Visualization and Forecast Analysis of Science and Technology Intelligence Based on Knowledge Graph

Rui Ji, Shiming Yu, Hongwei Yan, Suren Ding,Ben Wang,Hui Zong, *Quanyin Zhu

Faculty of Computer & Software Engineering, Huaiyin Institute of Technology, Huai'an, China
*Corresponding author's email: hyitzqy@126.com

***Abstract*—it is difficult for scientific researchers to extract strutted knowledge from massive data and clarify the process and complex connections between it. The proposed system collects scientific and technological intelligence via data mining. Then we conduct information extraction, knowledge fusion, and knowledge processing to build Knowledge Graphs of scientific and technological intelligence. Finally, we can realize visualization of scientific and technological intelligence. The knowledge base constructed by the system can help researchers more intuitively understand the history, status, and future of discipline development. It can help researchers better understand the classification of disciplines and deepen their understanding of knowledge. Also, the system can conduct empirical research through citation analysis, bibliographic coupling analysis and other bibliometric methods. Via trend extrapolation, Tri-training and other algorithms, we can predict the trend of the frontier.**

Keywords- Knowledge Graph; Science and Technology Intelligence; Visualization; Data mining; Hotspot Prediction

I. INTRODUCTION

Google formally proposes the concept vocabulary of the Knowledge Graph in 2012, aiming to achieve smarter search engines. Given the enormous advantage of Knowledge Graphs in intelligence analysis, building a high-quality knowledge database is an important measure that caters to current trends.

The Knowledge Graph is essentially the knowledge base of the Semantic Network. The Semantic Network [4] is proposed in the late 1950s and early 1960s. Representatives are M. Ross Quilling and Robert F. Simmons. Using Semantic Networks, sentences in natural language can easily express and store in graphs for machine translation [5], question answering systems [6] and natural language understanding [7]. With the development of Semantic Networks, the works of the Semantic Network more focus on the modeling of relationships between concepts. The most representative work during the period is the classic language proposed by Brachman et al. [8] and the Fact Inference Engine implemented by Horrock [9].

The use of Resource Description Framework (RDF) is very important in the establishment of Knowledge Graph. Entities in the database are annotated with the standard "RDF:type" predicate, while the canonical entities are unique. Then connect the canonical entity to the database and assign a unique Uniform Resource Identifier (URI) in the database. After connecting, the user can view information about S1 (S2) or S2 (S1) in the "RDF:seeAlso" predicate. If two entities are in different databases, the user can move to another database to continue the operation [10].

In China, Donghua University design and construct the Knowledge Graph of the electronic medical records in Shanghai Ruijing Hospital. They find that it has good stability and performance in the completion of knowledge base about diabetes [1].

Mayank Kejriwal and Pedro Szekely describe a new entity-centric Knowledge Graph work; use it as a semantic search engine to help analysts and survey experts in the HT field. Prototypes use open source components and extend to tb-level corpus. [2]

II. SYSTEM ARCHITECTURE

There are four modules in the system. (1) the crawling of data (2) the extraction of ontology and relationships (3) the establishment of knowledge base (4) the use of graph database for preliminary visualization and the forecast analysis of frontier hotspots.

In the first module, we mainly get data in CNKI. The main crawl here is structured data. Including title, abstract, author, keyword and so on. For crawled data, the resulting data format is different; there are json, text, Comma Separate Values (CSV), etc. Of course, the character encoding is not the same. However, the semantic web approach introduces Resource Description Format (RDF) to unify heterogeneous databases. Constructing Knowledge Graphs from biological RDF databases has a very important role in biology [3].

The second module is the extraction of ontology and relationships. Extracting relationship is to extract the attribute information of the entity from the text, such as "national area", "population quantity" and other attributes, since the attribute of the entity can be regarded as a noun relationship between the entity and the attribute. The attribute extraction problem can be regarded as a problem of extracting relationship. The initial judgment of the ontology is the article and the author. The attributes mainly include WRITE, ADDRESS, DETAIL, and KEYWORD, ABSTRACT....

Ontology construction: There are usually three construction methods: artificial, semi-automatic and automatic. A large number of related field experts usually complete the method of artificially constructing ontology. From a biological point of view, common methods of constructing ontology by man are skeleton method [11], TOVE method [12], SENSUS method [13], Methontology method [14], Ontology Development method (seven-step

978-1-5386-7445-1/18/$31.00 ©2018 IEEE
DOI 10.1109/DCABES.2018.00021

method) [15] and so on. Automatically constructing ontology is also called ontology learning. The purpose is to use the knowledge acquisition technology, machine-learning technology and statistical technology to acquire ontology knowledge from resources, thereby reducing the cost of ontology construction. The last semi-automatic construction ontology is between manual and automatic construction of ontology, reducing the difficulty of automatically constructing the ontology. At the same time, we train the neural network based on deep learning (word2vec model) and the framework, the ontology and relation extraction are automated, and the development efficiency of the model is rapidly improved. The framework is shown in Figure 2.

