IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 27, 2020, accepted September 9, 2020, date of publication September 14, 2020,
date of current version September 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3024070

Research on Information Extraction of Technical Documents and Construction of Domain Knowledge Graph

HUAXUAN ZHAO, YUELING PAN, AND FENG YANG
School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China

Corresponding author: Feng Yang (yangfeng@sdnu.edu.cn)

This work was supported by the Natural Science Foundation of China under Grant 61801277 and Grant 61373081.

ABSTRACT With the rapid development of knowledge graph related technologies, domain knowledge graph has become a research hotspot in academia and industry. However, the domain knowledge graph for technical documents is not mature enough, and the semantic information implicit in unstructured technical documents has not been fully tapped. Combining the characteristics of technical documents, the paper proposes a TextCNN-based topic information extraction model and constructs a domain knowledge graph for technical documents. It uses the graph database Neo4j for knowledge storage and visualization. The information extraction model based on TextCNN can automatically extract the subject information of the document and the summary information such as title, ID, status, meeting, organization, etc. Experiments show that the model has high accuracy on the technical document dataset, which can effectively reduce the cost of manual annotation and data collation. At the same time, knowledge graph visualization can facilitate scientific researchers to search, track and update technical documents, which can show the evolution of technology more clearly.

INDEX TERMS Domain knowledge graph, information extraction, graph database, TextCNN, Neo4j, resource retrieval.

I. INTRODUCTION

Knowledge is the cornerstone of cognitive intelligence, making it possible to explain artificial intelligence. As a large-scale semantic network, knowledge graph has become an important form of knowledge representation in the era of big data. The main goal of the knowledge graph is to describe the various entities and concepts that exist in the real world, and the various semantic relationships between them. Compared with traditional data management methods, knowledge graphs can efficiently obtain the logical relationship between knowledge. What's more, knowledge graphs can facilitate the use of visualization techniques to show semantic associations and have strong explanatory power. In recent years, intelligent question answering, search engines, precision marketing, and decision support based on the general knowledge graph have achieved rapid development. And the domain knowledge graph is oriented to a specific industry field, focusing on the accuracy and depth of knowledge. Therefore, the domain knowledge graph has a stronger professionalism. However, compared with the vigorous development of general knowledge graphs, it is not mature enough at the application level. The process of constructing professional knowledge graphs requires the participation of a large number of domain experts. Therefore, the construction of knowledge graphs for specific professional fields has great research value.

With the evolution of the new generation of information and communication technologies, the key technologies of 5G have been rapidly developed. As the world's most important mobile communication technology standardization organization, 3GPP (The 3rd Generation Partnership Project) is becoming larger and larger as the technology evolves. A technical standard was born when 3GPP screened out valuable proposals from tens of thousands of proposals, and then the working group decided on a time for collective discussion. From the 4G to 5G era, the number of key technical document contributions has more than quadrupled. However, not all technical proposals have equal value and the impact of a single technical proposal is difficult to

The associate editor coordinating the review of this manuscript and approving it for publication was Gianmaria Silvello.

assess. Standardization work of 3GPP is inseparable from the follow-up of technical proposals, thus it is particularly important to process and integrate massive amounts of unstructured text data. However, manual data labeling and sorting are time-consuming and labor-intensive. The classification of topics is more dependent on the expertise of experts and the traditional classification method can not capture the local semantic information of the context well. How to mine the semantic information hidden in the text and then extract the topic has become an urgent problem to be solved in the information extraction part. At the same time, tens of thousands of technical proposals are difficult to search and query. Therefore, it is urgent to build a domain knowledge graph based on the technical proposal documents.

This paper combines the characteristics of technical proposal documents to process unstructured proposal document data, and builds an information extraction model based on TextCNN. The model can realize the automatic extraction of summary information such as subject, key technology, title, proposal status, document source, and agenda item of technical proposal documents, which can effectively reduce the cost of manual annotation and data collation. At the same time, a domain knowledge graph for technical documents is designed and Neo4j is selected as the graph storage tool for visualization, which is convenient for researchers to optimize queries and update proposals. The constructed academic knowledge graph can provide accurate information resource retrieval and query recommendation for scholars and researchers, which can more clearly and intuitively show the association of technical proposals and the evolution of key technologies. This paper is organized as follows. Section II introduces related work. Section III describes the TextCNN-based topic information extraction model for the proposed document. Section IV displays the construction of domain knowledge graph for technical documents. And section V draws conclusions and describes avenues for future work.

II. RELATED WORK

Knowledge graph is a large-scale semantic network, and it is a representative product of knowledge engineering in the era of big data. The related research originated in the middle of the twentieth century and developed from knowledge engineering and semantic network. In 1955, Swanson [1] proposed the construction of a document map based on the co-citation relationship of documents, and the use of citation indexes to achieve document retrieval. This was the first time that a citation network was used to study the development of the literature. In 1997, Feigenbaum put forward the concept of knowledge engineering, using expert knowledge and reasoning ability to build an expert system [2]. The knowledge base system represented by the expert system has been widely studied and applied. In 1998, Semantic Web [3] was proposed by Tim Berners Lee, the father of the World Wide Web, using nodes and edges to describe the relationship between resources and data in the World Wide Web, providing a knowledge representation that can be understood and processed. In 2012, Google [4] took the lead in proposing the concept of ''knowledge graph'', building a search engine based on semantics, and truly proclaiming that knowledge engineering entered the era of big data. In recent years, with the increasing demand for Internet applications, knowledge graph technology driven by big data and featuring automated knowledge acquisition has developed rapidly, and more and more knowledge graphs have emerged.

Typical knowledge graphs Cyc [5] and WordNet [6] are manually edited and implemented by a team of experts. As a representative of early open domain knowledge graphs, they have high knowledge accuracy. Freebase [7] was founded by MetaWeb in 2005. Its knowledge is stored in the form of resource description framework (RDF) triples, and the knowledge graph is constructed using information from websites such as Wikipedia and NNDB. The multilingual knowledge graph DBpedia [8] and YAGO [9] successively proposed in 2007 to extract structured knowledge from Wikipedia for public use, but its scale is limited by website data and cannot meet the rapidly growing information update requirements. Google Knowledge Graph [4] was released in 2012, which can truly understand the user's search intent and find the results that best meet the user's needs. It is considered a major innovation of the search engine. In the same period, the general knowledge graph ''Sougou Zhicube'' and Baidu ''Zhixin'' were launched, and the knowledge graph module-Zhicube and Baidu Zhixin were added to the search engine, bringing a new user experience to the search engine. In 2015, the Knowledge Workshop of Fudan University developed a large-scale open Chinese general knowledge graph CN-DBpedia [10], which extracts information from the semi-structured pages of Chinese encyclopedia websites, and is formed through operations such as knowledge fusion, knowledge reasoning, and knowledge processing.

