International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 6669-6681

https://doi.org/10.70517/ijhsa463573

A Study on the Application of Multidimensional Data Analysis Model in Assessing the Teaching Effectiveness of Medical English

Di Wu[1,*]
[1] Public Foreign Language Teaching and Research Department, Heilongjiang University of Chinese Medicine, Harbin, Heilongjiang, 150040, China
Corresponding authors: (e-mail: Frankwudi@163.com).

Abstract Medical English teaching has problems such as language and content detachment and insufficient cultivation of practical application ability, which require innovative assessment methods. This study adopts clustering-based multidimensional data processing technology to construct CSSAQP algorithm to analyze and evaluate medical English teaching data. The algorithm deals with the extreme value problem through K-Means clustering, and adopts a two-phase strategy: the pre-constructed sample phase and the query execution phase for stratified sampling and precise analysis of teaching data. The experimental results show that on the medical English vocabulary teaching dataset (about 0.5 billion data), the query accuracy of the CSSAQP algorithm reaches 0.0122%, 0.0141%, and 0.0085% for the SUM, COUNT, and AVG metrics, respectively, which is better than the existing methods. Meanwhile, the algorithm excels in query response time, with SUM, COUNT, and AVG query times of 1.52 seconds, 1.24 seconds, and 1.34 seconds respectively, realizing real-time response. The research results provide an accurate assessment tool for medical English teaching, help optimize the structure of medical English courses and the construction of teaching resources, and provide data support for cultivating medical talents with a global perspective.

Index Terms medical English teaching, multidimensional data analysis, K-Means clustering, CSSAQP algorithm, teaching effect assessment, precise query

I. Introduction

The English for Specialized Purposes course, which is centered on students' professional and future career needs, is the development trend of university English teaching in the future and the key direction of university English teaching reform [1], [2]. Especially for medical students studying Western medicine, learning English well will be beneficial for students to learn subject knowledge from English originals and keep abreast of the latest scientific and technological achievements, broaden their professional horizons, and have the opportunity to engage in academic exchanges and cooperation with internationally renowned experts to obtain broader opportunities for medical development, as well as improve their scientific research, academic competitiveness, and competitiveness in the workplace [3]-[7]. Strengthening English teaching in medical education is very important, aiming to cultivate modern medical talents with comprehensive quality and international communication ability [8]. Therefore, medical English shoulders the important task of cultivating students' ability to effectively use English for academic communication, scientific research, medical treatment and related activities.

Nowadays, most higher medical schools and colleges have set up specialized medical English courses and have achieved good results. However, there are obvious shortcomings in the existing teaching assessment structure, which hinder the further development of medical English. The main manifestation is that the assessment is dominated by a single dimension, such as vocabulary, and lacks the assessment of clinical dialogues, case filling, academic communication and other scenarios. The integration and utilization of student-related data applied in the assessment system is low, and the assessment results are not comprehensive. The feedback cycle of paper-based test results is long, the time cost is large, and the timely feedback mechanism is missing [9]-[13]. Multidimensional data analysis is an important technology in the field of modern data analysis, which is characterized by multidimensionality, unstructuredness, high dimensionality and large scale. The multidimensional data analysis model, which combines data analysis and processing methods, can not only help us better understand the relationship between the data and visualize them, but also analyze the data across geography, time, events, behaviors, and other aspects to reveal the laws and trends hidden behind the data, and then carry out targeted optimization, making it possible to solve problems more accurately and efficiently [14]-[17].

Medical English, as an important course for cultivating the international communication ability of medical talents, is facing unprecedented challenges and opportunities. The globalized medical environment has put forward higher requirements for the professional quality of medical personnel, and the international development of the medical industry has made medical English proficiency a necessary quality for medical talents. The current situation of medical English teaching is not optimistic, and there is a disconnect between teaching content and practical needs. Most of the teachers in charge of medical English courses are English majors and lack medical professional background; some of them are foreign students majoring in medicine, who have medical knowledge but limited teaching experience and methods. Teachers generally follow the teaching mode of university English, mainly focusing on vocabulary learning and translation exercises, resulting in a serious disconnection between language learning and medical content. This teaching method is difficult to cultivate students' English application ability in medical contexts, and has limited effect on the improvement of medical professional ability, which fails to meet the actual needs of medical personnel training. The design of medical English course structure is not scientific enough, the content of the course lacks systematicity and coherence, the orientation of teaching objectives is vague, the construction of teaching resources is insufficient, and there is a single method for assessing the teaching effect, which makes it difficult to objectively reflect the students' learning results and ability enhancement. These problems seriously restrict the improvement of medical English teaching quality and the cultivation of international competitiveness of medical talents. Traditional teaching assessment methods cannot effectively deal with the complex data problems in medical English teaching, especially the multidimensional, heterogeneous and large-scale characteristics of students' learning effect data, which makes the teaching assessment results lack of precision and real-time. Facing the complexity of medical English teaching assessment and the challenges of data processing, the introduction of multidimensional data analysis technology has become an innovative solution. Multidimensional data analysis technology has the advantages of dealing with complex data structures, coping with massive information, and providing accurate assessment results, which is suitable for dealing with multivariate data in medical English teaching effectiveness assessment. By introducing the K-Means clustering algorithm and clustering-based multidimensional data query processing technology, the problem of extreme values in the assessment of medical English teaching effectiveness can be solved and the accuracy and efficiency of the assessment can be improved. The study proposes the Clustering plus Selective Estimation-based Stratified Sampling Approximate Query Processing (CSSAQP) algorithm, which handles massive medical English teaching data through a two-phase strategy: the pre-constructed sample phase utilizes the K-means clustering algorithm to divide the data into three clusters, and then performs stratified sampling; and the query execution phase provides accurate assessment results through query parsing and sample matching. In addition, the study constructed a complete structure system of specialized medical English courses, including three modules of general medical English, specialized medical English and academic medical English, and developed a medical English vocabulary teaching system based on digital resources. Through experiments, we verify the application value of CSSAQP algorithm in medical English teaching effect assessment, comparatively analyze its advantages in query precision and response time, provide data support for medical English teaching assessment, and promote the quality improvement of medical English teaching.

