International Journal of Data Mining and Bioinformatics

ISSN online: 1748-5681 - ISSN print: 1748-5673
https://www.inderscience.com/ijdmb

Fast retrieval method of biomedical literature based on feature mining

Duo Long, Yunxin Long, Fahuan Xie, Ping Yu, Hui Yan

DOI: 10.1504/IJDMB.2023.10058133

Article History:

Received:	14 February 2023
Last revised:	10 April 2023
Accepted:	19 June 2023
Published online:	17 October 2023

Fast retrieval method of biomedical literature based on feature mining

Duo Long

Suqian University,
Jiangsu, 223800, China
Email: Longduo7113@126.com

Yunxin Long

College of Traditional Chinese Medicine,
Changchun University of Chinese Medicine,
Changchun 130117, China
Email: 1273112667@qq.com

Fahuan Xie and Ping Yu

Jilin Province S&T Innovation Center for Physical Simulation and Security of Water Resources and Electric Power Engineering,
Changchun Institute of Technology,
Changchun, 130012, China
Email: 13596214960@163.com
Email: yuping@ccit.edu.cn

Hui Yan*

Suqian University,
Jiangsu, 223800, China
Email: Yanhui7125@126.com
*Corresponding author

Abstract: In order to solve the problems of large errors, low accuracy of feature mining and time-consuming traditional literature retrieval methods, this paper designs a fast retrieval method for biomedical literature based on feature mining. First, we simulate the document collection space, and collect documents according to the data centroid and probability density function. Secondly, the location of similar data is marked by mutual information method, and the hidden information of literature data is extracted after reducing the imbalance of dataset. Then, the Pearson correlation coefficient of the literature data is calculated and the key features of the literature are mined. Finally, we calculate the expected loss risk of literature data, design a fast retrieval algorithm for biomedical literature, and realise fast retrieval. The test results show that this method can reduce the retrieval error, improve the accuracy of document feature mining, and the retrieval time is shorter.

Keywords: feature mining; biomedical literature; quick search; probability density function; cost matrix.

Reference to this paper should be made as follows: Long, D., Long, Y., Xie, F., Yu, P. and Yan, H. (2023) 'Fast retrieval method of biomedical literature based on feature mining', *Int. J. Data Mining and Bioinformatics*, Vol. 27, No. 4, pp.297–311.

Biographical notes: Duo Long received his Bachelor's in Mechanical Engineering from Jilin Vocational Normal College, Changchun, China in 1993, and Master's in Optics from Changchun University of Science and Technology, Changchun, China in 2005, and PhD in Agricultural Mechanisation Engineering from Jilin University, Changchun, China in 2008, and Post-doctoral degree of Changchun University of Science and Technology, Changchun, China, in 2012. He is currently a Professor in the School of Management, Suqian University, Suqian City, China. He has published over ten international journal papers. His current research interests include smart agriculture, big data, soft science and artificial intelligence.

Yunxin Long currently studied in Changchun University of Chinese Medicine, Changchun, China, majoring in Traditional Chinese Medicine, She has published seven papers. Her current research interests are traditional Chinese medicine, artificial intelligence and big data.

Fahuan Xie received her Bachelor's in Electrical Engineering and Automation from Northeast Electric Power University, Jilin, China in 2021. She currently studied in Changchun Institute of Technology, Changchun, China, majoring in Water Conservancy and Hydropower Project. She has published three papers. Her current research interests are intelligent irrigation, artificial intelligence and big data.

Ping Yu received his Bachelor's in Computer Science and Technology from Harbin Institute of Technology, Harbin, China in 2004, and Master's in Computer Technology from Changchun University of Science and Technology, Changchun, China in 2010. She currently studied in Jilin University, Changchun, China, majoring in Computer Application Technology. She has published over ten international journal papers. She is currently a Lecturer in the School of Computer Technology and Engineering, Changchun Institute of Technology, Changchun, China. She has published over ten international journal papers. Her current research interests include edge computing, artificial intelligence, big data, and computer applications.

Hui Yan received her Bachelor's in Applied Electronics Technology from Jilin Vocational Normal College, Changchun, China in 1993, and Master's in Computer Application Technology from Changchun University of Science and Technology, Changchun, China in 2004, and PhD in Agricultural Mechanisation Engineering from Jilin University, Changchun, China in 2012. She is currently a Professor in the Jilin Province S&T Innovation Center for Physical Simulation and Security of Water Resources and Electric Power Engineering, Changchun Institute of Technology, Changchun, China. She has published over 20 international journal papers. Her current research interests include digital simulation, smart agriculture, big data, and artificial intelligence.