At present, the in-depth integration of the knowledge graph with various fields and industries has become an important trend. In terms of domain knowledge graphs, representative studies abroad include GeoNames [11], which was an open global geographic knowledge graph, covering more than 250 countries and more than 10 million pieces of geographic location information. It provides users with free API interfaces and is widely used in systems in various industries. At the same time, IMDB [12] in the film and television field and DBLife [13] in the academic field have been developed successively, demonstrating the application advantages of the knowledge graph in specific industries. Related research teams in China have also constructed some domain knowledge graph, including Alibaba's e-commerce cognitive map and Tsinghua University's AMiner. AMiner [14] builds an association relationship based on three types of data: scientific researcher, scientific research literature, and academic activities. It focuses on scholars' graphs, and provides academic information resource retrieval and semantic search for scholars, papers and literature. In addition, the construction of the domain knowledge graph has good application

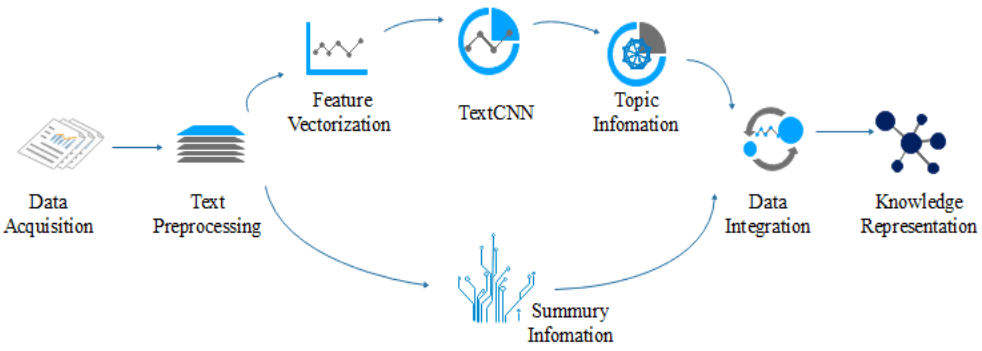


FIGURE 1. Information extraction model based on TextCNN.

prospects in the financial field, medical field, and agricultural field.

III. INFORMATION EXTRACTION FOR TECHNICAL PROPOSAL DOCUMENTS

Information extraction [15], also known as knowledge extraction, is an important part of constructing a knowledge graph. Information extraction is to extract the knowledge contained in the information source through the processes of identification, understanding, discovery of rules and screening. Based on the given data model, it extracts relevant entity attributes, relationships, etc. from the text to form knowledge and stores it in the database. It is an effective way to obtain knowledge. Because the domain knowledge graph has a high-quality vocabulary containing professional terms, it can reduce the high dependence on entity recognition and entity disambiguation when extracting information. High-quality technical proposal resources make automated information extraction possible.

The information extraction of the domain knowledge graph in this paper is based on the content of the technical proposal document, and the title, number, status, conference and source organization etc. are extracted as the basic summary information on the basis of the rule base. At the same time, the model extracts the topic category information from the body part as the semantic feature of the technical proposal, and attaches it as one of the attributes of the proposal entity, which is used to describe the semantic topic of a single technical proposal document.

TextCNN-based information extraction model includes five steps: data acquisition, text preprocessing, text feature vectorization, TextCNN topic classification, and data integration. It combines with the traditional rule-based knowledge extraction method to get a preliminary knowledge representation form, as shown in Figure 1. Among them, the text feature vectorization part uses Word2Vec's CBOW model for training to obtain the word vector representation, and the topic classification selects the TextCNN model. The specific experiments and results analysis section will be shown later.

A. TECHNICAL PROPOSAL DOCUMENT DATA ANALYSIS

As an important information carrier, the technical proposal is one of the important forms to organize the conference work.

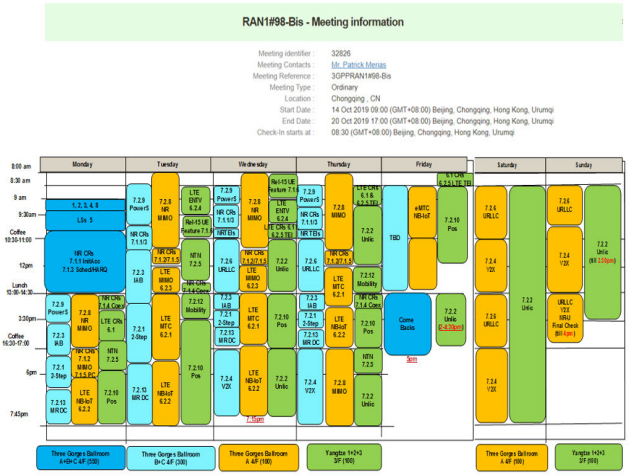


FIGURE 2. Agenda distribution of 3GPP RAN1#98-Bis.

The 3GPP technical proposal document is a series of related technical documents that are intensively discussed during the 3GPP conference. Submitting proposals is an effective way to lead the discussion direction of the meeting in advance and express the demands. As an important resource for technological development in the communications field, technical proposal documents play an indispensable role in promoting the formulation and improvement of industry standards.

Among them, the relevant technical proposal documents involved in the meeting described the problems of the key technologies of the communication protocol and their improvement measures. Taking the 3GPP RAN1 # 98-Bis meeting as an example, the meeting agenda was discussed based on the proposal documents under different topics, involving a total of 1812 technical proposal documents. The distribution of conference information and conference agenda topics is shown in Figure 2. Take the proposal numbered R1-1910434 as an example, the attributes of the proposal and the conference information to which it belongs are shown in Figure 3. The body part is generally composed of three elements: introduction, problem discussion, and conclusion, which are used to describe technical details, design implementation, and update and improve the program. Although

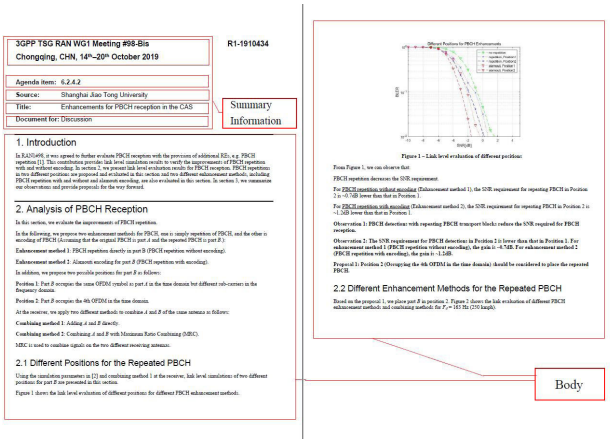


FIGURE 3. The structure of the proposal document.

there are subtle differences in the writing format of different researchers, they are all discussions on key technical topics around the core technical topics covering the comments and final suggestions. Therefore, the body part can be used as an important source of text semantic information and topic category extraction.

