Constructing government procurement knowledge graph based on crawler data

Yu Yaozu1, Zhang Jiangen2

1School of Civil Engineering, Tangshan University, Tangshan 063000, Hebei, China
2Information Center, Department of Finance of Hebei Province, Shijiazhuang 050051, Hebei, China
yzjlyu@126.com, zhangjg@hebcz.cn

Abstract. In the public information of government affairs, the public information of government procurement is more and more timely and complete, which makes it possible for the public to have a comprehensive understanding, in-depth analysis and intelligent use of government procurement information. This paper proposes a process for constructing government procurement knowledge graph: obtaining the bid-winning announcement information through web crawler technology, constructing the government procurement ontology by combining the professional knowledge of government procurement and professional standards and norms, and using Protégé, D2rq, Neo4j and other software to realize ontology instantiation, entity extraction, N-Triples storage. The establishment of knowledge graph for government procurement establishes a strong, convenient and efficient foundation for the public to inquire government procurement information, count government procurement content and analyze the relationship between government procurement.

1. Introduction

Within the implementation of the requirement about information disclosure in the whole process of government procurement, and the establishment of government procurement information (GPI) platform, positive progress has been made in the disclosure of GPI. The public has more and more convenient ways to obtain GPI data, and the amount of available data is also increasing quickly. Analyzing the GPI data can find a large number of useful value, which can help suppliers accurately formulate bidding plans and increase the probability of winning the bid; the agency fully analyzes the procurement content and improves the service level; the procurement unit reasonably sets procurement indicators and improves procurement quality; the public quickly find clues about procurement violations and enhance the quality and effectiveness of supervision. However, in general, the public can only access GPI through web browsing. This has brought great difficulties to the comprehensive utilization and the enhancement on information value of GPI data. On the one hand, it is difficult for the public to obtain a large number of GPI data quickly, and on the other hand, it is difficult for the public to query, summarize, statistics and analyze the unstructured GPI data.

A web crawler is a program or script that automatically grabs information from the Internet according to certain rules, which can capture GPI data from the GPI website and save them to the local database. Thus, it solves the difficulty of obtaining large amounts of GPI data. Knowledge graph is an expression of knowledge, which can be used as the basis for knowledge retrieval, correlation, analysis, synthesis and reasoning, and provides strong theoretical support for the storage and application of
knowledge. In order to improve the application value of GPI data, this paper proposes a solution to construct government procurement knowledge graph (GPKG) based on web crawler data. The solution gets GPI data by web crawler technology, utilizes the expertise of field experts, establishes the scheme for automatic constructing process of GPKG. The results of the solution can expand the depth and breadth of the application of GPI.

2. Related works

2.1. Python web crawler
In today's information age, those who get the data get the world. As an important means to obtaining Internet data, web crawler is widely concerned and used in various applications. However, the use of a large number of web crawlers has caused considerable problems for some Internet sites. For example, using web crawlers to "swipe tickets" on the concert website will occupy a large amount of site resources, or even bring down the server. However, because of its higher success rate of swiping votes, it has provided the scalpers with a living soil. Therefore, some websites take measures to prevent web crawlers, such as verification codes, IP restrictions, JS rendering, etc. Web crawler technology is constantly developing under the environment of explosive growth of data volume and restriction of anti-crawler, which promotes the birth of a large number of application frameworks.

Python is a cross-platform computer programming language. It is a high-level scripting language that combines interpretation, compilation, interactivity, and object orientation [1]. Due to its features of open source, simple, easy to learn, portable, extensible and easy to maintain, many kinds of web crawler frames have been developed, such as Scrapy, Crawley, Portia, Newspaper and Python-Goose. Among them, Scrapy is the most popular open source framework, which can crawl Web pages and extract structured data quickly, simply and efficiently [2].

2.2. Knowledge graph
In 2012, Google proposed the concept of knowledge graph [3] for improving the quality of its search engine. Knowledge graph connects the massive information in the real world to form a network, so it is a kind of semantic network [4], which can be represented by graph structure. The nodes and edges of the structure represent the concepts (entities) and the relationships between concepts (entities) respectively. Knowledge graph is very helpful for understanding and structuring, and builds a good communication bridge between human understanding and computer operation in massive information area. Therefore, it has quickly become the focus of artificial intelligence research and has been rapidly developed in semantic search, intelligent recommendation, decision support and other computer engineering field [5-6].

