A Government Policy Analysis Platform Based on Knowledge Graph

Peng Wang, Zesong Li, Zeyuan Li and Xin Fang
CETC Big Data Research Institute Co., Ltd.
Guiyang, China
e-mail: {wangpeng, lizesongcd, lizeyuan, fangxin}@cetcbigdata.com

***Abstract*—To develop government governance, government departments promote government information publicity actively in recent years. As an important part of government information, policies are published in the information disclosure guide on government's websites, which provide convenience for citizens and organizations to look up. However, when the government provides massive multi-source policy information, it also brings the problem of information overload. In view of the difficulty in the public inquiry, comprehension, and analysis of policy, this paper designed and developed a policy analysis platform based on the knowledge graph, aiming to provide information support for the government information disclosure. Employing technology of knowledge graph, the platform can process vast amounts of data. We also build a resource description system and a graph model to extract and store the feature of policy information. Then using visualizing human-computer interaction mode of exploratory analysis, we can achieve accurate and efficient policy information search, and policy analysis. The platform provides more convenience for the public to fully understand the content of policies by comprehensively combing the content, revealing policies' characteristics and mining their relationship.**

Keywords-knowledge graph; connections analysis; policy big data; knowledge extraction; knowledge graph visualization

I. INTRODUCTION

Policy in this paper refers to public policies, which are regulations and choices adopted by the government or other public authorities to solve public affairs or problems in a specific period. They are mainly expressed in government laws, regulations, decisions and actions [1].

On one hand, the better understanding and analysis of government policies can assist policymakers to adhere to governance objectives, promoting the development of our society and meeting the fundamental interests of our people; on the other hand, it can help the public to understand the government and take active participation in state governance.

Along with the disclosure of government information, a large amount of policy information is open to the public through government websites, new media, newspapers, radio, television, and other channels. However, whether the administrative organizations or the public are still facing various problems in understanding and utilizing policy information.

1) It is difficult for policymakers represented by party committees and government departments to grasp the needs of social issues in a timely manner. This leads to the lack of directivity in policy formulation, the lag of policy promulgation, and the complexity in data management.

2) Information users represented by public servants, enterprises and the general public are faced with difficulties in policy inquiries, understanding, and analysis, which impede the effective using of policy information.

3) The wide distribution and the Non-uniform format of the policy information also creates difficulties for the extraction and utilization of valuable information.

Although government departments, propaganda media, research institutions, and even consulting companies have already done a lot of work on policy information release, policy interpretation, and analysis. There are official platforms such as government websites at all levels, as well as communication platforms for news media, including Chinese Government Public Information Online, and China Knowledge Network's Party and Government Science Decision Support Service Platform (CNKI-Party and Government scientific decision Online) and other comprehensive knowledge service platforms and various policy industry consulting services. These platforms and services have established a new channel for Internet-based government and public information exchange, allowing the public to obtain authoritative policy information timely. However, the policy information is fragmented severely, the associated information is missing, and the phenomenon of information fragmentation is still outstanding. All of the information search functions are weak, and it is generally difficult to meet the needs of accurate retrieval of policy information. At the same time, subject to organizational hierarchy, departmental business responsibilities, geographic regions, and user permissions, it is difficult for users to use the website platform or consulting services to easily understand the overall understanding of policy information and the correlation between individuals. In the absence of more mature policy information automation correlation analysis tools or platforms, it still takes a lot of time and effort to deal with the manual experience.

In the era of economic globalization, to enhance the vitality of the digital economy, it is necessary to change the traditional thinking. The timely grasp of the ever-changing policy information needs to get rid of the long-term policy analysis mainly relies on manual processing. Using computer information technology to obtain policy information across organizational levels, inter-agency departments, and regions in a timely manner, use the big data platform to integrate information in real time and use policy knowledge extraction, understanding, and mapping results based on natural

978-1-7281-0831-5/19/$31.00 ©2019 IEEE

language processing technology. Presenting and other ways to solve the problem, speed up data circulation and knowledge output, and achieve the purpose of using artificial intelligence to improve productivity.

In recent years, the central government calls on local governments to build an open database. The database is open to the public and provides a document retrieval function to facilitate public access. This system, which is mainly based on the central, provincial and municipal levels, insists on the transmission of decrees and propaganda policies.

