Construction and application of power grid patent knowledge map

Zhicheng Ma[a], Jinxiong Zhao[a], Xun Zhang[a], Siheng Zhai[b] , Baohui Wang[b,*], Guangyuan Zheng[b] and Henan Zhang[b]

[a]State Grid Gansu Electric Power Research Institute, Lanzhou 730070, China; [b]College of Software, Beihang University, Beijing 100191,China.

ABSTRACT

At present, scientific and technological personnel in electric power enterprises mainly rely on patent search websites to obtain cutting-edge electric patent information. But these websites are mainly based on string matching, which fails to capture the connection between patents, making the recommend result unsatisfying. In view of the above problems, we first constructed a grid patent knowledge map, where knowledge extraction was carried out for entities such as title, abstract and applicant in the patent text, and the entity and defined relational data were stored in the Neo4j graph database.Secondly, the Transe-SNS algorithm with optimized negative sampling was used for vectorization of the graph entity relationship.Experiments showed that MeanRank and hit@10 improved by 27.5 and 2.2% respectively compared with the traditional TransE algorithm.Finally, the similarity between patents was calculated by combining the results of knowledge graph vector embedding of patent entities with the word vector embedding of patent title abstraction,and top-k patents in similar fields were recommended for users. The experiment proved that the proposed method is superior to the traditional text embedding method in recommending similar patent technology fields.

Keywords: Knowledge graph, knowledge graph embedding, recommendation algorithm

1. INTRODUCTION

In the face of the rapid development trend of scientific and technological knowledge, understanding and mastering the scientific and technological information contained in patent data has an important impact on promoting entrepreneurship and innovation of enterprises and individual. Patent-related data contains many innovative cutting-edge technologies and have relatively high value in the industry. Patents and the technology contained in patents have become an important resource to promote the progress and innovation of enterprises, mastering the most advanced technology in the industry can improve the competitiveness of enterprises, so mastering the development of patents in different technical fields is of great significance to enterprises and even countries. For China's power grid science and technology staff,the current access to the latest patent science and technology information mainly relies on patent search websites, which only present patent search results to users through string matching, making it difficult to explore the connection and similarity between patents. When a patent is retrieved, the power grid scientists hope to recommend a patent that is as like the patent as possible, to broaden their research ideas and give them a general understanding of the latest patent situation in a subfield. Based on these problems, this paper uses knowledge graph to model and store power grid patent data, since knowledge graph can better capture the connection between patents and realize the calculation of similarity between patents and patent recommendation according to the constructed knowledge graph.

*Baohui Wang :wangbh@buaa.edu.cn;
Zhicheng Ma: 2633116578@qq.com; Jinxiong Zhao: 1304518920@qq.com;
Xun Zhang: 3508788445@qq.com; Siheng Zhai: 396848580@qq.com;
Guangyuan Zheng: zhengguangyuan@buaa.edu.cn; Henan Zhang: 84868218@qq.com.

Seventh International Conference on Mechatronics and Intelligent Robotics (ICMIR 2023), edited by Srikanta Patnaik,
Tao Shen, Proc. of SPIE Vol. 12779, 1277926 · © 2023 SPIE · 0277-786X · Published under a
Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.2688885

2. KNOWLEDGE GRAPH CONSTRUCTION

2.1 Introduction to the knowledge graph

In 2012, Google first proposed the knowledge graph, and they immediately began using the knowledge graph technology to improve the core search engine[1]. Knowledge graph is essentially a semantic network that describes objective things in the form of a graph, composed of nodes and edges. Each knowledge point represents a triad (Subject-Predicate-Object, SPO), which can also be recorded as HRT (Head-Relation-Table).[2]

The knowledge graph consists of triples in the following two forms: ① "Entity-relationship-entity" describes the relationship between entities, such as "Guo Qilin-Father-Guo Degang"; ② "Entity-attribute-attribute value" describes the relationship between entities and their attribute values, such as "Guo Qilin-Gender-male" [3]. The general construction flow of the knowledge graph is shown in Figure 1[4].