Knowledge graph is composed of knowledge sets, including abstract knowledge (ontology layer) and concrete knowledge (entity layer). Abstract knowledge represents concepts and attributes abstracted from specific knowledge content, usually expressed in RDFS or OWL[7] standard formulated by W3C. Concrete knowledge represents entities (objects that exist in reality and are distinct from other objects), relationships between entities, or properties of entities, in the form of triples (entity-relationship-entity) or (entity-property-attribute values), usually expressed in the W3C RDF [8] standard.

3. Overall architecture design
According to the data flow in the construction and application process of the knowledge graph, and the layered design ideas, the overall architecture design of GPKG application is divided into four levels: data acquisition layer, data storage layer, knowledge graph layer and business application layer, as shown in Fig.1. Among them, data acquisition layer, data storage layer and knowledge graph layer are the construction process of GPKG, which is the focus of this paper. The business application layer is the application process of GPKG, which is not mentioned in this paper.
Data acquisition layer: It is the basis for constructing GPKG. On the one hand, it collects the information of government procurement announcement on the Internet and converts it into structured data; on the other hand, it pre-processes the data, such as cleaning abnormal data, normalizing data format, extracting data dictionary, etc.

Data storage layer: It provides technical support for data persistent storage, rapid retrieval and processing transformation for data acquisition layer and knowledge graph layer. It includes a relational database for storing collected and transformed structured data and a graph database for storing knowledge graph data.

Knowledge graph layer: It is the core part of constructing GPKG. Ontology construction refers to the construction of government procurement ontology from top to bottom by government procurement experts, who combines government procurement expertise and the ontology construction process. Entity extraction refers to the progress of extracting government procurement entities from government procurement data according to the established ontology;

Business application layer: It is the final business foothold of the government procurement knowledge graph, used for human-computer interaction. Information display is used to display government procurement content to users in a visual manner; information retrieval is used to semantically retrieve government procurement information for users.

4. System implementation

4.1. Collection and pre-processing of government procurement data

The government procurement data collected to construct GPKG in this paper is the announcement information of winning the bid on the government procurement website of Hebei Province. The web page of bid-winning announcement is http://www.ccgp-hebei.gov.cn/province/cggg/zhbgg/, and an example of announcement details is http://www.ccgp-hebei.gov.cn/xt/xt_wx/cggg/zhbggAAAA/201909/t20190925_1114497.html.

Through the analysis of the page source files, it can be found that most of the contents to be crawled are contained in the Html static code, such as the announcement list in the bid-winning announcement page, the announcement code, the way of procurement, the project name, administrative divisions, the name of the purchaser, the agency, review experts and most other contents in the announcement details. But there are still a few bits of data rendered by JavaScript code, such as the list of winning bids in announcement details. For the static data that directly responds to the web, this article uses the Scrapy framework to crawl according to regular expression matching, while for the data generated by JavaScript rendering, this article relies on the Selenium library that can simulate the running behaviour of the browser. Finally, the crawled data is stored in the MySQL database, and the main fields of the table structure designed according to the content of the crawling government procurement bid announcement are shown in Tab.1.
Tab. 1 The main fields of the table structure

| No | Field name               | Field identifier | Type       |
|----|--------------------------|------------------|------------|
| 1  | Announcement code        | ID               | varchar(32)|
| 2  | Purchase’s type          | Purchase_type    | varchar(50)|
| 3  | Administrative divisions | Adm_Div          | varchar(9) |
| 4  | Project’s code           | Project_Code     | varchar(50)|
| 5  | Project’s name           | Project_name     | varchar(100)|
| 6  | Purchaser’s name         | Purchaser_name   | varchar(50)|
| 7  | Agency’s name            | GP_Agency_name   | varchar(50)|
| 8  | Notice date              | Notice_date      | varchar(10)|
| 9  | Item                     | Item             | varchar(10)|
| 10 | Supplier’s unified credit code | Uscc | varchar(19)|
| 11 | Supplier’s name          | Supplier_Name    | varchar(50)|
| 12 | Bid amount               | AMT              | double(15,2)|
| 13 | Exporter’s Name          | Exporter_Name    | varchar(20)|

The data obtained by the web crawler is stored in a table, which is not conducive to the construction of the knowledge graph entity in the later period. Therefore, this article makes the necessary cleaning and conversion of the crawled data. The first is to split the joint bidder into independent individuals. The second is to extract and generate several types of data dictionaries, including procurement methods, purchasers, administrative divisions, review experts, agencies, item classifications, etc. according to the government procurement ontology (see section 4.2). Each type of data dictionary contains at least two basic data items: ID and name. The third is to construct the relationship table between tables, such as the relationship between the purchaser and the procurement project, the agency structure and the procurement project.

