Entity Recognition and Talent Profiles in Digital Industry Based on BERT and BiGRU

Yanwu Yang, Panyu Zhu, Ying Jin
[a]School of Management, Huazhong University of Science and Technology, Wuhan 43004, China, {yangyanwu.isec, zhaipanyu.isec, jinying.isec }@gmail.com

Abstract: In this study, we propose an entity recognition and talent profile model (BERT-BiGRU-Attention-CRF) to identify talent requirements for different jobs in the digital industry. This model combines the two representation learning methods (i.e., BERT and BiGRU) to represent textual contents according to contextual semantics in sentences, uses multi-head attention to strengthen the key semantic information, and then employs conditional random field (CRF) to classify texts to identify talent requirements. Moreover, we conduct empirical research by collecting the latest recruitment information about jobs in the digital industry from three major recruiting websites. Experimental results indicate that BERT-BiGRU-Attention-CRF significantly outperforms baseline models in terms of Precision, Recall, and F1 metrics. Based on the relationships between jobs and requirements, we construct knowledge graph and job-specific talent profiles for the digital industry using the proposed model, reveal similarities and differences between talent requirements for various jobs by analyzing job-related nodes in the knowledge graph. This study presents a feasible method for entity recognition and talent profiles construction for jobs in the digital industry based on the real-world data, which gives practical guidance for majors educating digital talents, and provides theoretical basis and methodological support for the optimization and upgrading of the digital industry in China.

Keywords: digital industry; entity recognition; talent profiles; BERT-BiGRU-Attention-CRF; digital talent

1. Introduction

Digital industry is a strategic emerging industry in China, exerting a profound influence on economic growth, people's livelihood, and modernization development. By 2023, among the 3,072 colleges and universities across the country, 757 have established the data science and big data technology major, and 220 have set up the big data management and application major, aiming to cultivate digital talents required for the innovation and development of the digital industry. Constructing, analyzing, and mining the knowledge graph of digital industry talents and the talent profiles of related positions play a crucial guiding role in the cultivation and employment of digital talents. It is conducive to optimizing the curriculum system and talent cultivation path settings of relevant majors in the digital industry, realizing a digital talent cultivation mechanism oriented by social needs. Moreover, it facilitates the structural adjustment of the digital industry and the sustainable development of the digital talent echelon, providing fundamental theoretical basis and methodological support for the optimization and upgrading of China's digital industry.

The talent profile of digital industry positions refers to extracting the multi-dimensional and highly refined key characteristics and ability standards that position talents should possess based on a series of real data of position qualifications. These include explicit characteristics that can be directly observed, such as educational background and major, and implicit characteristics that cannot be directly observed, such as traits and values.

Existing talent profile research can be categorized into qualitative and quantitative research methods. Qualitative research methods, such as expert evaluation [1] and grounded

theory [2], fail to accurately describe the details and characteristics of talent profiles. They overly rely on researchers' subjective interpretations and lack objective theoretical basis and real data support. Quantitative research mainly encompasses descriptive statistical analysis methods [3] and text mining methods [4]. Descriptive statistical analysis methods [2, 5, 6] conduct statistical descriptions of elements such as educational background and major to reveal the group characteristics of position talents. However, this method is limited to simple statistics such as percentages, means, and variances of collected data and can only characterize structured data. It is difficult to mine deep-seated characteristics such as attitudes and values from text. In many cases, the used data cannot reflect the actual talent demand situation, which is not conducive to result analysis and inference. Most text mining-based studies adopt the LDA (Latent Dirichlet Allocation) topic model to extract characteristic topic words [7-10]. Nevertheless, the descriptions of digital industry positions involve highly specialized terms, and the LDA topic model has difficulty handling deep semantic information in complex contexts and cannot accurately capture the dependency relationships between positions and skills.

In recent years, some studies have attempted to utilize natural language processing models to identify entities in text and classify the relationships between entities. They train Chinese word vectors using the pre-trained language model BERT and combine models such as Bi-LSTM or IDCNN (Iterated Dilated Convolutional Neural Networks) [11-13] to learn text context information. Literature [14] presents talent profiles through knowledge graphs and extracts the competencies required for different positions using a co-occurrence matrix. Similarly, literature [15] draws on the iceberg model to delineate the competency indicators required for data-related positions and employs the BERT-BiLSTM-CRF model to obtain the relationships between entities in recruitment information and construct the knowledge graph of data-related positions. To some extent, these models have improved the accuracy of entity recognition. However, they still struggle to mine deep semantic information in recruitment texts and focus on key semantics. Additionally, there is a lack of interpretable characterization methods for identified position talent requirements, making their results difficult to generalize and apply.

This study collects vast amounts of recruitment data and correlates multi-source heterogeneous data. It constructs an entity recognition and talent profile model (BERT-BiGRU-Attention-CRF) to analyze digital industry-related positions and their talent requirements. Subsequently, based on the relationships between positions and requirements, it constructs a digital talent knowledge graph and position talent profiles. This model combines BERT's deep semantic representation ability and BiGRU's advantages in processing text data. It utilizes the attention mechanism to highlight key semantic information in sentences and adopts CRF for text classification, enabling more accurate text representation and identification of positions and their talent requirements.

This paper makes the following contributions: 1) Introduce BiGRU and Attention into entity recognition for talent profiles. BiGRU is used to learn long-distance context dependency relationships, and the Attention mechanism is employed to highlight key semantic information in text. This optimizes the model structure, reduces the number of parameters, and enhances the model's recognition accuracy. 2) Apply the constructed model to build a digital industry talent knowledge graph and related position talent profiles from four dimensions: knowledge, skills, traits, attitudes and values. Analyze the similarities and differences between talent profiles of different positions and the relationships between skills and positions. Provide practical support and effective suggestions for talent cultivation, student employment, skill improvement, and career development in majors related to the digital industry.

2. Name Entity Recognition

Entity recognition is a key technology for constructing a talent knowledge graph. Effectively identifying key entities and their relationships in recruitment text can accurately and comprehensively reveal position talent requirements.

Named entity recognition identifies entities with specific meanings from natural language text and classifies them into predefined categories. Then, the entities are mapped to corresponding nodes in the knowledge graph. Current research on named entity recognition mainly includes rule-based and dictionary-based methods and machine learning-based methods. Rule-based and dictionary-based methods construct a dictionary of entities to be recognized and match them in text according to predefined rules [16-18]. These rules can be defined based on features such as grammar, context, and part of speech. If a word in the text satisfies the rule, it is marked as an entity. Although such methods are simple and easy to understand, they are limited by the frequent need to update the dictionary and rules. They lack good adaptability and are difficult to handle new entity types or complex scenarios. Machine learning-based methods include feature-based methods and deep learning methods [19]. Feature-based methods inductively learn from training corpora based on manually selected features and then apply them to new entity recognition tasks. Common methods include the Maximum Entropy Markov Model (MEMM), decision trees, and Support Vector Machines (SVM). The feature selection and extraction of these methods are relatively flexible, but they require expertise in the field and manual feature extraction. Deep learning technology can handle large amounts of natural language text and fully utilize context information to improve the accuracy of entity recognition. Deep learning views named entity recognition as a sequence tagging problem and can automatically learn useful features from raw data. Recurrent Neural Networks (RNN) have been widely studied in named entity recognition. Among them, the Bidirectional Long Short-Term Memory Network (BiLSTM) effectively processes complex sentence structures using forward and backward context information, improving recognition accuracy [20, 21]. Literature [22] applies the BiLSTM-CRF model to sequence tagging tasks, where BiLSTM is used to extract context features and CRF globally constrains the labels in the sequence. Similarly, literature [23] and [24] use BiLSTM to model the context information of each word and use CRF to output the results of entity recognition. These studies attempt to use natural language processing models to identify entities in text and classify the relationships between entities. They train Chinese word vectors using the pre-trained language model BERT and combine models such as Bi-LSTM or IDCNN [11-13] to learn text context information, improving the accuracy of entity recognition to some extent. However, when characterizing talent profiles, they often fail to fully mine the unstructured text in recruitment information. There are the following deficiencies: 1) They cannot effectively capture long-distance context dependency relationships and are difficult to mine deep semantic relationships. 2) They require a large number of model parameters, resulting in high computational costs and long training times. 3) They cannot distinguish the core semantics in text during the entity recognition process and are difficult to focus on key information. 4) For the identified position talent requirements, these studies do not provide interpretable characterization methods, making the results of talent profiles difficult to generalize and apply.

