Computer and Information Science; Vol. 11, No. 3; 2018
ISSN 1913-8989 E-ISSN 1913-8997
Published by Canadian Center of Science and Education

Design and Implementation of China HowNet Knowledge Map Generation Module Literature

Liu XiangWei[1]

[1] Luoyang campus of the University of Information Engineering, China

Correspondence: Liu XiangWei, Luoyang campus of the University of Information Engineering, Luoyang Henan 471003, China. Tel: 86-158-2490-9625. E-mail: liuxwletter@163.com

Received: April 20, 2018 Accepted: May 2, 2018 Online Published: July 28, 2018

doi:10.5539/cis.v11n3p89 URL: https://doi.org/10.5539/cis.v11n3p89

Abstract

Based on the analysis of the construction technology of the existing literature knowledge map, a China HowNet (CNKI) method for automatic generation of literature knowledge map. This paper according to the qualification to collect related data of the web page; Second original web data analysis, and design knowledge map structure; By determining the entity, after extracting properties and associated knowledge; And knowledge to the automatic mapping for secondary image database of Cypher code; Finally build the direction of our school language information processing of CNKI, included the literature of knowledge map. Automatic and efficient knowledge map construction in the evolution of discipline has great practical significance, methods and conclusions of this paper can provide literature research, the researchers in the field of knowledge map building and information visualization to provide enlightenment and reference.

Keywords: literature research, knowledge graph, information visualization, graph database

1. Introduction

Scientific and technological innovation is the driving force of social development, and the level of science and technology is the hard index to measure the soft power of the country. With the development of science and technology in China and the huge branches of science and technology, the academic literature of its by-products has been greatly expanded. The academic literature is not only the description and record of scientific research achievements, but also the epitome and auxiliary evidence of the development of the field discipline. Bibliometrics has been widely used and practiced in the literature since it was proposed by British scholar Alan pritchard in 1969. The visualization of knowledge mapping is applied to the research of literature to demonstrate the structure, relationship and evolution of the subject. At present, the theory of literature metrology and natural language processing, data mining technology, combining with term extraction, associated excavation, construction of the network and subject classification processing auxiliary statistical modeling, and subject knowledge map construction has become a popular method of literature research. But existing research more focus on the automatic extraction of knowledge, and knowledge map construction is done manually, more waste of resources, therefore, the text put forward a kind of automatic generation method of knowledge map based on CNKI literature, and on the basis of constructing the direction of literature language information processing knowledge map in our school.

2. Module Design Overview

2.1 Module Development Background and System Status

Literature research has realized the analysis of literature from qualitative to quantitative analysis, which is the routine method and important means to track the history of subject evolution, to explore the frontier development of the subject, and to predict the development trend of the subject. Literature research is based on bibliometrics, which can grasp both the macro and the details of the changes. Therefore, the analysis and prediction of literature research in various fields are endless. To this end, combining the Java language and Neo4j graph database to develop the academic literature research system is of great application value and practical significance.

This module is the basis of the academic literature research system module and the core module, the main work includes the CNKI literature data collection, the design of knowledge map structure, literature knowledge extraction and Cypher code automatically generated, and so on. This module provides data support and technology foundation for the system, and completes the leap of knowledge map from scratch. On the basis of the development of this module, it is easy to realize the query, display and other system functions of the knowledge map.

2.2 Module Workflow

According to functional requirements, the module is divided into the following working steps:

(1) Data collection of CNKI literature: the original data of relevant web pages were obtained from CNKI website according to the qualification criteria of the input;

(2) Knowledge map structure design: analyze the data content of literature web page, design the node, attribute and correlation of knowledge map;

(3) Knowledge extraction of CNKI literature: according to the structure design of the knowledge map, processing the data format of the literature webpage and extracting the literature knowledge;

(4) Cypher code is generated automatically: based on the construction rules of knowledge graph, the Cypher code of knowledge map is automatically generated by using the extracted knowledge;

(5) Knowledge map display: running Cypher code on the neo4j graph database platform to complete the construction and display of knowledge map.

