Classification and Informatization of Subject Resources for English Language Teaching in the Context of Data Theory

Bin Yin†

1. English Department, School of Foreign Languages, Hengyang Normal University, Hengyang, Hunan, 421002, China.

Abstract

The construction of information resources based on the development needs of disciplines is the fundamental guarantee for carrying out disciplinary service work. An English teaching resource guarantee system is created by combining the OAI protocol with Open URL technology in this paper. Furthermore, the VIPS algorithm is used for style tree construction, combined with split bar weights and mixed text density for extracting web page body information. Then, the weight factor is introduced to improve the PageRank algorithm, and the critical resource selection strategy is combined to achieve the mining and classification of English teaching resources. In addition, the study introduces a knowledge graph for entity recognition and combines graph convolutional networks with user preference weights for optimizing teaching resource recommendations. Based on this basis, the performance of the constructed English teaching resources assurance system is verified for webpage information extraction, teaching resources classification, and personalized recommendations. It is found that the correct rate of English webpage information extraction is 96.42% when the text density of the preceding and following text is 0.334 and 0.527, respectively, and the personalized recommendation performance of English teaching resources is 4.12% higher than the best-performing KGCN model in terms of AUC index. The English teaching resources guarantee system can effectively ensure the precise classification of teaching resources in the English language and provide new guidance for furthering the automation of English teaching resources.

Keywords: VIPS algorithm; Segmentation bar weights; PageRank algorithm; Knowledge graph; User preference weights; English teaching resources.

AMS 2010 codes: 97P10
1 Introduction

With the in-depth application of modern information technology in education and teaching, information technology teaching has become the primary medium to help achieve the goal of modernization of education. The effective classification of subject resources is the central core of information technology teaching, which needs to be comprehensively and scientifically constructed and systematically and in-depth applied in subject teaching [1-3]. Relying on the teaching platform constructed by information technology, the sharing of subject-teaching resources can be realized through interconnection so as to enhance the high quality and efficient development of subject teaching and learning [4-5]. English subject resources, as a kind of existence, are used to create an English nurturing atmosphere, reflecting their spirit, English education philosophy, and other values, reflecting the culture and historical mission of the subject of the organic component of the university resources [6-7]. It can not only promote the realization of the goals of the school system facilitate the performance of teachers, but also enhance students’ satisfaction with subject education, open up students’ innovative thinking and awareness, and promote the achievement of students’ knowledge goals [8].

An essential part of the resources of English subjects is English curriculum resources, which is a strong guarantee for improving the quality of education and teaching. English curriculum resources are the necessary conditions for the implementation of the elements of the English curriculum and the English curriculum, which include knowledge, experience, skills, ways and means of activities, values, attitudes, as well as the objectives of cultivation, etc. [9-11]. The conditions for curriculum implementation include human, material, and financial resources, time, space, equipment, media, and awareness of the curriculum [12-13]. Relying on information technology to carry out the classification of English subject resources can help users access relevant English learning resources more directly in the teaching platform, promote the user’s understanding of the degree of knowledge of the English subject, and promote the high-quality development of the English subject in colleges and universities [14-15].

For the informatization construction of English teaching, literature [16] introduces the technical support for English learning based on the RBF neural network model, synthesizes a variety of algorithms to improve the RBF neural network, realizes the construction of the English learning platform, and uses the platform to better assist the students in the English learning process, uses diversified teaching means to promote the enhancement of the student’s independent inquiry ability, and also effectively enhances the students’ information technology application ability. Literature [17] believes that the development of education informatization is a significant change brought about by the development of the information age, and combining university English teaching with information technology can promote the improvement of the English informatization teaching system, which effectively promotes the improvement of the quality of university English teaching. However, there is a corresponding ecological imbalance in English informatization teaching, which leads to the limitation of English informatization development, and it needs to be optimized to enhance the development of English informatization. Literature [18] analyzed the role of a wireless communication microprocessor combined with a virtual environment applied in English listening teaching, integrated VR teaching with English teaching, and designed a listening teaching experiment. It was found that the students of VR informational scenario teaching were able to improve their performance in vocabulary judgment and conversation completion, and also effectively enhanced their interest in English learning and the effect of English teaching. Literature [19] utilizes Internet technology for the management and distribution of network teaching resources, realizes the full integration of English teaching resources, classifies English teaching resources through the TF-IDF weighting method and KNN algorithm, promotes a more scientific and reasonable distribution of teaching resources in English disciplines in colleges and universities, and provides guidance for the
storage and sharing of English teaching resources. Literature [20] constructs a multimodal English teaching mode based on the information technology of Internet+, classifies resources from multiple modalities such as text and images of English teaching resources, and carries out multimodal informatization teaching of English in combination with the MOOC platform. Relying on informatization teaching, it provides a new reference for the innovative development of English teaching in colleges and universities and promotes the effective classification of English subject resources.