The research goal of this paper is to systematically collect policy raw data, use distributed big data management methods to clean and filter a large number of policy data with uneven data quality, and construct a resource description system for different forms of policy documents, using unified The metadata framework, after the policy feature information extraction and storage, is supplemented by a visualization module of human-computer symbiosis map interaction, and builds a policy analysis platform based on knowledge graph. In The Research of Policy Big Data Retrieval and Analysis Based on Elastic Search[2], a policy analysis system based on Elastic Search is built. However, ES only provides simple and basic searching and indexing ability. Using a graph based system would further improve the ability for analysis.

This paper include the collection of policy data, the cleansing of policy data, the construction of knowledge graph, the policy analysis and the visual presentation of policy data.

The process of policy data in our platform is shown in Fig. 1.

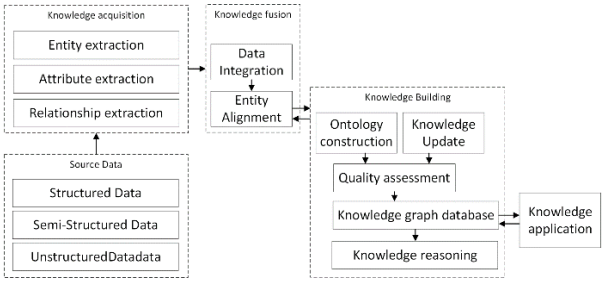

Figure 1. The data flow diagram for knowledge graph.

Based on our research, the platform provides users with convenient policy service, fast and accurate search, in-depth analysis of target policies, access to policy content, attributes, relationships, and historical development. It provides information support for government departments to formulate policies.

II. THE DESIGN OF PLATFORM

As a new method of knowledge organization and retrieval, knowledge graph provides a way for extracting knowledge from massive data and expressing semantic relatedness and relationship between various entities. We plan to extract policy information from various sources, including unstructured information (such as policy documents, pictures in policy illustrations), structured policy information, and semi-structured policy information pages to build an all-sided policy knowledge graph. The knowledge graph constructed by us mainly includes two parts: policy content graph and policy issuer graph. The construction of the two graphs are similar, and here we take the content graph as an example for detailed introduction.

We design the following system platform framework, which is divided into five layers and shown as below. The data collection layer completes the collection and aggregation of the original data; the data cleaning layer is responsible for the data's filter repetition and field completion, then the data is stored in the big data platform; in the construction layer, through the ontology establishment and concept extraction, we can get the knowledge elements and store them to form a policy knowledge graph; the relationship analysis layer serves as a bridge connecting the visual application layer and the knowledge graph construction layer, responding to user's request by reading the knowledge graph database; visualization layer is an interface for platform human-computer interaction, providing platform services conveniently and quickly.

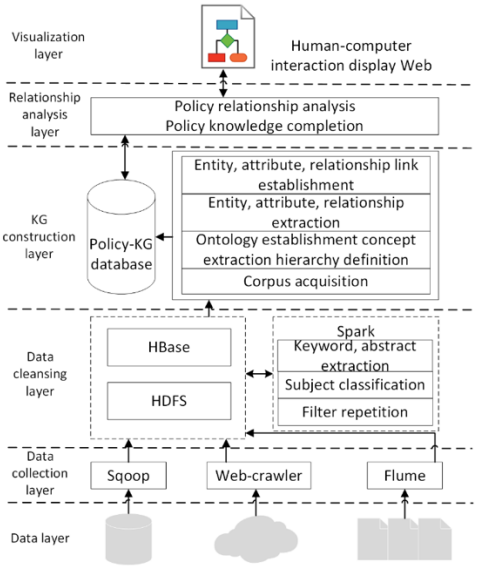

Figure 2. System structure.

A. Data Layer

The sources of data collected for policy information include government official websites, third-party integrated service platforms, and policy databases which are characterized by unstructured, semi-structured, and structured.

B. Data Collection Layer

In order to adapt to different corpus sources of policy knowledge graph, we design a variety of adaptation interfaces based on different data characteristics, Web-crawler for web policy data, Sqoop collection service for policy relational database, and Flume for document collection. The details about the techniques are introduced in Section 3 of this paper.

