Research Article

Construction of Alumni Information Analysis Model Based on Big Data

Xue Wang

Foreign Language Department, Shanghai Xingjian College, Shanghai 200072, China

Correspondence should be addressed to Xue Wang; wangxue6636@163.com

Received 25 May 2022; Revised 27 June 2022; Accepted 2 July 2022; Published 9 August 2022

Academic Editor: Man Fai Leung

Copyright © 2022 Xue Wang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In order to integrate and utilize alumni resources in a better way, big data is utilized to construct alumni information analysis model based on improved hierarchical clustering algorithm, so as to realize mining and retrieval of alumni information. First, the basic principle of hierarchical clustering algorithm is analyzed concretely. Moreover, the improvement is performed on this basis, and a method of calculating the distance between class clusters based on the ant colony optimization is proposed, which uses the shortest distance of the ant colony algorithm to optimally solve the distance between hierarchical class clusters, so as to improve the clustering accuracy. Then, the alumni information analysis model based on improved hierarchical clustering algorithm is constructed, and the model is divided into text preprocessing, keyword extraction, text feature vector generation, name disambiguation, and alumni recognition modules. Finally, the improved hierarchical clustering algorithm and construction of model are verified by experiments. The results show that the accuracy of the improved agglomerative hierarchical clustering algorithm is as high as 86.4% on average and 3.8% and 4.8% more than the two traditional algorithms. Thus, the clustering effect of the algorithm is better, and the proposed alumni analysis model can effectively process text disambiguation of web pages and identification of alumni information, which has certain effectiveness.

1. Related Work

For school, alumni resources are the favorable support and key to school development and construction. The effective integration and tracking of alumni information plays a crucial role for schools. For example, schools can evaluate teaching quality according to the alumni information, so as to continuously improve and perfect their teaching methods and concepts. Furthermore, the latest trends of alumni may be able to provide support and help for the development and construction of school. Therefore, it is necessary for school to manage alumni resources and information effectively. However, there are so many graduates, that it seems impractical to keep track of every alumni.

In recent years, with the development and application of big data and Internet, it has become possible to obtain alumni information and update it in real time. At the same time, the Internet information is very large, so how to accurately identify alumni-related information is the current challenge. In addition, how to quickly and accurately identify alumni information from many pieces of information and exclude the people with the same name to obtain the final target is the current urgent problem to be solved. Barnea Avner et al. constructed an emergency intelligence analysis model based on big data and utilized big data analysis technology to analyze and predict various emergencies in advance, which has certain effectiveness [1–3]. Li Hong et al. studied the information analysis model based on complex network deeply and then used Internet technology to classify and identify complex information on the network. Thus, the accuracy of information analysis is improved [4–6]. Jiho Lee et al. proposed a hierarchical clustering analysis algorithm to construct a learner emotion analysis model for learning experience text and classified learners through hierarchies. Thus, the emotion analysis of each learner is realized [7, 8]. Agnivesh et al. applied clustering algorithm to data classification, which showed good performance and effect.
Hierarchical Clustering Algorithm. Hierarchical clustering algorithm is one of the common algorithms in clustering algorithm, and its basic principle is clustering through the distance between objects [11]. This algorithm can also be called tree clustering, which is mainly divided into two forms of agglomeration and split.

Among them, agglomerative clustering means to calculate the distance between class clusters and merge its various categories. The clustering form is bottom-up. Multiple iterations and merges are carried out on a cluster. When all the data are concentrated in one cluster and the set standard is reached, the calculation can be completed [12]. The basic principle is shown in Figure 1.

The clustering form of split cluster is opposite to that of agglomerative clustering, and its split form is top-down. New clusters are obtained by repeatedly decomposing a data set.

In hierarchical clustering, each cluster is classified by evaluation criteria and usually by means of distance between points and distance between class clusters [13]. The calculation formulas of the two methods are as follows.

