Research on Adaptive Learning Path of Chinese Cultural English from the Perspective of Artificial Intelligence

Ruijun Duan*

Jinhua Polytechnic, Jinhua 321007, China

*Corresponding author: ruijunduan@qq.com

Keywords: Graph, Adaptive Learning, English.

Abstract: The educational thought of personalized education and adaptive learning has been widely advocated for many years. However, in the traditional classroom teaching, due to the shortage of teachers and the disorder of teaching resources, it is difficult to realize adaptive and personalized teaching. The paper is intended to create a path to cater for the personalized learning and adaptive learning in English learning, especial for the Chinese cultural English. Through the collection of data online and establishment of the English knowledge graph, personalized knowledge graph can be generated based on graph convolution neural network. In this way, all the nods of the knowledge graph tend to be presented. And at last personalized learning path is generated based on various algorithms. Consequently, the personalized learning path can be adapted according to the students' exercise situation and the difficulty level of themes and knowledge points. Meanwhile, it can be automatically revised as situation changes in the future.

1. Introduction

The educational thought of personalized education has been a hot topic for many years. However, in the traditional classroom teaching, due to the shortage of teachers and the disorder of teaching resources, it is difficult to realize personalized teaching. In the traditional classroom, teachers often rely on subjective experience to choose an appropriate speed to meet the needs of classroom teaching. The teaching steps accepted by all students are consistent.

The way of teaching has changed to a diversified modern way of teaching. In recent years, the research on knowledge graph is a heated research area of artificial intelligence, which has received a lot of attention in academia and industry. The generalized knowledge graph generally refers to all kinds of knowledge base projects. Formally, knowledge graph is a network structure composed of many interconnected entities and their attributes. Relevant knowledge questions and exercises in education can also be abstracted as nodes in the knowledge graph. These nodes can be used to form a knowledge graph. On the basis of this knowledge graph, personalized learning can be well studied. Computer technology can be combined with educational practice to realize personalized education and solve the problem of "teaching students according to their aptitude" One of the significance of this study is to provide different teaching services for different learners.

This paper establishes a knowledge graph for Chinese Cultural English exercises, generates a personalized knowledge graph for each student, recommends different exercises for different students, and fully reflects the idea of personalized education. By construction of the CHINESE CULTURAL ENGLISH knowledge graph based on NEO4J, personalized knowledge graph is generated based on graph convolution neural network. At last, personalized learning path is generated based on Prim algorithm and Kruskal algorithm.

2. Construction of English Knowledge Graph in Chinese Cultural English

Constructing knowledge graph cannot only help to integrate complex knowledge, but also clarify the complex links between knowledge points. When the amount of data is large and the relationship between data is complex, it is a good choice to express these knowledge integration as a knowledge

graph. Here, taking Chinese Cultural English exercises as an example, this paper studies the generation of personalized learning path. The final personalized learning path is composed of multiple Chinese Cultural English exercises.

