Application of the LDA Model to Semantic Annotation of Web-based English Educational Resources

Wei Du[1,2], Haiyan Zhu[1,*] and Teeraporn Saeheaw[2]

[1]*School of Foreign Languages and Culture, Ningxia University, Ningxia, China, 750021*
[2]*College of Arts, Media and Technology, Chiang Mai University, Chiang Mai, Thailand, 50200*
E-mail: PAPER202103@163.com
[*]*Corresponding Author*

Received 25 March 2021; Accepted 09 April 2021;
Publication 24 June 2021

Abstract

Based on the LDA model, this paper builds a three-layer semantic model of Web English educational resources "document-topic-keyword", models the semantic topics of resource documents, and obtains the semantic topics and keywords of document resources as the semantic labels of resources. The experimental results show that document LDA topic modeling is beneficial to the macroscopic classification of Web English educational resources. The experimental results show that LDA topic modeling of documents is useful for macroscopic cataloging of Web English educational resources, highlighting teaching priorities, difficulties, and interrelationships, while LDA modeling of teaching topics with the same teaching content expands the metadata generation method of resource description based on the basic education metadata standard and provides more information about the inherent characteristics of resources. The semantic information can be used to mine the semantic thematic features and detailed differences inherent in the resources, and the final performance analysis verifies the parallel computing advantages of the LDA model in a big data environment.

Journal of Web Engineering, Vol. 20_3, 1053–1076.
doi: 10.13052/jwe1540-9589.2047

Keywords: LDA model, Web English, educational resources, semantic annotation.

1 Introduction

In the context of big data in education, new resource integration technologies and platforms are needed as well as new governance mechanisms for Web English educational resources. Not only resource construction management specifications, sharing rights, and intellectual property rights are needed, but also technologies for resource collection, storage, labeling, and sharing under the big data environment [1]. The development and application of metadata standards for Web English education resources standardize the practice of Web English education resource construction and lay a metadata description framework for Web English education resource storage and sharing [2]. Based on the resource aggregation and sharing technique of web crawler, we use the Web English educational resources' titles, keywords, hyperlinks, and other information obtained during the analysis of Web English educational resources as the markers of Web English educational resources and provide students with keyword-based search services [3].

With the development and application of Web 3.0 semantic annotation technology, students are no longer satisfied with basic metadata descriptions such as titles, linked text, and string-matching search results; they expect resource descriptions to provide potential semantic information about the topic ideas, teaching methods, and knowledge-concept relationships within the document, such as the basic methods of language teaching, the central idea of the text, and the specific knowledge points [4]. Semantic information is hidden inside resources such as teaching methods. The semantic annotation of these resources can better reflect the characteristics of the resources themselves, and the descriptions can be more accurate and easier for students to select and use [5]. When Web-based English education resources are extremely rich and growing rapidly, an automatic and intelligent method of resource annotation and aggregation is also needed.

Information labeling of Web English educational resources is the most complicated and important part of Web English educational resources construction [6, 7]. On this basis, it is stipulated in a series of domestic standards, such as the Web Specification for the Construction of English Educational Resources, that the attribute labeling of Web English educational resources includes mandatory and optional attribute data, and the mandatory raw data includes representation, title, language, description, keywords, etc. The

metadata is directly related to the content of resources and teaching applications [8]. Manual annotation has its precision, but there are also problems such as individual understanding bias, selective filtering, and a large amount of annotation work, etc. In today's extremely rich web resources, there is an urgent need to annotate metadata by automated means. With the rise of semantic annotation technology, i.e., Web 3.0, and the linear increase in the number of web resources, the intrinsic semantic analysis of document resources has become popular among students, and its research and application have become a hot topic nowadays [9]. The strategy of openly shared subject resource catalogs automatically analyzes and labels the title, text, and links of Web pages semantically, and then builds a catalog index. Based on the linked data semantic distance (LDSD) technique, the researcher proposes a resource similarity calculation method, which takes into account the attributes of the resource and the satisfaction of popular students in the calculation process, and the method has shown better performance than similar resources in DBpedia repository testing and music recommendation system. Semantic annotation technology provides technical support for automated metadata annotation of English educational resources on a large-scale web [10].

The LDA model excels in large-scale topic modeling of document resources and has a wide range of applications in document semantic analysis, topic modeling, and resource recommendation. A comprehensive collaborative resource filtering algorithm and a probabilistic topic model based on LDA are used to develop a scientific and technical document recommendation system to recommend old and new scientific and technical documents that may be of interest to students [11]. The topic model for semantic relationship constraint is used to process a large amount of product review test data to better discover fine-grained feature words, sentimental words, and semantic correlations between product features, and thus obtain product feature levels and student sentiment preferences [12]. A new approach to quantitative analysis and evaluation of social networking sites is proposed from three perspectives: student, topic, and community [13]. As shown in Figure 1, the LDA model is often used for text classification processing, resource topic modeling, resource recommendation, and so on. The CA-LDA model combines the probabilistic topic extraction method and network analysis method of the traditional LDA model and add the network analysis method to the traditional LDA model.