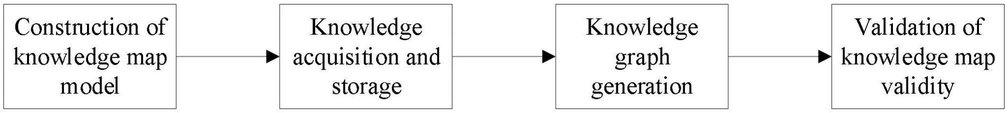
Construction of knowledge map model
Knowledge acquisition and storage
Knowledge graph generation
Validation of knowledge map validity

Figure 1. General construction process of the knowledge graph

At present, the mainstream knowledge graph is generally divided into two types, one is the knowledge graph of the general domain, such as DBpedia[5], Freebase, Wikidata, etc. They are mainly oriented to the general common-sense domain, and the search and question answering system of the underlying technology of these knowledge graphs can answer most general domain knowledge. There is also a vertical knowledge graph, which aims to build a knowledge base in a specific field. Compared with the general knowledge graph, the domain knowledge graph requires a more rigorous definition of data patterns and requires the person who constructs the map to have a clear understanding of the knowledge framework of the field. Domain knowledge graph is generally used as the vertical search and recommendation of the closed field, such as the e-commerce knowledge graph built by Alibaba mainly serves the product search and recommendation behavior of Taobao users. At present, there are very few knowledge maps applied to the patent field, especially in the field of power grid patents, and it is difficult for power grid scientists to find their correlation from a large number of patent data. The construction of knowledge graph is usually divided into the following steps: ontology definition->knowledge extraction->knowledge fusion-> knowledge storage.

2.2 Patent ontology construction

In the process of constructing a knowledge graph, ontology construction is an important link, and patent-related professional terms defined in the ontology will affect the quality of the graph[6].Ontology realizes the abstract representation of patent-related text data, completes and represents the relevant concepts of patents and the relationships between them, in order to effectively correlate patent data and standardize the representation of entities in the data layer. Considering that the scope of knowledge in the patent field is relatively fixed, this paper adopts a top-down approach to construct the patent ontology library[7],After analyzing the importance of each entity attribute of the original text of the patent, the schema that determines the entity and relationship of the patent is shown in Table 1 and Table 2：

Table 1. Patent Entity Definition.

Entity name	**Entity description**
Patent ID	A unique invention release number corresponding to each patent
Application	Patent filing companies/schools, usually several
Inventor	The inventor of a patent, usually several
IPC classification numberIPC	The IPC main classification number of a patent, which identifies the technical field of the patent
Keyword	Key terms extracted from patent texts

Table 2. Patent Relationship Definition.

Relationship name	Relationship description
Application	(Applicant, Application, Patent)
Invent	(Inventor, Invention, Patent)
Contain	(Patents, Inclusions, Keywords)
Technical field	(Patents, technical fields, IPC classification symbols)

After defining schemas for entities and relationships, the ontology can be instantiated by knowledge extraction from the original data.

2.3 Entity relationship extraction

Patent ID, patent applicant, inventor and IPC classification number can be obtained by simple script processing of the original data. We design an algorithm to extract the keywords of the patent. The steps of the keyword extraction algorithm proposed in this paper for patent text are as follows:

Stitch together the title and abstract text of the patent to form a piece of text and remove meaningless stop words from the text by customizing the stop word list.

2)Add a glossary of professional terms in the field of electric power based on jieba participle to make word segmentation more accurate and prevent wrongly segmenting electric power terms.

3)Extract patent text based on TextRank keyword extraction algorithm and return the Top-5 keyword list.Some of the keywords extracted using the improved keyword extraction algorithm are shown in the following Table 3:

Table 3. Examples of patented keywords.