In addition, literature [21] explored the design and realization method of a project-based English learning system, which enhances students’ subjective initiative in English learning through project-based learning, fully utilizes computer information technology to enhance student’s ability to analyze and solve problems, and self-exploration, and provides help to strengthen the level of English informatization through this system. Literature [22] established a university English online learning system with the B/S structure of information technology, through which the system realized the innovation of students’ English learning mode and effectively enhanced the efficiency of English teaching in colleges and universities, and also promoted the reform of English informatization teaching in colleges and universities. Literature [23] established an English culture teaching system by combining real-time video image processing technology and chose the method based on an octree structure to realize the development of informatization of English culture teaching, which provides support for improving the teaching effect of English culture in colleges and universities. Literature [24] believes that the purpose of teaching English courses in colleges and universities is to enhance students’ comprehensive application ability of English, and a university English informatization teaching system has been established by combining network technology, and the use of this system can provide support and services for university English teaching. Combined with the English teaching system can optimize the English informatization teaching method and promote the improvement of students’ intercultural communication ability and independent learning ability. Literature [25] combines artificial intelligence technology with English teaching to construct an English auxiliary teaching system in which only diagnosis of English teaching can be carried out, and the data of students in the process of English informative learning can be fully explored. From the results, the English informatization teaching system can enhance the students’ English performance and improve the student’s English independent learning ability. Literature [26] analyzes the role of modern multimedia technology combined with information technology applied in the integration of English information resources. It uses a fuzzy clustering algorithm to effectively classify the educational information resources of the English discipline, which can be combined with the disciplinary information resources to promote the enhancement of the quality of university English teaching.

This paper first establishes a distributed storage structure model of digital learning resources for English teaching based on the technical standards of education informatization with the technical support of OAI protocol and Open URL and designs the general framework of the English teaching resources guarantee system. Secondly, for the collection and mining function of English teaching resources, this paper utilizes the VIPS algorithm for style tree construction and segmentation bar weights and mixed text density for English web page body information extraction. The PageRank link analysis algorithm is improved by using the weight factor and combined with the critical resource selection strategy to mine English teaching resources in order to obtain more authoritative and relevant English teaching resources. The recognition of named entities in English teaching resources is carried out using knowledge graphs, and personalized recommendations are made by combining graph convolutional networks and user preferences. Finally, this paper carries out a validation analysis of webpage information extraction, teaching resource classification, and personalized resource recommendation through data to assess the effectiveness of these methods.
2 English Teaching Resource Guarantee System

Integration of disciplinary information resources is to follow certain principles, norms, and standards in accordance with the characteristics of the information needs of a particular disciplinary field, to centralize the scattered information resources of various channels, carriers, and media, to turn disordered resources into orderly resources, and to form a more efficient information resource system that is convenient for users to retrieve and use. The multidisciplinary and interdisciplinary nature of English studies determines the uniqueness of English teaching information resource integration. Relying on information technology to carry out English teaching resource integration, its purpose is to better realize the precise classification of English teaching disciplinary resources, to provide a more accurate resource supply for English teaching, and to promote the joint construction and sharing of English teaching resources.

2.1 OAI Protocol and Open URL Technology

1) OAI protocol operation framework

The efficient dissemination of bibliographies is made possible by OAI technology, which has its current main application in interactive search information systems. The purpose of the protocol is to provide and facilitate application-independent interactions between multiple groups of content publishers on the interconnected network, enabling access to “deep network” resources through OAI. The OAI protocol consists of two main aspects of interactions: the publishing metadata party and the service provider. The Publishing Metadata Party owns the information repository and publishes the data so that end-users or service providers can browse it. In order to construct additional services, the service provider sends requests to the publishing metadata party and receives the returned metadata.

2) Open URL Framework

Open URL is a context-sensitive open linking framework that separates the information service provider from the information provider and realizes unified retrieval of multiple databases or information resources from different organizations at the same time.

Open URL overcomes the limitations of traditional linking frameworks to provide users with context-sensitive link delivery services. Open URL is a Web-based standard for delivering metadata using a specified syntax to create a Web delivery package for metadata or an identifier of an information object. Open URL consists of three parts, namely the base URL, the source identifier, and the object metadata area. In the Open URL approach, instead of generating a direct connection to the reference target, the party referencing the other resource generates an HTTP request through a hookup point, which is the Open URL. The Open URL submits the context object to a third-party linking server using the HTTP protocol’s get/post methods. The linking server accepts the link server receives the Open URL request and dynamically calculates the target of the link based on the context object.

2.2 Classification of integration of digital learning resources

In a ubiquitous learning environment, digital teaching and digital learning are a unified concept, referring to the general term for activities on both the teaching and learning sides that are conducted at any time and any place using smart devices. However, in the field of education, the teaching mode of the transfer-receiving paradigm will continue to exist. Digital teaching and digital learning, on the other hand, refer specifically to traditional classroom teaching and extracurricular learning, which belong to two kinds of teaching and learning behaviors of teachers and students, respectively, and are distinctly different. Based on the technical standards of education informatization, this paper
establishes a distributed storage structure model of digital learning resources for English teaching, as shown in Figure 1, which divides English teaching resources into five categories: subject knowledge ontology resources, process resources, social cognitive network resources, teaching resources and learning resources.