B. DATA ACQUISITION AND TEXT PREPROCESSING

The source of the corpus data in the data acquisition part is the unstructured technical proposal document. This paper selects all the technical proposal documents involved in the 3GPP RAN1 # 98-Bis conference in the field of communications. The total number of documents is 1812. The resources of the technical proposal documents are downloaded from the 3GPP website.

The text preprocessing section includes common text preprocessing steps such as batch document format conversion, encoding format conversion, and stem extraction. Due to the inconsistency of the data source text format, batch format conversion operations must be performed on the obtained pdf and word documents, and the two formats of documents should be converted into txt documents that are easy to read. These documents will be saved in a unified UTF-8 encoding format. Then there are processing steps such as case normalization, removal of stop words, stem extraction, and word form reduction to facilitate the next step of extracting summary information and information in the body part. Among them, the word stem extraction and word form reduction use the NLTK to find the original form of the word.

C. TEXT VECTORIZATION

The purpose of text representation is to convert the pre-processed text into a computer-understandable way. This is the most important part of determining the quality of text classification. The distributed representation of the word (word embedding) is an important foundation for deep learning methods. The word2vec proposed by Mikolov [16] in 2013 is a neural network probabilistic language model that embeds text data from one-hot sparse high-latitude vectors into low-dimensional continuous dense features by means of word vector representation Space, and make the word vector carry certain context information. Word2vec includes two neural network structures, continuous bag of word (CBOW) and Skip-Gram model. The idea of the CBOW model is to predict the central word through the context. The input is the context of a certain central word, and the output is the word vector of the central word predicted by the model. The CBOW model calculates the parameter matrix of the hidden layer through the training of Deep Neural Networks (DNN) with a hidden layer, so as to extract the word vector representation corresponding to each word from the parameter matrix. The goal of the training process is to maximize the log-likelihood function, as shown by the following equation.

$$L = \sum_{w \in c} \log P(w|\ \text{content}\ (w)) \tag{1}$$

where $w(t)$ represents the current word, taking (context(w), w) as an example, context(w) is the sequence of adjacent words of $w(t)$, which length is c. The number of input words is determined by the hyperparameters related to the size of the sliding window. The Skip-gram model is similar to the CBOW model and it uses the central word to predict the context, which is structurally opposite to the CBOW model. The network structure of the two models is shown in Figure 4.

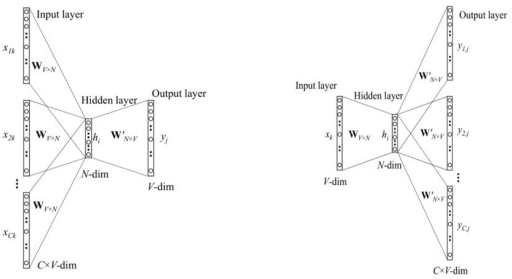


FIGURE 4. CBOW model and Skip-gram model.

In addition, Word2Vec proposed two strategies for Huffman coding: Hierarchical Softmax and Negative Sampling, and optimized the time complexity and efficiency, which solved the problem of computational validity very well. This papper uses word2vec in Google and the continuous word bag CBOW model for training. The word vector dimension is set to 128, and words that do not exist in the pre-training process are randomly initialized.

D. CLASSIFICATION ALGORITHM BASED ON TEXTCNN

Convolutional Neural Network [17] initially achieved great success in the field of image. Its core point is that it can capture local correlations. Subsequently, more neural networks were derived, such as MSIN [18], which realized multi-sample classification. Therefore, in the text classification task, it is of great significance to consider the local sequence information of words. In the text classification

problem of modeling conversion in this paper, we can use TextCNN [19] (Convolutional Neural Network for Sentence Classification) to extract the key information in the text sequence. Particularly, Multiple convolution kernels are used to obtain the semantic local correlation between the words.

The structure of TextCNN is divided into: input layer, convolutional layer, pooling layer, fully connected layer and output layer. Its principle is to convert each word of the original input text into a fixed-length vector representation, and convert the one-dimensional text sequence into a two-dimensional input matrix through the word vector. Then, through the sliding window of the convolution kernel to perform convolution to extract local features, we can design the convolution kernel with different convolution kernel size (Filter_size) to obtain the visual field of different widths, which is used to capture the original text Information on the local characteristics and order. Finally, the most important features are extracted by polling, and softmax is used to classify at the fully connected layer and output the probability of each category.

1) INPUT LAYER

Represent each word as a $k-$dimensional word vector, where k is the dimension of the word vector corresponding to each word. $xi \in \mathbb{R}^k$ represents the k-dimensional word vector of the i-th word in the corresponding sequence. Then the corresponding text sequence of length n can be expressed as

$$x_{1:n} = x_1 \oplus x_2 \oplus \ldots \oplus x_n \tag{2}$$

where $\oplus$ is the concatenation operator. Therefore, a text sequence composed of n words can be converted into an input matrix through word embedding, and the size of the input matrix is n^*k. If the number of words corresponding to the text sequence is less than n, it will be filled with 0, and the excess will be discarded.

$$c_i = f\left(w \cdot x_{i:i+h-1} + b\right) \tag{3}$$

2) CONVOLUTION LAYER

Convolution kernel of size h^*k and data matrix of input layer are used for convolution. The convolution kernel is the basic unit of feature extraction, and its size determines the size of the sliding window during feature extraction which represents the length of the text sequence. It can be used to capture the correlation between words in the text sequence and plays a vital role in the accuracy of the model. Each time the convolution kernel $w \in \mathbb{R}^{hk}$ acts on the input data matrix $\mathbf{x}_{i:i+h-1}$ to generate the feature C_i, as shown by following equation.

Where h represents the sliding window range, and $\mathbf{x}_{i:i+h-1}$ represents the sliding matrix window of size h^*k composed of the i-th row to the $n-h+1$ row of the input matrix. w has the same dimensions as $\mathbf{x}_{i:i+h-1}$, which is also h^*k, and b is the offset parameter. f is a nonlinear activation function, and c_i is a scalar. Therefore, the convolution kernel acts on the text sequence of length n to obtain $n-h+1$ results, and the set of feature vectors can be expressed as c, as shown in the following formula, where $c \in i^{n-h+1}$, c_i represent the i-th feature vector.

$$c = [c_1, c_2, \cdots, c_u, \cdots, c_{n-h+1}] \tag{4}$$

It is worth noting that the activation function, as an important part of the neural network, introduces nonlinear factors to the neural network. At present, the activation functions commonly used in convolutional neural networks are sigmoid function, tanh function and ReLU function. The ReLU function can effectively avoid the problems of gradient explosion and gradient disappearance, so the ReLU function is selected as the nonlinear activation function, and its formula is as follows.

$$f(x) = \max(0, x) \tag{5}$$

3) POOLING LAYER

We have adopted the max-polling method. That is, the largest one is selected from the feature vector values generated by each convolution kernel in the upper layer as the feature value, as shown by following equation.

$$\hat{c} = \max\{c\} \tag{6}$$

Finally, all the values are spliced to form a one-dimensional feature vector, and the maximum pooling results of each convolution kernel are combined for the next step.

