Journal of Intelligent & Fuzzy Systems xx (20xx) x–xx
DOI:10.3233/JIFS-189366
IOS Press

Research on data mining equipment for teaching English writing based on application

Bo Li[a] and Zheqian Su[b,*]
[a]*Foreign Language Teaching and Research Institute, Jilin University of Finance and Economics, Changchun, Jilin, China*
[b]*Foreign Language Department, Changchun University of Chinese Medicine, Changchun, Jilin, China*

Abstract. “College English teaching guide” puts in forth unique challenges and needs for College teaching skills of English. It is pressing to cultivate innovative talents with quality writing in English. Teaching in English, as a subject to check the mastery of students’ English knowledge. The successful instructional project of English writing is an assurance and support smooth growth of English writing. It can make education get double the efforts, and allow the students development.’ writing ability y, but in fact, the current situation of structural design of writing English in college is not optimistic. With the rise of “Web in addition to”, varying backgrounds have experienced emotional changes. “Web in addition to training” has become a pattern of advancement, which has brought new chances and difficulties for the educating and learning of College English composition. This research first describes the research status of data mining and College English writing at places of abroad and input the future the content of research and methods of research. Taking college English writing teaching as the research object, the association rules algorithm in data mining is applied to analyze the correlation factors of students’ writing performance and provide decision-making suggestions for teachers’ teaching.

Keywords: Data mining, English writing, teaching, association rules, internet plus

1. Introduction

With increased internet usage, the social life and social economy of human have entered the new era of information, digitalization and globalization, which makes the importance of English more and more outstanding. English is no longer just a language, a communication tool, but one of the necessary skills. As one of the most important international common languages in the world [1–3]. At present, there are many countries in the world that have taken English education as an important course and incorporated it into the school’s basic education curriculum system. English teaching has formed as vital portion of citizen quality education and is in a very important position. In our country, English is one of the required courses during the study of middle school, University, graduate student and so on. In the college entrance examination, college English test CET-4, CET-6, and postgraduate entrance examination, English scores have a very important proportion. So how to learn English well is becoming more and more important. But in the process of writing English textbooks at the present stage, they pay more attention to the students’ reading ability and lack the attention to the knowledge of writing [4]. In other words, in strategy of writing college teaching, the compiling of English writing is not systematic. Therefore, at this stage, we pay more

*Corresponding author. Zheqian Su, Foreign Language Department, Changchun University of Chinese Medicine, Changchun, Jilin, China. E-mail: zheqiansu@tutanota.com.

attention to how to use the scientific and technological means of mining data to recover the quality of education in English and enhance students' English writing ability [5, 6].

2. Related research based on data mining

2.1. Technical features of data mining

Information excavating is the way toward extricating concealed in-development and information from the huge, inadequate, loud, fluffy, and arbitrary crude information, which individuals don't know ahead of time, yet possibly helpful and believable. The characteristics of data mining include: massive data, large data sets, discrete variables, available rules, dynamic rules, and the relativity of rules. Data mining technology mainly includes two basic functions: description function and prediction function. At present, data mining has integrated various technologies and is divided into more functions. The functions of data mining technology can be more subdivided into the following functions [7–15]:

Data summary, data summary is derived from the Statistical score analysis. Compression of data is the main purpose of data summary, and a brief description of a data is given by summarizing the data.

Classification examination, the main attention of classification system is to construct a classification model or function. The classification model predicts categories by mapping each data item in the database.

Association analysis, association analysis is used to find the relevance of data in a database. The two main association analysis techniques are sequence patterns and association rules.

Forecast, forecast the trend of the future by analyzing the laws of the existing data.

Clustering analysis, cluster analysis is carried out by dividing the database into different groups. When grouping, the data of a group is as similar as possible, so there exist distinct difference amongst various groups. Cluster analysis is now widely used, but cluster analysis is different from classification analysis and is usually used for data subdivision.