C. Data Cleansing Layer

The raw data is processed by the distributed real-time flow calculation engine Spark to complete the Extract-Transform-Load (ETL) process of dirty data. The main process includes the completion of attribute fields of policy documents, such as: the extraction of policies, abstract and keyword, the classification of their subjects, and filter repetition. Then the cleansed data is stored in HBase and HDFS. Technical details are in Section 4 of this paper.

D. KG Construction Layer

Firstly, we establish the ontology database of the policy text, in which clarify the attributes and relationship existing in the policy. On this basis, the research is based on rule template matching and Bi-LSTM-CRF named entity recognition technology, which can automatically recognize and extracts entities from the policy text, as well as their attributes. Through the relation extraction algorithm of the learning method such as the convolutional neural networks or bidirectional cyclic neural networks, the layer extracts the relationship in the text based on former result of the entity recognition and shows the relationship types between them. Now we have the data of the "entity", the entity "attribute", and the "relationship" between the entities. Then we apply the RDF (Resource Description Framework) model in the Semantic Web Framework to show the knowledge extracted and store them in the graph database Neo4j. Above all, the policy knowledge graph has been fundamentally established.

E. Relationship Analysis Layer

In order to solve the problem of the incompleteness of the relationship between the policy entities and the lack of the attribute in the process of constructing, this paper conducts the reasoning research on the completed policy knowledge graph and deduces the relationship between the new entity and the entity. This paper designs a method based on first-order predicate logic to complete the policy knowledge. At the same time support the knowledge graph application of relationship analysis.

F. Knowledge Application Layer

Based on the actual needs of the public for policy information, this paper designs the knowledge application of policy information analysis. Relying on the policy information map knowledge base, through the map visualization technology, the policy development, policy evolution path, policy interpretation, through the node exploration, path discovery, association exploration visual analysis technology to display the comprehensive information of the policy, notify the knowledge graph, timing map, etc. The form interprets the geographical distribution and development trends and provides support for policy analysis.

The knowledge analysis platform based on knowledge graph designed in this paper is different from the traditional policy information service platform. It is not only the simple collection and integration of policy information, but also provides policy information retrieval and browsing, which can clearly and intuitively present the intrinsic attributes and policies of policies. The various relationships that help users of policy information better understand the various real-world policy entities, policy attributes and their connections. The platform design uses rich big data visualization components. Various visual maps such as scatter plots, maps, network maps, hot maps, and graphs combine with the rapid display and smooth interaction of policy entities, attributes, relationships, and provide users with a large policy. A comprehensive visualization of the data subject analysis and dynamic adjustment and real-time response as policy data sources or analysis conditions change.

III. Policy Data Collection

The policy information has the feature of multi-source heterogeneity. Not only it needs to read the structured data stored in the relational database, but also it needs to crawl the semi-structured data scattered on the government websites. Read unstructured data such as text. Considering the various structure type of data and different scenario, our platform are designed with different Application Programming Interface (API).

A. Structured Data Access

Structured data is generally stored in a relational database. This paper uses Sqoop (SQL-to-Hadoop) tool to implement data interaction between relational database (RDBMS) server and HBase and HDFS. During the data collection process, Sqoop will enable a MapReduce task to perform acquisition tasks and automatically and efficiently transfer large amounts of structured data. The implementation principle is shown in Fig. 3.

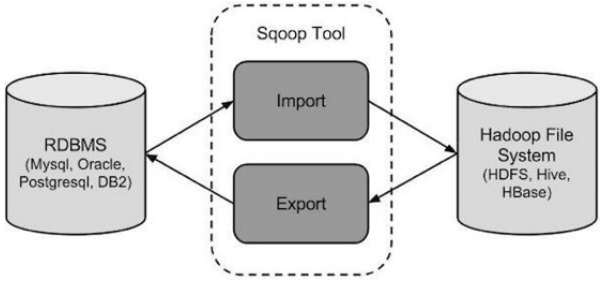

Figure 3. Sqoop schematic.

