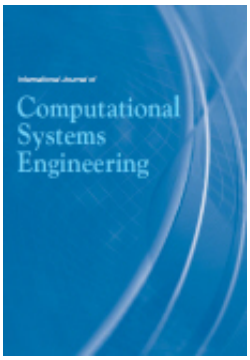
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Computational Systems Engineering

ISSN online: 2046-3405 - ISSN print: 2046-3391
https://www.inderscience.com/ijcsyse

Online education resource integration method for painting teaching of art majors based on cloud platform

Muchao Zhang

DOI: 10.1504/IJCSYSE.2027.10066544

Article History:

Received:	14 June 2023
Last revised:	15 January 2024
Accepted:	03 July 2024
Published online:	10 February 2025

Online education resource integration method for painting teaching of art majors based on cloud platform

Muchao Zhang

Academy of Fine Arts,
Nanjing Xiaozhuang University,
211171, Nanjing, China
Email: zhangmuchao063@outlook.com

Abstract: A method for integrating educational resource data is proposed to address the issue of redundant and low utilisation of educational resources in current cloud platforms. Firstly, the PMI algorithm is combined with Simhash algorithm to construct the PMI Simhash algorithm. Secondly, it is combined with mixed similarity to establish a BSM multimodal data matching model. Finally, an unsupervised self-learning entity matching algorithm based on Euclidean locally sensitive hashing algorithm and correlation vector machine is proposed, and a data integration method for cloud platform educational resources is constructed. During the test, the precision of BSM model was 0.953, recall was 0.962, F1 was 0.929, overall was 90.72%, and AUC value was 0.971. The precision of US-EM model was 97.78%, recall was 94.62%, F1 was 95.14%, and AUC value was 0.968. The above data validates the effectiveness of the research method and indicates that the study has positive implications for online education.

Keywords: cloud platform; art major; painting teaching; online education; resource integration.

Reference to this paper should be made as follows: Zhang, M. (2025) 'Online education resource integration method for painting teaching of art majors based on cloud platform', *Int. J. Computational Systems Engineering*, Vol. 9, No. 5, pp.1–10.

Biographical notes: Muchao Zhang obtained her BA in Fine Arts from Nanjing Normal University in 2002. She obtained her MA in Fine Arts from Nanjing Normal University in 2009. Currently, she is working as a Lecturer in the Department of Fine Arts, Nanjing Xiaozhuang University. Her areas of interest are art education, comprehensive painting.

1 Introduction

Cloud platform is a network platform based on cloud computing, which has the characteristics of scale, distribution, virtualisation, high availability, high scalability, economy and high security (Jin et al., 2020; Huang et al., 2020). At present, the online education model based on cloud platform has become mature. This enables students and teachers to achieve online learning and improve their knowledge and skills through the network, such as online teaching of painting for art majors (Castro and Tumibay, 2021; Verawardina et al., 2020). At present, online education based on cloud platform is very common, but its data heterogeneity problem has not been solved. This makes the use of educational resources in the cloud platform inefficient and affects the teaching effect (Luecken et al., 2022). Therefore, taking the online education of painting teaching for art majors based on cloud platform as an example, this paper proposes a data integration method of education resources. In data integration, data schema conflict and data redundancy are the most intuitive problems. To solve the above two problems, a BSM multi-pattern data matching model combining PMI-Simhash algorithm with mixed similarity and an unsupervised self-learning entity matching algorithm are proposed. This can achieve data integration, eliminate data pattern conflict and data redundancy, improve the quality of education resources data, and then improve the education effect. The research has promoted the development of data integration technology and online education model to a certain extent, and also widened the application scope of data integration technology to a certain extent.

2 Related works

Cloud platform was a network platform based on cloud computing, which was characterised by scale, distribution, virtualisation, high availability, high scalability, economy and high security. It was widely used in various fields, including data storage, resource sharing, etc. To maximise the potential of cloud platforms and promote the development of cloud platforms, many scholars had discussed and analysed cloud platforms and cloud computing. Chen et al. (2021) found in the survey that some cloud providers would launch preemptive services for cloud platforms to attract customers. In view of this preemptive service, this study proposed a discount scheme and compared it with the conventional discount scheme. The comparison results showed that the effect of the discount

scheme was better than that of the conventional discount scheme. Robertson et al. (2020) designed a drug supervision cloud platform, which deepened the information exchange between drug companies and regulatory agencies, thus having positive significance for drug development and regulatory strengthening. Vanama et al. (2020) used Google Earth Cloud Platform as a tool to achieve efficient mapping of flood areas through its Time Sentinel-1 SAR image function. This provided support and guarantee for the follow-up rescue and reconstruction work. Jingzhu and Yuanyue (2020) carried out an empirical research on the national public culture cloud platform. The main purpose of the research was to investigate the satisfaction of users with the cloud platform, so as to provide suggestions and ideas for the improvement and optimisation of the national public culture cloud platform. Rostami and others (2022) used a large number of remote sensing time series data stored in the cloud platform and built a flood warning model based on fuzzy comprehensive evaluation to achieve flood prediction and reduce economic losses. The experimental results verified the accuracy of the model. Yao et al. (2021) proposed a scheduling method based on task repetition for the workflow scheduling problem in the cloud platform to improve the use time of workflow. Experiments showed that this method could improve the efficiency of workflow, reduce time by 17.4%, and improve resource utilisation by 31.6%. Cao et al. (2020) proposed a digital watermarking encryption algorithm based on a cloud platform combining CPU and FPGA for the batch and cloud processing requirements of digital watermarking. The simulation results showed that this encryption method could improve the throughput and processing speed of the digital watermark encryption process. Based on cloud platform technology, Yao (2021) proposed an English hybrid learning method and applied it to English teaching in vocational high schools. This could improve the effect of English teaching and the English level of vocational high school students, and had a positive effect on the improvement of students' comprehensive quality and ability. For the service system management platform for employment and entrepreneurship of college students, Ping (2023) combines IoT technology, collaborative filtering algorithm, and binary K-means algorithm to construct a personalised recommendation model for graduates. This provides technical optimisation for the service platform for entrepreneurship and employment of college students, which to some extent alleviates employment pressure. In addition, for personalised recommendation information services, Ganesan et al. (2023) found in their exploration of recommendation systems that recommendation systems mainly solve the problem of information overload and exhibit significant advantages in the field of enterprise e-commerce. And combined with collaborative filtering algorithms to demonstrate its good performance in personalised recommendations in other fields.

