ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20240111001262
Original scientific paper

Efficient Decomposition Method for Similar Text Data in Large Corpora

Yun HE

Abstract: In order to solve the problems that the decomposition results of current similar data decomposition methods are inconsistent with the actual text quantity, the increase of sensitive data is not significant, and the absolute error mean and normalized root mean square error are high, a large-scale real text corpus similar data decomposition method is proposed. Dividing into a plurality of minority sub-clusters, determining the probability distribution of the minority sub-clusters in the similar data set of the text corpus, and sampling the data in the similar data set of the text corpus. On the basis of data ontology structure mapping model and text big data analysis model, tag semantics are generated to realize similar data decomposition of text corpus. The experimental results show that this method improves the category imbalance of the original data set, can decompose the text accurately, and the decomposition results are basically consistent with the actual text quantity, with lower absolute error mean and normalized root mean square error, and have better similar data decomposition ability.

Keywords: corpus; data decomposition; real text; similar data

1 INTRODUCTION

With the rapid development of information technology, large-scale real text corpus has gradually become an indispensable resource in natural language processing, machine learning, artificial intelligence and other fields [1]. These corpora contain various types, themes and sources of text data, such as news reports, social media, scientific papers, advertisements, user comments and so on [2]. They provide valuable resources for natural language processing, information retrieval, sentiment analysis, recommendation system and other research with their huge scale, rich diversity and dynamic updating characteristics. However, large-scale real text corpus brings a series of challenges [3], including data complexity, redundancy, noise interference and other issues, which seriously affect the quality and effective use of data. The existence of similar data is a common phenomenon in large-scale real text corpus [4]. Similar data may be caused by repeated publishing, reprinting, plagiarism, etc., or by the similarity of text theme and writing style. The existence of similar data brings great troubles to data processing and analysis, such as the interference of redundant information, the waste of computing resources and the decline of model performance. Therefore, how to effectively decompose similar data and extract useful information is an urgent problem in current research. In order to recognize, remove and cluster similar data in different degrees, the efficiency and quality of data processing are improved. In order to solve the problems of inaccurate measurement of similarity, insufficient processing ability for large-scale data, and weak adaptability to complex data types, this paper proposes a similar data decomposition method for large-scale real text corpus. The innovations of this method are as follows: (1) Divide it into several minority sub-clusters, determine the probability distribution of minority sub-clusters in similar data sets of text corpus according to the misclassification rate and sampling weight of sample sub-clusters, and sample the data in similar data sets of text corpus; (2) Construct a data ontology structure mapping model and a text big data analysis model, and on the basis of these models, generate the semantics of similar data tags in the text corpus, thus realizing the decomposition of similar data in the text corpus.

2 LITERATURE REVIEW

As one of the key technologies of natural language processing and machine learning, similar data decomposition method has been widely concerned and studied in recent years. At present, many similar data decomposition methods have been proposed. Reference [5] proposes a data-driven mode decomposition method for venturi nozzle cavitation flow. Proper orthogonal decomposition and dynamic mode decomposition are introduced to study the three-dimensional coherent structure of cavitation and flow field, and the cavitation flow in Venturi nozzle is measured experimentally. Then the simulation is based on Zwart cavitation model, and the experimental data are verified. Reference [6] proposes a rolling decomposition method driven by data features based on the integrated gasoline consumption forecasting model; it includes five steps: data feature testing, data decomposition, component feature analysis, component prediction and integrated output. In the data characteristic test and component characteristic analysis, the original time series and each decomposed component are thoroughly analyzed to explore the hidden data characteristics. According to these results, the decomposition model and prediction model are selected to complete the original time series data decomposition and decomposition component prediction. In the integration output, the integration method corresponding to the decomposition method is used for final aggregation. It can be seen that the decomposition-integration model with data feature-driven modeling idea and rolling decomposition prediction mechanism has superiority and robustness in the evaluation criteria of horizontal and directional prediction. Reference [7] proposes a multi-component submarine seismic data decomposition method using independent calibration filters. A separate calibration filter is designed to separate the calibration process of pressure and particle velocity (or displacement) into time difference elimination and amplitude compensation. Then the dynamic characteristics of particle velocity or displacement components are calibrated, and the energy distribution inside each component is adjusted. The proposed decomposition technique is applied to the active deep-water multi-component submarine seismic data set to