*Figure 1. Digital learning resource classification model*

In this model, the dynamics, evolution, complexity, openness, and contextualization of ELT learning resources are fully reflected. The model stores the process-based information of ELT resources on local cloud servers, which is not a structural part of the metadata of learning resources - “learning elements” and fully reflects the characteristics of the process-based resources such as the survival cycle, social context, openness to evolution, quality control, and so on. The learning element or metadata is the smallest unit for constructing subject knowledge ontology resources, which is no longer attached to some process-based information and is only generated when the knowledge ontology resource interacts with the social cognitive network, which fully guarantees the granularity of reuse or division of “learning element.”
2.3 Framework of the English Teaching Resource Assurance System (ETRAS)

The framework of the English teaching resources guarantee system constructed in this paper is shown in Fig. 2, which supports the retrieval of static resources browsing, and publishing of dynamic resources. Also, it provides personalized services for users so that it can better guarantee the construction of resources for English teaching disciplines in colleges and universities.

![Figure 2. English teaching resource protection system](image)

The framework of the assurance system based on the integration of English teaching resources can be divided into a unified retrieval module, a dynamic resource browsing and publishing module, a personalized service module, and other service modules.

1) Unified Search Module

This module is the most critical in the resource guarantee system for key disciplines of English teaching. It provides users with a single access point to search for all static resources related to the key disciplines. The majority of the resources required for key disciplines are static resources, which fully realize effective classification and navigation of disciplinary resources and provide support for accurate retrieval of English teaching resources.

2) Dynamic Resource Publishing and Browsing Module

The dynamic resources in the construction of English teaching discipline resources come from a wide range of sources. They are collected through many channels, so it is a waste of time to rely on...
discipline librarians to collect such information, as well as not being able to collect them ultimately. By adding domain experts and ordinary users to the collection of dynamic resources in ELT disciplines, it is possible to increase the collection channels of dynamic resources and enrich the content of dynamic resources in the resource guarantee system.

3) Personalized Service Module

Personalized service mainly refers to the fact that this resource guarantee system can provide users with the research-related materials they need according to their different needs, which is primarily realized through personalized customization and RSS information push function.

4) Other Service Modules

It mainly includes the introduction of disciplines and achievements, friendly links, and feedback services. It provides users with the development status of the English discipline, helps experts in the field to designate the development strategy suitable for the English discipline, and promotes the continuous development of the English discipline in colleges and universities.

3 Collecting and Mining of English Teaching Resources

Classification of subject resources is carried out with the intrinsic regularity of knowledge relationship and, based on subject classification can better reflect the connection between knowledge, which is more conducive to the efficient use of subject resources. The purpose of subject resource classification and informationization construction based on the English Teaching Resource Assurance System is to enhance the personalized retrieval and mining of English teaching resources, provide users with more diverse English teaching resources, and promote the high-quality development of the English subject.

3.1 Web Page Body Information Extraction

3.1.1 Style Tree and Split Bar Weights

1) Style tree model based on VIPS algorithm

The VIPS algorithm is roughly divided into three steps, i.e., Extracting page blocks, detecting separator bars, and reconstructing semantic blocks. The algorithm is based on the DOM tree combined with visual information such as fonts, white space area, background color, and other visual information on the visual information page. It gives some rules to split the page into different visual information blocks. The algorithm utilizes the layout features of the page to form the tree of the page and then find out the eligible information blocks from the tree. Secondly, it detects these blocks of information and identifies the vertical and horizontal separator bars. Finally, based on these separator bars, the information chunks are recombined into final semantic chunks.

The VIPS algorithm defines a single web page as:

\[ \Omega = (O, \Phi, \delta) \]  \hspace{1cm} (1)
Where \( O = (VB_1, VB_2, \ldots, VB_n) \) denotes the set of visual blocks in a web page, \( \Phi = \{\varphi_1, \varphi_2, \ldots, \varphi_n\} \) denotes the set of horizontal and vertical separator bars that divide the page into blocks, and \( \delta = O \times O \to \Phi \cup \{\text{NULL}\} \) denotes the relationship of each visual block.

The VIPS algorithm sets a value in advance, for each sizeable semantic block can be segmented into smaller semantic blocks, and then each iteration of the semantic block segmented into a semantic block, a value is set for it until the value of this block reaches the value, the segmentation process ends.

2) Segmentation bar weight generation

We propose to convert the style coefficient differences between different webpage blocks into partition bars and use the weight of the partition bars as the standard for segmentation between information blocks.

