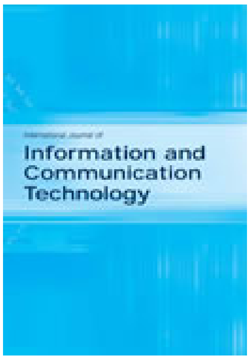
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



Information and Communication Technology

International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Knowledge creation in vocational education using multi-source data fusion under big data environment

Dongfang Zhu

Article History:

Received:	31 December 2024
Last revised:	13 January 2025
Accepted:	14 January 2025
Published online:	21 March 2025

Knowledge creation in vocational education using multi-source data fusion under big data environment

Dongfang Zhu

School of Marxism,
Xinxiang Vocational and Technical College,
Xinxiang 453000, China
Email: ggjgxx@163.com

Abstract: The development of big data technology has brought new challenges to the construction of vocational education (VE), knowledge graph (KG), and single data source does not fully capture data characterisation information. Therefore, this paper crawls internet text data from multiple sources VEs. The improved BiLSTM-CRF model is utilised to recognise the entities, and the context-aware location information is introduced into the BERT model to obtain the entity vectors containing context-aware semantics. The similarity function is used to realise entity alignment, TEXTCNN is used to extract semantic features of entity context-awareness, and the entity embedding vectors are obtained through graph annotation network, and the two are fused to obtain a more accurate representation of entity embedding vectors. The experimental results show that the entity recognition accuracy of the proposed method is improved by 3.47%–12.56%, and more accurate VEKG can be constructed.

Keywords: vocational education knowledge graph; VEKG; multi-source data fusion; context awareness; BiLSTM-CRF; graph attention network; GAT.

Reference to this paper should be made as follows: Zhu, D. (2025) 'Knowledge creation in vocational education using multi-source data fusion under big data environment', *Int. J. Information and Communication Technology*, Vol. 26, No. 5, pp.69–83.

Biographical notes: Dongfang Zhu received her Master's degree from Henan Normal University in 2010, working in the School of Marxism of Xinxiang Vocational and Technical College. Her research interests include pedagogy, ideological and political education.

1 Introduction

As the information technology rapidly growing, big data has become an important force driving change in the field of vocational education (VE). The needs of modern VE drive the continuous updating of teaching contents and methods, not only to strengthen the combination of theoretical teaching and practical skills, but also to develop innovative thinking and problem-solving skills in order to adapt to the rapid changes in technological development (Platonova et al., 2019). In the current critical period of economic restructuring and industrial transformation and upgrading, the social demand for skilled personnel has reached an unprecedented level, so it is necessary to speed up the establishment of a modern VE system to meet the diversified needs of the public for VE (Kovalchuk et al., 2022). In the field of VE, knowledge graph (KG), as an effective knowledge organisation and management tool, is of great significance for improving teaching quality and optimising the allocation of learning resources. However, traditional KG construction methods are often limited to a single data source, making it difficult to fully reflect the complexity and diversity of VE (Zheng, 2023). Therefore how to utilise multi-source data to efficiently construct vocational education knowledge graph (VEKG) has important practical value.

Guangfen and Dongke (2017) suggested a set of rules for constructing KGs in the field of VE and constructed a 'data structure' course KG based on the characteristics of computer courses. Angioni et al. (2021) constructed a KG based on the ontology of the industrial engineering subject area, guided by cognitive structural learning theory and constructivist learning theory. Chen and Song (2023) constructed the VEKG by analysing the attributes of the curriculum and its antecedent requirements and extracting the relationship between antecedents and successors, but the accuracy of the ontology is often insufficient due to the subjective factors of the experts.

When the amount of data is large, it is time-consuming to construct KGs manually. For this reason many scholars have started trying to go for developing methods to create KGs automatically. Sun and Gu (2021) automatically constructed a KG to represent a concept map of topics used in VE by using deep learning techniques to identify categorical and semantic relationships between concepts in certain specific domains. Wei et al. (2023) proposed a new method to extract course concepts from course video subtitles for automatic construction of course KGs, but the construction is time-consuming. Li et al. (2022) constructed a VEKG using open data in the field of VE and using the pre-trained model BiLSTM-CRF for named entity recognition of event records, but the accuracy of entity recognition was not high. Sun et al. (2022) extracted data from an online education platform and utilised BERT-CNN for entity-relationship extraction and TRansE for knowledge-embedded representation for KG construction.

Most of the current VEKGs are based on a single source of data, which cannot be constructed efficiently and accurately at the same time, and do not adequately cover the needs of learners. Tang et al. (2023) constructed a student-centred VEKG with relational data as the primary source, supplemented by semi-structured and unstructured data. Liu et al. (2021) introduced the EduCOR KG, an educational, career-oriented KG that provides the basis for online learning resources representing personalised learning systems. Yang and Tan (2022) collected data from multiple data sources, including MOOC websites, textbook data, and Baidu wikipedia, constructed a corpus, and constructed a course KG with multi-source data fusion by extracting the content entities of the corpus. Chang et al. (2022) obtained data related to occupational competence from multiple sources to construct the ontology of KG and realised the extraction of entity relations through BERT-BiLSTM. In order to ensure the accuracy of KG construction, scholars introduce context-awareness of text to fully consider semantic features. Yu (2022) utilised an ontology tool to construct the classic CONON model, which effectively portrays the knowledge context through 'location, task, user, and device', and based on this, constructed the KG. Wu and Jia (2022) incorporated context-aware information of text into the construction of KG and introduced location factors in the knowledge embedding representation to improve the construction efficiency.

