# Analysis of college oral English test scores based on data mining technology

**Yuanyuan Tai**

Heze University
Heze, Shandong, People's Republic of China

ABSTRACT: Data mining is a process of extracting information or knowledge, which cannot be known explicitly and has potential significance, from large quantities of data through computer and database technology. Data mining is able to extract significant data from mass data. It is a reliable tool and can resolve the awkward problem of having *rich data but poor knowledge*. Reported in this article is the use of data mining technology to analyse test scores of spoken English using cluster and correlation analysis of daily use of English to determine the relevant factors for students' oral English scores. The results show that English study requires diligent training and, in particular, frequent training in spoken English. Students should take the initiative in using English. Teachers should make opportunities and conditions for students to speak up in the classroom as much as possible.

## INTRODUCTION

Data mining (DM) derived from KDD (Knowledge Discovery in Database) was coined at the 11th International Joint Conference on Artificial Intelligence in 1989 [1]. Data mining is oriented towards extracting intelligence from information [2]. Generally speaking, it applies computer technology and database technology to extract information or knowledge, which is not explicit, from large quantities of data and from incomplete data [3].

English is the first language of international exchange. Given China's international development, English study has become especially important [4]. For a long time, China's examination-oriented education has neglected students' ability to apply learnt knowledge [5]. In the tests for undergraduate English competence, the most common CET-4 and CET-6, do not test reading and speaking [6]. Especially for oral English, autonomous behaviour is often the basis of the college English ability test [7]. These oral language tests are based on a student's speech and dialogue. Scoring is mainly based on testers' subjective judgment [8]. Thus, a student's oral English competence cannot be assessed frequently, let alone lead to rational and effective training methods to improve oral English through daily practice.

DM has been successfully applied by numerous domestic researchers in marketing management, insurance, finance, medical and health care, etc [9]. It has superiority for mining relations and determining rules of the noumena [10]. Data mining can analyse data from different perspectives, to find patterns and information, which cannot be found by other data analysis methods and allow perfected or improved strategies to deal with this [9].

The enhancement of students' oral English scores cannot be separated from the training and assessment methods used to measure the level and improvement of students' oral English. Data mining was used in this study to analyse students' oral English scores using an improved DM algorithm. Some important information is reported in this study, related to students' oral English.

## ANALYSIS OF STUDENTS' ORAL SCORES WITH DM

The DM process to analyse students' oral English scores is shown in Figure 1. First, construct a student information warehouse; pre-process the data in the warehouse; apply the improved DM algorithm to mine and analyse the preprocessed data; finally, gain relevant knowledge.

### Data Collection and Integration

Data analysis uses large quantities of data. On the basis of collecting and analysing large quantities of information about low-grade students, the following data are gained through data integration.

*Students' Basic Information Data*

Students' basic information data contain students' personal information and scores for each subject before they go to college. To some extent, college entrance examination score information can well reflect students' personal learning habits and confirm the level of students' scores compared to all students. A sample of students' basic information, after integration, is shown in Table 1.

Table 1: Table of students' basic information.

| Student number | Name | Gender | Chinese score | Math score | English score |
|---|---|---|---|---|---|
| **001 | Ding Hui | Female | 117 | 109 | 107 |
| **002 | Zhang Di | Female | 109 | 93 | 93 |
| **003 | Cao Meng | Female | 111 | 120 | 108 |
| **004 | Ma Xiao | Male | 97 | 132 | 110 |
| **005 | Liu Jie | Male | 117 | 105 | 109 |
| …… | …… | …… | …… | …… | …… |

*Students' College Score Data*

Students' college score data were obtained from the college. Student majors, course duration and nature of the course differ. The scores selected exclude professional courses and mainly include the scores of public courses. The scores are out of 100. The scores gained through rating systems are transformed to a range 0 to 100. The data sheet including students' college scores as set up, is shown in Table 2.

Table 2: Students' score table.

| Students number | Course number | Course score |
|---|---|---|
| **001 | 0001 | 89 |
| **001 | 0002 | 94 |
| **001 | 0003 | 90 |
| **002 | 0001 | 81 |
| **002 | 0002 | 77 |
| …… | …… | …… |

*Students' Comprehensive English Competence Table*

The college cultivates competency in English using English study software. Every student is tested in English listening comprehension, reading ability, oral English ability, vocabulary and writing ability. Students' English competence table is set up according to the test results, as shown in Table 3.

