on Management of Data. Proc ACM SIGMOD Int Conf Manag Data.

[6] Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th conference on Computational linguistics - Volume 2 (COLING '92), Vol. 2. Association for Computational Linguistics, Stroudsburg, PA, USA, 539-545.

[7] Snow, R., Jurafsky, D., and Ng, A. Y. (2005). Learning syntactic patterns for automatic hypernym discovery. In Saul, L. K., Weiss, Y., and Bottou, L. (Eds.), NIPS 17, pp. 1297 – 1304. MIT Press.

[8] Ji, L., Wang, Y.J., Shi, B., Zhang, D.W., Wang, Z.Y., & Yan, J.(2019). Microsoft concept graph: Mining semantic concepts for short text understanding. Data Intelligence. pp. 1-33. doi: 10.1162/dint_a_00013