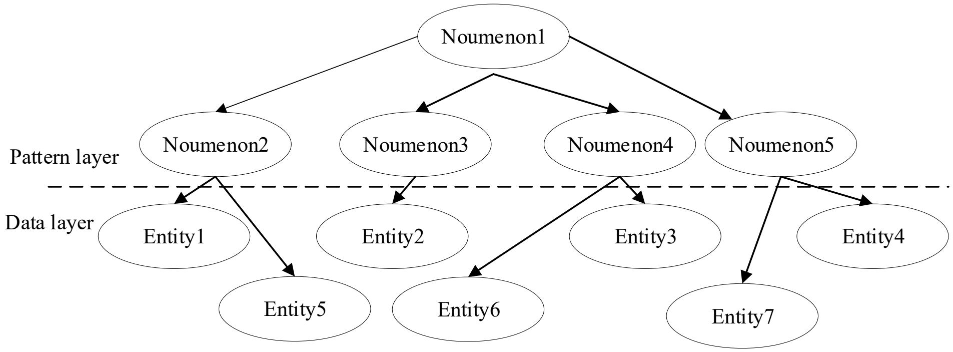
Through the above comprehensive analysis of existing research on VEKG construction methods, it is found that the VEKG construction method suffers from the problem of data source singularity as well as the phenomenon of rough knowledge granularity. In order to improve the construction efficiency of VEKG, this paper constructs a VEKG based on multi-source data fusion and context-awareness. This paper first Scrapy quickly scrapes multi-source VE text data, develops matching rules from database data to ontology data, and completes the mapping of different data to ontology data. Then the improved BiLSTM-CRF model is used to recognise the entities in the ontology, and the context-aware position information is introduced into the BERT model, which is used to label the relative position information of the words in the sentence using the position coding technique to obtain the entity vectors containing the context-aware semantic information. Attention mechanism (AM) is also utilised to enhance the word vector representation. Secondly, the feature matching algorithm using similarity function is utilised to achieve entity alignment. TEXTCNN is used to extract semantic features for entity context awareness; entities and relationships in KG are represented and learned by graph attention network (GAT), and entity embedding vectors with certain semantic information are obtained. And the entity context-aware semantic features are fused with the entity embedding vectors to get a more accurate representation of the entity embedding vectors. The experimental results imply that the offered method not only improves the accuracy of entity recognition, but also has better prediction performance on the link prediction task.

2 Relevant theoretical foundations

2.1 Basic theory of KG

KG is a semantic network that represents entities and their relationships in the objective world in the form of graphs. The concept was proposed by Google to enhance the knowledge base of its search engine functionality (Chen et al., 2018). The KG is logically composed of a data layer and a schema layer, as shown in Figure 1.

Figure 1 KG composition

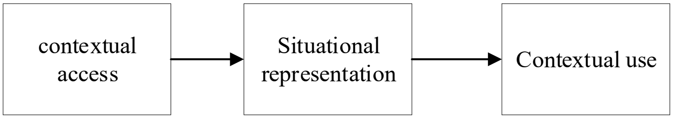


The schema layer mainly regulates the data layer through ontology library. Ontology is a kind of specification obtained by abstracting and constraining the concepts, attributes and their interrelationships of different entities in the domain, and it is the basis for constructing KG. The data layer consists of the smallest unit of knowledge storage, the ‘fact’, which is expressed as a triad of ‘entity-relationship-entity’ or 'entity-attribute-attribute-value'. An 'entity' describes an individual in a conceptual classification; a 'relationship' is an instance of a relationship defined at the schema level; and an 'attribute' is a mapping between an entity and an attribute value.

2.2 Basic concepts of situational awareness

Contextual information refers to the information obtained in a specific environment, including the physical characteristics of the environment, cultural background, knowledge characteristics and other aspects of the information. Contextual information can influence people's thinking, perception, language, and behaviour, and understanding contextual information is important for problem solving and goal achievement, as well as for constructing KGs that contain context-aware semantics. The composition of the situational awareness system is shown in Figure 2.

Figure 2 The constitution of situational awareness system



Situational awareness refers to the ability of a person or machine to perceive and understand the surrounding environment. It involves the collection, processing, and comprehensive analysis of multi-sensory information about objects, people, and locations in the environment to form a holistic perception of the environment (Endsley and Garland, 2000). Context-awareness is the representation of information about the user's context through a deep learning model that represents the acquired contextual information in such a way that it can be understood by a computer for use in the construction of KGs.

2.3 Basic concepts of situational awareness

BiLSTM-CRF is the most representative entity recognition model. BiLSTM learns long-term semantic dependencies by introducing gating and memorisation units where historical contextual information is selectively forgotten and updated, with forgetting gate f_t, input gate i_t and output gate o_t as shown below:

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_i\right) \tag{1}$$

$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \tag{2}$$

$$o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_i\right) \tag{3}$$

where h_{t-1} is the output of the hidden layer at the previous moment, x_t is the input at the current moment, σ is the activation function, b is the weight and bias term, and W_f, w_i, and W_o are the weights of the hidden, input, and output levels, respectively.

Although BiLSTM can capture contextual information, it does not take into account the transfer dependencies between labels. Conditional random fields (CRFs) (Yu and Fan, 2020) are Markovian random fields that represent a set of output random variables Y

given a set of random variables X. When the observation sequence is $X = (X_1, \ldots, X_n)$, and the probability of the state sequence $Y = (Y_1, \ldots, Y_n)$ is defined as shown in equation (4) and equation (5) for a given linear chain CRF $P(Y \mid X)$.

$$P(y \mid x) = \frac{1}{Z(x)} \prod_{t=1}^{y} \exp\left\{ \sum_{n=1}^{n} \theta_n \left(y_i, y_{i-1}, x_i \right) \right\} \tag{4}$$

$$Z(x) = \sum_{y} \prod_{t=1}^{y} \exp\left\{ \sum_{n=1}^{n} \theta_n f_n \left(y_i, y_{i-1}, x_i \right) \right\} \tag{5}$$

where f_n is the feature function, θ_n is the weight information, x_i is the input at time t, y_i is the output at time t, and $Z(x)$ is the normalisation factor.