The third module is the establishment of the knowledge base: collate the crawled data into three csv files that containing title, detail, keyword, abstract, Uniform Resource Location (URL) and author, address and author, title. Then import files into the neo4j database. Using Cypher batch data, construct nodes and relationships. Initially form a certain scale of knowledge base.

The fourth module is to use html, javascript...to make the data in the knowledge base into its own visualization site, allowing researchers to conduct visual search. In the end, the research method is used to study the history and status quo of related research. We can also use it to study the definition of frontier logic analysis. Using citation analysis, literature coupling and web-based analysis and other literature methods to conduct empirical research, combine with the expert consultation method to study the forefront of various fields. Use the trend extrapolation method, TRI-training and other methods to try to study the development trend of the frontier to predict. The system structure is shown in Figure 1.

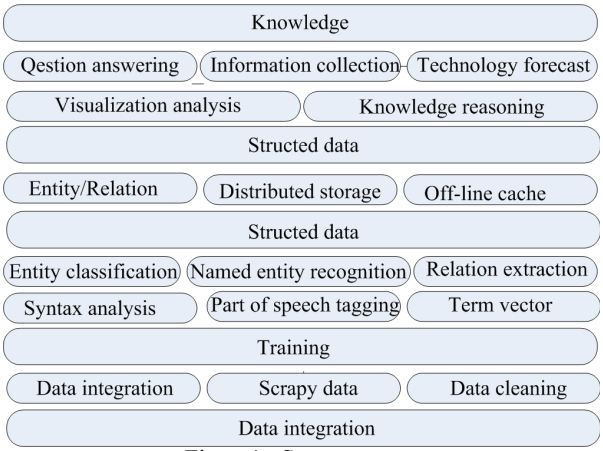

Figure1. System structure

III. DATA CRAWLED

The data crawling module indexes the keywords from the CNKI, uses the scrapy framework to implement the timed crawling of the CNKI, and achieves data crawling optimization. It turns out that the efficiency of crawling data through scrapy framework is greatly improved. Structured data crawled from CNKI, including title, url, detail, abstract, keyword, address, teacher, breaking, time, opennum, openday…

If the crawling speed is too rapid during crawling, CNKI will send a verification code to identify it. Even if a cookie is added, the disguised browser will not work. The initial crawl will be blocked on page 10. Once the verification code is identified, it will be blocked on page 12, and the cycle will be repeated. A single page crawl should be recommended for the issue. By simulating browser operations, the article information and author information are crawled at the same time, including references to articles and so on.

Of course, reading the contents of the verification code directly can also crack similar anti-crawler mechanisms. First, we extract the picture of the verification code from the result of the response and perform grayscale processing on the picture. That is to say, the picture is changed from color to gray. Then, the picture is binarization so that the picture is only black and white color. If the verification code has a border, we can first remove the border to improve the recognition accuracy. After getting preliminary processing results, we perform noise reduction on the image. The specific method of noise reduction processing is to select a point and then determine the four points adjacent to him. If two of these four points are white pixels, it is considered that the point is white, thereby removing the entire interference line. If the characters in the verification code adhere, because the glued characters are not well recognized, we must cut the glued characters into single characters and then identify the characters. The idea is to find one of the black points, and then iterate over all the points connected to it, including the highest point, the lowest point, the rightmost point, and the leftmost point, assuming that these four points form a character. Until all the points are found.... Finally, we call the typesseract library and execute the image_to_string function to get the result. Then, in the form of a form, post the result to the site server, you can crack.

Because the website uses javascript to dynamic loading, the simulated browser operation will be easier. By simulating the pull-down operation of the mouse, the asynchronous loaded data is obtained, including similar documents, similar papers, and the like.

Since the knowledge base is time-sensitive, updating the knowledge base is extremely important. Considering the difficulty of getting the website, the breadth of knowledge required, and the speed of crawling, we use the scrapy framework to crawl the CNKI from time to time. At the same time, the crawled data is processed automatically. The structured data is formed to the csv file of the ontology and the relationship and establishes nodes and relationships through the Cypher statement. For unstructured data, word segmentation and part-of-speech

tagging are used. Heuristic rules are used to filter. Finally, entities are selected and classified based on KNN to extract entities. The data classification is shown in Figure 3.