This study starts from the need for semantic annotation of resources in the process of building and sharing repositories, uses the LDA model to semantically model the document resources in repositories, explores potential

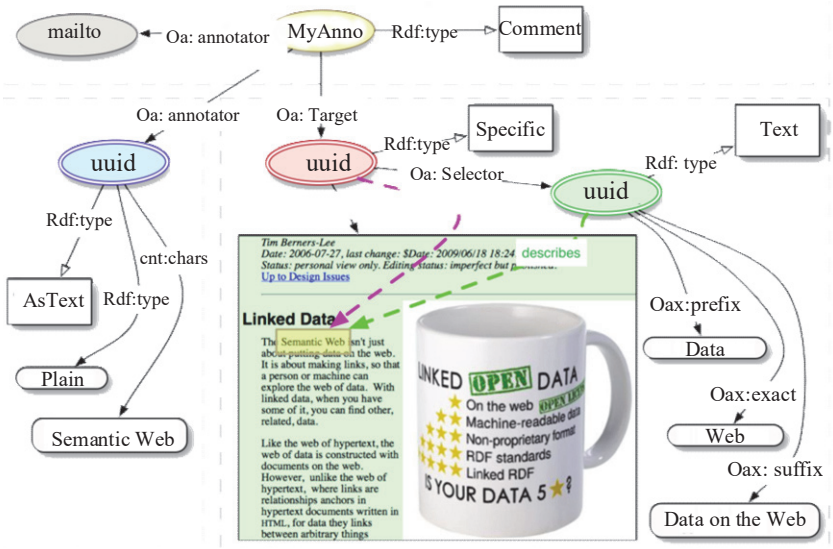

Figure 1 Structure of semantic annotation of common Web-based English education resources.

topics in the documents, and analyzes the document-topic-keyword tagging from three different levels. Semantic descriptions of Web-based English education resources enrich the metadata attributes and content of resources, adding more topics related to subjects, teaching contents, teaching methods, etc.

2 Semantic Annotation Modeling of Web-based English Education Resources

2.1 Multi-site Feature Fusions Based on Topic Area Extraction Algorithm

Web English educational resources are information carriers that contain not the only key, topical information about student concerns, but also a lot of noise, such as Web English educational resources, navigation bars, and copyright notices. Woodman calls the area shaped like area 2, which contains the topic information from Web English Educational Resources, the Web English Educational Resources topic area. Although this noisy information is useful for website owners and viewers, it often hinders the Web English educational resource analysis program from extracting and semantically annotating the

topic information from Web English educational resource content [14]. For this reason, identifying topic regions and filtering out noisy information can improve the clustering, classification, and topic extraction performance of Web English educational resources, which are important for the semantic annotation of Web English educational resource content [15].

For a particular Web site, most pages are dynamically generated using modular programs (JSP, ASP, PHP, etc.), so they generally have some fixed page layout. Thus, if the layout and content of some page sections appear frequently on other pages, they are likely to be noise, whereas the real content of interest to students (e.g., text, images, etc.) is usually obtained from a database with different content and style. The distinction between subject areas and noise areas is often very clear in some cases [16]. For example, in a news site, the subject area tends to contain more text than the noise area and occupies a central position in the Web English education resources. In e-commerce sites, for example, subject areas are arranged regularly and can often be divided into smaller areas by TR.

Based on the above analysis of Web English educational resources, this paper proposes a topic area extraction algorithm based on a site style tree and multi-feature fusion, which formalize the topic area problem into a classification problem [17]. The algorithm first constructs a SiteStyleTree based on the Web English education resource set, and then calculates the importance of each node based on the information, and obtains the candidate topic areas based on the node importance. Candidate topic regions are coarse-grained and contain a large number of pseudo-topic regions, which need to be further filtered and validated. In this section, in the topic region validation phase, we extract 10 characteristics of candidate topic regions including the number of texts, the number of external links, the region height, the width, etc., and use these characteristics as classification features in the topic regional recognition phase [18]. Due to the good performance of SVM on two-class learning problems, this section uses this method to train the classifiers on each feature. The fusion of the classifiers is achieved by a boosting algorithm proposed in this section, which adjusts the classifier weights in each round of iteration based on the classifier accuracy on the sample distribution that is dynamically adjusted on each round of weight. Based on the classifier weights obtained from training, some of the lowest weighted classifiers and corresponding features can be discarded, making the testing complexity reduced [19]. Moreover, the experimental results in this section show that the appropriate discarding of some features does not significantly affect the final performance.

2.2 Web English Educational Resources Site Style Tree Construction Extraction Algorithm

Although the DOM uses a tree structure to organize the nodes of HTML elements in Web English educational resources, it is weak in topic area recognition and noise detection, because the DOM tree has neither semantic information nor any statistical features. However, noise handling based on the style tree can be a good solution to the above problems, as it uses the style tree as the core data structure. As shown in Figure 2, the style tree is a data structure that can easily and accurately describe the style and content distribution of the entire site, and it compresses the common parts of the entire site page set.

A style node represents a layout or appearance style and consists of two parts, denoted by the symbols (Es, n), where Es is a string of element nodes and n is the number of pages at the node level that contain this particular style. The element node is composed of three parts, represented by the symbol (TAG, Attr, Ss) where TAG is the tag name, Attr is the set of attributes of the tag, and Ss is the set of style nodes in the context under the element node. An example of a site style tree is obtained by merging the two d1 and d2 style trees. It is clear to see that in d1 and d2, the markers are the same except for the bottom marker string, which is different, d1 is (P, IMG, PA) and d2 is (P, BR, P). When merging dl and d2, the number of pages with different styles below the tag is indicated by the number on the arc from the table tag. The construction process of the site style tree generated from the DOM is in a top-down model, and the resulting SST describes the site style distribution.