3. Introduction to Core Technologies

3.1 CNKI Literature Data Collection

China hownet (CNKI) is the largest and most widely used in domestic academic database, although you can't include all the journals and literature information, but to analyze the data of CNKI, generally reflect the whole situation of academic literature in China.

This module supports users to customize the qualified conditions of literature collection according to the object and purpose of the study, such as author, subject, publication journal, publication time and combination. The module searched the relevant literatures on CNKI site according to the qualification criteria, and determined the collection scope. According to the research experiment, the CNKI website has set up a number of anti-data crawling security measures due to the working mechanism of paid download. First of all, the URL list of the retrieved literature is not directly displayed in the source code of the webpage. Instead, it is hidden in the form of variables. The specific value of the URL is embedded in the website database, which is not seen by the visitors. Second, the keywords submitted by the user request are only used for identification, and the server dynamically generates digital sequences based on the keywords, as the credentials for accessing the resources. It is not only the technical difficulties but also the infringement of intellectual property rights.

3.2 Structure Design of Knowledge Map

Literature research of knowledge map with literature as the core object, in addition to the basic information of the clear reflection of literature, also should from the local to the overall integrity of knowledge of give attention to two or morethings, which can show the connection between the literature. Accordingly, the title, publication unit, publication time, author, summary and key words are the core elements of the literature.

According to the system functional requirements and the modeling specification of Neo4j, the node classification and its properties are described as follows:

(1) Root node r1, the label is root: the root node is unique, the name attribute value is timeLine, as the query handle of the time axis;

(2) Literature node ti, the label is title: I is the serial number created for the document node, the title attribute value is the literature title, the summary attribute value is the literature summary, and the keywords attribute value is the collection of literature keywords.

(3) The author node ai, labeled as author: I is the number created by the author node, and the name attribute value is the author's name;

(4) Time node yi, the label is year: I is the serial number created by the time node, and the value attribute value is the time year;

(5) Publish unit node PI, labeled publish: I is the serial number created by the publishing unit node, value attribute value is the name of the publishing house;

According to the relationship between the literature information and the subsequent query display and other functional requirements, the relationship between nodes is created, and the classification is as follows:

(1) Year correlation between root node and time node;

(2) The document node is associated with written_by of the author node;

(3) The literature node is associated with published_by of the publishing unit node;

(4) Association of the literature node to the published_in time node;

3.3 CNKI Literature Knowledge Extraction

Literature knowledge extraction is the key step of the module, and the accuracy of knowledge extraction directly determines the mapping level of Cypher code and the construction quality of knowledge map. The key function of the core function getInfo is set as follows:

Pattern Ptitle："<span id=\"chTitle\">(.+)</span>"；

Pattern Pauthors： "sfield=au&amp.+>(.+)</a>"；

Pattern Ppost： "sfield=inst&amp.+>(.+)</a>"；

Pattern Psummary："<spanid=\"ChDivSummary\".+>(.+)</span><span>"