1 Introduction

Biomedical literature is the key reference data that covers a variety of biomedicine. It can quickly and effectively explain the natural language text of biomedicine in a specific way and recognition mode (Zhao et al., 2022). Biomedical literature can be said to be the cornerstone of biomedical progress in human society. The exploration and research of the miracle of life based on the biomedical literature has contributed a key force to the benefit of humankind (Mounica and Lavanya, 2022). With the change of time and the continuous change of social science and technology, the volume of biomedical literature data also shows explosive growth. Biomedical literature covers more and more contents, providing more data for human beings. In this context, rapid retrieval of biomedical literature is the key to its effective use. However, due to the limitation of data volume and external technical conditions, biomedical literature has the problems of slow response and unsatisfactory retrieval results in the retrieval process, which has seriously affected biomedical research and innovation (Yin and Chen, 2022). For this reason, researchers in this field have done a lot of research on the retrieval of relevant biomedical literature and designed many retrieval methods.

In Xie et al. (2021), the research on biomedical literature text joint embedding cross-pattern retrieval method based on deep feature engineering is proposed. In the research of this method, an efficient learning two-stage deep feature engineering framework for semantic enhancement joint embedding is introduced, which separates the deep feature engineering in data preprocessing from the training of text joint embedding model. In the preprocessing, the depth feature engineering is performed by combining the depth feature engineering with the semantic context features derived from the original text input data. Use LSTM to identify key terms, extract deep NLP models from BERT family, TextRank or TF-IDF, and generate ranking scores for key terms before using word2vec to generate vector representation of each key term. WideResNet50 and word2vec are used to extract and encode to help semantic alignment in the joint potential space, and retrieval is realised on this basis. This method has good accuracy for biomedical literature retrieval, but due to the need to not only export semantic contextual features, but also use LSTM to identify key terms, the retrieval process is more complex and the retrieval speed is slower. Chen et al. (2021) proposes a new method for retrieving biomedical literature. Combine Shannon's information theory with antagonistic learning. In the aspect of heterogeneity gap, modal classification and information entropy maximisation are combined. For this reason, a modal classifier is constructed to distinguish according to the different statistical characteristics of biomedical literature modality. The discriminator uses its output probability to calculate Shannon information entropy. Maximising information entropy gradually reduces the distribution difference of cross modal features, and realises the domain confusion state. Finally, Kullback Leibler divergence and bidirectional triplet loss were used to associate intra modal and inter modal similarity between features in the shared space, thus achieving effective retrieval research of biomedical literature data. However, in practical applications, this method places more emphasis on data classification processing, but pays less attention to the features of the data, which affects its retrieval performance. Mafla et al. (2021) designed a real-time non-dictionary scene retrieval method for biomedical literature retrieval. This method solves the task of text retrieval of medical biological literature: given the medical biological literature query, the system returns the query text. The proposed model uses the single-shot CNN architecture to predict the bounding box and construct a compact

representation of speckled words. In this way, the problem can be modelled as the nearest neighbour search for the text representation of the query output by CNN collected from the database as a whole. This research has achieved rapid retrieval of medical biological literature, but there are many irrelevant data in the retrieval results, which need further improvement.

In view of the shortcomings of the above methods, in order to reduce retrieval errors, improve the accuracy of literature feature mining and shorten retrieval time, this paper designs a new fast retrieval method based on the results of feature mining and biomedical literature as the object. The design idea is as follows:

1 After simulating the collection space of biomedical literature, set data constraints and collection expected distance, and collect biomedical literature according to the data centroid and probability density function.

2 Calculate the similarity between random medical literature data, and label the similar literature by mutual information method. At the label, the cost matrix is used to reduce the imbalance of the dataset and ensure the integrity of the data information.

3 The hidden information in the document data is extracted by using the agent model, and the linear relationship between the parameters of different variables is determined by calculating the Pearson correlation coefficient of the data, so as to determine the correlation between the document data.

4 After setting the risk threshold of expected loss of data, mining the characteristics of biomedical literature. According to the results of feature mining, feature classification and directional retrieval are carried out to realise the rapid retrieval of biomedical literature.

2 Biomedical literature data collection and preprocessing

Biomedical literature data is different from ordinary data. It contains a lot of scientific information and is a valuable resource of human society. Therefore, the data retrieval is also different from other general data retrieval methods. In order to better design an effective retrieval method, this paper first collects biomedical literature data and preprocesses these data to lay the foundation for subsequent retrieval.

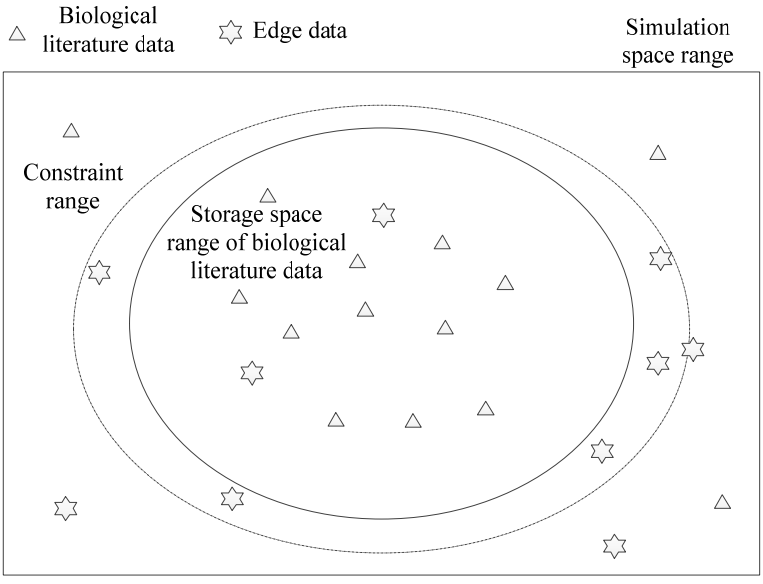
2.1 Research on biomedical literature data collection

