A Knowledge Map Mining-Based Personalized Learning Path Recommendation Solution for English Learning

Duong Thien Nguyen[1,2][a] and Thu Minh Tran Nguyen[1,2,*][b]
[1]*Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam*
[2]*Viet Nam National University, Ho Chi Minh City, Vietnam*

Keywords: Knowledge Graph, Personalized Learning Path, Recommendation, English Learning, Graph Database.

Abstract: Recommendation systems (RS) have been widely utilized across various fields, particularly in education, where smart e-learning systems recommend personalized learning paths (PLP) based on the characteristics of learners and learning resources. Despite efforts to provide highly personalized recommendations, challenges such as data sparsity and cold-start issues persist. Recently, knowledge graph (KG)-based RS development has garnered significant interest. KGs can leverage the properties of users and items within a unified graph structure, utilizing semantic relationships among entities to address these challenges and offer more relevant recommendations than traditional methods. In this paper, we propose a KG-based PLP recommendation solution to support English learning by generating a sequence of lessons designed to guide learners effectively from their current English level to their target level. We built a domain KG architecture specifically for studying English certification exams, incorporating key concept classes and their relationships. We then researched and applied graph data mining algorithms (GAs) to create an effective PLP recommendation solution. Using consistent experimental conditions and a selected set of weights, along with our collected dataset, we evaluated our solution based on criteria such as accuracy, efficiency, stability, and execution time.

1 INTRODUCTION

The amount of data has increased dramatically along with the internet's quick development. Users find it challenging to select what interests them from a wide range of options due to the information overload. RS has been developed to enhance the user experience and aid in making decisions. Personalized recommendations are generated by RS based on user behaviour and preferences, which increases user engagement and happiness on a variety of online platforms, such as music recommendations, movie recommendations, learning path recommendations, online shopping recommendations, etc. (Q. Guo et al., 2020). Despite significant progress, creating a RS specifically suited to provide suitable content remains a challenge. Accurately predicting user and content characteristics and their complex interrelationships is one of these issues. Thus, researchers' attention has been drawn in recent years to the introduction of KG as side information in the RS. A heterogeneous graph with nodes signifying entities and edges denoting relationships between entities is called a KG. To comprehend the relationships between objects, items and their properties can be mapped into the KG. Additionally, users and user-side data may be included in the KG, improving the accuracy of capturing user preferences and relationships with items (Q. Guo et al., 2020).

RS has also benefited academic sectors in many ways since it has driven the creation of smart learning systems. The aims, interests, and abilities of each learner are the learning criteria that these techniques adjust PLP to. Utilizing data-driven monitoring to make sure that learners' parameters are fulfilled, they change the content and order of learning materials, marking the shift from a one-size-fits-all approach to customized learning methodologies (M. Abed, 2023).

Among today's subjects of study, we give special attention to foreign language learning since it plays a

[a] https://orcid.org/0009-0002-2906-8423
[b] https://orcid.org/0009-0009-6961-3976
* Corresponding author

Nguyen, D. and Nguyen, T.
A Knowledge Map Mining-Based Personalized Learning Path Recommendation Solution for English Learning.
DOI: 10.5220/0012945200003838

In *Proceedings of the 16th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2024) - Volume 1: KDIR*, pages 48-59
ISBN: 978-989-758-716-0; ISSN: 2184-3228

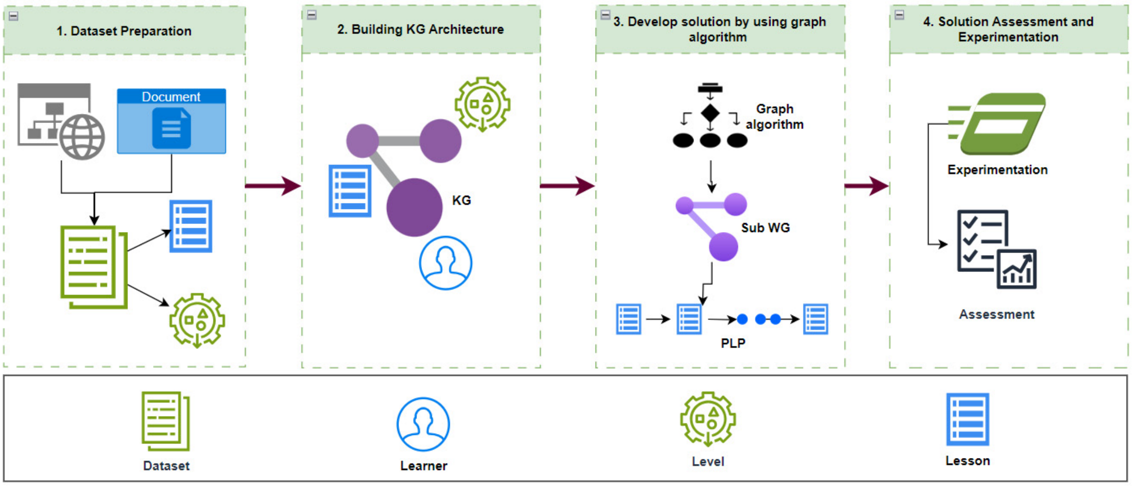
1. Dataset Preparation
Document
2. Building KG Architecture
KG
3. Develop solution by using graph algorithm
Graph algorithm
Sub WG
PLP
4. Solution Assessment and Experimentation
Experimentation
Assessment
Dataset
Learner
Level
Lesson

Figure 1: An overview flowchart illustrating the execution process for the proposed solution.

critical role in job application, study and research, travel, participation in global interchange, etc. (Ilyosovna, N. A., 2020). According to a Statista report published in 2023, English is presently the most common language in the world, with almost 1.5 billion users (E Dyvik, 2023), and certification of competency in English with four primary skills—listening, speaking, reading, and writing—is also frequently expected in job applications and university output standards. As a result, to demonstrate their ability to use English fluently, many individuals must study and prepare for tests to get international English credentials such as TOEIC (M. Schedl, 2010), IELTS, TOEFL (GH Sulistyo, 2009), and others. Learning these qualifications is now much more convenient, owing to the support of smart English learning applications and systems such as Duolingo, Elsa, and others, which allow learners to study more successfully while saving money and time (X. Fan et al., 2023). According to our survey results, these applications generally guide learners to learn in a pre-set sequence based on their goals and current level; however, to guide learners along a suitable learning path (LP) that meets their other personalized requirements, such as time, cost, progress, and learning outcomes, they are also being researched to apply the PLP recommendation (PLPR) models to advise learners on a suitable LP and fulfil the aforementioned aims.