B. Semi-Structured Data Access

Many policy information content on the government website is carried by the webpage and marked according to the webpage format. Since the content of the web page carries information about its style, and the structure is presented, the website data is self-describing semi-structured data. Web crawlers are one of the most effective methods for collecting semi-structured html data.

This paper designs and develops a set of crawler tools to respond to the webpage format of different government websites by means of template combination, which reduces the development of crawler, and obtains the original data of the policy, and finally inputs the raw data into the storage system. The principle is illustrated in Fig. 4.

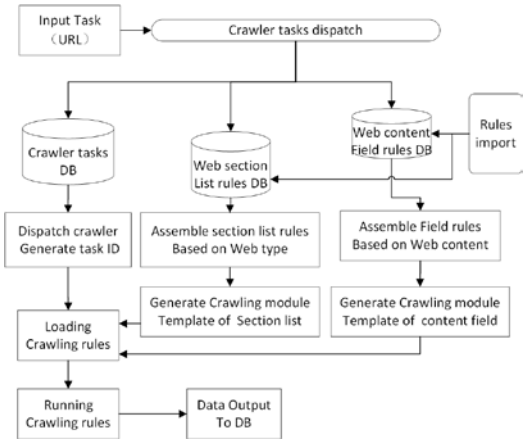

Figure 4. Crawler schematic.

C. *Unstructured Data Access*

Policy information is generally stored in the form of a document, such as word, pdf, in a folder on a computer. These document files may be stored on multiple computers in different geographical locations. In order to collect these unstructured document data centrally, Flume is employed. Flume is a distributed, reliable and highly available document collection system, which can flexibly and efficiently support the collection of document data in multiple computers in different places.

IV. POLICY DATA CLEANSING

Data cleaning in this paper refers to the process of merging and cleansing duplicate, erroneous, and missing data records when integrating multi-source heterogeneous data. That is to say, the policy data cleaning is mainly for the problem of the data itself, that is, the instance-level data quality problem is processed, thereby improving the data quality. The main cleaning methods used in this paper are: attribute cleaning method and repeated record cleaning method. Details are as follows:

A. *Cleansing of Invalid, Null and Erroneous Values*

When cleansing the data, we find that certain field may be missing, especially the publishing date of the document. There information is in the document, however, the date may not be expressed in Arabic numerals but in Chinese characters. In such scenarios, regular expression with certain pattern could extract the information appropriately.

Other property of document, such as the abstract, keywords and domain are difficult to obtain from the document directly. These kinds of information can be extract employing various NLP techniques. Due to the lack of annotated data, it is not possible to use a supervised algorithm for abstracts and keyword extraction. Graph-based method is a commonly used unsupervised method. Graph-based method is to construct a topological structure graph with sentences or words as nodes and co-occurrence relationship between the sentences or words as edges according to the sentences or words in the article. Globally sort sentences or words to determine abstracts or keywords for the text. TextRank [3] is one of the most famous graph-based keyword extraction methods, which itself is improved from PageRank [4]. In order to classify the theme, the training data set is automatically constructed by using the document issuing agency, because most of the documents issued by a document issuing agency belong to only one theme, and then the TextCNN algorithm is used to train on the training data set. The trained model is used for the classification of official theme.

Except the invalid and null value, in the data crawled from the websites, the date information can be wrong. In some documents, the publishing date of document is shown to be in year 1970 or 2090. In these cases, the erroneous dates are discarded, and the true publishing date can be extract using regular expression.

B. *Cleansing of Duplicate Values*

Sometimes, some important policy documents will be first published on the national government website, and then the provincial and municipal local administrative departments will reprint and publicize them. This will cause the crawler program to crawl to the same policy document on different government websites, resulting in duplicate data. In order to eliminate duplicate data, if the simplest method is used, all data are compared in pairs to determine whether the data match, and the computational complexity is O(N^2), where N is the number of data in the database. For the application scenarios involved in this article, such time overhead is unbearable. Therefore, this paper adopts the Sorted Neighborhood Method [5] (SNM), which first divides the independent subset by using the policy's posting time, and then uses the policy title to match the repeated data in the subset. This strategy of eliminating duplicate data can compare all the data in the database to determine whether the time complexity of the match is reduced to $O(N\log(N/C))$ (N is the amount of data in the database, the data set is divided into C independent sub-data sets)[6].The processing flow is as shown in Fig. 5.