In the collection of biomedical literature data, effective collection should be carried out according to the key characteristics of the literature data. From the relevant research, we can see that the way and characteristics of biomedical literature data are different from common data. The key characteristics of biomedical literature can be summarised as follows: large amount of data, fast production speed, many biological types, high value density and authenticity and objectivity (Arts et al., 2021). Through this universal feature, this paper first collects biomedical literature data from different platforms.

Assuming that the platform or database where biomedical literature data is located is a simulated space, biomedical literature can be stored only if the space constraints are met (Karatzas et al., 2022). Therefore, this paper collects biomedical literature data in this

constrained space. The limited storage space of biomedical literature data is shown in Figure 1.

Figure 1 Schematic diagram of limited space for biomedical literature data storage



Within this spatial range, set the data type to text data with a maximum data length of 16,384 bytes, and then use geometric distance measures to determine the relevant biological literature data. Suppose that an uncertain biomedical literature data point and a cluster centre are set within the storage space (Cox et al., 2020), and the expected distance between the biomedical literature data and the cluster centre point is expressed as:

$$D(a, b) = \int d(a, b) f_i(x) dx \tag{1}$$

In formula (1), D(a, b) represents the expected distance between the biomedical literature data and the cluster centre point, a represents the uncertain biomedical literature data point, b represents the cluster centre, and $f_i(x)$ represents the probability density function.

Based on the expected distance, the centroid of the biomedical literature data point is determined (Feng and Gao, 2022), and the result is:

$$c(a) = \int v f_i(x) dx \tag{2}$$

In formula (2), c(a) represents the centroid of biomedical literature data points, and v represents the variance of uncertain biomedical literature data.

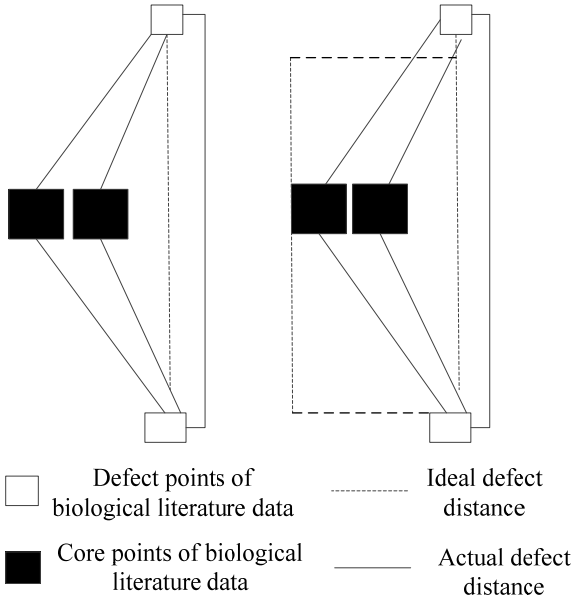
At this time, in the simulated space, set the constraints for biomedical literature data collection, as follows:

$$c_i = \arg\min c(a)\sqrt{g} \tag{3}$$

In formula (3), c_i represents the constraint conditions of biomedical literature data collection, and g represents the edge data excluded in the collection.

Due to the continuous expansion of the scope of this space, there is an uncertainty distance between data when collecting biomedical literature data here. Therefore, the existence of this distance defect cannot be ignored when collecting biomedical literature data (Alzoubi, 2020). The schematic diagram of uncertain distance defect is shown in Figure 2.

Figure 2 Schematic diagram of uncertain distance defect in biomedical literature data



On the basis of the above constraints and uncertainty distance defects, biomedical literature data can be quickly collected in this space, and the results are as follows:

$$Q(x) = \frac{1}{|p| \prod_x^n c_i} \sum c(a) k\left(\frac{c_i}{c(a)}\right) \tag{4}$$

In formula (4), Q(x) represents the result of collected biomedical literature data, k represents the probability density function value of data collection, and p represents Euclidean distance.

The above process first simulates the biomedical literature collection space, sets data constraints, calculates the expected collection distance according to the characteristics of the data, and determines the uncertain distance defect, and collects the literature data according to the data centroid and probability density function.

2.2 Research on biomedical literature data preprocessing

Based on the biomedical literature data collected above, the main method to remove the interference factors in the data is to effectively preprocess these data (Peng et al., 2022).