According to that motivation, this study proposes a PLPR solution for English learners using GAs on the KG architecture, which we have developed in the English learning domain. As per the process illustrated in Fig. 1, the proposed solution will proceed through four primary phases of development. Initially, a dataset pertaining to the format, content, and assessment methods of international English certifications and associated exams will be compiled by referencing many websites and official publications. In the second phase, a KG architecture will be constructed to present key concepts about English proficiency levels, knowledge, and skills that are needed, as well as the lessons that correlate to those levels based on the built dataset and the learners' learning requirements. To optimize execution time, in the third phase, we will next create a weighted subgraph (WG) based on the learner's requested learning information in the KG. This created WG will only contain entities and weighted relationship edges related to the target level, as well as the lessons or competencies the learner possesses that correspond to the current level. Next, leveraging the GAs from Neo4j's GDS library (Hodler et al., 2022), we recommend the most effective initial PLP for learners. To identify the optimal set of weights for our solution that meets all LP assessment criteria, we conducted extensive experiments with various weight configurations. Subsequently, employing consistent experimental conditions and a selected set of weights, along with our dataset, we evaluated our solution based on criteria including accuracy, efficiency, stability, and execution time.

This paper is organized as follows: Section 1 outlines our work and its contributions. Section 2 provides an evaluation of the current state of the art in the research field. The KG architectural development process is described in depth in Section 3. Section 4 describes the steps involved in creating a PLPR solution for learners. Section 5 describes the experiments, evaluations, and data collection methods. Finally, Section 6 concludes with suggestions for future research directions.

2 RELATED WORKS

A lot of approaches have been put forth by researchers to increase learning efficiency, and one of the most cutting-edge areas of study these days is developing systems to recommend LPs to e-learners as a chain of learning materials. As a result, numerous studies have been conducted to create RSs for LP recommendations that use semantic dependency links between learning objects (LOs) and learning materials that are simultaneously stored on a variety of data types to recommend LPs to learners, then utilize data mining models to arrange learning materials into learner-recommended LPs. These RS systems do (D. Shi et al., 2020), however, still adhere to the notion of a single learning path that is applicable to all learners and are not actually tailored to the unique learning characteristics of each learner, which results in the recommendation of LPs to learners with limited suitability.

As the research by D. Shi et al. (2020) illustrates, KG has been employed recently for LP recommendation as a prominent research domain since it may eliminate ambiguities in learning content and learner’s learning characteristics descriptions. Motivated by this feature, a few researchers attempted to develop learning systems for KG-based LP recommendation and were successful in resolving the issues raised. Huang and Xiangli (2011) used AI, data mining, and database technology to create a PLP recommendation system (PLP-RS). By fine-tuning learner models with learning history data, it improves specialized services and assesses improvement using customized Knowledge Structural Graphs. Zhang et al. (2023) provide a PLP-RS for e-learning that uses a KG structure by creating a multidimensional course KG and applying graph convolutional networks (GCN) to properly represent learner preferences. The algorithm recommends ideal courses based on both learner preferences and the significance of learning resources, decreasing the need for manual planning and increasing learner satisfaction. Shi et al. (2020) construct a multidimensional KG by connecting learning elements semantically. Their algorithms provide customized LP creation and suggestions, meeting each learner’s unique e-learning demands. Static code analysis is used by H. Yin et al. (2021) to build a structural KG program for open-source projects. Through depth-first and Dijkstra search algorithms, their deep learning model, which integrates this with multi-source data and an LP recommendation mechanism, helps developers quickly learn important functions. Using GCN on Junior High School English exercises, Y. Sun et al. (2021) in the field of English education generate individualized KG for pupils, creating PLP with the aid of Prim and Kruskal algorithms. In their work on computer-assisted vocabulary acquisition, F. Sun et al. (2020) develop a recommendation engine for Chinese vocabulary learning materials utilizing a KG. Hanyu Shuiping Kaoshi (HSK) three-level language resources and ten types of relations are integrated into the system, which was created using Protégé, Apache Jena, and Python. Chen et al. (2021) use course similarity computation and pre-knowledge annotation to automate the creation of Massive Open Online Courses on KG. They use rule-based and machine learning techniques to classify courses, improve TF-IDF computation, and build a network that integrates knowledge and course nodes. The knowledge network and learner data are then used to provide personalized suggestions. Z. Yan et al. (2023) suggest a technique that makes use of a course knowledge network to suggest customized activities. The method entails building the graph using deep knowledge tracing, producing individual knowledge structure diagrams, and producing a Q-matrix from learners’ responses. The model chooses tailored assignments based on factors such as complexity, individuality, and variance, which is consistent with constructivist learning theory.

The aforementioned studies all share the same goal of investigating LP recommendation models for every learner utilizing KG by incorporating concepts such as goals, learner behaviors, LOs, and learning resources, etc., along with their interrelationships, into the architecture of the KG. It has been demonstrated that this method outperforms conventional ones in personalized recommendation outcomes. However, based on the research of M. Abed et al. (2023), we think that other aspects of the learner's learning characteristics, such as the learner's current level of knowledge and skills, desired learning time and cost, etc., must be considered to recommend a more appropriate PLP for each learner. Moreover, little study has been done on learning foreign languages like English. Our study focuses on using KG-based data modelling and processing to develop a solution that recommends a PLP for English language learners. This solution will account for the learners' current knowledge and skills, target level, and desired learning time, enabling them to achieve their goals efficiently within the shortest possible time while adhering to an appropriate learning path.

Section 3 will provide a detailed presentation of the steps involved in developing the KG architecture as well as the proposed solution for this research.

3 DOMAIN KG CONSTRUCTION

We examined the vocabulary topics, grammar themes, scoring scale, format, assessed skill, and evaluation criteria of the TOEIC, IELTS, and TOEFL test components to develop a KG architecture for presenting the concept and learning material along with their relation in preparation for the English certification examinations, as shown in Fig. 1 for the second phase. Furthermore, in accordance with European norms, we investigated the Common European Framework of Reference (CEFR) (B. North et al., 2019) to assess the link between certificate scores and English competencies.

As far as we know, people who want to get an international standard English certificate have to first complete a competence exam and receive the certificate along with a score demonstrating their ability. The score on these certificates does not indicate whether the individual passed or failed, but it does demonstrate their level of English ability. The outcomes can be transferred to the CEFR to standardize English proficiency levels across European and other regional nations. As a result, to manage learners' test information for international English certificates, the KG architecture will include a *Level class* that is focused on storing score information from the current English certificate of the learner and the target score of the certificate that the learner hopes to attain in the future. Each certificate's score information, together with qualification information based on the associated CEFR framework, will be saved as a benchmark to examine the correlation of English proficiency to scores between various certificate types.

Besides, success in international English certificate exams necessitates skills in speaking, listening, reading, and writing, as well as mastery of vocabulary and grammar knowledge. Therefore, the Competency class contained in the KG architecture will cover all the necessary skills and knowledge. However, we will construct specific pronunciation and vocabulary knowledge in a separate Lesson class since we understand that this knowledge is only tested in certain parts of the exam. This personalized approach guarantees that important abilities are covered in every segment of the test.