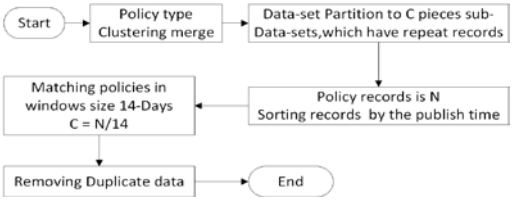

Figure 5. Procedure of matching duplicate data.

V. POLICY KNOWLEDGE GRAPH CONSTRUCTION

The policy knowledge graph is intended to describe the entities, attributes, and relationships that exist in policy documents. Its essence is a complex network that reveals the relationship between policy and document entities. For the policy document, the knowledge graph is defined by the domain experts. Based on the defined model of ontology, the collected is cleansed accordingly. Then the cleansed data is stored in the graph database Neo4j. Finally, the construction of the entire graph is achieved. The technical framework of the policy knowledge graph is as shown in Fig. 6.

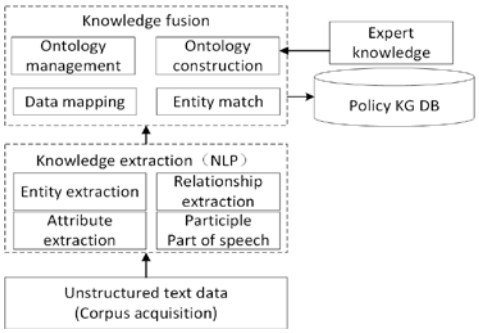

Figure 6. Structure of knowledge graph.

A. Ontology Construction

To extract the knowledge of policy data, this paper first defines the ontology model, the type of entities, and their corresponding attributes, relations, etc. In this way, the knowledge graph is standardized and extendable. In the policy graph, the entity types in the ontology model mainly include policy regulations, laws and regulations, policy interpretation, organization, subject domain, and so on. Each entity has detailed attribute descriptions, including the genre of the policy, the publishing time, the abstract, the keywords, the publishing authority, etc. Inter-entity relations include policy citation, transfer, forwarding, release, revision, abolition, etc.

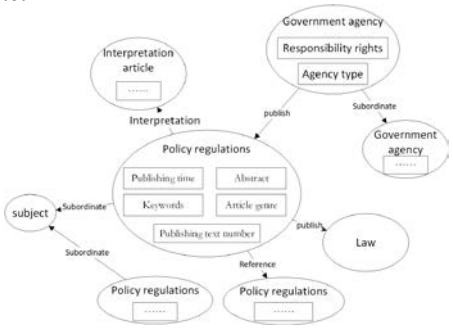

Figure 7. Ontology structure.

B. Knowledge Extraction

The knowledge, such as the entity, relations, attributes are obtained through supervised learning methods and patterns.

1) Entity extraction

The rule-based extraction is to identify the proper nouns that represent entities from the texts. Certain symbols can be employed as patterns that can be written regular expression.

In machine learning based approaches, entity extraction is performed from text in the form of supervised learning. This paper uses a named entity recognition technology based on Bi-LSTM+CRF [7].

2) Attribute extraction

In this paper, the extraction of attributes mainly adopts regular expression. The attributes of policy document are defined by domain experts. The attributes in policy document are "Arabic numerals", "Chinese numerals", "Year", "Month", "Day", and other statistical features.

3) Relation extraction

Relationship extraction, this article uses the method of rule matching, and the neural network-based relationship extraction algorithm complements each other. Because the rules rely heavily on expert experience and statistical results, it is inevitable that it cannot cover all aspects, but neural network-based relationship extraction can help support the problem of relation extraction. This paper mainly uses a relation extraction model based on local features and convolutional neural networks [8].

The method uses the vocabulary vector and the position vector of the word as the input of the convolutional neural network, and obtains the sentence representation through the convolutional layer, the pooling layer and the nonlinear layer. By considering the location vector of the entity and other related lexical features, the entity information in the sentence can be better considered in the relationship extraction.