To address problems 1) and 2), this paper adopts BiGRU to learn long-distance context dependency relationships. Its advantage lies in having a simpler model structure and fewer parameters than previous BiLSTM models, enabling more efficient and rapid learning of text semantic information [25, 26]. To address problem 3), this paper incorporates an attention mechanism, which highlights key semantic information in text to describe and present talent requirements more accurately by learning the weights of word vectors in text [13]. To address problem 4), this paper uses a knowledge graph to characterize the relationships between

positions and talent requirements. It portrays the talent requirements of each position more clearly. By comparing the similarities and differences of talent profiles of various positions and the correlations between skills in the knowledge graph, the research results of this paper's talent profiles can be extended to talent cultivation in colleges and universities, individual career planning of students, and talent selection in enterprises.

3. Model

The framework of the BERT-BiGRU-Attention-CRF model is shown in Figure 1. It mainly includes a text input layer, a word embedding layer, a sentence encoding layer, an attention layer, a text classification layer, and a knowledge graph layer.

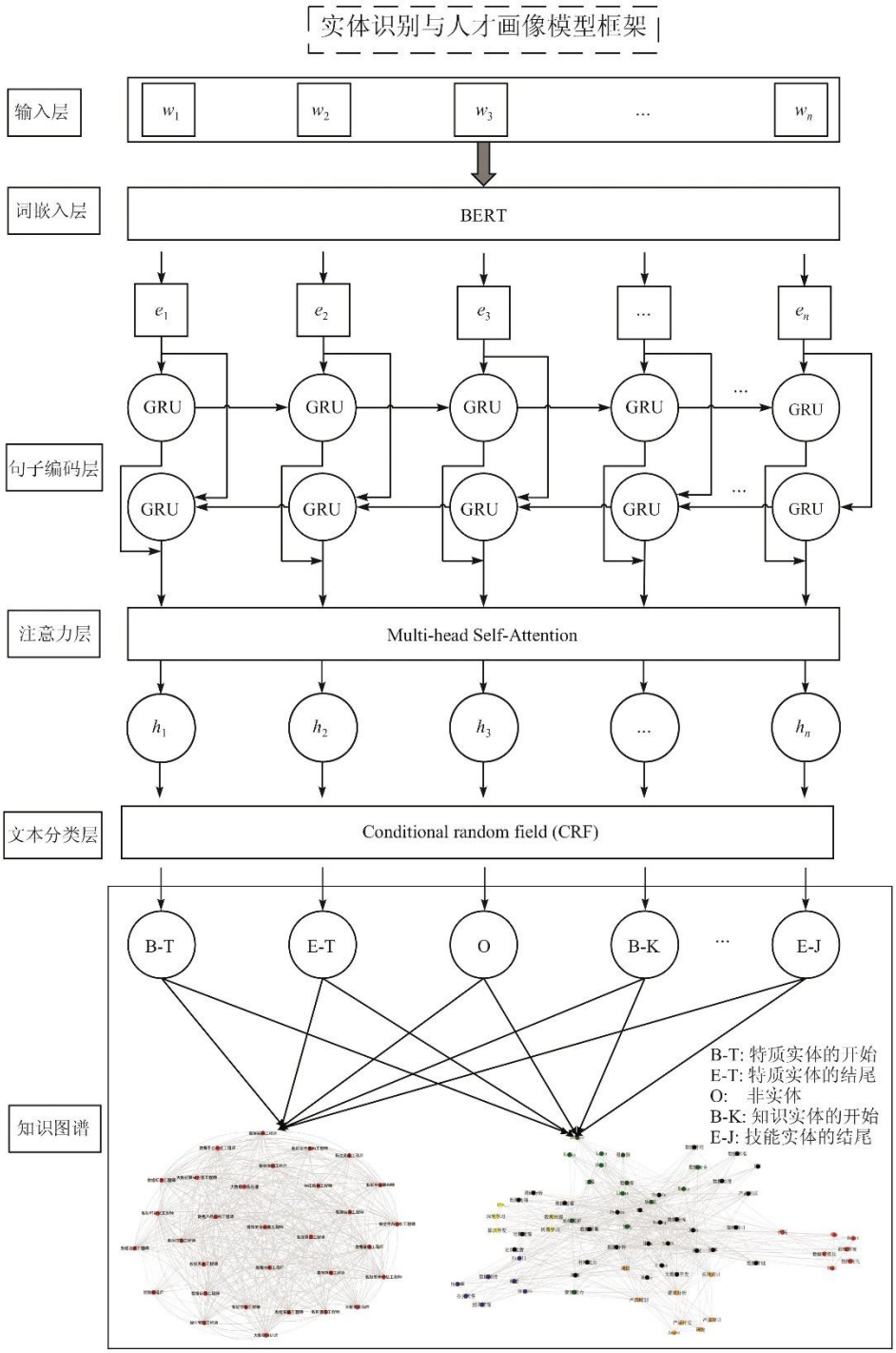

Fig.1 The modeling framework of entity recognition and talent profiling.

1) Text input layer: The unstructured data in recruitment information serves as the input of the model. Preprocessing such as deduplication, normalization, unification, and fixing sentence length is performed on the original data to organize the original data into a format that can be input into the model.

2) Word embedding layer: The pre-trained natural language model BERT [27] is used to obtain the embedding representation of each word in the unstructured data. Most traditional natural language processing models process text unidirectionally from left to right. That is, each word in a sentence is only affected by the words before it. This framework limits the text representation at the sentence level. The BERT model is a pre-trained language representation model based on the Bidirectional Transformer encoder [28]. It aims to jointly adjust the bidirectional semantic pre-training of words in all layers to deeply represent unlabeled text bidirectionally. Inspired by the masked language model [30], BERT randomly masks the tokens of some words in the input to alleviate the limitation of unidirectional text processing in traditional natural language models. When using BERT for natural language representation, only the corpus sequence needs to be input into the model. There is no need to pre-train the model. BERT will automatically extract the word representation in the language sequence according to the downstream task labels during the fine-tuning stage.