### 2.1.1. Calculation Method of the Distance between Points.
If there are two $n$-dimensional vector data points $A, B, A = \{A_1, A_2, \ldots, A_n\}$ and $B = \{B_1, B_2, \ldots, B_n\}$, then the distance between two points can be calculated by cosine similarity, that is, to figure out the cosine value of the included angle [14]. The calculation formula is as follows:

$$\cos \theta = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}.$$ (1)

Euclidean distance can measure the absolute distance between two vectors. The solution formula is as follows:

$$d(A, B) = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}.$$ (2)

### 2.1.2. Measuring the Distance between Class Clusters.
At present, clustering is mainly achieved by calculating the distance between class clusters. The common method for measuring the distance is to solve the minimum, maximum, average value, and average distance of class clusters, which are shown in formulas (3–6):

$$d_{\min}(c_i, c_j) = \min \{\|p_i - p_j\|\},$$ (3)

$$d_{\max}(c_i, c_j) = \max \{\|p_i - p_j\|\}.$$ (4)

Here, $(p_i \in c_i, p_j \in c_j)$.

$$d_{\text{mean}}(c_i, c_j) = \max \{|m_i - n_j|\},$$ (5)

where $m_i, n_j$ and $m_i, n_j$ are the centroids of $c_i, c_j$ and $c_i, c_j$, respectively.

$$d_{\text{avg}}(c_i, c_j) = \frac{1}{n_i, n_j} \sum \|p_i - p_j\|,$$ (6)

where $(p_i \in c_i, p_j \in c_j)$, $n_i, n_j$, and $n_i, n_j$ are the sample numbers of class $c_i, c_j$ and $c_i, c_j$, respectively.

### 2.2. Algorithm Improvement.
The traditional hierarchical clustering algorithm is simple, and the accuracy of distance calculation is not high, which is easy to be interfered by outliers in the class cluster. Therefore, to better represent the distance between clusters, avoid interference, and improve the clustering effect, the ant colony optimization algorithm is added to the agglomerative hierarchical algorithm. Based on the pheromone characteristics of the ant colony optimization algorithm, the optimal path is solved, so as to improve the clustering accuracy and achieve global optimization [15].

#### 2.2.1. Ant Colony Optimization Algorithm.
The basic principle of the ant colony optimization is global search, and it is an intelligent optimization algorithm. Simulating real ant behavior, the algorithm is constantly improved and optimized to identify the direction through pheromone concentration, so as to find the optimal path and achieve global optimization [16].

The ant colony optimization algorithm can also be called traveling salesman problem (TSP). If there are $n$ cities, TSP will implement the algorithm.
(1) When the ant colony reaches \( n \) cities, ants will update pheromones in the path, and the update formulas are as follows:

\[
\tau_{ij}(t+n) = \rho \tau_{ij}(t) + \Delta \tau_{ij}
\]

\[
\Delta \tau_{ij} = \sum_{k=1}^{m} Q \Delta \tau_{ij}^k,
\]

where \( m \) represents the total number of ants and \( Q \) represents the constant. \( \rho < 1 \), \( L_k \) is the distance between two points; \( \tau_{ij}(t) \) represents the pheromone on edge \((i, j)\) at time \( t \); \( \Delta \tau_{ij}^k \) is the pheromone quantity generated by ant \( k \) on edge \((i, j)\) between time \( t \) and \( t + n \) [17].

(2) The ant colony cannot return to the previous city until it reaches \( n \) cities.

(3) The ant colony starts from a certain point, and the probability of selecting the next target is

\[
p_{ij}^k(t) = \frac{[\tau_{ij}(t)]^{\alpha} [\eta_{ij}(t)]^{\beta}}{\sum_{l \in T_k} [\tau_{il}(t)]^{\alpha} [\eta_{il}(t)]^{\beta}},
\]

where \( T_k \) is the city that ants can choose; \( n_{ij} \) is the heuristic information; \( \alpha \) is the relative importance of pheromone heuristic factor; and \( \beta \) is the relative importance of heuristic information.

2.2.2. Agglomerative Hierarchical Clustering Algorithm Based on the Ant Colony Optimization

(1) Standard Distance. Euclidean distance is used to figure out the distance between two data points, and the solution formula is as follows:

\[
d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{im} - x_{jm})^2},
\]

where \( x_i \) and \( x_j \) represent two \( m \)-dimensional data points.

The similarity can be measured by calculating the distance between two clusters, and the distance between clusters can be calculated by the minimum distance formula (3) in agglomerative hierarchical clustering.

(2) Objective Function. Objective function is set as the clustering error square sum (suppose there are \( c \) clustering centers after clustering is completed):

\[
E = \sum_{l=1}^{c} \sum_{x \in c_l} (x - c_l)^2, c_j = \frac{1}{m_j} \sum_{i=1}^{m_j} x_i,
\]

where \( C_j \) represents the centroid, which is calculated by a specific cluster \( j \) and \( m_j \) represents the amount of data in the cluster.