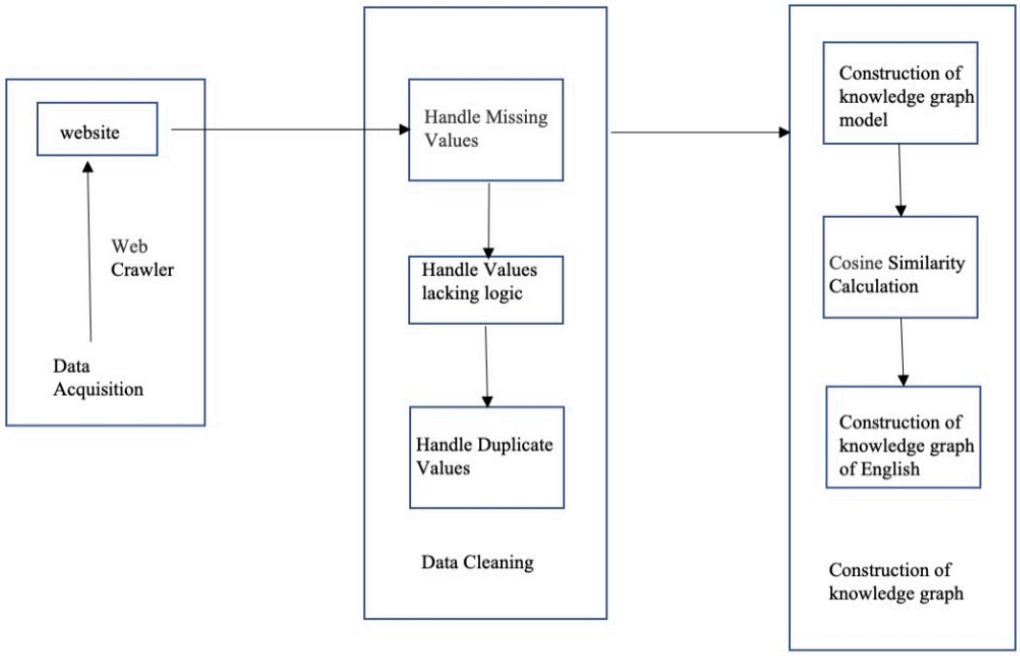

Fig 1. Overview of construction of knowledge graph.

2.1 Data Acquisition

The training of neural network can be viewed as a process of feature extraction. The so-called feature extraction is to obtain some common features from a large number of the same type of data through a model. When the new data belonging to this type is introduced into the learned model, the model can successfully judge the category of data. When training the model, the amount of data is very important, which determines whether the features learned by the model from the data are comprehensive, and further determines the final prediction result of the model. There are a lot of exercise data on the Internet and offline. Offline data collection needs manual input, which is time-consuming and laborious. The data on the network can be easily obtained by using web crawlers. This paper selects web crawlers to crawl the exercise data on the network. This paper crawls several web pages by scrapy, and finally obtains 12809 exercises. The number of exercises for investigating lexical and other knowledge points is 6589; the number of exercises to investigate syntax and other knowledge points is 6057; the number of exercises for comprehensive problems is 163. Most of the exercises taken this time are multiple-choice questions, followed by blank filling questions, reading comprehension, etc. multiple-choice questions include question stems, options, and analysis.

2.2 Data Cleaning

The data from different web pages are not always correct. These problematic data are called dirty data. The effect of neural network model will be directly affected by the data quality.

In the data preprocessing stage, we mainly do two things: data storage and data observation. After the data is stored, select several pieces randomly from the original data to observe the format and content of the data, so as to facilitate the subsequent processing of the data. There are 12809 exercise data crawled from the Internet in this paper, and it is planned to use Python code to process the crawled data later. Therefore, text files are selected for storage, and the data are stored in JSON format according to the information contained in the exercise. The JSON format data has a concise and clear hierarchical structure, which is not only convenient for reading, and it is convenient for machine analysis and generation.

The first step is to remove or complete the missing data. Then, the data with wrong format and content are modified by means of character replacement, manual proofreading and removing useless spaces. After cleaning the format and content, the next step is to modify and remove the data with logical errors. After a series of data cleaning operations, the current data quality has reached the requirements of the training model and can be used for model training.

After data cleaning, the amount of data changed from 12809 to 5931. Among them, there are 2992 exercises to investigate morphology, including modal verbs, nouns, conjunctions, prepositions and prepositional phrases, articles, non-finite verbs, verb voice, verbs and verb phrases, verb tenses, pronouns, adverbs and adjectives, numerals and similar words. There are 2725 exercises to investigate syntax, The knowledge points involved include subjunctive mood, imperative sentences, emphasis sentences, Nominal Clauses, sentence components, simple sentences, compound sentences, exclamatory sentences, antonym questions, object clauses, inversion sentences, paratactic sentences, attributive clauses, subject predicate consistency, adverbial clauses and general questions. There are 214 exercise data of comprehensive questions, including Cloze filling, reading comprehension Translation and composition.