Pattern Psummary1：" 【Abstract】 <span>(.+)</span>"；

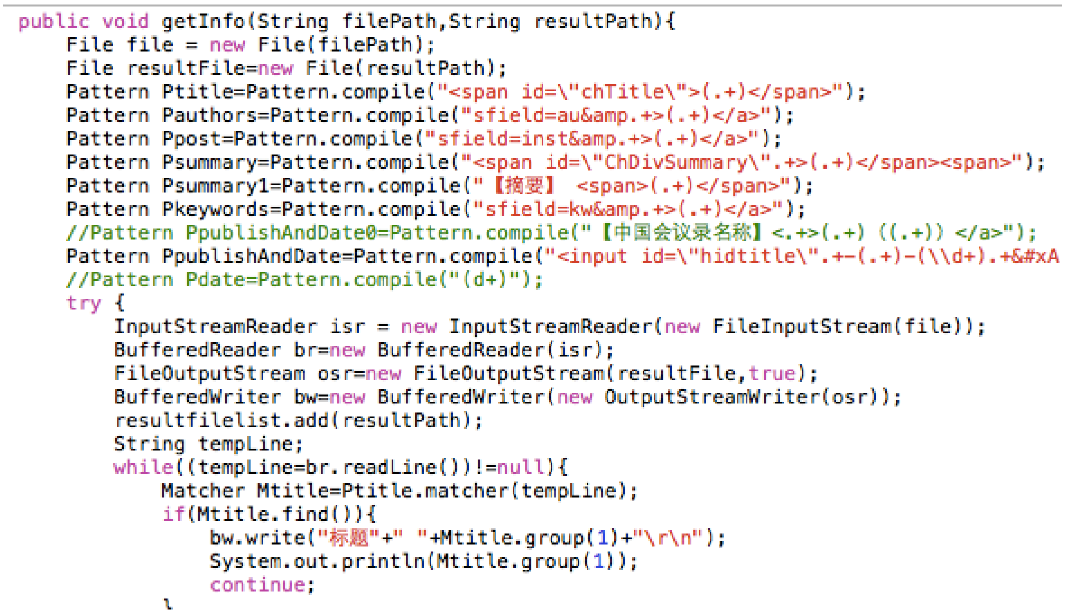
Pattern Pkeywords："sfield=kw&amp.+>(.+)</a>"；

Pattern PpublishAndDate："<input id=\"hidtitle\".+-(.+)-(\\d+).+&#xA;

The model corresponds to the title, author, author unit, page format 1 summary, page format 2 abstract, keyword, publication unit and publication time. In this paper, two matching modes are set up for the abstract because of the different webpage formats caused by the different sources of CNKI. As the publishing unit and publication time in the web format are stored in the same row, the matching extraction of the two is merged into one pattern.

Efficient multithreading concurrency to accommodate large scale document processing. The absolute address of the document to be processed is stored in the filelist table, and each document is processed by a single thread to avoid reading and writing conflicts and ensuring the accuracy of the results. Multiple threads process multiple documents in parallel, improving the efficiency of systemoperation.

The core code is as follows:



```
public void getInfo(String filePath,String resultPath){
    File file = new File(filePath);
    File resultFile=new File(resultPath);
    Pattern Ptitle=Pattern.compile("<span id=\"chTitle\">(.+)</span>");
    Pattern Pauthors=Pattern.compile("sfield=au&amp.+>(.+)</a>");
    Pattern Ppost=Pattern.compile("sfield=inst&amp.+>(.+)</a>");
    Pattern Psummary=Pattern.compile("<span id=\"ChDivSummary\".+>(.+)</span><span>");
    Pattern Psummary1=Pattern.compile(" 【摘要】 <span>(.+)</span>");
    Pattern Pkeywords=Pattern.compile("sfield=kw&amp.+>(.+)</a>");
    //Pattern PpublishAndDate0=Pattern.compile(" 【中国会议录名称】<.+>(.+) ((.+)) </a>");
    Pattern PpublishAndDate=Pattern.compile("<input id=\"hidtitle\".+-(.+)-(\\d+).+&#xA
    //Pattern Pdate=Pattern.compile("(d+)");
    try {
        InputStreamReader isr = new InputStreamReader(new FileInputStream(file));
        BufferedReader br=new BufferedReader(isr);
        FileOutputStream osr=new FileOutputStream(resultFile,true);
        BufferedWriter bw=new BufferedWriter(new OutputStreamWriter(osr));
        resultfilelist.add(resultPath);
        String tempLine;
        while((tempLine=br.readLine())!=null){
            Matcher Mtitle=Ptitle.matcher(tempLine);
            if(Mtitle.find()){
                bw.write("标题"+" "+Mtitle.group(1)+"\r\n");
                System.out.println(Mtitle.group(1));
                continue;
            }
```