Moreover, learners must fully comprehend the sorts of questions, subjects, and settings that will be posed in each part of the exam. Combine knowledge of grammar, vocabulary, pronunciation, and English abilities to create the ideal test-taking plan. That is, for each level of English that a learner wishes to achieve, the learner must be provided with knowledge from specific lessons on clearly understanding the structure, question type, topic context, strategies, and test-taking experience in that skills test, as well as knowledge from related grammar and vocabulary lessons. As a result, in the KG architecture, an extra Lesson class will be created to manage information about lesson entities that must be learned to pass the exams. Our proposed comprehensive KG, which is represented in Fig. 2, consists of three fundamental concept classes: Level, Competency, and Lesson

For the Level concept class, it will include entity categories such as *Current_Score* and *Target_Score*, which indicate the learners' current score via the *HAS_CURRENT_SCORE* relationship and target score via the *WANT_TARGET_SCORE* connection for the same certificate type with the same properties: *score, certificate*. These certificate's score entity nodes will be referenced to the CEFR competence framework entity nodes, which have properties such as *from_score*, *to_score*, and *certificate*, via the

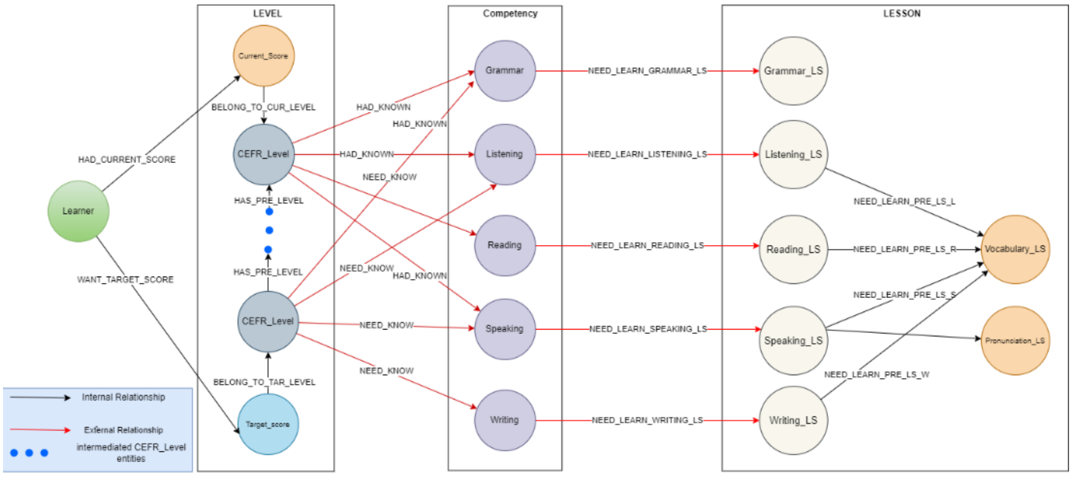

Figure 2: Complete KG architecture utilized in the proposed solution.

internal relationships *BELONG_TO_CUR_LEVEL* and *BELONG_TO_TAR_LEVEL*. Additionally, a *HAS_PRE_LEVEL* relationship will connect CEFR_Level nodes, e.g., a 'B1' level will precede 'A2' according to CEFR.

The Lesson concept class will manage entities comprising main lesson content related to grammar, listening, reading, speaking, and writing skills for each test section. Every lesson has common properties, including *title*, *category*, and *study time* (in days). Additionally, certain preparatory lessons with assigned pronunciation or vocabulary knowledge must be completed before advancing to the main lessons via system linkages like *NEED_LEARN_PRE_LESSON_L,* etc.

Finally, the Competency concept class will signify the learner's current proficiency level on the English certificate and identify acquired skills or knowledge through the *HAD_KNOWN* external relationship with entities like grammar, listening, reading, speaking, and writing lessons. Correspondingly, in alignment with the learner's target level, it outlines the requisite skills and knowledge through the *NEED_KNOW* relationship. Each skill and knowledge entity within the *Competency* class denotes the associated lessons required from the *Lesson* class, establishing external relationships like *NEED_LEARN_GRAMMAR_LS, NEED_LEARN_LISTENING_LS*, and others.

4 A SOLUTION FOR PLPR

4.1 Description of PLPR Problem

The main goal of our proposed solution is to recommend an appropriate PLP for each learner as they go toward preparing for international English certifications like the TOEFL, IELTS, and TOEIC (which include the Speaking-Writing and Listening-Reading combinations). The scores that correspond to the learners' current certificates, information about their English proficiency or lessons that correspond to their current level (which will be raised for the learner to choose based on our developed dataset), the desired study time, and the score that corresponds to the desired certificate are the first inputs of the solution. These inputs will be stored in the KG architecture (as shown in phase 2 of Fig. 1). Our system will then produce an initial PLP for each learner as indicated in phase 3 of Fig. 1, which will include a list of lessons to be learned and progress the learners from their current English level to their target level while accommodating their desired learning schedule.

4.2 End-to-End Solution Processing

As was indicated in Part 3, KG would house all data pertaining to the learning characteristics of learners as well as data on the acquisition and evaluation of English certifications. Simultaneously, we want to provide solutions using a novel approach that is simple to implement while maintaining natural logic and science. Because of this, we have examined and assessed GAs according to several factors, including the KG architecture, the issue that has to be addressed along with the intended outcomes at each stage of the solution's execution, and the fundamentals of how each algorithm works. In particular, we select the graph traversal algorithm BFS (section 4.2.1) for stage 1 of the solution in order to be able to create a subgraph with only entities connected to the learner's target-level entity. The PageRank algorithm is then combined with the LPA_NI algorithm in stage 2 of the solution with the aim of determining the importance of each lesson entity on WG and clustering these entities into clusters corresponding to the list of lessons to be learned from the current level to the target level, then merging into the original LP (section 4.2.2). Lastly, we use the Min Weighted Sum (MWS) approach to assess and determine which LP is the most appropriate as a recommended PLP for learners and meet the optimization objectives in stage 3 (section 4.2.3). Fig. 3 will provide details of the processing flow for each stage, precisely as follows:

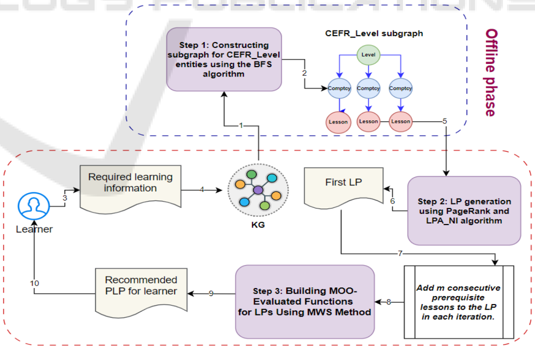
Step 1: Constructing subgraph for CEFR_Level entities using the BFS algorithm
CEFR_Level subgraph
Level
Comptcy
Comptcy
Comptcy
Lesson
Lesson
Lesson
Offline phase
1
2
5
Required learning information
3
4
Learner
KG
First LP
6
Step 2: LP generation using PageRank and LPA_NI algorithm
7
Online phase
Recommended PLP for learner
10
9
Step 3: Building MOO-Evaluated Functions for LPs Using MWS Method
8
Add m consecutive prerequisite lessons to the LP in each iteration.