2.3 Construction of English Knowledge Graph in Chinese Cultural English

After data cleaning, this paper has sorted the data into structured data, and then the construction of Chinese Cultural English knowledge graph. Knowledge graph adds rich entity information on the basis of ontology. The knowledge graph falls into two parts: vertical domain knowledge graph and open knowledge graph. The vertical domain knowledge graph has high requirements for the accuracy of knowledge, and the open knowledge graph has a wide coverage of knowledge. There are many ways to construct Chinese Cultural English knowledge graph. We choose to use Python expansion package cyphon and NEO4J graphic database to construct Chinese Cultural English knowledge graph. NEO4J is a graphical database of high-performance unstructured query language. The database structure is graphical, and the data is not stored in the table, but on the network. Using the expansion package in NEO4J and python, we can easily construct the Chinese Cultural English knowledge graph. In the process of constructing the Chinese Cultural English knowledge graph, word2vec is also used to generate sentence vectors for English exercises to determine the relationship between nodes.

The knowledge graph involved in this paper only contains Chinese Cultural English exercises, which is constructed from top to bottom. The central node of the knowledge graph is English. The second node is all the knowledge points involved in Chinese Cultural English. The categories of these knowledge points are consistent with those contained in Chinese Cultural English at the present stage. These knowledge points include lexical knowledge such as nouns and pronouns, as well as syntactic knowledge points such as attributive clauses and adverbial clauses. The third layer of the Chinese Cultural English knowledge graph is composed of exercises belonging to morphology and syntax and knowledge points such as reading comprehension and translation contained in the comprehensive questions. The questions contained under reading comprehension include daily life, biography, social news and other categories, and translation includes two types: English Chinese and Chinese English. The fourth layer consists of the analysis and answers of exercises in morphology and syntax and the exercises. The last layer consists of the answers and analysis.

The five layers contain two types of nodes: exercises and knowledge points. The nodes without text in the middle are knowledge point nodes, and the rest are exercise nodes. There are edge connections and similar relationships between exercise nodes and knowledge point nodes, and the relationship between exercise nodes and knowledge point nodes is belongs, that is, the exercise node belongs to a knowledge point.