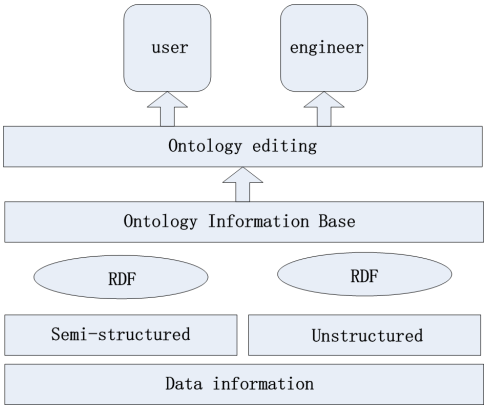

Figure2. Semantic Web Management

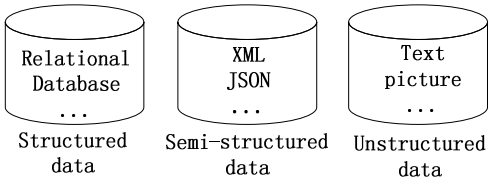

Figure3. Data Format

IV. FEATURE ENGINEERING

The operation parameters include the cosine similarity of the word vectors between the titles, the comparison of the preselected similarity of the document vector between the details, the number of identical or similar keywords, and the values are standardized so that they all obey the mean zero, and the variance a distribution of 1. Each value is weighted and summed. The weights are obtained by cross-validation + grid search. The framework is as follows:

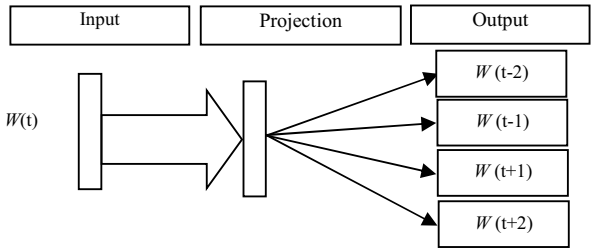

Figure4. Word2vec framework

Word vector generation formula based on KNN (1) algorithm:

$$\log \sigma(v_{WO}^{'T} v_{WI}) + \sum_{i=1}^{k} IE_{wi \sim p_{n(w)}} \left[\log \sigma(-v_{Wi}^{'T} v_{WI}) \right] \quad (1)$$

The development of Semantic Web brings great impetus to Knowledge Service System. A World Wide Web Consortium (W3C) technology includes RDF, Web Ontology Language (OWL) and Protocol and RDF Query Language (SPARQL). RDF is a way to describe resources. In simple terms, each description is a phrase consisting of a subject-predicate triple. However, the words we describe cannot be ambiguous. We must use the unique symbol method. Consequently, we must use URI to uniquely mark each RDF triple. With many such triples, we can get a Knowledge Network. Putting multiple such nets together forms the prototype of the World Wide Web. RDFS/OWL is used to describe RDF data. RDFS/OWL is essentially a collection of predefined vocabulary used to define RDF classes and attributes. Resource Description Framework Schema (RDFS) is the most basic pattern language. Because expressive ability of RDF is still quite limited, we also experiment with the OWL storage format. We find that using OWL to represent structured knowledge, data modeling becomes more flexible and rapid. At the same time, automatic reasoning also becomes more efficient. The Semantic Network management framework is shown in Figure 2:

V. QUERY MODULE DATA

The basic data query module is based on the neo4j database. The initial visualization of nodes and relationships is achieved through the node and relationship diagrams automatically generated by the neo4j diagram database.

However, during the experiment process, when the csv file imported to the graph database, it found that the nodes and relationships in the visualization are too cluttered because the relationship between nodes and nodes is too complicated and fail to meet the visual requirements.

At the same time as the author of the ontology, it should contain more information. Such as the author's institution, the research direction of the institution, the research direction of the author, the reference cited by the author at the time of writing, the number of times the article cited, tracking the trends of the authors and getting the conferences attended by the authors. The initial visualization results are shown in the figure 5:

In the visualization process, data grows exponentially. Even if you index only one article, the resulting divergent relationship diagram may never end. The users can click on the node endlessly and get a divergence relationship. To solve the problem, cluster analysis is used to achieve data clustering and compress the range of index results. At the same time, we should stratify the data. We can use a hierarchical tree-like graph; making the screening results, best fit the user's indexing intent.

In the process of data processing, because the data is too large, data coupling is likely to occur. For example, the names of people may be the same, but may come from different organizations. Therefore, for the ontology, the ontology should be given a unique label. The multi-label query on the ontology and the use of multiple labels to search at the same time make the relationship between the ontologies in the database uniquely meaningful.