Patent ID	Patent Keywords		
CN111555281B	emulation	flexibility	Power system
CN110299717B	energy storage	subsystem	power
CN112100829B	bridge following	cable	matrix
CN112531986B	stator	Winding	winding
CN113032984B	module	power	Fourier
CN106026322B	energy storage	charge	Electric vehicle

2.4 Knowledge storage

After extracting the patent knowledge of the power grid, the triples formed by the extraction are stored in the NEO4J database. Neo4j is an open-source NoSQL graphics database that started development in 2003, using the scala and java languages, and was published in 2007.[8]It stores the data and its attribute associations in the structure of a graph. The Neo4j graph database consists of three basic elements: nodes (nodes), relationships (relationships) and attributes (properties), each of which can be stored independently.All of its relationships and nodes are available to create properties, represented as key-value pairs, similar to the hashMap data structure.[9].Figure 2 takes patent CN108962515B as an example to show the relationship data information between entities and entities extracted from the patent. According to the structure diagram of patent entity relationship, the nodes of different colors represent the patent publication number, keywords, inventor, applicant and IPC classification number, and the relationship between each entity can also be displayed intuitively.

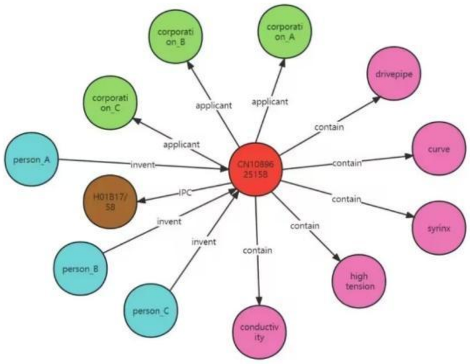

Figure 2 . Relational data information.

3. OPTIMIZE THE KNOWLEDGE GRAPH EMBEDDING ALGORITHM FOR NEGATIVE SAMPLING

3.1 Patent vectorized representation based on knowledge graph embedding

The vectorized representation of the knowledge graph, also known as knowledge representation learning、knowledge graph embedding, refers to the method of representing entities and relationships in the knowledge graph as real-valued numeric vectors which have certain semantic expression capabilities. After the storage of the knowledge graph, the similarity calculation and recommendation between entities cannot be achieved by relying solely on the text information of the triplet, and the entities and relationships need to be vectorized to calculate the similarity between entities.The TransE algorithm[10] is a classical knowledge graph embedding method, which regards a knowledge triplet (h,r,t) as a translation operation from its head entity h through the relation r to the tail entity t, that is, simply consider the sum of the head entity vector and the relationship vector $\vec{h}+\vec{r}$ should be as equal as possible $\vec{t}$ .The TransE algorithm can be very handy implemented with the OpenKE tool and is accurate and efficient in representing 1-1 relationships. However, the original random negative sampling algorithm of the TransE algorithm is easy to produce many false negative samples when facing the relationship between 1-N and N-N, which affects the model training effect and the final entity relationship representation. Therefore, this paper proposes a TransE-SNS model which optimize the negative sampling process of TransE. The optimized negative sampling algorithm replaces entities with a certain probability for triples other than one-to-one. For each relationship in the patent knowledge graph, according to the existing triplet data information, the average number of tail entities corresponding to the head entity under this relationship is counted separately N_{tp} ,The average of the number of tail entities corresponding to the number of head entities under this relationship N_{hp},Then the probability of replacing the entity p is calculated as :

$$\mathrm{p} = \frac{\mathrm{N_{tp}}}{\mathrm{N_{tp}+N_{hp}}} \tag{1}$$

In TransE-SNS, the entity data is no longer randomly replaced, which can avoid excessive false negative samples in the negative sampling process and retain the complex semantic correlation between the original correct triples, which makes the TransE model more realistic in the vectorization process. The improved algorithm flow is shown below：

Input Knowledge graph G=(E,R,P),E=(e_1,e_2,…,e_n),R=(r_1,r_2,…,r_m), P=(p_1,p_2,…,p_l)

Output Entity vector ($\vec{e_1}$, $\vec{e_2}$,…, $\vec{e_n}$) and relation vectors ($\vec{r_1}$, $\vec{r_2}$,…, $\vec{r_m}$)

Parameters entity and vector dimension d, learning rate λ, distance adjustment parameter γ , size b of each batch, number of training batches k.