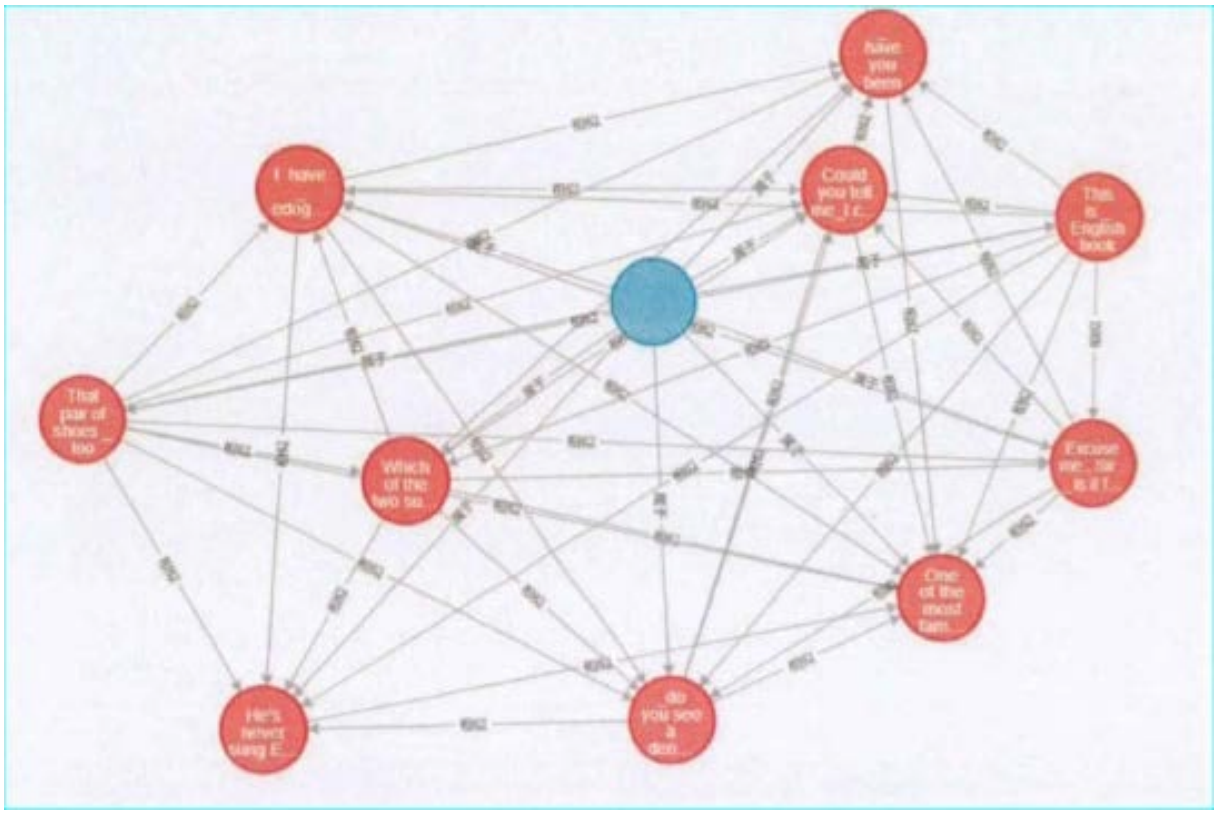

Fig 2. Knowledge graph.

3. Construction of personalized Chinese Cultural English Knowledge Graph

The generation of personalized Chinese Cultural English knowledge graph needs to take the original Chinese Cultural English knowledge graph as the input of graph convolution neural network. The first step is to generate a set of identical exercises according to the original knowledge graph. The exercises cover all knowledge points, and the number of exercises is 100. Through this set of exercises, we can obtain the mastery of knowledge points by different students. Then mark this part of the exercises in the knowledge graph according to the students' problem-solving situation. The questions that students do right are recorded as mastered, and the label is set to 1. The questions that students do wrong are recorded as not mastered, and the label is set to 0. The feature vector with dimension 512 is generated for each exercise by using word2vec. Then, the whole exercise data set is cut into training set and test set, in which the number of training sets is 9414, the topics of test set are 35, and all exercises in test set have labels. Adjacency list is a combination of sequential storage and chain storage. First, each node in the graph is numbered, and the number corresponding to the neighbor node of each node is stored in the chain list, and then all nodes are stored in the sequence list in numbered order. When using adjacency list to store weighted graphs, the weights on the edges between nodes are saved in the linked list. When storing non weighted graphs, there is no need to store any data. When an adjacency list is used to store an undirected graph, each node connected to the node is stored in the linked list corresponding to the node. When a directed graph is stored, only the nodes pointed from the node are kept.