The purpose of data preprocessing is to label similar data positions by mutual information method, and extract the hidden information of literature data after reducing the imbalance of datasets. By preprocessing, the integrity of data information can be effectively ensured. First, find the similar data between biomedical literature data, which can reduce the tedious process of repeatedly retrieving the same attribute data. Calculate the similarity between two random medical literature data, and the calculation formula is:

$$\text{same}(p_1, p_3) = \frac{\beta}{e+\beta} \tag{5}$$

In formula (5), p_1 and p_3 represent the semantics of two random biological literature data, e represents any positive integer, and β represents the similarity adjustment parameter.

After determining the similarity of two random biomedical literature data, the similar data are effectively labelled, so that similar literature data can be easily found (Saraswathi and Malarvizhi, 2021; Moon et al., 2020). At this time, it is marked by mutual information, and the result is:

$$h(x) = \text{info}(D) - \text{info}_a(D) \tag{6}$$

In formula (6), h(x) represents the labelled similar biomedical literature data, info(D) represents the entropy of random data, and $\text{info}_a(D)$ represents the expected entropy of data with the same attribute.

After labelling similar biomedical literature data, due to the imbalance of the dataset in the collection process of biomedical literature data, more data lost some information due to over-training. In order to ensure the information integrity of the collected biomedical literature data, this paper uses the cost matrix to reduce the imbalance of the dataset (Wang et al., 2021). Set the minority literature data as r_0 and r_1 as the majority biomedical literature data, where F(i, j) represents the generation value lost by the literature data when the dataset is unbalanced. The cost matrix constructed is shown in Table 1.

Table 1 Biomedical literature data cost matrix

Category	r_0	r_1
r_0	$F(r_0, r_0)$	$F(r_0, r_1)$
r_1	$F(r_0, r_1)$	$F(r_1, r_1)$

According to the constructed cost matrix, the probability results of retaining complete biomedical literature data can be obtained as follows:

$$R(r_0, r_1) = s\sum F(r_0, r_1) \tag{7}$$

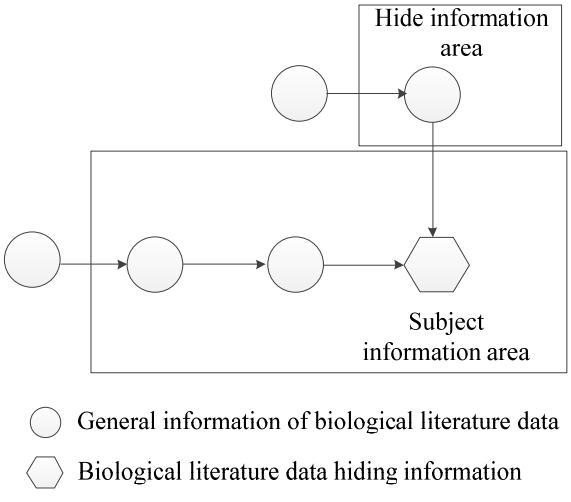
In formula (7), $R(r_0, r_1)$ represents the retention of complete biomedical literature data, and s represents a posterior probability.

Finally, identify the hidden information in the biomedical literature data to further improve the quality of the collected literature data. In this hidden information mining, we preprocess according to the agent model, which is generally used in the research of hidden information in text data (Steffen et al., 2020).

In this model, a number of biomedical literature data are formed into a set, and the set is selected as an agent, and the hidden information is determined according to the

selection of the agent. This model is an abstract research method. For the information body in which the data office is lack of information, these biomedical data should be aligned in the model and then processed uniformly (Sun et al., 2021). The basic structure of the model is shown in Figure 3.

Figure 3 Schematic diagram of basic structure of main model



The hidden information in biomedical literature data is processed according to the agent model, and the results are as follows:

$$U(x) = u\int \sum R\left(r_0, r_1\right) \tag{8}$$

In formula (8), U(x) represents the result of the hidden information in the biomedical literature data processed by the agent model, and u represents the convergence coefficient of the hidden information.

In the preprocessing of biomedical literature data, the similarity between random medical literature data is calculated, the position of similar literature data is marked by mutual information method, the imbalance of medical literature dataset is reduced by cost matrix, the integrity of literature data information is guaranteed, and the hidden information of literature data is extracted by the agent model to realise the preprocessing of biomedical literature data.

2.3 Fast retrieval of biomedical documents based on feature mining

On the basis of the pre-processed biomedical literature data, in order to realise the rapid retrieval research of biomedical literature, this paper carries out feature mining of biomedical literature data, and realises the final retrieval according to the mined features (Wang et al., 2022). The feature mining algorithm covers a lot of data mining algorithms, which are mainly used to determine the characteristics of the research object. This method can quickly extract the characteristics of the research object, and improve the accuracy of the retrieval in biomedical literature data retrieval.