(3) Agglomerative Hierarchical Clustering Based on the Ant Colony Optimization. The objective of this algorithm is to find a shortest path in all the data to improve the clustering efficiency and accuracy. Utilizing the ant optimization, ants are taken as the research object, and food is taken as the clustering center. The probability of ants searching for food is put into the clustering algorithm, and data are classified by probability [18].

There are six steps improving algorithm, and the specific steps are as follows.

It can be seen from Figure 2 that the process of improving algorithm is mainly divided into six steps, which are as follows:

(1) Initialize parameters, such as the number of ants \( m \), weight parameter \( \alpha \), and volatile factor \( \rho \).

(2) Set an ant as \( m \), calculate the distance and pheromone between data points, evaluate the transition probability, and determine the merger probability between data points and alternative points. The merger probability formula is as follows:

\[
p_{ij}^m = \frac{(\tau_{ij})^\alpha (\eta_{ij})^\beta}{\sum_{l \in N_j} (\tau_{il})^\alpha (\eta_{il})^\beta},
\]

where \( d_{ij} \) is the distance between two data points; \( n_{ij} \) is the distance-based heuristic information; \( \alpha \) and \( \beta \) are weight parameters, which have a great influence on pheromones [19].

If \( p_{ij}^m \geq p_0 \), \( x_j \) merges with \( x_i \); otherwise, there is no merger.

(3) Judge whether the number of ants \( k \) reaches the total number of ants; if not, set \( k = k + 1 \), and go back to Step (2) for calculation.

(4) After ants complete clustering, the clustering center and pheromone will be updated, where the expression of clustering center is

\[
c_j = \frac{1}{m_j} \sum_{i=1}^{m_j} x_i,
\]

where \( m_j \) is the total amount of data points classified in \( C_j \).

In the process of optimization, the pheromone concentrations in the paths passed by ant are different [20]. After clustering, ants’ pheromones will constantly evaporate, and the evaporation formula can be expressed as follows:

\[
\tau_{ij} = (1 - \rho) \tau_{ij},
\]
where $\rho$ represents pheromone evaporation rate. When $0 < \rho \leq 1$, it can enable the algorithm to delete long paths and avoid pheromone accumulation.

After completing evaporation, all ants will leave pheromones again:

$$
\tau_{ij} = \tau_{ij} + \sum_{k=1}^{m} \Delta \tau_{ij}^k
$$

where $\Delta \tau_{ij}^k$ represents the amount of information from data $x_i$ to class cluster $c_j$ of the $k$th ant.

$$
\Delta \tau_{ij}^k = \frac{1}{d(x_i, c_j)}.
$$

Figure 2: Flow chart of improving algorithm.

(5) Figure out the solution of objective function.

(6) Return to steps (2)–(4) until the minimum target is calculated, and then the calculation can be completed.

3. Model Construction

3.1. Design of the Alumni Recognition Model Based on Improved Algorithm. There are too much alumni information on the Internet. The structure of the web page is complex, and the data format is not standard, which means that it is difficult to extract features of alumni information and fail to accurately identify alumni information. Therefore, an alumni recognition model based on the improved agglomerative hierarchical clustering algorithm is proposed, which is shown in Figure 3. Firstly, alumni information is collected and preprocessed. Secondly, text features are extracted by embedding technology. Moreover, text representation model and feature vector are constructed. Finally, name disambiguation and alumni identification are carried out. Thus alumni information analysis is realized.

As can be seen, alumni information identification process is mainly divided into six steps, as follows:

1. Classify the text data and name entity recognition
2. Filter stop words and delete useless interjections, modal particle, personal pronouns, and so on [21]
3. Extract keywords and use TF-IDF algorithm to select key information
4. For text representation, use the word embedding tool to complete the word embedding of keywords and obtain the vectorization expression of keywords
5. For text clustering, adopt cosine similarity calculation method to cluster document vector
6. Analyze clustering results

3.2. Data Collection and Preprocessing. To perform the data collection and preprocessing, the first step is to collect the web page documents, and the original alumni data are collected by using the Python programming of Baidu’s search engine API. The second step is to specify a person’s name as a search term for the web page retrieval; thus, the web document is obtained and it is stored in a local folder. Then, the web page file is processed by the Python programming, and the web page information is extracted. Furthermore, the text content is extracted by regular expression. The specific work includes tag attribute extraction, tag filtering, and character filtering [22].

