# Construction of English Teaching Corpus Based on DM Technology in "Internet+" Era

Jianbo Xu, Hao Tang, and Zhihui Liu(✉)

College of Foreign Languages, Xiangnan University, Chenzhou 423000, Hunan, China
tonyxu5132@126.com

**Abstract.** With the rapid popularization of the Internet, a new project-based teaching of college English can be carried out in the form of "internet plus Corpus" to meet the learning needs. However. There are countless online learning platforms for English, but the learning content provided by them is very single. All learners, regardless of their learning purpose, see the same learning content. In addition, because the differences in learning basis, ability and interest of teaching objects are in great contradiction with the singleness of the traditional auxiliary teaching system, it is very important to realize the personalization of the system. Network-assisted teaching is a new educational technology. This new learning method breaks the time and space restrictions and geographical restrictions of traditional education, and can make full use of educational resources. These systems not only improve the level of teaching and management to a great extent, but also accumulate a large number of teaching and management data. However, at present, most of these information systems are online transaction processing systems, which lack the ability of comprehensive analysis and auxiliary decision-making. This paper introduces the application of data mining (DM) in English teaching corpus, analyzes and introduces the application of DM technology in the construction of teaching corpus, which is of great value to students' learning and teachers' teaching.

**Keywords:** DM · English teaching corpus

## 1 Introduction

Due to the continuous progress of IT, under the background of "internet plus Corpus", college students are not satisfied with simple textbook teaching in the process of English learning [1]. Their learning state is more inclined to autonomous, skilled and flexible learning [2]. China's Ministry of Education has put forward opinions on the reform of instructional mode that are in line due to the continuous progress of the present class, requiring universities to change the traditional instructional method that teachers give priority to teaching and students passively accept learning, and more importantly, make full use of computer network and multimedia technology in instructional methods [3]. As a result, there are numerous online English learning systems, including English courses of online education institutes set up by various universities, learning platforms provided


Y. Zhang and N. Shah (Eds.): BigIoT-EDU 2023, LNICST 580, pp. 512–519, 2024.
https://doi.org/10.1007/978-3-031-63130-6_57

by various English training institutions in the society, and so on [4]. English teaching corpus based on data mining (DM) technology is one of the hot research directions in DM at present, and it has a wide application range and great potential [5].

However, the existing network-based teaching system can't prepare the most suitable learning content for students according to their cognitive model, and can't teach students in accordance with their aptitude [6]. Without individuation, students can't learn according to their needs, that is, no matter which student visits the site, the learning content they see is almost the same, and they can't provide their own learning content or learning progress according to their own conditions, so they can't achieve individualized teaching and heuristic teaching required by pedagogy [7]. Because DM can cross and interact with other disciplines, it can transform simple data in the database into knowledge and information of some industries, and save a lot of manpower and material costs, so this new technology of DM has aroused great concern from all walks of life. Therefore, it is necessary to construct an English teaching corpus based on DM technology in the era of "internet plus" [8]. Universities have made gratifying achievements in MIS (Management Information System) [9]. At the same time, it provides a new opportunity and platform for the renewal of learning ideas and the transformation of instructional mode.

If the DM technology is introduced into the instructional model of universities, using the DM technology to analyze the existing data of the teaching management system, and then provide services for students, teachers and managers, it has become a new topic faced by teaching management [10]. Therefore, it is of great educational significance to apply corpus in English teaching and make use of its teaching auxiliary role to improve the efficiency of classroom teaching. At the same time, multi angle analysis and DM of the data in the existing teaching management system are carried out in order to find useful knowledge for school teaching management and student management. This knowledge can undoubtedly assist school managers in decision-making and improve the competitiveness of the school. It will also provide a practical basis for the decision-making of leading departments at all levels of the school to improve teaching quality and optimize teaching resources. For the school to take the initiative in the fierce competition, provide a broader space and play an important role in the future development.

## 2 Application of DM in English Teaching Corpus

### 2.1 DM Method

DM is the process of extracting information and knowledge that is hidden, unknown to people in advance, but potentially useful and ultimately understandable from a large number of incomplete, noisy, fuzzy and random practical application data [11]. DM is also called knowledge discovery in the field of artificial intelligence. At the same time, DM is also a basic step in the process of knowledge discovery. In this content library, in order to organize the content more conveniently, the contents of the same module are stored in the data table of the same database, and the specific design of the database is shown in Fig. 1.