In the style tree model, the visual information on the web page is ultimately added to the node. When going through the style coefficient transformation strategy, the original style node \( A \) is transformed into node \( T_\Lambda (S_{bc}, S_{bd}, S_p, S_m) \), where \( S_{bc} \) is the background color value coefficient, \( S_{bd} \) is the boundary coefficient, \( S_p \) is the inner edge coefficient and \( S_m \) is the outer edge coefficient. The above coefficients can be used to obtain the total number of styles of the current node information block \( T \). i.e.:

\[
T = (S_{bd} + S_m + S_p) \times S_{bc}
\]  

(2)

The separator bar is between two different nets, and its weight can be formally represented by utilizing the difference in the total number of styles between them. Then:

\[
N_{ij} = |T_i - T_j|
\]  

(3)

\( N_{ij} \) denotes the weight of the separator bar between the \( i \) nd block node and the \( j \) rd node, and \( T_i \) denotes the total number of styles in the \( i \) th block node.

For a preliminary style tree, the initial nodes do not have a complete split, and each node in the style tree corresponds to a block of regions on the web page. A large web page block is formed by one or more child nodes in each parent node. Visually, it is possible to subdivide each large web page block into small chunks of information, which are represented by separator bars between web page blocks. Splitting operation is carried out for the nodes whose weight of the separating bars is greater than the threshold value. Semantic reconstruction is carried out after the node iteration splitting is completed, and merging operation is carried out for the nodes with partially broken semantics to finally complete the composition of the visual block tree as the data source of the body text recognition algorithm.

3.1.2 Link Density Ratio Extraction Text

Hybrid text density is used in this paper to extract information from web pages because the content information required is relatively centralized.
Firstly, the text density is calculated, which refers to the ratio of the text length corresponding to a text node on a node path and the sum of the text lengths corresponding to text nodes on all paths in the web page. The density of text can be expressed as:

\[ TD = \frac{\text{text}_i}{\sum_{i=1}^{n} \text{text}_i} \]  

(4)

Where \(\text{text}_i\) represents the number of text characters contained in the subtree represented by node \(i\), and \(\sum_{i=1}^{n} \text{text}_i\) represents the total number of text contained in the subtree represented by all nodes in the web page.

If only the text density of a web page node is taken into account, other web page contents with shorter text are often forgotten. Therefore, link density should also be considered. Link density refers to the ratio of the number of links corresponding to link nodes on a node path and the sum of the number of links corresponding to link nodes on all paths in the web page. The density of links is defined as:

\[ HD = \frac{\text{hsyperlink}_i + 1}{\sum_{i=1}^{n} \text{hsyperlink}_i + 1} \]  

(5)

Where \(\text{hsyperlink}_i\) represents the number of links contained in the subtree represented by node \(i\), and \(\sum_{i=1}^{n} \text{hsyperlink}_i\) represents the total number of links contained in the web page.

If only the text and link density of a web page are considered, the content information of some short texts will be ignored. Therefore, the link text density also needs to be considered. Link text density refers to the ratio of the number of link texts corresponding to link nodes on a node path and the sum of the number of link texts corresponding to link nodes on all paths in a web page. Link text density can be expressed in terms of:

\[ LTD = \frac{\text{ltext}_i + 1}{\sum_{i=1}^{n} \text{ltext}_i + 1} \]  

(6)

Where \(\text{ltext}_i\) represents the number of link text characters contained in the subtree represented by node \(i\), and \(\sum_{i=1}^{n} \text{ltext}_i\) represents the total number of link characters contained in the whole webpage.

The hybrid text density is used to fuse the above three densities, and the fusion is carried out through the power relation fusion algorithm, which transforms the node features in the webpage into calculable values, increases the differentiation between the body information content and the noise content, so as to better access the webpage related body teaching resources.
3.2 English Teaching Resources Mining

3.2.1 PageRank Link Analysis

The PageRank algorithm is designed to link the importance of a web page to the number of links to it and the quality of its neighboring pages. Simply put, as long as the number of links to the web page is high and the quality of the web pages linking to the web page is high, then the current web page can be considered relatively important.

The PageRank algorithm is calculated as follows, assuming that the applicable network is a directed network:

1) Assigning initial values. Assign each node in the network model the same initial $PR$ value, in order to satisfy the condition of normalization because the $PR$ value will change with the number of iterations $t$, so when $t=0$, you can initialize $PR(0)$ to $\frac{1}{N}$, where $N$ is the total number of nodes in the network, that is, the initialization satisfies the following conditions:

$$\begin{align*}
PR_i(0) &= \frac{1}{N} \\
\sum_{i=1}^{N} PR_i(0) &= 1
\end{align*}$$

(7)

2) Iterative update. In each step of the update, the $PR$ value of the current node can be considered as the influence that can be propagated and then distributed equally to the nodes pointed to by the node. Assuming that the current node $i$ has $n_{iout}$ outgoing links, the value passed by the current node to the node pointed to by the current node is $\frac{PR(k-1)}{n_{iout}}$, where $k$ is the current number of iterations. At the same time, the current node receives the PR value passed by the lead node. It is clear that if there is an isolated node in the network, the isolated node will not update the $PR$ value, while the $PR$ value of the non-isolated node will be updated in the current iteration. The updated policy is as follows:

$$PR_i(k) = \sum_{j=1}^{N} a_{ji} \frac{PR_j(k-1)}{k_{jout}}, i=1,2,\ldots,N$$

(8)