C. Knowledge Fusion and Storage

After completing the entity, attribute, and relationship extraction, an RDF-based triple is formed and stored in the graph database Neo4j to form an entity network of the policy knowledge graph. But it is also necessary to standardize the data. In this project, the combination of automatic detection and manual assistance is used to verify the entity in the knowledge graph. The semantic library of policy is designed first, including common policy names, special nouns, government department names, such as “The belt and road”, “Internet plus”. And in this database, entities like "National Ministry of Science and Technology", "Ministry of Science and Technology" can be mapped to a single entity.

Since the attribute information of each policy entity in the policy knowledge graph is various, the retrieval requirement of the entity attribute information is also tremendous in practical application. To ensure the efficiency of the relation graph calculation, the graph database and the elastic storage entity and relationship are supported by the graph query language and algorithm. Based on this, this paper designs the storage mode as shown in Fig. 8.

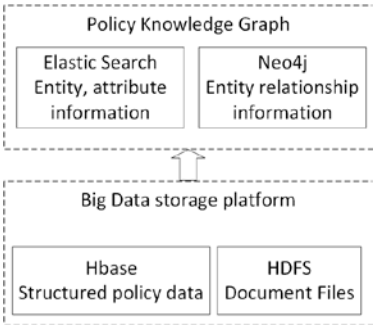

Figure 8. Storage of knowledge graph.

The knowledge graph data is divided into two parts: the entity attribute information and the entity relationship, and the attributes of the entity are stored as the publishing time of the policy document, the keyword, the topic, the abstract, the issuing department, etc., and the Elasticsearch index is created for storage. And the entity relationship is stored in

the map database Neo4j, and the storage, management and query of the knowledge graph data are completed.

VI. POLICY RELATIONSHIP ANALYSIS

Relational reasoning technology is an effective method to solve the problem of information loss and knowledge quality assessment in knowledge map. Because the research on policy knowledge map is still in the early stage of statistical relationship learning research, the main means of knowledge map reasoning in this paper is to combine with various entities through manually defined first-order predicate logic rules, and then build according to The logical network uses the entity as the basic unit for relational reasoning and taps potential knowledge to acquire new knowledge. In order to make the relationship analysis efficient and scalable, Hadoop's Spark parallel framework is used in the work to realize parallel reasoning.

The relationship reasoning process is as shown in the following figure. First, the entity involved is queried from the graph database. If the entity exists, then the relationship between the entity and other entities is queried, the relationship exists, and the relationship between the entity and the entity is established. Otherwise skip, so loop until all entity-relational-entity triples are completed. Return to the app.

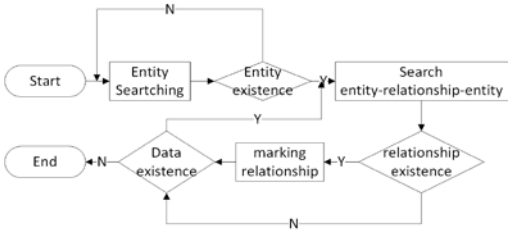

Figure 9. Procedure for relationship analysis.

VII. VISUALIZATION

Based on the collected items of millions of policy data, building a policy knowledge graph, our platform provides an accurate searching, indexing and relational visualization of policy data for the public and policy researchers. Our website is now online for probation. The URL is https://zwenzyong.com/home . There are two functions in our platform. One is searching and the other is knowledge graph visualization.

A. Policy Semantic Intelligent Retrieval and Sorting

This paper builds a semantic search model based on the constructed knowledge graph index, and provides semantic retrieval, intelligent sorting and related recommendation functions for policy entities by providing accurate and efficient keyword matching.

Firstly, the policy keyword and policy semantic library are used as candidate words, and the cosine similarity algorithm is used to calculate the policy of candidate words [9]. Finally, the relevant policies are recommended according to the relevance ranking to help users find out the meaning they want to express.

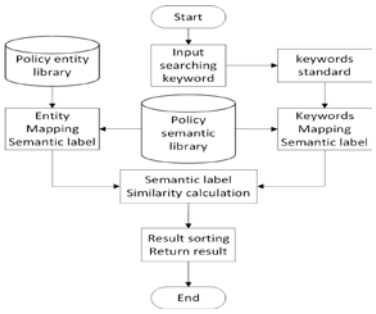

Figure 10. Procedure of policy indexing and workflow.