obtain high-quality upstream and downstream P waves and S waves. The decomposition results show that the expected effect of decomposition, such as water column-related multiple attenuation has been achieved. Reference [8] proposes a corpus-based crowdsourcing enhanced data decomposition method. In this article, we have presented the details of the crowdsourcing platform named "Konkani Shabdarth" (kōṁkanīś abdārth). Konkani Shabdarth attempts to use the knowledge of Konkani speaking people for creating new synsets and perform the quantitative enhancement of the word net. It also intends to work toward enhancing the overall quality of the Konkani WordNet by validating the existing synsets, and adding the missing words to the existing synsets. A text corpus named "Konkani Shabdarth Corpus", has been created from the Konkani literature while implementing the Konkani Shabdarth tool. Using this corpus, 572 root words that are missing from the Konkani WordNet have been identified which are given as input to Konkani Shabdarth. The expected increase in the percentage coverage of Konkani WordNet has been found to be in the range 20 - 27 after adding the missing words from the Konkani Shabdarth corpus in comparison to the other corpora for which the increase is in the range 1 - 10.

3 RESEARCH METHODOLOGY

3.1 Large-scale Real Text Corpus Similar Data Sampling

In order to make the sample information of similar data in text corpus more effective, it is necessary to sample the similar data in text corpus. The misclassification rate is used to measure the ratio of misclassification data to all sub-cluster data in each sub-cluster divided by classifier pairs in the integrated decomposition method [9-10]. The calculation formula of data misclassification rate is as follows:

$$D = \frac{D_1}{D_Z} \cdot 100\% \tag{1}$$

In Eq. (1), D_1 represents the number of misclassified samples in each sub-cluster; D_Z represents the total number of samples in each sub-cluster. Sampling weight is the importance of the text data in the sample in the whole sample research, which reflects the representativeness and importance of each sample point to the whole data set. The sampling weight of similar data in the text corpus is expressed by W, and its expression is as follows:

$$W = D \cdot \left(R_a - R_b \right) \cdot \lambda \tag{2}$$

In Eq. (2), R_a represents the number of samples in most categories; R_b represents the number of minority samples; β stands for sampling rate. The concept of probability distribution of sub-clusters is introduced. Sub-clusters usually refer to data subsets with some similarity or common characteristics in data sets. The probability distribution of sub-clusters refers to the distribution of these sub-clusters in the whole data set, that is, the number and density of data points contained in each sub-cluster and the relationship between them. This distribution describes the relative importance and frequency of different sub-clusters in the corpus. Assuming that there is a sample x in a few kinds of sub-clusters, and using the probability that the sample x belongs to the seed sample, the expression of probability distribution P of sub-clusters is constructed as follows:

$$P = \frac{p_i}{\sum p_i} \cdot x \cdot W \tag{3}$$

In Eq. (3), p_i represents the seed sample probability. Among them, the formula for calculating the probability p_i that the sample x belongs to the seed sample is as follows [11-12]:

$$p_i = \frac{1}{M_{xy}} \tag{4}$$

In Eq. (4), M_{xy} represents the Euclidean distance between the sample x and the nearest neighbor y of most classes of samples. Select the seed samples in each sub-cluster to simulate the original data distribution in similar data sets of the original text corpus [13-15]. All the minority sub-clusters in the data set are given corresponding weights, and data sampling is carried out to ensure data balance. Similar data sampling of large-scale real text corpus includes the following processes:

Step 1: Dividing minority samples in similar data sets of a large-scale real text corpus into a plurality of minority sub-clusters;

Step 2: According to the misclassification rate and sampling weight of sample sub-clusters, determine the probability distribution of minority sub-clusters in similar data sets of text corpus.

Step 3: Synthesize sample sub-cluster, which processes minority samples and their nearest neighbors by linear interpolation method to synthesize new data samples. The composite expression is as follows:

$$C = \theta \cdot \left(C_{i,j} - C_i \right) \cdot p_i \tag{5}$$

In Eq. (5), C represents a new minority sample synthesized; C_i and $C_{i,j}$ respectively represent the i minority sample and its j neighbor sample; θ stands for interpolation coefficient.

Step 4: Synthesize new samples repeatedly according to the probability distribution until the number of iterations meets the threshold of sampling weight setting, and end the large-scale real text corpus similar data sampling.

Using the above process, the data in the similar data set of the text corpus are sampled, and the sampled data set is used as the basis for generating the semantics of similar data labels of the text corpus.