4) FULLY CONNECTED LAYER

The spliced features will be transferred to the fully connected softmax layer. The output of the softmax layer is the probability distribution on the label, which acts as a classifier. The fully connected layer function ϕ is given by

$$\phi\left(x_i, W_Z\right) = W_Z \hat{f} \tag{7}$$

where W_z is the weight of the fully connected layer and $\hat{f}$ is the output of the activation function of the previous layer.

Finally, the softmax function converts the classification result into a probability distribution as follows, where c_j is the $j-$th classification category.

$$P\left(c_j | x_i, W_Z\right) = \frac{\exp\left(\phi\left(x_i, W_Z\right)\right)}{\sum_{j=1}^{|c|} \exp\left(\phi\left(x_i, W_Z\right)\right)} \tag{8}$$

The convolutional neural network used for text classification [20] is shown in Figure 5. From the figure, the principle of the text classification convolutional neural network with two channels and three types of convolution kernels can be further clearly and specifically shown.

The text of the input layer undergoes word2vec pre-trained word vectors to form an input matrix, and then passes through a one-dimensional convolution layer with Filter_size= (2, 3, 4). Each convolution kernel has two output channels, and different results of convolution kernels of different sizes are placed separately after the convolution operation. The third layer is a maximum pooling layer, and sentences of different lengths will become fixed-length representations

TABLE 1. Distribution of subject categories in the dataset.

Item	Category	Agenda item	Number
0	5G V2X with NR sidelink	7.2.4	266
1	NR-based Access to Unlicensed Spectrum	7.2.2	257
2	Physical Layer Enhancements for NR URLLC	7.2.6	208
3	Maintenance of Release 15 NR	7.1	137
4	Enhancements on MIMO for NR	7.2.8	136
5	Additional MTC Enhancements	6.2.1	100
6	UE Power Saving for NR	7.2.9	97
7	Incoming Liaison Statements	5	93
8	Study on solutions for NR to support Non Terrestrial Network (NTN)	7.2.5	87
9	Multi-RAT Dual-Connectivity and Carrier Aggregation enhancements (LTE, NR)	7.2.13	87
10	NR positioning support	7.2.10	85
11	Two step RACH for NR	7.2.1	72
		Total	1625

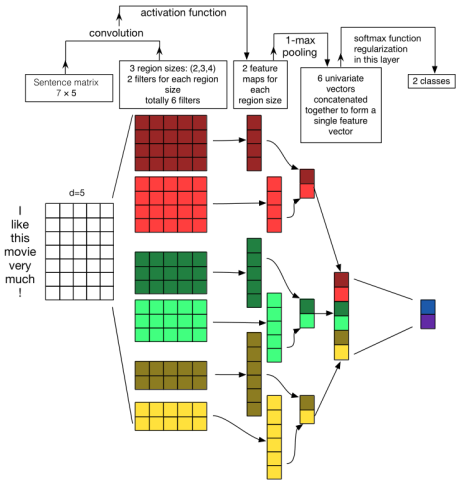


FIGURE 5. The Structure of TextCNN.

after passing through this layer. The last layer is the fully connected layer, which uses the Softmax function to output the probability of each text category.

Due to the large number of parameters when training a small data set, it is easy to produce overfitting. Therefore, dropping can reduce the interaction effect of the neurons in the network, thereby improving the generalization ability of the neural network. When the network is performing the forward propagation in training, certain neurons stop working with a certain probability (dropout), so that each training only updates a part of the neuron parameters, thereby reducing the network's dependence on local features to prevent overfitting problems. As one of the tuning parameters, the discarding rate macroscopically determines the proportion of neurons updated in each training. Therefore, in the model, some feature neurons are randomly discarded with a probability of $p = 0.5$, and the discarding rate is 0.5.

E. EXPERIMENTS

1) DATASET

In order to test the performance of TextCNN-based information extraction model when applied to the topic classification of proposal documents, this paper collects all the technical proposal documents involved in the 3GPP RAN1 # 98-Bis conference to construct a corpus. The unstructured proposal document resources are downloaded from the 3GPP website for a total of 1812 articles. This paper also obtains the report of the meeting, which covers the category information of the meeting agenda. According to different levels of fine-grained division, 117 categories are divided into finer granularity, and 21 categories are divided into coarser granularity.

Due to the uneven distribution of data, we must first filter the technical proposal documents in the corpus and data that does not significantly contribute to the topic classification. We have filtered the proposals classified as ''other'' whose semantic themes are not yet clear, and the topic categories where the number of proposals involved in the conference is too small. This paper finally selects twelve categories and a total of 1625 proposals to construct our proposal data set. For the sake of clarity, Table 1 statistics the various theme categories and numbers on the entire data set. The distribution of topic categories is shown in Figure 6.

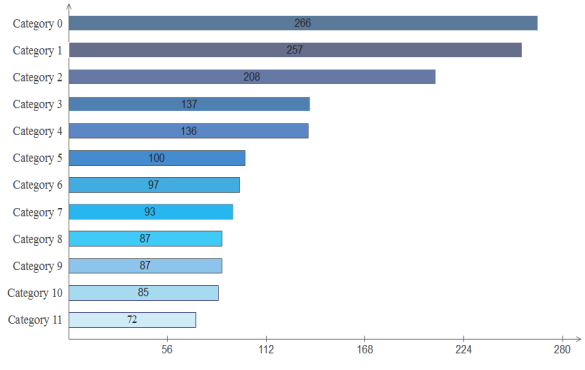


FIGURE 6. Distribution of subject categories in the dataset.

2) EVALUATION INDICATORS

Classification models usually use factors such as precision, recall and F1 score (harmonic average) as the evaluation

indicators to measure the model performance. Among them, the F1 score is a weighted average of precision and recall, which can be regarded as a comprehensive consideration of precision and recall. Its mathematical definition is shown in following formulas.

$$\text{precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{recall} = \frac{TP}{TP + FN} \tag{10}$$

$$F1_{\text{score}} = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{11}$$

Among them, P (Positive) and N (Negative) represent the number of positive and negative samples in the sample. The meanings of TP, FP, TN and FN are shown in the following table.

TABLE 2. Confusion matrix.