The database system has also developed from the early layered and meshed database system to relational database system. Structured query language, online transaction processing, multidimensional database and other technologies make it possible to effectively

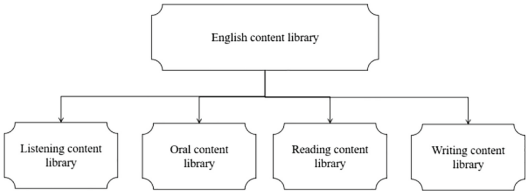


**Fig. 1.** Content library design

store, retrieve and manage a large amount of data [12]. When using machine learning model algorithm to train data sets, the features or attributes of sample data can be divided into local features and global features. When using model algorithm to mine the features of sample data, it can be divided into global features and local features according to the machines corresponding to the mined targets. According to the principle of topic-related, data sampling takes out a subset of data related to the exploration target from a large amount of data, and provides materials and resources for the following DM. When using the algorithm to mine data features, the algorithm can't distinguish global features from local features according to people's intuitive understanding. Besides getting all the features of the data, some local features of the data may also be learned by the algorithm.

The advantages of data warehouse are also the necessary supporting points of DM, such as distributed processing of massive data, high-performance parallel processing technology and so on. OLAP (On-Line Analytical Processing) uses multidimensional data set and data aggregation technology to organize and summarize the data in the data warehouse, and uses online analysis and visualization tools to quickly evaluate these data. When more local features are mined, the proportion of new samples without local features will become more, and the algorithm will not be able to correctly identify these new samples, resulting in poor prediction effect, that is, the so-called "generalization ability" is poor. Different applications usually need to integrate methods that are particularly effective for the application. In an industry, there can be different mining methods according to different departments and mining tasks.

## 3 Application of Cluster Analysis

First of all, when building a corpus, we should comprehensively consider the present situation and teaching conditions of English teaching, adopt the principle of targeted practice, and fully embody the auxiliary and service functions of the corpus. The content organization system is divided into three functional modules, namely, data input, data organization and data output. The data organization module adopts genetic algorithm. Figure 2 is the functional structure diagram of the content organization system.

Multimedia technology can be divided into two different concepts: monitoring system and information. Multimedia is developed on the basis of computer mathematics technology. It realizes the combination of digital control and digital media [13]. Using the organic combination of sound, image and communication technology of computer technology, multimedia technology has been widely used in many industries in modern society, and has played a great role in promoting science and technology, education, production and other industries [14].

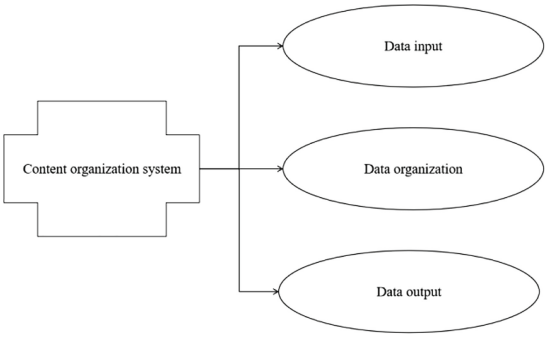


**Fig. 2.** Functional modules of content organization system

At present, CAI (Computer Aided Instruction) software based on behaviorist learning theory is still relatively mature and occupying the market [15]. The form of this software is familiar to everyone. Their common technical characteristics are that it is based on frame surface, adopts small-step branch programming, and students, as users of the software, passively accept knowledge instillation. Although these systems can efficiently realize data entry, modification, statistics, query and other functions, they can't find the relationships and rules existing in more and more data accumulated in daily work, and can't predict the future development trend according to the existing data. Therefore, improving the traditional instructional mode, reshaping the new learning concept and promoting quality education in an all-round way have become the top priority in the process of today's teaching reform. Therefore, the construction of corpus should be combined with the rules and characteristics of English teaching.

## 4 Application Analysis of DM Technology in the Construction of Teaching Corpus

### 4.1 Artificial Neural Network Mining Analysis

Artificial neural network is an information processing method developed under the inspiration of studying biological neural system. The core is the step-by-step recursive algorithm, which is based on the idea of two-stage frequency set. There are many successful models of artificial neural network, but the error back propagation (BP) model is the most perfect in theory and widely used. According to the requirements of system functions, artificial neural network mining mainly includes four levels: content database layer, student data layer, content organization layer and content display layer. Figure 3 is the architecture diagram of classical DM algorithm.