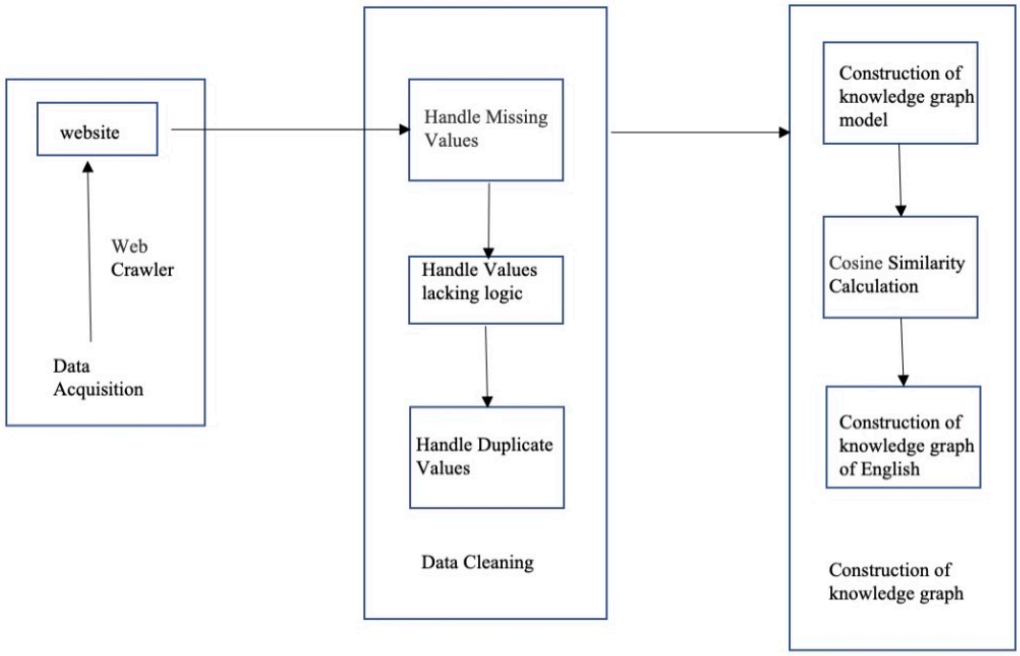

Fig 3. Construction process of personalized knowledge graph.

Eight different files are generated according to the knowledge graph, test set and training set, which are necessary for training the graph convolution neural network. The eight files are: the adjacency matrix corresponding to the knowledge graph, the eigenvector corresponding to the exercise data in the test set, the eigenvector corresponding to the exercise data in the training set, the eigenvector corresponding to the labeled exercise data in the training set, the label information of the labeled exercise data in the training set, the label information of the exercise data in the test set, label information of exercise data in training set and number of exercise data in test set in knowledge graph. When generating knowledge graph, the corresponding adjacency matrix can be directly generated by using Python's numpy expansion package and according to the definition of adjacency matrix. Numpy is a scientific computing tool that supports python programming language and can easily perform matrix operation and storage. When generating the adjacency matrix with numpy, firstly, a two-dimensional matrix with the size of 5931 * 5931 and the value of 0 is defined.

4. Personalized Learning Path Generation

After the personalized Chinese Cultural English knowledge graph is generated, the personalized learning path is created by referring to the graph path generation algorithm in the data structure according to the category of exercise nodes, the difficulty of exercises and the similarity between exercises in the personalized knowledge graph. And the personalized learning path is continuously modified.

4.1 Difficulty Evaluation of Exercises

When generating personalized learning path, the exercises are sorted according to the difficulty value of the exercises. When evaluating exercise difficulty, this paper investigates the difficulty of exercise from three aspects: whether there are hyper-syllabic words in exercise, the similarity of answers, and the number of knowledge points in exercise investigation. Among them, lexical similarity is divided into synonyms and shape similarity. For composition, translation and other question types, the difficulty degree cannot be evaluated according to the above aspects. For such question types, the difficulty degree value is set as the average value of the difficulty degree value of all exercises. This paper crawls some exercise data with difficulty value from the Internet, analyzes these exercises from the above three aspects, summarizes the conclusions from different evaluation

angles, and corresponds the three difficulties of simple, medium and difficult with the scores of 1, 2 and 3 respectively.

4.2 Path Generation Based on Personalized Knowledge Graph

The exercises constituting the personalized learning path are selected in the personalized knowledge graph through the students' problem-solving situation and the similarity between exercise nodes. Then the personalized learning path is generated according to the difficulty of the exercise. Before generating the personalized learning path, the exercises constituting the personalized learning path need to be selected. The vertices are divided into two categories: included and not included in the tree in the search process. When generating the minimum spanning tree, all vertices in the initial state are classified as the second category. Then, select any vertex as the starting point and move it from the second category to the first category, then find the vertex with the lowest weight between the vertices in the second category and the first category, and move it from the second category to the first category.

This is repeated until there are no vertices in the second class. The vertices and edges traversed are the minimum spanning tree corresponding to the graph. When generating the minimum spanning tree, kruska algorithm first constructs a subgraph with only n vertices and empty edge set. The exercise path created in this paper requires the similarity and maximum which is consistent with the idea of generating the minimum spanning tree. In addition, in this paper, it is not required that all the selected exercise nodes are connected.