	Predition Value=0	Predition Value=1
Actual Value=0	TN	FP
Actual Value=1	FN	TP

3) EXPERIMENTAL DESIGN

In order to effectively verify the performance of TextCNN-based information extraction model, especially the effectiveness of topic classification and applicability on the proposed data set, this paper designs experiments and evaluates the classification results on the above-mentioned constructed dataset. In order to ensure the validity of the experiment, we randomly divided the entire data set into five parts, of which four were selected as the training set and one as the validation set. We also need to use the above three evaluation indicators of precision, recall, and F1 score to conduct experiments, and take the average of five classifications results as the final result. At the same time, three kinds of classification algorithms commonly used in machine learning are selected for comparison experiments with TextCNN, namely K-nearest neighbor, logistic regression algorithm and SVM classification algorithm. The input word vector is unified as Word2Vec, which is obtained by Gensim module in Python. And we compared the classification performance of TextCNN and FastCNN as well. FastCNN was run with the same parameters in [21], which has 10 hidden units with bigrams.

This paper uses Tensorflow to complete the construction of TextCNN. The experiment was carried out on Pycharm2017.3.3 (Community Edition), and the PC used was Intel (R) Core (TM) i5-4200H CPU 2.80 GHz, memory 4.00 GB.

4) EXPERIMENTAL PARAMETER SETTINGS

Since TextCNN requires the input text data to have a fixed length, the maximum length is set to 600. When it is less than the maximum length, a zero vector is used for completion. The length of the word vector in this paper is 128 dimensions.

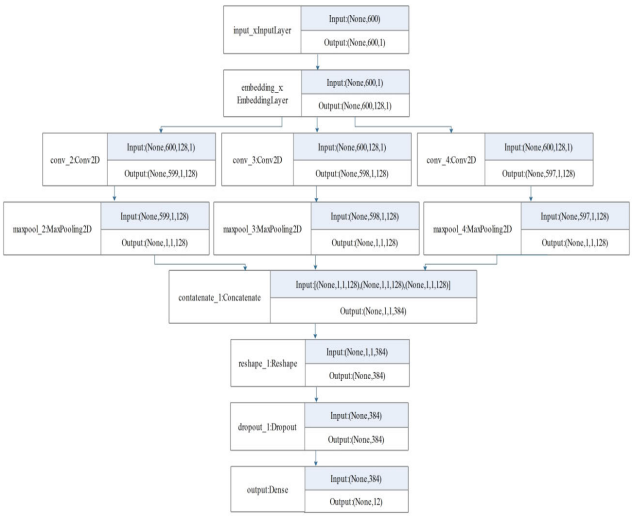


FIGURE 7. Data dimension of input and output of each layer of TextCNN.

In order to reduce the number of iterations during training, we train by training one batch at a time, where the batch size is 64 and the maximum number of iterations epoch is set to 30. Figure 7 shows the data input and output dimensions of each layer. And the parameter settings are shown in Table 3.

5) ANALYSIS OF RESULTS

According to the above experimental design, the experimental results of the three traditional machine learning classification algorithms, FastCNN and TextCNN on the data set are shown in Table 4, which details precision, recall and F1 score of the five algorithms. Among the three machine learning algorithms, SVM algorithm and logistic regression algorithm have higher precision and F1 score than K-nearest neighbor algorithm. TextCNN's three classification evaluation indicators all maintain the highest, especially in terms of precision, which has a 4.55% improvement over the logistic regression algorithm. In addition, we compare FastCNN and TextCNN. Although the training time of FastCNN is short, the classification result can be obtained quickly. For our fine-grained proposal dataset, the precision, recall and F1 score of TextCNN are higher. Compared with traditional machine learning classification algorithms and FastCNN, TextCNN shows better classification performance on the proposed dataset.

IV. CONSTRUCTION OF DOMAIN KNOWLEDGE GRAPH FOR TECHNICAL DOCUMENTS

Knowledge graph technology describes the concepts, entities and their relationships in the objective world in a structured way, and uses relationships to describe the association with two entities. As an important part of artificial intelligence, knowledge graph can provide a better way to organize and understand management information and knowledge, and transform the information expression on the Internet into a form closer to the human cognitive world.

TABLE 3. Parameters of TextCNN.

Parameter Name	Parameter Value	Parameter Description
Max_sentence_LENGTH	600	Maximum sentence length
Embedding_DIM	128	Dimension of word vector
Classes	12	Number of categories
Filter_size	[2,3,4]	The size of the filter
Num_fiters	128	The number of filters of each size
Batch_size	64	Number of training samples per batch
Epoch	30	Number of iterations
Dropout	0.5	Probability of randomly discarded neurons

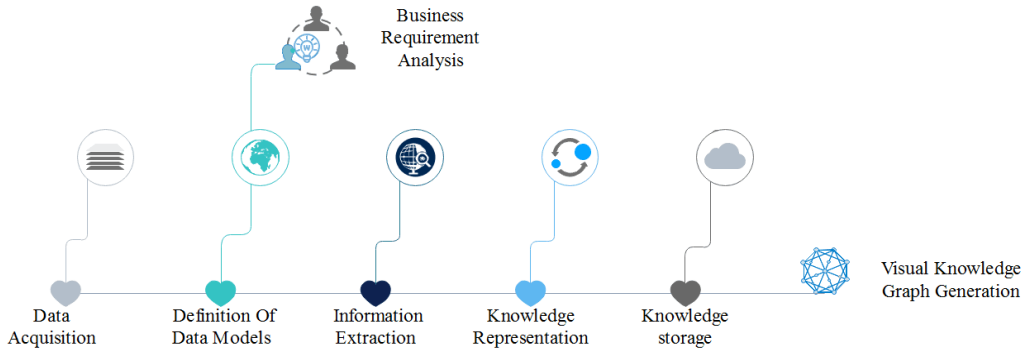


FIGURE 8. Construction of domain knowledge graph.

In this paper, the proposal-oriented domain knowledge graph aims to use the knowledge graph to describe the objective world, to model and process 3GPP related technical proposal resources in the communications field, and to make full use of the semantic processing of the knowledge graph. The main goal of constructing this knowledge graph in this paper is to realize the definition of the proposed entity and the multi-dimensional attributes attached to the entity, and demonstrate relationships with other related entities.

The construction process of domain knowledge graph includes 7 steps including data acquisition, business requirement analysis, data model definition, information extraction, knowledge representation, knowledge storage and visual knowledge graph generation. The flow chart is shown in Figure 8.

A. DEFINITION OF DATA MODELS

The knowledge representation of the knowledge graph is embodied through the data model, which mainly includes entities, entity attributes, and relationships between entities. In terms of domain knowledge graphs, knowledge representation needs to meet actual business needs and industry standards, and it can also solve the problems of multi-source heterogeneous data sources and poor compatibility in the process of data mapping. As a data organization framework for knowledge graphs, the definition of data models is particularly important. The entity is the core of the data model, and its definition and naming play a vital role in the process of constructing the knowledge graph model.

TABLE 4. Experimental comparison results.