II. Building the development of medical English language teaching

II. A. Problems in the Characterization and Practical Application of Medical English

As an English language course, Medical English has its own unique field of application. The teaching content of the course includes both theoretical medical knowledge and practical terms. Its teaching goal is to cultivate students' ability to translate Chinese and English in medical-related fields, and at the same time, it is necessary to improve students' practical application of English, which can meet the needs of medical personnel in their practical work [18]-[20].

In the process of teaching medical English, teachers usually take the language knowledge as the teaching focus, and focus on knowledge inculcation. In addition, the asymmetry of teachers' specialties limits the teaching activities, and most of the teachers responsible for the teaching of medical English courses are unable to comprehensively cover both medical knowledge and English knowledge.

At present, the requirements for the professionalism of medical personnel are increasing, and the medical industry is constantly developing towards globalization, which requires their mastery of medical English. A detailed understanding of the teaching of medical English courses in medical schools reveals that most of the teachers teaching medical English courses were originally English professionals, and some of them may be foreign students majoring in medicine.

In the process of teaching activities, teachers more often follow the English teaching mode of college students, mainly learning vocabulary and practicing translation. For medical English teaching, this teaching mode leads to a

serious detachment between language and content, which is of limited help to cultivate students' English application ability in medical situations, and it is almost impossible to realize the demand of cultivating students' medical professional ability through English teaching. This teaching mode is of little help to the application of medical English in practice, and cannot meet the demand for the cultivation of real medical talents.

II. B.Course Structure Design for Medical English Majors

Medical English is an encompassing term that includes both academic medical English and vocational medical English. Within these two broad segments, specific courses can be offered that are targeted. All majors should offer general medical English, different majors should offer different medical professional English, and the focus of each major should be clear.

The series of medical English courses should be based on cultivating the overall quality of students, and strive to embody the guiding ideology of ability-based and vocational ability cultivation. Students should continuously improve their English language ability, master the basic features and expression habits of medical English, and use medical English as a tool to engage in study, research, communication and practice of medical-related work.

Based on the characteristics of the institution, the medical English specialized courses in this paper are divided into three categories: general medical English module, professional medical English module, and academic medical English module. The structure of specialized medical English courses is shown in Figure 1.

(1) General Medical English

Module is a basic course and an elective course. The general medical courses include Introduction to Medical Fundamentals, Humanistic Literacy in Medicine, Medical Sociology, and Medical Psychology, and 2 courses are offered in each semester of 1~2 semesters for 1 year. These general courses are designed to understand the medical English curriculum, medical psychology, medical sociology, and to master general medical vocabulary, basic medical knowledge, and to lay the knowledge and language foundation for the study of professional medical English and academic medical English. Students choose at least 1 course per semester according to their actual needs and English level.

(2) The module of Professional Medical English is the main course, which consists of compulsory courses and elective courses. The compulsory main courses include Medical English Listening and Speaking, Medical English Reading, Medical English Writing, and Medical English Translation, with 1 course in each semester from 1 to 4 semesters for two years. The elective main courses are English for Clinical Medicine, English for Pediatrics, English for Preventive Medicine, English for Nursing Medicine, English for Laboratory Medicine, English for Traditional Chinese Medicine, and English for Rehabilitation, which are offered for 1 year in 3~4 semesters, and the students choose 1 course per semester according to their own specialties and interests.

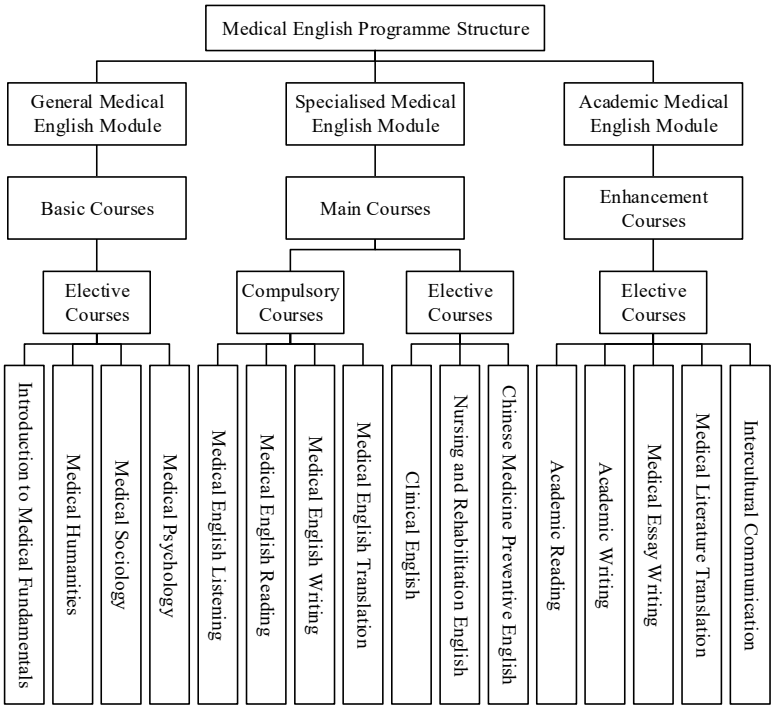


Figure 1: The structure of the medical English course

(3) Academic Medical English Module is an improvement course and is an elective course. It mainly includes academic reading, academic writing, translation of medical literature, writing of medical thesis, and cross-cultural communication courses, etc. It is offered for one year in 5~6 semesters. Students choose at least 1 suitable course to study each semester according to their medical specialties, to meet the demand for innovative applied research and the demand for students' academic advancement, aiming to improve students' professional academic quality, to meet the higher-order academic requirements, and to meet the demand for advanced medical foreign affairs.