2.2. Problems in English writing teaching

English composing capacity is a significant piece of English language capacity, yet the English composing capacity of Chinese understudies has not been improved successfully for quite a while. In the national college English test band 4 and 6, students' performance in listening and reading has improved significantly in recent years, but there is little improvement in writing performance. On the one hand, this aspect may be due to the demand for writing ability in the "College English syllabus" is relatively low, and on various portion, it is also related to the traditional teaching models of English writing. In general, education of our university is to append significance to the investigation of information and detest the yield of information. To a large extent, the process of learning is the process of accumulation and absorption of existing knowledge. The teaching form mainly adopts the method of teacher teaching and students' extracurricular practice to consolidate the knowledge acquired in the course. From every point of development of essentials of language abilities, the educating of "tuning in" and "perusing" can frequently be instructed by educators in enormous classes and understudies' extracurricular practice. However, the educating of "talking" and "stating" has higher prerequisites for the cooperation among understudies and instructors, and it is hard to execute in bigger classes. In the educating of English writing in China, there is a typical issue that educators are not ready to "instruct" and understudies would prefer not to "practice". From the perspective of teachers, many language rules can't be mastered by students through classroom teaching, which can only be achieved through students' extensive use of English language. These uses not only refer to the "writing" itself, but also the "listening", "speaking" and "reading". From the perspective of students, because writing involves both language and content, students have difficulties in language expression and lack of timely feedback. If students do not receive timely and targeted feedback, it will further frustrate their enthusiasm for improving their English writing ability [16–18].

3. The application and research of data mining knowledge in college English writing teaching

When all is said in done, the capacity of information mining is to locate the relating design type as per the target the chance to be mined. As a rule, the undertaking of information mining is separated into two classes: enlightening data mining and prescient information mining. Enlightening information mining task is to describe the general characteristics and properties of data information. The predictive data mining task is to analyze and infer the current data, find the rules, generate the rules, establish the model, and then use the established model to predict the new data. This paper applies the first function in data mining, that is, descriptive data mining. The algorithm used is the Apriori algorithm in the association rules. This paper uses association rules to analyze the factors related to English writing, assisting English teaching and enhancing the teaching quality.

3.1. Algorithm analysis

Association rule mining involves early data excavating methods in data mining research, and so far, researchers are still exploring. Association rule mining is to find hidden associations or relationships between data objects from large-scale data sets, also known as association analysis or association rules learning. The most commonly used association algorithm rule is the Apriori algorithm and the association rules are also commonly used. For example, in the process of learning English, we use association rule mining algorithm to find out some correlations between students' learning contents, and between each knowledge point or between English listening, speaking, reading and writing skills, so as input the guidance and help for learning students and teachers' teaching.

Association rule algorithm is a common algorithm used in data mining, which has two measures of support level and confidence level. At the point when the affiliation rules fulfill both the base certainty and the base help, it is thought that the association rules are interesting.

Expression of association rule $A \Rightarrow B$,

$$A \subset I, B \subset I \tag{1}$$

Expression of support:

$$\sup port(A \Rightarrow B) = P(A \cup B) \tag{2}$$

Expression of confidence:

$$confidence(A \Rightarrow B) = P(B|A) \tag{3}$$

The simplified of confidence:

$$confidence(A \Rightarrow B) = P(B|A) = \frac{\sup port - count(A \cup B)}{\sup port - count(A)} \tag{4}$$

The Apriori algorithm is an original algorithm produced by two scientists of R. Agrawal and R. Srikant specially for excavating frequent association rules item set. It is a more classical algorithm in association rules. The algorithm is easy to operate and easy to understand. The representation diagram of the Apriori algorithm is shown in Fig. 1.

The stages of the Apriori algorithm are mainly divided into two shares: intensity and frequency. The frequency is to excavate all the frequent items according to the minimum support. The transaction A contains the K element, which is called the K entry set. It is called a recurrent K set when it satisfies the minimum support degree simultaneously. So not every item set is a K item set, there are two steps in finding frequent item sets: self – assembly and clipping, the purpose of self – joining is to ensure that the previous K-2 items are the same and are sorted in the dictionary order. The purpose of pruning is to remove the missing items, ensuring that all non-empty subsets of any frequent item are frequent; Intensity is a strong association rule based on minimum confidence. By repeating these two steps, strong association rules are finally obtained.