In the initial process, entities and relationships that are connected to the ontology are hidden. Through the operator's click event, all connected hidden relationships and entities are displayed. The final figure is shown in Figure 6:

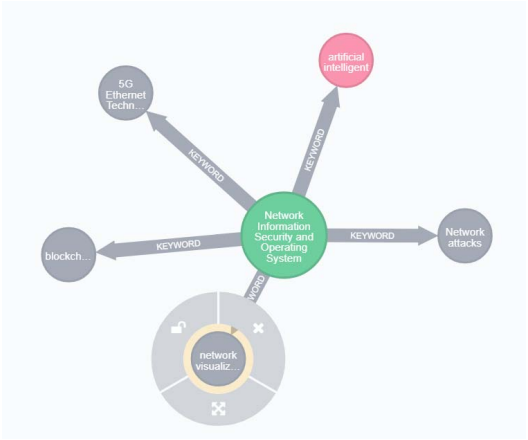

Figure5. Initial visualization

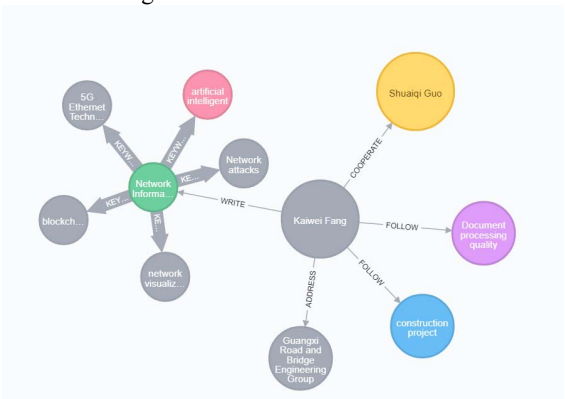

Figure6. Final visualization

VI. CONCLUSION

Through the establishment of the system, we understand the knowledge about Semantics Web. The expression of knowledge in network includes three kinds. These are RDF, RDFS, and OWL. Via data extraction and data cleaning, we realize the representation and the storage of the strutted data. Also, the proposed system realizes the visualization of knowledge in the neo4j database through basic nodes and relationships. With Knowledge Graph, we can mine the hidden relationships between nodes. When we combine experimental results with related algorithms, then the system can understand knowledge, infer knowledge, and even can predict the public hotspot.

ACKNOWLEDGMENTS

The system described in the paper is supported by fund from National Undergraduate Innovation & entrepreneurship training program (201811049022z) and The Provincial Key Research and Development Program of Jiangsu (BE2015127).

REFERENCES

[1] Suna Yin, Dehua Chen*, Jiajin Le. Deep Neural Network Based on Translation Model for Diabetes Knowledge Graph Domain [J].2017 Fifth International Conference on Advanced Cloud and Big Data (CBD), 2017: 318-323.

[2] Mayank Kejriwal, Pedro SzekelyKnowledge. Graphs for Social Good: An Entity-centric Search Engine for the Human Trafficking [J]. IEEE Transactions on Big Data, 2017, PP(99): 1-1.

[3] Nazar Zaki, Chandana Tennakoon and Hany Al Ashwal. Knowledge Graph Construction and Search for Biological Databases [J]. 2017 International Conference on Research and Innovation in Information Systems (ICRIIS), 2017: 1-6.

[4] Robert F. Simmons.Technologies for machine translation [J]. Futur e Generation Comp. Syst, 1986,2(2): 83-94.

[5] Robert F. Simmons.Technologies for machine translation [J]. Futur e Generation Comp. Syst,1986, 2(2): 83-94.

[6] Robert F. Simmons.Natural language question-answering systems: 1969 [J].Commun. ACM ,1970,13(1): 15-30.

[7] Yeong-Ho Yu, Robert F. Simmons:Truly Parallel Understanding of Text [C].AAAI 1990: 996-1001.

[8] Ronald J. Brachman, Deborah L. McGuinness, Peter F. Patel-Schneider, Alexander Borgida: "Reducing" CLASSIC to Practice: Knowledge Representation Theory Meets Reality [J].Artif. Intell.1 999, 114(1-2): 203-237.

[9] Ian Horrocks. The FaCT System [C]. TABLEAUX 1998: 307-312.

[10] Nazar Zaki, Chandana Tennakoon and Hany Al Ashwal. Knowledge Graph Construction and Search for Biological Databases [J]. 2017 International Conference on Research and Innovation in Information Systems (ICRIIS), 2017: 1-6.

[11] Uschold M, King M. Towards a methodology for building ontologies. Edinburgh: Artificial Intelligence Applications Institute, University of Edinburgh, 1995.

[12] Fox M S. The TOVE Project Towards a Common-Sense Model of the Enterprise. Industrial and Engineering Applications of Artificial Intelligence and Expert Systems, International Conference, Iea/aie - 92, Paderborn, Germany, June 9-12, 1992, Proceedings. 1992: 25-34.

[13] Swartout B, Patil R, Knight K, et al. Toward distributed use of large-scale ontologies. Proc. of the Tenth Workshop on Knowledge Acquisition for Knowledge-Based Systems. 1996: 138-148.

[14] Fernández-López M, Gómez-Pérez A, Juristo N. METHONTOLOGY: from ontological art towards ontological engineering. Proceedings of the AAAI97. 1997.

[15] Noy N F, McGuinness D L. Ontology development 101: A guide to creating your first ontology. 2001.