II. C.Teaching Objective Formulation for Medical English Course

The medical English series of courses consists of three modules, namely general, specialized and academic, to achieve the level-by-level training objectives in a gradual manner. The goal of cultivating applied medical English talents is to enable students to master the English language, be proficient in medical knowledge and professional skills, have excellent political quality and high moral character, and be able to apply their knowledge and skills in the social and medical environments to relieve people's pains and suffering, deal with public medical affairs, and disseminate Chinese medical culture. The three-dimensional synergistic education goal of “knowledge transfer, value shaping and ability cultivation” is established in the teaching of the program.

III. Multidimensional data retrieval technology for medical English teaching resources

III. A. Construction of Digital Teaching Resources for Medical English

Medical English vocabulary is a language knockout for medical students' professional learning. The data analysis of medical students' English learning demand shows that students' demand for English courses of medical specialty-related categories, especially medical English vocabulary courses, reaches 80%. Medical English Vocabulary Advanced” catechism course follows the law of medical course learning. Starting from anatomy vocabulary which is the first thing that medical students learn, it analyzes and summarizes the law of medical English vocabulary construction according to the learning vein of organ, root, diagnosis and treatment, from shallow to deep, from easy to difficult, from basic to clinical, and gradually develops step by step. The book is designed to solve the problem of medical students falling into professional learning difficulties due to the large number of medical English vocabulary, and to lay a solid vocabulary and language foundation for their professional learning and medical research. Medical English Vocabulary Advanced is positioned as a public basic course, completely open to undergraduates, postgraduates and social learners, and belongs to the credited courses of universities and social learners.

The construction of “Advanced Medical English Vocabulary” catechism course follows the principles of science, systematicity and appropriateness in the construction of digital teaching resources for medical English, and goes through a five-step cycle of analysis, design, production, application and optimization, and is continuously optimized.

(1) Analyze the positioning of the course, which introduces the English vocabulary formation rules of human body tissues and systems in an all-round way, and helps students to grasp the vocabulary formation characteristics of medical English in general.

(2) Design the three main carriers of the course, i.e. the three-dimensional medical English vocabulary series paper-based textbook, the short medical English vocabulary series microclasses, and the open medical English vocabulary catechism, and design the course content and select the high-quality on-line platform “Xuedang Online”.

(3) Based on the idea of curriculum framework design, compile the three-dimensional paper textbook of Medical English Vocabulary Series “New Ways of Learning Medical English Vocabulary” and “Advanced Medical English Vocabulary Tutorial”, create the supporting e-teaching plan and multimedia courseware, develop the framework of Medical English Vocabulary Microclasses and “Medical English Vocabulary Advancement” Catechism, write the script, record the video, and design the Medical English Vocabulary Exercises and the humanistic and ethical discursive discussion questions. Discussion questions.

(4) Integrate the teaching resources of the paper textbook and the “Advanced Medical English Vocabulary” catechism class, carry out online and offline blended teaching before, during and after the class, cultivate students' linguistic ability to raise and analyze questions, and collect data on the use of the catechism class resources in the process, including the number of students' participation, posting, replying to posting, completion of exercises, and problems found in the course of using the catechism class resources.

(5) Based on the data feedback, supplement and improve the digital teaching resources to address the teaching problems, and continue to optimize the paper textbook, catechism and microteaching digital teaching resources in the new round of teaching practice of Medical English Vocabulary Advancement.

The construction of medical English catechism and microteaching resources follows the three principles of scientificity, systematicity and appropriateness, and through the five-step construction path of analysis, design, production, application and optimization, the digital teaching resources are initially constructed, and then after several rounds of iterative construction, the medical English catechism and microteaching digital teaching resources are created to achieve the continuous optimization of the medical English digital teaching resources.

III. B. Techniques related to indexing multidimensional data

III. B. 1) Massive data storage in a big data environment

In the face of the explosive growth of massive multidimensional data, there are many computer clusters using large-scale distributed storage, which connect thousands of computers through cloud storage technology to form a cloud storage system. These storage spaces are shared by many users, and the storage services they need can be dynamically formulated for users according to their actual storage needs. The existing cloud storage system mainly adopts Master/Slave architecture and pure distributed architecture.

(1) Master-Slave Architecture

The main data organization mode of master-slave architecture is based on the metadata server, which is the description of other relevant information about the data, including the storage address, dimension, attributes, data size and other information of the multidimensional data.

(2) Purely Distributed Architecture

Cloud storage systems with purely distributed architecture need to manage and organize massive data based on some kind of P2P protocols. Although these P2P protocols have different topologies, they mainly use the distributed hash table (DHT) method.

Relational databases do not have enough flexibility and scalability for diverse and complex massive data storage, and the speed of data processing is affected by data volume and data dimensions, so NoSQL databases with high scalability and applicability have emerged. More and more NoSQL databases are now able to address the challenges posed by the diverse and complex data types in the current big data collections, especially the application problems in big data.

Classified by data model, existing NoSQL databases can be categorized into four types: key-value databases, document databases, column-store databases, and graph databases.

III. B. 2) Multidimensional data

The so-called multidimensional data refers to data in multidimensional space, such as points, line segments, rectangles in two-dimensional space, and cubes, spheres, etc. in three-dimensional space. Generally speaking multidimensional data is characterized by complex structure, large amount of data, dynamic changes, etc. Therefore, certain features are also needed in designing and constructing the multidimensional index structure:

(1) Dynamic construction: after the storage of multidimensional data, it is necessary to support stable insertion, deletion and other update operations, therefore, after the index is constructed, in order to maintain the consistency of the data index, it is also necessary to support stable update operations.

