Automated construction technology of the government agencies knowledge graph based on the topical crawler

Ziqi Xiong1*, Dezhi Kong1, Zhichao Xia1, Yankai Xue1, Ziyu Song1 and Peng Wang1

1 CETC Big Data Research Institute Co., Ltd., Guiyang, Guizhou, 550000, China.

*Corresponding author’s e-mail: xiongziqi@cetcbigdata.com

Abstract. Government agencies are the skeleton of the knowledge graph in the field of government affairs; it is the starting and ending point of various government affairs. This paper proposes an automated construction technology of the government agencies knowledge graph based on the topical crawler, which collects government and department information from relevant government websites and constructs the knowledge graph automatically. Sampling verification shows that the knowledge precision of this government agencies knowledge graph exceeds 96%, and it covers nearly 99% of municipal government information and more than 92% of government department information across the country. Furthermore, through the tracking of government websites, the knowledge graph of government agencies can be automatically updated. As far as our understanding is concerned, this is the first time to explore the technology of automated construction of knowledge graph in the related fields of government affairs combined with the collection system of the web crawler.

1. Introduction

The construction of knowledge graph in the field of government affairs belongs to the category of domain knowledge graph construction. In recent years, the domain knowledge graph has made great progress in practical applications. However, due to the high complexity and wide coverage of the affairs in the government, the construction of the knowledge graph in the government has been slow. Government agencies occupy the dominant position in the field of government affairs and run through all aspects of the field of government affairs. They are policy-makers and promoters as well as issuers and receivers of government affairs. This paper uses government agencies as the main framework for knowledge graph in government affairs, sorts out the context of the field of government affairs, and connects the various contents involved in government affairs. The knowledge graph of government agencies proposed in this paper automatically collects government and department information from relevant government websites and automatically constructs a knowledge graph.

The main contributions of this paper include:

- Constructed a government agencies knowledge graph that includes governments and departments, providing framework support for the expansion of the knowledge graph in the field of government affairs;
- Explored and verified the feasibility of the automated construction technology of knowledge graphs in the related fields of government affairs, and provided specific operational suggestions for the construction and expansion of knowledge graphs in other government affairs fields;

* Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
2. Related work

2.1. Topical Crawler
General web crawler, the crawling target is in the entire Internet, this kind of retrieval tool for the whole network cannot accurately provide the user's specific needs [1], so the topical crawler was born. The topical crawler can crawl purposefully according to the corresponding topic. For the topical crawler to find the URL related to the topic, it is necessary to accurately describe the topic, and then parse the URL in the webpage to determine the relevance of the webpage and the topic, according to the webpage search strategy predict the topic relevance of links and determine the URL priority, thereby reducing web crawling of unrelated topics.

The topic search strategy is the core of the topical crawler. The existing mainstream topic search strategies are based on three methods:

- The search strategy is based on the text content of web pages, such as Fish-Search [2], Shark-Search [3], and best-first-search [4], etc. This type of strategy predicts the relevance of web page’s topics to be crawled based on the topic, so that the crawler only selects URLs with higher predicted similarity for crawling, but blindly pursues the optimal relevance, which results in local optimal solutions.
- The search strategy based on the link of webpages, used links to analyze and predict the topic of webpage, finally, evaluates the priority of URL. Such as the PageRank algorithm [5] which is based on whether be linked by an authoritative website and the link's pointing to evaluate the degree of relevance, HITS algorithm [6] measures the value of the web page through two dimensions which authority and hub.
- Using only the relationship of links without considering text contents to judge the criticality of the webpage will cause the crawling result to be irrelevant to the topic, so a strategy of mixing text contents and links is generally adopted. Cai et al.[7] combined the Fish-Search and the PageRank algorithm which consider the contents and links between webpages to calculate the correlation of webpage contents and topics; Liu et al.[8] used Best-First and HITS algorithm, designed a mixed link selection strategy, and the content and link of webpages are merged, which effectively improves the topic relevance of the topic crawler; Qiu et al.[9] merged Shark-Search and PageRank algorithm, on the one hand, it used Shark-Search algorithm to calculate score of webpages’ contents, on the other it used PageRank to calculate the weight value of URL links between webpages. This method defined the importance of the webpage while making up for two traditions Algorithm flaws.