Indicators	Method				
	KNN	LR	SVM	FastCNN	TextCNN
Precision	0.7788	0.8649	0.8333	0.8947	0.9104
Recall	0.8605	0.8421	0.8572	0.8641	0.8696
F1 score	0.8132	0.8533	0.8511	0.8694	0.8777

The data model defines five types of entities, namely TDoc, Member, Organization, Meeting, and Category. The model also introduces the attribute information contained in the entity. For example, the attributes included in the proposed entity are title, TDoc_ID, status, type, release, agenda item, etc. As the core entity type, the proposed entity and the other four types of entities should establish contacts. It is also worth noting that there is also a subordinate relationship between members and organizations. The definition and display of various entities, attributes and relationships are given in Tables 5 and 6.

B. KNOWLEDGE STORAGE

Graph database, also known as graph database management system, refers to a database that uses graph theory to store entities. The use of a simple and intuitive graph management system to uniformly manage and use knowledge graphs can effectively reduce the application thresh-

TABLE 5. Display of various entities and attributes.

Entity	Attributes
TDoc	TDoc_ID Title Status Type Release Agenda_item For
Member	Member_name Contact_ID
Orgnization	Organization_name
Category	Category_description
Meeting	Meeting_name Meeting_place Meeting_date

TABLE 6. Display of relationships between various entities.

Relationship	Entity1	Entity2	Description of relationship
REVISED_OF	TDoc	TDoc	Revision relationship between two TDocs
CONTACT_OF	TDoc	Member	Relationship between TDoc and Member
ORGANIZATION_OF	TDoc	Organization	Relationship between TDoc and Organization
CATEGORY_OF	TDoc	Category	Relationship between TDoc and Category
MEETING_OF	TDoc	Meeting	Relationship between TDoc and Meeting
BELONGING_TO	Member	Organization	Relationship between Member and Organization

old of knowledge graphs. At the same time, it can achieve higher query efficiency when processing complex relationships and entity mappings, and it is more humanize. Common graph databases include Neo4j [22], AllegroGraph, FlockDB, and GraphDB, among which Neo4j has become one of the most popular graph data management systems. Neo4j is a high-performance NOSQL graph database that stores structured data in the form of native graphs, enabling efficient graph data operations. At the same time Neo4j supports low-latency online query, which can be better suited for real-time applications. Therefore, this paper selects Neo4j to visualize the domain knowledge graph, which can more clearly and intuitively show the association of technical proposal documents in the professional field. It can also facilitate researchers to optimize query and retrieval, and has high application feasibility.

Neo4j tool is used to perform data storage based on the graph database. Neo4j organizes data with nodes (Vertex) and edges (Edge). The data structure can be expressed as a graph G = (V, E), where V represents a node set and E represents a set of edges. We use Neo4j to create entities and relationships, with nodes representing entities in the knowledge graph and edges representing relationships between entities. The relationship edge is directed in Neo4j. We need to pay attention to specify the start node and end node when creating a relationship.

Neo4j uses Cypher language for database storage and data query in knowledge graph [23]. As a descriptive graph query language, Cypher has become the standard of graph query language due to its simple syntax and powerful functions. This paper compares several ways of importing data, such as Neo4j-import, admin-import, Load CSV. Among them, the Load CSV module is suitable for small and medium-sized databases, without offline import. Data import involves five entity node corresponding attribute tables and six inter-entity relationship tables.

The knowledge graph MyGraph after loading data for knowledge storage which has 5 types of entity nodes, a total of 2295 nodes, 6 relationship types, and a total of 7683 relationships.

C. VISUAL DISPLAY

Information visualization is based on computer technology to visually express and display abstract data. It supports interaction and can help people quickly understand the meaning behind the data. Visually displaying the entities, attributes and relationships of the knowledge graph is an important part of the domain knowledge graph. The use of visualization technology can improve the user experience and show the data connection. Especially the display of multi-dimensional related information between multi-dimensional data is more clear and intuitive. Neo4j can provide a visualization function while storing knowledge. The constructed knowledge graph is shown in Figure 9, in which different entities are displayed with nodes of different colors. Here we added constraints to show only the relevant attributes and relationships of 500 entity nodes.

To query the data attributes and relationships of entity nodes in the graph database Neo4j, we take the proposal numbered R1-1910434 as an example. The display interface is shown in Figure 10. After inquiring about the splicing relationship of proposal R1-1910434, we can retrieve the proposal entity R1-1911348 which has a ''revision relationship'' with it. According to the links between the proposals, we can further analyze the development and evolution of key technologies. At the same time, various entities related to proposal R1-1910434 and their relationships can also be clearly displayed.

Finally, we still take the proposal numbered R1-1910434 as an example, the corresponding entity relationship of the domain knowledge graph is shown in Figure 11. The proposal title is ''PBCH repetition for enhancement in the CAS''. In addition to the number and title, it also contains four data-type attributes: its proposal status is ''revised'', the release version is ''Rel-16'', the agenda item is

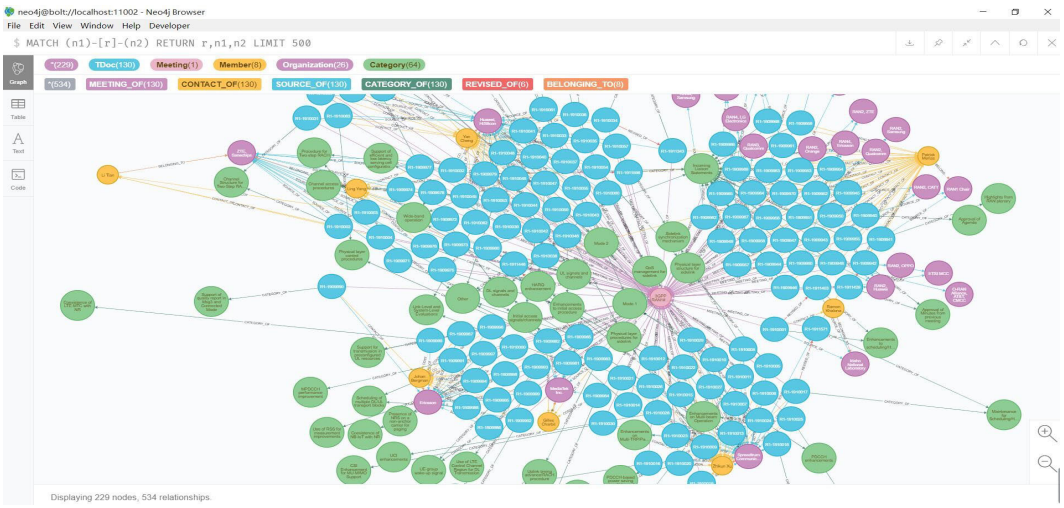


FIGURE 9. Neo4j knowledge graph visualization display.