3.3. Text Feature Extraction. The python package pynlpir based on mlpir natural language processing system of Chinese Academy of Sciences is utilized to annotate text sequences and complete text classification.
The TF-IDF algorithm is used to extract keywords, each information is named, and its entity is taken as the final text keyword. The algorithm flow is in Figure 4.

The TF-IDF algorithm is used to calculate word frequency of text and inverse document frequency; thus, TF-IDF value is obtained to measure the importance of words. The solution formula is as follows:

$$tf - idf = tf^*idf,$$  \hspace{1cm} (16)

where $tf$ (text frequency) represents the occurrence frequency of words in the document, and the solution formula is as follows:

$$tf_{i,j} = \frac{n_{ij}}{\sum_k n_{k,j}},$$ \hspace{1cm} (17)

where $n_{ij}$ is the number of occurrences of word $i$ in document $j$; $\sum k \cap k, j$ is the sum of occurrences of all words in document $j$; and IDF (inverse document frequency) indicates the rarity of the word in the document, namely, the importance of the word. The expression is as follows:

$$idf_i = \log \frac{|D|}{\left| \{ j : c_i \in d_j \} \right|},$$ \hspace{1cm} (18)

Here, $|D|$ is the number of all documents in the corpus and $\left| \{ j : c_i \in d_j \} \right|$ is the number of documents containing word $c_i$. 

![Figure 3: Alumni recognition model based on improved agglomerative hierarchical clustering algorithm.](image-url)
3.4. Construction of Text Representation Model and Feature Vector

3.4.1. Construction of Word2Vec Text Representation Model. Before classifying text, it is necessary to select an appropriate model to represent text. The quality of text representation has a great influence on the text clustering result. However, many text models are deficient in information and incomplete, which leads to poor clustering effect in the later stage. Therefore, word embedding model is used to represent text content, so as to improve the effect of text clustering.

Word embedding model improves text quality by transforming text dimensions and filling in missing information. On this basis, Skip-gram, which has a good application effect at present, is used to train model and generate word vector.

3.4.2. Construction of Text Feature Vector. The construction process of text feature vector is shown in Figure 5. The first step is to use trained Word2Vec model to generate word vector. The second step is to find the average value of keywords; thus, the text feature vector is obtained.

If there are \( n \) keywords in the text, the text representation model \( d = \{c_1, c_2, \ldots, c_n\} \), and the word vector \( v(c_i) \) can be obtained by training. Thus, the feature vector expression of document is

\[
v(d) = \frac{1}{n} \sum_{i=1}^{n} v(c_i),
\]

where \( v(d) \) is the feature vector representation of a document and \( v(c_i) \) is the word vector of the \( i \) th feature word \( c_i \), namely, the average value of sum of the document vector and all the keyword vectors.

The personnel text feature vector to be disambiguated and the text vector of alumni attribute in the knowledge base can be obtained by the above model.

3.5. Name Disambiguation and Alumni Identification. Using text feature vector, text is clustered based on the improved agglomerative hierarchical clustering algorithm. Through clustering, people with the same name is distinguished, and the texts related to the same person are divided into the same category, which achieves name disambiguation. Then, the knowledge base information is used to assist identification to find the class cluster to which a specific person belongs, so as to realize the information identification of specific person. The disambiguation process is shown in Figure 6.

The improved condensed hierarchical clustering algorithm proposed in this paper is utilized to cluster text information, and the central point of each cluster is figured out. The calculation formula is shown in formula (22) [23–27]:

\[
v(C_i) = \frac{1}{m} \sum_{j=1}^{m} v(d)_j,
\]

where \( v(C_i) \) represents the feature vector of the center point of the \( C_i \) th class cluster; \( m \) represents the number of data points in class cluster \( C_i \); and \( v(d)_j \) represents the feature vector of the \( j \) th data point.

4. Experimental Results and Analysis

4.1. Experimental Data. To verify the effectiveness of the improved method, there are 8000 text corpus about sports, government, commerce, culture and scientific research obtained from the Python web crawler as experimental data, and the data set D1 and D2 of alumni with the same name are divided into 5000 training sets and 3000 testing sets.

