Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 11, No. 1, 2024

Research on Knowledge Graph Construction Methods for News Domain

Peiyi Yang*, Bin Song and Zhiyong Zhang

Information Engineering College, Henan University of Science and Technology, Luoyang 471023, China
* Corresponding author

Abstract: With the rapid development of the Internet and artificial intelligence technology, massive text data is growing exponentially, and the news industry is generating massive text data every moment. Most of the existing knowledge graphs focus on tourism, healthcare, finance and other fields, while there are relatively few in the news field, which is not well constructed. Aiming at the above problems, this paper firstly devotes to the construction of ontology library in news domain, and precisely defines 8 types of entities and 9 types of relations, which lays a solid foundation for the construction of news knowledge graph. Subsequently, with the help of web crawler technology, we extensively collected news text and image data, and carried out rigorous knowledge cleaning, extraction and fusion processing on these data to ensure the accuracy and completeness of the data. Finally, with the help of Neo4j graph database, the effective storage of news knowledge is realised, and the news domain knowledge graph is successfully constructed. It provides new ideas and means for information mining and utilisation in the news industry, and also provides rich, high-quality data support for downstream applications, which is expected to promote the intelligent development of the news field.

Keywords: News domain; knowledge graph; knowledge extraction; knowledge storage; visualisation.

1. Introduction

On May 17, 2012, Google officially launched the Knowledge Graph project [1], which has become an important cornerstone for Google to build a new intelligent search engine. Google has successfully built a new generation of search engine with the help of knowledge graph project, so that it has a higher level of intelligence and can better meet the user's search needs. The application of knowledge graph technology in Google search platform significantly improves its functional performance and empowers the system to provide detailed knowledge background and accurate answers to user queries. Knowledge graph is a knowledge representation method based on semantic network, its main role is to transform unstructured and semi-structured data into structured data [2], through which knowledge graph can accurately present complex and diverse knowledge relationships. It successfully integrates the scattered entities and their relationships in the text into a scaled and systematic knowledge network by abstracting the entities and relationships into nodes and edges in the graph. This network not only accurately reflects the existence of various types of entities and concepts in the real world, but also digs deeply into the intrinsic connections and mutual influences between them. As a powerful data model, it has the ability to support the relational representation of complex and massive data, and provides a new perspective to deeply understand the world [3].

With the rapid development of the Internet and the explosive growth of news information, people are faced with a large amount of fragmented news information and information overload [4]. Traditional news processing methods often fail to fully mine and integrate this information, making it difficult for users to accurately access the desired news content. In this context, the research of news knowledge graph has emerged. News Knowledge Graph is a knowledge graph with structured and semantic associations constructed based on news content. It provides a more comprehensive, accurate and comprehensible representation of news information by extracting and organising the entities, events, relationships and other elements in the news, presenting them in the form of a graph, and connecting different knowledge points through semantic associations, structuring and semantically associating the fragmented news information, so that the user can understand the correlation and internal logic of the news events more clearly, and thus understand the news content better. The news content can be better understood by the users. However, due to the complex relational attributes of news texts, knowledge graphs in the news domain are still rare and the related datasets are extremely scarce. Meanwhile, most of the knowledge graph construction work focuses on text data as a single modality, mainly focusing on the organisational and textual knowledge presented in a structured form, but often neglecting the rich image information contained in image data [5]. This paper proposes a knowledge graph construction framework for the news domain based on the research value and information characteristics, constructs a news domain dataset containing text and image modalities, makes great use of multi-source heterogeneous information, adopts deep learning and other techniques to complete the extraction, and expresses the news knowledge in the data in a structured way, provides a knowledge base for relevant practitioners and researchers, and provides a reference for the subsequent research in this field. Provide reference for the subsequent research in this field.