Figure 3: The execution flow with applied algorithms.

Step 1: Constructing a subgraph for CEFR_Level entities using the BFS algorithm: This initial step involves offline processing to traverse the KG using the BFS algorithm (S. Huan, 2014). The goal is to generate subgraphs for each CEFR_Level entity. By doing so, we create a comprehensive list of all entity categories that are directly or indirectly connected to each CEFR_Level entity. This approach reduces the number of entity interactions, thereby optimizing the execution time for subsequent steps.

Step 2: LP generation using PageRank and the LPA_NI algorithms: This step occurs during the online processing phase. It starts by transforming the CEFR_Level entity subgraph into a WG tailored to the learner's target level, based on information provided by the learner. Next, the PageRank algorithm (C. Tulu, 2020) is employed to evaluate the relevance of each node within the WG. Following this, the LPA_NI algorithm (Zhang, 2017) groups significant nodes into clusters, reflecting the competencies and lessons required for each CEFR_Level entity in the WG. By merging these clusters and sorting them in ascending order according to the CEFR_Level entity values within each cluster, the initial learning path comprising the primary lessons is obtained.

Step 3: Building Multi-Objective Optimization (MOO)-Evaluated Functions for LPs Using the MWS Method: This step will also be completed online. The LP made in step 2 is to keep adding *m* significant nodes in the WG as prerequisite lessons as nodes in the Vocabulary or Pronunciation entity category and then utilize the developed evaluation function to gauge the LP's satisfaction at each k^{th} iteration by using the MWS method. Then, as the PLP to counsel the learner, select the LP that produces the most optimal outcome while meeting all stated optimization objectives. In the following sections, we will present the details of these main processing steps.

4.2.1 Constructing Subgraph for CEFR_Level Entities

The implementation procedure for step 1 is detailed in Algorithm 1.

Algorithm 1: Constructing subgraph for each CEFR Level entities in KG.

```
Input:
    ▪ G (V, E): The KG includes V vertices and E relationship edges.
    ▪ LVL = {LVL_i | i = $\overline{1,n}$}: set of the $i^{th}$ CEFR_Level entity denoted as $LVL_i$ contained in G (V, E).
    ▪ n: number of elements in the LVL set.
Output: $LV\_EN_i$ (set of entities related to each $i^{th}$ CEFR_Level entity).
$LV\_EN_i \leftarrow \emptyset$ , $i \leftarrow 1$
while $i \le n$:
    Apply the BFS with each $LVL_i$ as the source vertex → Obtain a set containing k nodes {$v_1$, $v_2$, …, $v_k$}
    $LV\_EN_i \leftarrow$ {$v_1$, $v_2$, …, $v_k$}
    $i \leftarrow i + 1$
End while
```

For example, after executing this algorithm, as shown in Fig. 4, we obtain a subgraph of the CEFR_Level node with the value "CEFR_B2," representing learner X's target level. This subgraph includes nodes related to the learner's current level, their existing competencies, the skills they need to acquire, and the lessons that they might have to learn.

4.2.2 LP Generation Using PageRank and LPA_NI Algorithms

To clearly explain the implementation process in step 2, we introduce the notations outlined in Table 1 and describe the two primary tasks. The first task involves using the PageRank algorithm, detailed in Algorithm 2. Additionally, we use Eq. 1 (C. Tulu et al., 2020) to calculate the PageRank score (PR_score) for each node in the WG derived from the subgraph defined in step 1:

$$PR(i) = (1-d) + d\sum_{j\to i}\frac{W_{ji}PR(j)}{\sum_k W_{kj}} + PR'(i) \qquad (1)$$

Algorithm 2: Determine each node's significance within the WG.

```
Input: TAR, CUR, subgraph of TAR as G' (v, e), CPT_HAD, LS_KNOWN.
Output: IPT set.
$WG \leftarrow G'$
For each edge e point to node u in WG:
    If (u ∈ CPT_HAD)||(u ∈ LS_KNOWN)
        e.weight ← 0
    Else e.weight ← 1
For each node u in WG:
    If (u == CEFR_Level entity)
        PR_score(u) ← 1
    Else PR_score(u) ← 0
While not converged:
    For each node u in WG:
        PR_old ← PR_score(u)
        Using Eq.1 to calculate PR_score(u)
        If |PR_score(u) - PR_old| < threshold:
            break loop
IPT = {u | PR_score(u) > 0 and sort by PR_score(u) decreasing}
```

Note that in Eq. 1 (E. Turan et al., 2020), *PR(i)* denotes the PR_score calculated for each node *i* in the *LV_EN* set during the current iteration, while *PR'(i)* represents the existing PR_score of node *i* from the previous iteration, indicating the spread of points among related nodes. *PR(j)* refers to the current PR_score of nodes *j* in the *LV_EN* set linked to node *i*. The weight of the edge from node *j* to node *i* is denoted as W_{ji}. Similarly, W_{kj} represents the weight of nodes *k* in the *LV_EN* set, pointing away from node *j*. The damping factor *d*, set as 1, reflects the probability of the learner accessing node *i* from node *j*, ensuring