At present, with a highly developed degree of informatisation, all walks of life were gradually moving towards the direction of intelligence, digitalisation and automation. Therefore, there was a huge amount of data information in various fields, including a large amount of data stored on various cloud platforms. However, due to the different sources of these data, there were problems in their data format and structure, which had seriously affected the efficiency and effectiveness of the use of these data. Data integration was the main way to solve the above problems, which had attracted the attention of various research fields. Argelaguet et al. (2021) analysed data integration in multimodal analysis of single cell. They believed that the effect of the existing data integration strategy was not ideal enough, and needed to develop new and more efficient data integration strategy. At present, most of the quantitative biomarker methods only aimed at a single data model, and perform poorly in cross-data model. Boehm et al. (2022) proposed a multimodal data integration method based on artificial intelligence for these problems, thus promoting the development of accurate oncology. Canzler et al. (2020) discussed and analysed the integration of multi-omics data in toxicology, and studied its development prospects and existing problems based on the existing results of multi-omics data integration. Younas et al. (2020) proposed a data integration method that combined display features to solve the problem of inaccurate nursing due to different data sources and formats of patients in the nursing process. The experiment showed that this method can improve the nursing effect. There was a confounding effect in the identification of biomarkers of intestinal microorganisms, resulting in conflicting research results. Xiao et al. (2022) proposed a NetMoss algorithm to solve this problem, so as to achieve data integration of large-scale microbiota, thus improving the recognition stability. In the process of abstract information extraction, the data structure and data model were different due to different abstract sources. Zhai and Han (2022) proposed a heterogeneous data integration method to solve this problem, and carried out experimental verification on the performance of the method. In data analysis, data from a single source could not meet the distribution requirements of the research group, so data from multiple sources was often used, resulting in data heterogeneity. Nargesian et al. (2021) proposed a data integration strategy to solve this problem. The experiment showed that the data integration strategy had excellent performance. Raghavan and others (2019) analysed the problem of large amount of sensor data and different sources in the construction of smart cities. They believed that data integration was an important means to achieve smart city data analysis. It also analysed the challenges faced by data integration technology and the development prospects of data integration in the current process of smart city construction.

As can be seen from the above contents, cloud platform and data integration technology had a wide range of applications, and scholars in various fields had also conducted in-depth research on them. In addition, online education based on cloud platform had become very popular. However, the problem of data heterogeneity had not been solved, and the relevant research was also less,

resulting in a certain impact on the effect of online education. Therefore, taking the online education of painting teaching for art majors based on cloud platform as an example, this paper proposed a data integration method of education resources. This method could eliminate the problem of data pattern conflict and data redundancy, improve the quality of education resource data, and then improve the education effect.

3 Research on education resource data integration based on cloud platform

3.1 Multi-mode data matching based on Simhash and mixed similarity

In recent years, due to the high development of information technology and the impact of the COVID-19, online education based on cloud platforms has also developed and become one of the mainstream education methods. However, in the cloud platform, the sources of education resources are different, resulting in data mode conflicts. The data of education resources in different modes cannot be interconnected, affecting the effect of online education. Therefore, it is necessary to carry out multi-pattern matching on these data to eliminate the impact of data pattern heterogeneity. Pattern matching is realised by calculating the similarity between attributes. Simhash algorithm is a common text similarity calculation method, and Figure 1 is the process.

Attribute columns, that is, the characteristics of instance data in attributes can effectively reflect attributes. When extracting attribute column features in general methods, the vector dimension is too high, resulting in increased computational difficulty. To solve this problem, a feature extraction algorithm combining point mutual information (PMI) and Simhash algorithm is proposed. In the PMI-Simhash algorithm, a fixed-digit signature can be produced, and the signature can be used to represent the characteristics of the attribute column, thus reducing the dimension of the feature vector. For numeric attribute columns, the equidistant division method is used to extract features and express them as key-value pairs, such as equation (1).

$$a = \left\{\left\langle u_1, ta(u_1)\right\rangle, \left\langle u_2, ta(u_2)\right\rangle, \ldots, \left\langle u_i, ta(u_i)\right\rangle\right\} \tag{1}$$

a is an attribute column in equation (1). u_i is the i characteristic unit of a, that is, the i contained in a has the numerical value or string that actually contains and can be used to represent the a characteristic. $ta(u_i)$ represents the number of times u_i has occurred in a. If there are n attribute columns, the feature unit intersection of these attribute columns is expressed as equation (2), that is, the feature set of all attribute columns.

$$U = \{u_1, u_2, \ldots, u_m\} \tag{2}$$

The amount of information contained in a random feature unit u_y can be evaluated through the mutual information between points of PMI. At the same time, the difference between the amount of information contained in a random attribute column a_k can also be evaluated, expressed as $pmi(a_k, u_y)$ in equation (3).

$$pmi\left(a_k, u_y\right) = lb\frac{ta_k\left(u_y\right)/T}{\left(\sum_{i=1}^{n} ta_i\left(u_y\right)/T\right)\cdot\left(\sum_{y=1}^{m} ta_k\left(u_y\right)/T\right)} \tag{3}$$

$\sum_{i=1}^{n} ta_i(u_y)/T$ is the sum of the times that u_y appears in the attribute column in equation (3). $\sum_{y=1}^{m} ta_k(u_y)/T$ is the sum of the frequency of all characteristic units in U in a_k. T is the total number of occurrences of all characteristic units in all attribute columns. The larger the $pmi(a_k, u_y)$, the greater the correlation between the attribute column and the feature unit. The more identical and similar feature units between two attribute columns, the greater the probability of the two attribute columns matching successfully. Therefore, the PMI value is taken as the weight value in Figure 1, and the signature of the feature unit is generated to represent the attribute to complete the feature dimension reduction. The signature is clustered and the attributes with similar characteristics are divided into the same class to achieve the matching between attributes. K-means algorithm is used to achieve clustering analysis, and Figure 2 shows the clustering process.