CRF can add more constraints to the predicted labels, as the transfer probabilities between labels can be learned The BiLSTM-CRF model combines the advantages of BiLSTM and CRF, and is easier to train for convergence than other large models.

3 Ontology modelling of VE based on multi-source data fusion

There is a large amount of VE data in the internet, which comes from a wide range of sources, including structured data, semi-structured data and unstructured data. In this paper, we use Scrapy to quickly crawl the website and extract all kinds of VE data from it, and the overall architecture of Scrapy is shown in Figure 3.

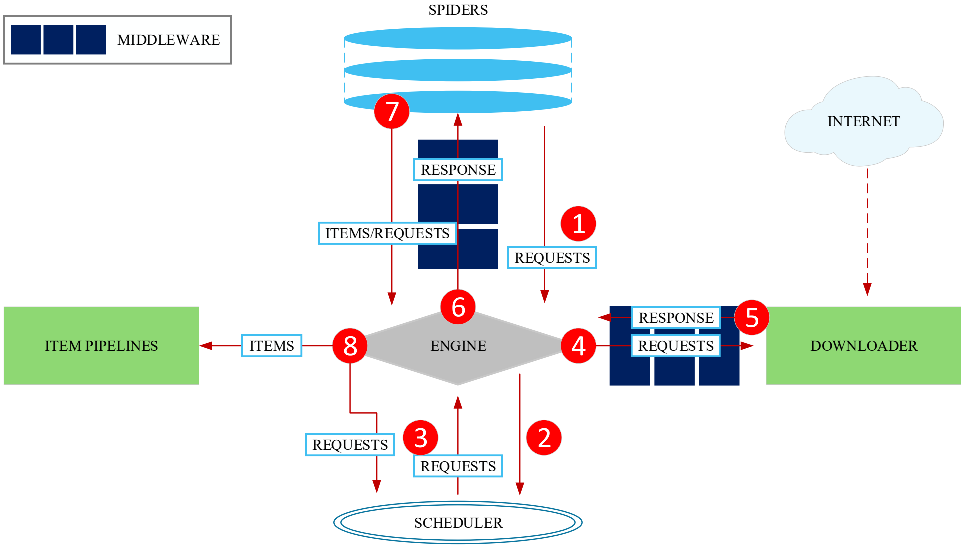
First, identify the VE-related websites to be crawled, and use this as a starting point to place them in the queue to be crawled. Then, according to the task requirements, all the data related to VE are extracted accordingly, the data request is sent to the server, the server responds to the request, the browser obtains the response content, parses the content of the web page, extracts the target data from it, and finally stores the obtained data in the database. The URL of the queue to be crawled will continue to loop until the waiting data queue is empty, then the loop will end.

VE data from different sources are heterogeneous, and the crawled data are stored in MySQL, where the database schema with entities, relationships, attributes, constraints, primary/foreign keys, etc. happen to provide a description of the entities and their relationships for the construction of the ontology. For two different data schemas, the database schema and the ontology schema, a data schema matching mapping technique is required to consider the degree of matching by analysing and calculating the similarity between entities in different schemas. Given a particular relational database schema S and an ontology model O, the matching relationship map between S and O is represented by the set $\{m\}$ as shown in equation (6).

$$Map : \{m\} = \left\{ \left\langle u, e, v, rel, f \right\rangle \right\} \tag{6}$$

where m is a mapping unit consisting of five elements; u is the identifier of a mapping unit, which serves as the unique identifier of the mapping unit; e and v are the elements in S and O, respectively, and satisfy $map(e) = v$. rel is used to describe the semantic matching relationship between elements e and v; f is the similarity of the mapping ontology, ranging from 0 to 1, used to determine whether there is a conflict or not.

Figure 3 The overall architecture of Scrapy (see online version for colours)



4 Entity relationship recognition based on improved BiLSTM-CRF model and context awareness

Based on the heterogeneous data from multiple sources obtained by web crawlers, this paper uses the improved BiLSTM-CRF model to recognise entities in the ontology based on the aforementioned modelling of VE ontology. Existing BiLSTM-CRF entity recognition methods based on BiLSTM-CRF rely on manually labelled data and suffer from the problem of overlapping labels, which does not take into account the positional information in context-awareness, resulting in unclear global semantics. Therefore, in this paper, we add additional position encoding vectors after each word through the position encoding technique, so as to label the relative position information of the word in the sentence, in order that the improved entity recognition model can fully understand the context-aware information of different words, and the computational process is as follows.

$$X_{ij} = e^{w}\left(LS_{ij}\right) \tag{7}$$

$$P_i = p_i + q_i \tag{8}$$

$$x'_{ij} = x_{ij} + P_i \tag{9}$$

where LS_{ij} is the potential word obtained for the i^{th} word based on the j^{th} set of word sequences, e^w is the word vector, d_e is the word vector size, i is the position in the sentence that expresses the context-aware semantics, P_i is the vector encoded at each position, $p_i = \cos(i/1000^{2i/d_e})$, $q_i = \sin(i/1000^{2i/d_e})$. To align the word vectors with the hidden layer output of the BERT model, a nonlinear transformation is used on the word vectors to obtain the hidden vector v_{ij}^{h} for each word.

$$v_{ij}^{h} = W_2\left(\tanh\left(W_1 x_{ij}' + b_1\right)\right) + b_2 \tag{10}$$

where W_1 is a matrix of dimension $d_h \times d_e$, W_2 is a matrix of dimension $d_h \times d_h$, and d_h is the size of the BERT hidden level.