Where $a_{ji}$ is an element of the adjacency matrix $A=\left(a_{ji}\right)_{N \times N}$ in the network, and the original definition of $a_{ji}$ in a directed unweighted network is as follows:

$$a_{ji} = \begin{cases} 1 & \text{There exists an edge pointing from point } j \text{ to point } i \\ 0 & \text{There is no edge pointing from point } j \text{ to point } i \end{cases}$$

(9)

It follows that the state transfer matrix $M$ can be derived from the adjacency matrix and the elements in matrix $M$ are defined as follows:
$$m_{ji} = \begin{cases} 
1/k^\text{out}_{ji} & \text{There exists an edge pointing from point } j \text{ to point } i \\
0 & \text{There is no edge pointing from point } j \text{ to point } i 
\end{cases} \quad (10)$$

The probability distribution obtained by the random walk model at a certain iteration $$t$$ through each node in the network is equivalent to the state distribution of the Markov chain at moment $$t$$, which can be represented by a $$N$$-dimensional vector $$\mathbf{PR}_t$$, and thus the simplest matrix form of the PageRank algorithm can be obtained as follows:

$$\mathbf{PR}_{t+1} = M \cdot \mathbf{PR}_t \quad (11)$$

It follows that the state transfer matrix can be derived from the adjacency matrix, and the elements in the matrix are defined as follows: For English teaching, obtaining more authoritative teaching resources is an inevitable requirement for improving the quality of English teaching and enriching students’ English learning experience. Therefore, this paper proposes to improve the PageRank algorithm by analyzing web content and links, considering the influence of web links and web content on page ranking. Namely:

$$PR(k) = dD(k) + (1-d) \sum_{v \in B(k)} \left[ \frac{\lambda \cdot \text{LinkSim}(k,v)}{\sum_{w \in F(v)} \text{LinkSim}(v,m)} + (1-\lambda) \frac{\text{ConSim}(k,v)}{\sum_{w \in F(v)} \text{ConSim}(v,m)} \right] \quad (12)$$

in which $$k$$ is the current page, $$B(k)$$ is the set of all pointing pages $$k$$, $$v$$ belongs to any one of the pages in the set of pages $$B(k)$$, $$F(v)$$ is the set of all pages pointed to by page $$v$$, and $$w$$ belongs to any one of the pages in the set of pages $$F(v)$$. $$\text{LinkSim}(k,v)$$ is the link structure similarity between page $$k$$ and page $$v$$, $$\text{ConSim}(k,v)$$ is the content similarity between page $$k$$ and page $$v$$, $$\frac{\text{LinkSim}(k,v)}{\sum_{w \in F(v)} \text{LinkSim}(v,m)}$$ is the authoritativeness ratio of the page in the set of forward links, $$\frac{\text{ConSim}(k,v)}{\sum_{w \in F(v)} \text{ConSim}(v,m)}$$ is the relevance ratio of the page in the set of forward links, and $$\lambda$$ is the weighting factor to control the authoritativeness and relevance.

When $$\lambda = 1$$, it means that the assignment of PageRank values between pages depends only on the links to the pages and does not take into account the effect of the content of the pages on the PageRank values. When $$\lambda = 0$$, it means that the allocation of PageRank between pages depends only on the content of the pages and does not take into account the effect of the links on the PageRank values. When $$\lambda = 0.5$$, it means that the content of the web page and the links to the web page play equal roles in assigning PageRank values to the web pages.

### 3.2.2 Strategy for Selecting Essential Resources

The PageRank algorithm views the entire Internet as a directed graph and solves the problem of page ranking through link analysis. The core idea of the PageRank algorithm is that each link to a page is a vote for that page, and the more links to other pages, the more votes it receives. The same rule applies to internal links on a website. If the PageRank algorithm is applied to the internal pages of a
website, it is possible to arrange the pages within the website. The sequence shows how important each page on the site is, with higher-ranking pages being more critical and reflecting the site’s theme.

From the above two points, we can derive the strategy for selecting essential resources. Firstly, the topology of the website is processed, from which a subgraph with a tight link structure is selected, and the nodes in this subgraph should be related to the theme of the website. The PageRank algorithm is used to rank the pages of this subgraph and identify those pages that are structurally important as critical resources of the website.

3.3 Experimental results and analysis

3.3.1 Web Page Information Extraction Experiments

This section mainly analyzes the impact of the parameter adjustments involved in each link on the overall English teaching resource assurance system in the Web page body text extraction experiment, which is divided explicitly into the impact of the text link density ratio on the correct rate of information extraction as well as the analysis of the effect of the segmentation bar weight value on the proper rate of information extraction.

1) Impact of Text Link Mixing Density on the Correct Rate of Information Extraction

In order to verify and examine the actual impact of the mixed density of text links on the extraction algorithm, this paper has done the following experiments: i.e., A batch of Web pages were selected from mainstream websites. In order to ensure that the algorithm’s requirements for topic-based pages are met, we try to choose topic-based Web pages that contain textual information content. Figure 3 shows the results of the information extraction.