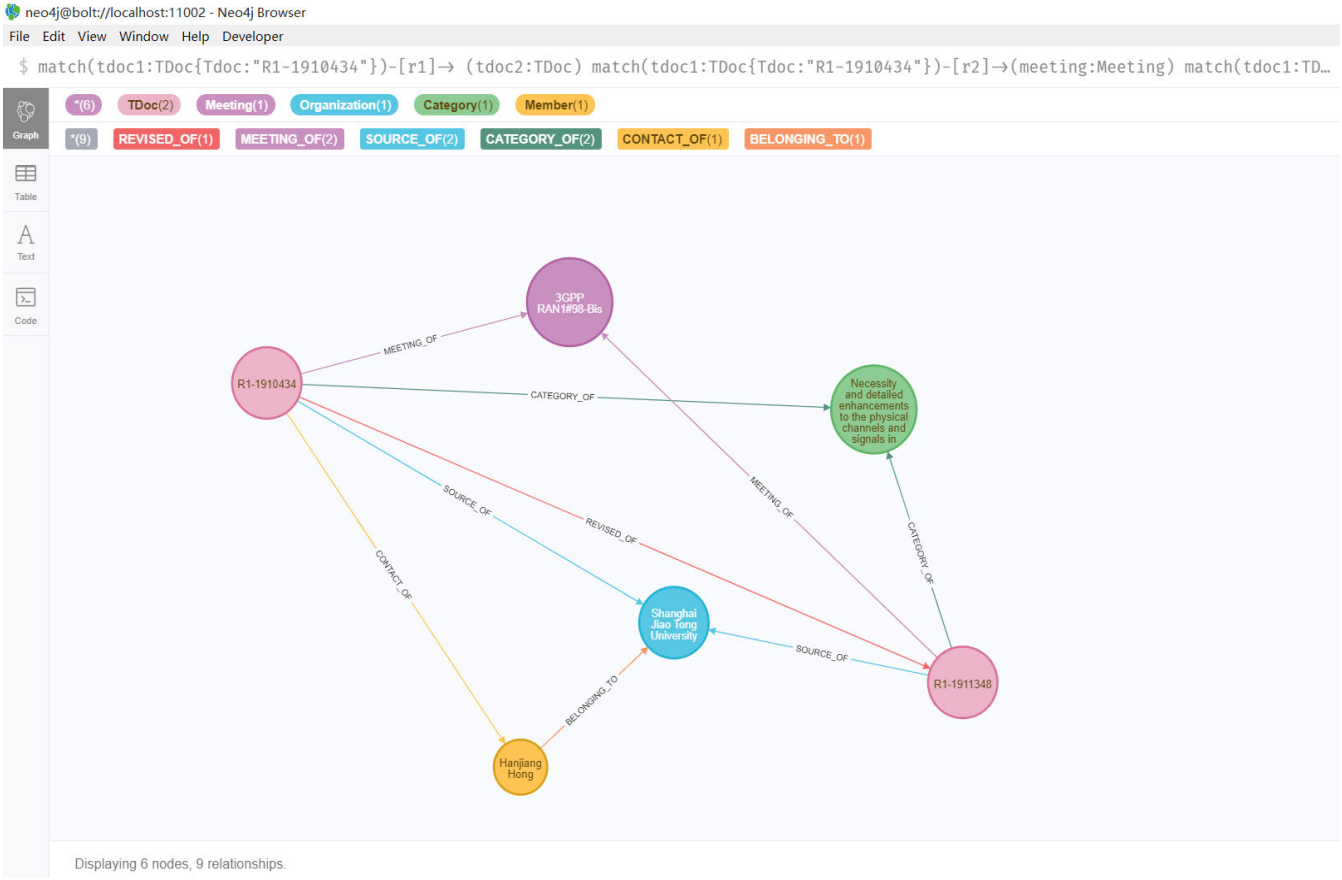


FIGURE 10. Entity and corresponding relationship of proposal R1-1910434.

''6.2.4.2'', and the purpose of the document is ''Discussion''. In addition, there are four relational attributes of the proposal: the affiliated organization is ''Shanghai Jiao Tong University'', the proposed meeting is ''3GPP RAN1 # 98-Bis'', the proposed member is ''Hanjiang Hong'', and the subject category is '' Necessity and detailed enhancements to the physical channels and signals in the CAS ''. They are connected with the corresponding entities respectively to establish the source of the organization relationship(SOURCE_OF), meeting subordination relationship(MEETING_OF), liaison relationship(CONTACT_OF) and category subordinate relationship(CATEGORY_OF).

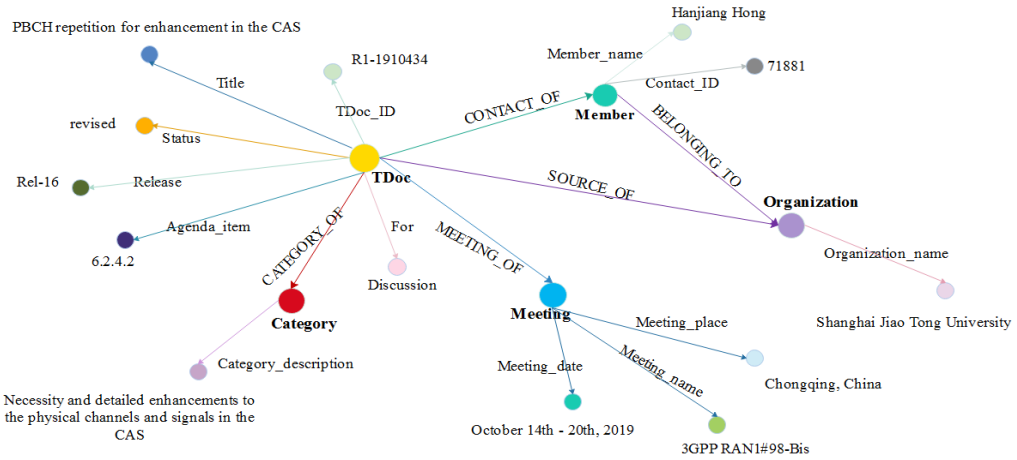


FIGURE 11. Entity and corresponding relationship of proposal R1-1910434.

In addition, the member ''Hanjiang Hong'' and the organization ''Shanghai Jiao Tong University'' also established a subordinate relationship(BELONGING_TO), which further formed the ternary relationship of ''proposal''-''member''-''institution'', which is more beneficial to the data query and analysis in the knowledge graph.

V. CONCLUSION

With the rapid development of knowledge graphs in various industries, domain knowledge graphs have shown unparalleled advantages over general graphs. However, the research on the domain knowledge graph for technical proposal documents is not mature enough. Massive unstructured text data still needs to be manually annotated and integrated. The semantic information implicit in the proposal documents has not been fully excavated, which gives researchers' work brings great inconvenience.