3.2. The submission of association rules in English writing teaching

This paper mainly covers the design of two parts of database storage data, data mining and data analysis based on association rules algorithm. The database storage data is the data source of this paper as well as the basic content. The application of association rules algorithm is the core content of this paper and directly related to decision maker's decision. Finally, the research applications of association rules algorithm in the system are completed, and structure of system, diagram is exposed in Fig. 2.

Data storage based on MySql database

In this paper, the Mysql database is used to preserve the students' English writing performance in real time. Mysql simpler database to operate and calmer to handle than different databases. And we can operate the Mysql database directly with the help of the Nevicat for Mysql tool. The data is presented in the form of a table directly, which ensures the persistence of the data. But in the process of storing data to the database, due to the limited capacity of Mysql database, a large number of data can also cause the database to crash, so set the storage cycle of the data. The data storage flowchart is shown in Fig. 3.

Data cleaning is the primary step and the most important step in the procedure of processing data.

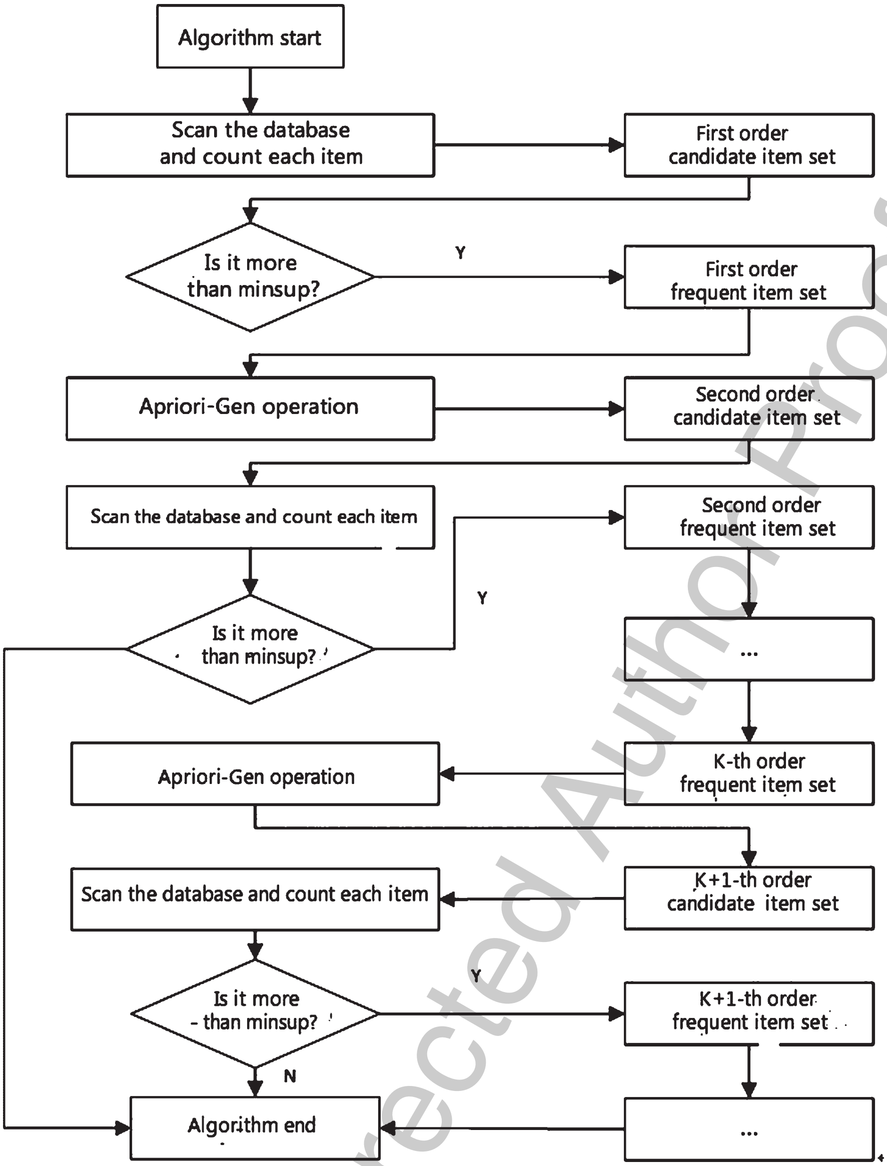
Algorithm start
Scan the database and count each item
First order candidate item set
Is it more than minsup?
Y
First order frequent item set
Apriori-Gen operation
Second order candidate item set
Scan the database and count each item
Second order frequent item set
Y
Is it more than minsup?
...
Apriori-Gen operation
K-th order frequent item set
Scan the database and count each item
K+1-th order candidate item set
Is it more than minsup?
Y
K+1-th order frequent item set
N
Algorithm end
...