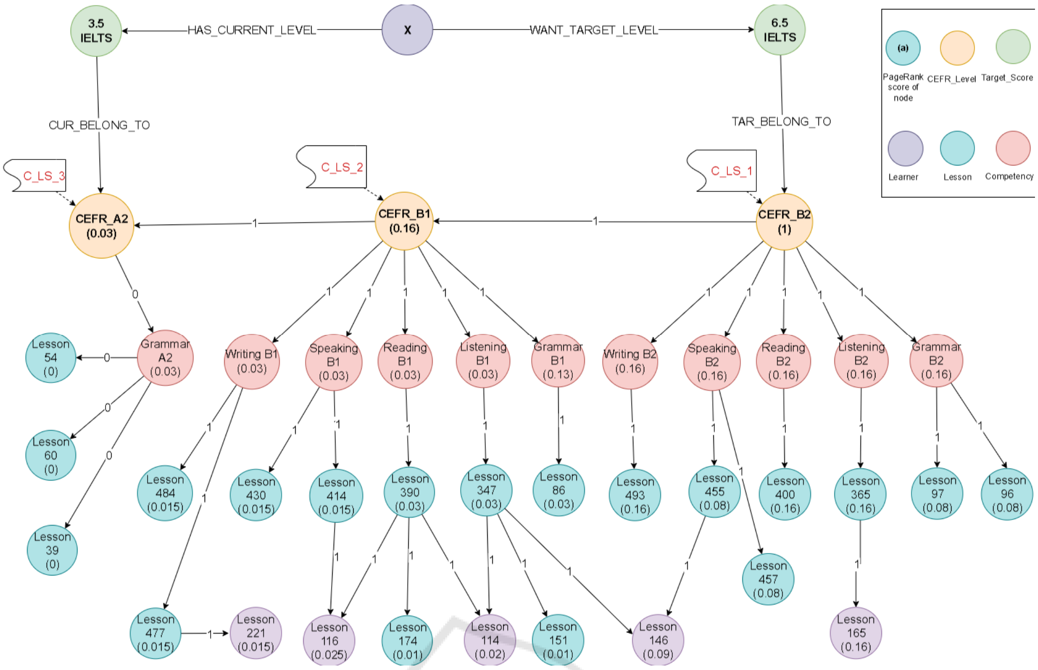

Figure 4: Weighted subgraph for the learner's target level completed after executing step 1 and 2.

Table 1: The meaning of the signs used in Step 2.

Signs	Meaning
LV_EN	Set of entities (competence, previous CEFR level, lesson) related to the learner's target CEFR level.
TAR	English proficiency according to the CEFR framework on the target certificate that the learner wants to achieve.
CUR	English level on the current certificate according to the CEFR framework that the learner currently has.
CPT_HAD	Set of competencies that the learner already has. Equivalent to a competency number belonging to CUR.
LS_KNOWN	Set of lessons that the learner has learned before (lessons that the learner can optionally learn) belongs to the competencies of CUR.
EN_IPT	Set of CEFR_Level, Competency, and Lesson entities has decreasing importance to learners according to their PR score, which is greater than 0.
INTM_LV	The set contains intermediate CEFR_Level nodes between TAR and CUR.
CL_EN_u	The u^{th} cluster contains nodes with the same label after each label propagation step.
LN_LS_u	The set contains only lesson entities filtered from the corresponding CL_EN_u clusters.

learning from node *j* to node *i* for pairs with $W_{ji} = 1$. For instance, in Fig. 4, based on the learner's input data, each edge pointing to a node in the subgraph is given a weight value of either 0 or 1. The subgraph will be transformed into the WG following this weight assignment procedure. Once Algorithm 2 has run on this WG and assigned a PR_score to each node, we will add these nodes to the *EN_IPT* set in decreasing order of their PR_scores. Fig. 5 presents the *EN_IPT* set as an example.

Based on the WG architecture designed in Fig. 4, when learner X wants to achieve a TAR (e.g., level "B2") from a CUR (e.g., level "A2"), learners must also achieve the Competencies of the intermediate

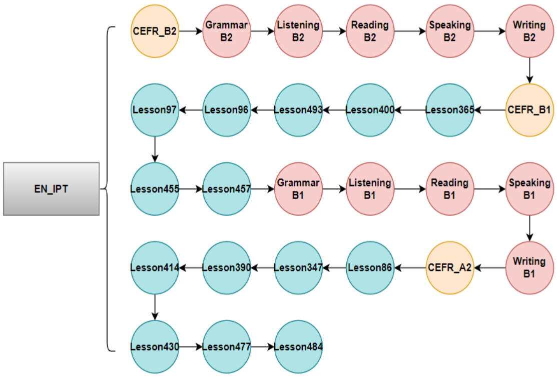

Figure 5: Entity nodes included in the EN_IPT set.

levels (INTM_LV) (for example, “B1”). To guarantee that learner X studies enough lessons for the needed Competencies from CUR to TAR, the second task in this step will cluster the most critical lessons to learn (according to the PR_score of each node in the WG) into each cluster at each level of proficiency. LPA, a well-liked clustering algorithm (Čížková, K. 2022), uses graph design to build a label propagation mechanism for random nodes. Nevertheless, we will use the method developed by Zhang et al., which is called LPA_NI, to boost second task efficiency. When propagated, LPA_NI has been demonstrated to provide superior clustering results over regular LPA based on node importance and label influence. Eqs. 2 and 3 (Zhang et al., 2017) are used by the LPA_NI for this step, where *LI (i, lb)* denotes the label's influence *(lb)* on node *i*, *d(j)* denotes the outdegree of node *j*, $N^l(i)$ denotes the set of labels *lb* surrounding node *i*, c_i denotes the most influential label that will be assigned to node *i*, and *l_max* denotes the sets of the maximum number of labels.

```
Algorithm 3: Building the first LP.
Input: WG of TAR, CUR, EN_IPT, MaxIter
       (Maximum number of execution loops)
Output: LN_LP.
 1: Initialize seedLabel for CEFR_Level nodes in
    WG.
 2: t ← 1
 3: For each node x ∈ EN_IPT:
 4:     Assign label of most represented connected
        node.
 5: If connected nodes' labels to x are all different:
 6:     Calculate viral influence using Eq. 2.
 7: Choose label satisfying Eq. 3 to update node x.
 8: If t = MaxIter or labels of node x match majority
    connected nodes' labels:
 9:     Assign nodes x to CL_EN_1, CL_EN_2, ...,
        CL_EN_k with specified labels.
10:     End.
11: Else
12:     t ← t + 1;
13:     Repeat steps 3 – 10.
14: For each CL_EN_1, CL_EN_2, ..., CL_EN_k:
15:     Initialize LN_LS_u (u = $\overline{1,k}$) containing
        Lesson entities for each cluster.
16: Create set $LN\_LP = LN\_LS\_1 \cup \ldots \cup LN\_LS\_u$
    containing required lessons.
```

The following algorithm 3 describes in detail the idea of this task. Furthermore, in accordance with the example shown in Figure 6, the nodes on the WG will be split into two clusters, *C_LS_1* and *C_LS_2*, following the completion of algorithm 3. Next, entities of the type of Lesson will be chosen from each cluster to create the appropriate *LN_LS_1* and *LN_LS_2* additional clusters. Finally, we will combine the entities in the aforementioned two clusters and rearrange them in the order of rising PR_score values to construct an initial LP, known as the *LN_LP* set, which will include the key lessons to be learned from the current level to the target level.

$$LI(i, lb) = \sum_{j \in N^l i} \frac{PR(j)}{d(j)} \qquad (2)$$

$$c_i = \underbrace{argmax\ LI(i, lb)}_{lb \in l\_max} \qquad (3)$$

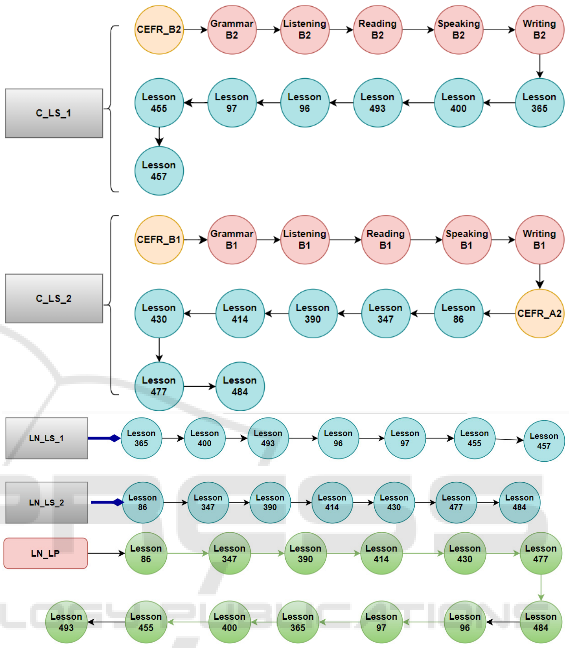

Figure 6: The process of building the LN_LP set in step 2.