3) Sentence encoding layer: Based on the word embedding representation by BERT, BiGRU is used to extract the context information of words in the sentence. BiGRU reads the sentence from both the forward and backward directions at the sentence level. Based on the dependencies between words in the sentence, it extracts the context, semantic, and grammatical features of words [29]. BiGRU is composed of a forward gated recurrent unit and a backward gated recurrent unit. It can simultaneously use the forward and backward gated recurrent units to represent word vectors from the forward and backward directions of the sentence, capturing long-distance dependencies in text. The design of the reset gate and the update gate enables BiGRU to adaptively control the flow of information. With fewer parameters than BiLSTM, it can achieve similar model performance and reduce the risk of overfitting while maintaining high computational efficiency [25, 26, 30].

4) Attention layer: The multi-head attention [28, 31] is used to measure the importance of words in the sentence to strengthen the key semantic information in the sentence. The multi-head attention is the core component of the Transformer. By executing attention calculations in parallel with multiple attention heads and combining the outputs of multiple heads, a more comprehensive context representation of words in the sentence is obtained. The multi-head attention assigns weights to each output of the BiGRU layer. The greater the weight of a word, the more important its semantics, which means that the key semantics are concentrated here, thus affecting the final recognition result of entities in the text. By assigning different attention weights to word groups in the text, the key skills, experiences, and traits required in the position description are dynamically emphasized, enabling the drawing of a more accurate talent profile [13].

As shown in Figure 2, for the text “A data platform development engineer requires a bachelor’s degree or above, be familiar with front-end and back-end development, and have good learning ability”, this paper uses BiGRU to learn the deep context dependency relationships between “data platform development engineer” and “bachelor’s degree or above”, “front-end and back-end development”, and “good learning ability” from both directions of the text. The multi-head attention mechanism assigns different importances to each word vector of the sentence, enabling the full acquisition of key semantic information in the sentence.

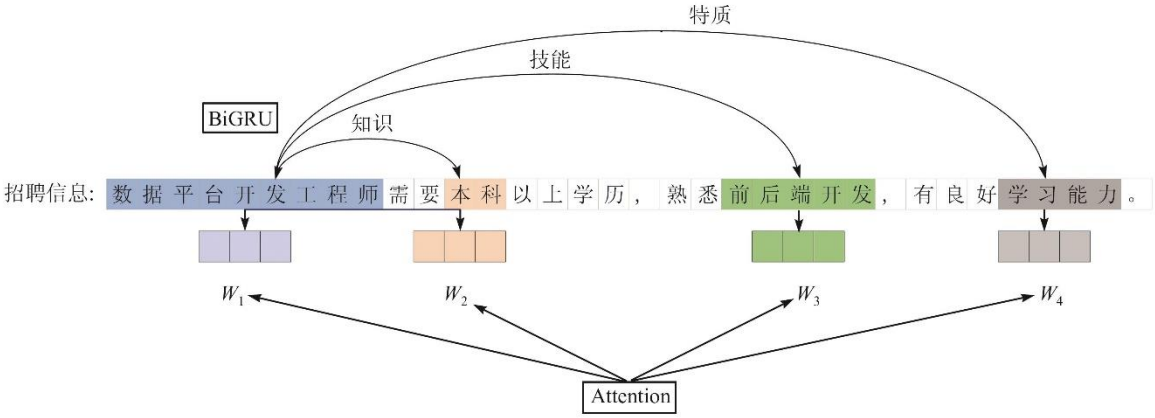

Fig.2 Using BiGRU to extract dependency relations from sentences and Attention to obtain weights for word vectors.

5) Text classification layer: CRF is used to identify the optimal tagging order of words in the entity to form the entity. BiGRU can extract long-distance text information but cannot well recognize the sequence information between adjacent tags. CRF can predict the optimal tag sequence based on the dependencies between tags. To make the prediction result more accurate, this study uses cross-entropy as the loss function and trains the model by the probability that the predicted tag approaches the real tag.

6) Knowledge graph layer: Based on the identified entities and their relationships, a series of (position, relationship, talent requirement) triples are formed and stored. According to these triples, a digital industry talent knowledge graph and related position talent profiles are constructed, comprehensively showing the knowledge (K), skills (S), traits (T), and attitudes and values (V) required by talents for digital industry-related positions.

4. Data Collection and Experiments

4.1 DataCollection and Data Processing

This study crawls recruitment information published online by digital industry-related enterprises and institutions and cleans and encodes the original data for entity recognition and talent profile construction. The specific steps are as follows.

1) Data collection: Recruitment information on the current three mainstream recruitment websites in China (Boss Zhipin, 51job, and 58.com) is crawled as the original data. First, the fields to be extracted, including position names and position descriptions, are selected. The navigation structure and search parameters of the website are analyzed to construct a URL that can access the target position list. Second, Python crawler code is written to crawl the recruitment information of 31 specific positions related to 10 directions (including data preprocessing, data annotation, data analysis, product development, project implementation and operation and maintenance, platform construction, data security, data management, operation and application, and consulting services) listed in “Big Data Industry Talent Position Competency Requirements” (http://www.mitec.cn/home/index/detail?id = 2676) on each recruitment website, totaling 15,980 pieces. Each recruitment information contains structured data (such as position names, work locations, salaries, etc.) and unstructured text data (such as position qualification requirements). Finally, the recruitment information of the same position is merged and summarized.

2) Data processing: The crawled data is cleaned. First, deduplication is performed to ensure the uniqueness of each recruitment information. Second, some irrelevant information such as company profiles and company benefits is deleted, and talent demand information such as position requirements and qualification requirements in the samples is extracted. Next, the samples are standardized and unified, including the unification of the case of English

words, the consistency of specific phrases, and the consistency of number formats. For example, different expressions such as “position requirements”, “qualification requirements”, and “job requirements” are unified as “qualification requirements”. Then, the text in the recruitment information is segmented into individual sentences to facilitate subsequent feature extraction by the model. Finally, the maximum length of each sentence is determined according to the characteristics of the data set and the requirements of the model. This length should cover the length of most sentences to avoid unnecessary computational burden caused by overly long sentences. At the same time, sentences longer than the maximum length are truncated, and sentences shorter than the maximum length are filled with fixed characters.

3) Manual coding: The main work of manual coding is data annotation. This study extracts 10% of the samples and annotates relevant entities according to the four dimensions of talent profiles (knowledge, skills, traits, attitudes and values). The annotation standards are shown in Table 1. The BiOSE annotation method is used, where B represents the start of an entity, I represents the inside of an entity, 0 represents a non-entity, S represents a single entity, and E represents the end of an entity. Each sample is annotated by two staff members. If there are differences in annotations, they are discussed and re-annotated.