Fig. 1. The schematic diagram of the Apriori algorithm.

In the original research data, there are missing values in some fields, and the main methods to deal with the missing values are as follows: (1) the missing value is filled with special values by manual filling. (2) A continuous feature that uses the average value of the feature to fill the missing value. (3) Such samples can be discarded directly in cases where the eigenvalues are not too important or not essential. (4) The average values of similar samples are used to fill the missing values. (5) The algorithm of machine learning predict model the missing value. In this experiment, for example, “reading ability”, these missing values are filled in by means of the average value method. For example, Table 1 is part of the data.

Data mining based on Apriori

The original data sheet of English writing. The key factors in this research were reading ability, vocabulary knowledge, grasp the textual knowledge and listening ability, and analyzed only 80–100 points, which is the data of the superior level. A representative sample of nine students was selected for analysis, as shown in Table 2.

The minimum support frequency of this paper is 2, and the first frequent item usual table is exposed in

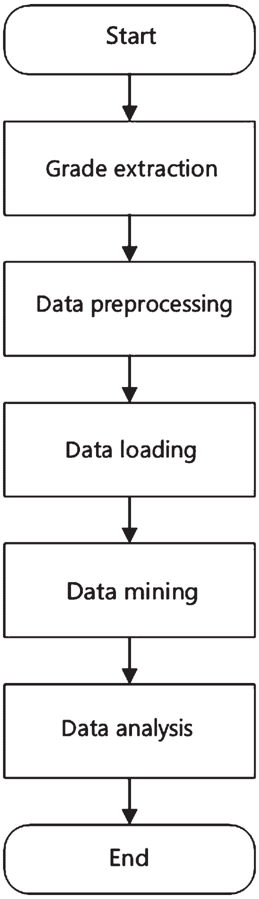

Fig. 2. The system structure diagram.

Table 3, and the second item usual table is exposed in Table 4. According to the minimum support degree principle, the item sets which are not conforming to the conditions are deleted, and the second recurrent item sets are obtained, as shown in Table 5. Third item set table as shown in Table 6, according to the principle, the third frequent item set table are obtained, as shown in Table 7. The fourth item set table is shown in Table 8. Since the support degree is less than 2, the frequent item sets will not be found, and the algorithm is over.

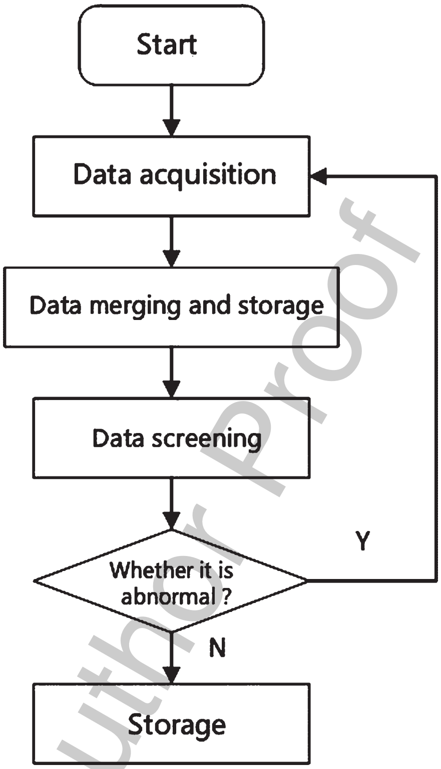

Fig. 3. The data storage process.