4.2.3 Building a MOO Function Using the MWS Method

Not only should the PLP that is recommended to learners be a collection of lessons that are taught in a sequential manner and cover the competencies that the learner needs to master, but it should also contain a number of prerequisite lessons, or lessons that must be studied prior to studying the main lessons that are directly taught to achieve competencies. The current solution will be to provide an LP so that learners only need to learn a minimal amount of vocabulary, covering as many required main lessons as possible, as there is currently no specific statistical report on the amount of vocabulary required to be learned at each level of English certification exams.

In light of these remarks, in this step, our solution will develop an evaluation function based on the MWS method (N. Gunantara, 2018) to assess each *LN_LP's* optimization objectives in each iteration.

We set parameter *m* as a fixed number of consecutive prerequisite lessons taken from the *PRE_LS* set and then added to the existing *LN_LP* in each iteration. The proposed weights for each objective to be optimized in the evaluation functions are described in Table 2. Finally, based on the MWS formula, utilizing information from the weight set and value function for each objective (refer to table 2), let *x* represent the existing *LN_LP* in the k^{th} iteration. The MWS formula for the *LN_LP* evaluating function is expressed as in Eq. 4., which states that the LP with the lowest overall optimization score for all objectives will be deemed to be the most optimum LP when each LP in each loop has four goals that need to be optimized and each goal has a weight indicating the attached priority. The implementation procedure of step 3 is shown in Algorithm 4, and the phases are illustrated in Fig. 7 to illustrate how they are carried out. Specifically, at every k^{th} iteration, we will progressively add one required lesson to the LP that existed in the *(k-1)* th iteration. Concurrently, we utilize Eq. 4 to determine the evaluated score for each LP. In the end, only the LP found in the fifth loop will be chosen as the PLP to recommend to the learner since it fulfills Eq. 4.

Table 2: The weights and value functions of objectives.

Weight	Function	Meaning
w_1	$f_1(x)$	Maximize the number of competency entity types present in the LN_LP set.
w_2	$f_2(x)$	Minimize the number of prerequisite lesson entities (which are vocabulary or pronunciation lessons) learned enough for the required lessons in LN_LP.
w_3	$f_3(x)$	Minimize the inverse sum of the PageRank (PR) scores of lessons in LN_LP.
w_4	$f_4(x)$	Minimize the number of lessons left over in LN_LP after being evaluated.

Algorithm 4: Building the completed LP as PLP.

```
Input: WG, LN_LP, EN_IPT, m
Output: LN_LP.
    1: Initialize PRE_LS = ∅.
    2: For each node u in WG:
    3:    If ((u == Vocabulary entity || u == Pronunciation
          entity) && u ∈ EN_IPT:
    4:          PRE_LS ← PRE_LS ∪ {u}.
    5: Initialize LN_LP_L = {LN_LP}.
    6: While (PRE_LS ≠ ∅)
    7:      Last_LN_LP = GetLastElement (LN_LP_L).
    8:      Add m Lessons entities category from
            LN_PRE_LS to Last_LN_LP.
    9:      Calculate Evaluation Score for Last_LN_LP
            using MWS with Eq.4.
   10:      Add Last_LN_LP to LN_LP_L.
   11:      |PRE_LS| = |PRE_LS| - m.
   12: Select the best LN_LP from LN_LP_L based on
          optimal evaluation score.
```

$$minF(x) = \sum_{i=1}^{4} f_i(x).w_i \mid \begin{cases} \sum_{i=1}^{4} w_i = 1 \\ w_3 > 0 \\ 0 \leq w_i \leq 1 \; với \; i = \overline{1,4} \end{cases} \quad (4)$$

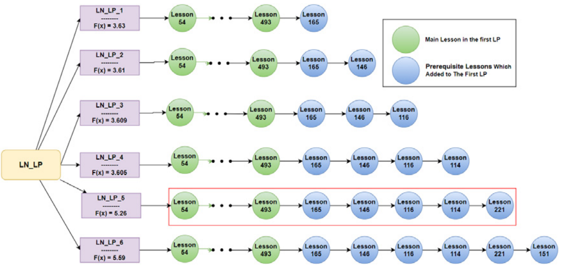

Figure 7: Development of the complete LN_LP in step 3.

5 EXPERIMENTATION AND EVALUATION

We conducted the experimentation process by considering the learner's aspiration to advance from the lowest current level and all proficiency levels of the learner's knowledge and skills, which are not yet there, to the highest target level (aligned with the CEFR competency framework: 'TOEIC (L-R)-C1', 'TOEIC (S-W)-C1', IELTS-C2', 'TOEFL-C2'). Specifically, we focused on step 3 of the solution, varying the chosen weight sets and adding a fixed number of consecutive prerequisite lessons *(m = 5)* to the LP in each k^{th} iteration. Using Eq. 4 and the completed dataset, we identified the optimal weight set for this step. We then compare approaches using the combined PageRank algorithm with the traditional LPA algorithm (PR_LPA for short), and the approach used in solution development applies the PageRank algorithm combined with the LPA_NI method (PR+LPA_NI for short) to assess the performance, stability, accuracy, and efficiency of our proposed solution (M. Abed et al., 2023) (Nabizadeh et al., 2020). The identical experimental dataset and weight set that were established following the experiment will be used for this comparison.

5.1 Dataset Building

There is currently hardly any standardized dataset that announces the learning content and skills required for

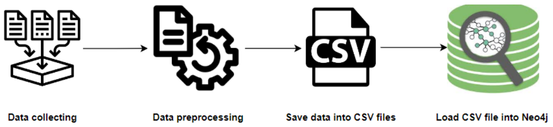

Figure 8: The process of building experimental dataset.