Figure 1 Schematic diagram of Simhash algorithm (see online version for colours)

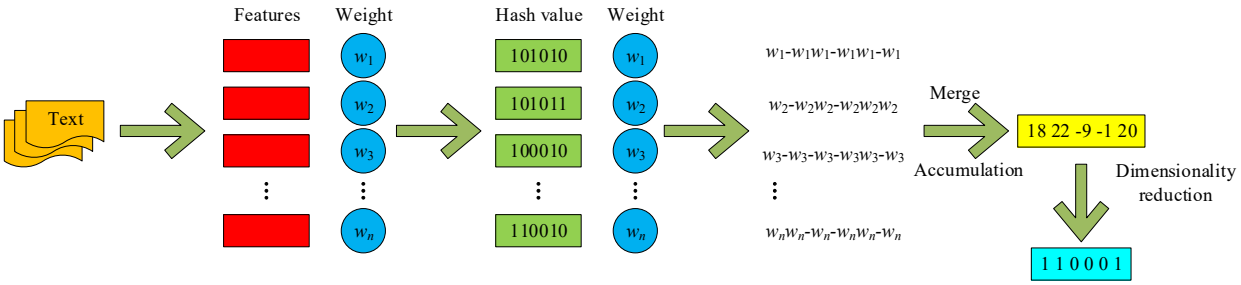


Figure 2 K-means algorithm for signature clustering (see online version for colours)

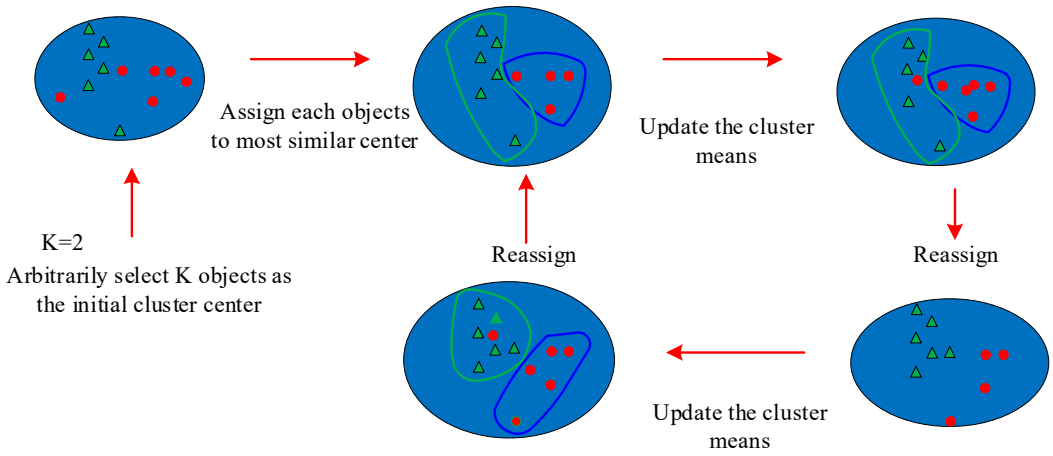
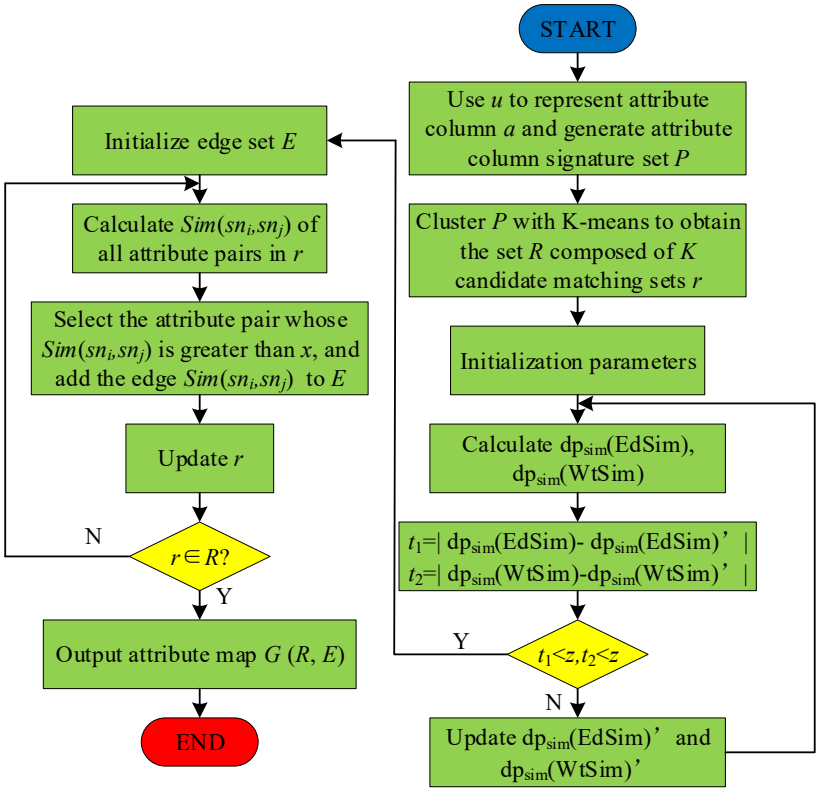


Figure 3 Basic process of multi-pattern data matching model (see online version for colours)



Mismatch may occur during clustering. There are generally two kinds of mismatches. The first is that attributes with different semantics are grouped into one class. The second is that attributes with the same semantics are in different classes. In view of the above situation, to remove mismatching items, the matching accuracy between attributes is improved by calculating the semantic similarity of attributes. In educational resources, the calculation of grammatical and semantic similarity alone cannot accurately reflect the relationship between various attributes. Therefore, a hybrid similarity calculation method considering syntax and semantics is proposed. Equation (4) is the definition of the similarity distinguishing ability of the grammar or semantic similarity calculation method.

$$dp_{sim} = \frac{\sum_{sim \in SIM_{X^m}} sim_i + \sum_{sim \in SIM_{X^m}} 1 - sim_i}{|X^m| + |X^u|} \tag{4}$$

In equation (4), dp_{sim} represents the similarity discrimination ability. sim_i represents the similarity of the i attribute pair in the attribute pair set. X^m and X^u represent a set of attribute pairs that are marked as matched and unmatched respectively. $|X^m|$ and $|X^u|$ represent the number of attribute pairs in X^m and X^u, respectively. Based on equation (4), the hybrid similarity calculation model proposed as follows.

$$sim\left(sn_i, sn_j\right) = \left(\begin{array}{c} \dfrac{dp_{sim}(EdSim)}{dp_{sim}(EdSim)+dp_{sim}(WtSim)} \\ \cdot EdSim\left(sn_i, sn_j\right)^p \\ + \dfrac{dp_{sim}(WtSim)}{dp_{sim}(EdSim)+dp_{sim}(WtSim)} \\ \cdot WtSim\left(sn_i, sn_j\right)^p \end{array} \right)^{1/p} \tag{5}$$

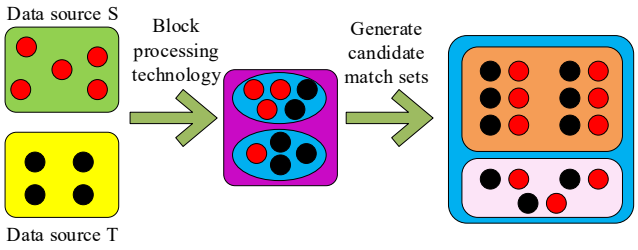
$dp_{sim}(EdSim)$ and $dp_{sim}(WtSim)$ represent the similarity differentiation ability of grammatical similarity and semantic similarity in equation (5), respectively. p is a parameter greater than 0. $EdSim(sn_i, sn_j)$ and $WtSim(sn_i, sn_j)$ represent the syntactic similarity and semantic similarity between attributes sn_i and sn_j. Based on the above contents, in Figure 3, the basic flow of the research on the multi-pattern data matching model (BSM) based on Simhash and mixed similarity is shown.

x, z are the two thresholds set in Figure 3. $dp_{sim}(WdSim)'$ and $dp_{sim}(WtSim)'$ are changed $dp_{sim}(EdSim)$ and $dp_{sim}(WtSim)$. Through the multi-pattern data matching model, the matching accuracy of multi-pattern data can be effectively improved, thus realising the integration of education resources in the cloud platform.