3.2 Generate Similar Data Label Semantics of Text Corpus

Generating similar data tag semantics of text corpus. By analyzing large-scale real text corpus, similar text data are divided into data sets with common semantics, and corresponding semantic tags are generated for each data set [16]. Considering the semantic inconsistency of similar data tags in each text corpus, it shows that the data

categories in the tags are different. In this paper, semantic cleaning is proposed for all subsets of the tag class, and the semantic similarity weight matrix between data samples in the subset is calculated. In the process of actual semantic clarity, samples are divided into two groups, and the intermediate semantics of labels is determined by inter-group cleaning. Then, the results of other inter-group cleaning are synthesized, and the common semantics is retained, which serves as the similar data label semantics of text corpus. In the process of multi-label text big data analysis, the context mapping principle is introduced to construct a data ontology structure mapping model; this model is used to describe and construct the corresponding relationship of ontology structure between different data sources in large-scale real text corpus similar data sampling. Specifically, it focuses on how to map and align ontology concepts, attributes and their relationships in different data sets, so as to realize similar data sampling across data sets and the expression is:

$$\delta = \frac{C}{1+e^{n}} \tag{6}$$

In Eq. (6), δ represents the original text data; n represents mapping data; e represents the similarity of semantic ontology fusion. Based on the mapping model, the semantic mapping relationship of large text data is established, and the fuzzy correlation coefficient between similar data in different text corpora is determined to obtain the association rule set. On this basis, combining the strategy of semantic ontology feature reconstruction, this strategy mainly focuses on how to reconstruct or adjust the semantic ontology features of text data to more accurately reflect the internal meaning and similarity of data, thus optimizing the effect of similar data sampling. The core idea is to deeply analyze and process the original text data and extract more representative and differentiated semantic features. Construct a text big data analysis model, which refers to the model of in-depth analysis and mining of large-scale real text corpus, and the expression is:

$$E_Q = \eta \cdot \mu + \sum_{i=1}^{n} \left(b_i, d_i, t_i\right) \cdot \delta \tag{7}$$

In Eq. (7), E_Q stands for multi-label text big data analysis model; η stands for support; μ stands for confidence; i represents a data node; n represents the total number of data nodes; b_i represents the entity set of multi-label text data; d_i represents the entity set of similarity information; t_i represents the entity set of sampling interval time. Through the data ontology structure mapping model and text big data analysis model, the multi-label text big data is deeply analyzed, and the relational knowledge between the data is clarified, and then the relational knowledge is logically reasoned, and the statistical characteristics such as semantic attribute characteristics and autocorrelation characteristics contained in the text big data are detected. Combined with the principle of linear regression analysis, this principle is a method to analyze the relationship between two variables when one response variable is considered as a linear function of another explanatory variable. The autocorrelation features are described as triplets, and on this basis, the semantic feature extraction parameters of big data are determined as follows:

$$\tau = \frac{1}{n} \sum_{i=1}^{n} W_i \cdot \alpha_i \cdot \beta_i \tag{8}$$

In Eq. (8), τ represents the feature extraction result; W_i stands for triplet; α_i and β_i represent autocorrelation feature and semantic attribute feature respectively. Through the above-mentioned feature extraction parameters, the features of text big data can be obtained, which can be used as the basis for subsequent classification processing. Taking the word frequency of the text as the basic index to generate the tag semantics of similar data in the text corpus, the tag semantic feature weight is determined, the weight reflects the importance and influence of semantic features corresponding to different tags in the sampling process. Simply put, the weight of tag semantic features is used to measure the proportion of semantic features of a particular tag in determining the similarity of data points, and the calculation formula is as follows:

$$L_{TJ} = K_f \cdot K_s \cdot \tau \tag{9}$$

In Eq. (9), L_{TJ} represents the tag semantic feature weight; K_f indicates the number of occurrences of the selected words in the text; K_s represents the total word frequency in the text. By calculate that weight of the obtained tag semantic feature, data support is provided for generating tag semantics of similar data in a text corpus. Generate similar data tag semantics of a text corpus from a plurality of subsets under a tag, as shown in Eq. (10):

$$G = L_{TJ} \cdot \left(g_1, g_2, g_3, \cdots, g_m\right) \tag{10}$$

In Eq. (10), G represents tag semantics; g_1, g_2, g_3, …, and g_m represent the semantics of the tag; m represents the number of labels. The specific tag semantic structure is shown in Fig. 1.

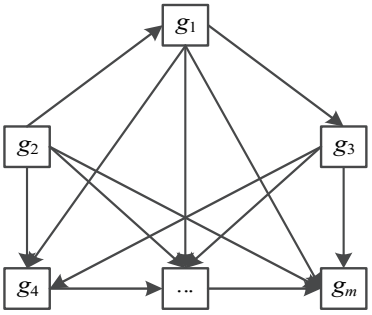

Figure 1 Semantic Structure Diagram of Tags

Through the tag generation method, each data set is given a representative and indicative semantic tag for better information organization, retrieval and understanding.