Table 1 Labeling standards

Dimensions	Descriptions	Keywords
Knowledge	Basic knowledge and professional knowledge required by the applicant to complete the job	Computer, work experience, mathematics, statistics, software engineering, communication, statistics, finance, experience, electronics, automation, applied mathematics, machine learning, information management, college, undergraduate, master’s, development experience, data structures, software, data warehouse, electronic engineering, operating systems, data security, knowledge graph, etc.
Skills	Applicant’s level of application of professional knowledge and the ability to use special tools when completing the job	Python, SQL, Hadoop, Spark, Java, C++, Flink, Linux, Hive, Kafka, Oracle, MySQL, ETL, Redis, Hadoop, HBase, Scala, Visio, Windows, Excel, PyTorch, Tensorflow, Docker, PPT, data warehouse, data governance, data mining, architecture design, operation and maintenance, data processing, data modeling, product design, data analytics, data security, data management, data cleaning, data collection, demand analysis, data lake visualization, performance optimization, testing, data storage, data structure, performance tuning, data annotation, data operation, document writing, software development, demand research, data development, data service, document writing, product planning, metadata management, distributed computing, fault handling, user profile, system operation and maintenance, database design, data transmission service, business analysis, data migration, operation and maintenance management, troubleshooting, etc.
Traits	Internal qualities that a candidate needs to possess	Communication ability, learning ability, logical thinking, data sensitivity, stress resistance, execution, coordination ability, self-drive, understanding ability, expression ability, innovation ability, independent analysis, etc.
Attitudes and values	Work attitude and values that the applicant should possess	Responsibility, teamwork, team spirit, initiative, team awareness, service awareness, etc.

4.2 Parameter Settings

The annotated sample data is divided into a training set and a test set in an 8:2 ratio for model learning. In the following experiments, to ensure a fair comparison, both the model constructed in this paper and the baseline models use Adam as the optimizer, and the batch size is set to 15. To obtain stable results, each model is experimented three times, with each

experiment including 5 epochs. The average of the three experimental results is taken as the final result of the model. To obtain the optimal performance of the entity recognition and talent profile model (BERT-BiGRU-Attention-CRF) constructed in this study, the grid search method is used to determine the hyperparameters of the model. The learning rate is searched in the interval [0.0001, 0.0005, 0.001, 0.0015, 0.002, 0.0025, 0.003], and the optimal learning rate is determined as 0.0001. The word embedding dimension of the BERT model is searched in the interval [16, 32, 64, 128], and 64 is obtained as the optimal word embedding. The embedding dimension of BiGRU is searched in the interval [32, 64, 128, 256], and 256 is obtained as the optimal dimension.

4.3 Evaluation Metrics

Precision, Recall, and F1 score (F1) are used as evaluation criteria [33]. Precision refers to the proportion of true positive samples in all samples judged as positive samples by the model. The calculation formula is as follows:

$$Precision = \frac{TP}{(TP+FP)}, \tag{1}$$

where TP (True Positive) is the number of true positive samples correctly classified as positive samples, and FP (False Positive) is the number of negative samples misclassified as positive samples.

Recall refers to the proportion of samples correctly judged as positive samples by the model in all true positive samples. The calculation formula is as follows:

$$Recall = \frac{TP}{(TP+FN)}, \tag{2}$$

where FN (False Negative) is the number of true positive samples misclassified as negative samples.

The F1 value is a harmonic mean index of precision and recall, used to measure the comprehensive performance of the evaluation model. The F1 value can avoid the drawbacks of using only precision or recall as a single evaluation index and is particularly effective for handling imbalanced data set tasks. The calculation formula is as follows:

$$F1 = \frac{2*(Precision*Recall)}{(Precision+Recall)}, \tag{3}$$

The values of precision, recall, and F1 range from 0 to 1, and the larger the value, the better the prediction effect of the model.

4.4 Model Comparison

This paper adopts common entity recognition models as baselines, namely BERT-CRF [22], BERT-IDCNN-CRF [34], and BERT-BiLSTM-CRF [11]. Table 2 shows the experimental comparison results between the BERT-BiGRU-Attention-CRF model constructed in this paper and the baseline models.

Table 2 Comparison of model performance

Model	Precision	Recall	F1	Parameters	Time (s)
BERT-CRF	0.7517	**0.8932**	0.8163	103047143	146.05
BERT-IDCNN-CRF	0.7683	0.8800	0.8203	103054951	168.26
BERT-BiLSTM-CRF	0.7848	0.8748	0.8273	103212839	279.88
BERT-BiGRU-Attention-CRF	**0.8017**	0.8825	**0.8401**	103262503	280.77

It can be seen from Table 2 that the precision, recall, and F1 values of the BERT-BiGRU-Attention-CRF model are all higher than those of all baseline methods. This indicates that the BERT-BiGRU-Attention-CRF model can more accurately identify the talent requirements of positions related to the digital industry. By comparing the experimental performances of the BERT-IDCNN-CRF [34] and BERT-BiLSTM-CRF [11] models, it can

be found that the ability of IDCNN to learn text context information is inferior to that of BiLSTM and BiGRU.

This paper's data set covers 31 job positions related to the digital industry on Boss Zhipin, 51job, and 58.com. Empirical evidence shows that, regardless of the diversity of recruitment information expression styles on different platforms or the significant differences in professional terms and demand characteristics of different positions, the model in this paper can effectively identify entities and construct talent profiles. This indicates that the BERT-BiGRU-Attention-CRF model has strong adaptability and generalization ability. Therefore, the results of this study not only have high credibility but also have wide applicability in practical application scenarios and can provide strong technical support for the accurate identification of digital talent requirements.

4.5 Ablation study

An ablation experiment was conducted on the BERT-BiGRU-Attention-CRF model to verify the effectiveness of the BiGRU and Attention layers. The results are shown in Table 3.

By comparing the BERT-BiGRU-Attention-CRF and BERT-Attention-CRF models in Table 3, it can be found that while using the pre-trained BERT model to learn the word vector representations in recruitment information, it is also necessary to use BiGRU to learn the text context information. By comparing the experimental performances of the BERT-BiGRU-Attention-CRF and BERT-BiGRU-CRF models, it can be seen that the attention mechanism plays an important role in fully recognizing the semantic information in the position description and highlighting the key semantics. By comparing Tables 2 and 3, the BERT-BiGRU-CRF model performs better than the BERT-BiLSTM-CRF model in the three indicators, indicating that BiGRU can better capture the context dependency relationships between words in the sentence in the entity recognition task while using fewer model parameters.

Tables 2 and 3 record the number of parameters and time required for the training of the BERT-BiGRU-Attention-CRF and baseline models. It can be found that the BERT-BiGRU-CRF requires fewer parameters and time than the BERT-BiLSTM-CRF. This shows that the structure of BiGRU is simpler than that of Bi-LSTM and can efficiently learn the long-distance context dependency relationships in the sentence with fewer parameters. At the same time, by comparing the BERT-BiGRU-Attention-CRF and BERT-BiLSTM-CRF, it can be seen that the number of parameters and training time of the two models are relatively close. This indicates that the BERT-BiGRU-Attention-CRF can achieve better entity recognition results than the BERT-BiLSTM-CRF without increasing the complexity of the model. In summary, the BiGRU and Attention layers in the BERT-BiGRU-Attention-CRF model constructed in this study can better learn the context and semantic information of the recruitment text.

Table 3 Ablation study

Model	Precision	Recall	F1	Parameters	Time (s)
BERT-BiGRU-CRF	0.791 3	**0.884 4**	0.835 3	103196455	270.63
BERT-Attention-CRF	0.787 2	0.870 3	0.826 7	103069351	154.06
BERT-BiGRU-Attention-CRF	**0.801 7**	0.882 5	**0.840 1**	103262503	280.77

5. Talent Knowledge Graph for Digital Industry

In this section, the BERT-BiGRU-Attention-CRF model is used to analyze and mine the recruitment information crawled in Section 4 to identify the requirements of each position in terms of knowledge, skills, traits, attitudes and values. A digital industry talent knowledge graph is constructed, and the talent profiles of related positions are depicted. The relationships between positions and the relationships between skills are also analyzed.