3.2 *Construction of entity matching model based on RVM*

In the education resources of the cloud platform, in addition to the problem of pattern conflict, there is also the problem of data redundancy. Two sets $R_1 = \{t_1, t_2, \ldots, t_n\}$ and $R_2 = \{s_1, s_2, \ldots, s_n\}$ are supposed. If there are two sample data in these two datasets, namely t_i and s_j. The description of these two sample data is different, but they represent the same entity, which will lead to data redundancy. The entity matching technology can solve the problem of data redundancy through the matching relationship between sample data and (t_i, s_j). Before entity matching, entity segmentation is required. That is, the sample data with high matching probability in the original dataset is allocated to the same entity block. The data with mismatched probability will be filtered out, so as to simplify the candidate matching set and improve the matching efficiency. Figure 4 is the entity block.

Figure 4 Entity blocking process (see online version for colours)



If there are two sample data pairs set X^M and X^U, the sample data pairs in X^M are labelled as matching. The sample data in X^U is labelled as non-matching. For a common attribute p_k of sample data t_i and s_j in these two datasets, its attribute differentiation ability needs to be calculated. That is, the ability to distinguish whether sample data pairs match in equation (6).

$$df_{pk} = \frac{\sum_{x_i \in X^M} x_i + \sum_{x_j \in X^U} 1 - x_i}{|X^M| + |X^U|} \tag{6}$$

x is the similarity $sim(t_i \cdot p_k, s_j \cdot p_k)$ of the attribute value corresponding to p_k in equation (6). In entity matching, a classification model is usually needed. However, when the number of tagged data samples is too large, the complexity of obtaining tagged data will greatly increase. To solve this problem, an unsupervised self-learning entity matching algorithm (US-EM) based on Euclidean locally sensitive hash (E2LSH) algorithm and correlation vector machine (RVM) is proposed in Figure 5.

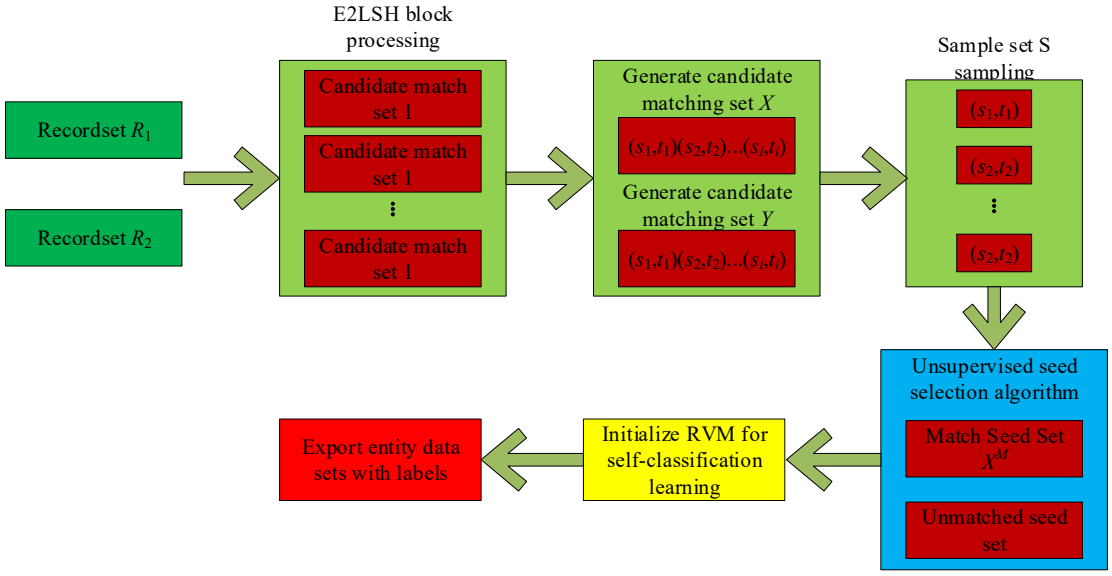
Before using the E2LSH algorithm to block entity records, it is necessary to extract their feature vectors in Figure 5. In this process, an entity signature will be produced, as shown in equation (7).

$$v = \sum_{k=1}^{m} hash\left(s \cdot p_k\right) \tag{7}$$

v is a signature with dimension d in equation (7). $s \cdot p_k$ is the attribute value corresponding to the k attribute p_k of entity record s. $hash(s \cdot p_k)$ is a vector with dimension d. m is the number of attributes of the entity. After the block processing, the entity records that are higher than the set threshold similarity are divided into the same block and the candidate matching set is produced. From this set, the training sample set S is obtained through random sampling. In the experiment, matching pairs were selected as matching seeds in S. Among them, the matching seed set with all dimensions of similarity vector close to 1 is marked as X^M. The matching seed set with all dimensions of similarity vector close to 0 is marked as X^U. Different attributes have different attribute differentiation capabilities. Therefore, when selecting seeds, it should consider the distinguishing ability of attributes and equate them to attribute weights. The weight calculation and distribution of attributes are realised through K-means. Its basic principle is to divide the unlabeled data in S into two categories. One is the matching seed set X^M, which takes the complete matching as the centroid. The other is the unmatched seed set X^U, which takes complete unmatched as the centroid. The smaller the sum of distances within the cluster, the greater the weight of the attribute. Equation (8) shows the calculation of attribute weight.

$$w_j = \begin{cases} \dfrac{1}{\left|\left\{p_i : D_i = 0\right\}\right|} & D_j = 0 \\ 0 & D_j \neq 0 \wedge \left|\left\{p_{i \neq j} : D_i = 0\right\}\right| \neq 0 \\ \dfrac{1}{\sum_{k=1}^{n}\left(D_j / D_k\right)} & \text{else} \end{cases} \tag{8}$$

Figure 5 Unsupervised self-learning entity matching algorithm (see online version for colours)