2.2. Knowledge graph
Knowledge graph is a knowledge base proposed by Google in 2012 to enhance the capabilities of search engines. It is now widely used in intelligent search, intelligent question answering, personalized recommendation and other fields. The general representation of the knowledge graph is 3-tuple. The basic form of 3-tuple mainly includes entity 1, relationship, entity 2, or concept, attribute, attribute value. Entities are the basic elements in the knowledge graph. There are different relationships between different entities, and every relationship can connect two entities [10].

At present, large-scale knowledge graphs in European and American countries mainly include YAGO, DBpedia, NELL, etc. They contain a large amount of semi-structured and unstructured data, and cover a wide range of fields. In China, knowledge graphs mainly include CN-DBpedia, zhishi.me, etc., and there are few knowledge graphs in vertical fields. Typical like, in the field of Chinese medicine, Ruan et al.[11] constructed a Chinese medicine knowledge graph which through a series of information technologies such as text extraction technology, multi-strategy learning methods and relational data to RDF conversion (D2R), and the constructed process is semi-automated; In the field of film and television, Wang et al.[12] constructed a film knowledge graph semi-automatically, they used entity

- Without the guidance of expert knowledge, save a lot of manual construction costs.
similarity calculation methods based on word2vec and TFIDF to complete the knowledge link of the knowledge graph; In the field of pets, the pet knowledge graph constructed by Yuan et al.[13], they designed the schema layer in a top-down manner, and extracted knowledge from semi-structured and unstructured data to construct the data layer, so as to construct the pet knowledge graph semi-automatically. In the field of government affairs, there is currently no complete knowledge graph of government affairs in China. We are committed to the application of big data in the field of government affairs, and through the construction of a complete knowledge graph in the field of government affairs, we can improve the government's governance and decision-making capabilities. In this paper, we propose an automated construction technique for knowledge graphs of government agencies, which can be used to supplement and expand knowledge graphs in the field of government affairs.

3. Method

![Figure 1. The process of automated construction of government agencies knowledge graph based on the topical crawler](image)

The knowledge graph automated construction technique proposed in this paper is mainly based on topical crawler. We identify specific links that meet the conditions by analyzing the structure of the web page and using the navigation information under the guidance of the administrative division base and the department regulation base. This technique traverses through the specified websites link information, and uses the data processing method to build hierarchical relationships among government departments through link information, and then construct the government agencies knowledge graph. The process is shown as Figure 1.

3.1. Data sources
The government data used in the government agencies knowledge graph comes from the website of The State Council People's Republic of China (www.gov.cn). By analyzing the website structure and data, we found that the homepage of the website provides links between all provincial governments and they also provide links between all municipal governments and departments. Through the navigation information of all levels of websites, we can build the general framework of the government and departments in the knowledge graph from top to bottom. At the same time, the websites of departments at all levels contain information such as responsibility introductions and leadership introductions, which provide attribute descriptions for government agencies in the knowledge graph.
This paper adopts the topical crawler technology, taking www.gov.cn as the entrance and combined with relevant page recognition technologies (heuristic methods, XGBoost or fastText algorithm and other technologies) to crawl the information we need from websites at all levels. Since the data sources on the Internet are constantly being updated, the web crawler technology can be used to monitor the relevant changes in the information of government departments and institutions in real time, update the various types of Internet data promptly and then update the knowledge unit in the government agencies knowledge graph.

3.2. Government homepage identification
In order to construct the government agencies knowledge graph, the establishment of government systems at all levels needs to be completed. On the one hand, the government system constitutes the skeleton of the knowledge graph; on the other hand, the crawling logic of the entire crawler system is built based on the government system. This paper uses the national administrative division information released by the National Bureau of Statistics (http://www.stats.gov.cn/tjjs/tjbz/tjyqhdmhcxhfdm/2019/index.html) as a priori knowledge to complete the lower levels of the State Council Government identification. Since there is usually a one-to-one correspondence between government names and administrative divisions (such as "the People’s Government of Sichuan Province"."Sichuan Province"), no distinction will be made below.
This paper builds an administrative division database based on the information of the national administrative division. The division library mainly uses key-value pairs to store subordinate division information. The heuristic rules are used to expand the administrative division database, which increases the mapping relationship between the division and its abbreviation. At the same time, the government that is not in the division (such as the "Management Committee") and the division that is nominally directly managed by the central government (such as the "Xiongan New Area") have been dealt with.