To enhance the accuracy of BiLSTM-CRF recognition, this paper introduces word sequence AM after BiLSTM to select the most relevant words in the vocabulary set, using $A = (a_1, a_2, \ldots, a_j)$ to denote the weight vector, and a_j to be the attentional weight of the j^{th} word sequence. Firstly, the vector of each group of word sequences is obtained by weighted average, and then the weights of each group of sequences are calculated by using the self-consciousness mechanism, and the specific calculation process is as follows.

$$a_j = softmax\left(\frac{QK^T}{\sqrt{d_e}}\right) \tag{11}$$

where $Q = s_j W_Q$, $K = s_j W_K$, $s_j = (1/n)\sum_{i=1}^{n} x_{ij}'$, W_Q and W_K are weight matrices.

Finally, the lexical information injection is realised by weighting the lexical hidden vectors to get z_i^h, which is associated with the word vector implicit vector h_i^c. Use $H^l = (h_1^l, h_2^l, \ldots, h_n^l)$ to represent the output of the transformer at level l in the BERT model at the word level. $\tilde{h}_i^l$ is the vector output of layer l after incorporating the lexical information.

$$z_i^h = \sum_{j=1}^{TOP} a_j v_{ij}^h \tag{12}$$

$$\tilde{h}_i^l = \tilde{h}_i^l + z_i^h \tag{13}$$

The final decoding of the model employs a CRF layer to capture the dependencies between consecutive labels and generate the related label sequences. Given a word sequence $C = (c_1, c_2, \ldots, c_n)$ of a sentence, and the real label sequence Y corresponding to the word sequence is $Y = (y_1, y_2, \ldots, y_n)$. After going through all the transformer layers to get the hidden output $H^l = (h_1^l, h_2^l, \ldots, h_n^l)$ in the last level, the probability of labelling is calculated as follows.

$$O = W_0 H_L + b_0 \tag{14}$$

$$p(y \mid c) = \frac{\exp\left(\sum_i \left(O_{i,y_i} + T_{y_{i-1},y_i}\right)\right)}{\sum_{\tilde{y}} \exp\left(\sum_i \left(O_{i,\tilde{y}_i} + T_{\tilde{y}_{i-1},\tilde{y}_i}\right)\right)} \tag{15}$$

where W_0 and b_0 are matrix parameters, T_{y_{i-1},y_i}, $T_{\tilde{y}_{i-1},\tilde{y}_i}$, O_{i,y_i}, and $O_{i,\tilde{y}_i}$ are subterms of T and O, respectively; T is the transpose of the score matrix O; and $\tilde{y}$ is the sequence of all possible labels.

5 Construction of VEKG based on multi-source data fusion and context awareness

5.1 Multi-source VE knowledge integration

After recognising the entities, the entity dataset and raw corpus for VEKG construction were obtained. To improve the efficiency of VEKG construction, this paper firstly utilises the feature matching algorithm of similarity function for multi-source knowledge fusion to achieve entity alignment, based on which TEXTCNN is adopted to extract entity context-aware semantic features. The entities and relations in the KG are represented and learned through GAT to obtain entity embedding vectors with certain semantic information; and entity context-aware semantic features are fused with the entity embedding vectors to obtain more accurate entity embedding vectors, so as to construct a more efficient VEKG. The model structure of the suggested VEKG is indicated in Figure 4.

Knowledge fusion is first needed to disambiguate concepts and eliminate redundant and erroneous data, thus ensuring the quality of the final knowledge. For the set of entities to be aligned, this paper designs a feature matching algorithm based on similarity function to realise entity alignment. For the two entities E_1 and E_2 in the entity alignment process, the similarity function is defined as follows.

$$sim\left(E_1, E_2\right) = (1-\alpha)sim_{Attr}\left(E_1, E_2\right) + \alpha sim_{Stru}\left(E_1, E_2\right) \tag{16}$$

where $sim_{Attr}(E_1, E_2)$ is the attribute similarity function corresponding to the entity pair, $sim_{Stru}(E_1, E_2)$ is the structural similarity function corresponding to the entity pair, and $\alpha \in (0, 1)$ is the conditioning parameter.

For the attribute set U of E_1 and the attribute set V of E_2, the similarity is computed using the similarity function based on the edit distance for the attributes u and v, respectively, which are contained in both entities. The solution procedure is: Initialise a matrix A of $(|u| + 1) \times (|v| + 1)$. Denote the element of the i^{th} row and j^{th} column of A as $A_{i,j}$, where the values of $0 \le i \le |u|$, $0 \le j \le |v|$, and A are as follows:

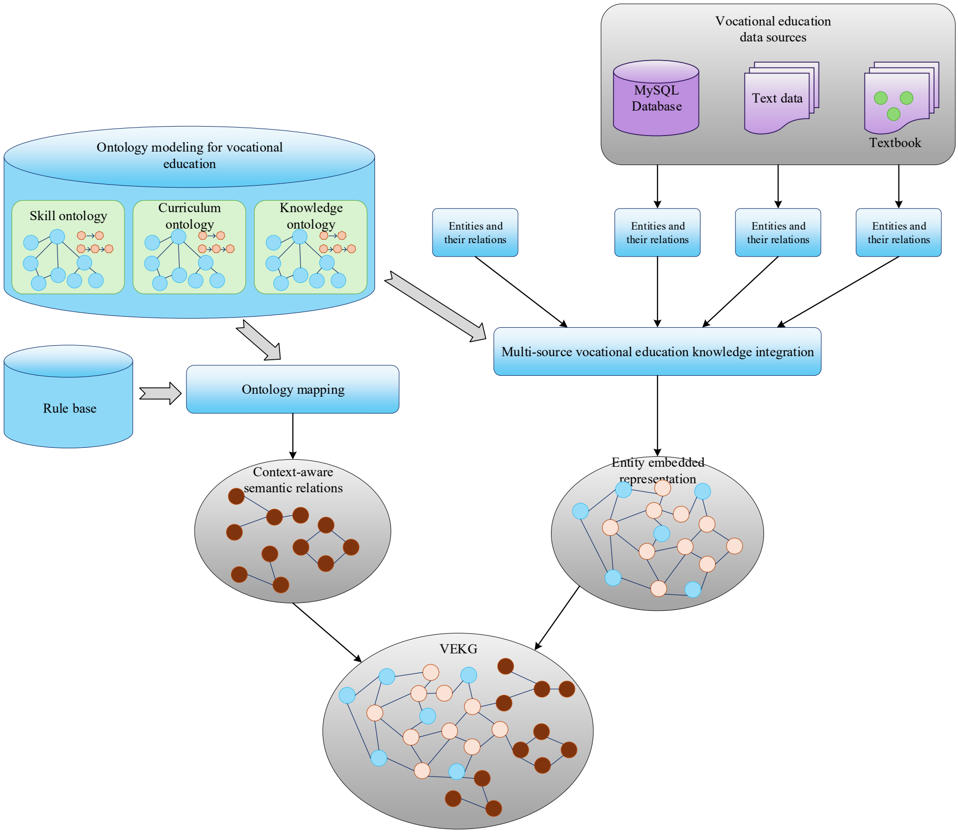
$$A_{i,j} = \begin{cases} A_{i-1,j-1}, u_i = v_j \\ 1 + \min\left(A_{i-1}, A_{i,j-1}, A_{i-1,j-1}\right), \text{otherwise} \end{cases} \tag{17}$$

Structural similarity is an important component of the entity pair similarity measure, which is measured by using the Jaccard correlation coefficient (Bag et al., 2019) of the entities' common neighbours. The advantage of the similarity function based on Jaccard coefficients is that the intersection operation of the sets is order-independent, i.e., the order of the different entities has no effect on the result, as shown below:

$$sim_{Stru}\left(E_1, E_2\right) = sim_{JaccardCoeff}\left(E_1, E_2\right) = \frac{\left|Stru\left(E_1\right) \cap Stru\left(E_2\right)\right|}{\left|Stru\left(E_1\right) \cup Stru\left(E_2\right)\right|} \tag{18}$$

where $Stru(E_1)$ and $Stru(E_2)$ are sets of common neighbours of entities.

Figure 4 The model structure of the suggested VEKG (see online version for colours)

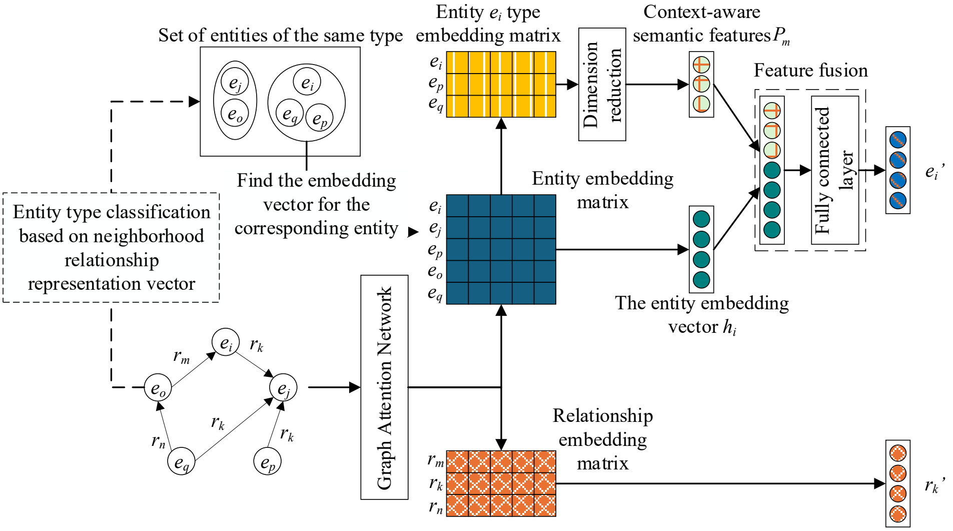

Vocational education data sources
MySQL Database
Text data
Textbook
Ontology modeling for vocational education
Skill ontology
Curriculum ontology
Knowledge ontology
Entities and their relations
Entities and their relations
Entities and their relations
Entities and their relations
Multi-source vocational education knowledge integration
Rule base
Ontology mapping
Context-aware semantic relations
Entity embedded representation
VEKG

Finally, the similarity of two entities can be obtained according to equation (16), and the entity pairs are filtered according to the set threshold, retaining the entity pairs within the threshold range, and manually checking the entity pairs exceeding the threshold, to confirm whether they are different expressions of the same entity, and completing the entity alignment.

5.2 *Embedded representation of KGs fusing context awareness and entity relationships*

After the integration of multi-source VE knowledge, the entity and relationship are embedded through GAT. Obtain the set of same-type entities through entity characterisation and type classification, and combine the entity embedding matrix to obtain the entity type embedding matrix, and perform entity context-aware semantic feature extraction on it to obtain the context-aware features of the entities. Finally, entity context-aware features are fused with entity embedding vectors, which are used to update the embedding vector representations of entities and relationships to construct more accurate VEKGs. The overall process is shown in Figure 5.