4.2. Experimental Environment and Parameter Setting. To obtain better experimental results, the experimental processor is the 9th generation of Intel core i5, and the operating system is windows 64-bit. In addition, cas NLPIR word segmentation system and python language are adopted in the experiment.

The parameter value setting of the algorithm is shown in Table 1.

4.3. Evaluation Indicators. To evaluate the experiment more objectively, accuracy, recall rate, and F1 value are adopted as evaluation indicators. The accuracy is the ratio of correct number to total number, and the recall rate is the ratio of correct number to real number of texts. The expression of accuracy and recall rate is as follows:
precision = \frac{a}{a + b}, \quad (21)

\text{recall} = \frac{a}{a + c}, \quad (22)

where precision represents accuracy; recall stands for recall rate; \( a \) is the number of text correctly identified by algorithm; \( b \) is the number of text incorrectly identified by algorithm; and \( c \) is the number of text belonging to a certain category that algorithm does not recognize.

\( F_1 \) value is the comprehensive analysis standard of the clustering algorithm. The higher the value of \( F_1 \) is, the better the clustering effect of algorithm is. Formula of \( F_1 \) value is as follows:

\[ F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{(\text{precision} + \text{recall})} \quad (23) \]

4.4. Improved Algorithm. To test the effectiveness of the proposed improved algorithm, the experiment compares the algorithm before and after improvement with the K-means algorithm. The experimental results are shown in Table 2.

It can be seen from Table 2 that the accuracy of the improved algorithm is 84.2% and 85.6% in D1 and D2 data sets, respectively, which is higher compared to the accuracy of the algorithm before improved and the traditional clustering algorithm, which indicates that the improved algorithm proposed in the paper is effective.

To test the performance of the improved algorithm, the comparison curves of accuracy, recall rate, and \( F_1 \) value of the three algorithms are as follows.

As can be seen from Figure 7, the highest accuracy of the improved algorithm is 89% and the average rate is 86.4%, which are higher compared to those of the other two algorithms by 3.8% and 4.5% respectively, which means that adding the ant colony optimization algorithm can improve the clustering accuracy.

As can be seen from Figures 8 and 9, the highest recall rate of the improved algorithm is 89% and the average rate is 86.5%, which are 3% and 3.5% higher compared to those of the other two algorithms. The \( F_1 \) value of the improved algorithm is as high as 88% and as low as 86%, which is obviously higher compared to that of the other two algorithms. Thus, the improved algorithm has better effect on text classification and better algorithm performance.

4.5. Alumni Recognition Model Based on the Improved Algorithm. To verify the validity of the constructed alumni information recognition and analysis model, there are 5 alumni information selected from the above data set for experimental verification. The Word2Vec tool is adopted to train word vector, and the hyperparameter settings of the model during training are in Table 3.

After name disambiguation and recognition are performed by the alumni information recognition model, the statistical results obtained are shown in Table 4.

As can be seen from Table 4, using this model to identify and classify 5 alumni, alumni with the same name is accurately distinguished, name disambiguation is realized, and the class cluster of 5 alumni is also accurately
Table 2: Comparison results of the accuracy of the three algorithms.

| Data set | Number of samples | Agglomerative hierarchical clustering [28] (%) | K-means (%) | The proposed algorithm (%) |
|----------|-------------------|-----------------------------------------------|-------------|---------------------------|
| D1       | 4000              | 76.3                                          | 79.4        | 84.2                      |
| D2       | 4000              | 78.7                                          | 80.1        | 85.6                      |

Figure 7: Algorithm accuracy.

Figure 8: Recall rate of algorithm.

Figure 9: F1 value of algorithm.
located. Thus, alumni information identification is realized, which shows that the constructed recognition model is feasible.

5. Conclusion

The proposed ant colony optimization clustering algorithm can achieve accurate clustering of alumni information, and the clustering effect is good. Moreover, the constructed model can disambiguate alumni name information and improve the accuracy of alumni information identification. The results show that the improved agglomerative hierarchical clustering algorithm has higher recognition accuracy, recall rate, and F1 value compared to traditional agglomerative hierarchical clustering algorithm and K-means algorithm. The corresponding average rates are 86.4%, 86.5%, and 87%, which shows that adding the ant colony algorithm into traditional agglomerative hierarchical clustering algorithm can figure out the optimal solution. Applying the modified algorithm to the alumni analysis model can further improve the recognition accuracy and clustering effect, which makes the model be extended and applied in the field of alumni analysis. The contribution of this study is to improve hierarchical clustering by using the ant colony algorithm; thus, the accuracy of massive data classification is improved. This method is applied to the analysis of alumni information, which provides a reference for the extension of informatization in various fields.