In the feature mining algorithm, the Pearson correlation coefficient is calculated to reflect the linear relationship between the parameters of two different variables, and the expression of the coefficient can improve the correlation between the research objects. Therefore, before this rapid retrieval of biomedical literature data, the Pearson correlation coefficient of biomedical literature data is calculated by this coefficient. Assume that any set of random variables in biomedical literature data is:

$$Z = \{z_1, z_2, \ldots, z_n\}\{x_1, x_2, \ldots, x_i\} \tag{9}$$

In formula (9), Z represents any two sets of random variables in biomedical literature data.

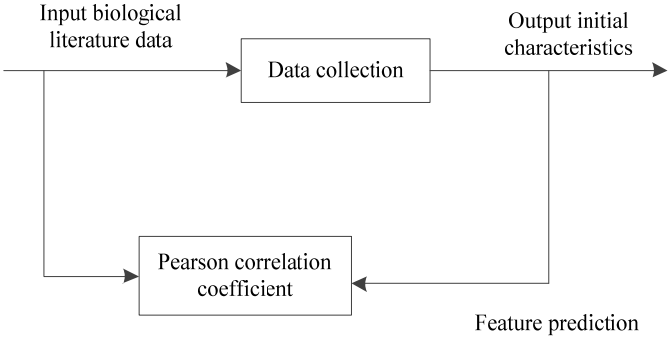
At this time, the Pearson correlation coefficient of the set is defined as:

$$V(Z) = \frac{\sum_i (x_i - x_i')(z_i - z_i')}{\sqrt{\sum_i (x_i - x_i')^2 (z_i - z')}} \tag{10}$$

In formula (10), V(Z) represents the value of Pearson correlation coefficient, x_i' and z_i' represent the mean value of random biomedical literature data. The value range of this value is [–1, 1]. When the value is close to 0, it indicates that the correlation between the two biomedical literature data is low and the setting is not relevant.

Set the biomedical literature set with n features as L, and determine the correlation coefficient threshold between the literature data in the set through formula (10). At this time, document data with multiple key features can be obtained, and the process of output key features is shown in Figure 4.

Figure 4 Schematic diagram of biomedical document feature mining output process



Suppose there are m independent biomedical literature data feature samples, which are expressed as $(d_1, f_1,)$, (d_2, f_2),(d_3, f_3), ⋯, (d_n, f_n). On this basis, the above independent sample data is used to calculate the expected loss risk, and the result is:

$$L(I) = \int h\left(d(x, w)df(x, w)\right) \tag{11}$$

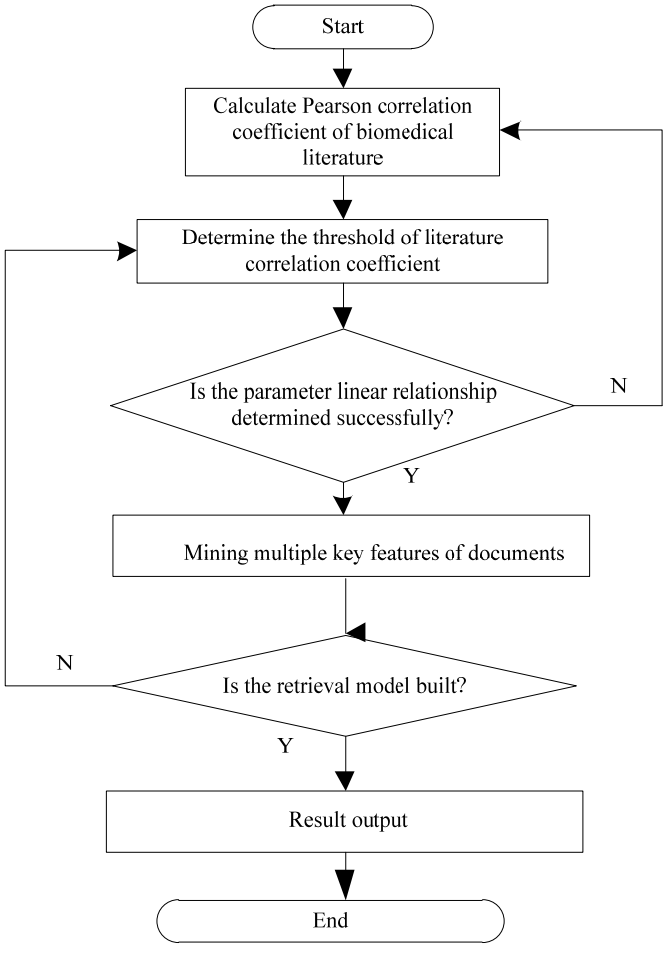
In formula (11), L(I) represents the expected loss risk value of biomedical literature data characteristics, h represents the risk coefficient, and df(x, w) represents the expected key value of characteristics.

On this basis, complete the feature mining of biomedical literature, and the key feature results are:

$$\vartheta_i(x) = \frac{1}{n}\sum L(I)\left[\frac{x - x_i}{h}\right] \tag{12}$$

In formula (12), $\vartheta_i(x)$ represents the feature result of biomedical literature, and n represents the number of feature mining.