The calculation of D_j in equation (8) is shown as follows.

$$D_j = D_j^M + D_j^U \tag{9}$$

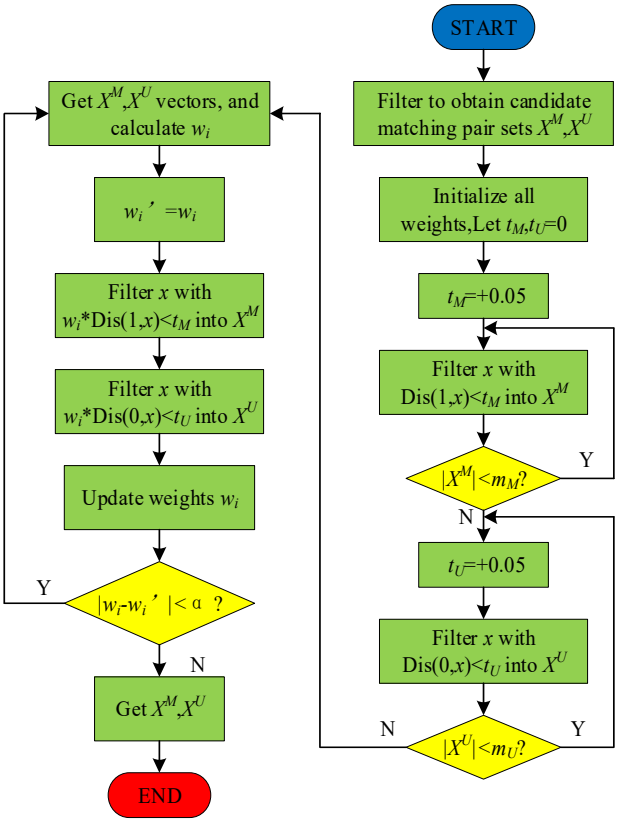
D_j^M is the sum of the similarity of all records in X^M to the attribute value on attribute p_j and the distance of 1 in equation (9). D_j^U is the sum of the similarity of all records in X^U to the attribute value on attribute p_j and the distance of 0. Based on the above, an unsupervised seed selection algorithm is proposed in Figure 6.

α is the set threshold in Figure 6. A sample set that meets the requirements can be obtained through Figure 6. Finally, the data in the sample set is input into an initialised RVM, and self-learning is performed on the RVM, so as to classify these tagged samples. In the process of self-learning, when the number of seed samples used for training is small, the recognition and matching accuracy of RVM classification model will be low. To solve this problem, it is necessary to obtain the matching reliability of RVM model and improve the accuracy of classification and matching according to the matching reliability. See equation (10) for the calculation method of matching reliability.

$$c = \frac{1}{\sqrt{\dfrac{\sum_{x^{m'} \in X^{M'}} \left(x^{m'} - 1\right)^2 + \sum_{x^{u'} \in X^{U'}} \left(x^{u'}\right)^2}{|X^{M'}| + |X^{U'}|}}} \tag{10}$$

$X^{M'}$ and $X^{U'}$ represent the labelled entity pair set after classification of the unlabeled subsequent matching set in equation (10). $|X^{M'}|$ and $|X^{U'}|$ represent the number of samples present in $X^{M'}$ and $X^{U'}$, respectively. $x^{m'}$ and $x^{u'}$ represent the similarity measure of record pairs in $X^{M'}$ and $X^{U'}$, respectively.

Figure 6 Unsupervised seed selection algorithm (see online version for colours)



4 Effectiveness analysis of data integration method of education resources

4.1 Effect analysis of multimodal data matching model

The BSM model is studied and constructed to realise the pattern matching of heterogeneous data. The experimental data comes from the relevant data of painting teaching on the online education cloud platform for art majors in a university. This includes students' learning data (data source 1), teachers' teaching data (data source 2), students' evaluation data (data source 3), and teachers' evaluation data (data source 4). Among these data sources, the data schema is different, and there are redundant attributes with different names and the same semantics in each system. The performance of BSM model is evaluated by accuracy, recall, F1, running time, overall and other indicators. The BSM model is compared with the two most widely used multi-pattern data matching models, including pattern information based matching model (BATT) and data instance based matching model (BINS). In Table 1, the average value is taken from four tests of BSM model. The precision of BSM model is 0.953, 0.021 and 0.058 higher than BATT and BINS respectively. The recall of BSM model is 0.962, which is 0.030 and 0.049 higher than BATT and BINS respectively. F1 of BSM model is 0.929, 0.020 and 0.037 higher than BATT and BINS respectively. The performance of BSM model in heterogeneous data pattern matching is better than BATT model and BINS model.

The comprehensiveness and runtime of BSM model, BATT model and BINS model are shown in Figure 7. The overall of BSM model reached the highest level after 102 iterations, and it is 90.72% in Figure 7(a). The BATT model has the highest overall after 164 iterations, 62 times more than the BSM model. And it is 90.31% overall, 0.41% lower than the BSM model. After 165 iterations, BINS model has the highest overall, 63 times more than BSM model. And it is 89.68% overall, 1.04% lower than BSM model. In Figure 7(b), when the amount of data reaches 10,000, the time required by BSM model is 0.60 s, which is 0.63 s and 0.78 s less than that of BATT and BINS respectively. These data can show that the BSM model proposed in the study is more comprehensive and efficient than BATT model and BINS model in heterogeneous data pattern matching.

Table 1 Precision, recall and F1 values of BSM model.

Model	*Index*	*Number of experiments*				*Average*
		1	*2*	*3*	*4*	
BSM	Precision	0.966	0.963	0.935	0.947	0.953
	Recall	0.956	0.981	0.952	0.958	0.962
	F1	0.920	0.930	0.932	0.933	0.929
BATT	Precision	0.926	0.919	0.923	0.920	0.922
	Recall	0.938	0.940	0.925	0.927	0.932
	F1	0.911	0.910	0.907	0.909	0.909
BINS	Precision	0.902	0.898	0.893	0.886	0.895
	Recall	0.914	0.915	0.910	0.911	0.913
	F1	0.887	0.908	0.892	0.882	0.892

The ROC of the BSM model is shown in Figure 8. The AUC value of BSM model is 0.971, which is 0.08 and 0.15 higher than BATT and BINS respectively.