Figure 5 VEKG embedded representation process (see online version for colours)


Set of entities of the same type
Entity e_i type embedding matrix
Context-aware semantic features P_m
Dimension reduction
Feature fusion
Fully connected layer
e_i'
Find the embedding vector for the corresponding entity
Entity type classification based on neighborhood relationship representation vector
Entity embedding matrix
The entity embedding vector h_i
Graph Attention Network
Relationship embedding matrix
r_k'
e_i
e_j
e_o
e_p
e_q
r_m
r_k
r_n

1 GAT-based entity embedding vector representation. The embedding vector representations of the entities and relations of the triad (e_i, r_k, e_j) associated with entity e_i are firstly spliced and manipulated and linearly transformed using the linear transformation matrix as follows, where c_{ijk} is the semantic feature representation of (e_i, r_k, e_j) and W_1 is the linear transformation matrix.

$$c_{ijk} = W_1 \left[e_i, r_k, e_j \right] \tag{19}$$

To obtain the attention value of (e_i, r_k, e_j) for e_i, c_{ijk} is linearly transformed using the weight matrix W_2 and its attention value is obtained using the LeakyReLU activation function as follows:

$$b_{ijk} = Leaky\,\mathrm{Re}\,LU\left(W_2 c_{ijk} \right) \tag{20}$$

The weighted summation of the attention values of each node is then used to update the embedding vector representation of e_i.

$$h_i = \sigma\left(\sum_{j \in N_i} \sum_{k \in R_{ij}} b_{ijk} c_{ijk} \right) \tag{21}$$

The input initialised entity embedding matrix E_0 is linearly transformed with the help of the residual idea, where W^E is the linear transformation matrix, and is fused with the entity embedding matrix E' after the GAT update, as shown below:

$$E = W^E E_0 + E' \tag{22}$$

The relational embedding matrix G_0 is linearly transformed so as to update the relational embedding vector as follows, where W^R is the linear transformation matrix.

$$G = G_0 W^R \tag{23}$$

2 Entity context-aware semantic feature extraction. Dynamic word vector v_{ij}^{h} containing rich information is generated by the BERT model incorporating context-aware information in Section 4. Then v_{ij}^{h} is passed through the convolutional layers of TextCNN with three different sizes of 1D convolutional kernels to extract the feature vectors at different levels, which are calculated as follows.

$$c_i = f\left(W_m \cdot H_{i:i+h-1} + b\right) \tag{24}$$

where c_i is the feature vector obtained after the convolution operation, f is the nonlinear activation function, b is the bias term, W_m is the convolution kernel of several different sizes, and $H_{i:i+h-1}$ is the word vector matrix H consisting of submatrices from row i to row $C_m = [c_1, c_2, \ldots, c_{n-h+1}]$. Therefore, for sentences of length n, the corresponding set of feature mappings is formed, and the maximum value in the feature matrix is extracted, thus replacing the entire feature matrix and capturing the most important context-aware features for each feature matrix.

$$P_m = \max\left(C_m\right) \tag{25}$$

3 Fusion of entity context-aware semantic features and entity embedding vectors. To consider the interaction of feature information and map it to the same feature space, two different sources of information are fully integrated to obtain a more effective feature representation. Firstly, the two features are fused serially, and then the fused feature vectors are input to the fully connected layer for full fusion, and finally the vector expression combining the two features is obtained, as shown below:

$$e_i' = concat\left(h_i, P_m\right)W + b \tag{26}$$

where *concat* is the operation of splicing, h_i is the embedding vector of e_i, and e_i' is the embedding vector representation of e_i after fusion of h_i and P_m. W not only assigns weights to the fused feature vectors, but also changes the output dimensions of the feature vectors so that the dimensions of the fused feature vectors are consistent with those of the entity embedding vector representations for subsequent training in VEKG construction.

