A Personalized English Learning Material Recommendation System Based on Knowledge Graph

https://doi.org/10.3991/ijet.v16i11.23317

Yiqin Huang (✉), Jiang Zhu
Zhejiang Business College, Hangzhou, China
hyqing329@163.com

Abstract—The world has rushed into the information age. As the lingua franca, English shapes the global landscape of information transmission and exchange. Mastering English is equivalent to possessing an important tool for acquiring precious information. Therefore, it is very necessary to improve English teaching. This paper analyzes the problems in traditional classroom teaching and online learning of English, and discusses how to keep students unconfused in online learning, improve their English learning efficiency, and satisfy their personalized demands. Specifically, the relevant data were characterized by knowledge points in English teaching, and used to formulate a knowledge graph. Then, the related knowledge points were labeled in online learning data on learning platforms. After that, user portrait was created by analyzing the data on daily learning behaviors. Finally, collaborative filtering was coupled with content-based recommendation to push English learning resources to students, which meet their personalized demands.

Keywords—Knowledge graph, personalized English learning, learning resources recommendation, collaborative filtering, recommendation system

1 Introduction

As digital economy and artificial intelligence are developing rapidly, our world has rushed into the information age. As the lingua franca, English shapes the global landscape of information transmission and exchange. Mastering English is equivalent to possessing an important tool for acquiring precious information [1]. No doubt, English is now an important information transmission tool for fast acquisition of various cutting-edge technologies and professional knowledges. In the field of education, English teaching runs from elementary schools to colleges and universities. Nowadays, English is no longer just a language tool for ordinary people to communicate with each other, but also a useful means for people to acquire knowledge and information from various media [2]. Therefore, the quality and effect of English teaching are of crucial significance. The traditional English classroom teaching has very limited time and space, within which teachers are not able to provide students with sufficient English learning materials or language application opportunities, let alone the one-on-one personal tutoring, and the learning demands of students of different levels
cannot be met either. In terms of online platforms and mobile APPs, although there are massive learning resources available for students, two common problems often accompany online English learning, namely: (1) Lacking the professional guidance from teachers, students are often confused when facing such massive learning data; they might lose their learning interest and get lost by distractions, or even feel befuddled and bewildered with the knowledge they learnt; (2) Students are usually of different learning and comprehension ability levels; even for students in a same class, there are learning level hierarchies, and different students have different demands for knowledge learning. Existing online learning platforms mostly divide knowledge content based on students’ grades, they generally unify and organize the knowledge without considering students’ individual demands, which has resulted in issues such as insufficient learning materials for good students, while average and poor students never learn [3, 4]. Therefore, in view of these two issues, how to achieve personalized recommendation of learning resources based on the characteristics of English knowledge and the individual demands of learners is a question worthy careful consideration.

Regarding the first issue mentioned above, with English course of a certain grade as an example, this paper analyzed the characteristics of English as a linguistic tool and its knowledge attributes, sorted out the knowledge system structure, and built knowledge graph of the course to describe the knowledge points. In such case, students would be able to learn the knowledge structure of the course according to the knowledge graph without the professional guidance of teachers, avoiding them from getting lost in the massive learning resources. As for the second issue, this paper analyzed the user portrait of students in the grade, the learning behavior paths of these students, as well as the specific situations of the online teaching platforms; then with the help of knowledge graph, it labeled the related knowledge points of learning data on online learning platforms, constructed the personalized English learning material recommendation system, and researched a few key points such as the system recommendation flow, user portrait, and recommendation algorithm, etc.

2 Construction of Knowledge Graph

With the advent of the era of big data, data are accumulated continuously and hardware is upgrading and iterating at an astonishing speed [5], on this basis, schools and various educational institutions have developed a series online English learning platforms and APPs, which contain a huge amount of English learning resource libraries that accumulated massive education data resources (such as audio files for English listening, teaching videos, courseware, exercise questions, test papers, and interactive games, etc.), however, these resources generally have the characteristics of unstructured or semi-structured, and diverse in forms, the traditional data mining methods or machine learning methods cannot use them directly [6]; and the processing of these education data has determined the effectiveness of the personalized recommendation system in subsequent researches. In English teaching, the teaching content is usually formulated centered on the knowledge points that should be mastered by students in a
certain grade, therefore, knowledge points are the focus of teaching and testing, and they should be taken as the main input information of the intelligent recommendation algorithms. For this reason, this paper attempted to construct the knowledge graph based on knowledge points of a certain grade, and then applied technologies related to knowledge graph and network to describe the nodes in the corresponding graphs, and the advantages lie in that [7]: (1) knowledge graph can well reflect the system structure of subjects and visualize the internal structure of knowledge and the connections between knowledge nodes; (2) knowledge graph has good scalability and mobility, and can well adapt to the subsequent expansion of the knowledge system and the improvement of the algorithm.