From the data in the figure, it can be seen that some specific thresholds have a significant impact on the topic extraction results of English teaching resources when performing information extraction from the body of the Web page, and these thresholds include the text density of the preceding text and the text density of the following text. When extracting from the English web pages for the topics of news, entertainment, finance, real estate, politics, economy, culture, etc., the overall correct extraction rate is between 94.38% and 96.42%, with an extreme difference of 2.04 percentage points, which indicates that this paper’s web page body information extraction method can obtain more complete and correct English-related teaching resources. In addition, according to the changes of different text mixing densities, when the text density of the preceding text and the text density of the following text are 0.334 and 0.527, respectively, the correct rate of web page body information extraction is optimized at 96.42%. To achieve the optimal information extraction effect, it is essential to set the text mixing density to a suitable threshold when extracting topic information from English teaching resource web pages.
In this paper, the VIPS algorithm is used to segment the body resources of English teaching resource web pages, and the segmentation bar weight values are designed to realize the correct classification of English teaching resource web pages. When encountering some exceptional cases, the classifier may classify the information that should be recognized as noise as the body information content and classify the information that should be classified as the body as noise content, which results in the impact on the extraction accuracy, so the selection of the segmentation bar weighting degree has a significant effect on the correct rate of information extraction. Therefore, this subsection is to argue that, given different weighting degrees, the impact of the classifier on the system’s extraction correctness, and at the same time, in the process of argumentation, to find and discover the most suitable weighting degree for extraction.

In this experiment, 50 random body information contents of English teaching-related news-like Web pages were sampled for extraction, and their accuracy for body text extraction was compared. The results of extracting body text using different segmentation bar weight values are shown in Figure 4.

As can be seen from the figure, in the process of extracting the body information of English teaching resources Web pages, the correct rate of content extraction will show not an increasing trend but an increasing and then decreasing trend with the gradual increase of the segmentation bar weight value. When the segmentation bar weight value increases from 1 to 5, the correct rate of English teaching resources information extraction reaches a peak of 0.96. Then, it decreases to 0.42, which indicates that the choice of segmentation bar weight value has a more significant impact on the correct rate of English teaching resources Web page body information extraction. It can also be found that the proper rate of English teaching resources body information extraction is the highest when the value of segmentation bar weight is chosen as 5.

Figure 3. The results of the text of the web page

2) Impact of Segmentation Bar Weights on the Correct Rate of Body Extraction
3.3.2 Classification of English Language Teaching Resources

In this paper, the improved PageRank algorithm is used to analyze the authority and relevance of English teaching resources so as to ensure that the informationized teaching resources obtained from the English teaching resources guarantee system can promote the development of English teaching with high quality. The method is affected by the weighting factor of relevance and authority, i.e., the proportion of webpage content and webpage links is different depending on the value of $\lambda$.

In this paper, Macro Average (Macro) is chosen as the evaluation index for the categorization of English teaching resource web pages, and different weighting factors are chosen for experiments according to the previous description of the improved PageRank algorithm, i.e. $\lambda = 0, 0.5, 1$. Using Google’s website navigation as a source of data and choosing 20 categories as the classification system for website categorization, a total of 1,365 English teaching-related resource websites were acquired. According to the principle of two-eight, 80% is selected as the training set, and the remaining 20% is selected as the test set. Fig. 5 shows the change in the accuracy rate of English teaching resource websites under different weighting factors, where Figs. 5(a) and (b) show the comparison results of MacroP and MacroF1, respectively.

Based on the data distribution in the figure, it can be seen that when the value of the weighting factor is taken as 0.5, the mean values of MacroP and MacroF1 for the classification of English teaching resource web pages are 0.859 and 0.901, respectively, which are better than the MacroP and MacroF1 obtained from the weighting factor 0 and 1. It shows that when both web content and web links are considered for their influence on the PageRank value of web pages, the macro mean value achieves the optimal value, and only considering web links or only considering web content cannot effectively realize the accurate classification of English teaching resources web pages.
4 Personalized Recommendation of English Teaching Resources

The purpose of classifying subject teaching resources is to make the organization of information resources more regular, improve the use of resources, and reduce the time cost of user retrieval. The construction of a subject teaching resources security system and the establishment of a subject resources classification system based on it will significantly improve the current status quo of vague categorization and inconsistent standards of digital subject teaching resources. Through the comprehensive collection of similar resources, it is of great practical significance to carry out proactive, personalized, and targeted disciplinary services and to improve the guarantee rate and utilization rate of teaching resources in critical disciplines.

4.1 Resource Recommendations for Knowledge Graphs

4.1.1 Named Entity Knowledge Reasoning

Reasoning about named entities in knowledge graphs involves obtaining new inter-entity associations that meet the semantic requirements. The purpose of entity-relationship mapping in a knowledge graph is to vectorize the representation of entity and relationship mapping so as to simplify the computation and inference of knowledge graph relationships. The general steps are to first carry out the representation of entities and relationships in the graph, then to define a loss function, and finally to carry out the vector representation of the learned entities and relationships.