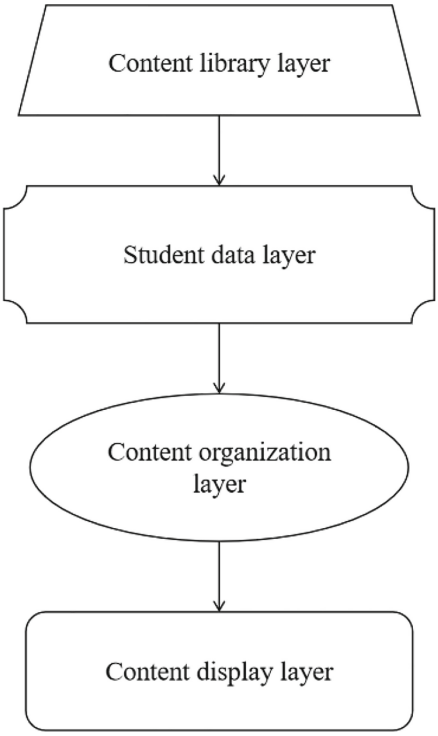


**Fig. 3.** Artificial neural network mining algorithm

The relationship between these frequency sets and the predefined minimum support is that the frequency of frequency sets must be greater than the minimum support. The purpose is to construct an accurate classifier by analyzing the characteristics of the training data set. The classifier can be used to judge the category of samples of unknown category. Therefore, it is necessary to reset the clustering center. The formula is as follows:

$$Z_j = \left( \sum_{i=1}^{n} (X_i)^j \right) / n \tag{1}$$

$Z_j$ – Output k cluster centers;
$C_j$ – Set of k cluster data objects;
$X_i$ – Set of n data objects.

The result of transaction execution should be a project that makes the database change from one consistent state to another, which is very important in relational database. According to the standard requirement of greater than the minimum support and greater than the predefined minimum reliability, strong association rules are generated between frequency sets. Most DM methods discard outliers as noises or anomalies. Over-fitting of the model means that when the model learns the features in the samples of training

data, it is too thorough to learn, which leads to the mining of more local features. These local features will lead to the prediction error of the model when the model is predicted, and finally it will be found that the model effect is not good when the model is evaluated.

When the highest rule found is applied to deal with practical problems, the rules in the rule set may overlap, so some records may have more than one rule, but these rule sets are more general than the decision tree. The deviation based method identifies outliers by examining the differences in the main features of a group of objects, rather than using statistics or distance measurement. Deviation analysis can detect credit card fraud. It can detect the fraudulent use of credit card by detecting that the payment amount of a given account is particularly large compared with the normal payment.

## 5 Analysis on the Construction of English Teaching Corpus Based on Internet Plus

The construction of English teaching corpus based on "Internet+" is not a simple text stacking. It needs to consider many factors and make a systematic planning according to certain principles. The so-called project-based instructional method refers to the process of arranging projects in the teaching process, allowing students to carry out project implementation, and then making summary and evaluation to expand project training. The rules were screened according to the research objectives and English knowledge. Figure 4 is the flow chart of applying relevance analysis in English online learning platform.

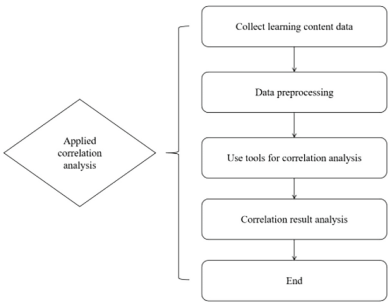


**Fig. 4** Flow Chart of Application Association Analysis

In order to speed up the mining and reduce the amount of data mined, the session set should be purified before mining, requiring the same user to contain at least two pages, and deleting users and sessions containing only one page. First of all, when scanning the database for the first time, the initial cluster is created according to the criterion function, and the total probability of the population is calculated to construct a roulette. The calculation formula is as follows:

$$P = \sum_{i=1}^{N} p_i \tag{2}$$

$p_i$ – Selection probability;
N – Total number of individuals in the population.

Then, scan the database for thousands of times and adjust the clustering results. If the clustering results are not changed by the two database scans, stop the cyclic results. These nodes are interconnected with each other through the network. If there is data input, they can determine the data mode. The reliability of the combined node is calculated by the following formula:

$$CF(A) = \sum_{i=1}^{n} w_i \times CF(A_i) \tag{3}$$

CF ($A_i$) – the credibility of each word condition $A_i$;
$w_i$ – weighting factor (1, 2, …, n).

Establishing a standardized corpus involves corpus design, corpus acquisition, corpus processing, corpus storage and corpus copyright, etc. The essence of college English education is to enable students to master the knowledge needed for their major or non-major occupations, and the purpose is to meet the needs of society and achieve certain professional skills or non-professional skills. Users' access has obvious preference bias: general users will have a targeted tendency when visiting, such as C#.net courses, Java basic courses and so on. The rule base contains those transformation rules that transform the problem from the initial state to the target state (or solution state). This iterative layer by layer search method has a large amount of access to the database. Each search has to scan the whole database. It is feasible for small sample data, but for a large amount of data, this scanning is a waste of system resources. Therefore, when designing the database, we should consider some conditions that will lead to data damage, and take corresponding preventive measures.