3.4 Automatically Generates Cypher Code

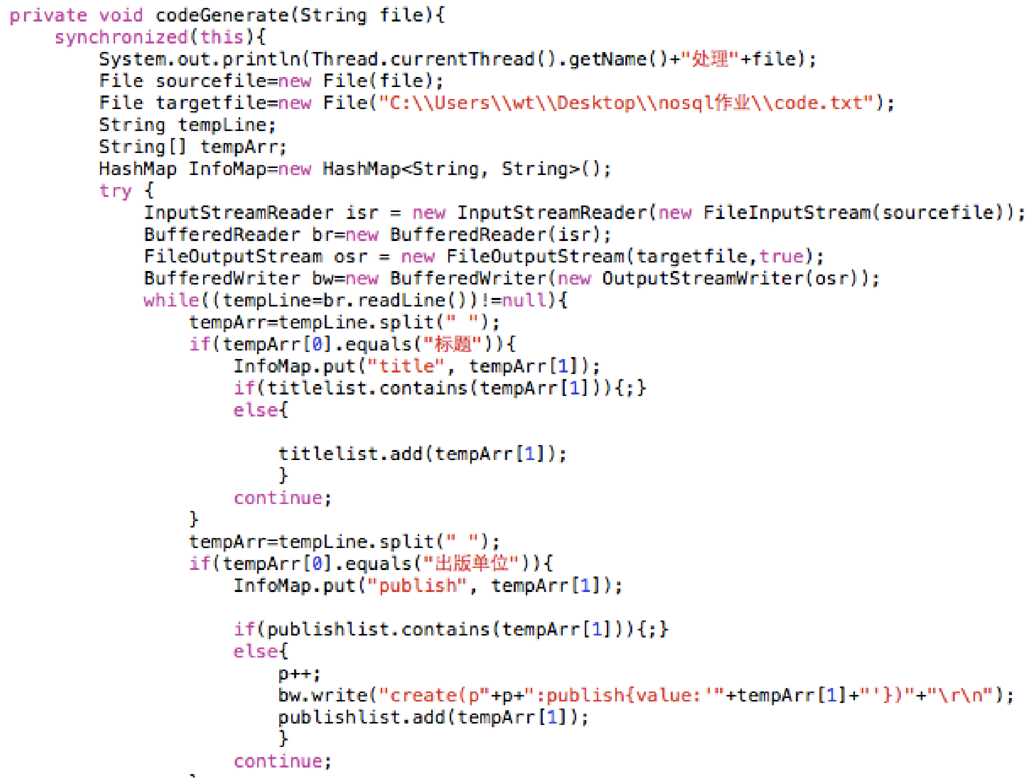
Cypher is a simple and concise graph database query language, which supports the neo4j graph database, which is very suitable to describe the graph programmatically in a precise manner. Cypher code is automatically based on knowledge map nodes, attributes, relationships and other structural design and literature knowledge extraction results.

The mapping format of the code is as follows:

```
create(pi:publish{value:'PPP'})
create(yj:year{value:'YYY'})
create(r1)-[:Year]->(yj)
create(am:author{name:'NNN'})
create(tn:title{name:'TTT',summary:'SSS',keywords:'Key1  Key2 ……'})
create(tn)-[:published_by]->(pi)
create(tn)-[:published_in]->(yj)
create(tn)-[:written_by]->(am)
```

Among them, p, r, j, a, t, for the knowledge map nodes, i, j, a, m, n to the corresponding node number, due to the author of a paper may have one or more, so the author node creation code: the create (am: author {name: 'NNN}) and the relationship between literature and the author create code: the create (tn) - [: written_by] - > (am) ,up to a dynamically generated according to the actual situation. ArrayList: Titlelist, authorlist, datelist and publishlist are created for the literature, author, publication time and publication. Node corresponding ArrayList table will be created before the find, to determine whether a node has been in existence in figure database, if there is no criterion to create nodes, or skip the creation process, to avoid repetition and create errors.