1.Random initialization ($\vec{e_1}$, $\vec{e_2}$,…, $\vec{e_n}$) and ($\vec{r_1}$, $\vec{r_2}$,…, $\vec{r_m}$) , normalize the relation vector

2.Loop1(Cycle k times，k)：

normalize the entity vector for each entity, $\overrightarrow{e_{\mathrm{n}}}$ =$\overrightarrow{e_{\mathrm{n}}}$ /‖$\overrightarrow{e_{\mathrm{n}}}$‖$_2$

Loop2(Cycle $\frac{1}{b}$ times$\frac{1}{b}$)

(1) Randomly select b triples from the knowledge graph to form a set P_b.

(2) For each triplet in P_b,Pb， construct the corresponding negative case triplet to form the negative case triplet set P_b′ corresponding to P.

(3) Calculate the loss function $L = \frac{1}{l}\sum_{p \in P_b'}[f(p) - f(\acute{p}) + \gamma]_+$, update entity and relationship vectors with stochastic gradient descent.

For the evaluation of the quality of the knowledge graph vectorization representation algorithm, link prediction is commonly used as a test task, and hit@10 and MeanRank[11] are used as indicators for evaluation.The TransE-SNS model was used to represent the semantic information in the patent entity in a vector space, and the original data was 1000 patent data related to the power industry during the experiment. According to the patent knowledge graph constructed in Chapter 2, a total of 12,328 pieces of application, invention and inclusion relationship data in the triplet (h,r,t) are obtained, namely 1985 application relationship data between applicant and patent, 5343 invention relationship data between inventor and patent, and 5000 inclusion relationship data between patent and keyword. During the TransE-SNS model training process, 12328 relational data pieces were trained and tested at a 4:1 ratio. Take the learning rate λ=0.01, the boundary value γ =2, the dimension of entity embedding d={20,50,70,100,150}, and finally determine the optimal embedding dimension according to the MeanRank value and Hist10 value. To optimize the effectiveness of the negative sampling algorithm, the TransE algorithm and the improved TransE-SNS algorithm are used to embed the patent entity relationship and calculate the corresponding MeanRank and hit@10 values.

The smaller the MeanRank value and the higher the Hist10 value, the better the model. The MeanRank values and hit@10 values of the TransE model and the improved TransE-SNS model are shown in Table 4 and Table 5, and their scoring trends are shown in Figure 3 and Figure 4 .

Table 4. The MeanRank value of the TransE algorithm and the TransE-SNS algorithm in different embedding dimensions.

Mean Rank					
	d=20	d=50	d=70	d=100	d=200
TranSE	420	250.4	220.3	188.1	202.4
TranSE-SNS	370.1	215.3	184.4	160.6	172.1

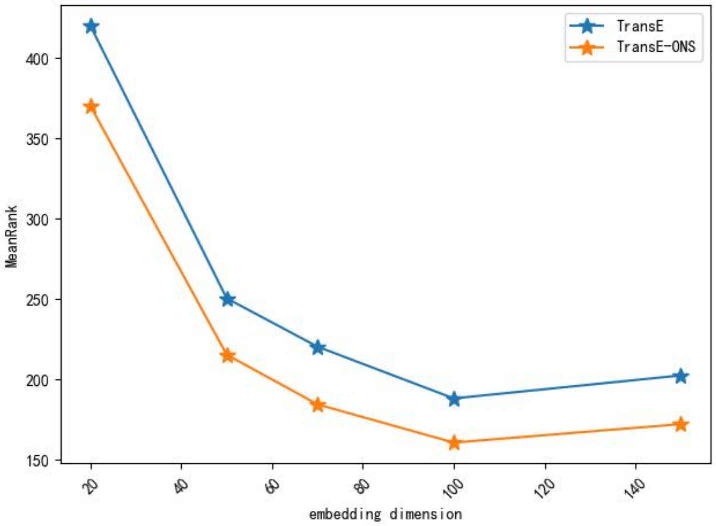

Figure 3. Comparison of MeanRank values of TransE algorithm and TransE-SNS algorithm in different dimensions.