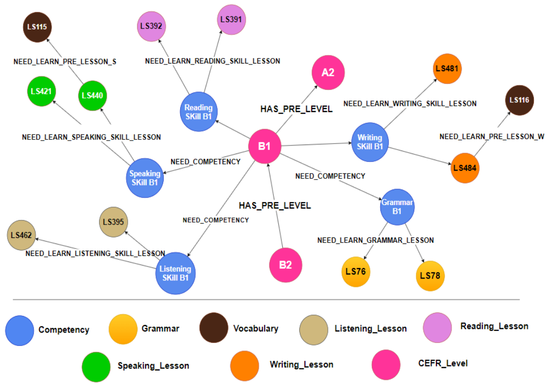

Figure 9: A part of the nodes and their relations in KG.

Table 3: Statistics on the number of entities and relationships in the KG.

Entities	Amount	Relations	Amount
		NEED_KNOWN / HAD_KNOW	63
CEFR_Level	22	NEED_LEARN_GRAMMAR	98
Competency	49	NEED_LEARN_LISTEN_SKILL	102
		NEED_LEAN_READING_SKIL	68
Grammar_LS	98	NEED_LEAN_SPEAKING_SKL	115
Listening_LS	79	NEED_LEARN_WRITING_SKIL	54
Pronunciation_LS	7	NEED_LEAN_PRE_LESSON_L	103
Reading_LS	54	NEED_LEARN_PRE_LESSON	59
Speaking_LS	91	NEED_LEAN_PRE_LESSON_S	66
Vocabulary_LS	124	NEED_LEAN_PRE_LESSON_W	12
Writing_LS	39	NEED_LEN_PRONUNCIATION	7

these English certificates according to each level, according to our survey conducted on various websites, official reference documents, and the organizations that organize these exams. Therefore, we followed the procedure outlined in Fig. 8 to produce a data set appropriate for the experimental and assessment phases.

Steps 1 and 2: Data Collection and Processing: We gather information on English certification exam formats, knowledge matter, and evaluation standards from the official homepage of ETS, the British Council, and some standard documents about these exams. Convert this information into English lesson units, detailing certification levels, required competencies, and specific lessons in grammar, vocabulary, pronunciation, listening, speaking, reading, and writing. Transform the collected data into entities, relationships, and properties matching the KG architecture.

Steps 3 and 4: Saving reprocessed data as a CSV file and importing it into Neo4j: Create CSV files containing entity and relationship data from step 2. Import these files into Neo4j using its import function to generate a comprehensive graph database schema aligned with the KG architecture. The number of entities and relationships in the KG architecture is presented in Table 3, and a part of the data set in the KG architecture is shown in Fig. 9. A full experimental dataset is now available on Kaggle.

5.2 Experimental Results and Evaluation Findings

Following the experimentation method outlined above, we discovered that all of the sets of weights tested indicated that the number of lessons was adequate to meet the necessary knowledge, skills, and lessons. Additionally, we discovered that the number of lessons—the number of prerequisite lessons that are redundant in the PLP recommended—remained unchanged in all four types of English qualification certificates. Simultaneously, the LP's evaluating function score tends to drop while the objectives' weights exhibit a significant value difference. This implies that when the objectives' weights are nearly equal, the best LP will be guaranteed when the optimal goals are deemed nearly equally important. Ultimately, we concluded that the set of weights *{w1 = 0.28, w2 = 0.27, w3 = 0.25, w4 = 0.2}* is the most ideal one to employ for this solution since it fits the requirements of the evaluation function as in Eq. 4.

As previously said, we compare the accuracy, efficiency, stability, and performance of the two approaches to the solution PR+LPA_NI and PR+LPA to assess our solution. Effectiveness is illustrated by presenting a PLP with scores from the evaluating function that conforms to the requirements in Eq. 4 and has the lowest score. Accuracy is determined by the number of lessons in PLPs that are sufficient for the number of Competency types required, and the two values of this quantity must be smaller or equivalent to the number of entities in the Lesson and Competency classes in the original KG architecture. The constancy of the PLP output across several runs with the same input data is known as stability, and the suggested PLP is used to assess the performance.

When it comes to efficiency, Fig. 10 demonstrates that the PLP's evaluating score recommended by the solution for implementation in the PR+LPA_NI or PR+LPA approach consistently satisfies Eq. 4, yet the PR+LPA_NI approach almost produces the optimal

Table 4: The percentage of Lesson entities that meet the Competency entities needed to learn in the recommended PLP.

Type of Recommendation (1)	**Number of Competency entities required (2)**	**Number of Lesson entities to learn (3)**	**Number of Lesson entities in PLP (4)**	**Number of Competency entities learned in PLP (5)**	**Competency entity rate is met** **(6) = (5)/ (2)**	**Lesson entity rate is met** **(7) = (4)/ (3)**
TOEIC L-R (A1-C1)	12	198	161	11	91,67%	81,31%
TOEIC S-W (A1-C1)	12	161	124	11	91,67%	77,02%
IELTS (A1-C2)	20	309	272	19	95%	88,03%
TOEFL (A1 – C2)	16	148	111	15	93,76%	75%

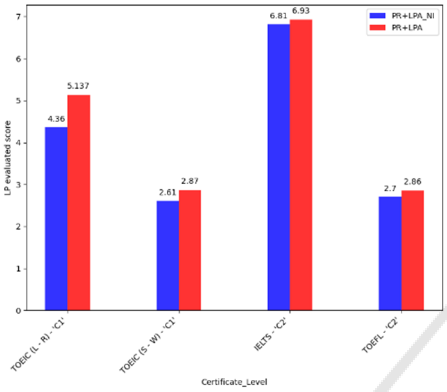

Figure 10: Evaluation scores of the recommend PLP on two algorithms.

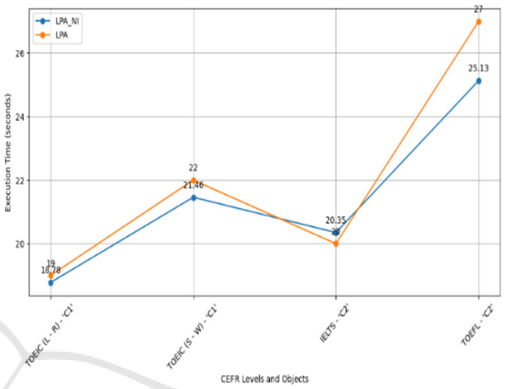

Figure 12: Results when executed on two algorithms in multiple executions.

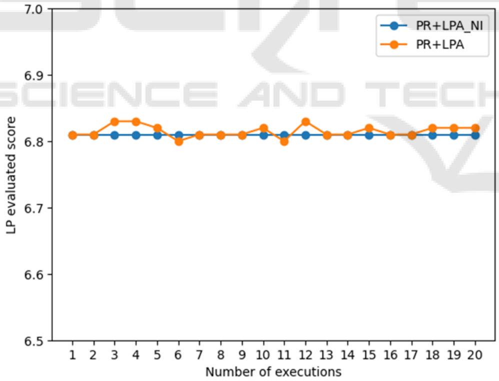

Figure 11: Execution time when executing on two algorithms for making PLP in "IELTS - C2".