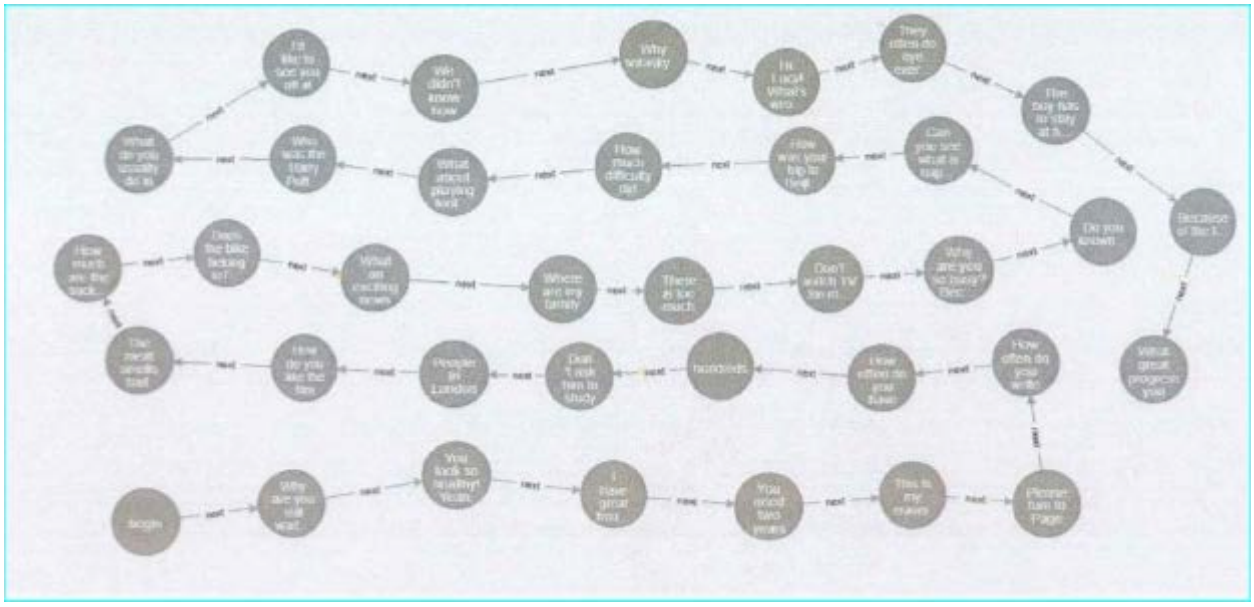

Fig 4. Personalized learning path.

Before selecting the exercise node, you need to obtain the students' problem-solving situation of the last set of questions, and select the exercise according to the students' problem-solving situation. Considering the actual situation of students' psychological state, learning ability and learning burden, and referring to the number of exercises in the current Chinese Cultural English test paper, the number of exercises in the personalized learning path is set to 35. When selecting the exercise node in the personalized learning path, first find the exercises that the students did wrong according to the students' last exercise results, and count the number of these exercises.

Acknowledgements

This work was financially supported by fund of visiting scholars and teachers' professional development project in Colleges and Universities and the work is the result of the project “On the path of integrating Chinese traditional culture into higher vocational English teaching from the perspective of system function”.

References

[1] Ashraf, B., Doaa, S.: The need for a paradigm shift in CSCL. In: The Computing Conference 2017. IEEE, London (2017)

[2] Marcheggiani D, Titov I. Encoding sentences with graph convolutional networks for semantic role labeling [J]. arXiv preprint arXiv:1703.04826, 2017.

[3] McCarthy, B., et al.: Journey to Personalized Learning (2017)

[4] Ryu S, Kwon Y, Kim W Y. A Bayesian graph convolutional network for reliable prediction of molecular properties with uncertainty quantification [J]. Chemical science, 2019, 10(36): 8438-8446.

[5] Scarselli F, Gori M, Tsoi A C, et al. The graph neural network model [J]. IEEE transactions on neural networks, 2008, 20(1): 61-80.