NodeImportance For each element node E in the SST set m the number of pages of the site containing E and l to the number of child style nodes. Then the importance NodeImportance(E) of element node E is defined as follows:

$$\mathrm{NodeImP(E)} = -\sum_{i=1}^{i} \frac{p_i}{\log_m p_0} \tag{1}$$

CompositeImportance is expressed as CompImp(E) for the intermediate element node E and is defined as follows:

$$\mathrm{CompImP(E)} = \frac{(1-\gamma)\mathrm{NodeImp(E)}}{\gamma^2 p_i} \tag{2}$$

In the above equation, pi is the probability that a page will use the *i* style in E.Ss, and γ is typically chosen to be 0. CompImp(Si) is the importance of

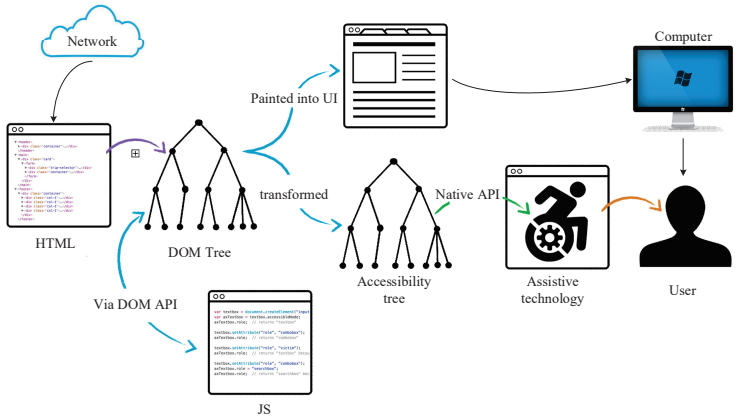

Figure 2 DOM tree and style tree ST.

the style node Si(E.Ss), defined as:

$$\mathrm{CompImP}(S_i) = k \cdot \sum_{j=1}^{k} \mathrm{CompImp}(E_j) \tag{3}$$

To mark a candidate topic area, after calculating the node importance, you can define a candidate topic area. The CandidateTopicArea (CTA), which is determined by such an element node E, satisfies the following rules.

$$\mathrm{NodeImP(e)} < m, \mathrm{CompImp(e)} > m, \nabla e \in parent(E) \tag{4}$$

2.3 Topic Classifications of Web-based English Educational Resources Based on the LDA Topic Model

For online Web-based marketing of Web educational resources, the Web educational resource loading and matching strategy based on Web educational resource themes has been one of the hot topics of research, because theme-based matching can improve the efficiency of Web educational resource placement compared with keyword-based matching. To achieve this form of efficient matching, it is a key step to classify and extract the topics of Web-based English education resources.

In the structure of human cognition and experience, there are a large number of concepts that are similar or equal in meaning. When matching Web English educational resources, only matching based on individual phrases is likely to produce partial understanding. Since most of the current Web

English education resource loading systems are based on keywords, they do not solve the problem of Web English education resource-related matching fundamentally. To solve this problem, the Web English educational resource loading system must analyze the Web English educational resource from the perspective of the topic, because the topic can express the descriptive intent of Web English educational resources more accurately than phrases [20]. To be more specific, we first extract keywords of Web English educational resources, then analyze the semantic contextual relationship between the keywords, that is, analyze the theme expressed by the combination of the keywords, and finally combine the similarity calculation of the theme to extract the most similar and matching Web English educational resources from the candidate Web English educational resources database. The key question here is how to organize the semantic context from the extracted Web English educational resources keywords and extract the themes embedded in them [21]. This section discusses issues related to the topic classification of Web English educational resources.

The mapping rules for topic classification of Web English educational resources are discriminating formulas and discriminating rules established by the classification system to summarize the patterns of topics based on information from several samples of each topic class. Then, when encountering untagged topic categories of Web English educational resources, it will determine the topics related to the content of Web English educational resources according to the discriminant rules established. Extraction and classification of topics are mainly solved by the hierarchical mapping idea of "Web English educational resources - words – topics". Web English educational resources are composed of words according to lexical and semantic rules, while themes are composed of words according to their relevance.

The probabilistic topic model can be interpreted from two perspectives, corresponding to different applications, i.e. document generation and topic extraction. The topic model can be viewed as a model of document generation, i.e., the process of generating a document as a simple probabilistic process based on the topic model [22]. When a new document is to be generated, the first distribution of topics is obtained, then for each word in the document, a topic is randomly obtained from the topic distribution, and then a word is randomly obtained from the distribution of words in the topic. The document generation process can be reversed to extract the topic of the document, that is, the content of the document can be used to calculate the distribution of topics and topics of the distribution of a word. Currently, it is common practice to indirectly calculate the topic-word distribution and topic

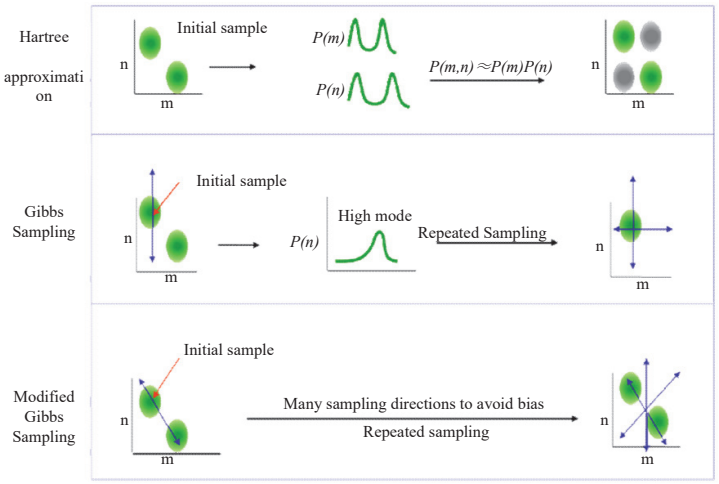

Figure 3 Structure of Gibbs sampling algorithm.

distribution by finding the a posteriori distribution of potential topics (i.e., assigning each word in a document to a topic) based on the visible word sequences in the document.