## 6 Conclusions

Due to the continuous progress of information technology, all fields of modern society have developed rapidly, and education is also facing new challenges and opportunities. College English teaching and evaluation system trains and evaluates English through online learning, so that students have made great progress in English listening, speaking, reading and writing. As a new data analysis technology, DM has made remarkable achievements and has been successfully applied to many fields. This paper attempts to use DM discovery method to mine English teaching corpus based on DM technology in the era of "internet plus", generate effective models and association rules, and analyze the models and discovered rules, so as to provide some useful reference for the teaching management decision-making in our school. Facts have proved that the new project-based instructional model under the background of "internet plus Corpus" is feasible and suitable for the needs of college English teaching today.

**Acknowledgments.** This work is supported by Hunan Academic Degree and Postgraduate Teaching Reform Project: *Research on the Teaching Ability Training Mode of English Teaching (No.2022JGYB227)*, Hunan Basic Education Teaching Reform Project: *Research on the Current Situation and Model Innovation of English Teachers' Professional Development in Rural*

*Primary and Secondary Schools in Chenzhou in the New Era* (No.Y20230752), and Chenzhou Chinese-English Bilingual Linguistic Landscape Corpus Technology R & D Center (No. 2022-39).

**Declaration of Interest Statement.** Author Jianbo Xu receives research fee from Chenzhou Chinese-English Bilingual Linguistic Landscape Corpus Technology R & D Center (2022-39), and the detailed listing is available at http://kjj.czs.gov.cn/zwgk/tzgg/content_3521589.html.

Jianbo Xu also receives research fee from Hunan Academic Degree and Postgraduate Teaching Reform Project: *Research on the Teaching Ability Training Mode of English Teaching* (No. 2022JGYB227).

And Jianbo Xu also receives research fee from Hunan Basic Education Teaching Reform Project: *Research on the Current Situation and Model Innovation of English Teachers' Professional Development in Rural Primary and Secondary Schools in Chenzhou in the New Era* (No.Y20230752).

No other author has reported a potential conflict of interest relevant to this article.

## References

1. Lu, D.: Corpus-based research and practice of college English for special purposes teaching. Northern Literature: China, no. 7, p. 2 (2018)
2. Qin, M.: Construction of English teaching corpus in the "Internet+" Era. J. Liaoning Radio Telev. Univ. (2), 2 (2018)
3. Yang, K.: An analysis of the effectiveness of "Internet+" English teaching. English Teach. **21**(15), 3 (2021)
4. Wang, Y.: Research on the innovation of college English project-based teaching mode based on "Internet + Corpus". J. Zhengzhou Railw. Vocat. Tech. College **030**(002), 75–77 (2018)
5. Li, Y.: The creation of multimodal corpus MCCT for Foreign language classroom teaching. J. Univ. Shanghai Sci. Technol.: Soc. Sci. Edn. **1**, 11 (2019)
6. Han, H., Jiang, Y., Yuan, X.: A study on corpus translator style in the era of big data. Foreign Lang. Teach. **40**(2), 6 (2019)
7. Yu, J., Fu, J., Bai, T., et al.: Data mining research on modality and context interaction based on unique attribute features. J. Yanshan Univ. **43**(5), 462–470 (2019)
8. Tang, J., Zhou, G.: Research report on the independent development of English teachers based on data mining. Women's Square (4), 2 (2020)
9. Fei, L.: Training critical thinking ability in problem-oriented teaching—Taking the teaching of network big data mining as an example. Modern Educ. Forum **3**(5) (2020)
10. Dong, J., Wu, H.: A corpus-based study on the effectiveness of English for special purposes teaching. Educ. Heilongjiang: Higher Educ. Res. Eval. **2**, 3 (2018)
11. Han, H., Jiang, Y., Yuan, X.: A study on the style of corpus translators in the era of big data. Foreign Lang. Teach. **40**(6), 2 (2019)
12. Zhang, S.: Research on the online corpus of agricultural English based on "Internet+". J. Hubei Correspond. Univ. **032**(169–171), 019 (2019)
13. Wang, Y.: Research on the innovation of college English project-based teaching mode based on "Internet + Corpus". J. Zhengzhou Railw. Vocat. Tech. College **30**(3), 2 (2018)
14. Cai, Y., Yu, X.: Research on corpus-based translation teaching mode under the background of "Internet+". J. Yichun Univ. **43**(4), 5 (2021)
15. Zhang, X., Ren, X.: A preliminary study on the construction of corpus-based college English vocabulary curriculum in the Internet+ era. Jiangsu Foreign Lang. Teach. Res (4), 2 (2018)