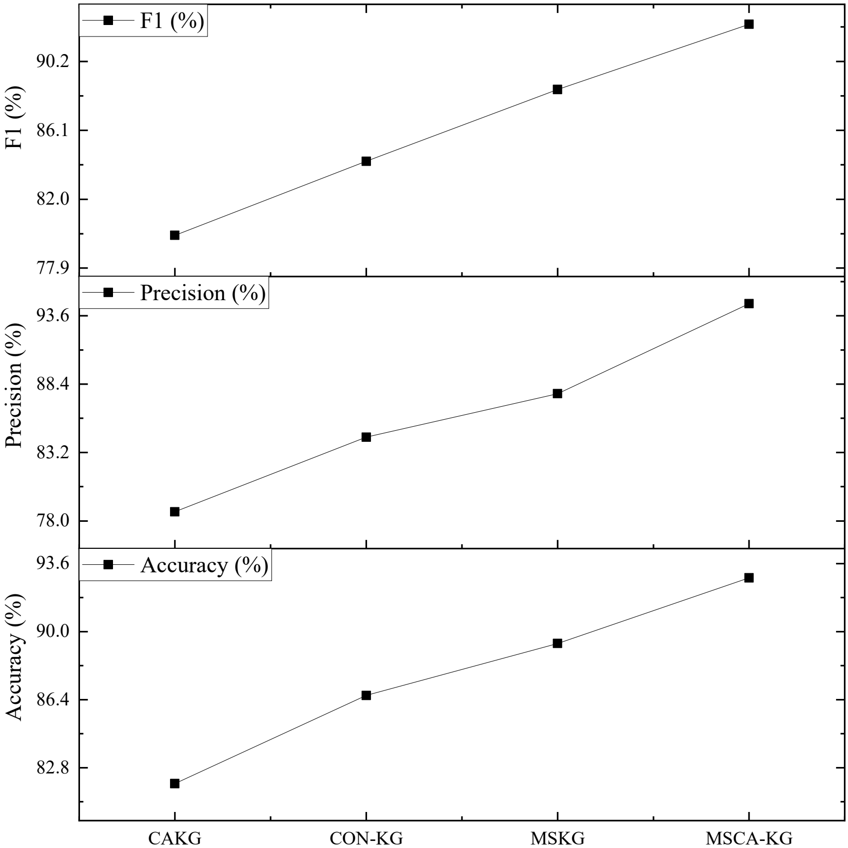
6 Experimental results and analyses

In this paper, through the crawler program on the national wisdom education platform, Baidu encyclopedia, MOOC wisdom education platform multiple data sources, the list of engineering and technology knowledge in each online VE platform to obtain, a total of 8,193 data records of hierarchical relationships between knowledge points, the number of entities 40,943, entity categories 2,596. The dataset is divided into training set and test set according to the ratio of 8:2. The initial input embedding vector dimensions for entities and relations are set to 50, the Dropout ratio is 0.5, the initial learning rate is 0.001, and the batch size is 128. The experiments were conducted with Python 3.5 and torch 0.4.1 on a PC server configured with an Intel(R) Core(TM) i7-8700 CPU 3.20 GHz, a 12 G NVIDIA Tesla K80C GPU, and 128 GB of RAM.

Table 1 Recognition accuracy for different entity classes (%)

Entity category	*OG*	*EI*	*EG*	*SG*	*EF*
CAKG	80.17	83.58	80.57	79.92	84.58
CON-KG	85.74	87.41	85.74	82.69	88.54
MSKG	88.69	89.17	88.15	87.59	90.78
MSCA-KG	92.01	94.33	90.11	92.58	94.57

Figure 6 Comparison of entity recognition performance of different methods



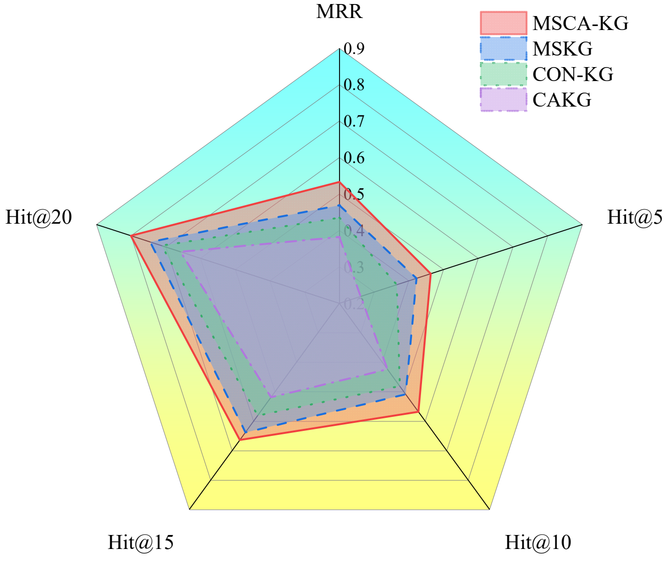
To comprehensively evaluate the effectiveness of the constructed VEKG (MSCA-KG), the recognition accuracies of five types of entities, namely, occupational category (OG), educational institution (EI), educational course (EG), skill category (SG), and employment information (EF) are analysed, and the comparative methods are selected as CAKG (Yang and Tan, 2022), CON-KG (Yu, 2022), MSKG (Wu and Jia, 2022), and the experimental results are shown in Table 1. CAKG performed the worst, with the lowest recognition accuracy for the five types of entities, and its BiLSTM-CRF was not optimised, although it was collected from multiple VE data sources. Although CON-KG and MSKG consider context-aware features, they do not collect data from multiple sources, and their recognition accuracy is not as good as that of MSCA-KG.

MSCA-KG demonstrated the best recognition accuracy above 90% for all five entity types.

In addition, the entity recognition accuracy, recall, and F1 value of the above four methods are also compared and experimented, and the experimental results are shown in Figure 6. The entity recognition accuracy and F1 value of MSCA-KG are 92.84% and 92.41%, respectively, which are improved by 10.89% and 12.56% compared to CAKG, 6.22% and 8.15% compared to CON-KG, and 3.47% and 3.88% compared to MSKG. After MSCA-KG introduces context-aware position vectors in entity recognition, the vector representation is more reasonable, and higher entity recognition accuracy is obtained by eliminating irrelevant words and retaining key words through word vector AM.

In addition to evaluating the accuracy of entity recognition in VEKG, this paper also compares the prediction performance of MSCA-KG with other models on the dataset through experiments based on the construction of KG embedded representations with the link prediction task as an application scenario, and the evaluation metrics are selected as mean reciprocal rank (MRR) and Hit@N, where N stands for the number of entities, and the results of the comparison are shown in Figure 7.

Figure 7 Prediction performance comparison on the link prediction task (see online version for colours)



The experimental outcome implies that the MRR of MSCA-KG is 0.533, which is 39.53%, 22.52% and 13.65% improved compared to CAKG, CON-KG and MSKG, respectively. When N is taken as 10, MSCA-KG improves 11.59%–33.96% compared to CAKG, CON-KG and MSKG. CAKG is an embedded representation of KG using a bilinear-based model, but the model has a high number of parameters, leading to high computational complexity. CON-KG is an embedded representation of KGs using TransE, but it is difficult to handle complex entity-relationship models and performs poorly in handling entity semantics. MSKG is modelling KGs through the ConvKB model, which is capable of handling complex relational models, but is the embedded representation capability is insufficient. MSCA-KG passes AM to compute the relationships between nodes and is able to automatically learn the correlations between nodes without the need to manually preset the relationships. This ability to learn automatically allows GAT to more accurately capture the complex connections between entities and relationships when processing KGs, thus improving the quality of the embedded representation.