However, due to the lack of experimental conditions and research experience, there are still certain limitations. The future research is to improve the execution efficiency of the improved hierarchical clustering algorithm, so as to complete text data clustering and analysis in a shorter time.

Data Availability

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Table 3: Hyperparameter settings.

| Hyperparameter name | Hyperparameter description        | Hyperparameter values |
|---------------------|----------------------------------|-----------------------|
| Size                | Adjust the dimension of word vector | 100                   |
| Min-count           | Adjust the minimum frequency value of word | 2                     |
| cbow_mean           | Select the training model         | 1                     |
| Window              | Adjust the size of context window | 5                     |
| Worker              | Adjust the number of computing cores | 10                    |

Table 4: Statistical table of clustering results.

| Personnel | Number of class clusters | Class clusters of alumni |
|-----------|--------------------------|--------------------------|
| Alumni 1  | 22                       | 3                        |
| Alumni 2  | 6                        | 5                        |
| Alumni 3  | 13                       | 7                        |
| Alumni 4  | 34                       | 26                       |
| Alumni 5  | 11                       | 11                       |

Conflicts of Interest

The author declares that there are no conflicts of interest regarding this work.

Acknowledgments

This study was supported by Research Project on the Work of the Communist Youth League in Shanghai Schools in 2020, the Research on Tracking Service Mechanism of the Student Leaders after Graduate by Searching the Past Ones, taking Shanghai Xingjian College as an example, Project Approval No. 2020YB116.

References

[1] B. Meshulach and M. Avi, “Forecasting for intelligence analysis: scenarios to abort strategic surprise,” *International Journal of Intelligence & Counter Intelligence*, vol. 34, no. 1, pp. 106–133, 2021.

[2] S. José Ramón and R. Dag, “A three-stage method for data text mining: using UGC in business intelligence analysis,” *Symmetry*, vol. 11, no. 4, p. 519, 2019.

[3] C. Zhang, “User multi-modal emotional intelligence analysis method based on deep learning in social network big data environment,” *IEEE Access*, vol. 7, pp. 181758–181766, 2019.

[4] Li Hong, Z. Chen, and Q. Guo, “The evolution and approaches of information analysis service of university libraries in China,” *Science & Technology Libraries*, vol. 40, no. 1, pp. 52–64, 2021.

[5] B. Azeddine, M. A. Qureshi, S. A. Amirshahi, A. Chetouani, and M. Pedersen, “A critical analysis on perceptual contrast and its use in visual information analysis and processing,” *IEEE Access*, vol. 8, pp. 156929–156953, 2020.

[6] W. Zhang, K. Yang, and X. Guo, “Design and implementation of the pedestrian information analysis system,” *IOP Conference Series: Materials Science and Engineering*, vol. 439, no. 3, 2018.

[7] J. Lee, H. S. Yang, H. Jeong, J. H. Kim, and H. Yang, “Targeted isolation of lignans from trachelospermum asiaticum using molecular networking and hierarchical clustering analysis,” *Biomolecules*, vol. 10, no. 3, p. 378, 2020.

[8] H. A. H. Naji, Q. Xue, K. Zheng, and N. Lyu, “Investigating the significant individual historical factors of driving risk using hierarchical clustering analysis and quasi-Poisson regression model,” *Sensors*, vol. 20, no. 8, p. 2331, 2020.

[9] A. Agnivesh, R. Pandey, and A. Singh, “Improved K-means map reduce algorithm for big data cluster Analysis[,]” *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 8, pp. 1796–1802, 2019.
[10] S. Ko, “Predicting learning achievement using big data cluster Analysis: focusing on longitudinal study,” *Journal of Digital Contents Society*, vol. 19, no. 9, pp. 1769–1778, 2018.