The mapping format of the code is as follows:



```
private void codeGenerate(String file){
    synchronized(this){
        System.out.println(Thread.currentThread().getName()+"处理"+file);
        File sourcefile=new File(file);
        File targetfile=new File("C:\\Users\\wt\\Desktop\\nosql作业\\code.txt");
        String tempLine;
        String[] tempArr;
        HashMap InfoMap=new HashMap<String, String>();
        try {
            InputStreamReader isr = new InputStreamReader(new FileInputStream(sourcefile));
            BufferedReader br=new BufferedReader(isr);
            FileOutputStream osr = new FileOutputStream(targetfile,true);
            BufferedWriter bw=new BufferedWriter(new OutputStreamWriter(osr));
            while((tempLine=br.readLine())!=null){
                tempArr=tempLine.split(" ");
                if(tempArr[0].equals("标题")){
                    InfoMap.put("title", tempArr[1]);
                    if(titlelist.contains(tempArr[1])){;}
                    else{

                        titlelist.add(tempArr[1]);
                        }
                    continue;
                }
                tempArr=tempLine.split(" ");
                if(tempArr[0].equals("出版单位")){
                    InfoMap.put("publish", tempArr[1]);

                    if(publishlist.contains(tempArr[1])){;}
                    else{
                        p++;
                        bw.write("create(p"+p+":publish{value:'"+tempArr[1]+"'})"+"\r\n");
                        publishlist.add(tempArr[1]);
                        }
                    continue;
```

4. Experimental Results Show and Analyze

4.1 Knowledge Map Building Platform Introduction

In this paper, the neo4j graph database is selected as the building platform of knowledge map. The reason is:

(1) Different from the relational database, which is used as a means of connecting tables, neo4j is a typical representative of a non-relational database, and it is also associated with both entities and associations in storage. Avoid the relational database by scale and the data from the group of the macroscopic structure complex, thin line, and not empty check logic problem, to pay attention to the associated applications such as literature research is easy to be a global analysis.

(2) Compared with the relational database and other non-relational databases, the neo4j diagram data will be limited to the part of the molecular graph, and the performance stability will not cause the efficiency bottleneck due to the expansion of the data scale; The graph database can be extended to add different kinds of connections,

new nodes and new subgraphs to the existing structure, and the construction is more flexible. Neo4j supports incremental iterative development of software, which conforms to the need for tracking research and so on, and provides elegant system maintenance management.

(3) Neo4j abstracts the entities in the domain into circles, and the associations are abstracted as directed arrows, and the arrows connect the circles to indicate the connection between the nodes. The graph database reduces the impedance mismatch between analysis and implementation, and the description of nodes, attributes and relationships is vivid and expressive.

4.2 Knowledge Map of Language Information Processing Direction Literature

4.2.1 Description of Knowledge Map

In this experiment, five leading scholars in the direction of language information processing in our school were selected as the constraint conditions for literature collection, and 104 articles were collected from CNKI. According to the module workflow, extract the literature knowledge and generate Cypher code. According to Cypher code, there are 1 root node, 104 literature nodes, 79 author nodes, 45 publishing unit nodes and 24 time nodes. There are 460 properties of each node. Create the root node to the time node 24 Year relationship and literature node to the node written_by relationship between 277 and 104 document node to time published_in relations, to the publisher literatures node 104 published_by relations.

4.2.2 Knowledge Map Display

The knowledge map of the author is shown in Figure 6. The time knowledge map is shown in Figure 7