3.3 Realize the Task of Decomposing Similar Data in Text Corpus

Realize the decomposition of similar data in text

corpus, and divide the text data in large-scale real text corpus into similar data sets. This process involves the selection of similarity measurement methods, and the calculation is based on word overlap, semantic relevance or contextual information. Text is grouped by graph structure method to form similar data sets. This decomposition process is helpful to better organize, manage and understand text data to support the subsequent information retrieval, classification and mining tasks. Assuming D_j represents the user's satisfaction with the data decomposition task, the user's j satisfaction is quantified and transformed into (0, 0.25, 0.5, 0.75, 1), where 1 represents that the user is very satisfied with the data decomposition task and 0 represents that the user is very dissatisfied with the data decomposition task. Use $\overline{D}_j$ to express the average satisfaction of users. It is used to measure the satisfaction of sampling results to users' needs, and its calculation formula is as follows:

$$\overline{D}_j = \frac{\sum_{j=1}^{m} D_j}{m} \cdot G \tag{11}$$

In Eq. (11), m represents the maximum number of users. Decomposing the similar data decomposition task according to the structure and function to obtain a plurality of subtasks, and obtaining the smallest task when the subtasks cannot be decomposed any more, and stopping decomposing the similar data decomposition task at this time, if the subtasks can be further decomposed, the user's satisfaction with the subtasks can be measured, set a threshold μ, and stop the decomposition of similar data in the text corpus when the user's satisfaction reaches the threshold μ. In the process of similar data decomposition of large-scale real text corpus, the threshold μ needs to meet the following conditions:

(1) When D_j belongs to the smallest similar data decomposition task, the decomposition operation of similar data decomposition task is stopped, and when D_j does not belong to the smallest similar data decomposition task, the user's satisfaction with task D_j is evaluated;

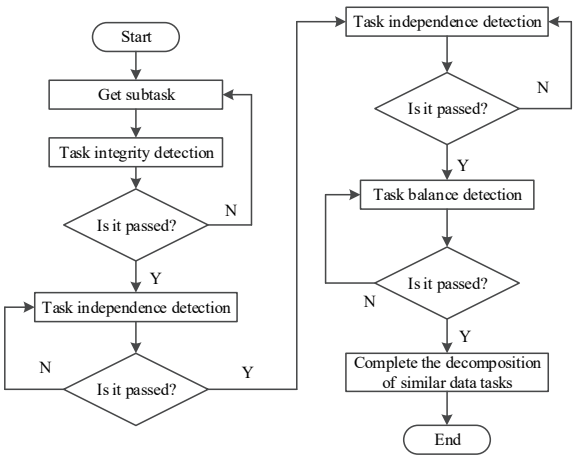

Figure 2 is a flowchart of similar data decomposition of text corpus

(2) When D_j is greater than the threshold μ, stop the decomposition operation of similar data decomposition tasks, and when D_j is less than the threshold μ, continue to decompose task D_j.

The specific process of similar data decomposition of large-scale real text corpus is shown in Fig. 2.

According to Fig. 2, the specific steps to realize decomposition are as follows:

Step 1: based on the main activities and contents, according to the characteristics of similar data decomposition tasks, obtain subtasks d_1, d_2, …, d_n of similar data decomposition tasks;

Step 2: Check the completeness of subtask d_1, d_2, …, d_n obtained in the above process, that is, whether subtask can express the original similar data decomposition task. If you can go to the next step, if you can't go back to the previous step.

Step 3: Check whether subtasks d_1, d_2, …, d_n are related, that is, whether subtasks are independent or not, and whether the execution of subtasks will affect the execution of other subtasks. If subtasks are not related, proceed to the next step, and if subtasks are related, return to the previous step.

Step 4: In the execution of the task, check the feasibility of the task, that is, whether the subtask D_j can be completed. The feasibility of the subtask has a direct impact on the rationality of its decomposition process, which is an important step in the similar data decomposition process of the sub-text corpus of cloud computing.

Step 5: Check the completeness of similar data decomposition of the text corpus. If the subtasks are decomposed completely, proceed to the next step. If the similar data decomposition tasks are not completely decomposed, return to the previous step.

Step 6: Check the balance of similar data decomposition tasks during execution. If the task pressure is unbalanced, it is necessary to decompose similar data decomposition tasks again, otherwise the work will be unbalanced. If the pressure is balanced, proceed to the next step.

Step 7: After the above processing, each subtask obtains the corresponding output and input, and has a corresponding execution task to complete the decomposition of similar data decomposition tasks.