(2) storage management: with the emergence of distributed storage technology, multidimensional data began to be gradually stored in various types of cloud databases, and cloud storage databases often have two or even three levels of management, the index structure needs to take into account the relationship between multi-level management.

(3) Growthability: the multidimensional index structure can be adapted to the growth of the multidimensional database size.

(4) Spatial effectiveness: Although the multidimensional index is much smaller than large-scale multidimensional data in space, but still need to take up a certain amount of storage space, so the index in the design and construction of the index structure needs to be fully considered in terms of space utilization.

(5) Time complexity: the existence of multidimensional indexes is to improve the query performance of multidimensional data, so in the design of query algorithms need to fully consider the query efficiency.

The query of multidimensional data refers to the process of finding data from a distributed database that satisfies one or more conditions. Because the data volume of multidimensional data is often large and the dimensions are complex, different query methods have different processing methods. According to the needs of the actual application system, distributed databases need to deal with different query requests. According to the different query requests of the user, the query methods for multidimensional data include precise matching query, point query, range query, intersection query, K-nearest neighbor query and so on. The more query methods used to query data in meteorological systems are Boolean query, point query, range query.

III. C. Clustering-based query processing techniques for multidimensional data

K-Means algorithm is an unsupervised clustering algorithm [21], [22]. Clustering refers to partitioning a dataset into different classes or clusters according to a specific criterion (e.g., distance criterion), so that data in the same class after clustering are aggregated together as much as possible, and different data are separated as much as possible.K-Means refers to constructing feature indicator data based on the business requirements or modeling needs in a two-dimensional or higher dimensional space, and through the pre-set K value and the initial center of

mass of each class, similar data points by pre-set K-value and initial center of mass of each category. The optimal clustering results are obtained through iterative optimization of the mean values after the division to maximize the intra-class similarity and minimize the inter-class similarity clustering results.

In the field of machine learning, model feature term dimensionality reduction is used to reduce the redundant and noisy information in each feature set in the model in order to improve the accuracy of the model in applications.

There are many commonly used dimensionality reduction methods, among which are Pearson correlation coefficient.Pearson correlation coefficient is used to reflect the closeness of the correlation between two variables. It takes values in the range of [-1,1]. When the coefficient is close to 1, it indicates that the two variables are positively correlated. When the coefficient is close to -1, it indicates that the two variables are negatively correlated. When the coefficient is close to 0, it means that there is no linear relationship between the two variables.

The role of model feature scaling is to eliminate the bias caused by the different scales of the features in the model, and the common methods to achieve feature scaling homogenization are standardization and normalization methods. The formula (1) of the normalization method is as follows:

$$Z = \frac{\chi - \mu}{\sigma} \tag{1}$$

In the K-Means algorithm, the feature indicator data generally have units, such as the life cycle unit is the number of months, the number of application products is the number of products (a), the actual business share ratio is the ratio, itself a different unit, the value of the size of the different, positive and negative. The different orders of magnitude of the feature items of the model will affect the clustering effect, and it is necessary to standardize the processing of such feature items, so that the items are in a uniform order of magnitude.

K-Means algorithm first randomly generated k an initial clustering center, and then calculate the distance between each data object and the clustering center, generally using the European distance see formula (2) shown:

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{2}$$

If the data object is closest to a certain clustering center, the data object will be grouped into this class and the average value of each cluster will be recalculated. The convergence is finally achieved through continuous calculation, so that the data objects within the same class cluster have high feature similarity, while the similarity between different class clusters is low, forming the final clustering center and achieving the purpose of cluster analysis.

Using the K-Means algorithm, the optimal number of clusters k value will affect the clustering effect, need to determine the best value of k through the research trial. In this paper, the contour coefficient method is used to determine the value of k.

The contour coefficient is affected by the degree of cohesion and separation, the degree of cohesion reflects the degree of closeness of the sample points to the elements within the cluster, and the degree of separation reflects the degree of closeness of the sample points to the elements outside the cluster. The data is divided into k clusters by clustering, and for each vector in the cluster, its contour coefficient is calculated. The formula for any point i is:

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{3}$$

$$S(i) = \begin{cases} 1 - \dfrac{a(i)}{b(i)}, & a(i) < b(i) \\ 0, & a(i) = b(i) \\ \dfrac{b(i)}{a(i)} - 1, & a(i) > b(i) \end{cases} \tag{4}$$

where $a(i)$ is the degree of cohesion denotes the mean of the sum of the distances from the point to the other samples in its same cluster. $b(i)$ is the minimum value of separation denoting the average distance of the point to all the points in each of the other clusters. When $a(i) < b(i)$, that is, when the distance within the class is smaller than the distance between the classes, the result of the calculation of the contour coefficient tends to 1, and the clustering result is more compact, with a clear contour of each class.

III. C. 1) Extreme value problem description

Sampling-based approximate query processing techniques can effectively shorten the response time of multidimensional analysis queries on massive data. However, there are usually many cases of extreme values in massive data, which seriously affect the accuracy of approximation results. In order to further improve the accuracy of aggregated queries, this chapter proposes the Clustering + Selective Estimation based Stratified Sampling Approximate Query Processing (CSSAQP) algorithm.

Multidimensional analysis of massive data is an extreme resource-consuming operation. The sampling-based approximate query processing technique, on the other hand, greatly reduces the amount of data to be processed by extracting a small amount of sample data that is representative of the full set of data, and therefore can drastically reduce the execution time of the query. Compared to other approximate query processing techniques with parameters, it does not need to predict the data distribution of the query dataset in advance, which is simpler and easier to realize, and becomes the most commonly used approximate query processing technique.