Knowledge graph is an important concept developed with the use of mind graph in recent years [8], it mainly consists two parts: (1) the “Entity-Relation-Entity” triplets; (2) the entities and their corresponding attributes, they are usually value pairs or relationship expressions that are similar to keys/values in programming languages. Taking the “Topic 1–Nouns” in the junior high school English course as an example, the knowledge graph of the topic can be plotted as a network-shaped knowledge library structure, as shown in Figure 1 [9]. In subsequent works, it was converted into structured language that can be understood by computers, for instance, a triple of the entity relation of English teaching can be <teacher, lecture, course>, <student, grade, first grade of junior high school>, <course resource, include, knowledge point label>. At the same time, this paper took the knowledge points of the course as the entities of the knowledge graph, then, for a specific knowledge point "noun", the triple can be expressed as <noun, classify, proper noun>, or <noun, classify, common noun>, and so on.

![Figure 1. Description of knowledge of an English course topic in junior high school](http://www.i-jet.org)

Constructing knowledge graph can help students expand and transfer course knowledge and cultivate their independent learning ability; at the same time, it can better solve their problem of “lost in online resources”, and it facilitates the personalized recommendation. Based on the knowledge graph constructed in previous text, the resources of online platforms need to be integrated with knowledge points as the main
content of the characterization technique, then the related knowledge points are used to normalize the specific content of the network resources and process them into characteristics that can be understood by the machine and the algorithm (such as classification, labels, knowledge library). Labels are highly refined characteristics identifications obtained through the analysis of resource information, and multimedia contents such as images, videos, and audios in the learning resources often involve image recognition and processing, and it’s quite difficult for machine to comprehend the contents directly and label them, therefore, for resources in these formats, a set of user labeling mechanisms could be established and manually fill out and make the labels. For example, when uploading resources to the teaching platforms or the back-end database of APPs, the system can remind teachers to manually label the related knowledge points, with the "Lesson 55 The Sawyer family" in the text book *New Concept English Volume 1* as an example, the involved knowledge points include the phrases of the verb “go”, the grammar knowledge of general present tense, and the changes of third-person predicate verbs, etc. When teachers have uploaded DOC files, PPT files, audios, videos, exercises and other related learning materials, they can label the related knowledge points in the software interface. The names of these related knowledge points are specific entities in the content of the knowledge graph. Teachers do not have to input by themselves, but can choose in the graphical interface, as shown in Figure 2 below.

![Fig. 2. Labels of related knowledge points of documents](image)

### 3 Technical Framework of Knowledge Graph

After the knowledge graph is constructed and the related knowledge points of the resources are labeled, the unstructured data resources have been preprocessed into structured data that can be easily recognized by computers [10]; and the next thing to
be considered is using database to store them. Nowadays, relational database and graph database are the mostly commonly-used database; relatively speaking, graph database is much more flexible than relational database. When the content volume of knowledge graph increases, the number of system layer increases as well, and the efficiency of graph database will be hundreds or even thousands of times higher than that of relational database. For example, the commonly-used Neo4j graph database [11] is a high-performance NOSQL graph database, instead of storing structured data in tables, it stores structured data on the network to form graphs. Neo4j graph database has a high-performance graph engine [12] and the important features of a mature database, such as transactions and indexes. A graph generally contains two basic data types: Nodes and Relationships; Nodes and Relationships contain attributes in the form of key/value. Nodes are connected by the connections defined by Relationships to form a relational network structure [13]. For instance, for the knowledge graph entity relationship triples <teacher, lecture, course>, <teacher, lecture, student> <course resource, include, knowledge point label> mentioned in previous text, taking a student as an example, the relationship creation steps of teachers, students, and courses stored in the Neo4j database are as follows:

1. Create a student node, including node name, label name, attribute name, and attribute value, etc.:

   ```
   create (<node-name>:<label-name>{
   <property1-name>:<property1-Value>,
   <property2-name>:<property2-Value>,
   ...
   <property3-name>:<property3-Value>
   })
   ```

   For example, if a student’s student ID is 200610201, his name is Alex, his age is 13 years old, and his gender is male, then the student node can be described as: create (s:Student{id: 200610201, name:"Alex",age:13,sex:1})

2. Create relationship: create teacher relationship for the student. For example, for an English teacher named Tom with an age of 35, there is:

   ```
   create(t:Teacher{id:2000000121,name:"Tom",age:35,sex:1,teach:"English"})return t
   ```

   Assume the class taught by teacher Tom has 3 students: Alex, Jane, and Lucky, here we need to create the relationships between the teacher and the 3 students. The first is to create the relationship between Tom and Alex:

   ```
   match (t:Teacher),(s:Student) where t.id=2000000121 and s.id=200610201
   create (t)-[teach:Teach]->(s)
   return t,teach,s
   ```
Then, we continue to create the relationships between teacher Tom and students Jane and Lucky:

```cypher
match (t:Teacher),(s:Student) where t.id=2000000121 and s.id=200610202
create (t)-[teach:Teach]->(s)
return t,teach,s
match (t:Teacher),(s:Student) where t.id=2000000121 and s.id=200610203
create (t)-[teach:Teach]->(s)
return t,teach,s
```

The relationships between the teacher and the students are stored in the Neo4j graph database as the relationship description shown in Figure 3 below. Subsequently, for the knowledge point labels of the courses taught by teachers, the course learnt by students, and the course resources, the same steps are taken to create nodes and relationships, and finally forming a knowledge graph-based database network structure.

![Teacher-student relationships created in the Neo4j graph database](image)

Knowledge graph is not constructed at a time, but accumulated slowly through constant repetitions, iterations, and updates [14]. As shown in Figure 4 below, to generate a complete knowledge graph structure, the construction process mainly includes 4 stages: (1) Data collection: the resources on the teaching platform are mainly composed of unstructured data (such as images, audios, videos, and texts), and some semi-structured data (such as XML, JSON, and Encyclopedia); these two kinds of data cannot be well recognized by computers, and need to be converted into structured data to perform subsequent algorithmic calculations; (2) Information extraction: extract the three content items of attributes, relationships, and entities from the content data source, and form knowledge expressions on this basis; (3) Knowledge fusion:
after obtaining the new knowledge expressions, use anaphora resolution and entity disambiguation methods to process them to eliminate the contradictions and ambiguities of knowledge points. For example, some knowledge in English may have multiple expressions, and a specific appellation may correspond to different entities; (4) knowledge processing: the fused new knowledge needs to be evaluated for its quality by the system or professional English teachers, then the qualified parts are added to the knowledge library to ensure that the generated knowledge graph is of high quality.

Fig. 4. Flow of knowledge graph construction and updating

4 Creation of User Portrait

An accurate and real-time personalized learning recommendation system needs to help the web page-end or the APP-end of the platform to create accurate user portrait, analyze the students’ online and offline learning behaviors, capture the correct learning scenarios, and find the right “person”, only in this way can the system realize effective personalized commendation [15]. In this paper, the user portrait technology refers to the modeling of student users of the English teaching platform based on student data: first, the system collects data such as students’ school attributes, learning habits, and learning behaviors, etc.; then, using the user portrait technology based on knowledge graph, deeper level information that is related to students’ learning requirements are extracted from the existing student data, and the different data are abstracted to form a student learning model with labels. Just like labeling the learning resources with related knowledge points mentioned in previous text, each platform student user also needs to be labeled to facilitate using recommendation algorithms to classify student users and extract information in subsequent research; a student’s user portrait can be created as shown in Figure 5 below. The triple entity relationship of
the knowledge graph can be expressed as <student A, classmate, student B>, <student A, gender, male>, etc., in this way, the knowledge graph structures of specific user portraits could be created, and the knowledge graph information could be stored in the Neoj graph database.