Figure 5 Biomedical literature fast retrieval process based on feature mining



After the feature mining of the above biomedical literature, effective retrieval is carried out according to different features. At this time, the retrieval model is constructed as follows:

$$\mu\left(\vartheta_i(x)\right) = \frac{1}{2}w - \sum_{i=1}^{n}\alpha_i\left[p_i\right]L(I) - 1 \tag{13}$$

In formula (13), $\mu(\vartheta_i(x))$ represents the description of the retrieval model, w represents the weight value of key features, α_i represents the expected retrieval result, and ρ_i represents the probability of the occurrence of the retrieved object.

In the process of rapid retrieval of biomedical documents, the Pearson correlation coefficient of biomedical documents is calculated to reflect the linear relationship of different variable parameters, determine the threshold of correlation coefficient between document data, mine multiple key features of documents, calculate the expected loss risk of document data, and realise the feature mining of biomedical documents. On this basis, the rapid retrieval algorithm of biomedical documents is designed to achieve rapid retrieval.

The rapid retrieval process of biomedical literature based on feature mining is shown in Figure 5.

The above process first collected biomedical literature data and preprocessed it to lay the foundation for subsequent retrieval. Then, calculate the Pearson correlation coefficient of the literature data and mine the key features of the literature. Finally, the expected loss risk of literature data was calculated and a fast biomedical literature retrieval algorithm was designed.

3 Experiment and result display

3.1 Experimental parameter setting

In the test, the literature database in a professional biomedical platform is taken as the research object, and the biomedical literature data tested in the database is selected as the research sample for test analysis. In order to ensure the effectiveness of the experiment, the test was conducted in a unified experimental environment. The specific experimental parameters set are shown in Table 2.

Table 2 Experimental parameters

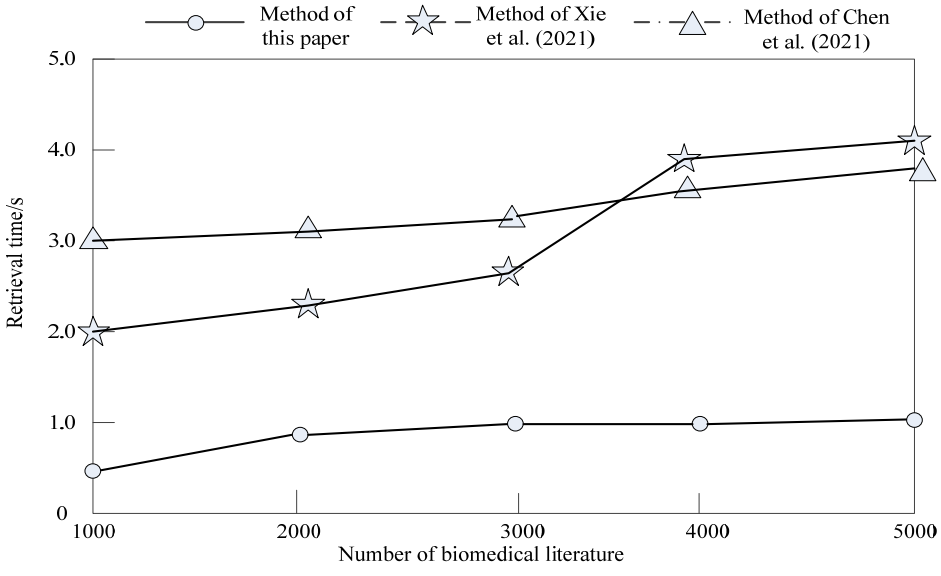
Parameter	*Value*
Biomedical database / GB	16
F sample biomedical data / piece	5,000
Data noise / dB	[–2, 2]
Sample data retrieval interval / s	0.5
Characteristic quantity of sample data / piece	5,000
Operating system	Windows xp
Data statistics software	SPSS 13.0
Data type	Text data
Maximum data length	16,384 bytes

According to the above determined experimental environment and other data, the test was conducted by comparing the method of this paper with the method of Xie et al. (2021) and the method of Chen et al. (2021). The main test is to test and analyse the error of sample biomedical literature retrieval, the accuracy of feature mining and the retrieval time.

3.2 Result comparison and analysis

The retrieval errors of method of this paper, method of Xie et al. (2021) and method of Chen et al. (2021) are compared in the test, and the results are shown in Figure 6.

Figure 6 Retrieval error of biomedical literature



By analysing the test results in Figure 6, we can see that the retrieval error results of method of this paper, method of Xie et al. (2021) and method of Chen et al. (2021) for sample biomedical literature are changing with the number of literature. The retrieval error of method of this paper is the lowest, always lower than 0.2%. While the retrieval error results of method of Xie et al. (2021) and method of Chen et al. (2021) show a downward trend, they are always higher than that of method of this paper. Therefore, we can see that using method of this paper can reduce the error of biomedical literature retrieval and improve the feasibility of retrieval.