Suppose that the relation R contains the character attribute ItemNO, and the aggregation attribute Profit with a total of 4 records as shown below:

Relationship R:

$$<\text{ItemNO,Profit}> \tag{5}$$

Record:

$$\{<1,800>,<2,900>,<3,800>,<4,10000>\} \tag{6}$$

Query Q: SELECTAVG (Profit) FROMR

A simple random sampling technique is used to construct a sample set S by taking two records and executing Q on S. Suppose the constructed random sample S is: $\{<1000>,<4000>\}$. The actual result of executing Q query on the original record Result=4528, and the approximate result of executing the query on S $Result'$=1200.

Due to the existence of extreme values, the approximate results obtained by applying the random sampling technique have a large deviation from the actual results, which seriously affects the results of the approximate query.

III. C. 2) Definitions and design

In this paper, the relative error rate (hereinafter referred to as error rate) is used to measure the level of error generated by each grouped query in the aggregation query containing Group-by. Assuming that the exact aggregation value of subgroup i is c_i and its approximation is $c_i^{'}$, the error rate of subgroup i, ε_i, is defined as follows:

$$\varepsilon_i = \frac{|ci - ci'|}{ci} * 1000\% \tag{7}$$

For the aggregation operation with Group-by in the multidimensional analysis of massive data, the use of approximate query processing technique for it needs to satisfy the following two requirements: first, the result of the approximate query should contain all the groupings. Since CSSAQP will stratify the Group-by attribute first, and then implement stratified sampling for each stratum, it is assumed that this requirement has been satisfied under the premise of large data volume.

Second, the query results of each grouping should be as accurate as possible. In this paper, the maximum error rate caused in the approximation results of all subgroups is chosen to measure the overall approximation results. For the Group-by aggregation query containing n subgroups, the following definition is quoted for the index to measure the overall result of its goodness:

$$\varepsilon_\infty = MAX_i^n \varepsilon_i \tag{8}$$

Pre-construction of the sample table, first of all, it is necessary to clarify what kind of samples to build, need to make some assumptions about the future analysis task, usually assuming that the future query and the historical query similar, in the construction of the sample table can be based on the historical query prediction of the future query.

According to the predictability of future queries, the types of queries can be categorized into four types of query hypothesis models: query completely knowable, query predicate knowable, query column set knowable and query completely unknown.

It is found that the columns involved in a query (e.g., Group-by attribute columns) are usually stable in realistic analysis tasks, and the set of columns accessed in a historical query has a high probability of being used to construct

queries in future queries as well. The query-based column set agnostic model, on the other hand, can effectively precompute samples precisely by assuming that the information about the set of columns likely to be involved in future queries is known, and thus can be considered as a basis for constructing sample tables.

III. C. 3) CSSAQP algorithm design

In order to facilitate the elaboration of this algorithm, the attributes of the relational table OrgTable are divided into character attribute sets C and aggregated attribute sets V, where Group-by attribute sets $G \subset C$. When performing data analysis, the model is known based on the query column set, assuming that the Group-by attribute set and the aggregated attributes to be calculated are known.

If the relational table OrgTable, Group-by attribute set G, an indicator attribute $v \in V$, and sampling rate f are specified, CSSAQP can constitute the overall sample table SampleTable, and when the query arrives the query is executed by query rewriting on the resulting sample table SampleTable, and an approximation of the true value is returned . The specific steps are as follows:

CSSAQP is divided into two phases, the preconstruction sample phase and the query execution phase.

Phase 1: Pre-constructed sample phase, CSSAQP constructs two kinds of sampling tables. Firstly, K-means clustering algorithm is utilized to divide OrgTable into 3 clusters according to v. Based on the clustering results, CSSAQP directly conducts random sampling within each cluster at the sampling rate f and constructs the total sample table SampleTable, denoted as CSTable. And then divides the clusters into a number of disjoint strata according to G, then randomly draws samples within each stratum at the sampling rate f and constructs the total sample table SampleTable, denoted as CSSTable.

Phase 2: Query execution phase. When the query Q arrives, go through the query parsing layer to obtain the G, v of the query statement Q. Then through sample matching, select the corresponding SampleTable to replace OrgTable and execute the query to output the approximate result of the query. If the sample matching fails, the original query is executed directly and the exact value of the query is output.

IV. Experimental evaluation

The experimental environment is based on Hadoop 2.7.7 and Spark 2.4.4, using five AliCloud servers configured with Centos 7.4 operating system, one of which serves as a master node and four as slave nodes. Each server is configured with 32G RAM, 8-core processor and 1TB hard disk. The servers are connected to each other with Gigabit Ethernet.

IV. A. Experimental design

(1) Data sets

The experiments were conducted on 2 datasets, containing the TPC-H test dataset and the Medical English Vocabulary Teaching dataset.

The TPC-H test dataset was generated from a Zipfian distribution with a skewness of 1 and a scale factor of 10. The result table has about 600 million rows and occupies about 54 GB of storage space. The size of each chunk is 128 MB, totaling 450 partitions.

The Medical English Vocabulary Teaching dataset contains 10 columns of data, totaling about 0.5 billion pieces of data. It occupies about 90GB of storage space. Each chunk size is 128MB, totaling 510 partitions.

(2) Methods involved in the comparison

Random sampling: chunks are uniformly sampled randomly and the final query results are scaled according to the sampling rate. Since the query results of random sampling fluctuate greatly due to the influence of chunk selection, the final comparison of the query result error takes the average result of 10 queries.

Random sampling + selective estimation: the same as random sampling, filtering the chunks with selective estimation of 0.