Figure 7 Completeness and running time of the model, (a) overall (b) time consuming (s) (see online version for colours)

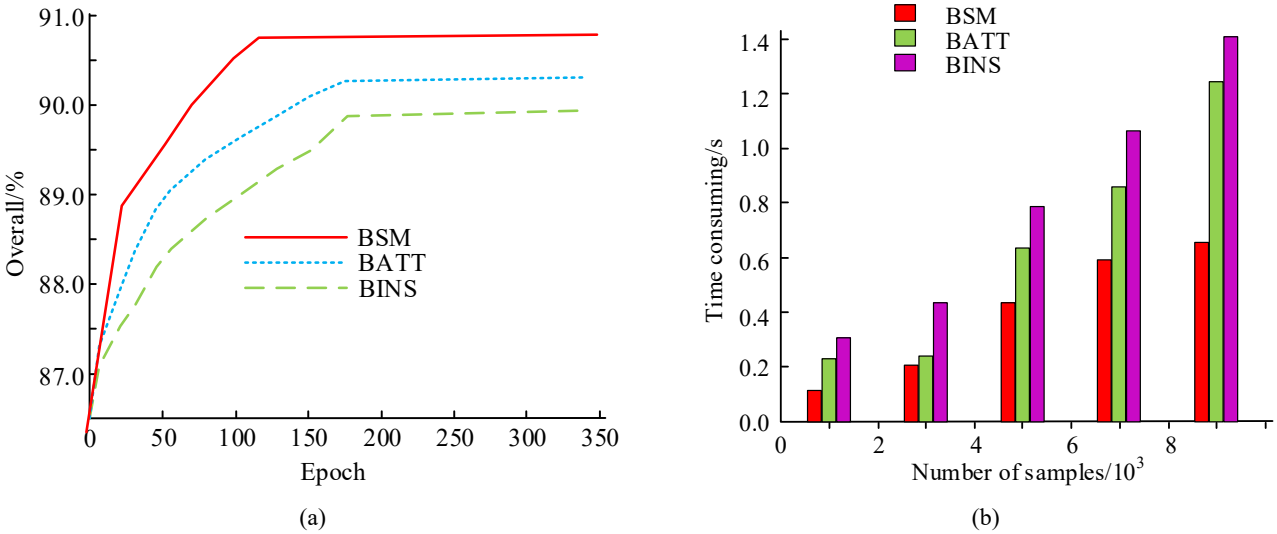
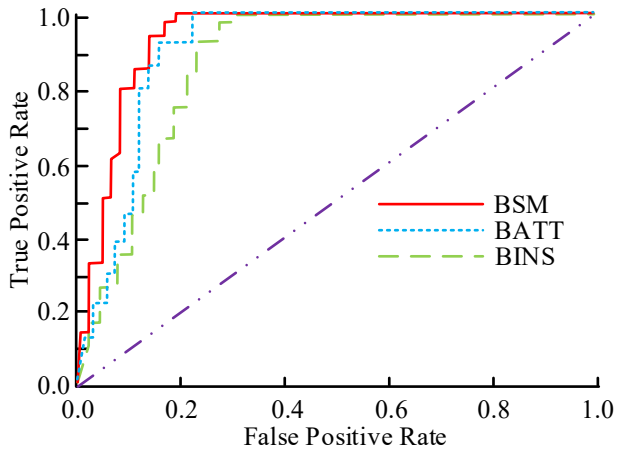


Figure 8 Comprehensiveness and uptime of BSM, BATT and BINS (see online version for colours)



4.2 Performance analysis of entity matching model

The US-EM model is researched and constructed to realise the entity matching of online education resources of painting teaching for art majors in the cloud platform. The data from the online education cloud platform of a university is used for the experiment. The obtained data is divided into two datasets, one of which contains 3,680 records, 18 attributes and 2,260 entities. The other dataset contains 3,032 records, 18 attributes and 2,260 entities. In the experiment, the US-EM model is compared with the two most advanced entity matching models at present, including the entity matching algorithm based on mixed similarity measure (MSM-EM) and the entity matching model based on integrated learning CatBoost method (CatBoost-EM). The precision and recall of the above three models are shown in Figure 9. The US-EM model has the highest precision in 60 iterations, which is 36 and 41 times less than MSM-EM model and CatBoost-EM model respectively in Figure 9(a). At this time, the precision of US-EM model is 97.78%, which is 0.11% and 0.26% higher than MSM-EM model and CatBoost-EM model, respectively. In Figure 9(b), when the US-EM model iterates 64 times, the recall reaches the highest, which is 22 and 28 times less than the MSM-EM model and CatBoost-EM model, respectively. At this time, the recall of US-EM model is 94.62%, which is 0.24% and 0.30% higher than MSM-EM model and CatBoost-EM model respectively. The above results can prove that the US-EM model based on RVM and E2LSH has better performance.

F1 and running time of US-EM model, MSM-EM model and CatBoost-EM model are shown in Figure 10. F1 of US-EM model, MSM-EM model and CatBoost-EM model is positively correlated with the number of iterations in Figure 10(a). However, after the number of iterations reaches a certain number, F1 stops growing and tends to stabilise. When the US-EM model is iterated to 52 times, its F1 reaches the highest and tends to be stable, and finally it is stable at 95.14%. When the MSM-EM model iterates to 92 times, its F1 reaches the highest and tends to be stable, which is 40 times more than the US-EM model. Finally, the F1 of MSM-EM model was stable at 94.48%, 0.66% lower than that of US-EM model. When CatBoost-EM model iterates to 103 times, its F1 reaches the highest and tends to be stable, which is 51 times more than US-EM model. Finally, the F1 of CatBoost-EM model is stable at 94.20%, which is 0.94% lower than that of US-EM model. In Figure 10(b), when the sample data reaches 14,000, the US-EM model takes 1.4 s, which is 0.2 s and 0.4 s less than MSM-EM model and CatBoost-EM model, respectively. The above data can reflect the performance of entity matching models such as US-EM model, MSM-EM model and CatBoost-EM model to a certain extent. This proves that compared with the other two models, the US-EM model proposed in the study has better performance in entity matching.