Fig. 5. Creation of user portrait

5 Architecture of the Personalized English Learning Material Recommendation System

Personalized recommendation of English learning resources means to discovering a student’s English learning interest or similar learning content based on his/her past learning behaviors or learning records, and then providing the student with learning resources that meet his/her individual needs [16]. The architecture of the recommendation system ensures the automatic and real-time operation of the entire recommendation processes; it receives student requests, collects, processes, and stores student data, performs calculations using recommendation algorithms, and returns the recommendation results. The architecture of the personalized recommendation system built in this study is shown in Figure 6 below.
As shown in the Figure, a complete recommendation system should include core modules such as content source, content processing, user mining, and recommendation algorithm, etc.

1. Content source: to establish a comprehensive personalized recommendation system, the data collection sources should be wide and deep enough. First, collect the buried data points from web pages and APPs of the platform, embed invisible probes in web pages, collect user learning behavior data and platform operation log, store data, and then the server is responsible for putting the collected log information into the storage device.

2. Content processing: if the collected basic data are structured data, then they can be processed directly by computer language; for unstructured and semi-structured data, as discussed in previous text, the content needs to be sorted out into specific knowledge points, the content is constructed by knowledge graph, marked by labels, and stored into the Neoj graph database for efficient reading and storage.

3. User mining: this step mainly includes four contents: user behavior log collection, transmission, mining, and storage. The collection of behavior log can still adopt the front-end data-point burying method to report the users’ clicks, sharing, collection and other behaviors. Transmission requires stable transmission and update of the log, so that a certain operation of the user can be quickly fed back to the next recommendation [16]. The mining process is to calculate and mine user data into the features we want, namely the "user portraits". Here we need to add attribute labels, interest labels, behavior labels and customized labels to users.
4. Recommendation algorithms: after completing the first three steps, the system has got the content and user data; after that, algorithms can be used to match the two. For recommendation systems, one of the most classic algorithms is the user-based collaborative filtering algorithm [17-20], and the purpose of the algorithm is to recommend courses or knowledge points that are similar to student users’ interest, and this algorithm has two steps: (1) The courses or knowledge points learnt by the object student user are taken as the characteristics of the learner to calculate the learner similarity matrix; (2) According to the learner similarity matrix, other K students who are similar to the object student user are obtained, then with similarity added as a weight, from courses/knowledge points learnt by other similar students, the top N courses/knowledge points are recommended to the object student. However, the biggest flaw of the collaborative filtering algorithm is the cold start problem. When a new user has just entered the platform, there is no history information of the user, so the system cannot perceive its existence and it is difficult to make personalized recommendations. Another classic recommendation algorithm is the content-based recommendation algorithm. This algorithm creates configuration files for users and learning materials respectively; then by analyzing the content browsed by a user, a configuration file of the user is created; after that, by comparing the similarity between the user and the learning materials, the most matching learning materials are recommended to the user. The English materials of the learning platforms are mostly text and multimedia data. Previous text has introduced how to use knowledge graph to describe course content and knowledge system and how to create user portrait; the labeling method is adopted to make marks using the labels, which are then stored in the Neo4j graph database, therefore, the content-based recommendation can make up for the defects of the user-based collaborative filtering algorithm and solve the user cold start problem. As a result, the mixed recommendation technology combining user-based collaborative filtering algorithm and content-based recommendation algorithm can improve the accuracy of the recommendation system.

6 Mixed Recommendation Technology

In the system architecture shown above, we have mentioned that the mixed recommendation technology adopted by the algorithm layer is the core content of the entire system. The accuracy and real-time performance of the system are closely related to it and the effective mixed recommendation technology is important for improving the performance of the recommendation system and enhancing the recommendation effect. As the recommendation technology field has been developed so far, the user-based collaborative filtering is a common method that has certain advantages in terms of the novelty of recommendation results, but the relevance of recommendation results is relatively weak, popular learning resources are recommended more frequently, and thus losing the meaning of personalized learning, moreover, as mentioned above, there’s also the problem of cold start with it. The content-based recommendation algorithm is the simplest and most intuitive recommendation algorithm, it
doesn’t have the cold start problem and it judges whether to recommend a content based on the similarity, therefore, its accuracy depends on the descriptions of the learning resources, namely the labels made during content mining, and this has determined that this algorithm is limited by the in-depth analysis of the content of texts, images, audios, and videos. The two methods have their respective pros and cons, and both have certain limitations. Simply applying one recommendation technology cannot guarantee high-quality recommendation results; therefore, this study proposed to combine the two methods to make up for their respective shortcomings. About how to mix the two to truly maximize their strengths and avoid weaknesses, two measures were proposed as follows:

First, in terms of the mixed recommendation framework, the algorithm’s calculation cycle, online response time, and resource consumption should be considered, and the accuracy of the recommendation results needs to be taken as a reference, we need to find a balance between the two. A lot of practical applications have proved that, by applying the three-staged system, namely the Online-Nearline-Offline three-layer mixed mechanism, a relative ideal balance between the two could be found. In this mechanism, the Online module is used to quickly capture the behavior information of students or platform users, such as what courses have been studied, what videos have been clicked and scored, etc., and this module requires a Cache system to process and respond to API requests. The Nearline module makes use of the user events and uses stream-oriented computation to obtain intermediate results, on the one hand, these intermediate results are sent to the online part to update the recommendation model in real-time; on the other hand, these intermediate results are stored as well. The Offline module mines the long-term and massive user behavior logs, undertakes high-load complex algorithms, and regularly takes out the data stored in the data warehouse for batch operations and model updates; therefore, it consumes large resources and requires a long mining cycle, but the quality of the calculation results of the Offline recommendation system is often the highest, and these results will be transmitted online through the Nearline system. These three layers need to work collaboratively to bear high-load and lightweight algorithms to quickly respond to user behaviors, ensuring the high reliability and high concurrency performance of the recommendation system and the high accuracy of recommendation results.

Second, in terms of the algorithms, applying the weighted mixed recommendation technology can yield better recommendation results. That is, run the two recommendation algorithms separately to generate some candidate results, then perform weighting and get the final recommendation ranks and results. A relatively simple weighting method is to give the same weight value to the collaborative filtering recommendation result and the content-based recommendation result, and then generate the recommendation results. However, the fixed weighting system ignores that the quality of the algorithm varies according to the different learning objects and under different learning scenarios, therefore, it needs to consider setting training samples and compare whether the user’s evaluation of the recommendation results is consistent with the system’s prediction, then, the system generates the weighted model based on the training results, and dynamically adjusts the weights to generate recommendation results, the flow of the algorithm is shown in Figure 7.
7 Conclusion

Combining with the characteristics of English knowledge, this paper analyzed the problem that students often get lost in the massive online learning materials and built an efficient and accurate personalized English learning material recommendation system. The proposed system used knowledge points of English course to characterize the data and create the knowledge graph, and the relevant technologies of the proposed system had been introduced in detail. To realize personalized and customized recommendations, the learners’ behavior data were taken into consideration to create complete user portrait and the specific knowledge graph structure; and user label data was combined with the mixed recommendation technology (integrating the user-based collaborative filtering algorithm and the content-based recommendation algorithm) to make recommendations. The operation and implementation of the proposed system can help learners save a lot time spent on data search and query, and it can improve the efficiency of English learning.

8 Acknowledgement

This article is the research result of the Vocational Education Special Research Project (No. 2018ZSZJZD08) funded by The Commerce Economy Association of China.
References

[1] Warschauer, M. (2000). The changing global economy and the future of English teaching. TESOL Quarterly, 34(3): 511-535. https://doi.org/10.2307/3587741

[2] Fan, J. (2011). Constructing web-based learning environment for college English teaching. International Conference on Information and Management Engineering, 1(233): 515-528. http://dx.doi.org/10.1007/978-3-642-24010-2_69

[3] Alsubhi, M.A., Sahari, N. (2020). A Conceptual Engagement Framework for Gamified E-Learning Platform Activities, International Journal of Emerging Technologies in Learning, 15(22): 4-23, https://doi.org/10.3991/ijet.v15i22.15443

[4] Harvey Arce, N.P., Cuadros Valdivia, A.M. (2020). Adapting Competitiveness and Gamification to a Digital Platform for Foreign Language Learning, International Journal of Emerging Technologies in Learning, 15(20): 194-209. https://doi.org/10.3991/ijet.v15i20.16135