4 RESULTS AND DISCUSSION

In order to verify the overall effectiveness of the research on similar data decomposition method of large-scale real text corpus, it is necessary to test it. Taking the decomposition effect of similar data texts, the decomposition changes of sensitive item data hiding, adding and modifying sensitive item data, the mean value of absolute error and normalized root mean square error as indicators, the methods of this paper, reference [5], reference [6] and reference [7] are used for comparative testing. Through the correlation analysis of simulation experiments, MYSQL is selected as the data set of this real text corpus, and the files are divided into private data and non-private data according to their attribute types. According to the basic parameter types, it is divided into 1 MB, 10 MB and 100 MB data. The experiment collected 80000 pieces of data from a website, including 30000 pieces of private data and 50000 pieces of non-private data. In order to effectively simulate the simulation experiment environment, considering that there is a high probability of similar data in the real text corpus in the network, a

dynamic social network was established, and the dynamic updating process was simulated in an iterative way to ensure the reference of the experiment. The sampling frequency of similar data in text corpus is 18 kHz, and the sampling size is 20 bits. The details of similar data in large-scale real text corpus are shown in Tab. 1.

Table 1 Details of Similar Data in Text Corpus

Corpus coding	Sampling duration / s	Number of texts / Number	Data type
Q11	20	7000	Privacy data
Q12	24	11000	Does not contain private data.
Q13	22	9000	Privacy data
Q14	21	8000	Does not contain private data.
Q15	20	14000	Privacy data
Q16	28	31000	Does not contain private data.

Similar data types of large-scale real text corpus are mainly shown in Tab. 2.

Table 2 Similar Data Types of Text Corpus

Chunk type	Content	Mark symbol
Universal text corpus	News, blogs, forums, social media, etc.	H0001
Domain-specific text corpus	Medical care, finance, law, science and technology, etc.	H0002
Task-specific text corpus	Emotional analysis, question answering system, machine translation, etc.	H0003
Social media text corpus	Weibo, Twitter, Reddit, etc.	H0004

After using the corpus, the decomposition effect of similar data texts in the large-scale real text corpus shown in Tab. 1 is shown in Tab. 3.

Table 3 Decomposition Effect of Similar Data Text in Large-scale Real Text Corpus

Mark symbol	H0001	H0002	H0003	H0004
Q11 Actual text quantity/piece	1000	2500	2000	1500
The method in this paper decomposes results/piece.	1000	2500	2000	1500
Q12 Actual text quantity/piece	1600	1400	5000	3000
The method in this paper decomposes results/piece.	1600	1400	5000	3000
Q13 Actual text quantity/piece	1200	2300	2400	3100
The method in this paper decomposes results/piece.	1200	2300	2400	3100
Q14 Actual text quantity/piece	2000	3000	1100	1900
The method in this paper decomposes results/piece.	2000	3000	1100	1900
Q15 Actual text quantity/piece	2800	3000	5000	3200
The method in this paper decomposes results/piece.	2800	3000	5000	3200
Q16 Actual text quantity/piece	9000	8500	7500	6000
The method in this paper decomposes results/piece.	9000	8500	7500	6000

From Tab. 3, it can be seen that the decomposition effect of this method in various large-scale real text corpora can accurately decompose text for different scale text data sets, and the decomposition results are basically consistent with the actual text quantity, showing its high-precision decomposition performance, which shows that this method has a good application effect in large-scale real text corpora and can effectively decompose text data. Choose MYSQL as the real text corpus data set, and the original data sample distribution of MYSQL data set is shown in Fig. 3.

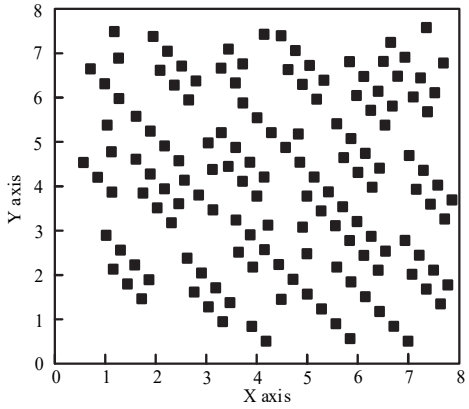

Figure 3 Distribution of Original Data Samples of MySQL Dataset

After sampling the MYSQL data set selected by this method, the data decomposition results in the data set are shown in Fig. 4.