2. Related Work

Knowledge graphs are now widely used to process structured and textual data, and have been applied in many fields such as finance, medicine, education, and agriculture [6]. Miao et al [7] proposed a dynamic financial knowledge

graph construction method that integrates reinforcement learning and migration learning techniques. Migration learning algorithms are applied to train models based on BERT, Bi-LSTM and CBE, which are fine-tuned for the financial domain to recognise various financial entities in the text. In addition, a display website was designed and implemented to visualise the structural changes of the knowledge graph over time in real time.Li et al [8] proposed to learn the embedding vectors using PrTransH, performed graph embedding, and finally constructed a knowledge graph with nine relationships defined with diseases at the core, which optimised the construction of the knowledge graph. Literature [9] proposes a method for automatic construction of educational knowledge graphs based on word embedding techniques, using advanced word and sentence embedding techniques to improve concept extraction and weighting mechanisms. Specifically, the SIFRank keyword extraction method is improved using SqueezeBERT, which effectively improves the accuracy and efficiency of keyword extraction. And a concept weighting strategy based on SBERT is proposed, which provides theoretical support and practical paths to achieve efficient integration and intelligent application of educational resources.Qin et al [10] proposed a set of methods for constructing agricultural knowledge graphs, which successfully applied knowledge graphs to the field of agriculture.

At the same time, the modality of knowledge graph is also continuously enriched and expanded. Initially, the knowledge graph is mainly dominated by a single textual modality, but with the continuous progress of technology and the continuous expansion of application scenarios, the knowledge graph gradually realises the coexistence and integration of multiple modalities. Images, audio, video and other modal data are incorporated into the knowledge graph, which enables the knowledge graph to present the diversity of the world in a more comprehensive and three-dimensional way. Several large-scale knowledge graph projects have attempted to integrate visual elements, and IMGpedia [11] has successfully built a system containing over 15 million visual description records using high-quality visualisation resources from the Wikimedia Commons database. At the same time, the system also establishes up to 450 million visual similarity relations between images, which effectively facilitates the deep fusion of visual and textual information, thus achieving deeper information integration. However, it still faces the problems of sparse relationship types, small number of relationships, unclear classification, etc., and has not fully explored the potential entity relationships in images.Liu et al [12] constructed a collection named MMKG, which aggregates three knowledge graphs, each of which fuses the digital features of entities and image information, effectively integrating diverse entity and image information from multiple knowledge bases, so as to perform more complex relationship reasoning. This effectively integrates diverse entity and image information from multiple knowledge bases to perform more complex relational reasoning tasks. This approach breaks through the limitations of traditional visual reasoning and achieves more comprehensive and in-depth information processing and analysis. However, it is mainly constructed for small datasets, so it may have some limitations when dealing with large-scale data.

3. Ontological Construction of the News Domain

Ontology construction methods can be divided into manual and semi-automatic construction. Manual ontology construction relies on the cognitive wisdom and practical experience of domain experts, which has the advantage of ensuring that the constructed ontology closely matches the characteristics and practical needs of a specific domain, and accurately captures various concepts and their interconnections within the domain. However, the disadvantage is that it requires a lot of manpower, time and professionalism. Semi-automated ontology construction is a collaborative effort between computer technology and human expert knowledge. In this process, the computer can intelligently extract concepts, associations and attributes from the data source, which improves the efficiency and accuracy to a certain extent, while the professional insight and judgement of human experts can still be effectively integrated to ensure the quality of the construction results.

(1) Entity Definition

Given that news data covers a rich variety of entity categories, in order to ensure the accurate expression and efficient integration of news domain knowledge, it is especially crucial to define these entity types and their meanings in the process of constructing the news domain knowledge graph. After exhaustive analysis and collation of a large amount of news data, eight entity types are identified, with detailed information shown in Table 1.

Table 1. Entity type definitions

Number	Entity Type	Examples of entities
1	character	Wu Dajing, Yao Ming
2	Location	Beijing, London, New York
3	Organisation	United Nations, World Health Organisation
4	Time	2022, December, Thursday
5	Event	Olympics, Election
6	Position	President, Professor
7	Topic	Politics, Education
8	product	Huawei mobile phones, Tesla Motors

(2) Definition of entity attributes

Each entity represents a concrete object in the real world, and it is especially critical to define the types of attributes of these entities. In news data, entities usually have multiple attributes. Take the person entity as an example, it contains rich attribute information such as gender, age, position, place of birth, etc. These attributes not only help to characterise the entity more comprehensively, but also provide clues for explorations that reveal potential relationships between entities. If there are associations, they are presented in the form of graph structures for visual query and display, so as to more intuitively understand and analyse the complex relationships among entities.