Figure 9 (a) Precision and (b) recall of the model (see online version for colours)

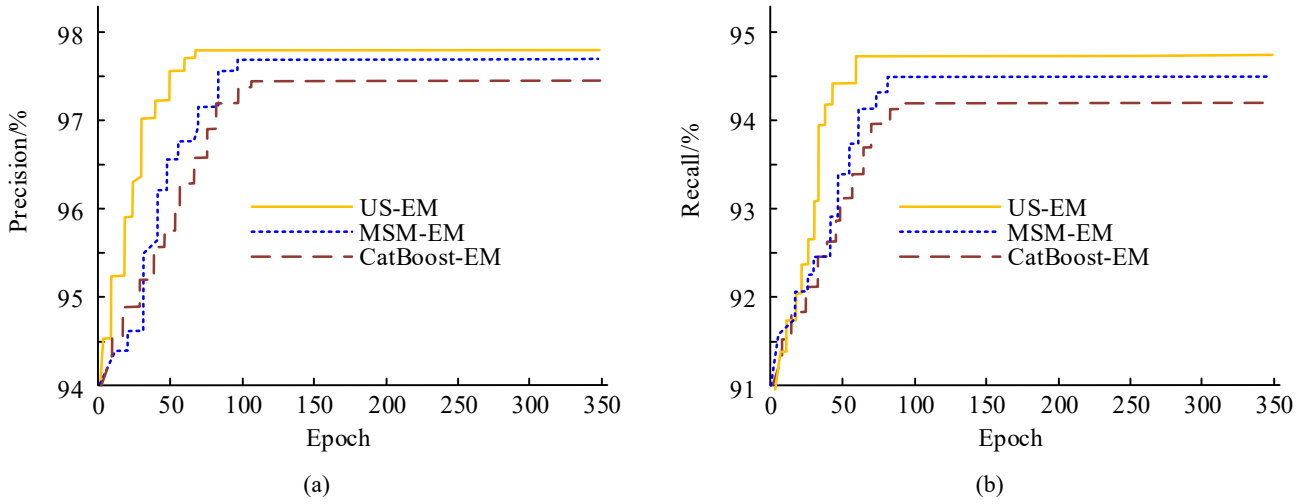
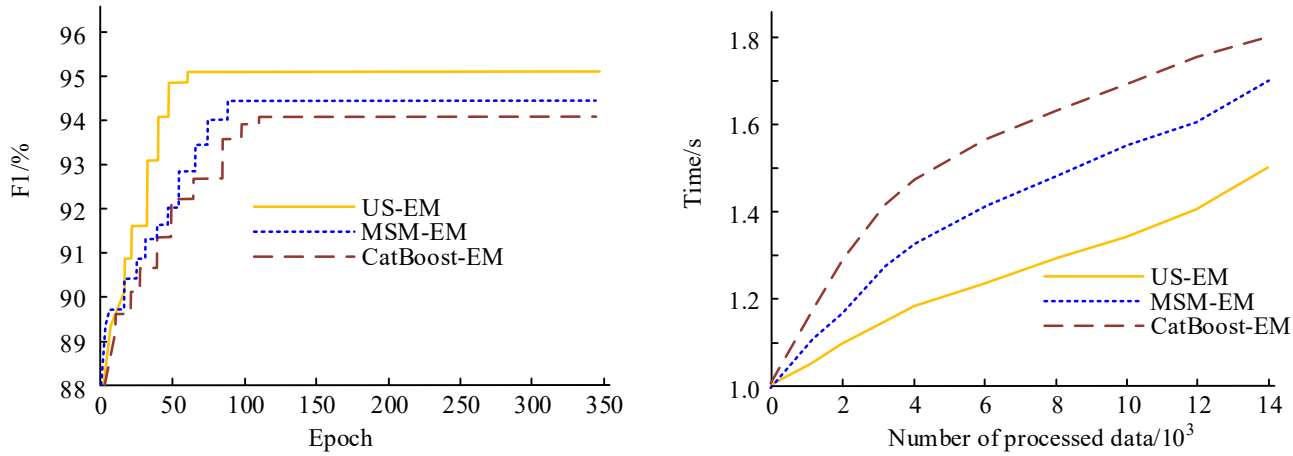


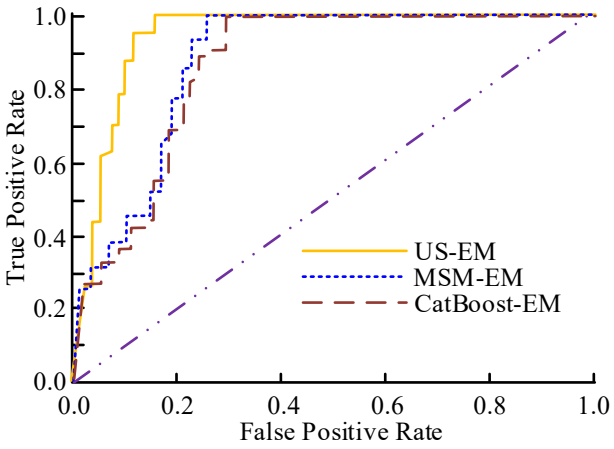
Figure 10 (a) F1 and (b) running time of US-EM model, MSM-EM model and CatBoost-EM model (see online version for colours)



(a) (b)

ROC curves of US-EM model, MSM-EM model and CatBoost-EM model are shown in Figure 11. The AUC value of US-EM model is 0.968 in Figure 11, which is 0.083 and 0.112 higher than MSM-EM model and CatBoost-EM model, respectively. Based on the above contents, the integration method has very excellent results, which is proposed for online education resources of painting teaching for art majors based on cloud platform. This method can complete data integration with high efficiency and precision, thus improving the teaching effect rate and improving the students' performance.

Figure 11 ROC curve of US-EM model, MSM-EM model and CatBoost-EM model (see online version for colours)



5 Conclusions

In the online education of painting teaching for art majors based on cloud platform, there is a large amount of education resource data. However, these data sources, models and structures are different and difficult to be effectively used. Therefore, the data integration method is proposed. Firstly, a PMI-Simhash algorithm combined with a BSM model of mixed similarity is proposed to achieve multi-pattern data matching. Then a US-EM model based on E2LSH algorithm and RVM is proposed to realise entity matching. The precision of BSM model is 0.953, which is 0.021 and 0.058 higher than BATT and BINS respectively. Its recall is 0.962, which is 0.030 and 0.049 higher than BATT and BINS respectively. Its F1 is 0.929, which is 0.020 and 0.037 higher than BATT and BINS respectively. The overall value of BSM model is 90.72%, which is 0.41% higher than BATT model and 1.04% higher than BINS model. When the data volume reaches 10,000, the BSM model takes 0.60 s, which is 0.63 s and 0.78 s less than BATT and BINS respectively. The AUC value of BSM model is 0.971, which is 0.08 and 0.15 higher than BATT and BINS respectively. The precision of US-EM model is 97.78%, which is 0.11% and 0.26% higher than MSM-EM model and CatBoost-EM model respectively. Its recall is 94.62%, which is 0.24% and 0.30% higher than MSM-EM model and CatBoost-EM model respectively. F1 of US-EM model is 95.14%, which is 0.66% higher than MSM-EM and 0.94% higher than CatBoost-EM. When the sample data reaches 14,000, the US-EM model takes 1.4 s, which is 0.2 s and 0.4 s less than MSM-EM model and CatBoost-EM model, respectively. The AUC value of US-EM model is 0.968, which is 0.083 and 0.112 higher than MSM-EM model and CatBoost-EM model respectively. To sum up, the data integration method proposed in the study has a good effect, which can improve the utilisation effect of educational resources and improve student performance. The study did not discuss the application effect of this data integration method in other scenarios, and the test scope needs to be expanded in the follow-up study.