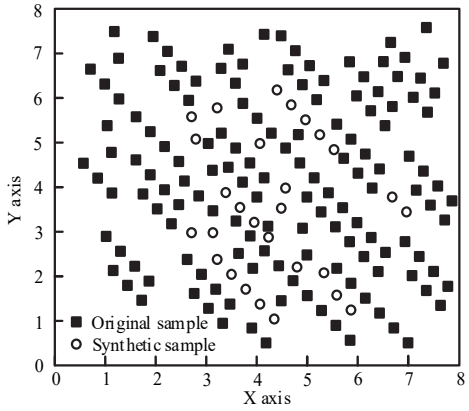

Figure 4 Sampling results of sub-data set

By comparing the experimental results of Fig. 3 and Fig. 4, it can be seen that the new samples are successfully synthesized by using the similar data decomposition task of the text corpus, and these new samples are mainly concentrated in the middle area of the data set. This sampling method not only improves the category imbalance of the original data set, but also makes the sampled samples reflect the data distribution more effectively. When dealing with unbalanced data sets, category imbalance is a common problem, which leads to the model being biased towards most categories in the training process, thus ignoring a few categories. Through sampling processing, the method in this paper increases the number of samples in a few categories, so that the decomposition method can learn data distribution more comprehensively and improve the generalization ability and accuracy of decomposition tasks. In a large-scale real text corpus, the most easily leaked and representative similar data is sensitive data. Usually, the changes of sensitive data include options such as hiding, adding and modifying. Hiding means that sensitive data is hidden or encrypted to protect the security and privacy of data. Adding refers to adding sensitive data, which usually involves adding, updating or creating data, including adding new personal data, updating existing data or

creating new data items. Modification refers to the modification of sensitive data, which involves the change of existing data, including the change of the address, telephone number or email address of personal information. The change of sensitive data can also reflect the similar data decomposition effect. The better the decomposition effect, the lower the degree of nudity representing data attributes. Analyze the data decomposition changes of hidden, added and modified sensitive items under the protection of four methods, and the results are shown in Fig. 5.