(3) Entity Relationship Definition

After an in-depth analysis of the types of entities covered by news data, the following definitions are set for the types of inter-entity relationships in the news domain. These

relationships aim to reflect the association and interaction between entities more accurately and provide a solid foundation for the subsequent construction of the knowledge graph. The entity relationship definitions are shown in Table 2.

Table 2. Definitions of entity relationships

Number	Type of entity relationship	Entity-relationship example
1	Position	[Zhang Guimei, Position, Principal]
2	Country	[Event, Country, China]
3	Located	[Baima Temple, Located, Luoyang]
4	Time	[Beijing Winter Olympics, Time, 2022]
5	Membership	[Communist Party, Member, Wang Yang]
6	People involved	[Asian Games, People Involved, Ma Long]
7	Source	[Event, Source, People's Daily]
8	Report Time	[Event, Report Time, 26 November 2022]
9	Topic	[Event, Topic, Politics]

4. News Domain Knowledge Graph Construction

4.1. Overall Architecture

The overall system architecture of multimodal knowledge graph construction for news domain is mainly divided into data extraction, knowledge extraction, knowledge fusion and knowledge storage, as shown in Figure 1. The overall process is as follows:(1) Data extraction. Using web crawler technology, news data are crawled from official websites such as Xinhua, People's Daily, China News, Baidu encyclopaedia-type websites and so on. And pre-process it with data cleaning, de-weighting and so on. (2) Knowledge extraction, using the entity recognition model proposed in Chapter 3 to extract the acquired data according to the defined entity types, after obtaining the entities and their category labels, based on the defined nine relationship extraction guidelines, the associated entity pairs are extracted and transformed into the form of a relationship triad. (3) Knowledge fusion. The systematic convergence, organic articulation and consistent construction of knowledge units of different sources and forms aims to form a more complete, precise and practical integrated knowledge architecture. (4) Knowledge storage. The fused triad data is stored in the Neo4j graph database, so as to display and query the relational information intuitively.

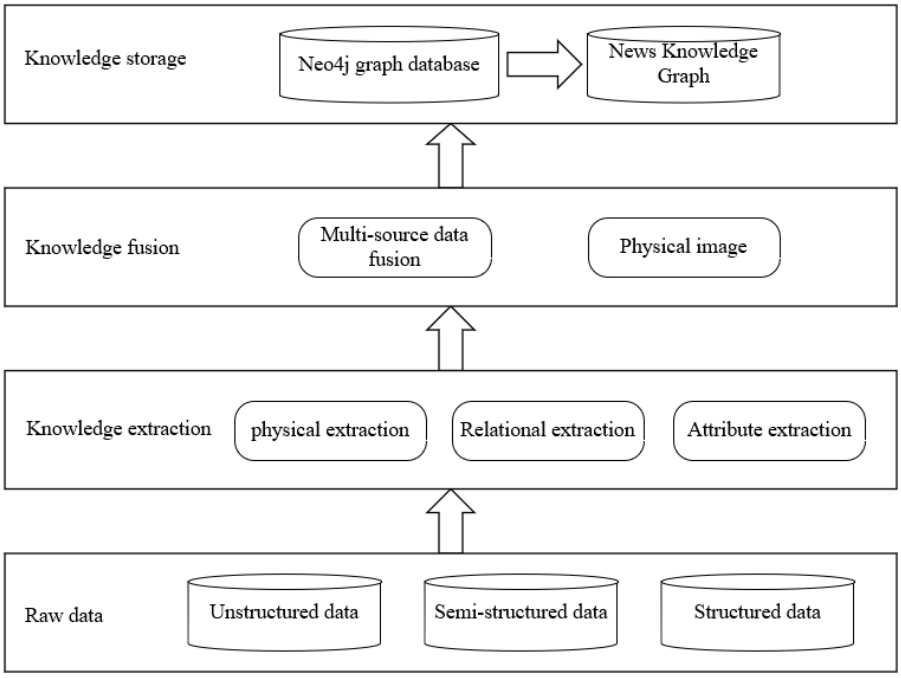

Figure 1. Architecture of knowledge graph construction for news domain

4.2. Knowledge Acquisition and Preprocessing

4.2.1. Text data

In view of the unique data organisation characteristics of news web pages, this paper selects the Request library and Selenium library in Python language as the technical means for news data capture. Selenium can accurately capture the rapidly changing information content when facing dynamic page scenarios that contain real-time update elements, such as scrolling news display area, to ensure the timeliness and completeness of data capture.Requests library can efficiently obtain the HTML document of the server's response by sending HTTP requests directly to the server, thus realising the comprehensive capture of the content of static and unchanging pages. Crawling. The crawling process is shown in Figure 2.