Table 5. The hit@10 values of the TransE algorithm and the TransE-SNS algorithm in different embedding dimensions.

hit@10(centesimal system)					
	d=20	d=50	d=70	d=100	d=200
TranSE	25.4	28.1	29.4	30.2	29.5
TranSE-SNS	27.5	30.4	31.5	32.4	31.4

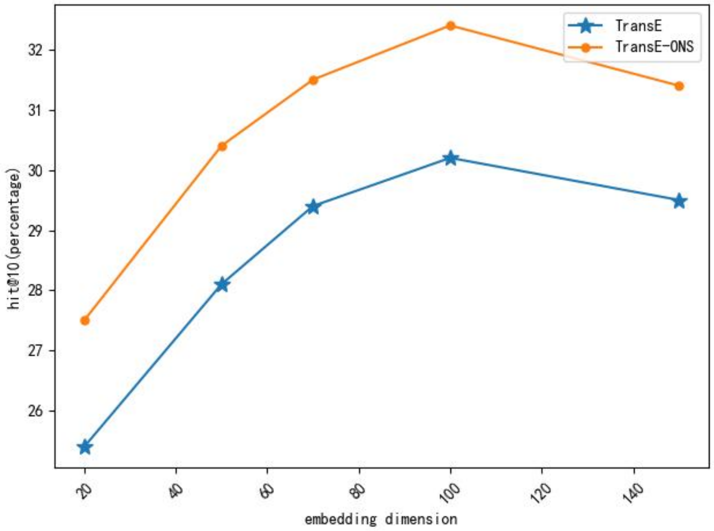

Figure 4. Comparison of hit@10 values of TransE algorithm and TransE-SNS algorithm in different dimensions.

Through the experimental results of the average rank MeanRank value and the top ten hit rate hit@10 value, when the embedded dimension of the patent entity vector is 100, its training effect is relatively optimal, and the improved negative sampling algorithm can make the model achieve a better effect.

4. PATENT RECOMMENDATION THAT INTEGRATES KNOWLEDGE GRAPH EMBEDDING AND WORD EMBEDDING

At present, in the field of patent recommendation, it is mainly based on content-based recommendation algorithms. Based on the content of patent texts, patent recommendation is realized by calculating the semantic similarity between texts, but this method does not consider the interrelationship between multiple patents, and it is difficult to recommend patents in similar fields. By integrating the recommendation algorithm based on patent knowledge graph and content, this paper can better model the key between patents while retaining the semantic information of patent text, so as to achieve more accurate patent recommendation.

4.1 Patent similarity calculation based on word embedding

Content-based recommendation algorithm has been applied in major fields. This paper simply calculates the text semantic similarity between the patents through title and abstract to recommend patent. We need to vectorize the text content to obtain the text vectorization result before calculating the similarity. The word embedding[12] method of converting natural language text into vectors is very mature, and the mainstream word embedding methods are distributed representation methods based on word2vec.

This paper uses the tokenizer library in Baidu paddle framework PaddleNLP to realize the word vector embedding of patent title and abstract text, because Baidu's natural language processing framework has been pre-trained on large-scale Chinese corpus in advance, and can better identify entities in long text.

After setting the embedding dim to 300, we can get the word vector converted by the title and abstract text of each patent. Then we can calculate the textual similarity between patents by cosine similarity.

$$\text{similarity} = \cos\theta = \frac{A \cdot B}{\|A\|\|B\|} \tag{2}$$

A, B are the word vectors of any two patent texts.