[5] Maddumala, V.R., R.A. (2020). A weight baseweight-based transformation model on multifaceted multimedia bigdata using convolutional neural network. Ingénierie des Systèmes d’Information, 25(6): 729-735, https://doi.org/10.18280/isi.250603

[6] Liu, Q., Li, Y., Duan, H., Liu, Y., Qin, Z.G. (2016). Knowledge graph construction techniques. Journal of Computer Research and Development, 53(3): 582-600. http://dx.doi.org/10.7544/issn1000-1239.2016.20148228

[7] Liu, Q., Wang, F., Wang, X. (2019). Deep representation and intelligent application of educational resources. artificial intelligence, (3): 44-54. http://dx.doi.org/10.16453/j.cnki.issn2096-5036.2019.03.007

[8] Ge, Q.P., Zhong, Y.R. (2019). Construction of Knowledge Atlas Based on Mathematics Teaching. Computer Technology and Development, 29(3): 193-195. http://dx.doi.org/10.3969/j.issn.1673-629X.2019.03.039

[9] Hao, P.H., Gao, J. (2018). Visualization Analysis of Police Security Knowledge Map Based on Neo4j Graph Database. Modern Computer, (35): 8-11. http://dx.doi.org/10.3969/j.issn.1007-1423.2018.35.003

[10] Xiao, Q.D., Qi, L.L., Hou, X. (2017). Design and implementation of knowledge mapping system of curriculum system based on neo4j graph database. Computer Knowledge and Technology, 13(36): 130-132.

[11] Patel, J., Strickman, M. (2006). Statistical personalized recommendation system. US.

[12] Zhang, Z.K., Zhou, T., Zhang, Y. C. (2010). Personalized recommendation via integrated diffusion on user–item–tag tripartite graphs. Physica A: Statistical Mechanics and its Applications, 389(1): 179-186. https://doi.org/10.1016/j.physa.2009.08.036

[13] Guo, X., Lu, J. (2007). Intelligent e-government services with personalized recommendation techniques. International Journal of Intelligent Systems, 22(5), 401-417. https://doi.org/10.1002/int.20206

[14] Gan, M., Jiang, R. (2013). Improving accuracy and diversity of personalized recommendation through power law adjustments of user similarities. Decision Support Systems, 55(3): 811-821. https://doi.org/10.1016/j.dss.2013.03.006

[15] Zeng, W., Shang, M.S., Zhang, Q.M., Lü, L., Zhou, T. (2010). Can dissimilar users contribute to accuracy and diversity of personalized recommendation? International Journal of Modern Physics C, 21(10): 1217-1227. https://doi.org/10.1142/s0129183110015786

[16] Shi, Y., Larson, M., Hanjalic, A. (2014). Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. ACM Computing Surveys (CSUR), 47(1): 1-45. https://doi.org/10.1145/2556270
[17] Gholami, A., Forghani, Y. (2020). Improving multi-class Co-Clustering-based collaborative recommendation using item tags. Revue d'Intelligence Artificielle, 34(1): 59-65. https://doi.org/10.18280/ria.340108

[18] Xiong, C.P., Sun, H., Pan, D., Li, Y. (2019). A personalized collaborative filtering recommendation algorithm based on linear regression. Mathematical Modelling of Engineering Problems, 6(3): 363-368. https://doi.org/10.18280/mnep.060307

[19] Liu, Y., Yang, H., Sun, G.X., Bin, S. (2020). Collaborative filtering recommendation algorithm based on multi-relationship social network. Ingénierie des Systèmes d’Information, 25(3): 359-364. https://doi.org/10.18280/isi.250310

[20] An, Q.Q. (2019). A novel recommendation algorithm considering average similarity and user-based collaborative filtering. Mathematical Modelling of Engineering Problems, 6(3): 390-396. https://doi.org/10.18280/mnep.060310

10 Authors

Yiqin Huang is a member of School of International Studies, Zhejiang Business College, Hangzhou, China, Email: hyqing329@163.com. She is an associate researcher and mainly engaged in vocational education management and education internationalization research.

Jiang Zhu is a lecturer of Zhejiang Business College, Hangzhou, China, Email: 36946278@qq.com. He mainly researches in information technology and ideological and political education.

Article submitted 2021-03-14. Resubmitted 2021-04-26. Final acceptance 2021-04-27. Final version published as submitted by the authors.