In the test, the feature mining accuracy of method of this paper, method of Xie et al. (2021) and method of Chen et al. (2021) is compared, and the results are shown in Table 3.

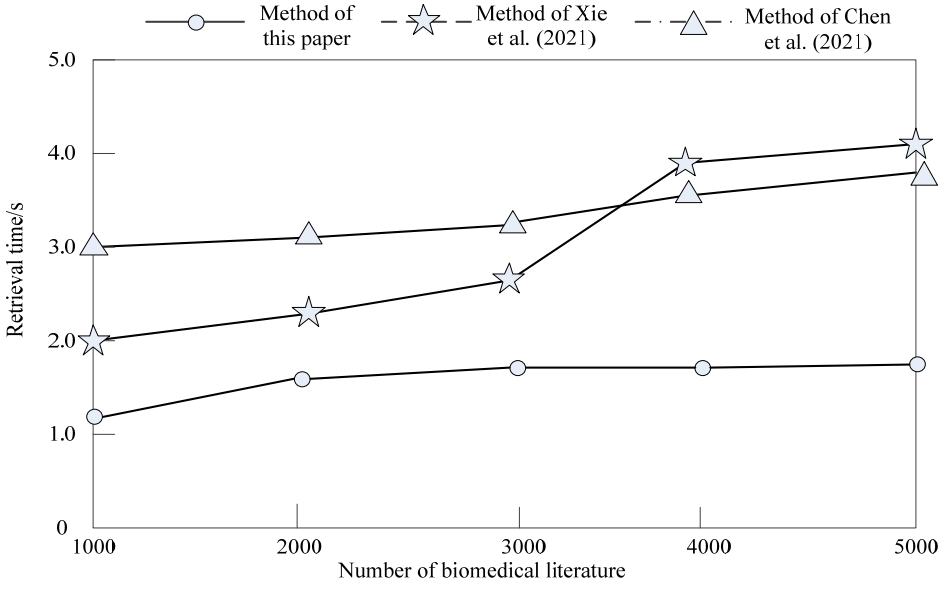
Table 3 Analysis of biomedical literature feature mining precision results (%)

Biomedical literature / piece	*Method of this paper*	*Method of Xie et al. (2021)*	*Method of Chen et al. (2021)*
1,000	99	92	90
2,000	98	90	87
3,000	98	88	86
4,000	97	87	84
5,000	97	85	80

According to the analysis of the experimental results in Table 3, there are certain differences in the accuracy of feature mining for biomedical literature retrieval between method of this paper, method of Xie et al. (2021) and method of Chen et al. (2021). Among them, when the sample biomedical literature is 1,000, the accuracy of feature mining of the three methods is 99%, 92% and 90%, respectively; when the sample biomedical literature is 2,000, the accuracy of feature mining of the three methods is 98%, 90% and 87%, respectively; when the sample biomedical literature is 5,000, the accuracy of feature mining of the three methods is 97%, 85% and 80%, respectively. From the comparison of data results, it can be seen that the method of this paper has a high accuracy of feature mining, with a maximum of about 99%.

At the end of the test, the retrieval time of method of this paper, method of Xie et al. (2021) and method of Chen et al. (2021) for sample biomedical literature was analysed, and the results are shown in Figure 7.

Figure 7 Analysis of retrieval time consuming results of biomedical literature



According to the analysis of the experimental results in Figure 7, there is a large difference in the time consumption of method of this paper, method of Xie et al. (2021) and method of Chen et al. (2021) for the retrieval of sample biomedical literature. From the curve analysis of the experimental results, we can see that the retrieval time of the three methods shows an upward trend. Among them, the rising trend of method of this paper is small, and the retrieval time is less than 2 s. The retrieval time of the other two methods is always higher than that of the method of this paper, which shows that the method of this paper has better timeliness.

In summary, compared with traditional retrieval methods, this method can reduce retrieval errors, improve the accuracy of literature feature mining, and effectively shorten retrieval time.

4 Conclusions

In this paper, a new fast retrieval method of biomedical literature is designed. After simulating the collection space of biomedical literature, the expected collection distance is calculated according to the characteristics of data, and the uncertain distance defect is determined. The collection research is realised according to the data centroid and probability density function. Then, calculate the similarity between random medical literature data, mark the position of similar data through mutual information method, reduce the imbalance of medical literature dataset through cost matrix, ensure the integrity of data information, and extract the hidden information of literature data with the help of agent model to realise the preprocessing of biomedical literature data. On this basis, by calculating the Pearson correlation coefficient of biomedical literature, reflecting the linear relationship of different variable parameters, determining the threshold of correlation coefficient between the literature data, mining multiple key features of the literature, calculating the expected loss risk of the literature data, realising the feature mining of biomedical literature, and then designing the rapid retrieval algorithm of biomedical literature to achieve fast retrieval.