The key to the analysis and extraction of the thematic probabilistic model LDA is to compute and evaluate the distribution of themes on the document. Current methods for extracting LDA models are the variational method, expectation-maximization EM algorithm, and the Gibbs sampling algorithm. The model obtained by the variational method deviates from the real situation, while the EM algorithm often fails to find the optimal solution due to the local maximization of the likelihood function. However, Gibbs sampling is a Markov-chain Monte Carlo (MCMC) method that is very efficient in extracting topics from large document sets and is relatively simple to implement. As shown in Figure 3, Gibbs sampling is the most popular and commonly used LDA model extraction algorithm.

3 Experimental and Analytical Text Annotation in English Education Based on the LDA Model

3.1 Web English Education Page Interest Quantification Method Evaluation

When academic students are browsing Web English educational resources, their interest level changes in a similar way, which can be roughly divided into

three stages. In the beginning, students do not know much about the contents of Web English educational resources and they are in the stage of understanding the contents of Web English educational resources. At this stage, students' interest in the Web educational resources is very low and changes slowly with browsing time. When students have a general understanding of the contents of the resources, they will decide whether to continue browsing according to their preferences. If they are interested, they will choose to continue browsing more carefully, and if not, they will stop browsing. At this stage, as the browsing time changes, the student's interest in Web-based English education resources varies considerably. After a certain period, students have already detailed the contents of the Web English educational resources and their interest in them has grown to a certain degree. However, if students are interested in the Web-English educational resources, they may choose to browse the resources again. At this stage, students have already shown a great interest in the Web educational resources, and their interest starts to slow down as the browsing time increases. From the above analysis, we can see that the change of students' interest in Web educational resources concerning browsing time should be an S-shaped curve, so the quantification of interest should be considered using a mathematical model more in line with the change.

To conduct the parameter evaluation experiment, this paper establishes an experimental evaluation system that collects news, sport, technology, and other information from various web portals and displays it on the site according to the time of collection. Besides, four test students browse news and other information through the site every day, and the system records the time and number of times the students browse the site. The interface of the experimental system is shown in Figure 4. In this paper, the sample data of four students are collected through the experimental platform described above. Sixty percent of the data are used for parameter estimation and 40% for quantitative testing [23]. The scatter plot of the sample data is used for parameter estimation. The horizontal coordinate represents the unit browsing time of the students browsing the Web English educational resources, and the vertical coordinate represents the rating value of the students' interest in the Web English educational resources. After completing the parameter estimation and applying it to the interest quantification formula, we quantified the interest of some of the sample data of the four students who participated in the experiment. Quantified interest values were compared with the students' ratings to assess the quantitative effect. After quantifying the interest in the pages using the LDA method, the absolute value of the residuals was

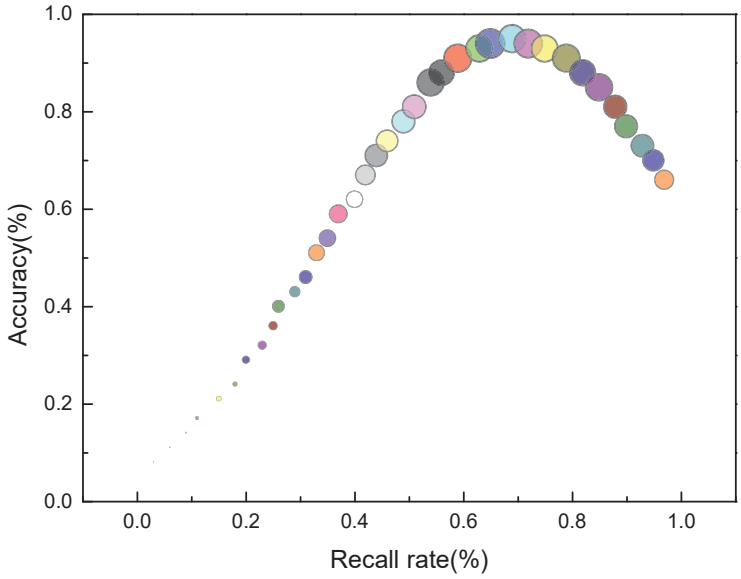

Figure 4 Fit on the web english education resource interest page of the LDA model.

calculated based on the quantified interest, and the interest ratings given by the students. The mean absolute residual values of the logistic quantification method for the four students were 0.0362, 0.03444, 0.03326, and 0.03432 in order, which showed that the logistic-based page interest quantification method could fit the students' interests very well.

We first obtain access records of the test students over some time, then divide them into groups by period and chronological order, cluster the first group and the cluster interest and then add the subsequent data in turn. To accurately examine the stability of the cluster interests of Web English educational resources, the newly added Web English educational resources are classified into the original clustering results according to the KNN classification method after the new data are added. After adding all the data, the computational results of different clusters are counted according to the number of Web English educational resources in the clusters, and the experimental results are shown in Figure 5. From the experimental results, it can be seen that the cluster interest quantification methods based on the cluster prime center and Gaussian model are both stable and can be used to represent the students' stable interest, but in contrast, the Gaussian model has a higher stability performance.