3.3. Experiment and analysis

Through the algorithm, the final set of item group is $\{I_1, I_2, I_3\}$, $\{I_1, I_2, I_5\}$. Therefore, the association rules in college students' English writing are: writing ability, reading ability, vocabulary knowledge and

Table 1
Part of the data

Score people	Writing ability	Reading knowledge	Vocabulary ability	Listening knowledge.	Grasp the textual	…	…	… expressive ability	Language
1	80	88	70	75	80	…	…	..	82
2	87	80	90	86	80	…	…	..	91
3	67	59	48	52	61	…	…	..	81
4	83					…	…	…	79
…	…	…	…	…	…	…	…	…	…
..	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…
1000	90	85	87.5	88	78	…	…	…	89

Table 2
The original data

Id	I1 (writing)	I2 (reading ability)	I3 (vocabulary knowledge)	I4 (grasp the textual knowledge)	I5 (listening ability)
1	√	√			√
2		√		√	
3		√	√		
4	√	√		√	
5	√		√		
6		√	√		
7	√		√		
8	√	√	√		√
9	√	√	√		

Table 3
The first frequent item set table

Frequent item sets	Support count
$\{I_1\}$	6
$\{I_2\}$	7
$\{I_3\}$	6
$\{I_4\}$	2
$\{I_5\}$	2

Table 4
Second item set table

Frequent item sets	Support count
$\{I_1, I_2\}$	4
$\{I_1, I_3\}$	4
$\{I_1, I_4\}$	1
$\{I_1, I_5\}$	2
$\{I_2, I_3\}$	4
$\{I_2, I_4\}$	2
$\{I_2, I_5\}$	2
$\{I_3, I_4\}$	0
$\{I_3, I_5\}$	1
$\{I_4, I_5\}$	0

Table 5
The second frequent item set table

Frequent item sets	Support count
$\{I_1, I_2\}$	4
$\{I_1, I_3\}$	4
$\{I_1, I_5\}$	2
$\{I_2, I_3\}$	4
$\{I_2, I_4\}$	2
$\{I_2, I_5\}$	2

Table 6
Third item set table

Frequent item sets	Support count
$\{I_1, I_2, I_3\}$	2
$\{I_1, I_2, I_5\}$	2
$\{I_1, I_3, I_5\}$	1
$\{I_2, I_3, I_4\}$	0
$\{I_2, I_3, I_5\}$	1
$\{I_2, I_4, I_5\}$	0

Table 7
The third frequent item set table

Frequent item sets	Support count
$\{I_1, I_2, I_3\}$	2
$\{I_1, I_2, I_5\}$	2

Table 8
The fourth item set table

Frequent item sets	Support count
$\{I_1, I_2, I_3, I_5\}$	1

writing ability, reading ability and listening ability. It can be seen from the rules that the key factors affecting the writing ability are reading ability, vocabulary knowledge and listening. Through the Apriori algorithm to excavate the factors that the teacher can't intuitively judge, especially the influence of English listening on English writing cannot be ignored. Therefore, in the course of teaching English writing, it should provide consideration amount of students' reading, the accumulation of vocabulary and the exercise of listening.

4. Conclusion

Broad enhancement in field of internet, "Internet plus education" has become a trend. College English teaching makes students make great progress in all aspects of English reading, listening, writing and speaking. After years of accumulation, a great number of valuable education data are stored in the system, and these source data provide a basis for mining work. Through deep study of mining and other relevant theories and technologies, this paper analyzes the influencing factors in college English writing and provides suggestions for college English writing teaching. Therefore, this research has important research value for the advancement of College English composing instructing, and to a limited degree, it advances the degree of English educating and the capacity of information mining in China.