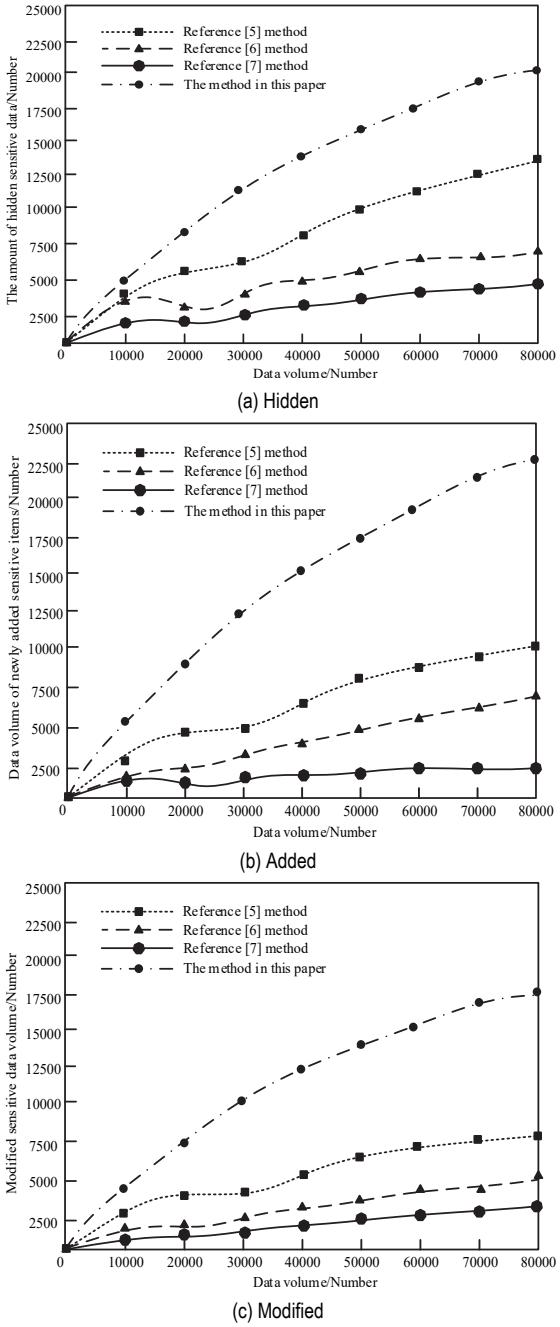

Figure 5 Data decomposition results of sensitive items under four methods

According to Fig. 5, with the increase of the number of data sets, the four different methods all show an overall upward trend in hiding, adding and modifying sensitive data. However, the increase of this method in this process is the most significant, which shows that this method adopts a gradual protection strategy, that is, real-time updating and encryption protection according to the matching degree of private data. This method not only helps to prevent the loss of important data, but also has high applicability to large-scale data environments. When faced with the experimental situation of adding data or modifying data, this method still shows the best effect of hiding sensitive data, thus ensuring high protection quality. In contrast, there is a big gap between the protection quality of the other three methods in the literature with new data and the environment without new data, which further highlights the superiority and stability of this method in dealing with dynamic data sets. In order to further verify the effectiveness of the similar data decomposition method of large-scale real text corpus, MYSQL data set was used in the experiment, and reference [5] method, reference [6] method and reference [7] method were used as the control group to test the similar data decomposition performance of the four methods. Using the mean absolute error M_{AE} and normalized root mean square error N_{RMS} as indicators to evaluate the decomposition performance of the methods proposed in this paper, reference [5], reference [6], and reference [7], $\hat{x}_i$ represents the ith similar data decomposition result, x_i represents the ith actual load value, m represents the total number of data, and $i = 1, 2, \ldots, m$. The calculation formulas for M_{AE} and N_{RMS} are as follows:

$$\begin{cases} M_{AE} = \dfrac{1}{m} \cdot \left| \hat{x}_i - x_i \right| \\ N_{RMS} = \sqrt{\dfrac{\left(\hat{x}_i - x_i \right)^2}{x_i^2}} \end{cases} \tag{12}$$

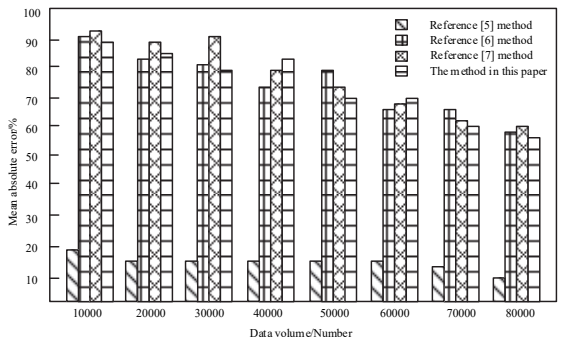

Figure 6 Absolute error mean detection results

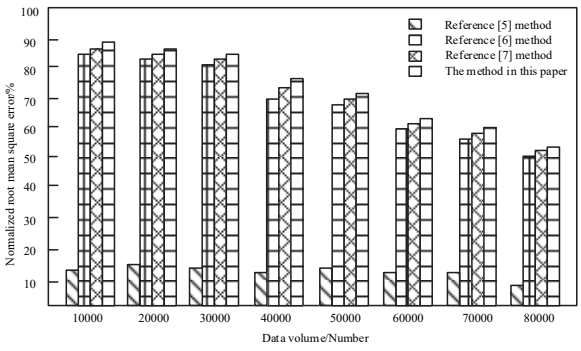

Figure 7 Normalized root mean square error detection results

According to Eq. (12), the absolute error mean M_{AE} and normalized root mean square error value N_{RMS}, M_{AE} and N_{RMS} of the four methods after similar data decomposition are calculated, and the lower the values DD and FF, the

better the similar data decomposition performance of the corresponding methods, and the obtained results are shown in Fig. 6 and Fig. 7 respectively.

As can be seen from Fig. 6 and Fig. 7, the method in this paper shows excellent performance in similar data decomposition tasks. Through in-depth comparison and analysis, it is found that this method is superior to the reference [5] method, the reference [6] method and the reference [7] method in many key evaluation indexes. First of all, from the index of absolute error mean, this method is significantly lower than the other three methods. The mean absolute error is an important criterion to measure the accuracy of the algorithm, which reflects the deviation degree of the algorithm in practical application. This method can achieve a lower mean absolute error, and the key lies in its careful design according to the characteristics of similar data decomposition tasks. According to the task function, the decomposition task is refined to ensure that each subtask can accurately reflect the similarity between data, so that the method in this paper can capture the subtle differences of data more accurately in the decomposition process, thus reducing the overall error. Secondly, judging from the normalized root mean square error, this method also performs well. The normalized root mean square error not only considers the size of the error, but also considers the distribution of the error, so it can evaluate the performance of the algorithm more comprehensively. The method in this paper performs similar data decomposition tasks through parallel processing rules, which effectively improves the calculation efficiency and ensures the accuracy and stability of decomposition results, so that the method in this paper can still remain efficient and accurate when dealing with large-scale data sets, thus achieving lower normalized root mean square error. To sum up, this method shows excellent performance in similar data decomposition tasks, with higher accuracy and stability, because this method refines the decomposition tasks according to the task function and adopts efficient parallel processing rules. These innovative designs make this method have a wider application prospect and higher practical value in practical application.