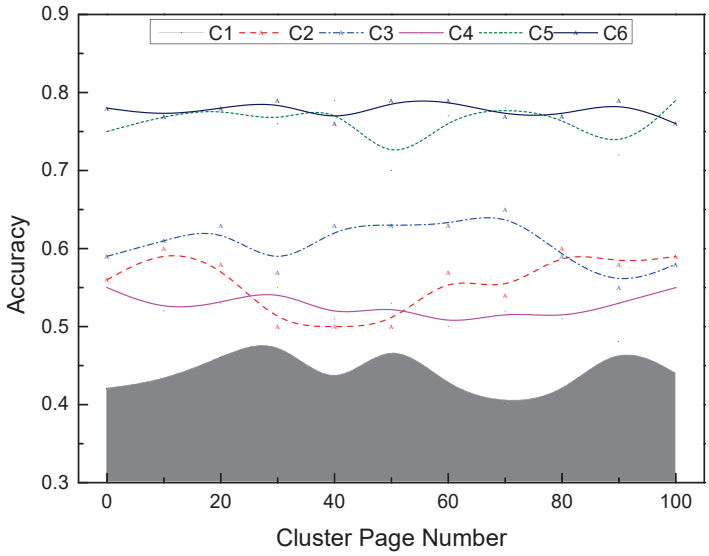

Figure 5 Comparison of quantitative stability of interesting clusters in Web English education resources.

3.2 Web-based English Education Resource Standard Details Variance Marking Assessment

The research experiment is to perform semantic modeling on all documents in the set to compare the effects of LDA on document semantic classification and keyword distribution. Secondly, multiple documents with the same educational content are selected for semantic modeling to examine the detailed differences in the semantic modeling of documents with the same educational content by LDA. Finally, the performance of LDA in large-scale document semantic modeling under Map/Reduce parallel computing framework is verified to meet the practical application requirements for semantic modeling and annotation of English educational resources on the Web.

From the perspective of the role of LDA in the classification of document resources, the feature keywords in the "topic-keyword" distribution obtained by LDA modeling can be used as semantic metadata of document resources to classify and retrieve catalogs on resource sharing platforms and to label the common features and teaching priorities and difficulties of Web-based English educational resources. It provides richer resource retrieval and navigation services for users. To further verify the semantic mining effect of the LDA model on the detailed themes of Web-based English educational

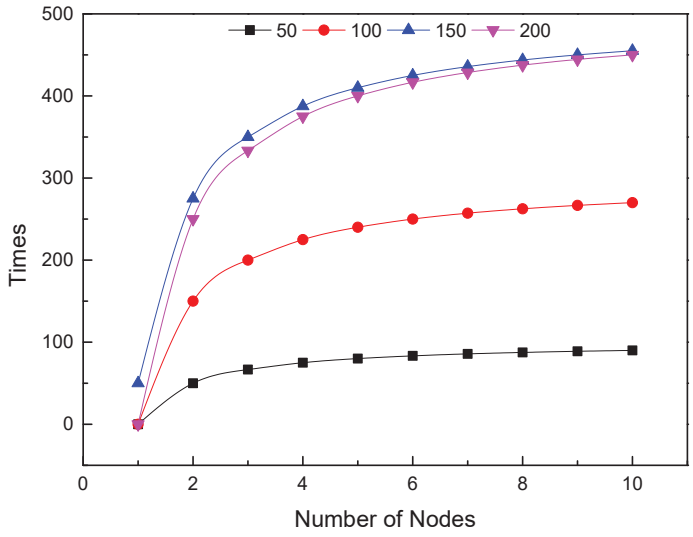

Figure 6 Calculation time of LDA at different nodes and number of documents.

resources, the study counted the number of lesson plans of six randomly selected texts in each junior high school volume and conducted LDA modeling experiments on those with more than 50 textbooks. The top four themes and their corresponding keywords were obtained after the modeling of the themes of the lesson plans. It can be seen that Topic0 is the noun kinship, Topic1 is the description related to the teaching objectives and requirements and the learning requirements that students should achieve, Topic2 is the series of actions and performance of kinship expressed in the form of verbs, and Topic3 is the feature mining for the text genre and description style. Small themes and keywords within these lesson plans describe the characteristics of each lesson plan in multiple details, providing a richer and more accurate internal semantic description of the resource metadata.

The "topic-keyword" correspondence matrix obtained from the LDA topic modeling of other textbook lesson plans reveals similar topic features and corresponding keyword distribution features, which capture the teaching requirements, teaching methods, central ideas, and other intrinsic details of Web English educational resources and provide rich semantic information. Both teachers and students can find the resources they need from these detailed "themes and keywords" to grasp the key points of teaching and learning.

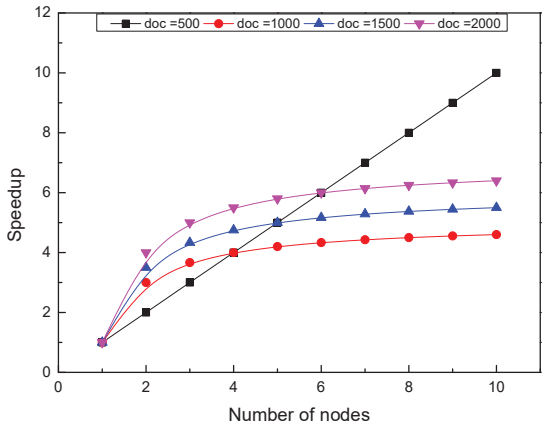

Figure 7 Acceleration ratio of LDA at different nodes and number of documents

Another advantage of the LDA topic model is its ability to model large-scale document topics. The experiment randomly selected 5000, 15000, and 40,000 documents from the lesson plan document library, and then selected a different number of parallel computing nodes, ran it three times, and averaged it to get the time used for LDA modeling in multiple parallel computing nodes for different size datasets. As shown in Figure 6, the ratio of the LDA modeling time and the benchmark time in a multi-node environment is used as the acceleration ratio. It shows the time and acceleration ratios for the LDA modeling of different corpus sizes for multiple nodes in parallel.

It is obvious from Figure 7 that the LDA modeling time for a large corpus set is longer for the same number of parallel nodes, and as the number of parallel computation nodes increases, the document modeling time is shorter and the time difference between the three corpus sets is smaller, which means that the LDA modeling acceleration ratio increases as the number of parallel computation nodes increases. This means that the LDA model has superior computational performance with larger datasets.