Taking patent 'an electric vehicle charging plug-and-play system control method equipped with energy storage battery' as an example, the IPC classification number is H02J7/02, and the top 10 patents with the title and abstract similarity of the patent are shown in Table 6.

Table 6. Patent similarity calculation based on Word Embedding.

Similar Patent Titles	Similarity	IPC
High-power charging cable for new energy vehicle charging and how to use it	0.90376	H01B7/42
Control method and control device of energy storage charging pile system	0.899404	H02J3/32
A fire extinguishing system for energy storage battery device based on temperature switch	0.896481	A62C3/16
A battery module that uses a stiffener beam to improve the fastness of the cooling plate and a battery pack that includes the battery module	0.896186	H01M10/613
A battery balancing control method and device based on MPC	0.88657	H02J7/00
A SOC immunity evaluation method for new energy vehicle battery for DC charging pile charging monitoring data	0.885712	G01R31/382
Control methods, devices and electronic equipment based on flywheel energy storage system	0.884083	H02J3/30
Automatic switching control method of charge and discharge mode of energy storage converter	0.88359	H02J3/32
Devices, methods, and batteries for measuring the internal temperature and strain of batteries	0.882987	H01M10/48
An integrated wind, solar and storage electric vehicle charging system based on model predictive control	0.881613	H06L53/51

Most of the patents obtained by this method are calculated based on the number of words co-existing between the two patents, which only considers the textual information, but ignore the possible connection between the patents and the potential attributes of the patent, making the recommended patents have little connection with the original patent in content. For example, we can roughly see the technical field of the recommended patent through the IPC of the similar patent. Although most of them start with H electricity, there are two patents belonging to parts A and G respectively, and some recommended patents and the H02J7 small part of the original patent do not match, indicating that their research fields are not identical, and the possibility of similarity in the patented technologies is very low.

4.2 Patent similarity calculation combining knowledge graph embedding and word embedding

After using the TransE-SNS model to vectorize the entities in the patent knowledge graph, the vectorized representation of the patent title entities is obtained, which can better contain the rich semantic relationships in the map. This paper proposes a patent recommendation algorithm that integrates knowledge graph embedding and word embedding and splices the patent entity knowledge graph embedding vector obtained in Chapter 3 with the patent text word embedding vector. The spliced 400-dimensional vector contains the entity relationship information in the text and knowledge graph at the same time, and then calculates the similarity to obtain a better recommendation effect.

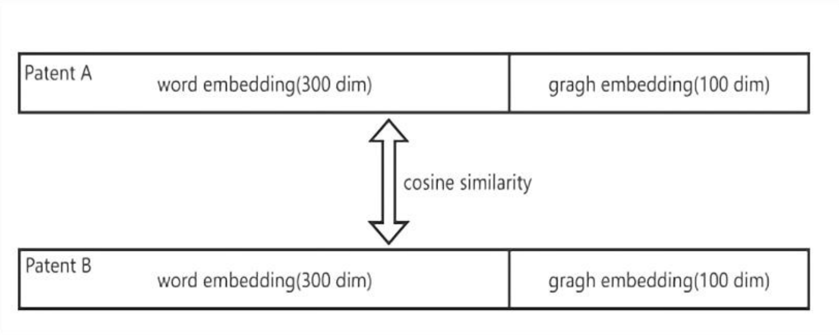

Figure 5. Cosine similarity of word vectors with fused knowledge graph embeddings and word embeddings.

Using this algorithm, the TOP10 similar patents of the example patent in the previous section are obtained as shown in Table 7.

Table 7. Entity similarity of patent titles integrating knowledge graph embeddings and word embeddings.