The policies and interpretations of this paper can be sorted according to the single dimension such as time and relevance. Because of the large amount of policy documents, the data overload is serious. The search results obtained by users after entering keywords are overwhelmingly large. To accurately find the policy document in the database, the following three factors are integrated into the sorting algorithm: time factor, policy agency impact factor and number of times the policy is cited. The calculation formula is

$$\text{Score} = W_t\, C/\big(C - T(time)\big) \times W_{if} IF \times W_{fr} FR$$

where C is the timestamp, T(time) represents the publication timestamp of the policy document, IF represents the policy agency impact factor, FR represents the number of times the policy is cited, and W represents the weight of each factor. The impact factor of the policy and the number of citations of the policy are positively correlated with the ranking score, while the publication time of the policy is negatively correlated with the ranking score.

Using the semantic intelligent retrieval technology of research and development, intelligently identify the true intention of user, accurately return the search results that best meet their needs to the user, and then use the intelligent sorting algorithm to recommend the high-quality policies published recently to the users.

Figure 11. The searching service in policy analysis platform.

From the perspective of user feedback, after considering the time factor, policy citation frequency and impact factor, the policy search results can be sorted well, which is beneficial to reduce information redundancy and improve retrieval efficiency. The details of the searching are shown in Fig. 11.

B. Entity Aggregation Exploration Analysis

By exhibiting the policy data in a graph, we provide another angle for the public, company, and policy researcher to understand the policy document. The graph is shown in

Fig. 12. And it reveals the relationship of an entity with other entities, and the relationship between policy documents.

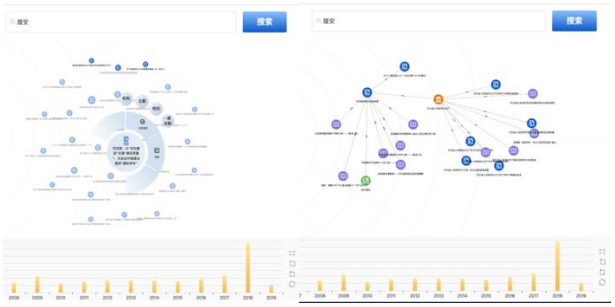

Figure 12. The policy relation in kowledge graph.

VIII. CONCLUSION

Based on the policy knowledge graph of policy data, a complex network of entities, policies, official documents, laws and regulations, interpretation and other entities is constructed. In this platform, we provide an accurate semantic searching tool, a statistic visualization of the collected data and a graph-based visualization of the relation of the policy data.

REFERENCES

[1] M. N. I. Sarker, M. Wu and M. A. Hossin, "Smart governance through bigdata: Digital transformation of public agencies," *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, Chengdu, 2018, pp. 62-70.

[2] A. Yang, S. Zhu, X. Li, J. Yu, M. Wei and C. Li, "The research of policy big data retrieval and analysis based on elastic search," *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)*, Chengdu, 2018, pp. 43-46.

[3] Mihalcea, Rada & , Rada & Tarau, Paul & , Paul. (2004). TextRank: Bringing Order into Texts.

[4] Page, Larry, "PageRank: Bringing Order to the Web". Archived from the original on May 6, 2002. Retrieved 2016-09-11., Stanford Digital Library Project, talk. August 18, 1997.

[5] Hernandez MA. Stolfo JS. Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem. Journal of Data Mining Knowledge Disco vo12. 1998. pp.9-37.

[6] Guha S.Jagadish H V.Koudas N.et al. Approximate XML joins[A]. In: Proceedings of the 2002 ACM SIGMOD International Conference on Management of Data[C]. Madison: ACM Press.2002:287-298.

[7] Ma, Xuezhe , and E. Hovy . "End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF." (2016).

[8] Daojian Zeng; Kang Liu; yubo Chen; Jun Zhao. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. EMNLP2015.

[9] Zeyuan Li, Jie He, Dagang Chen, Xin Fang, Yajun Song, and Zesong Li. 2018. A Hybrid Approach for Measuring Similarity between Government Documents of China. In *Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence* (CSAI '18). ACM, New York, USA, 431-435.