| Table 1. Government Homepage Identification Algorithm. |
|---------------------------------------------------------|
| **Input:** Listsup, urlcontent, url, dichtadm          |
| **Step 1:** clean urlcontent, remove spaces and suffixes|
| **Step 2:** if urlcontent in dichtadm, go to Step 3    |
|            else go to Step 5                             |
| **Step 3:** if urlcontent is divisionshort: output     |
|            else go to Step 4                             |
| **Step 4:** Judging the administrative level according to the length of Listsup |
| **Step 5:** According to the administrative level, use Listsup, dichtadm to get the area pathfull |
| **Output:** divisionfull                                 |

Listsup represents the superior government list, urlcontent represents the link text, url represents the link, dichtadm represents the administrative division database, divisionshort represents the division abbreviation, divisionfull represents the division full name, pathfull represents the division complete path. To obtain complete administrative divisions, the government home page identification algorithm proposed in this paper is shown in table 1.

3.3. Department homepage identification
Given a government homepage, its homepage may have links to its subordinate governments, or it may contain links to its jurisdiction or constituent departments. Therefore, the identification of the home page of the government department and the identification of the home page of the subordinate government are carried out simultaneously. The identification method of the home page of the government department is similar to the identification of the home page of the government, which is given a text to determine whether the link is a link to the home page of the government. The specific difference is that the government homepage identification is based on the administrative division
library, while the government department homepage identification is based on the department name and web page link characteristics.

This paper organizes the information of the provincial departments of 31 provinces (autonomous regions and municipalities directly under the Central Government) and Xinjiang Production and Construction Corps, and at the same time selects the homepage information of all departments in 33 prefecture-level cities including Guangzhou, Shenzhen, Hangzhou, Chengdu, Xi'an, Guiyang, Changsha, Xiamen, Lhasa, Urumqi, etc., and obtains the statistical results of the homepage URL of the government department is shown as Table 2.

| URL type            | Amount | End with gov.cn | End with other | Proportion of end with gov.cn |
|---------------------|--------|-----------------|----------------|-------------------------------|
| Provincial Government Department | 1229 | 1120 | 109 | 91.1% |
| Municipal government department | 1132 | 922 | 210 | 81.5% |
| Total               | 2361 | 2042 | 319 | 86.4% |

Among them, the proportion of URLs of provincial government departments ending in gov.cn is 91.1%, and the proportion of URLs of municipal government departments ending in gov.cn is 81.5%.

According to the data survey and analysis of the departments of the State Council and the departments of provinces, we extract the autonomous regions and municipalities directly under the Central Government, the regional characteristics (information of provinces, cities, etc.), agency characteristics (information of ministries, bureaus, departments, commissions, etc.) and link features of the link text (suffix information of the Link domain name), and comprehensively judge whether the obtained website link is the home link of the government department through the information above.

3.4. Department responsibility page identification and text extraction

The identification of the department's responsibility page depends on the body text of the web page. Department responsibilities are usually in the form of large text introductions, and their contents are highly recognizable. Based on this, this paper manually collected a large number of positive and negative web page text samples from the Internet, and used fastText algorithm to train a classification model to judge whether a web page is a department responsibility page.

From department homepage identification to department responsibility page identification, the conventional processing method is to traverse all web page links under the domain name of the department homepage, and then identify in turn. However, this processing method is very inefficient, because each page needs to go through the process of text extraction-word segmentation-prediction, which will become the bottleneck of the entire construction process. To solve this problem, this paper builds a link path library from the department home page to the department responsibility page link. When identifying department responsibility pages, the crawler uses this path library as a priori knowledge to guide the crawling logic, which greatly improves the efficiency of web page crawling.

Taking “The State Council People's Republic of China-the People’s Government of Sichuan Province-Science and Technology Department of Sichuan Province-Responsibility of Science and Technology Department of Sichuan Province” as an example, Finding the department responsibility page from the department homepage requires three times for page link traversals. Assuming an average number of links per page is m, the original processing needs to be executed m3 times. Assuming that the size of the link library is n and the link in the path library is l on each page, the process only needs to execute comparison m3 times and identification once. Since l<<m, the time complexity of each comparison operation is O(1).
After identifying the department responsibility page, the correct department introduction text needs to be extracted from the page. Because the body of the department's responsibility page usually contains noise, this paper uses the fastText algorithm to train another model for extracting the sentence (line) of the responsibility introduction, and then combine these lines to obtain the final responsibility text.