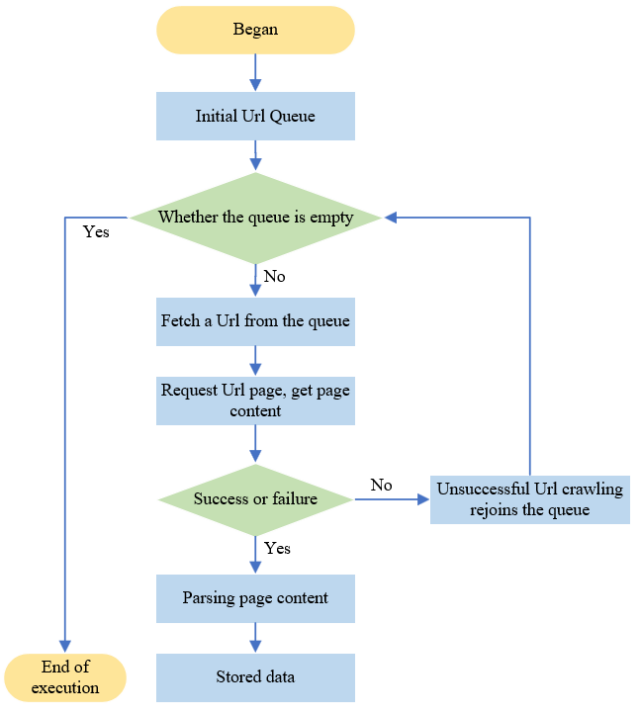

Figure 2. Data crawling process

In the process of improving the accuracy of knowledge extraction, it is crucial to perform the necessary pre-processing operations on the collected data. The following are the specific steps of this preprocessing process:

(1) De-duplication. In the collected raw data, there may be duplicate data entries. The de-duplication step is to identify and remove these duplicates.

(2) Data cleaning. There are non-substantive elements such as HTML tags, whitespace characters and line breaks mixed within the raw data set. In order to ensure the accuracy and reliability of the data, these irrelevant symbols should be rigorously cleaned up and eliminated, so as to ensure that the original data does not contain any interfering information.

(3) Text Sentencing: Excessively long sentences will adversely affect the extraction work, so these data need to be processed in sentences to improve the efficiency and accuracy of knowledge extraction.

4.2.2. Image data

Among the eight defined entity types, the three categories of people, places and products have detailed image information, so the acquisition of image information will be carried out for these three entity types. In this paper, the Internet search engine is chosen as the data source, and Baidu Encyclopaedia is used as the entity image acquisition tool. In the design process of the crawler programme, the list of extracted entities is first used as the input data. Then, Selenium, a browser automation testing framework, is used to achieve automated crawling and storage of the image entities and their corresponding URLs. Finally, this crawled image information is integrated into the knowledge graph as the attribute information of the entities in order to enrich and improve the content of the graph. In this process, the entity image attribute triad is represented in the following form: <entity, image_url, image URL>, which precisely describes the association between entities and image attributes.

After completing the acquisition of image entities, it is particularly important to fine-tune the processing and screening of high-quality images. Different images show the visual characteristics of entities from diverse perspectives, however, this method of supplementing visual information for entities based on Internet search engines often leads to a large number of duplicate images in the entity image set due to the lack of clear supervisory signals. These duplicate images not only increase the complexity of data processing, but also may adversely affect the subsequent knowledge graph construction and application. As shown in Figure 3, the picture collection of the entity Wu Dajing contains multiple similar pictures. Therefore, in the process of supplementing visual information, effective methods are needed to identify and remove duplicate images to ensure the quality and accuracy of image data.