Clustering + selective estimation: the method described in this paper is realized without using the model to determine the chunk selection.

(3) Evaluation metrics

Regarding the evaluation of the accuracy of the query results, the average relative error is used, i.e., the average of the relative errors of all the aggregate function results of the sampled query results compared to the full query results. Noting that the number of query aggregation functions is m, the result of the i th aggregation function of the sample query is P_i, and the result of the i th aggregation function of the full query is F_i, the average relative error of this sample query ARE is calculated as follows:

$$ARE = \frac{\sum_{i=1}^{m} \left| \frac{P_i - F_i}{F_i} \right|}{m} \tag{9}$$

IV. B. Analysis of experimental results

IV. B. 1) Query errors

The experiment is comparing the query error performance of various methods on different datasets and different sampling rates, the lower the curve as a whole, the better the method performs.

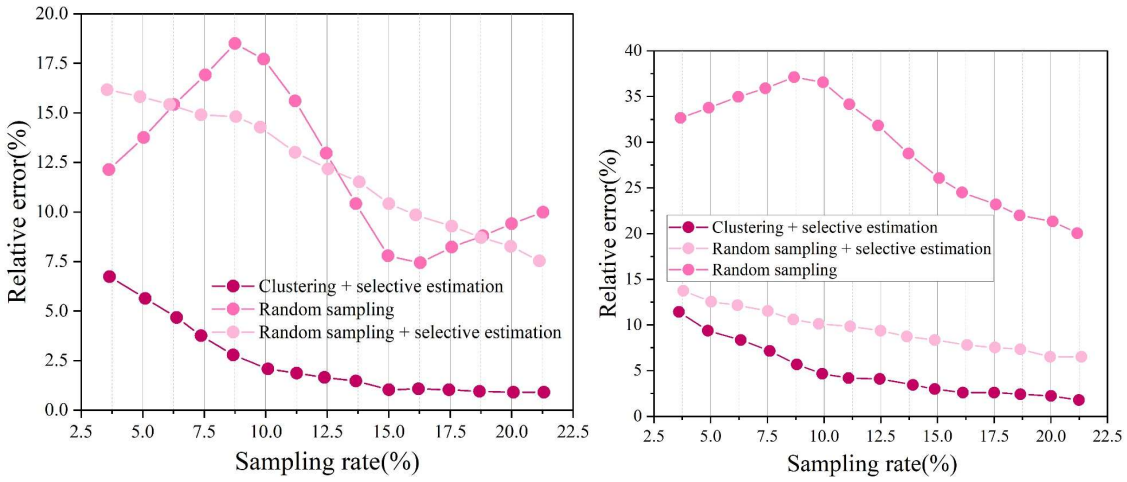
The comparison of query relative error is shown in Fig. 2, Fig. (a) and Fig. (b) show the query relative error of the algorithms on two datasets respectively. The overall performance of the method in this paper is optimal.

The results show:

① Due to the non-uniform data distribution, the results of randomly sampled queries depend on whether the chunks that effectively participate in the query are selected or not, and do not get better with the increase of the sampling rate, and the error fluctuates significantly.

② Filtering the non-participating query chunks using selective estimation of the query yields better and more stable error performance.

③ By clustering the chunks, better error performance can be obtained when few chunks are read. And as more chunks are read, the query error can be effectively reduced.



(a) TPC-H data set

(b) Medical English vocabulary teaching data set

Figure 2: Comparison of relative error of query

IV. B. 2) Query results

For single table single column query as well as range query, this paper uses TPC-DS dataset to generate store_sales dataset of 45G size and takes ss_wholesales_list column as aggregated query column.

The experiments are conducted by performing SUM, COUNT, and AVG aggregation operations several times and taking the average of the results of each experiment. In this paper, the comparison of accuracy and query time is done with BlinkDB as well as VerdictDB and Spark DataFrame queries.

The results of the query time comparison are shown in Figure 3. It can be seen that the BlinkDB method and the method of this paper and VerdictDB are able to complete the query in a shorter period of time, while the Spark DataFrame query takes a longer time. SUM=1.52s, COUNT=1.24s and AVG=1.34s for this paper's method.

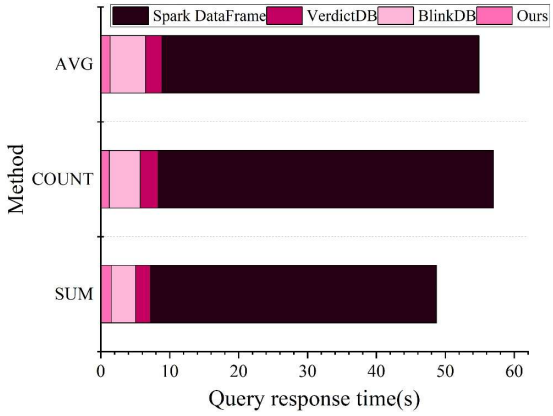


Figure 3: Comparison results of query time

Next is the precision comparison, at this point Spark DataFrame is no longer included, because the query results of Spark DataFrame are the precise results.

The query precision comparison results are shown in Figure 4.

The method in this paper improves 0.0087%, 0.0073%, and 0.0103% precision in SUM, COUNT, and AVG respectively than VerdictDB method. It can be found that the accuracy achieved by this paper's method is higher than that of BlinkDB and VerdictDB, in which the COUNT queries of this paper's method are accurate.