For each ternary instance \((h,r,t)\) in the dataset, the relation \(r\) in it is considered as a connection vector from the head entity \(h\) to the tail entity \(t\), and the values of \(h\), \(r\), and \(t\) are adjusted by training to make \((h+r)\) as close to \(t\) as possible.

Here, we are given a \(f(h,r,t)\) and denote the relation translation as a vector \(\vec{r}\). This can be done by picking up the vector \(\vec{h},\vec{r}\) of entities, i.e., entity \(\vec{h}\) is obtained through relation \(\vec{r}\), then:

\[
\vec{h} + \vec{r} \approx \vec{t}
\]

We, therefore, define the distance formula \(f(h,r,t)\) for the error between \(\approx\) two side vectors, then:
\[ f(h,r,t) = \| \tilde{h} + r - \tilde{t} \|_2 \]  \hspace{1cm} (14)

If ternary \((h,r,t)\) exists in the dataset, then \( f(h,r,t) \) is smaller.

### 4.1.2 Calculation of user preference weights

In this paper, a knowledge graph \( G \) consists of entity-relationship-entities \((h,r,t)\), where \( h \) and \( t \) denote the head entity and the tail entity of a triad in the knowledge graph, respectively, and \( r \) denotes the relationship between them. Where \( h \in E, r \in R, t \in E, E \) represents the set of entities in the knowledge graph and \( R \) represents the set of relationships in the knowledge graph. \( U = \{ u_1, u_2, \ldots, u_a \} \) represents the set of users, \( V = \{ v_1, v_2, \ldots, v_b \} \) represents the set of items, \( a \) and \( b \) are the number of users and items, respectively, and the user-item interaction. The matrix can be represented as \( Y = \{ y_{uv}, u \in U, v \in V \} \). Where:

\[
Y = \begin{cases} y_{uv} = 1 & \text{denotes that user } u \text{ and item } v \text{ have interaction} \\ y_{uv} = 0 & \text{denotes no interaction between user } u \text{ and item } v \end{cases}
\hspace{1cm} (15)

The problem to be solved for personalized recommendation in conjunction with knowledge graphs can be described as:

\text{Probability of interaction between user } u \text{ and item } v \text{ with which he has not interacted in his history. The probability of a user interacting with an item is obtained through equation (16). I.e.:

\[
y'_{uv} = F(u,v|\theta,Y,G)
\hspace{1cm} (16)
\]

Where \( y'_{uv} \) is the probability of interaction between user \( u \) and item \( v \), and \( \theta \) is a model parameter of function \( F \).

The user’s preference for relationships is used to judge the degree of influence of neighboring entities on the current entity. The attention mechanism calculates the user preference to calculate the weights of the relationships in the knowledge graph on the user, i.e., The degree of choice of the user for different relationships. The edges connected by the head and tail nodes in the knowledge graph are called weights. Then:

\[
S^u_r = f(u,r)
\hspace{1cm} (17)
\]

Where \( u,r \) is a \( d \)-dimensional vector representation of the user and the relationship, and \( r \) represents the relationship vector between two nodes in the knowledge graph. The function \( f(\cdot) \) represents an inner product operation on the two vectors, i.e., The user’s preference level for the relationship can be obtained. As with all attention mechanisms, the weights are normalized by a normalization operation. Thus:

\[
W_{u,(h,h_i)} = softmax\left( f(u,r_{(h,h_i)}) \right) = \frac{\exp\left( f\left( u,r_{(h,h_i)} \right) \right)}{\sum_{i \in M(h)} \exp\left( f\left( u,r_{(h,h_i)} \right) \right)}
\hspace{1cm} (18)
\]
Where \( r_{(h_i,h_j)} \) denotes the relationship between both entities \( h_i \) and \( h_j \) in the knowledge graph, \( f\left(u, r_{(h_i,h_j)}\right) \) denotes the degree of preference of user \( u \) for relationship \( r_{(h_i,h_j)} \), and \( M(h_i) \) denotes the set of first-order neighbors of node \( h_i \).

Combining the knowledge graph with graph convolutional networks, graph convolutional networks can handle non-Euclidean structures, i.e., Graph-structured data is very well-structured. Graph-structured data takes into account both node information and structural information between nodes, and graph convolutional networks can acquire both the features of nodes and the structural information between nodes. To generate new node representations, message propagation, and aggregation operations are carried out on nodes using the information of the edges in the graph, and user preference weights are taken into account to recommend personalized teaching resources.

4.2 Resource Personalization Recommendation Performance

4.2.1 Personalized Recommendation Cold Start Analysis

In order to verify the performance of this paper’s personalized recommendation algorithm for English teaching resources based on Knowledge Graph and Graph Convolutional Network, this experiment compares this paper’s method with SVD, LibFM, CKE, and KGCN baseline models. The data applied in this experiment comes from the dataset made from the relevant data extracted and mined through the web page body information in the previous paper, which contains 16,574 English teaching resources and is set as the training set and test set according to the ratio of 8:2. The AUC value is chosen as the evaluation index to analyze the effectiveness of the knowledge graph in mitigating the cold-start in making personalized recommendation. The performance of various methods in cold-start scenarios is shown in Figure 6.