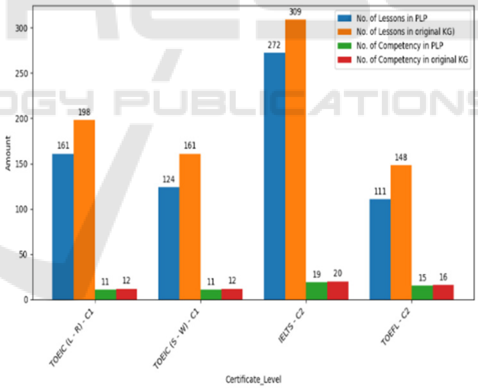

Figure 13: Number of lessons and competencies in PLP of PR+LPA_NI solution.

PLP evaluating score when compared to PR+LPA. Simultaneously, we examine the example with input data for the "IELTS-C2" certificate to recommend PLP to learners, as illustrated in Fig. 11. The solution implemented using the PR+LPA_NI approach yields the PLP evaluation function score nearly unchanged through multiple executions with the same input data in comparison to the PR+LPA approach, and the outcomes are also comparable when applied to other certificate types. Moreover, Fig. 12 indicates that PLP, as recommended by the PR+LPA_NI technique, has an approximately faster execution time than PR+LPA. Finally, when considering accuracy, based on Fig. 13 and Table 4, the solution proposed when developed in the direction of PR+LPA_NI or the PR+LPA approach all recommends being PLP for the proportion of Lesson entities almost learning enough for the required Competency entities, and the number of entities is smaller than the number of original KG.

Overall, the solution developed as a PR+LPA_NI approach route better met assessment requirements than the PR+LPA approach.

6 CONCLUSIONS

Our work developed a comprehensive solution for recommending PLPs in the English learning domain. First, we designed a KG architecture to represent key concept layers and their relationships for learning resources in international English certifications. Next, we utilized GAs and objective optimization techniques to generate the most suitable personalized learning paths. Through rigorous assessment and testing, our solution has proven to effectively generate PLPs that meet established evaluation standards and align with learners' consultation needs. To assist learners in completing their learning program as quickly and effectively as possible, future research will concentrate on developing an adaptive LP recommendation system (I. Katsaris, 2021) that modifies the original PLP in real-time after a predetermined amount of time by improving algorithms or technical processes for processing learners' learning progress data.

ACKNOWLEDGEMENTS

This research is partially supported by research funding from the Faculty of Information Technology, University of Science, VNU-HCM, Vietnam.

This research is funded by University of Science, VNU-HCM, Vietnam under grant number CNTT 2023-15.

This research is partially funded by the Vingroup Innovation Foundation (VINIF) under the grant number VINIF.2021.JM01.N2.

REFERENCES

Guo et al. (2020). A survey on knowledge graph-based recommender systems. IEEE Trans. Knowl. Data Eng., 34(8), 3549-3568.

Mansouri, N., Soui, M., & Abed, M. (2023, Sept). Full Personalized Learning Path Recommendation: A Literature Review. In AISI (pp. 185-195). Springer.

Ilyosovna, N. A. (2020). The importance of English language. *International Journal on Orange Technologies*, *2*(1), 22-24.

Dyvik, E. (2023). The most spoken languages worldwide 2023. *Statista. Retrieved.*

Fan, X., Liu, K., Wang, X., & Yu, J. (2023). Exploring mobile apps in English learning. *Journal of Education, Humanities and Social Sciences, 8*, 2367-2374.

Hodler, A. E., & Needham, M. (2022). Graph data science using Neo4j. In *Massive Graph Analytics* (pp. 433-457). Chapman and Hall/CRC.

Shi, D., Wang, T., Xing, H., & Xu, H. (2020). A learning path recommendation model based on a multidimensional knowledge graph. *Knowledge-Based Systems, 195*, 105618.

Huang, X. (2011). Study of personalized E-learning system based on knowledge structural graph. *Procedia Engineering*, *15*, 3366-3370.

Zhang, X., Liu, S., & Wang, H. (2023). Personalized learning path recommendation based on knowledge graph and graph convolutional network. *Int. J. Software Eng. Knowl. Eng.*, 33(01), 109-131.

Yin, H., Sun, Z., Sun, Y., & Huang, G. (2021). Automatic learning path recommendation for open-source projects using deep learning on knowledge graphs. In *2021 IEEE 45th Annual COMPSAC*, pp. 824-833.

Sun, Y., Liang, J., & Niu, P. (2021). Personalized recommendation of English learning based on knowledge graph and graph convolutional network. In ICAI Security (pp. 157-166). Springer.

Sun, F., Yu, M., Zhang, X., & Chang, T. W. (2020). A vocabulary recommendation system based on knowledge graph for Chinese learning. In 2020 IEEE 20th ICALT (pp. 210-212).

Chen, H., Yin, C., Fan, X., Qiao, L., Rong, W., & Zhang, X. (2021). Learning path recommendation for MOOC platforms based on a knowledge graph. In *KSEM 2021* (Vol. 14, pp. 600-611). Springer.

Yan, Z., Hongle, D., Lin, Z., & Jianhua, Z. (2023). Personalization exercise recommendation framework based on knowledge concept graph. *Computer Science & Information Systems, 20*(2).

Huang, S., Cheng, J., & Wu, H. (2014). Temporal graph traversals: Definitions, algorithms, and applications. *arXiv preprint arXiv:1401.1919*.

Turan, E., Arslan, E., Tülü, Ç., & Orhan, U. (2020). A comparison of graph centrality algorithms for semantic distance. *Lapseki Meslek Yüksekokulu Uygulamalı Araştırmalar Dergisi, 1*(2), 61–70.

Zhang, X. K., Ren, J., Song, C., Jia, J., & Zhang, Q. (2017). Label propagation algorithm for community detection. *Physics Letters A, 381*(33), 2691-2698.

Gunantara, N. (2018). A review of multi-objective optimization: Methods and its applications. Cogent Engineering, 5(1), 1502242.

Čížková, K. (2022). Comparing two community detection algorithms and their applications on human brains.

North, B., & Piccardo, E. (2019). Developing new CEFR descriptor scales and expanding the existing ones. *Zeitschrift Fremdsprachenforschung, 30*(2), 142-160.

Katsaris, I., & Vidakis, N. (2021). Adaptive e-learning systems through learning styles: A review of the literature. *Advances in Mobile Learning Educational Research, 1*(2), 124-145.

Nabizadeh, A. H., Leal, J. P., Rafsanjani, H. N., & Shah, R. R. (2020). Learning path personalization and recommendation methods: A survey of the state-of-the-art. *Expert Systems with Applications*, *159*, 113596.