![Image of a knowledge graph centered on the State Council]

**Figure 2.** Sub-graphs of the government agencies knowledge graph (centered on the State Council)

### 4. Experiment and analysis

In this section, we have constructed a knowledge graph of government agencies. Figure 2 is a part of the knowledge graph of government agencies centered on the State Council. This knowledge graph mainly contains two types of entities: government and department, with a total of 5 relationships.

#### 4.1. Precision

In order to verify the precision of the knowledge graph of government agencies constructed in this paper, it is necessary to compare the government and agencies in the knowledge graph with the actual government and agencies. In view of the particularity of the constructed knowledge graph, this paper adopts manual verification to calculate the precision of the knowledge graph. We randomly selected a certain amount of knowledge in the knowledge graph and judged it by multiple people. Due to the difference in personal knowledge reserves, we provided some additional knowledge (Baidu Encyclopedia) to assist in judgment.

The evaluation results (shown in table 3) verify the reliability of the knowledge graph in this paper. The knowledge graph we constructed has achieved 100% precision in the relationship of TYPE. In addition, the SUPERIOR and DEPARTTO have also achieved about 99% precision. The main reason is that we are guided by the administrative division library, there are few errors in the subordinate relationship between governments. At the same time, the name of the department carries administrative division information. The precision of HOMEPAGE is as expected, but in actual analysis, it is found that the format of the homepage of some government departments is not standardized, there are commercial website domain names such as "xxx.com", and directory suffixes such as "xxx.cn/ABCD/". The precision of RESP is slightly worse, the main reason is that the current algorithm has not been dynamically adjusted according to the government departments, resulting in incorrect identification.
Table 3. Sampling verification results of the government agencies knowledge graph.

| Relationship  | Mean         | Sample size | Precision       |
|---------------|--------------|-------------|-----------------|
| TYPE          | type         | 300         | 100% ± 0%       |
| HOMEPAGE      | home page    | 280         | 99.29%±0.07%    |
| RESP          | Statement of right | 260     | 96.25% ± 2.5%  |
| SUPERIOR      | business superior | 400      | 99.25% ± 2.5%  |

4.2. Coverage

Table 4 shows that the government agencies knowledge graph which we constructed contains information on 2,884 governments and 9,763 departments in central, provincial, and municipal (due to connection stability issues, the information of the district and county government departments was not crawled). Similarly, since there are no comparable objects, we adopted a sampling method to verify the coverage of the extracted entities to the real data. We manually verified all provincial and municipal government information (a total of 390) and corresponding government department information. The verification method is to compare the provincial and municipal governments in the knowledge graph with the provincial and municipal governments where appearing in existing public networks (such as Baidu Encyclopedia, National Bureau of Statistics website). The verification results found that the knowledge graph constructed in this paper, relative to the real data, there are only 4 municipal governments with missing information, and the relevant recall rate is 98.97%, and the reasons are shown in Table 5; A total of 29 municipal government departments have missing information, the relevant recall rate is 92.56%, and the reasons are shown in Table 6.

Table 4. Entity information of the government agencies knowledge graph.

| Entity name   | Explanation   | Total entity |
|---------------|---------------|--------------|
| GOVERNMENT    | Government    | 2884         |
| DEPARTMENT    | department    | 9763         |
| GOVERNMENT    | Government    | 2884         |
| DEPARTMENT    | department    | 9763         |

Table 5. reasons for the lack of governments.

| Government name | Superior government       | Cause of failure                                      |
|-----------------|---------------------------|-------------------------------------------------------|
| Laiwu City      | Shandong Province         | unknown reason                                       |
| Shihezi City    | Xinjiang Uygur Autonomous Region | Xinjiang government homepage does not provide a link to Shihezi City (with link text) |
| Hefei City      | Anhui Province            | HTTP return code that the crawler cannot support      |
| Hengyang City   | Hunan Province            | Only mobile site, can not request to the computer site, the mobile site no navigation information |
Table 6. reasons for the lack of government departments.

| Reasons for crawling failure | Scale (percentage) |
|-------------------------------|--------------------|
| No information on the homepage | 68%                |
| Crawler collection rules are not perfect | 28%                |
| Javascript rendering problem | 4%                 |

The government agencies knowledge graph contains five types of relationship and 56,404 facts, see Table 7 for details. The relationships of SUPERIOR and DEPARTTO do not cover all entities, the reason for DEPARTTO is that it affects departments only. Although the SUPERIOR affects both the government and the department, the total amount of fact differs significantly from the total amount of entities of the government and the department (12647-8700=3947). The analysis found that these entities lacking SUPERIOR are mainly divided into the following three reasons: first, there is no SUPERIOR, such as central ministries and commissions, there is no higher level; The second is the lower level governments set up their own department by themselves; The third is the processing error caused by the department name. Statistical results and reasons for the missing SUPERIOR are shown in Table 8.