7 Conclusions

To address the issues of single data source and ignoring the semantic features of context-awareness in the VEKG construction method, firstly, this article uses Scrapy to quickly crawl the internet data from multiple sources and extract various types of VE text data, analyse the ontology characteristics, and complete the mapping from database data to ontology data. The context-aware location information is then introduced into the BERT model to obtain an entity vector containing context-aware semantic information. And AM is introduced into the BiLSTM-CRF model to enhance the word vector representation. Second, the feature matching algorithm with similarity function is utilised for entity alignment, based on which TEXTCNN is adopted to extract entity context-aware semantic features. The entity embedding vectors with semantic information are obtained by GAT; and the semantic features of entity context-awareness are fused with the entity embedding vectors to obtain a more accurate representation of the entity embedding vectors. The experimental outcome implies that the entity recognition accuracy is 92.84% and the MRR on the link prediction task is 0.533, which greatly improves the construction efficiency of the VEKG.

Declarations

All authors declare that they have no conflicts of interest.

References

Angioni, S., Salatino, A., Osborne, F. et al. (2021) 'AIDA: a knowledge graph about research dynamics in academia and industry', *Quantitative Science Studies*, Vol. 2, No. 4, pp.1356–1398.

Bag, S., Kumar, S.K. and Tiwari, M.K. (2019) 'An efficient recommendation generation using relevant Jaccard similarity', *Information Sciences*, Vol. 483, pp.53–64.

Chang, C., Tang, Y., Long, Y. et al. (2022) 'Multi-information preprocessing event extraction with BiLSTM-CRF attention for academic knowledge graph construction', *IEEE Transactions on Computational Social Systems*, Vol. 10, No. 5, pp.2713–2724.

Chen, C. and Song, P. (2023) 'Evaluation method of diversified teaching effect of higher vocational education based on knowledge map', *International Journal of Sustainable Development*, Vol. 26, No. 4, pp.318–328.

Chen, P., Lu, Y., Zheng, V.W. et al. (2018) 'Knowedu: a system to construct knowledge graph for education', *IEEE Access*, Vol. 6, pp.31553–31563.

Endsley, M.R. and Garland, D.J. (2000) 'Theoretical underpinnings of situation awareness: a critical review', *Situation Awareness Analysis and Measurement*, Vol. 1, No. 1, pp.3–21.

Guangfen, Y. and Dongke, Z. (2017) 'An analysis based on Citespace III knowledge maps of Chinese vocational education research', *Chinese Education & Society*, Vol. 50, No. 5, pp.499–519.

Kovalchuk, V., Maslich, S., Tkachenko, N. et al. (2022) 'Vocational education in the context of modern problems and challenges', *Journal of Curriculum and Teaching*, Vol. 8, No. 11, pp.329–338.

Li, N., Shen, Q., Song, R. et al. (2022) 'MEduKG: a deep-learning-based approach for multi-modal educational knowledge graph construction', *Information*, Vol. 13, No. 2, p.91.

Liu, J., Li, T., Ji, S. et al. (2021) 'Urban flow pattern mining based on multi-source heterogeneous data fusion and knowledge graph embedding', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 35, No. 2, pp.2133–2146.

Platonova, R., Pankaj, V., Olesova, M. et al. (2019) 'Modernization of secondary vocational education system', *Journal of Environmental Treatment Techniques*, Vol. 7, No. 4, pp.562–565.

Sun, K., Liu, Y., Guo, Z. et al. (2016) 'Visualization for knowledge graph based on education data', *International Journal of Software & Informatics*, Vol. 10, No. 3, pp.13–17.

Sun, P. and Gu, L. (2021) 'Fuzzy knowledge graph system for artificial intelligence-based smart education', *Journal of Intelligent & Fuzzy Systems*, Vol. 40, No. 2, pp.2929–2940.

Tang, Y., Wu, X., Zhou, C. et al. (2023) 'Automatic schema construction of electrical graph data platform based on multi-source relational data models', *Data & Knowledge Engineering*, Vol. 145, p.102129.

Wei, Y.Y., Wei, Q.Y., Qin, C.Q. et al. (2023) 'Research on the construction and application of knowledge graph of digital resources in vocational colleges', *Journal of Computers*, Vol. 34, No. 4, pp.195–201.

Wu, Z. and Jia, F. (2022) 'Construction and application of a major-specific knowledge graph based on big data in education', *International Journal of Emerging Technologies in Learning (IJET)*, Vol. 17, No. 7, pp.64–79.

Yang, X. and Tan, L. (2022) 'The construction of accurate recommendation model of learning resources of knowledge graph under deep learning', *Scientific Programming*, Vol. 20, No. 1, pp.10–17.

Yu, B. and Fan, Z. (2020) 'A comprehensive review of conditional random fields: variants, hybrids and applications', *Artificial Intelligence Review*, Vol. 53, No. 6, pp.4289–4333.

Yu, Y. (2022) 'Research on the application of knowledge graph in constructing ecological chain of supply of lifelong learning resource base', *Open Access Library Journal*, Vol. 9, No. 9, pp.1–11.

Zheng, S. (2023) 'Combined application of employment education and big data internet technology based on the context of vocational education reform', *Applied Mathematics and Nonlinear Sciences*, Vol. 9, No. 1, pp.1–13.