3.3 Evaluation of Significant Text Recognition and Annotation of Graphic Web-based English Educational Resources

This paper takes advantage of the color, stroke, shape and other visual features of distinctive text to realize the effective complementation of different features in the detection and recognition of distinctive text. In the feature extraction and detection stage, the areas are clustered by color. The main purpose of clustering by color is to extract the color distribution features

of an area, which can grasp the color tone of the area as a whole. Similar regions generally have similar distribution characteristics. In the area fusion stage of the candidate text, the area energy map is used to fuse the edge sets, which transform the area fusion problem into a directional projection law detection problem and effectively improves the fusion performance. The regional energy map is constructed by wavelet transform. In the candidate text area validation stage, an LDA-based classification validation method is used to classify the text areas, which not only includes conventional visual features but also color distribution features. In the text recognition stage, the segmentation of the stroke map and text color layer was performed separately, and the text regions were baptized from the perspectives of strokes and colors, respectively. The two methods were analyzed and compared, and the two methods were integrated, which effectively improved the quality of minimization.

After the image is transformed by the wavelet boosting algorithm, the image is decomposed into several sub-images by direction, in which all sub-images, except for the low-frequency sub-image, contain edge features in a certain direction, representing the texture features of the image in that direction. For example, textural features of a horizontally aligned text region in the horizontal direction can be a good model of the horizontal curve of the font energy. However, simulating the whole region will reduce the performance of the algorithm, for this reason, this paper first clusters the region color, extracts the distribution of regional color features, and the distribution of similar characteristics of the region wavelet changes can be more effective in extracting the direction of the texture feature changes. The above algorithm is effective in detecting the directionality of text, but there is another problem, namely, the problem of text block construction.

In this paper, experiments are conducted on a large number of multimedia images of Web English educational resources. The experimental images contain high-quality training and testing data from commercial Web English educational resources crawled from the Internet. Each image contains one or more pieces of text in Web English educational resources. To train the SVM classifier, this paper arranges 200 image Web English educational resources in the training set and the test set respectively.

By extracting the set of edges, it is possible to identify connected fields that are regularly arranged in the image. As shown in Figure 8, the next step is to remove these areas that are mistaken for text (called pseudo-text areas in this paper). Besides, detecting text blocks may result in overlap, e.g., when two or more horizontally aligned texts are close enough to each other that

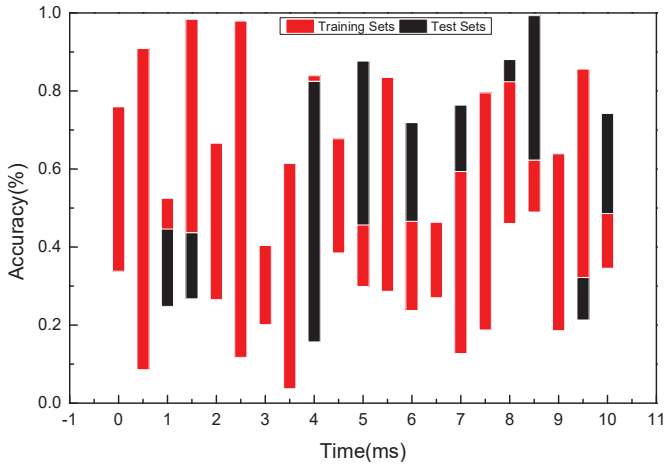

Figure 8 Comparative analyses of the effect of image resources on algorithm performance.

multiple vertically aligned texts may be constructed when detected vertically. The identification and elimination of pseudo-text regions and overlapping phenomena are the main tasks of Wenmu-area fusion. From the bionics point of view, since the text is arranged in a regular pattern, when people browse the text, the stimulus of the text to the human eye is always passed from one word to the next, which means that the stimulus of the text to the human eye is directional, and the direction of the stimulus is the direction of the text arrangement. Besides, people can always smoothly navigate through a line of text word by word, that is, without stopping between words. For this reason, the stimulation of text to the human eye is much larger than the stimulation of the space between negligible words. As shown in Figure 9, when we try to navigate horizontally aligned text in a vertical direction, we can always feel that the human eye is drifting, i.e., the stimulus is not in a straight line. Based on the above observations, this paper attempts to quantify the textual stimulus to the human eye, defined here as the energy of the font. The stimulus to the human eye can be simulated when the browsing direction and the text alignment are the same, and the stimulus to the human eye can be simulated when the browsing direction and the text alignment are not the same.

3.4 An Application of the LDA Model for Loading and Tagging Web-based English Education Resources

The prototype system of Web English education resource loading and annotation is designed to improve the mismatch of Web English education resources

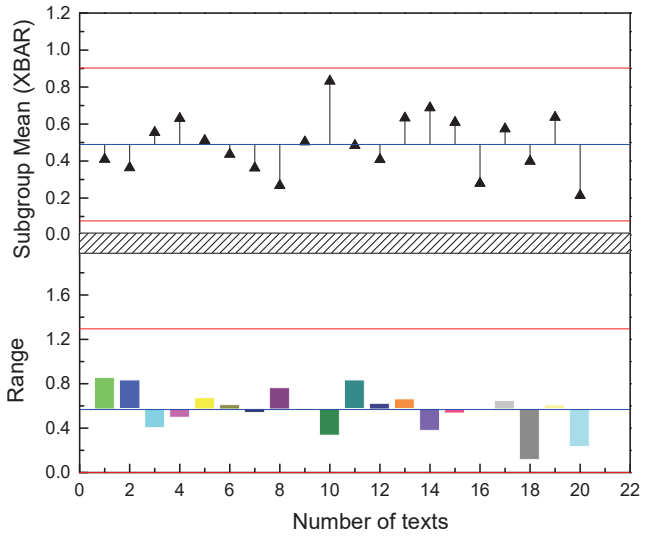

Figure 9 Image significant text recognition annotation evaluation results.

and improve the performance of Web English education resource service system, which includes online subsystem and offline subsystem. The online subsystem mainly obtains related information such as student ID, site ID, page ID, etc., and then gets the Web English education resources through the Web English education resources data access module and loads them into the Web English education resources. The offline part mainly provides semantic annotation and sorting functions for the Web English education resources.