5.1 Industry Related Job Profiles

Combined with the digital talent requirements identified in the positions and their recruitment texts, a series of corresponding triples are formed. For example, (Big Data Consultant, Skill, Java) indicates that a big data consultant needs to master Java programming skills. By integrating these triples, a digital industry talent knowledge graph can be constructed and the talent profiles of related positions can be depicted. The talent knowledge graph for digital industry includes the names of related positions and the knowledge they require (such as a bachelor's degree or above, a computer-related major, and statistical knowledge), skills (such as Java, Python, and data governance), traits (such as data sensitivity, communication ability, and logical thinking ability), and attitudes and values (such as a sense of responsibility and teamwork).

To deeply explore the talent profiles of each position, this study counts the top 20 requirements in terms of knowledge, skills, traits, and attitudes and values that the talents recruited for each position should possess. Through statistics, it is found that in terms of the required knowledge: 1) Among the 31 positions, 22 positions (such as big data consultants, data platform operation and maintenance engineers, and data platform architects) require the top three knowledge as a bachelor's degree or above, a computer-related major, and work experience; 2) Data analysis engineers have a higher requirement for statistical professional knowledge, while data algorithm engineers have a higher requirement for mathematical professional knowledge, and data platform architects tend to require software engineering professional knowledge; 3) Compared with the above positions, the educational requirements for annotation quality inspection engineers and annotation collection engineers are relatively low, and talents with a junior college degree or above can meet their job requirements.

In terms of the required skills: 1) Positions related to operation and maintenance (such as data platform operation and maintenance engineers, data security operation and maintenance engineers, and data operation and maintenance engineers) pay more attention to operation and maintenance capabilities, the ability to use Linux systems and system maintenance; 2) Data platform architects, data security architects, and data product architects all require applicants to be proficient in the Java programming language and be able to carry out architecture design; 3) Data platform architects also need a solid Python foundation, data security architects need strong development capabilities, and data product architects need strong data analysis capabilities; 4) Positions related to annotation (such as annotation collection engineers, annotation quality inspection engineers, annotation management engineers, and data annotation engineers) require applicants to be proficient in using annotation tools and have certain data analysis capabilities, but do not have high requirements for programming languages such as Python and Java; 5) Big data consultants require applicants to be familiar with distributed computing technologies such as Hadoop and Spark and the Java programming language; 6) Data platform development engineers require applicants to be proficient in the Java programming language and have a certain understanding of the architecture and working principle of the platform development and MySQL database system; 7) Data implementation engineers should be familiar with the project implementation process and the installation, debugging, and maintenance of SQLServer and Oracle databases; 8) Data security assessment engineers should have the skills to carry out data security assessments on products and have a certain understanding of databases and the Python programming language; 9) Data storage engineers should be proficient in the Java programming language, understand the basic principles of data storage, data management, and data processing, and also have the basic capabilities of designing, managing, and developing data warehouses; 10) Data visualization engineers should have front-end development and data visualization skills and be proficient in using HTML and JavaScript to build high-performance Web applications; 11) Data analysis engineers should

have data analysis capabilities, be proficient in using the Python programming language and the MySQL database management system.

In terms of the required traits: 1) Among the 31 positions, 17 positions require (ranked in the top three) talent traits including communication ability, learning ability, and logical thinking ability. Among them, good communication ability and logical thinking ability are helpful for communication with colleagues and cooperation with customers at work, and learning ability is helpful for improving and upgrading one's professional skills; 2) Data operation and maintenance engineers, data security assessment engineers, and data platform operation and maintenance engineers, in addition to emphasizing communication ability and learning ability, also pay attention to coordination ability; 3) Data security architects, big data trainers, and big data solution engineers, in addition to emphasizing learning ability, also pay attention to communication ability and expression ability to facilitate the transmission of their own ideas to others; 4) Annotation management engineers, data management engineers, and big data product managers, while emphasizing communication ability and logical thinking ability, also pay attention to coordination ability; 5) Data implementation engineers need to carry out on-site implementation and debugging work and be able to handle business trips; 6) Community managers and annotation quality inspection engineers, while having communication ability and learning ability, should also have strong execution ability and stress resistance respectively.

All positions require teamwork, a sense of responsibility, and a positive and active attitude and values. At the same time, most positions also require applicants to have a strong service awareness.

5.2 Relationships between Jobs in the Digital Industry

In this section, the relationships between related positions in the digital industry are explored based on the digital industry knowledge graph (as shown in Figure 3), and the similarities and differences between different positions are analyzed. First, each position is represented as a fixed-length one-hot vector according to the skills required by each position. Due to the large number of skill categories, the position vector representation is relatively sparse. This study uses the Singular Value Decomposition algorithm (SVD) to reduce the dimension of the position vector representation to obtain a dense representation of each position. Then, the cosine similarity is used to calculate the similarity between positions to construct a similarity matrix. The nodes in Figure 3 represent different positions, and the weight of the edges represents the similarity between the two connected positions.

Through Figure 3, the relationships between different positions can be further analyzed to identify sets of positions with high similarity and overlapping skills. For example, there is a close relationship between the four positions in the data analysis direction, namely "data algorithm engineer", "data mining engineer", "data analysis engineer", and "data visualization engineer". Among them, data algorithm engineers focus on developing and optimizing data analysis algorithms, data mining engineers use algorithms to mine potential values and trends in data, data analysis engineers are responsible for statistical analysis, in-depth mining analysis, and business prediction of data, and data visualization engineers are responsible for presenting data results to various stakeholders in a visual form. These four positions all involve analyzing and processing large amounts of data and require applicants to have similar data analysis and processing skills: be familiar with commonly used data analysis tools and programming languages such as Python and R, understand data structures and algorithms, and be able to perform data cleaning and data visualization. These four positions cooperate with each other to jointly complete complex data analysis and mining tasks, thus helping enterprises better mine and utilize data.