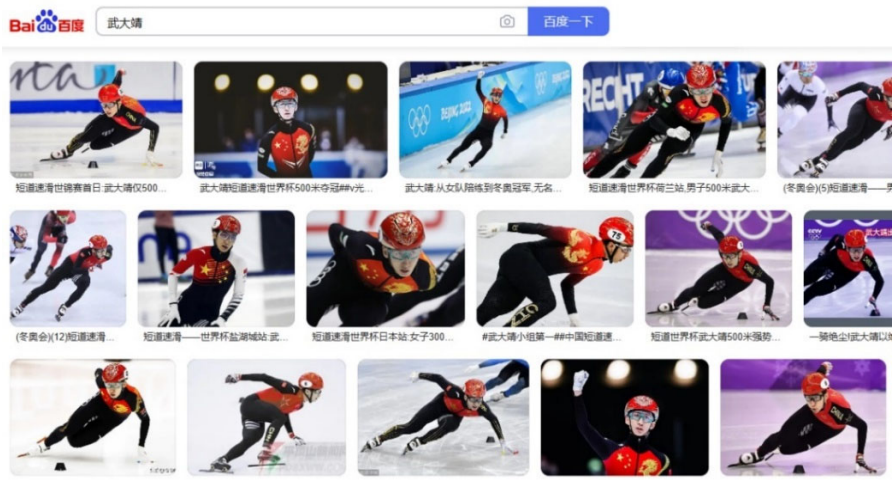

Figure 3. Collection of images returned by the search entity Wu Dajing

Firstly, each image is converted into a high-dimensional feature vector using the pre-trained ResNet model as a feature extractor. Next, the extracted feature vectors of the images are clustered using the DBSCAN algorithm, which is a density-based clustering method that is capable of discovering clusters of arbitrary shapes and efficiently handles noise points and outliers. The above steps enable to identify images with similar features and classify them into the same cluster. Eventually, a picture is randomly selected from each cluster

to form a collection of unduplicated and representative pictures. The specific flowchart is shown in Fig. 4.

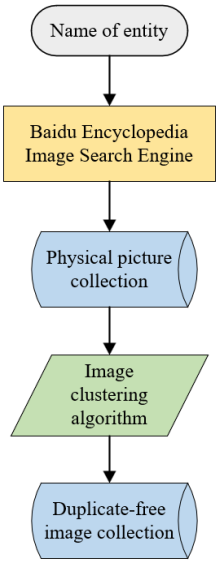

Figure 4. Flow of obtaining entity pictures

4.3. Knowledge Extraction

For semi-structured data with a relatively regular structure, entity information is relatively easy to extract. When storing such information, two types of triples, namely "entity-relationship-entity" and "entity-attribute-attribute-value", are commonly used to organise and represent the data efficiently. However, for unstructured data, the situation is relatively complex. In order to extract valuable entity information from the knowledge text in the news domain, this paper utilises the ALBERT pre-training model to dynamically weight and fuse the textual representations output from each layer of its encoder, and then adopts the BiLSTM-CRF model with the addition of an attention mechanism as a downstream task model, which focuses the limited information processing power on the effective entity information, as compared to the existing models, and It effectively solves the problems of multiple meanings of words and the cost of training time, and is able to accurately identify eight types of entities such as people, countries, organisations, events, time, etc. in the text. Through this technique, structured information can be effectively extracted from unstructured data, which provides strong support for constructing a knowledge graph in the field of news.

After completing the entity extraction of knowledge in the news domain, the next crucial step is to establish the semantic relationships between entities within the news domain. In the news domain, the rule-based approach tends to achieve better results because the text structure is relatively standardised and the relationship between events and entities is clearer. Specifically, after the entities were identified, the categories to which they belonged were further identified and the relationships between them were categorised on the basis of these categories. In this way, it was possible to clearly define what types of relationships existed between entities. This approach not only improves the accuracy of relationship extraction, but also provides strong support for the subsequent construction of the knowledge graph.