From the figure above, we can see that the semantic annotation of the Web English education resources and the Web English education resources sorting algorithm are the core modules of the Web English education resources loading system, which are the key and guarantee to reach the best matching among students, target Web English education resources and Web English education resources. We will analyze the contents of each module and the technical methods used by each module.

As shown in Figure 10, the Web English Education Resource Data Access module accepts the Web English Education Resource Service request sent by the student (the browser), which contains all the information of the publisher, and subsequent billing relies on this key information. The Web English Education Resource Service request is generally sent as JavaScra.pt, which is highly scalable and can be dynamically displayed using Javascript with a text-based approach. In this system, students are required to implant detailed

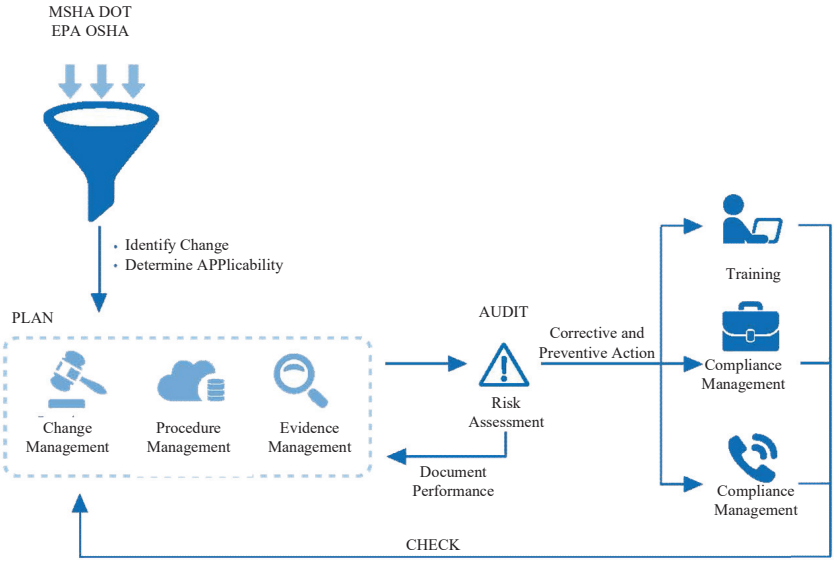

Figure 10 Schematic diagrams of the prototype architecture of the Web English educational resource loading an annotation system.

identifiers of the Web English educational resources display context when they send Web English educational resources service requests, including publishing site ID, the target page ID, and student ID.

Web Business Access to English Educational Resources module is responsible for developing the Web Business Access to English Educational Resources protocol for the Web English Educational Resources Loading and Marking Prototype System and Wedsite. In this system, this module is mainly used to enter the web site's usage logs to engage students. Before the Web site's usage log is used to load and mark the prototype system of Web English educational resources, it is necessary to use the relevant data protocol to filter out irrelevant records and incomplete records, and after entering the system, it is necessary to store them following the phased storage format and store them in the database of students' personalized information. When the student information is stored, the student's identifier should be checked or set to ensure the consistency and integrity of the student's personalized information. It is a key component of the offline subsystem, and the system doesn't need to capture student usage logs. It is also important to define protocols for how to update and maintain the logs when new ones are added to the system.

4 Conclusion

The development and application of semantic annotation technology provide new ideas and methods for the construction of basic Web-based English education resources. The use of LDA semantic modeling technology can my potential topics and keywords in document resources, provide semantic metadata in addition to basic metadata for the construction and sharing of resource repositories, and provide more reference information for teachers and students using Web-based English education resources. From the experimental results, the "document-topic-keyword" semantic model obtained from LDA modeling of Web English educational resources can enrich the metadata description of the resources from the document semantic level, and its computational performance is improved by the parallel Map/Reduce algorithm. This is a sufficient advantage for the computing environment.

The value of this paper is to use LDA topic modeling to automatically perform semantic analysis and topic modeling on documents in repositories, and provide technical support for semantic annotation of repository resources. In this paper, we propose a semantic annotation-based approach to solve the relevant matching problem among the subjects of Web English education resources in the loading process of Web English education resources. The procedure is as follows: Firstly, the target Web English education resources, student interests, and Web English education resources are semantically annotated, and their semantic features are extracted. The semantic features of these resources are extracted. The relevance features of the target Web English education resources and the Web English education resources are extracted and ranked according to their degree of relevance to get the first-round Web English education resources ranking results; finally, the relevance features of student interest and Web English education resources are extracted and the results of the first-round Web English education resources are re-ranked according to their degree of relevance to get the final Web English education resources ranking results. Based on the semantic annotation method, the correlation matching problem between subjects of Web English educational resources is transformed into a semantic correlation order problem, which can be analyzed and solved by using mature text processing technology. The shortcoming of this study is that the LDA semantic modeling technique is still at the experimental research stage and needs to be tested in practice. The resources involved in this study are lesson plan documents, while the LDA semantic modeling of short texts needs to be further solved for multimedia resources such as audio, video, and Flash animation. Besides, the

LDA model needs further attempts in timely data processing for the growing number of Web-based English education resources every day.