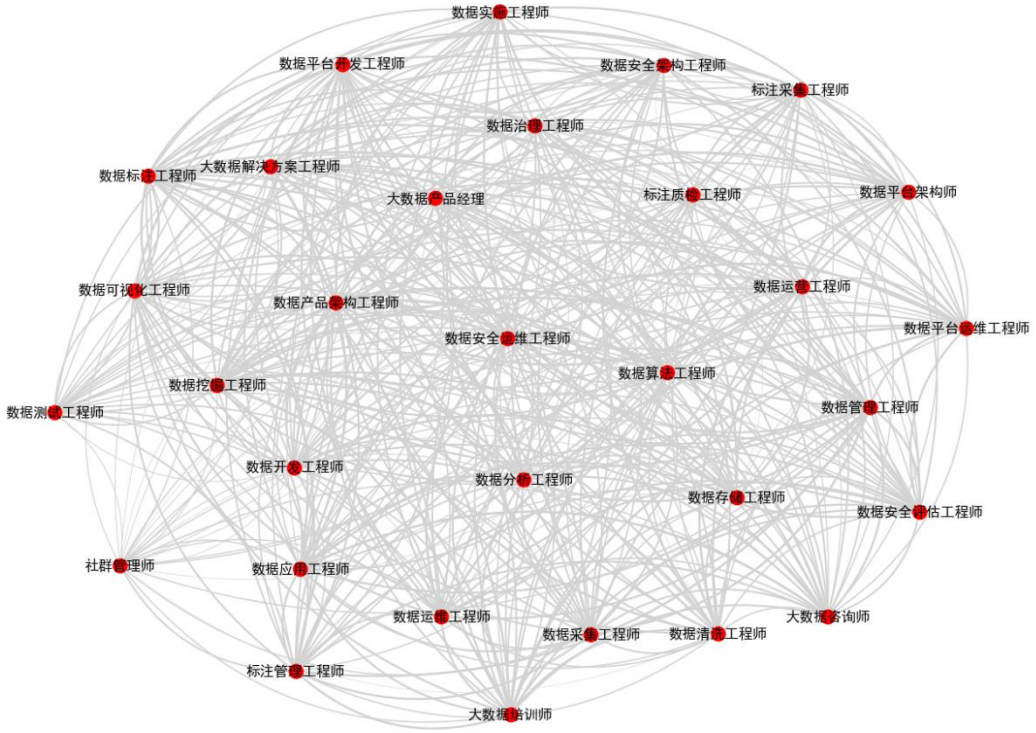

Fig. 3 Knowledge graph of relationships between jobs in the digital industry

5.3 Relationships between Skills in the Digital Industry

In this section, the relationships between the skills required for related positions in the digital industry are explored based on the digital industry knowledge graph (as shown in Figure 4). The construction method of the digital industry-related skills relationship knowledge graph is similar to that of the digital industry-related position relationship knowledge graph in Section 5.2 and will not be repeated here. The nodes in Figure 4 represent different skills, and the weight of the edges represents the similarity between the two connected skills. On this basis, the degree centrality and betweenness centrality of each skill can be calculated to measure the importance of various skills required for digital industry-related positions. The degree centrality of a skill represents the number of edges directly connected to other skills, that is, the degree. The three skills with the highest degree centrality are “data analysis”, “Python”, and “Java”, indicating that these three skills have more direct connections with other skills in the knowledge graph network and are the basis of other skills. The betweenness centrality of a skill describes the role of each skill as a technical intermediary in the ability structure of digital industry talents. The three skills with the highest betweenness centrality are “data analysis”, “product testing”, and “project”, indicating that these three skills play an important bridging role and can promote the transfer and cooperation between different skills. For example, the “data analysis” node is on the shortest path connecting “Python” and “Java”, indicating that these two skills serve data analysis capabilities.

Furthermore, based on the relationships between skills, the spectral clustering method is used to cluster the digital industry-related skills relationship knowledge graph, and the Calinski-Harabaz values (CH) corresponding to different numbers of clustering categories are calculated. The larger the CH value, the greater the variance between classes and the smaller the variance within classes, indicating a better clustering effect. Table 4 shows the clustering effects corresponding to different numbers of clustering categories. It can be seen from Table 4 that when the skills are divided into six categories, the clustering effect is the best.

The clustering results of the digital industry-related skills relationship knowledge graph are shown in Table 5. It can be seen that the first category of skills is mainly related to data processing, involving skills such as “big data development”, “data analysis”, “data storage”, “data management”, and “data governance”. Talents with these skills can apply for corresponding positions (such as data development engineers, data analysis engineers, and data storage engineers). The second category of skills is mainly related to data annotation, involving mastering basic office software and data annotation tools. Talents who are good at data annotation can apply for corresponding positions (such as annotation collection engineers, annotation quality inspection engineers, and annotation management engineers). The third category of skills is mainly related to products, mainly involving “product development”, “product planning”, and “product design” skills. Talents interested in products can apply for corresponding positions (such as big data product managers, data product architecture engineers, and big data solution engineers). The fourth category of skills is mainly related to front-end development, requiring mastering web page making tools such as HTML and visualization software. Talents who are good at front-end development and web page making can apply for corresponding positions (such as data visualization engineers and data application engineers). The fifth category of skills is mainly related to algorithms, requiring skills such as “data mining”, “machine learning”, and “algorithm development”. Talents with a good algorithm foundation can apply for corresponding positions (such as data mining engineers and data algorithm engineers). The sixth category is mainly related to operation and maintenance, and the corresponding positions include data operation and maintenance engineers and data platform operation and maintenance engineers.

Through the above analysis, the following conclusions can be obtained. First, most positions in the digital industry require a bachelor’s degree or above, a computer-related major, and work experience; annotation positions have relatively lower educational requirements, and talents with a junior college degree or above can be employed. Second, among the skills required for positions, Java, Python, and databases serve as common skill bases; different positions also require specific skills. For example, data platform architects need to have architecture design capabilities, and data visualization engineers need to have front-end development and data visualization skills. Third, among the traits required for positions, all positions require good communication ability, learning ability, logical thinking, and expression ability; at the same time, some positions require strong stress resistance and the ability to accept long-term business trips to implement projects. Fourth, in terms of the attitudes and values required for positions, all positions require the ability to actively participate in work, carry out teamwork, and have a strong sense of responsibility. Fifth, “data analysis”, “Python”, and “Java” are the basic skills that digital industry-related positions need to master; at the same time, “MySQL” and “data warehouse” and other database management capabilities play an important role in the knowledge structure of digital industry talents; in addition, “data analysis”, “product testing”, and “project” are bridges for promoting the transfer and cooperation between different skills. Sixth, the skills required for digital industry-related positions can be divided into six categories, including data processing, data annotation, products, front-end development, algorithms, and operation and maintenance.

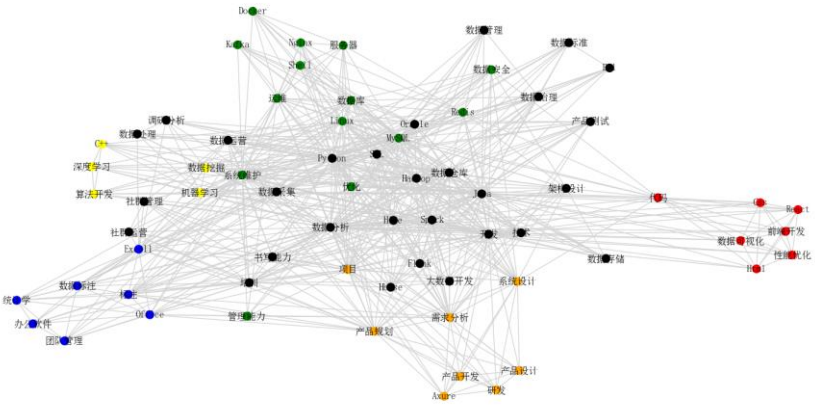

Fig. 4 Knowledge graph of relationships between skills in the digital industry

Table 4 Clustering performance with different number of clusters

Number of clusters	CH values
4	16.264
5	18.965
6	19.235
7	18.948

Table 5 Clustering results of knowledge graph of relationships between skills in the digital industry