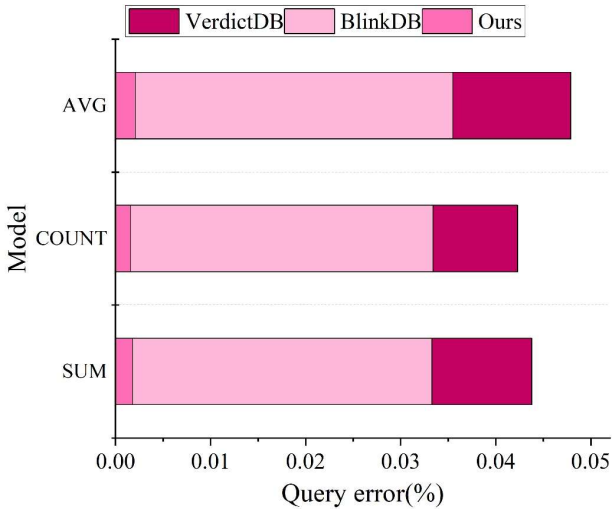


Figure 4: Comparison results of query accuracy

Next, this paper compares the range query accuracy of the three methods, whose response times are similar. The comparison of range query accuracy is shown in Fig. 5.

The accuracy is lost because the method in this paper involves scaling calculation during range query. But the method still outperforms BlinkDB as well as VerdictDB method. The SUM=0.0089%, COUNT=0.0095% and AVG=0.0112% for this paper's method.

In this paper, a total of about 54GB of data is generated using TPC-H dataset, in which lineitems table occupies 40G and orders table occupies 14G, and primary key foreign key join experiments are conducted using the two tables. And the experiments were compared with post-connection sampling, VerdictDB, and Spark DataFrame join methods.

The result of query time comparison is shown in Figure 6. From the bar presentation in the figure, it can be seen that the query time of Spark DataFrame method is longer, significantly more than this paper's method, VerdictDB, and post-connection sampling method. And the query time of this paper's method is better than VerdictDB, post connection sampling method, which can achieve real-time response.

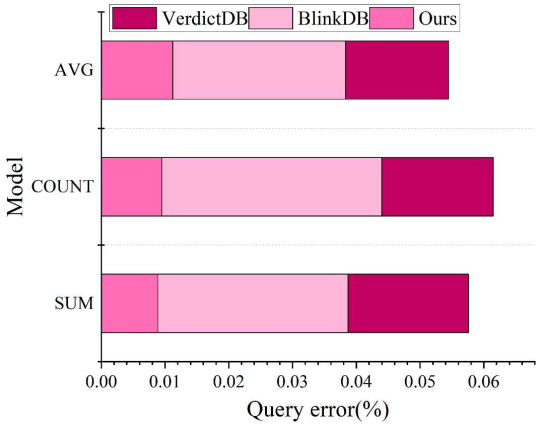


Figure 5: Range query accuracy contrast

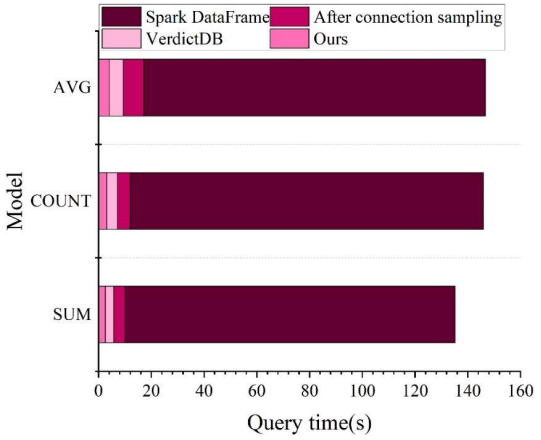


Figure 6: Query time comparison results

Next compare the query accuracy of the two sampling methods and the method of this paper, the connection accuracy comparison results are shown in Figure 7. The query accuracy of this paper's method is 0.0122%, 0.0141% and 0.0085% in SUM, COUNT and AVG respectively. It can be seen that the method of this paper can still get high accuracy, which can prove the feasibility of the method of this paper.

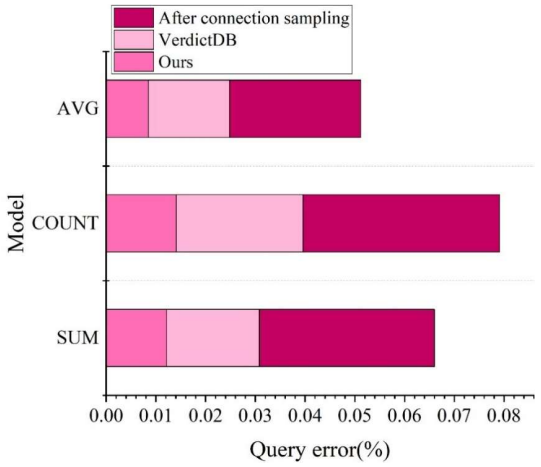


Figure 7: Comparison of connection accuracy

V. Conclusion

The application of multidimensional data analysis model in medical English teaching effectiveness assessment provides an effective way to solve the problems existing in traditional assessment methods.CSSAQP algorithm

successfully copes with the problems of massive data processing and extreme values in medical English teaching assessment, and experiments have proved that the method has significant advantages. In terms of query accuracy, the algorithm achieves 0.0122%, 0.0141%, and 0.0085% accuracy for SUM, COUNT, and AVG metrics in the connection query, which is better than the performance of VerdictDB and connection post-sampling methods. In terms of query efficiency, the query times of the three aggregation operations are SUM (1.52 seconds), COUNT (1.24 seconds), and AVG (1.34 seconds), respectively, which realizes real-time response to about 0.5 billion medical English teaching data, and meets the timeliness needs of teaching evaluation. The three-module medical English curriculum system (general, specialized, and academic) constructed in this study was combined with multidimensional data analysis technology to form a complete medical English teaching evaluation system. Through the five-step cycle of "analysis-design-production-application-optimization", the construction of medical English digital teaching resources continuously optimizes the teaching content and improves the relevance and effectiveness of medical English teaching. Future research should focus on the combination of multidimensional data analysis model and artificial intelligence technology to further improve the accuracy of medical English teaching assessment; explore the design of more personalized medical English learning paths; and expand the application of multidimensional data analysis in the assessment of medical humanities, so as to comprehensively improve the quality of medical English teaching.