The following experimental conclusions are obtained through the experimental test:

Conclusion 1 The error of using method of this paper to retrieve biomedical literature is lower than 0.2%.

Conclusion 2 Using method of this paper to mine the features of biomedical literature has high accuracy, up to 99%.

Conclusion 3 Using method of this paper to retrieve biomedical literature takes less than 2 s, and the retrieval speed is faster.

Acknowledgements

This work was supported by 2023 Research Start-up Fund of Suqian University.

References

Alzoubi, W.A. (2020) ‘Dynamic graph based method for mining text data’, *WSEAS Transactions on Systems and Control*, Vol. 15, No. 20, pp.453–458.

Arts, S., Hou, J. and Gomez, J.C. (2021) ‘Natural language processing to identify the creation and impact of new technologies in patent text: code, data, and new measures’, *Research Policy*, Vol. 50, No. 2, pp.3302–3315.

Chen, W., Liu, Y.y., Bakker, E.M. and Lew, M.S. (2021) ‘Integrating information theory and adversarial learning for cross-modal retrieval’, *Pattern Recognition*, Vol. 117, No. 1, pp.107–113.

Cox, J., McBeath, D. and Harper, C. (2020) ‘Co-occurrence of cell lines, basal media and supplementation in the biomedical research literature’, *Journal of Data and Information Science*, Vol. 3, No. 24, pp.4781–4792.

Feng, B. and Gao, J. (2022) ‘AnthraxKP: a knowledge graph-based, anthrax knowledge portal mined from biomedical literature’, *Database*, Vol. 11, No. 3, pp.63–71.

Karatzas, E., Baltoumas, F.A., Kasionis, I., Sanoudou, D., Eliopoulos, A.G., Theodosiou, T., Iliopoulos, I. and Pavlopoulos, G.A. (2022) 'Darling: a web application for detecting disease-related biomedical entity associations with literature mining', *Biomolecules*, Vol. 12, No. 4, pp.520–530.

Mafla, A., Tito, R., Dey, S., Gómez, L., Rusiñol, M., Valveny, E. and Karatzas, D. (2021) 'Real-time lexicon-free scene text retrieval', *Pattern Recognition*, Vol. 110, No. 53, pp.107–116.

Moon, K.R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D.B., Chen, W.S., Yim, K., van den Elzen, A., Hirn, M.J., Coifman, R.R., Ivanova, N.B., Wolf, G. and Krishnaswamy, S. (2020) 'Visualizing structure and transitions in high-dimensional biological data', *Nature Biotechnology*, Vol. 54, No. 37, pp.1482–1492.

Mounica, B. and Lavanya, K. (2022) 'Real time traffic prediction based on social media text data using deep learning', *Journal of Mobile Multimedia*, Vol. 18, No. 2, pp.373–391.

Peng, S., Yang, Z. and Ling, B.W-K. (2022) 'Dual semi-supervised convex nonnegative matrix factorization for data representation', *Information Sciences: An International Journal*, Vol. 12, No. 585, pp.571–593.

Saraswathi, S.S. and Malarvizhi, N. (2021) 'Block level time variant dynamic encryption algorithm for improved cloud security and de-duplication using block level topical similarity', *International Journal of Advanced Intelligence Paradigms*, Vol. 3, No. 4, pp.271–283.

Steffen, P., Wu, J. and Hariharan, S. (2020) 'OmixLitMiner: a bioinformatics tool for prioritizing biological leads from 'omics data using literature retrieval and data mining', *International Journal of Molecular Sciences*, Vol. 21, No. 4, pp.412–417.

Sun, X., Long, X., He, D., Wen, S. and Lian, Z. (2021) 'VSRNet: end-to-end video segment retrieval with text query', *Pattern Recognition*, Vol. 119, No. 4, pp.108–127.

Wang, L., Wang, Y. and Wen, D. (2021) 'Tunable biological nonvolatile multilevel data storage devices', *Physical Chemistry Chemical Physics*, Vol. 23, No. 10, pp.321–328.

Wang, Z., Zhu, A., Xue, J., Jiang, D., Liu, C., Li, Y. and Hu, F. (2022) 'SUM: serialized updating and matching for text-based person retrieval', *Knowledge-based Systems*, Vol. 248, No. 19, pp.1–16.

Xie, Z., Liu, L., Wu, Y., Zhong, L. and Li, L. (2021) 'Learning text-image joint embedding for efficient cross-modal retrieval with deep feature engineering', *ACM Transactions on Information Systems (TOIS)*, Vol. 22, No. 10, pp.1–7.

Yin, X. and Chen, L. (2022) 'A cross-modal image and text retrieval method based on efficient feature extraction and interactive learning CAE', *Scientific Programming*, Vol. 20, No. 12, pp.1–12.

Zhao, S., Wang, A. and Qin, B. (2022) 'Biomedical evidence engineering for data-driven discovery', *Bioinformatics*, Vol. 38, No. 23, pp.5270–5278.