Categories	Skills
The first category	ETL, Flink, Hadoop, Hbase, Hive, Java, Kafka, Oracle, Python, SQL, Spark, Writing Skills, Product Testing, Training, Big Data Development, Development, Technology, Data Warehouse, Data Analysis, Data Processing, Data Storage, Data Security, Database, Data Standardization, Data Governance, Data Management, Data Operation, Data Collection, Architecture Design, Community Management, Community Operation, Management Capability, Research and Analysis
The second category	Excel, Office, Office Software, Team Management, Data Annotation, Annotation, Statistics
The third category	Axure, Product Development, Product Planning, Product Design, Research and Development, System Design, Requirement Analysis, Project
The fourth category	Css, Html, React, Code, Front-end Development, Performance Optimization, Data Visualization
The fifth category	C++, Data Mining, Machine Learning, Deep Learning, Algorithm Development
The sixth category	Docker, Linux, MySQL, Nginx, Redis, Shell, Optimization, Server, System Maintenance, Operation and Maintenance

5.4 Suggestions

The relevant results of this study can provide effective suggestions for talent cultivation in colleges and universities, individual career planning of students, and talent selection in enterprises. In the process of cultivating talents in majors related to the digital industry, the digital industry talent profile helps to accurately grasp the industry needs, adjust curriculum settings and teaching content, and thus ensure that graduates have advantages in the highly competitive job market. First, schools should set Java and Python as high-priority compulsory courses and offer data analysis, MySQL, and data visualization as elective courses. Attention should be paid to the individualized cultivation of students so that students with different employment intentions can choose different courses. Second, corresponding career planning courses should also be offered. School teachers and enterprise-related

personnel should jointly be responsible for the development of the courses to provide students with employment guidance and employment training. Students should be allowed to understand the specific requirements and job responsibilities of positions in advance and be guided to carry out career planning as early as possible. Third, schools should pay attention to cultivating students' communication ability and expression ability, conduct group exercises, and regularly hold workplace simulation training. At the same time, attention should be paid to cultivating students' sense of unity and cooperation. In class and after class, students should be encouraged to complete team tasks and team competitions. Finally, schools should pay attention to school-enterprise cooperation, provide students with more opportunities to participate in projects and internships, obtain real practical experience and work experience, help enhance students' understanding of positions, increase students' employment advantages, and thus improve the employment rate and employment satisfaction of college students.

For enterprises, they can formulate recruitment strategies and talent team construction plans reasonably according to the digital industry talent profiles and knowledge graphs. First, enterprises can use artificial intelligence technology to quickly screen resumes based on the knowledge, skills, traits, attitudes, and values shown in the talent profiles of each position. In the interview process, focus on the key factors in the talent profiles to improve recruitment efficiency and reduce recruitment costs. Second, enterprises can identify sets of positions with high similarity and overlapping skills based on the relationships between positions and the relationships between skills. Promote close collaboration and resource sharing across positions, help enterprises formulate common skill training courses and personalized training plans, and thus reasonably formulate the promotion paths of employees. Ensure that employees have a clear development direction and growth opportunities in their career paths and cultivate a high-quality and stable talent team for the long-term development of enterprises.

6. Conclusions

This paper establishes an entity recognition and talent profile model (BERT-BiGRU-Attention-CRF) for the digital industry, depicting digital talent profiles from four dimensions (knowledge, skills, traits, and attitudes and values). The model combines the two representation learning methods of BERT and BiGRU, expresses text according to the context semantics in the sentence, and uses the multi-head attention to strengthen the key semantic information in the sentence. Then, the CRF method is used to classify the text to identify the talent requirements of the position. The latest recruitment information of digital industry-related positions on three mainstream recruitment websites in China is crawled to identify the specific requirements of each position in the four dimensions. Based on the identified position talent requirements, this study constructs a digital industry talent knowledge graph and related position talent profiles, deeply excavates the commonalities and differences of each position talent profile, and the importance and influence of various skills.

This study focuses on applying artificial intelligence technology and natural language methods to identify the talent requirements in the recruitment texts of digital industry-related positions, the similarities and differences between positions, and the connections between skills. In future research, efforts will be made to explore and optimize the entity recognition and talent profile model constructed in this paper to better analyze the complex semantics and associations in the talent requirements of digital industry-related positions from the position descriptions. At the same time, generative large models will be considered for data cleaning and preprocessing to generate high-quality automatically annotated labels and reduce the cost of manual annotation. Use large models to generate industry and position demand analysis reports in real-time, thereby increasing the data sources of digital industry talent profiles and constructing more comprehensive and accurate digital talent profiles. In addition, based on the digital industry talent profiles and the collected recruitment data constructed in this paper,

a digital industry talent demand prediction model will be constructed to further identify and predict the changing trends of the importance of various elements in the digital industry talent profiles and the changes in the talent demand of each position, accurately predicting the future development direction of the digital industry.