Acknowledgments

This work was supported by Educational Science Planning Project of Jilin Province "Research on the dilemma and countermeasures of college students' English vocabulary mobile learning"(GH19146); The Higher Education Scientific Research Project of Jilin Province in 2019(Health Education for Teachers and Students): Research on the Status Situation, Causes and Intervention Paths of Job Burnout of General English Teachers in Colleges and Universities of Jilin Province (Grants No. JGJXZ2019C9); This research was supported by 2019 Jilin Social Science Fund Project: Jilin Province Traditional Chinese medicine cultural identity survey and Promotion Path Research (Grant No. 2019B89); "Apricot Woods Scholar Project" 2019 – Young Scientist Project in Changchun University of Chinese Medicine.

References

[1] A. Kumari, S. Tanwar, S. Tyagi, N. Kumar, M. Maasberg and K.K.R. Choo, Multimedia big data computing and Internet of Things applications: A taxonomy and process model [J], *Journal of Network and Computer Applications* **124**, 169–195.

[2] C. Angeli, S. Howard, J. Ma, et al. Data mining in educational technology classroom research: Can it make a contribution?[J], *Computers & Education*, 2017.

[3] J. Cai, Challenges to Traditional College English Teaching Concepts: A Study of College English Teaching Guidelines[J], *Foreign Language Education*, 2017.

[4] D. Zhang, D. Zhang, H. Xiong, C.-H. Hsu and A.V. Vasilakos, BASA: Building mobile Ad-Hoc social networks on top of android [J], *IEEE Network* **28**(1), 4–9.

[5] J.Y. Feng, The application of intercultural communication in college English teaching[J], *Journal of Jiamusi Vocational Institute*, 2017.

[6] M. Khari, A.K. Garg, R. Gonzalez-Crespo and E. Verdú, Gesture Recognition of RGB and RGB-D Static Images Using Convolutional Neural Networks [J], *International Journal of Interactive Multimedia and Artificial Intelligence* **5** (2019),22–27.

[7] M. Khari, A.K. Garg, R. Gonzalez-Crespo and E. Verdú, Gesture Recognition of RGB and RGB-D Static Images Using Convolutional Neural Networks, *International Journal of Interactive Multimedia and Artificial Intelligence* **5** (2019), 22–27.

[8] V. Khosravi, F.D. Ardejani, S. Yousefi, et al. Monitoring soil lead and zinc contents via combination of spectroscopy with extreme learning machine and other data mining methods[J], *Geoderma* **318** (2018), 29–41.

[9] T. Li, Research on College English Classroom Teaching in the Perspective of Educational Ecology[J], *Theory & Practice of Education*, 2017.

[10] M. Elhoseny, A. Shehab and X. Yuan, Optimizing Robot Path in Dynamic Environments Using Genetic Algorithm and Bezier Curve [J], *Journal of Intelligent & Fuzzy Systems* **33**(4) (2017), 2305–2316.

[11] N. Yuri, G.-D. Vicente, M. Carlos and G.C. Rubén, Supporting academic decision making at higher educational institutions using machine learning-based algorithms [J], *Soft Computing* **23** (2019), 4145–4153.

[12] R. Nisbet, G. Miner and K. Yale, Chapter 5 – Feature Selection[J], *Handbook of Statistical Analysis & Data Mining Applications*, 2018, 83–97.

[13] A. Prathik, J. Anuradha and K. Uma, Survey on Spatial Data Mining, Challenges and Its Applications [J], *Journal of Computational and Theoretical Nanoscience* **15**(9-10) (2018), 2769–2776.

[14] A. Sepúlveda, A New Application for Data Mining and Analytics[J], 2017.

[15] C. Silva and J. Fonseca, Educational Data Mining: A Literature Review[J], 2017.

[16] A. Tharwat, H. Mahdi, M. Elhoseny and A.E. Hassanien, Recognizing human activity in mobile crowdsensing environment using optimized k-NN algorithm [J], *Expert Systems with Applications* **107** (2018), 32–44.

[17] M. Tiwari, Strategic decision making: a data mining approach[J], 2017.

[18] J. Zhao, G.H. Liu and L. Yang, Basic Connotation of College English Teachers' Professional Development and Promotion Strategies[J], *Journal of Northeast Normal University*, 2018.