Table 3: Comprehensive English competence table.

| Student number | Oral English | Listening | Reading | Writing | Vocabulary |
|---|---|---|---|---|---|
| **001 | 89 | 85 | 80 | 96 | 82 |
| **002 | 94 | 89 | 85 | 81 | 87 |
| **003 | 90 | 86 | 81 | 87 | 93 |
| …… | …… | …… | …… | …… | …… |

*Table for Students' English Study Attitude*

The main scores are students' oral English scores. Some research results indicate that the students' oral English is related to practising in their spare time and participation in club activities apart from the classroom. To gain relevant data, quantification of attitudes to English learning was carried out using teachers' evaluation, students' personal evaluation and the evaluation from English-related clubs. Finally, the learning attitude table was developed and includes the following: class attendance, seat selection in class, handing-in school assignment, class statement, after-class learning time and learning time in clubs.

Data Pre-processing

Data preprocessing is an important process prior to DM and mainly involves processing the integrated data required by DM [10]. These processes mainly include data cleaning, data conversion, and processing of incomplete and inconsistent data, including noise [11].

*Data Conversion*

Some data may not be data statistics and are inappropriate for DM. Thus, conversion is required [12]. This processing mainly concerns the normalised conversion of non-data-type data.

1. Conversion of students' examination scores: college entrance examination scores in the database mainly cover the scores for three subjects. Many provinces autonomously mark examination papers in English, mathematics and Chinese, and a consistent conversion to a standard mark is required, e.g. to convert a 150-mark system to a 100-mark system uses the following conversion formula:

$$M = (150 - K) / (150 - L) * 100 \tag{1}$$

Where, M is the score after conversion; K is the highest score in the database; L is the original score.

2. Quantification of students' seat selection: a student's seat selection, to a large extent, reflects a student's learning attitude. Seat selection is descriptive and quantification  is required. The quantification formula is as follows:

$$L = (\sum_{i=1}^{n} N_i / (k_1 * 1 + k_2 * 0.7 + k_3 * 0.3) * 100) / n \tag{2}$$

Where, L is the score for seat selection; n is the number of seats; $k_1$ means seats selected are in the first 5 rows; $k_2$ in rows 5-10; and $k_3$ means in the back row.

3. Learning time: after-class and team learning time were quantified using a 0 to 100 scale. Quantification assumes 2-hour English learning equals a score of 100. The quantification formula is as follows:

$$L = \left( \begin{array}{lr} 100 & 2<\mathrm{n} \\ 100 * \mathrm{n}/2 & 1.5<\mathrm{n} \leq 2 \\ 70 * \mathrm{n}/1.5 & 0.5<\mathrm{n} \leq 1.5 \\ 60 * \mathrm{n}/0.5 & \mathrm{n} \leq 0.5 \end{array} \right) \tag{3}$$

Where, L is the score for learning time; n is time (hours).

*Data Cleaning*

There are usually data either incomplete or unsuitable for statistics. A manual filling method was adopted to include average values for the same kind of samples. Some special or odd data may be present because of statistical error. These data can be eliminated, deleted or changed so that, finally, all data are complete and comprehensive.

*Data Discretisation*

Since data in a hundred-mark system cannot be compared according to the grade, discretisation was conducted to evaluate special scores. After discretisation, data are classified into five grades. The discretisation table is shown in Table 4.

Table 4: Discretisation table.

| Score section | | 0-40 | 40-60 | 60-75 | 75-85 | 85-100 |
|---|---|---|---|---|---|---|
| Grade | Test score A | A1 | A2 | A3 | A4 | A5 |
| | School assignment B | B1 | B2 | B3 | B4 | B5 |
| | Attendance C | C1 | C2 | C3 | C4 | C5 |
| | Class statement D | D1 | D2 | D3 | D4 | D5 |
| | After-class learning time E | E1 | E2 | E3 | E4 | E5 |
| | Club learning time F | F1 | F2 | F3 | F4 | F5 |

Since the test scores involve multiple courses and various aspects, in the analysis, the scores of the following courses were selected for use in data mining. The courses selected and the number of scores are shown in Table 5.

Table 5: Course and number of scores.

| Course | No. | Course | No. |
|---|---|---|---|
| College English | 11 | Oral English | 21 |
| College Chinese | 12 | Listening comprehension | 22 |
| Mandarin | 13 | Reading | 23 |
| Advanced mathematics | 14 | Writing | 24 |
| College physics | 15 | Vocabulary | 25 |
| PE | 16 | | |

Data Mining (DM)

The DM algorithm: DM algorithms include decision, cluster, regression analysis and neural network algorithms, etc. The Apriori algorithm put forward by Agrawal et al [5] is a commonly used algorithm for data relation analysis. The Apriori algorithm is an algorithm to mine frequent item sets using Boolean Association Rules. Its main thought is to apply prior knowledge of frequent item sets to find all the frequent item sets through an iterative method with layer-by-layer searching.

