ScholarGraph :a Chinese Knowledge Graph of Chinese Scholars

Shuo Wang, Zehui Hao, Xiaofeng Meng, Qiuyue Wang
Renmin University of China, Hebei University
Beijing, Baoding in Hebei
shuowang@ruc.edu.cn, jane0331@126.com, xfmeng@ruc.edu.cn, qiuyuew@ruc.edu.cn

Abstract
Scholars and their academic information are widely distributed on the Web. Integrating these information and making association between them can play a catalytic role in academic evaluation and research. Since 2008, Web and Mobile Data Management Laboratory (WAMDM) in Renmin University of China began to collect Chinese literatures in more than 20 academic domains, and build a data integration system called ScholarSpace to automatically integrate the relevant chinese academic information from the chinese scholars and science. Focusing on the chinese scholars, ScholarSpace can give you an academic portrait about a chinese scholar with the form of knowledge graph. So the ScholarSpace can be tranformed into a knowledge graph called ScholarGraph. It includes the scholar information such as the affiliation, publications, teacher-student relationship, etc. ScholarGraph is a subset of the whole knowledge graph generated from the ScholarSpace and is published on the web page of WAMDM. ScholarGraph consists of more than 10,000,000 triples, including more than 9,000,000 entities and 6 relations. It can support the search and query about portrait of Chinese scholars and other relevant applications.

Keywords: knowledge graph construction,chinese scholars, entity, entity relation

1. Introduction

On the Web, many information about any person such as working place, living addresss, friends, personal journey can be searched online. The same things happened in the academia. The academic information related to a scholar are already available on the Web, but they are not effectively linked together. As shown in Figure 1, the relations are actually existing among scholars, journals and academic institutions. In fact, these relations are a small part of the reality. So it is necessary to build an integration system for the academic field.

The main character of the academia information is centered on scholars. Because the academic actions are completed by scholars. For example, a professor can publish a paper in a journal or a coference, supervise postgraduate students, cooperate with other scholars, apply for fundings, etc. The relations corresponding to these actions are generated by scholars.

Figure 1: Example of scholars' information.

So the data integration system needs to be focused on scholars. According to DBLP, we firstly built a Chinese data integration system – CDBLP – to collect Chinese literature information of the computer science community in China. Based on CDBLP[1], we expanded the collection from a single domain of computer science (Chen Wei et al., 2011) to 25 domains including education, management, archaeology, economics, etc. On this basis, we propose ScholarSpace system which was constructed for Chinese scholars to support Chinese search, Chinese query, Chinese journals' evaluation, etc. There is a wealth of knowledge in these integrated data. Many entities (Professor /Ph.D. /Journal /Paper) and relations (Author /PublishedIN /Affiliation /Advisor) are shown in Figure 1. So a knowledge graph – ScholarGraph – is generated from ScholarSpace. We select a partial data from ScholarSpace and transform them into SPO triples to represent the knowledge.

In Section 2, we focus on the structure design and function implemention of ScholarSpace system. We present the transformation from the ScholarSpace to the ScholarGraph in Section 3. We show the statistic and examples about the application of ScholarGraph in Section 4. At last, the conclusion and perspectives about the ScholarGraph is described in details.

2. Data Collection and Integration

The data collection and integration is the first step for generating a knowledge graph (Cowie and Lehnert, 1996 ; Zhao Jun et al., 2011). As mentioned above, the ScholarSpace system is so important to ScholarGraph that we can not ignore the introduction about it.

2.1 ScholarSpace Structure

As shown in Figure2, the ScholarSpace system is divided into four levels by the logical design that describes the data processing flow from the Web source to the user interface.

At the bottom level, data sources include semi-structured data or unstructured data such as personal homepage, research institutes' web pages, coferences' web pages, journals' web pages.

Above the bottom level, the data integration function model aims at designing the configuration files for each

Corresponding author is Xiaofeng Meng.
[1] http://www.c-dblp.cn/

kind of data sources that can collect the data from the Web to the database.

On the third level, data processing maintains three function parts as follow. Data cleaning needs to remove the information that is not relevant to the literature (such as documents of Call For Papers) or duplicated records from the integrated data on the Web. Entity extraction means to extract the information about authors, journals,

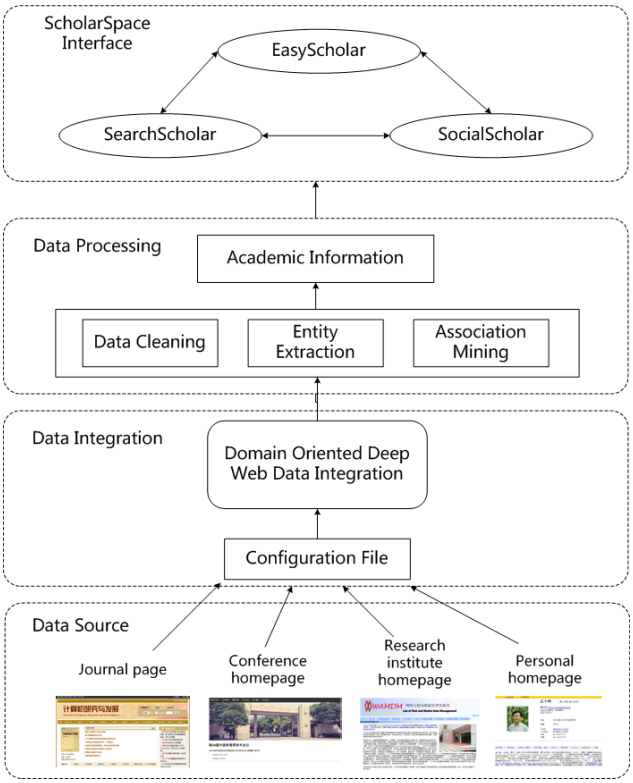

Figure 2: The logical design structure

coferences, institutes, etc. All of these objects can be described as entities with attributes and relations. Association mining can discover the associated entities and construct the relations between these entities.

On the top level, users can see three parts of ScholarSpace. The first and most important part is SearchScholar which can build the data set for the Chinese scholars and support the information query and retrieval. The second part is EasyScholar which can generate the scholar page based on SearchSholar and provide the association mining based on the author. The third part is named as SocialScholar. It builds a science community for academic communication and dissemination based on the SearchScholar and EasyScholar. The three parts support and promote each other.

2.2 ScholarSpace Storage

The ScholarSpace has more than thirty kinds of tables to store the information about the authors, the journals, the papers, etc. The kernel tables include c_papers, authortab, coauthor, disambiguation, affiliation, author_research_domain, clcs_intro, journals, projects, student, etc. The extraction and integration processing are shown in Figure 3. The author "杜治娟" is disambiguated through the table disambiguation and inserted properly into the table authortab. The paper "大数据融合研究：问题与挑战" is stored in the table c_papers. The institution "中国人民大学信息学院" is checked with the table affiliation and determined whether or not to insert it into the table. The project information about this paper are extracted and stored into the corresponding items of the table projects. There are more data extracted from the Web and stored into more tables for this paper.

3. ScholarGraph Construction

The ScholarGraph is a knowledge graph constructed based on ScholarSpace. SPO (subject-predicate-object) triples is a kind of decription forms for representing the human knowledge. For example, the sentence "姚明是篮球运动员" can be parsed into a triple "姚明-IsA-篮球运动员" as a SPO form in Chinese. ScholarGraph is a dataset which consists of SPO triples. As mentioned in Section 2, entities and their relations are extracted from the Web and stored into the tables of the database. So ScholarGraph system needs to transform the table into the triples describing the knowledge about the scholars. We need to manually define some extraction rules for recognizing the entities from database tables according to the values of database field and type. Firstly, the entities or properties are selected based on the rules. Secondly, the candidate entities and properties are manually filtered. Lastly, the relations between the entities are generated by the association mining method (Zen Wandan et al.,2006) and inspected manually.

In Figure 4, the transformation are shown from one or more tables to the triples. The paper "大数据融合研究：问题与挑战" is selected from the table c_papers. The properties of this paper include Title, Keywords, Keywords_en, Source_site, Year, CLC, etc. The triples about author "杜治娟" are generated from the table disambiguation and the table c_papers to describe the relation "Write" and the property "Name". The table affiliation and disambiguation can make up the triple for the relation "Affiliated".

4. Description and Application

ScholarGraph datasets are set up through processing databases as mentioned obove. The details of the datasets are described in in Section 4.1 and its application can be shown in Section 4.

4.1.1 Details of ScholarGraph Datasets

ScholarGraph datasets consist of 7 subsets for each scientific domain such as archaeology, computer,education, economy, geography, management and physics. The 7 subsets include Author, Affiliation, Journals,Name, Papers, and PublisheIN that can give the portraits for the Chinese scholars in one domain. We just select the necessary properties and relations from the database to construct the triples as shown in Table 1.

The value in the fourth column means that the triple does not represent relation but property. The statistics of the triple numbers are listed in Table 2. We select the data that we collected from 2013 to 2016. In fact, ScholarSpace began to integrate the Web data from 2000 and collected all papers published in the selected journals

Figure 3: Integration and storage processing in ScholarSpace

Subset	Subject Entity	Relation/Property	Object Entity/Value
Author	AuthorID	Write	PaperID
Affiliation	AuthorID	Affilated	Value
Journals	JournalID	Name	Value
	JournalID	BeginYear	Value
	JournalID	Cnki_code	Value
Name	AuthorID	Name	Value
Papers	PaperID	PaperTitle	Value
	PaperID	PaperKeywords	Value
	PaperID	PaperKeywords_en	Value
	PaperID	PaperYear	Value
	PaperID	PaperSource_site	Value
	PaperID	PaperCLC	Value
PublishedIN	PaperID	PublishedIn	JournalID
Advisors	AuthorID	GuideBy	AuthorID

Table 1: The patterns of the datasets

from the begining of the publication. So the complete dataset for the same journals has 10602497 triples. Table 2 just show 12 percent of the complete set published on the webpage of WAMDM. We do not mention the advisor subset because we just generate the teacher-student relation for computer science domain. And there are lots of authors that are unable to find their teachers or students. A data portrait for a Chinese scholar can be constructed from this dataset. For example, a scholar has relations with papers, journals and institutions. A scholar, an author entity to published papers in journals, includes relation « Write » as follows.

<http ://www.c-dblp.cn/computer/**AuthorID**/ 25447-0>

<http ://www.c-dblp.cn/computer/**Write**>

<http ://www.c-dblp.cn/computer/**PaperID**/99955>

And a paper entity can have a property triple to represent the title of this paper based on URI forms as follows.

Triple Number	Author	Journals	Name/ Affiliation	Papers	PublisheIN
Archaeology	5998	24	3661	22248	3708
Computer	39622	33	21025	84905	14128
Economics	69899	219	36167	301806	50301
Education	38371	108	20421	178620	29770
Geography	7006	18	3749	27817	4636
Management	58018	87	29715	189234	31539
Physics	94213	90	41237	151363	24802
Total	1584558				

Table 2: The statistics of the datasets

<http ://www.c-dblp.cn/computer/**PaperID**/99955>

<http ://www.c-dblp.cn/computer/Property/**PaperTitle**>

« 支持QoS保障的可信服务组合调度算法 »

A paper entity can also have a relation triple to represent the publishment of this paper as follow.

<http ://www.c-dblp.cn/computer/**PaperID**/99955>

<http ://www.c-dblp.cn/computer/**PublishedIn**>

<http ://www.c-dblp.cn/computer/**JournalID**/9>

As mentioned above, we got the information about the authors of a paper, the journal in which a paper is published, the institution of the author, etc. Further more, we infered the co-author, the interest domains of a scholar, the research topic of a scholar, etc. through the knowledge graph. So we can develop many applications with ScholarGraph (Li Hehan et al., 2015).

4.2 Applications of ScholarGraph

The application of ScholarGraph is constructed based on the knowledge graph of the scholar information. We have built an interactive interface for ScholarGraph through visualization tools. ScholarExplorer[2] is a webpage which can browse on the homepage WAMDM. In Figure 5, when you search the name "孟小峰", the results in five areas are shown on the webpage. In the middle of the page, a graph is centered by this author and links the author to the papers, journals, coauthors and institutions. In the left area, the students or teachers of this author are listed. In the right upper area, the academic statistics of this author and the properties are shown in text form. In the right lower area, user can find the word cloud of research domains about this author. In the bottom of the page, the publication years of the corresponding topic can be found. The highlight Chinese characters mean big data which is the topic, and the highlight spots are the papers related to the big data topic in the middle area. At the bottom, the highlight belts are the publication years of these papaers.

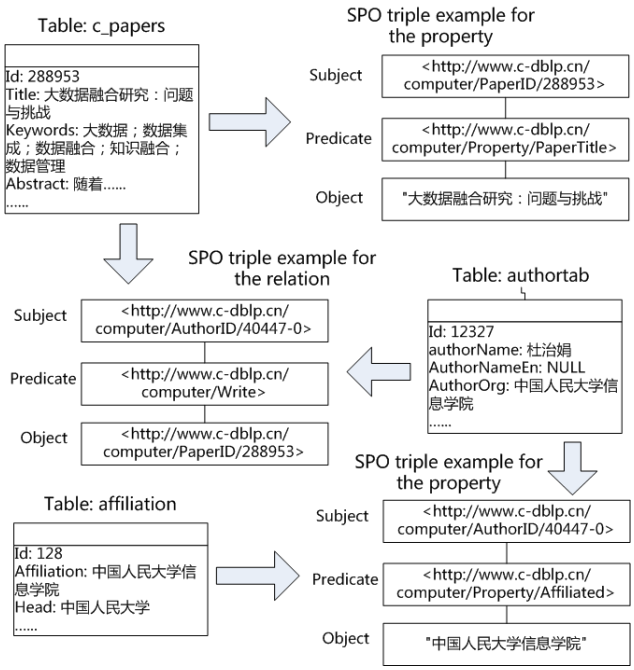

Figure 4: Examples for SPO triples generation

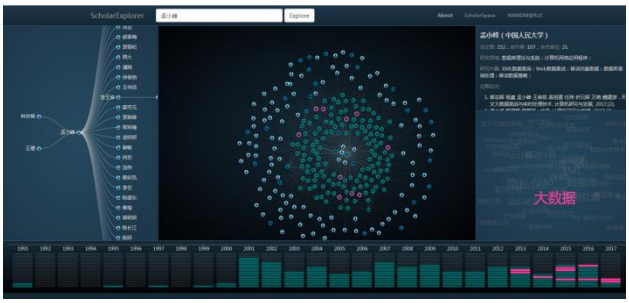

Figure 5: Application example of ScholarGraph - ScholarExplorer

Further more, we have taken adventage of the knowledge graph for the experts recommendation[3] when a paper or an item needs to review or check up. A user can input the information about the paper or the item, and the recommendation system can compute the suitability of experts in the related domain based on the ScholarGraph. Due to the word limit, we don't describe in details.

5. Conclusion and Future Work

ScholarGraph is built by the knowledge graph technique and firstly focuses on the Chinese scholars. We developed an automated system to integrate the Chinese academic information from Web and transform them into the knowledge graph for improving the application . The accuracy of ScholarGraph is relied on the manual evaluation. The triples are randomly selected to human judges and made the judgements. The average accuracy of thirteen relations or properties is 97.76%. We need to extend the domains and add the relations that exist in ScholarSpace but not in ScholarGraph in future.

6. Acknowledgements

This research was partially supported by the grants from the Natural Science Foundation of China (No. 61532010, 61379050, 61532016, 91646203, 61762082); the National Key Research and Development Program of China (No. 2016YFB1000602, 2016YFB1000603); the Fundamental Research Funds for the Central Universities, the Research Funds of Renmin University(No. 11XNL010); and the Science and Technology Opening up Cooperation project of Henan Province (172106000077) ; the Opening Project of State Key Laboratory of Digital Publishing Technology of Founder Group ; the Natural Science Foundation of Hebei Province (No. F2017201020）.

7. Bibliographical References

Chen Wei, Wang Zhongyuan, Yang Sen, Zhang Peng, Meng Xiaofeng. (2011). ScholarSpace : Computer Oriented Academic Space. Journal of Computer Research and Development, 48(S3), 395-399(in Chinese).

Cowie, J., & Lehnert, W. (1996). Information extraction : Natural language processing. Communications of The ACM, 39(1), 80-91.

Zhao Jun, Liu Kang, Zhou Guangyou, et al. (2011). Open information extraction. Journal of Chinese Information Processing, 25(6), 98-110(in Chinese).

Zeng Wandan, Zhou Xubo, Dai Bo, Chang Guiran, Li Chunping. (2006). Matrix Algorithm for Mining Association Rules. Computer Engineering, (02), 45-47(in Chinese).

Li Hehan, Meng Xiaofeng, Zou lei. (2015). Keyword query method oriented to the knowledge base of ScholarSpace. Journal on Communications, 36(12), 28-36(in Chinese).

[2] http://www.c-dblp.cn/scholarexplorer/

[3] http://www.c-dblp.cn/recommend.php