References

[1] 高扬,池雪花,章成志,等.杰出人才精准画像构建研究——以智能制造领域为例[J].图书馆论坛,2019,**39**(06):90-97.
GAO Y, CHI X H, ZHANG C Z, *et.al*. Precise user profile for domain-special talents: A case study of intelligent manufacturing [J]. *Library Tribune*,2019,39(06):90-97.
[2] 李勇,陈晓婷,黄格.基于招聘数据的人工智能人才画像与培养对策[J].重庆高教研究,2021, **9**(05):55-68.DOI:10.15998/j.cnki.issn1673-8012.2021.05.006.
LI Y, CHEN X T, HUANG G. Talent profiles of artificial intelligence and its training countermeasures based on recruitment [J]. *Chongqing Higher Education Research*, 2021, **9**(05):55-68. DOI:10.15998/j.cnki.issn1673-8012.2021.05.006.
[3] 陈明红,张倩琳,韩静.信息管理与信息系统专业人才招聘需求分析及培养启示[J].图书馆学研究,2021(20):9-20. DOI:10.15941/j.cnki.issn1001-0424.2021.20.004.
CHEN M H, ZHANG Q L, HAN J. Recruitment demand analysis and training inspirations of information management and information system professionals [J]. *Research on Library Science*, 2021(20):9-20. DOI:10.15941/j.cnki.issn1001-0424.2021.20.004.
[4] LIU Y, WEI S, HUANG H, et al. Naming entity recognition of citrus pests and diseases based on the BERT-BiLSTM-CRF model. *Expert Systems with Applications*, 2023, 234, 121103.
[5] WANG Z, HUANG M, LI C, et al. Intelligent recognition of key earthquake emergency Chinese information based on the optimized BERT-BiLSTM-CRF algorithm[J]. *Applied Sciences*, 2023, **13**(5): 3024.
[6] SHI Y, KIMURA M. BERT-Based models with attention mechanism and lambda layer for biomedical named entity recognition[C]//Proceedings of the 2024 16th International Conference on Machine Learning and Computing. 2024: 536-544.
[7] 茹宁,苏靖雅.人工智能人才画像与培养路径探析[J].天津市教科院学报,2021(02):5-11.
RU N, SU J Y. Analysis of Talent Training Path Based on Artificial Intelligence Talent Profile [J]. *Journal of Tianjin Academy of Educational*, 2021(02):5-11.
[8] SAPUTRA A, WANG G, ZHANG J Z, *et al*. The framework of talent analytics using big data[J]. *The TQM Journal*, 2022, **34**(1): 178-198. DOI：https://doi.org/10.1108/TQM-03-2021-0089.
[9] 宋培彦,龙晨翔,李怡然等.基于冰山模型的科研人员学术专长识别方法研究[J].数据分析与知识发现,2023,**7**(06):50-60.
SONG P Y, LONG C X, LI Y R, *et al*. Research on academic expertise recognition method for researchers based on iceberg model[J]. *Data Analysis and Knowledge Discovery*, 2023 ,**7**(06):50-60.
[10] 张俊峰. 国内网站招聘岗位需求特征挖掘及其应用研究[D]. 安徽财经大学, 2017.
ZHANG J F. Research on Demand Characteristics Mining and Application of Domestic Website Recruitment [D]. Anhui University of Finance and Economics, 2017.
[11] 陈明红,张倩琳,韩静.信息管理与信息系统专业人才招聘需求分析及培养启示[J].图书馆学研究,2021(20):9-20. DOI:10.15941/j.cnki.issn1001-0424.2021.20.004.
CHEN M H, ZHANG Q L, HAN J. Recruitment demand analysis and training inspirations of information management and information system professionals [J]. *Research on Library Science*, 2021(20):9-20. DOI:10.15941/j.cnki.issn1001-0424.2021.20.004.
[12] 杨静. 基于文本挖掘的网络招聘信息分析[D]. 山东师范大学, 2019.
YANG J. Analysis of Online Recruitment Information Based on Text Mining [D]. Shandong Normal University, 2019.
[13] 朱爱璐. 基于文本挖掘的数据分析岗位人才需求分析[D]. 江西财经大学, 2020. DOI:10.27175/d.cnki.gjxcu.2020.001524.
ZHU A L. A Study on the Recruitment Market of Data Analysis Based on Text Mining [D]. Jiangxi University of Finance and Economics, 2020. DOI:10.27175/d.cnki.gjxcu.2020.001524.
[14] 王一博. 基于知识图谱的计算机领域胜任力研究与应用[D]. 吉林大学, 2020. DOI:10.27162/d.cnki.gjlin.2020.003054.
WANG Y B. Research and Application of Competency in Computer Field Based on Knowledge Grap [D]. Jilin University, 2020. DOI:10.27162/d.cnki.gjlin.2020.003054.
[15] 张欣欣. 基于胜任力模型的数据类岗位知识图谱构建研究与应用[D]. 吉林大学,2022. DOI:10.27162/d.cnki.gjlin.2022.002029.
ZHANG X X. Research and Application of Data Post Knowledge Graph Construction Based on Competency Model [D]. Jilin University, 2022. DOI:10.27162/d.cnki.gjlin.2022.002029.
[16] 王宁,葛瑞芳,苑春法等.中文金融新闻中公司名的识别[J].中文信息学报,2002(02):1-6.
WANG N, GE R F, YUAN C F, *et al*. Company Name Identification in Chinese Financial Domain [J]. *Journal of Chinese Information Processing*, 2002(02):1-6.
[17] Hanisch D, Fundel K, Mevissen H T, *et al*. ProMiner: rule-based protein and gene entity recognition[J]. *BMC bioinformatics*, 2005, **6**(1): 1-9. DOI: ProMiner: rule-based protein and gene entity recognition.
[18] 黄诗琳,郑小林,陈德人.针对产品命名实体识别的半监督学习方法[J].北京邮电大学学报,2013,36(02):20-23+54.

HUANG S L, ZHENG X L, CHEN D R. A semi-supervised learning method for product named entity recognition [J]. Journal of Beijing University of Posts and Telecommunications,2013,**36**(02):20-23+54.
[19] LI J, SUN A X, HAN J L, *et al*. A survey on deep learning for named entity recognition[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2020, **34**(1): 50-70. DOI：https://doi.org/10.1109/TKDE.2020.2981314.
[20] YADAV, V, BETHARD, S. A survey on recent advances in named entity recognition from deep learning models [C]// Proceedings of the 27th International Conference on Computational Linguistics. New Mexico:Association for Computational Linguistics,2018:2145–2158.
[21] 李小龙，孙水发，唐庭龙，等. 基于超声检查报告的乳腺癌诊断知识图谱构建[J]. 武汉大学学报（理学版），2023，**69**（1）：69-78. DOI：10. 14188/j.1671-8836. 2022. 0005.
LI X L，SUN S F，TANG T L，*et al*. Construction of knowledge graph for breast cancer diagnosis based on ultrasound examination reports [J]. J Wuhan Univ （Nat Sci Ed），2023，**69**（1）：69-78. DOI：10. 14188/j. 1671-8836. 2022. 0005（Ch）.
[22] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint*, 2015.
[23] MA, X Z, HOVY, E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF [C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin:Association for Computational Linguistics, 2016, 1064–1074.
[24] DONG C H, ZHANG J J, Zong C Q, et al. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition [C]//Natural Language Understanding and Intelligent Applications: 5th CCF Conference on Natural Language Processing and Chinese Computing, and 24th International Conference on Computer Processing of Oriental Languages. Kunming, China:Springer International Publishing, 2016, 239-250. DOI：https://doi.org/10.1007/978-3-319-50496-4_20.
[25] CHO K, MERRIËNBOER V B, GULCEHRE C, *et al*. Learning phrase representations using RNN encoder–decoder for statistical machine translation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar:Association for Computational Linguistics, 2014:1724–1734. DOI：10.3115/v1/D14-1179.
[26] CHUNG J, GULCEHRE C, CHO K, *et al*. Empirical evaluation of gated recurrent neural networks on sequence modeling[C]// NIPS 2014 Workshop on Deep Learning. 2014.
[27] DEVLIN J, CHANG M W, LEE, K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota:Association for Computational Linguistics, 2018:4171–4186. DOI：10.18653/v1/N19-1423.
[28] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. New York:Curran Associates Inc., 2017:6000–6010.
[29] LIU, J, YANG, Y, LV S, et al. Attention-based BiGRU-CNN for Chinese question classification. *Journal of Ambient Intelligence and Humanized Computing*, 2019, 1-12.
[30] CAHUANTZI R, CHEN, X, GÜTTEL, S. A comparison of LSTM and GRU networks for learning symbolic sequences[C]// Science and Information Conference. Switzerland:Springer, 2023:771-785. DOI ： https://doi.org/10.1007/978-3-031-37963-5_53.
[31] ZHAI P Y, YANG Y W, ZHANG C J. (2023). Causality-based CTR prediction using graph neural networks. *Information Processing & Management*, **60**(1), 103137. DOI：https://doi.org/10.1016/j.ipm.2022.103137
[32] 翟社平,柏晓夏,张宇航,等.融合依存分析和图注意网络的三元组抽取[J].计算机工程与应用,2023,**59**(12):148-156.
ZHAI S P, BAI X X, ZHANG Y H, *et al*. Triple Extraction of combining dependency analysis and graph attention network [J]. *Computer Engineering and Applications*, 2023,**59**(12):148-156.
[33] YANG Y W, ZHAI P Y. Click-through rate prediction in online advertising: A literature review. *Information Processing & Management*, 2022, **59**(2), 102853.
[34] CAI X, SUN E, LEI J. Research on application of named entity recognition of electronic medical records based on BERT-IDCNN-CRF model[C]//Proceedings of the 6th International Conference on Graphics and Signal Processing. 2022: 80-85.