4.4. Knowledge Integration

In previous work, a large amount of data on entities and their relationships within the news domain has been successfully acquired. However, these data inevitably contain some redundant items, which need to be removed and processed in order to ensure the accuracy and efficiency of the data. At the same time, there are a large number of heterogeneous situations in the mapping, and these problems lead to ineffective interaction and integration of information between different entities or concepts. For example, "Peking University" is referred to as "Peking University", and "People's Bank of China" is referred to as "Central Bank". The complexity and diversity of news texts often lead to inconsistency in entity designation, which in turn affects the overall quality of the mapping. Multiple lexical representations need to be mapped to the same entity to ensure that they refer to the same entity.

By analysing the redundant data, it can be found that the text of some synonymous entities is similar, and entity alignment can be performed by calculating the proportion of the same characters in the entity's character sequence. For example, "General Office of the Central Committee of the Communist Party of China" and "General Office of the Central Committee" point to the same entity, and only two characters are different between them, so the Dice coefficient is used to calculate the textual similarity between entities, as shown in equation (1).

$$Sim_{Dice}(A,B)=\frac{2|A|\cap|B|}{|A|+|B|} \tag{1}$$

Where $|A|\cap|B|$ represents the intersection between A and B which is the number of characters that are the same between the two entities, and $|A|$ and $|B|$ represent the number of characters corresponding to the two entities A, B. Considering that the denominator covers the counts of these two characters at the same time, the numerator is doubled to ensure the balance. If the result is close to 1, it means that the textual features of the two entities have a high degree of similarity; on the contrary, if it is close to 0, it means that the textual features of the two entities are more different.

In addition, although some entity texts may appear similar, they do not actually point to the same entity. If the aforementioned method is used for calculation, it may trigger an entity alignment error. For example, for the entities "19th National Congress of the Communist Party of China" and "18th National Congress of the Communist Party of China", although their textual similarities are very high, they do not point to the same entity. Therefore, the cosine similarity is used to determine the distance between the word vectors of entities, as shown in equation (2).

$$Sim_{\cos}(A,B)=\frac{\sum_{i=1}^{n}A_i\times B_i}{\sqrt{\sum_{i=1}^{n}A_i^2}\times\sqrt{\sum_{i=1}^{n}B_i^2}} \tag{2}$$

Where A and B refer to the corresponding word vectors of the two entities, and A_i and B_i refer to the elements in the word vectors of the respective entities, the

larger the value of the publicity, the higher the semantic similarity between the entities.

To address the advantages of the two proposed similarity calculation methods, a comprehensive similarity value is obtained by assigning appropriate weights to each method. This method can more accurately judge whether the entities are similar or not, and is more in line with the needs of the actual task, in which the textual similarity accounts for 0.4 and the semantic similarity accounts for 0.6. The specific calculation is shown in equation (3).

$$Sim(A,B) = 0.4 \cdot Sim_{Dice}(A,B) + 0.6 \cdot Sim_{cos}(A,B) \quad (3)$$

4.5. Knowledge Storage and Visualisation

Neo4j is a powerful graphical database with a browser module called "Open Browser" that allows users to interact with the database intuitively. As a disk-based embedded persistence engine, it stores large amounts of structured data in an intricate network, a unique way of storing data that makes it excellent at handling relational data. In addition, the relationship between Neo4j nodes can be established directly , without additional connection or indexing process . Meanwhile, it supports multiple programming languages, such as Java, Python, etc., which further enhances flexibility and scalability. Table 3 shows the Cypher statement operation of Neo4j.

Table 3. Neo4j operation statements

Operation	Cypher statement
Creating nodes	CREATE (n {name:"zhangsan"})
Creating Relationships	MATCH (a:TEST),(b:TEST) WHERE a.name = 'TEST-NAME' AND b.name = 'TEST-NAME1' CREATE (a)-[r:RELTYPE]->(b) RETURN r
Delete nodes and their relationships	MATCH (n:Label {property: 'value'}) DETACH DELETE n
Adding node image properties	MATCH (p: Person {name: 'John Doe'}) SET p.image_url = 'https://example.com/image.jpg'

Once the knowledge storage is completed, the graphical data stored in the Neo4j database can be viewed and managed visually by visiting the URL "http://localhost:7474/". As shown in Figure 5, the visualisation of part of the news knowledge graph is implemented. Circular nodes are presented in different colors to identify different types of entities. These entity nodes are closely connected to each other by edges, which clearly shows the association and connection between them.