From different methods in the cold-start personalized recommendation, when the size of the training set is gradually reduced from 100% to 5%, the AUC values of the baseline models of SVD, LibFM, CKE, and KGCN decrease by 2.14%, 3.11%, 4.51%, and 5.57%, respectively, whereas that of this paper’s method decreases by only 0.95%. This shows that in the cold-start scenario of personalized recommendation of English teaching resources, combining a knowledge graph with a graph convolutional network can obtain a better resource recommendation effect, which provides support for improving the diversity of English teaching resources for users.

![Figure 6. Performance of different methods in cold start scenarios](image_url)
4.2.2 Personalized Resource Top-K Recommendation

In order to further analyze the performance of the method given in this paper in performing personalized recommendation of English teaching resources, in the experiments, the data is divided into a training set and test set in the ratio of 8:2. Two recommendation scenarios are considered: click-through prediction and Top-K recommendation. The selected model is the same as in Section 4.2.1, and the AUC value and F1 composite score are chosen as the evaluation indexes of the model of this paper. The selected methods are used to conduct CTR experiments on the dataset. The results of personalized recommendations for English teaching resources are shown in Fig. 7, where Figs. 7(a) and (b) show the results of the CTR experiment and the Top-K recommendation curve, respectively.

The conclusions can be summarized based on the experimental results:

1) In the CTR experiment, it can be found that the performance of this paper’s method is significantly improved over other baseline models. Specifically, the AUC metric on the dataset has been enhanced by 4.12% over the state-of-the-art KGCN model. It illustrates the effectiveness of the model to take full advantage of collaborative and knowledge-aware information when using a knowledge graph for named entity recognition followed by a graph convolutional neural network for English teaching resources recommendation.

2) The performance of this paper’s method and KGCN model with the introduction of the knowledge graph is much higher than that of such models as SVD, LibFM, and CKE, which are not combined with the knowledge graph, which indicates that the introduction of knowledge graph can indeed provide more auxiliary information for the recommender system, and can effectively improve the recommendation effect.

3) In the Top-K recommendation scenario, this paper’s method performs well in the dataset. With the increase of the K value, the RECALL@K value for Top-K recommendation increases gradually, and its RECALL@K value reaches 46% when the K value is 50. To analyze the reason, the model in this paper can effectively aggregate information in sparse data scenarios, even if more noise is introduced in massive data scenarios. However, its recommendation effect still maintains a high level.

To summarize the above, the knowledge graph is used to identify the named entities of the English teaching resources in the educational resource guarantee system and then combined with the user preference and graph convolutional network to solve the first-order neighbor set. Through the information in the knowledge graph to nodes for message propagation and aggregation, so as to realize the update of the recommendation results, the user in the English teaching resources guarantee system to obtain more in line with their needs of resources.

Figure 7. Personalized recommendations for English teaching resources
5 Conclusion

In this paper, the integration and classification of digital learning resources are carried out with OAI protocol and Open URL technology, and the English teaching resources guarantee system is constructed to help users better realize the personalized recommendation of English teaching resources through the collection and mining of English teaching resources, and the effectiveness of the system is analyzed. The following are the conclusions drawn from the experimental results:

1) When extracting information from the body of English teaching resources web pages, the correct rate of information extraction reaches 96.42% when the text densities of the preceding and following texts are 0.334 and 0.527, respectively. For the classification of English teaching resources, when the weighting factor is 0.5, the mean values of MacroP and MacroF1 for the classification of English teaching resources web pages are 0.859 and 0.901, respectively. The collection and mining technology of English online teaching resources can ensure the authority and relevance of English teaching resources, better ensure that the data in the English teaching resources guarantee system is more diversified, and provide users with English teaching resources in multiple fields. Offer English teaching resources for various fields.

2) In terms of the personalized recommendation of English teaching resources, the AUC index of this paper’s method on the dataset is improved by 4.12% compared with the best-performing KGCN model, and the RECALL@K value obtained by this paper’s method reaches 46% when the K-value is 50 in Top-K recommendation. The combination of a knowledge graph and a graph convolutional network can effectively enhance the personalized recommendation of English teaching resources and promote the acquisition of personalized English resource learning by users.

In summary, the resource guarantee system constructed on the basis of information technology can provide a new way to classify and informatize the teaching resources of the English subject collect and organize various types of teaching resources of the English subject in the network. In accordance with specific standards for the integration of English teaching resources, so that users can browse and retrieve the way, access to high-quality English teaching information resources can be the network of English teaching information resources from disorder to order, from dispersion to centralization, and provide users with comprehensive and fast information resource retrieval services.

Funding:

Project Source: Hunan Province Colleges and Universities Education Teaching Reform Research Project in 2023;

Project Title: Innovation of the “3-Stage, 4-Dimensionality” Evaluation System and Development of Evaluation Scales for Vocational Techniques of English Normal Students;

Project Number: HNJG-20230903

Host: Bin Yin.
References