The main process involves adding up items to determine the quantity of each item by scanning the database to find out the first frequent item set L1 with the minimum support degree; then, seek the second frequent item set L2 according to L1; then, apply L2 to seek frequent item set L3 meeting the minimum support degree until no frequent items can be found [6][7].

In this algorithm, the inputs are the transactional database D and minimum support degree minsup; the output is the frequent item set L.

1. Scan to gain 1- frequent item set:

$$L_1 = \text{quent\_frequent\_1-itemsets(D)} \tag{4}$$

2. Search new item candidate set $C_k$ with $L_l$:

$$C_K = apriori\_gen(L_{K-1}, \min\_\sup) \quad (k \geq 2, L_{K-1} \neq \varnothing) \tag{5}$$

3. Scan all transactions t in the business item and obtain the candidate set containing transaction T:

$$C_t = sutset(C_K, t) \quad (t \in D) \tag{6}$$

4. Obtain the frequent sets meeting requirements in the candidate set:

$$L_K = \{c \in C_K\} \tag{7}$$

Through the above data scanning mining, finally return $\cup_K L_K$ .

Data mining algorithm improvement: Relevant information can be gained through the application of the above algorithms, but this method will scan the database whenever a candidate item set is produced. The database used for DM may be very large, and the algorithm has low efficiency. In practice, a hash function was used to optimise the scan of the database. A hash function can reduce the quantity of candidate 2- item sets included in the second cycle $C_Z$ and, thus, effectively and rapidly produce the candidate set and so improve efficiency. The execution process is as follows:

1. Confirm transactional database to be mined: select other scores related to oral English score from the data collected and pre-processed as pre-selection database. This data transaction segment to be mined is shown in Table 6. The minimum support degree is set to 50%:

Table 6: Data transactions to be mined.

| Transaction No. | Item list |
|---|---|
| A | 21, 12, 13, 14 |
| B | 21, 14, 15, 16 |
| C | 21, 22, 23, 24 |
| D | 21, 13, 14 |
| E | 21, 22, 25 |
| F | 21, 12, 13, 25 |

2. Confirm the hash function: when hash technology is used to improve DM, the hash function should be confirmed prior to mining. Assuming candidate 1- item set in $C_l$ produces frequent 1- item set L, the hash function for all 2-item sets is as follows:

$$\text{h(x，y)=(10} * \text{(order of x)+(order of y))mod 7} \tag{8}$$

Where, (order of x) is the order item of X; (order of y) is the order item of y; is the value of item list in the transaction table of the order item set. If an item set of 2- is {11,13}, its hash function is as follows:

$$h(x，y)=(10*11+13))\bmod 7=4$$

Then, the hash table of Table 6 is shown in Table 7.

Table 7: Hash number of transaction items.

| Transaction No. | Hash number |
|---|---|
| A | h(21,12), h(21,13), h(21,14), h(12,13), h(12,14), h(13,14) |
| B | h(21,14), h(21,15), h(21,16), h(14,15), h(14,16), h(15,16) |
| C | h(21,22), h(21,23), h(21,24), h(22,23), h(22,24), h(23,24) |
| D | h(21,13), h(21,14), h(13,14) |
| E | h(21,22), h(21,25), h(22,25) |
| F | h(21,12), h(21,13), h(21,25), h(12,13), h(12,25), h(13,25) |

3. Execute mining analysis: take the hash number gained as the address list for DM. The association rule table finally obtained by taking the occurrence times of each transaction in the Hash Table and in combination of A1-A5 is as follows:

Table 8: Oral English association result.

| Rule No. | Item list |
|---|---|
| 1 | A521, A513 |
| 2 | A421, A521, A325 |
| 3 | A221, A125, A112 |
| 4 | A523, A524, A525 |

The above rules show, taking the test scores and grades as an example, students with good oral English performance and an outstanding Chinese vocabulary have excellent listening comprehension and reading ability; students with good oral English performance do not always have an outstanding vocabulary. So, vocabulary is not an important influencing factor for oral English, but a large vocabulary is largely associated with good reading ability and writing ability.

Mining analysis of all data. Based on the above improved algorithm, DM analysis was conducted for all data shown in Table 4. The minimum confidence coefficient of DM is 0.2 and the minimum support degree is 50%. The whole DM process is: based on the confirmed database, connect every two row items in the database and store the results in the auxiliary candidate set $C_k$; select the data stored and compare them with the existing data in the candidate set. The comparison algorithm is:

$$l_K = (l_K = l_i (i \in (1, k-1)) ? l_K \in C_k, l_1) \tag{9}$$

That is, there are K-1 elements in $C_k$ and there is an element equal to $l_k$. Then, elements of $C_K$ will add to 1, or else, $l_k$ serves as the first element and is directly stored in $C_k$. Then, confirm the hash function of addressing and conduct hash address marking for all elements stored; then, take the minimum support degree as the constraint to scan all hash addresses, judge relevance confidence coefficient among elements and, finally, generate the association rule table (remove the association results shown in Table 8) as follows:

Table 9: Association results of the whole database.

| Item | Rule | Support degree |
|---|---|---|
| 1 | A521, F5 | 98 |
| 2 | A5, B5 | 56 |
| 3 | A121, D1, E1 | 75 |

*Rule 1:* oral English score has strong association with club activities. Those with excellent oral English scores spend much time participating in club activities.