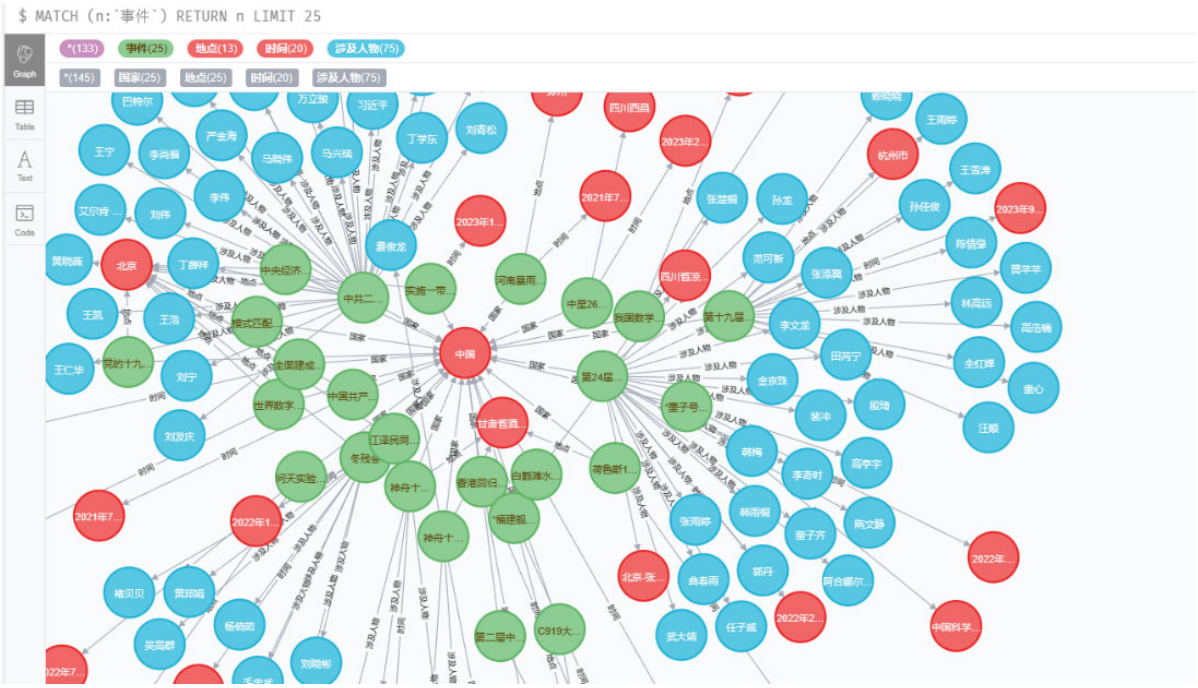

Figure 5. News Knowledge Graph

The image information related to the news is taken as a specific attribute and applied to the corresponding entity using the entity image attribute triad <entity, image_url, image URL> to achieve the supplementation and extension of its information. As shown in Figure 6, this expression not only enriches the display form of the graph, but also enables users to understand the news content more intuitively and enhances the user experience.