Similar Patent Titles	Similarity	IPC
Control method and control device of energy storage charging pile system	0.8845	H02J3/32
An energy storage power supply based on parallel output to increase output power	0.8635	H02J7/00
An automatic switching control method for charging and discharging mode of energy storage converter	0.8513	H02J3/32
A battery balancing control method and device based on MPC	0.8486	H02J7/00
A mobile self-circulating reserve integrated energy guarantee system	0.8456	H02J7/34
High-power charging cable for new energy vehicle charging and how to use it	0.8358	H01B7/42
Battery power system of new energy ship and its control method	0.8215	H02J7/00
A backup power circuit	0.8186	H02J7/34
A backup power supply, vehicle and vehicle control method	0.8064	H02J7/18
Control methods, devices and electronic equipment based on flywheel energy storage system	0.7934	H02J3/30

As we can see, this patent recommendation algorithm can more accurately recommend patents in similar fields, while considering the text similarity of patent titles.The algorithm fuses the text features of patent data titles and the entity relationship features in the patent knowledge graph and can more accurately identify patent data collections with similar research fields and similar patent texts, which is of great help to researchers in querying and recommending similar patents. The experimental results have also been externally verified. After consulting with patent agents, the patent recommendation results obtained in this paper are reasonable, and the scope of technical fields can be determined with enough data. From this perspective, it can save researchers' time in searching similar patents.

5. CONCLUSION

In this paper, we take the electric power as the research direction. By extracting the entity and relationship data of electric power field patents and storing them in the knowledge graph in the form of a triad, we can obtain better results by fusing the text vector features in the patent data and the patent relationship vector features in the knowledge graph, and then calculate the similarity between patents. The similar patent pairs calculated by this algorithm can be used for patent recommendation or real-time pushing, which will save the time of researchers in finding patents in similar fields. Also, this method can be used in the construction of patent maps in all fields, not only in the field of electricity. In the future work, we can consider increasing the types of entities and relationships in patent mapping to enrich the features of patents and thus obtain more accurate results.

REFERENCE

[1] Auer, S., Kovtun, V., Prinz, M., Kasprzik, A., Stocker, M., Vidal, M. E., Towards a knowledge graph for science. In Proceedings of the 8th international conference on web intelligence, mining and semantics (pp. 1-6) (2018).

[2] Deng, L., Cao, C. G., A method for building a patent knowledge graph," computer science, 49(11), (2022).

[3] Kejriwal, M., Sequeda, J. F., Lopez, V., Knowledge graphs: Construction, management and querying[J]. Semantic Web, 10(6): 961-962 (2019) .

[4] Lu, C. C., Liu, Y. S. Y., Gu, F., Gu, X. J., etc., Research on the construction of API international drug registration knowledge map based on Neo4j graph database. Group Technology and Production Modernization. 37(04),1-44 (2020).

[5] Yin, D. Y., Research on the construction of police patrol knowledge graph [D]. Beijing: People's Public Security University of China (2019).

[6] Nickel, M., Murphy, K., Tresp, V., Gabrilovich, E., A review of relational machine learning for knowledge graphs. Proceedings of the IEEE, 104(1), 11-33 (2015).

[7] Suchanek, F. M., Kasneci, G., Weikum, G., Yago: A large ontology from wikipedia and wordnet. Journal of Web Semantics, 6(3), 203-217(2008).

[8] Song. H. Y., Research and application of power system knowledge graph based on graph database. Master of Electronic Journal Publishing Information: Period: Issue 08 (2021).

[9] Wang, F., Yi, M. Z., Tan, X., et al. Research on the storage method of Tibet-related domain ontology based on Neo4j [J]. Journal of Zhengzhou University (Science edition), 51 (2): 60-65 (2019).

[10] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O., Translating embeddings for modeling multi-relational data. Advances in neural information processing systems, 26 (2013).

[11] Zhu, M., Zhen, D. S., Tao, R., Shi, Y. Q., Feng, X. Y., Wang, Q., Top-N collaborative filtering recommendation algorithm based on knowledge graph embedding. In Knowledge Management in Organizations: 14th International Conference, (pp. 122-134). Springer International Publishing (2019).

[12] Mikolov, T., Chen, K., Corrado, G., Dean, J., Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013).