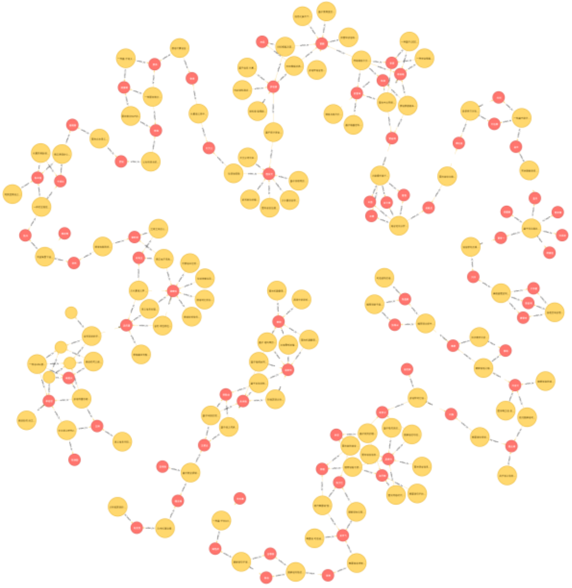


Figure 1. Author's knowledge map

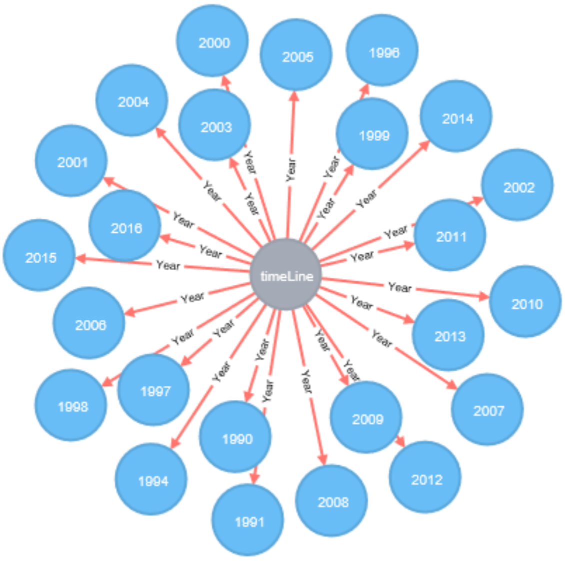


Figure 2. Time knowledge map

4.2.3 Knowledge Map Analysis

The knowledge map clearly and intuitively shows the literature publication of language information processing in our school. Diagram reflects the literature published literature published time according to the annual distribution of published literature on the annual average annual professional to undertake major projects subject for me or the next year, so the literature of the many. The schematic diagram of the literature publishing unit reflects the distribution of the literature in accordance with the publishing units, and the published units published in the literature can be regarded as the publishing institutions of my professional recognition and attention. The author has published a document showing the output of the author's literature, taking the leading figures of the language information processing of our school as the core, and the relevant scholars who have had academic cooperation with them through the submap of the joint writing of the papers.

Knowledge, knowledge and information in the knowledge map nodes contain information and knowledge about our language processing. Through Cypher query, we can return the sub graph satisfying the condition, which provides assistance for literature research.

5. Conclusion

In the era of big data arises at the historic moment under the information visualization demand of knowledge map, for knowledge has strong expressive force, make up for the deficiency of the literature metrology statistic result is too abstract, makes the literature research. In view of the existing literature research assistant system can automatically extract entities and associations, but the status quo of knowledge map building automation degree is not high, this paper proposes a against China hownet (CNKI) document automatic generation method of knowledge map. Experimental results show that the precision of a regular expression match template and scientific knowledge graph structure to ensure the quality of knowledge extraction, multithreading concurrent work mechanism and promoted module of Cypher code automatically generated rules work efficiency. The knowledge map is the basis for all subsequent query and display functions of the system, and the design and implementation of the author's literature publication timeline is the basis for system scholars to study the evolution function. In conclusion, this module lays a solid data foundation for the development of literature research systemand provides the core technical support.

Reference

Chen, F., & Zhu, T. X. (2018). A study on the scientific research of university libraries based on bibliometrics and knowledge map analysis. *Agricultural Book Intelligence Journal, 2018*(3).

Hu, Z. W., & Song, J. J. (2017). A review of domestic knowledge mapping application. *Book Intelligence, 2017*(3).

Jim, W. (2015). Emil Eifrem. Figure database. *People's post office, 2015*(2).

Onofrio Panzarin. (2014). Learning Cypher.[M]. *Packt Publishing, 2014*(5).

Tandan. (2014). Literature metrology analysis of Chinese periodical research since the reform and opening up. *Chinese Journal of Science and Technology, 2014*(11).

Tseng, Y., & Lin, Y. (2007). Text Mining Techniques for Patent Analysis [J]. *Inform Process Manag, 2007*(5), 1216-1247.

Yang, S. L., & Han, R. Z. (2013). Analysis on the application of foreign knowledge map. *Information Work, 2013*(6), 15-20.

Copyrights