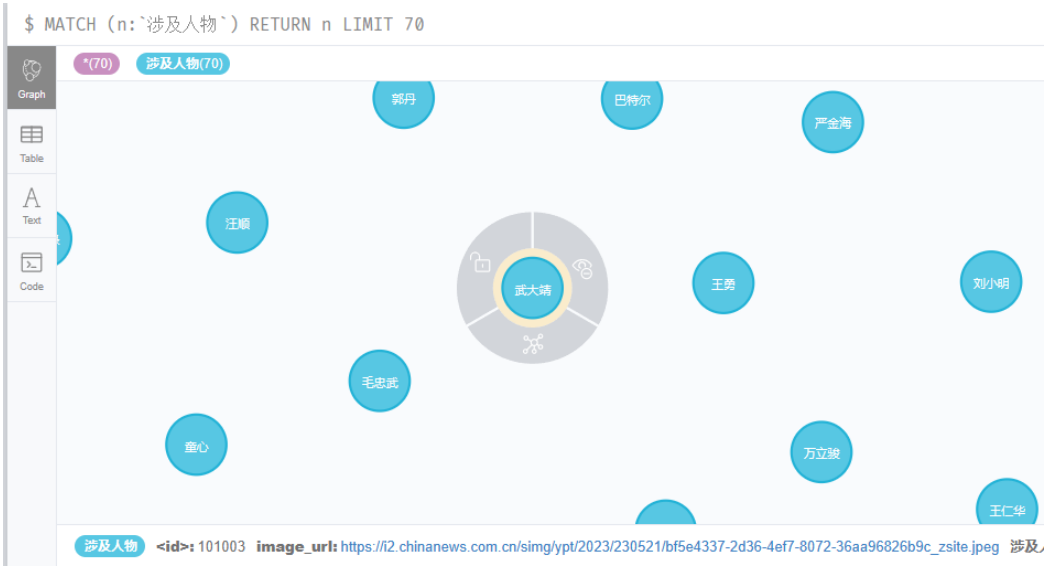

Figure 6. Picture property storage

5. Conclusion

In this paper, on the basis of fully explaining the development of knowledge graph, definition and its general construction process, we analyse the special characteristics of news knowledge graph construction, and put forward a knowledge graph construction method for the news domain, which first builds the ontology of the news domain, acquires high-quality knowledge from news websites by using web crawler technology, then removes redundant knowledge through knowledge fusion, and finally stores the obtained knowledge in the form of ternary groups. Finally, the knowledge obtained is stored in the Neo4j graph database to form a knowledge system with a mesh structure. In the future, the existing knowledge graph will be further enriched and improved by continuously refining the knowledge system in the news field to enhance its accuracy and completeness. At the same time, in the process of constructing the knowledge graph, information from other modalities such as video and voice can also be incorporated to enhance the comprehensiveness and usefulness of the knowledge graph and provide users with more accurate and comprehensive news knowledge services.

References

[1] Google Inside Search[EB/OL]. https://www.google.com/intl/es419/insidesearch/features/search/knowledge.html, [2016-02-10].

[2] Tamašauskaitė G, Groth P. Defining a knowledge graph development process through a systematic review[J]. ACM Transactions on Software Engineering and Methodology, 2023, 32(1) , p. 1-40.

[3] JI S, PAN S, CAMBRIA E, et al. A survey on knowledge graphs: Representation, acquisition, and applications[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(2) , p. 494-514.

[4] Yang Y , Li Y , Tung A K H .NewsLink: Empowering Intuitive News Search with Knowledge Graphs[C]// Proceedings of the 37th IEEE International Conference on Data Engineering (IEEE ICDE),2021, p. 876-887.

[5] Wu T, Qi G, Li C, et al. A survey of techniques for constructing Chinese knowledge graphs and their applications[J]. Sustainability, 2018, 10(9): 3245.

[6] Li J, Sun A, Han J, et al. A survey on Deep Learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 34(1) , p. 50-70.

[7] Miao R , Zhang X, Yan HFA, et al. Dynamic Financial Knowledge Graph Based on Reinforcement Learning and Transfer Learning[C]// Proceedings of the IEEE International Conference on Big Data, 2019, p. 5370-5378.

[8] LI L, WANG P, YAN J, et al. Real-world data medical knowledge graph: construction and applications[J]. Artificial Intelligence in Medicine, 2020, 103: 101817.

[9] Ain U Q ,Chatti A M ,Bakar C G K , et al.Automatic Construction of Educational Knowledge Graphs: A Word Embedding-Based Approach[J].Information,2023,14(10).

[10] QIN H, YAO Y. Agriculture Knowledge Graph Construction and Application[J]. Journal of Physics: Conference Series, 2021, 1756(1): 012010.

[11] Ferrada S, Bustos B, Hogan A. IMGpedia:a linked dataset with content-based analysis of Wikimedia images[C]// Proceedings of the International Semantic Web Conference. Cham:Springer, 2017, p. 84-93.

[12] Liu Ye, Li Hui, Garcia-Duran A, et al.MMKG:multi-modal knowledge graphs[C]// Proceedings of the European Semantic Web Conference.Cham:Springer,2019, p. 459-474.