[11] S. Balakrishna, M. Thirumaran, R. Padmanaban, and V. K. Solanki, “An efficient incremental clustering based improved K-Medoids for IoT multivariate data cluster analysis,” *Peer-to-Peer Networking and Applications*, vol. 13, no. prepub, pp. 1–24, 2019.

[12] F. Nan, Y. Li, X. Jia, L. Dong, and Y. Chen, “Application of improved SOM network in gene data cluster analysis,” *Measurement*, vol. 145, pp. 370–378, 2019.

[13] Y. Wang, “Application of ant colony algorithm and artificial intelligence in training simulation of athletes in sports arena,” *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–10, 2021.

[14] L. Niu and L. Xiong, “Optimisation and application research of ant colony algorithm in vehicle routing problem,” *International Journal of Computing Science and Mathematics*, vol. 13, no. 2, 2021.

[15] Y. Lin and Lu Han, “Diagnosis strategy design method based on ant colony algorithm,” *Journal of Physics: Conference Series*, vol. 1894, no. 1, 2021.

[16] J. Ge, D. Yu, and Y. Fang, “Multi-dimensional QoS cloud computing task scheduling strategy based on improved ant colony algorithm,” *Journal of Physics: Conference Series*, vol. 1848, no. 1, 2021.

[17] X. Tian, L. Li, L. Shuang, D. Zhiping, and P. Ming, “Path planning of mobile robot based on improved ant colony algorithm for logistics,” *Mathematical biosciences and engineering: MBE*, vol. 18, no. 4, pp. 3034–3045, 2021.

[18] K. Bilal, K. Ali, and A. Yassine, “Open capacitated ARC routing problem by hybridized ant colony algorithm,” *RAIRO - Operations Research*, vol. 55, no. 2, pp. 639–652, 2021.

[19] M. Khodabandeh, M. Ghassabzadeh Saryazdi, and A. Ohadi, “Multi-objective optimization of auto-body fixture layout based on an ant colony algorithm,” *Proceedings of the Institution of Mechanical Engineers - Part C: Journal of Mechanical Engineering Science*, vol. 234, no. 6, pp. 1137–1145, 2020.

[20] X. Wang, H. Li, J. Yang, C. Yang, and H. Gui, “Optimal path selection for logistics transportation based on an improved ant colony algorithm,” *International Journal of Embedded Systems*, vol. 13, no. 2, 2020.

[21] J. X. Liu, T. Z. Huang, and X. i. Duan, “Application of improved ant colony algorithm in optimal planning of dynamic distribution network,” *Journal of Physics: Conference Series*, vol. 1533, no. 3, 2020.

[22] L. i. Xiao, J. Wei Yang, and L. i. Nai Qian, “Multi-feature fusion energy-saving routing in internet of things based on hybrid ant colony algorithm,” *International Journal of Innovative Computing and Applications*, vol. 11, no. 2-3, pp. 103–108, 2020.

[23] P. Praveen, M. R. Kumar, M. A. Shaik, R. Ravikumar, and R. Kiran, “The comparative study on agglomerative hierarchical clustering using numerical data,” *IOP Conference Series: Materials Science and Engineering*, vol. 981, no. 2, Article ID 022071, 2020.

[24] P. Shi and Z. Zhao, H. Zhong, H. Shen, and L. Ding, An improved agglomerative hierarchical clustering anomaly detection method for scientific data,” *Concurrency and Computation: Practice and Experience*, vol. 33, no. 6, 2020.

[25] S. Lim, “Sequential batching with minimum quantity commitment in N-level non-exclusive agglomerative hierarchical clustering structures,” *International Journal of Industrial Engineering Computations*, vol. 11, no. 3, pp. 317–340, 2020.

[26] X. Cai and Q. Wang, “Educational tool and active-learning class Activity for teaching agglomerative hierarchical clustering,” *Journal of Statistics Education*, vol. 28, no. 3, pp. 1–15, 2020.

[27] S. Saurabh and S. Gautam, “Modelling and statistical analysis of YouTube’s educational videos: a channel Owner’s perspective,” *Computers & Education*, vol. 128, pp. 145–158, 2019.

[28] P. Garikapati, K. Balamurugan, T. P. Latchoumi, and R. Malkapuram, “A cluster-profile comparative study on machining AISi 7/63% of SiC hybrid composite using agglomerative hierarchical clustering and K-means,” *Silicon*, vol. 13, pp. 1–12, 2020.