Acknowledgements

The research is supported by: General Project of Philosophy and Social Science Research Project of Jiangsu Universities in 2021 Research on the Implementation Path and Strategy of Integrated Aesthetic Education in Colleges and Universities in the New Era under the idea of 'Life Aesthetic Education' Project No.: 2021SJA0491.

References

Argelaguet, R., Cuomo, A.S.E., Stegle, O. and Marioni, J.C. (2021) 'Computational principles and challenges in single-cell data integration', *Nature Biotechnology*, Vol. 39, No. 10, pp.1202–1215.

Boehm, K.M., Khosravi, P., Vanguri, R., Gao, J.J. and Shah, S.P. (2022) 'Harnessing multimodal data integration to advance precision oncology', *Nature Reviews Cancer*, Vol. 22, No. 2, pp.114–126.

Canzler, S., Schor, J., Busch, W., Schubert, K., Rolle-Kampczyk, U.E., Seitz, H., Kamp, H., Bergen, M.V., Buesen, R. and Hackermüller, J. (2020) 'Prospects and challenges of multi-omics data integration in toxicology', *Archives of Toxicology*, Vol. 94, No. 6, pp.371–388.

Cao, Y., Yu, F. and Tang, Y. (2020) 'A digital watermarking encryption technique based on FPGA cloud accelerator', *IEEE Access*, Vol. 8, No. 1, pp.11800–11814.

Castro, M.D.B. and Tumibay, G.M. (2021) 'A literature review: efficacy of online learning courses for higher education institution using meta-analysis', *Education and Information Technologies*, Vol. 26, No. 2, pp.1367–1385.

Chen, S., Moinzadeh, K. and Tan, Y. (2021) 'Discount schemes for the preemptible service of a cloud platform with unutilized capacity', *Information Systems Research*, Vol. 32, No. 3, pp.967–986.

Ganesan, T., Jothi, R.A. and Vellaiyan, P.. (2023) 'A comprehensive survey on recommender system techniques. *International Journal of Computational Systems Engineering*, Vol. 7, Nos. 2–4, pp.146–158.

Huang, J., Zhou, J., Luo, Y., Yan, G., Liu, Y., Li, H.L., Yan, L.B., Zhang, G.H., Fu, Y.Q. and Duan, H. (2020) 'Wrinkle-enabled highly stretchable strain sensors for wide-range health monitoring with a big data cloud platform', *ACS Applied Materials & Interfaces*, Vol. 12, No. 38, pp.43009–43017.

Jin, H., Fu, Y., Yang, G. and Zhu, X.L. (2020) 'An intelligent scheduling algorithm for resource management of cloud platform', *Multimedia Tools and Applications*, Vol. 79, No. 7, pp.5335–5353.

Jingzhu, W. and Yuanyue, W. (2020) 'Empirical research on user satisfaction of national public culture cloud platform', *Information and Documentation Services*, Vol. 41, No. 4, pp.30–38.

Luecken, M.D., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M.F., Strobl, D.C., Zappia, L., Dugas, M., Colomé-Tatché, M. and Theis, F.J. (2022) 'Benchmarking atlas-level data integration in single-cell genomics', *Nature Methods*, Vol. 19, No. 1, pp.41–50.

Nargesian, F., Asudeh, A. and Jagadish, H.V. (2021) 'Tailoring data source distributions for fairness-aware data integration', *Proceedings of the VLDB Endowment*, Vol. 14, No. 11, pp.2519–2532.

Ping, O. (2023) The construction of college students' job recommendation model based on improved k-means-CF', *International Journal of Computational Systems Engineering*, Vol.7, Nos. 2–4, pp.190–198.

Raghavan, S., Simon, B.Y.L., Lee, Y.L., Tan, W.L. and Kee, K.K. (2019) 'Data integration for smart cities: opportunities and challenges', *Computational Science and Technology: 6th ICCST 2019, Kota Kinabalu, Malaysia*, 29–30 August, Vol. 2020, pp.393–403.

Robertson, A.S., Malone, H., Bisordi, F., Fitton, H., Garner, C., Holdsworth, S., Honig, P., Mathias Hukkelhoven, M., Kowalski, R., Milligan, S., O'Dowd, L., Roberts, K., Rohrer, M., Stewart, J., Taisey, M., Thakkar, R., Baelen, K.V. and Wegner, M. (2020) 'Cloud-based data systems in drug regulation: an industry perspective', *Nature Reviews Drug Discovery*, Vol. 19, No. 6, pp.365–366.

Rostami, A., Akhoondzadeh, M. and Amani, M. (2022) 'A fuzzy-based flood warning system using 19-year remote sensing time series data in the Google Earth Engine cloud platform', *Advances in Space Research*, Vol. 70, No. 5, pp.1406–1428.

Vanama, V.S.K., Mandal, D. and Rao, Y.S. (2020) 'GEE4FLOOD: rapid mapping of flood areas using temporal Sentinel-1 SAR images with Google Earth Engine cloud platform', *Journal of Applied Remote Sensing*, Vol. 14, No. 3, pp. 034505-034505.

Verawardina, U., Asnur, L., Lubis, A.L. and Hendriyani, Y. (2020) 'Reviewing online learning facing the COVID-19 outbreak', *Journal of Talent Development and Excellence*, Vol. 12, No. 3s, pp.385–392.

Xiao, L., Zhang, F. and Zhao, F. (2022) 'Large-scale microbiome data integration enables robust biomarker identification', *Nature Computational Science*, Vol. 2, No. 5, pp.307–316.

Yao, F., Pu, C. and Zhang, Z. (2021) 'Task duplication-based scheduling algorithm for budget-constrained workflows in cloud computing', *IEEE Access*, Vol. 9, No. 3, pp.37262–37272.

Yao, Y. (2021) 'Research on blended learning of higher vocational English based on cloud platform', *Advances in Educational Technology and Psychology*, Vol. 5, No. 2, pp.74–79.

Younas, A., Pedersen, M. and Durante, A. (2020) 'Characteristics of joint displays illustrating data integration in mixed-methods nursing studies', *Journal of Advanced Nursing*, Vol. 76, No. 2, pp. 676-686.

Zhai, Y. and Han, P. (2022) 'Data integration with oracle use of external information from heterogeneous populations', *Journal of Computational and Graphical Statistics*, Vol. 31, No. 4, pp.1001–1012.