5 CONCLUSION

How to decompose similar data efficiently is an important research direction in large-scale real text corpus. Therefore, a similar data decomposition method for large-scale real text corpus is proposed, and the following conclusions are drawn through research:

(1) This method not only improves the category imbalance of the original data set, but also makes the sampled samples reflect the data distribution more effectively.

(2) The decomposition effect of this method in various large-scale real text corpora can accurately decompose the text for different size text data sets, and the decomposition results are basically consistent with the actual text quantity.

(3) With the increase of the number of data sets, the four different methods all show an overall upward trend in hiding, adding and modifying sensitive data, and this method has the most significant increase in this process.

(4) The mean absolute error and normalized root mean square error of this method are always lower than those of reference methods, and it has better similar data decomposition ability in practical application.

In large-scale real text corpus, similar data decomposition method aims to divide massive text data into subsets with similarity for further analysis and processing. However, this method faces some limitations in practical application, and there are also some data sets with weak performance. The following is a description of these limitations and a discussion of the future work.

(1) limitations

Data sparsity and dimension disaster: In a large-scale text corpus, due to the high dimension and sparsity of text data, similarity calculation becomes complicated and expensive, which leads to the decrease of the accuracy of similar data decomposition, especially when dealing with high-dimensional sparse data.

Limitations of semantic understanding: The existing similar data decomposition methods mainly calculate the similarity based on the surface features of texts (such as word frequency, TF-IDF, etc.), ignoring the deep semantic information of texts, resulting in some semantically similar texts with different surface features being wrongly divided into different subsets.

Influence of noise and outliers: There are often a lot of noise and outliers in real text corpus, which come from spelling mistakes, grammatical errors, non-standard terms and so on. These noises and outliers interfere with the accuracy of similar data decomposition, resulting in unstable decomposition results.

(2) Data sets with weak performance

In some specific fields or application scenarios, there are data sets with weak performance. Because of the small data scale, low data quality or uneven data distribution, the performance of similar data decomposition methods is not good.

(3) Future work

Solve the scalability problem: With the continuous expansion of text data scale, similar data decomposition methods need to have good scalability. Future work should focus on designing more efficient and scalable algorithms to deal with larger data sets.

Deepening semantic understanding: In order to improve the accuracy of similar data decomposition, future work should further explore the deep semantic information of the text. We can use the latest technologies in the field of natural language processing, such as deep learning and knowledge mapping, to enhance our understanding of text semantics.

Dealing with noise and outliers: In order to solve the problem of noise and outliers in real text corpus, more robust similarity calculation methods can be studied in future work to reduce the influence of noise and outliers on decomposition results.

Technology expansion and application exploration: Similar data decomposition method can be applied to many fields, such as information retrieval, recommendation system, text clustering and so on. Future work can explore the application of this technology in more fields, and expand and optimize the technology according to specific application scenarios.

6 REFERENCES

[1] Manerkar, S., Asnani, K., Khorjuvenkar, P. R., Desai, S., & Pawar, J. D. (2022). Konkani wordnet: corpus-based enhancement using crowdsourcing. *ACM transactions on Asian and low-resource language information processing*, 21(4), 1-18. https://doi.org/10.1145/3503156

[2] Li, S., Song, P., & Zhang, W. (2022). Transferable discriminant linear regression for cross-corpus speech emotion recognition. *Applied acoustics, 197*(8), 1-11. https://doi.org/10.1016/j.apacoust.2022.108919

[3] Liu, M. (2022). Stancetaking in hongkong political discourse:a corpus-assisted discourse study. *Chinese Language and Discourse, 13*(1), 79-98. https://doi.org/10.1075/cld.21001.liu

[4] Svetanant, C., Ballsun-Stanton, B., & Rutherford, A. T. (2022). Emotional engagement in thai and japanese insurance advertising: corpus-based keyword analysis. *Corpora, 17*(1), 69-96. https://doi.org/10.3366/cor.2022.0235

[5] Han, Y., Liu, M., & Tan, L. (2022). Method of data-driven mode decomposition for cavitating flow in a venturi nozzle. *Ocean engineering, 261*(1), 1-14. https://doi.org/10.1016/j.oceaneng.2022.112114

[6] Yu, L. & Ma, Y. (2021). A data-trait-driven rolling decomposition-ensemble model for gasoline consumption forecasting. *Energies, 14*(15), 1-26. https://doi.org/10.3390/en14154604

[7] Yu, P., Chu, M., & Jiang, J. (2023). Multicomponent ocean-bottom seismic data decomposition using separate calibration filters. *Pure and Applied Geophysics, 180*(1), 41-57. https://doi.org/10.1007/s00024-022-03196-5

[8] Manerkar, S., