Establishment of Packaging Knowledge Graph Based on Multiple Data Sources

Changyun Li¹,², Yuezhong Wu¹,²*, Fanghuai Hu³

¹ School of Computer Science, Hunan University of Technology, Zhuzhou 412007, China.
² Intelligent Information Perception and Processing Technology Hunan Province Key Laboratory, Zhuzhou 412007, China.
³ Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai 200237, China.

Abstract

The establishment of a knowledge graph manually is time-consuming, error-prone, and complicated process. Hence, this paper proposes an automatic knowledge graph establishment method based on multiple data sources; the data sources adopted include structured data like data in relational databases, semi-structured data like data from web page, and unstructured data like text. We have used different methods to extract knowledge from these three types of data sources respectively, including rule-based method, wrapper method, machine learning method, etc. We have tested our method through establishing a knowledge graph of packaging industry; and the established knowledge graph contains about 2,000 concepts and hyponymy among them, over 10,000 entities and their attributes, and more than 10,000 relations among the entities; and the average precision of these knowledge is above 0.9. The experimental result proves that the proposed method is precise and efficient.

Keywords: Domain Knowledge Graph; Packaging Industry; Multiple Data Sources.

1. INTRODUCTION

More and more linked open data and user generated contents are published on the internet after the Semantic Web is proposed; which is now changing from a web of documents to a web of data, and contains plenty of entities and relations. The knowledge graph was first proposed by Google in 2012 (Amit, 2012), which was focused on describing various entities and concepts in the real world, as well as their relations. Actually, the knowledge graph is a new vision of ontology and extends the ontology at the level of entity. The ontology generally focuses on concepts and their relations and specifies the schema of knowledge graph; while the knowledge graph adds a large number of entities into ontology. The knowledge graph enjoys widely application in semantic search, intelligent question-answering, knowledge engineering, data mining, and other artificial applications.

From the perspective of coverage, the knowledge graph can be divided into two categories, general knowledge graph and domain knowledge graph; and the knowledge graph described by the Google belongs to the former. A domain knowledge graph usually covers only a specific domain, and it is different from a general knowledge graph in several aspects, including the coverage, depth and applications. Due to these differences, the establishment methods of them also vary. When establishing a domain knowledge graph, we usually define the schema (ontology) first, fill knowledge, including entities, attributes, relations, events; and then get knowledge from some structural data sources followed by semi-structural and unstructured data sources. This paper focuses on the establishment of domain knowledge graph.

The Section 2 of this paper describes some related works; Section 3 gives a framework of the method; the detailed establishing process is specified in Section 4; the experiment is illustrated in Section 5, and finally we conclude this paper in Section 6.
2. RELATED WORKS

The knowledge graphs have attracted increasingly more attention from both academia and industry. The Linked Open Datasets, such as freebase, DBpedia and YAGO can be viewed as general knowledge graphs. The knowledge graphs are firstly applied by some Internet search engine companies, including Google, Microsoft and Baidu, and they have made efforts to build knowledge graphs so as to improve the capacity of their search engines, and Google Knowledge Graph, Satori, and Baidu Intimate are their products. Nguyen et al. (2015) analyzed the fitness for use of two encyclopedic datasets, namely, DBpedia and Freebase, in music recommendations. Pirro et al. (2015) provided a tool called RECAP to generate explanations of relatedness on entities in encyclopedic datasets, including DBpedia and Freebase. Moreover, there are also some large-scale ones in Chinese, such as, zhishi.me (Niu et al., 2011), SSCO (Harris and Gibbins, 2003), and CN-DBPedia (Xu et al., 2017).

In contrast to the general-purposed KGs, including DBpedia and YAGO, Szekely et al. (2015) presented a system called DIG, and discussed the procedures to establish a knowledge graph for combating human tracking with it. Ruan et al. (2014) presented us a solution on building marine-oriented knowledge graphs in; and Ruan et al. (2016) have also built an enterprise knowledge graph covering about 40,000,000 enterprises in China. The notion of enterprise knowledge graph is also used by IBM (Hu et al., 2014) and other information technology companies when they apply the knowledge graph technology to specific enterprises.

3. FRAMEWORK OF THE MULTIPLE DATA SOURCE-BASED METHOD

3.1 Lifecycle of domain knowledge graph

The lifecycle of domain knowledge graph is shown in Figure 1.; which mainly has five processes, namely, ontology definition, knowledge extraction, knowledge fusion, knowledge storage and knowledge application respectively. The first three processes are related to knowledge establishment, which will be elaborated in the following sections.

3.2 Framework

As shown in Figure 2., there are three major steps, namely, Ontology Definition, Knowledge Extraction, and Data Fusion with Entity Merging. The process is data-driven, which means we will use different methods for different data sources; and to be specific, whether the D2R Transformation step or the Information Extraction step should be initiated depends on the type of data source. Meanwhile, the process is also iterative, during which the basic knowledge graph are established first, and new iteration with the input of new data sources are initiated afterwards. Usually, multiple sources are used in each iteration; and those structured data sources are used in the first place.
Figure 2. The Framework of Domain Knowledge Graph Construction Method

4. ESTABLISHING PROCESS OF KNOWLEDGE GRAPHS

4.1 Ontology Definition

Most general knowledge graphs, such as, DBpedia and YAGO, are built in a bottom-up manner so as to realize wide coverage of cross-domain data, and while manually-established knowledge graphs are worked out in a top-down fashion. In our framework, we adopt a mixed approach (both top-down and bottom-up) for establishing our packaging knowledge graph to ensure the data quality and achieve a stricter schema. This approach is quite useful when the domain schema is complex; even though there is an access to methods for automatically extracting knowledge at schema-level, such as taxonomies and class definitions, from the Internet and databases. We manually design or extend the schema of the knowledge graph, since the subject of schema is changed when new data sources are added. At the first iteration, the knowledge graph includes the following basic concepts, namely, "packaging knowledge point", "Company", "Product", "Organization", "Patent", "Paper" and "Event". Major relations include "has product", "upstream", "downstream", "has patent", and "executive".

4.2 Knowledge Extraction

4.2.1 D2R for Structured Data

We obtained the original data tables from existing application systems and took three steps to transform RDBs into RDF, including table splitting, basic D2R transformation by D2RQ and post processing. We have several problems to figure out: first of all, these tables do not follow the basic design principles of databases (e.g., BNCF12); and some tables may contain multiple entities and relations. Furthermore, these tables may contain n-ary relations, or the same table column may refer to different entity types. Therefore, we split the original tables virtually into smaller ones, namely, atomic entity tables, atomic relation tables, complex entity tables and complex relation tables (e.g., relation table with meta properties) to make them be easier for comprehending and handling. An atomic entity table corresponds to a class, and an atomic relation table corresponds to a relation instance, where the domain and range are two different classes. We have used D2RQ to transform the atomic entity table and the atomic relation table into RDF; and wrote special programs to deal with complex relation tables, which may require meta property mapping, conditional taxonomy mapping, etc.

4.2.2 Wrapper for Semi-Structured Data

The wrapper (Muslea et al., 2001) in data mining is a program that extracts the contents of a particular information source, and translates them into a relational form. Many web pages present formatted structured data, including, telephone directories, product catalogs, etc. for browsing using HTML language. We use the wrapper to extract knowledge from semi-structured data, such as, HTML pages, office document, and other data with structured information. There are mainly four steps: firstly, we set the data source, then define the extracting target; thirdly, we define the extracting template (usually some rules), and finally execute some closeout processes.
4.2.3 Information Extraction from Text

There are various information extraction tasks; and we have drawn the tasks including entity recognition, concept inducing, relation extraction, and event extraction in this paper. We extracted targets of different type against the background that most information extraction research work focuses on extracting one particular kind of targets, such as entities or relations between entities. The extraction strategy varies significantly when different data sources and extraction targets are handled. We adopted a multi-strategy learning method to extract multiple types of knowledge from various data sources. The whole process is as follows:

a) Extracted the entities and attribute value pairs and product information of enterprises from "www.cpta.org.cn", "www.pack.cn" and "www.chndesign.com" respectively using HTML wrapper. We also used some classical NLP tools, such as, NLPIR, LTP and Stanford NLP to extract the name of people.

b) Extracted attribute value pairs from infoboxes of encyclopedic sites by HTML wrapper, as there are sometimes many web pages describe the same entity, and we evaluated corresponding information based on the pages' publishing time to determine which data source is up to date and correct.

c) With regard to the relation extraction from free texts, we usually apply pattern-based method first, and then try to induce patterns automatically using some seed annotations in sentences to learn the patterns. The quality of the extracted information largely depends on the number of annotated sentences; meanwhile, manual annotation requires too much effort. Thus, we defined a set of Hearst patterns to extract data from free texts of encyclopedic sites. We also used a remote learning method to minimize the manual work, the process of which is as follows: the extracted triples were fed as seeds automatically to label free texts; first of all, we collected sentences that contain seeds and labeled these sentences; then we generated extraction patterns based on these annotated sentences; finally, we used the generated extraction patterns to extract new information from other free texts. The additional information extracted is added to the seeds for bootstrapping. The whole process is iterated until no new information can be extracted.

d) In term of event extraction, there are three steps. Firstly, we selected the seed words, which summarize the events, for example, “product launch” and “strategic cooperation”. Then, we found those sentences containing the seed words, and considered them as candidate events. Finally, we extracted the elements of events separately with the methods described in c).

4.3 Knowledge Fusion

The knowledge fusion tasks, such as, entity merge and instance matching, are usually very difficult; and the establishing process of general knowledge graphs is not so reliable, with an accuracy of lower than 0.5 usually, which is unacceptable in domain knowledge graph. Therefore, we incorporated the knowledge fusion process into the knowledge extraction process in our process, and also defined the knowledge fusion rules to complete the knowledge fusion task while extracting knowledge with D2R and Wrappers.

5. EXPERIMENT

5.1 Data Analysis

The data sources adopted are as follows:

a) Structured data sources, including packaging industry taxonomy defined by industry experts, packaging products and enterprises stored in relation databases, and several patent and paper databases.

b) Semi-structured data sources, the first type is online encyclopedias, such as, Chinese-Wikipedia, HudongBaike and BaiduBaike, and we only referred to these pages related to packaging industry with filters. The second type is some online domain websites, including "www.cpta.org.cn", "www.pack.cn" and "www.chndesign.com".
c) Unstructured data sources, text from all data sources.

5.2 Knowledge Graph Establishing Result

As a whole, the established packaging knowledge graph contains about 2,000 concepts and hyponymy among them, over 10,000 entities and their attributes, more than 10,000 relations among the entities; and the average precision of the knowledge is above 0.9.

The defined concepts and taxonomy is presented in Fig. 3 (a); while (b) shows the visualization results.

![Packaging Knowledge Graph Result](image)

Figure 3. Packaging Knowledge Graph Result

6. CONCLUSION

This paper illustrates the whole process to establish a knowledge graph from scratch, including ontology definition, knowledge extraction and knowledge fusion; and the detailed methods of each process are also elaborated. We have established a packaging knowledge graph with 2,000 concepts, 10,000 entities, and more than 10,000 relations with this method.

In the future, we will continuously enrich the knowledge with more data sources; improve the related methods; and exploit more applications based the packaging knowledge graph.

ACKNOWLEDGEMENTS

This paper is funded by the Hunan Provincial Natural Science Foundation of China under grant numbers 2016JJ 5035; the Hunan Provincial Science and Technology Project (key project) under grant numbers 2015TP1003; the Scientific Research Project of Hunan Province Department of Education under grant numbers 14A037 (key project) and 16C0481; the Project of China packaging federation under Funding Support Number 2014GSJJWT0 01KT001, 2014GSJJWT001KT010, 17ZBLWT001KT010; the Scientific Research Project of Zhuzhou City Department of science and technology bureau under grant numbers 2016-68-12; The Project of National Packaging Advertising Research Base and Hunan Packaging Advertising Creative Base under Funding Support Number 17 JDXMA03.

REFERENCE