Funding

This work was supported by Heilongjiang Province Philosophy and Social Science Research Project (23YYC314); Heilongjiang Province Department of Education Language and Script Science Research Project (2024Y043); Heilongjiang Province General Higher Education Teaching Reform Project: "Research on the Construction and Enhancement Strategy of 'Competence' Model for Foreign Language Teachers in Traditional Chinese Medicine Colleges and Universities in the Context of Course Ideology and Politics".

References

[1] Hyland, K. (2022). English for specific purposes: What is it and where is it taking us?. ESP Today-Journal of English for Specific Purposes at Tertiary Level, 10(2), 202-220.
[2] Yu, X., & Liu, C. (2018). Curriculum reform of college English teaching in China: From English for general purposes to English for specific purposes. ESP Today, 6(2), 140-160.
[3] Li, Y., & Heron, M. (2021). English for general academic purposes or English for specific purposes? Language learning needs of medical students at a Chinese university. Theory and practice in language studies, 11(6), 621-631.
[4] Ferguson, G. (2025). English for medical purposes. The handbook of English for specific purposes, 343-362.
[5] Himmelstein, J., Wright, W. S., & Wiederman, M. W. (2018). US medical school curricula on working with medical interpreters and/or patients with limited English proficiency. Advances in medical education and practice, 729-733.
[6] Tomak, T., & Šendula-Pavelić, M. (2017). Motivation towards studying English for specific purposes among students of medical and healthcare studies. Jahr–European Journal of Bioethics, 8(2), 151-170.
[7] Outemzabet, B., & Sarnou, H. (2023). Exploring the significance of English-based communication for a community of medical academics in a public university teaching hospital in Algeria. English for Specific Purposes, 70, 116-130.
[8] Zhixiang, S. H. I., Yuhui, Z. H. A. N. G., & Jing, C. H. E. N. (2023). Cultivation of Cross-Disciplinary Talents of English for Pharmacy with Global Vision in the Context of New Liberal Arts Construction. Pharmaceutical Education, 39(6), 15-18.
[9] Khalili, S., & Tahririan, M. H. (2020). Deciphering challenges of teaching English for specific purposes to medical students: Needs, lacks, students' preferences, and efficacy of the courses. Teaching English Language, 14(1), 365-394.
[10] Han, L., Yao, X., & Yu, J. (2022). Application of deep learning in medical English teaching evaluation. Wireless Communications and Mobile Computing, 2022(1), 8671806.
[11] Wei, C., Gu, Y., Wang, W., Gu, B., Li, Q., & Wang, Z. (2025). Student-centered, humanities-guided teaching of the "Medical Practical English" course and its assessment. Global Medical Education, (0).
[12] Wang, X., Zhong, L., Li, Q., Chen, Y., Huang, T., Li, X., & Lin, H. (2024, November). Implementation of Innovative Medical English Teaching Model Based on Narrative Medicine and Evaluation System Design. In Proceedings of the 2024 7th International Conference on Educational Technology Management (pp. 539-545).
[13] Yan, S. J., Chen, Y. Q., Chen, J., & Lv, C. Z. (2017). Evaluation of teaching-in-English reform in five-year clinical tropical medicine program–Case analysis of curriculum reform of clinical medicine in Hainan Medical University. Asian Pacific Journal of Tropical Biomedicine, 7(12), 1125-1128.
[14] Zakharova, A. A., Vekhter, E. V., Shklyar, A. V., & Pak, A. J. (2018). Visual modeling in an analysis of multidimensional data. In Journal of Physics: Conference Series (Vol. 944, No. 1, p. 012127). IOP Publishing.
[15] Appah, B., & Amos, D. (2018). Multidimensional data model for health service decision making data. International Journal of Computer Science Engineering Techniques, 3(3), 1-6.
[16] Vintilă, G., Onofrei, M., & Gherghina, Ş. C. (2017). Multidimensional data analysis towards assessing the European education systems. European Journal of Sustainable Development, 6(2), 69-69.
[17] Francés, O., Abreu-Salas, J., Fernández, J., Gutiérrez, Y., & Palomar, M. (2023). Multidimensional data analysis for enhancing in-depth knowledge on the characteristics of science and technology parks. Applied Sciences, 13(23), 12595.
[18] Hongmei Cui & Naginder Kaur. (2025). Challenges in Teaching Medical English Vocabulary to Tertiary Students in China: A Systematic Literature Review. Theory and Practice in Language Studies,15(2),579-591.

[19] Peixin Lin,Yuexin Cai,Sijia He,Xinxin Li,Yuanke Liang,Tian Huang... & Haoyu Lin. (2025). Enhancing medical English proficiency: the current status and development potential of peer-assisted learning in medical education. BMC Medical Education,25(1),79-79.
[20] Zhang Bingru & Huang Manxi. (2025). The Integration Path and Theoretical Teaching Innovation of Ideological and Political Education in Medical English Courses. Journal of Contemporary Educational Research,8(12),42-48.
[21] Mohana Alanazi,Almoataz Y. Abdelaziz,Junhee Hong & Zong Woo Geem. (2025). Enhancing stochastic planning in autonomous hybrid energy systems through an advanced arithmetic optimization algorithm and K-means data clustering. Energy Reports,13,4375-4387.
[22] Jun Lu,Tingjin Luo & Kai Li. (2025). A forward k-means algorithm for regression clustering. Information Sciences,711,122105-122105.