Acknowledgments

This work was supported in part by the Talents Highland for Ningxia High-level Translation and International Publicity Translation.

References

[1] Goodman M G, Alvarez M, Halstead S B. Secondary infection as a risk factor for dengue hemorrhagic fever/dengue shock syndrome: a historical perspective and role of antibody-dependent enhancement of infection[J]. Archives of virology, 2013, 158(7): 1445–1459.

[2] Jackson P, Raiji M T. Evaluation and Mangement of Intestinal Obstruction[J]. American family physician, 2011, 83(2): 159–165.

[3] Vidal J C, Lama M, Otero-García E, et al. Graph-based semantic annotation for enriching educational content with linked data[J]. Knowledge-Based Systems, 2014, 55: 29–42.

[4] Xu H, Zhang R, Lin C, et al. Novel approach of semantic annotation by fuzzy ontology based on variable precision rough set and concept lattice[J]. Int. J. Hybrid Inf. Technol, 2016, 9(4): 25–40.

[5] Kadda B, Ahmed L. Semantic annotation of pedagogic documents[J]. International Journal of Modern Education and Computer Science, 2016, 8(6): 13.

[6] Vrablecová P, Šimko M. Supporting semantic annotation of educational content by automatic extraction of hierarchical domain relationships[J]. IEEE transactions on learning technologies, 2016, 9(3): 285–298.

[7] Koutsomitropoulos D A, Solomou G D. A learning object ontology repository to support annotation and discovery of educational resources using semantic thesauri[J]. IFLA journal, 2018, 44(1): 4–22.

[8] Jensen J. A systematic literature review of the use of semantic web technologies in formal education[J]. British Journal of Educational Technology, 2019, 50(2): 505–517.

[9] Sánchez-Nielsen E, Chávez-Gutiérrez F, Lorenzo-Navarro J. A semantic parliamentary multimedia approach for retrieval of video clips with content understanding[J]. Multimedia Systems, 2019, 25(4): 337–354.

[10] Balavivekanandhan A. A technique for semantic annotation and retrieval of e-learning objects[J]. International Journal of Business Intelligence and Data Mining, 2020, 17(1): 12–31.
[11] Zarzour H, Sellami M. A linked data-based collaborative annotation system for increasing learning achievements[J]. Educational Technology Research and Development, 2017, 65(2): 381–397.
[12] Chi Y, Qin Y, Song R, et al. Knowledge graph in smart education: A case study of entrepreneurship scientific publication management[J]. Sustainability, 2018, 10(4): 995.
[13] Stepanov E A, Chowdhury S A, Bayer A O, et al. Cross-language transfer of semantic annotation via targeted crowdsourcing: task design and evaluation[J]. Language Resources and Evaluation, 2018, 52(1): 341–364.
[14] Simko M, Bielikova M. Lightweight domain modeling for adaptive web-based educational system[J]. Journal of Intelligent Information Systems, 2019, 52(1): 165–190.
[15] Neal M L, König M, Nickerson D, et al. Harmonizing semantic annotations for computational models in biology[J]. Briefings in bioinformatics, 2019, 20(2): 540–550.
[16] Wongthongtham P, Chan K Y, Potdar V, et al. State-of-the-art ontology annotation for personalised teaching and learning and prospects for smart learning recommender based on multiple intelligence and fuzzy ontology[J]. International Journal of Fuzzy Systems, 2018, 20(4): 1357–1372.
[17] Zarzour H, Sellami M. Effects of a linked data-based annotation approach on students' learning achievement and cognitive load[J]. Interactive Learning Environments, 2018, 26(8): 1090–1099.
[18] Gayoso-Cabada J, Goicoechea-de-Jorge M, Gómez-Albarrán M, et al. Ontology-Enhanced Educational Annotation Activities[J]. Sustainability, 2019, 11(16): 4455.
[19] Gomes Jr J, Dias L L, Soares E R, et al. Framework for Knowledge Discovery in Educational Video Repositories[J]. Computing and Informatics, 2020, 38(6): 1375–1402.
[20] Al-Osta M, Ahmed B, Abdelouahed G. A lightweight semantic web-based approach for data annotation on IoT gateways[J]. Procedia computer science, 2017, 113(1): 186–193.
[21] Rezgui K, Mhiri H. Towards a Semantic Framework for Lifelong Integrated Competency Management and Development[J]. The Computer Journal, 2020, 63(7): 1004–1016.

[22] Yanchinda J, Yodmongkol P, Chakpitak N. Measurement of Learning Process by Semantic Annotation Technique on Bloom's Taxonomy Vocabulary[J]. International Education Studies, 2016, 9(1): 107–122.
[23] Pereira C K, Siqueira S W M, Nunes B P, et al. Linked data in Education: a survey and a synthesis of actual research and future challenges[J]. IEEE Transactions on Learning Technologies, 2017, 11(3): 400–412.

Biographies

Wei Du is an Associate Professor at School of Foreign Languages and Cultures, Ningxia University, China. She is currently doing her PhD in Knowledge Management at College of Arts, Media and Technology, Chiang Mai University, Thailand. Her research interests are related to translation teaching and language policy and language management.

Haiyan Zhu is an Associate Professor at School of Foreign Languages and Cultures, Ningxia University, China. She holds a PhD in Ethnology from Ningxia University, China. Her research interests are related to translation teaching, corpus-based translation, ethnology.

Teeraporn Saeheaw is currently an Assistant Professor at College of Arts, Media and Technology, Chiang Mai University, Thailand. She holds a PhD in Knowledge Management from the Chiang Mai University, Thailand. Her research interests are related to innovation and learning, knowledge management in cognitive science, and second language study.