Taking the relevant technical proposal documents of the 3GPP conference as an example, this paper proposes an information extraction model based on TextCNN, and constructs a domain knowledge graph which selects Neo4j for visual display. Based on TextCNN's information extraction model, the subject of the proposal and the summary information are automatically extracted to form a preliminary knowledge representation, so as to enrich the relevant attributes and the relationships of the domain knowledge graph. Combining the features of the proposal document, we used the TextCNN algorithm for classification in the topic information extraction model, and compared it with the traditional machine learning classification algorithm and FastCNN. Experiments show that the model has high accuracy on the technical document dataset, which can effectively reduce the cost of manual annotation and data collation.

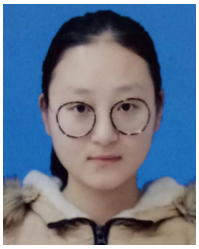
The high fragmentation and flexibility of the data in the knowledge graph brings new challenges to the construction of domain knowledge graphs.

For multi-source heterogeneous data sources, the future development direction of this paper is to expand the entities and the relationships, and further enrich the data model of the domain knowledge graph. At the same time, further improving the accuracy of the information extraction model and the efficiency of knowledge acquisition are also the important tasks in the future.

REFERENCES

[1] D. R. Swanson, ''Migraine and magnesium-11 neglected connections,'' *Perspect. Biol. Med.*, vol. 31, no. 4, pp. 526–557, Summer 1988, doi: 10.1353/pbm.1988.0009.
[2] H. Tan, ''A brief history and technical review of the expert system research,'' in *Proc. IOP Conf. Ser., Mater. Eng. Sci.*, vol. 242, 2017, Art. no. 012111.
[3] N. Shadbolt, T. Berners-Lee, and W. Hall, ''The semantic Web revisited,'' *IEEE Intell. Syst.*, vol. 21, no. 3, pp. 96–101, May 2006.
[4] A. Singhal. (May 2012). *Official Google Blog: Introducing the Knowledge Graph: Things, Not Strings*. Official Google Blog. [Online]. Available: http://googleblog.blogspot.pt/2012/05/introducing-knowledge-graph-things-not.html
[5] D. B. Lenat, ''CYC: A large-scale investment in knowledge infrastructure,'' *Commun. ACM*, vol. 38, no. 11, pp. 33–38, Nov. 1995.
[6] G. A. Miller, ''WordNet: A lexical database for English,'' *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.
[7] F. Abbas, M. K. Malik, M. U. Rashid, and R. Zafar, ''WikiQA—A question answering system on Wikipedia using freebase, DBpedia and infobox,'' in *Proc. 6th Int. Conf. Innov. Comput. Technol. (INTECH)*, Aug. 2016, pp. 185–193.
[8] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, ''DBpedia–a crystallization point for the Web of data,'' *J. Web Semantics*, vol. 7, no. 3, pp. 154–165, Sep. 2009.
[9] F. M. Suchanek, G. Kasneci, and G. Weikum, ''Yago: A core of semantic knowledge,'' in *Proc. 16th Int. Conf. World Wide Web (WWW)*, Banff, AB, Canada, 2007, pp. 697–706.
[10] X. Bo, X. Yong, J. Liang, C. Xie, and Y. Xiao, ''CN-DBpedia: A never-ending Chinese knowledge extraction system,'' in *Proc. Int. Conf. Ind., Eng. Appl. Appl. Intell. Syst.*, 2017, pp. 428–438.
[11] B. Regalia, K. Janowicz, G. Mai, D. Varanka, and E. L. Usery, ''GNISLD: Serving and visualizing the geographic names information system gazetteer as linked data,'' in *The Semantic Web*. Cham, Switzerland: Springer, 2018.
[12] P. Wang, S. Li, G. Sun, X. Wang, Y. Chen, H. Li, J. Cong, N. Xiao, and T. Zhang, ''RC-NVM: Enabling symmetric row and column memory accesses for in-memory databases,'' in *Proc. IEEE Int. Symp. High Perform. Comput. Archit. (HPCA)*, Feb. 2018, pp. 518–530.

[13] P. Derose, W. Shen, F. Chen, Y. Lee, and R. Ramakrishnan, "DBlife: A community information management platform for the database research community (demo)," in *Proc. CIDR*, Asilomar, CA, USA, 2007, pp. 169–172.
[14] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "ArnetMiner: Extraction and mining of academic social networks," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining (KDD)*, 2008, pp. 990–998.
[15] M. T. Pazienza, "Information retrieval: Still butting heads with natural language processing?" in *Proc. Int. Summer School Inf. Extraction, Multidisciplinary Approach Emerg. Inf. Technol.*, 1997, pp. 115–138.
[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Effificient estimation of word representations in vector space," Jan. 2013, *arXiv:1301.3781*. [Online]. Available: https://arxiv.org/abs/1301.3781
[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.
[18] D. Liang, F. Yang, X. Wang, and X. Ju, "Multi-sample inference network," *IET Comput. Vis.*, vol. 13, no. 6, pp. 605–613, Sep. 2019.
[19] Y. Kim, "Convolutional neural networks for sentence classification," 2014, *arXiv:1408.5882*. [Online]. Available: http://arxiv.org/abs/1408.5882
[20] Y. Zhang and B. Wallace, "A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification," Oct. 2015, *arXiv:1510.03820*. [Online]. Available: https://arxiv.org/abs/1510.03820
[21] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," 2016, *arXiv:1607.01759*. [Online]. Available: http://arxiv.org/abs/1607.01759
[22] J. Guia, V. Gonçalves Soares, and J. Bernardino, "Graph databases: Neo4j analysis," in *Proc. 19th Int. Conf. Enterprise Inf. Syst.*, 2017, pp. 351–356.
[23] F. Holzschuher and R. Peinl, "Performance of graph query languages: Comparison of cypher, gremlin and native access in Neo4j," in *Proc. Joint EDBT/ICDT Workshops (EDBT)*, 2013, pp. 195–204.



HUAXUAN ZHAO was born in Feicheng, Shandong, China. She is currently pursuing the degree with the School of Information Science and Engineering, Shandong Normal University, China. Her research interests include knowledge graph, natural language processing, recommender systems, and machine learning.

YUELING PAN was born in Rizhao, Shandong, China. She is currently pursuing the degree with the School of Information Science and Engineering, Shandong Normal University, China. Her research interests include machine learning, knowledge graph, and data management.

FENG YANG received the bachelor's and master's degrees in radio electronics from Shandong University, in 1985 and 1988, respectively. He is currently a Professor with the School of Information Science and Engineering and the Director of the Department of Communication Engineering, Shandong Normal University. He has published more than 30 articles, edited textbooks, and five books. He holds seven national invention patents and four utility model patents. He has presided over five provincial and university-level education reform projects. He has participated in one national fund project. He received one Provincial-Level Teaching Achievement Award, one School-Level Teaching Achievement Award, and three provincial awards for scientific and technological progress.