Table 7. Entity information of the government agencies knowledge graph.

| Relationship | area                              | Ranges          | Total facts |
|--------------|-----------------------------------|-----------------|-------------|
| TYPE         | Government, department            | class           | 12647       |
| HOMEPAGE     | Government, department            | str             | 12647       |
| RESP         | department                        | str             | 12647       |
| SUPERIOR     | Government, department            | Government,    | 8700        |
|              |                                   | department      |             |
| DEPARTTO     | department                        | Government      | 9763        |

Table 8. reasons for the missing SUPERIOR.

| Entity name                        | Explanation                        | Total entity                                                                 |
|------------------------------------|------------------------------------|------------------------------------------------------------------------------|
| Central Ministries                 | 2%                                 | National Code Administration, National Tobacco Monopoly Administration        |
| Subordinate government set up by itself | 92.5%                             | Enshi Public Resources Trading Center, Ganzhou Park Authority                 |
| Handling errors                    | 5.5%                               | Huaian Human Resources and Social Security Bureau                             |

5. Conclusion
This paper proposes an automated construction technology of the government agencies knowledge graph based on the topical crawler. This technology can construct the government agencies knowledge graph automatically, and without the knowledge of experts. The actual use also verified the availability of this knowledge graph and saved a lot of manual construction costs. Limited to resources and other factors, the knowledge map we have constructed currently only realizes the construction of departmental institutions at the prefectural level and the government at the district and county level. At the same time, there are still some deficiencies in data processing. In future work, we will Extend the knowledge graph to the next level of departmental institution construction (the data source analysis
part verifies the feasibility of this program) and the automated construction of specific government affairs, and at the same time improve the coverage of the knowledge graph by improving relevant processing techniques.

References

[1] Shkapenyuk, V. Suel, T. (2002) Design and implementation of a high-performance distributed Web crawler. In: International Conference on Data Engineering. CA, USA. pp. 357-368.

[2] Bra P, Post R. (1994) Information retrieval in the World-Wide Web: Making client-based searching feasible. J. Computer Networks & ISDN Systems, 27(2): 183-192.

[3] Hersovici M, Jacovi M, Marek Y, et al. (1998) The shark-search algorithm. An application: tailored Web site mapping. J. Computer Networks & ISDN Systems, 30(1-7): 317-326.

[4] Cho J, Garcia-Molina H, Page L. (2012) Reprint of: Efficient crawling through URL ordering. J. computer networks, 56(18): 3849-3858.

[5] PAGE.L. (1998) The PageRank citation ranking: bringing order to the web. J. Online manuscript, 9(1): 1-14.

[6] Jon M, Kleinberg. (1998) Authoritative sources in a hyperlinked environment. In: Ninth Annual ACM-SIAM Symposium on Discrete Algorithms. pp. 668-677.

[7] Cai, G. (2017) Design and implementation of topic-oriented multi-threaded web crawler. Lanzhou: Northwest University for Nationalities.

[8] Liu, Y. Li, H. (2017) The topical crawler algorithm based on fusion link structure. J. Journal of Huaqiao University: Natural Science, 38(2): 195-200.

[9] Lei, Q. Lou, Y. Min, C. (2016) Research on topical crawler based on Shark-Search and PageRank algorithm. In: International Conference on Cloud Computing and Intelligence Systems (CCIS). Beijing. pp.268-271.

[10] Xu, Z. Sheng, Y. He, L. Wang, Y. (2016) Overview of Knowledge Graph Technology. J. Journal of University of Electronic Science and Technology of China. 04 589-606.

[11] Ruan, T. Sun, C. Wang H. (2016) Construction of traditional Chinese medicine knowledge graph and its application. J. Journal of Medical Informatics. 37(4): 8-13.

[12] Weiwei, Wang. Zhigang, Wang. Liangming, Pan. (2016) Research on the construction of bilingual movie knowledge graph. J. Acta Scientiarum Naturalium Universitatis Pekinensis, 52(1): 25-34.

[13] Qi, Y. Liu, Y. Xie, Z. (2020) Semi-automated construction method of pet knowledge graph. J. Application Research of Computers, 37(1): 178-182.