4.2. Ontology design of government procurement

OWL has rich expressive ability and strong expansibility, and can completely express concepts, attributes, relationships and other knowledge. Besides it has high compatibility and compatibility, and can adapt to the mainstream software environment. Therefore, this paper adopts OWL as the expression language of government procurement ontology.

![Fig. 2 Structure of the GPKG.](image)
Combining the professional knowledge of government procurement experts, a seven-step method is used to construct the government procurement ontology from top to bottom. The process is described as follows:

1. Confirming that the ontology constructed in this paper is applied in the field of government procurement and used for the intelligent analysis of bid-winning information in government procurement;
2. Existing available government procurement domain ontology is found in literature [9], and parts of them are referenced;
3. Enumerate the important terms of government procurement and define the ontology of government procurement;
4. Define the hierarchy of government procurement ontology from top to bottom;
5. Define the attributes of the government procurement ontology;
6. Define constraints of ontology attributes (definition domain and range);
7. Protégé is used to instantiate the defined ontology.

Fig.2 and Fig.3 respectively show the structure and knowledge description of the GPKG.

4.3. Extraction and storage of government procurement entities

D2RQ is a tool for converting structured data into RDF, and the conversion process is divided into two steps. In the first step, each table with a primary key is mapped to one ontology where the primary key value is mapped to the unique identifier of the conceptual entity, the other table attributes are mapped to the concept attribute; and each table with foreign key constraint specifications is mapped to object attribute. In the second step, the table data is converted to RDF according to the mapping relationship determined in the first step. Based on the actual situation, this article makes necessary modifications to the mapping file automatically generated in the first step. Part of the mapping file content and part of the RDF content are shown in Fig.4 and Fig.5.
Fig. 5 Part of the RDF content

In order to improve query efficiency and data reasoning ability, this paper stores RDF data to Neo4j, a graph database with the highest utilization rate.

5. Conclusions

In this paper, knowledge graph is applied to the field of government procurement, and the winning announcement information of government procurement on the Internet is obtained by means of Scrapy framework and Selenium library. The government procurement ontology is constructed by seven-step method and instantiated by Protégé. RDF is extracted by D2RQ and stored in Neo4j database. The construction of GPKG has laid a solid foundation for the in-depth utilization of GPI. In the future work, on the one hand, we’ll expand the coverage of the GPKG (e.g. bidding announcements, complaints and reports); on the other hand, we’ll increase the development of business applications, and continuously improve the ability of intelligent analysis and decision support for the government procurement industry.

References

[1] Cui QingCa. Python 3 Web crawler development combat[M].Beijing:Post&Telecom press Co., LTD,2018(in Chinese).
[2] Du Penghui, Qiu Jiyang, et al. Design and implementation of Web crawler based on Scrapy [J]. Electronic Design Engineering, 2019, 27(22):120-123(in Chinese).
[3] Amit Singhal. Introducing the Knowledge Graph: Things, Not Strings[EB/OL].(2012-05-16)[2020-08-19]. https://www.blog.google/products/search/introducing-knowledge-graph-things-not/.
[4] Jeff Z. Pan, Guido Vetere, Jose Manuel Gomez-Perez, et.al. Exploiting Linked Data and Knowledge Graphs in Large Organisations[M]. Springer,2017.
[5] ZhouYuanchun, WangWeijun, et al. A survey on the construction methods and applications of sci-tech big data knowledge graph [J].Sci Sin Inform,2020,50(7): 957-987(in Chinese).
[6] Qin chuan, Zhu Hui, et al. A survey on knowledge graph-based recommender systems[J]. Sci Sin Inform,2020,50(7): 937-956(in Chinese).
[7] Learn OWL and RDFS [EB/OL].[2020-08-19].https://www.cambridgesemantics.com/blog/semantic-university/learn-owl-rdfs/.
[8] R. Cyganiak, D. Wood, Markus Lanthaler.RDF 1.1 Concepts and Abstract Syntax[EB/OL].(2014-05-25)[2020-08-19]. https://www.w3.org/TR/rdf11-concepts/.
[9] Sun Han. Knowledge Map Construction and Visualization System Design Based on Public Resource Transaction. [D].Taiyuan, North university of China,2018(in Chinese).