*Rule 2:* in the current educational system, handing in a school assignment largely decides students' test scores. This is related to the grading system where the final score is equal to the ordinary performance plus the test score.

*Rule 3:* students with poor oral English performance never take the initiative to make statements in class and spend less time on after-class study. This indicates students with poor academic records have a bad learning attitude and poor diligence in learning.

## DISCUSSION OF RESULTS

Improvement of oral English scores is the major objective of practical English education and quality-oriented education. The above results show oral English is somewhat related to Mandarin. This indicates the students with language talents usually have good oral English performance. Besides, students with good oral English are slightly better in listening comprehension.

Meanwhile, this analysis also shows that students with excellent vocabulary have strong writing and reading ability. At the same time, DM was further applied to the analysis of students' daily learning situations and oral English performance. The analysis shows oral English to a large extent comes from students' diligence and efforts in class and after class. Especially, extracurricular training plays an important role in improving oral English. Those who do not take the initiative to make a statement in class and do not study after class have poor oral English scores.

The above analysis suggests strengthening extracurricular activities and especially encouraging students to participate in clubs or training related using oral English. Also students should be encouraged to speak so as to avoid poor English and improve students' oral English performance.

## CONCLUSIONS

This article was based on an analysis of DM and its algorithm took college oral English as the object of study, improved the DM efficiency by improving the existing DM association algorithm and by introducing the concept of a hash function. The improved DM algorithm was used to analyse the correlation between students' Chinese Mandarin, English reading, listening, speaking, writing and oral English scores. As well, the relation was studied between students' daily study and oral English scores.

The conclusions show that English study is a process of diligent practice and frequent speaking and it is necessary for students to show initiative. Besides, teachers should make opportunities and conditions to encourage students to speak in the classroom as much as possible.

## REFERENCES

1. Hamilton, H.J., Geng, L., Findlater, L. and Randall, D.J., Efficient spatio-temporal data mining with GenSpace graphs. *J. of Applied Logic*, 4, **2**, 192-214 (2006).
2. Gibert, K., Spate, J., Sànchez-Marrè, M., Athanasiadis, I.N. and J. Comas, J., *Data Mining for Environmental Systems*. In: Developments in Integrated Environmental Assessment, 3, **12**, 205-228 (2008).
3. Li, X., A new clustering segmentation algorithm of 3D medical data field based on data mining. *J. of Digital Content Technol. and its Applications*, 4, **4**, 174-181 (2003).
4. Vanci-Osam, U. and Aksit, T., Do intentions and perceptions always meet? A case study regarding the use of a teacher appraisal scheme in an English language teaching environment. *Teaching and Teacher Educ.*, 16, **2**, 255-267 (2000).
5. Borg, S., Research engagement in English language teaching. *Teaching and Teacher Educ.*, 23, **5**, 731-747 (2007).
6. Leong, A.M.W. and Li, J.X., A study on English teaching improvement based on stakeholders' needs and wants: the case of the Faculty of International Tourism of the Macau University of Science and Technology (MUST). *J. of Hospitality, Leisure, Sport & Tourism Educ.*, 11, **1**, 67-78 (2012).
7. Hayes, D., Non-native English-speaking teachers. *Context and English Language Teaching System*, 37, **1**, 1-11 (2009).
8. Fareh, S., Challenges of teaching English in the Arab world: why can't EFL programs deliver as expected? *Procedia - Social and Behavioral Sciences*, 2, **2**, 3600-3604 (2010).
9. Xiaoan, B., Research on quick sort methods based on ID3 algorithm. *Modern Electronic Technique*, 27, **4**, 84-85 (2004).
10. Han, J., Nishio, S., Kawano, H. and Wang, W., Generalization-based data mining in object-oriented databases using an object cube model. *Data & Knowledge Engng.*, 25, **1-2**, 55-97 (1998).
11. Zhigang, Y. and Xu, J., Research on data preprocess in data mining and its application. *Application Research of Computers*, 21, **6**, 117-118 (2004).
12. Wang, Y. and Zheng, L., Endocrine hormones association rules mining based on improved Apriori Algorithm. *J. of Convergence Infor. Technol.*, 7, **7**, 72-82 (2012).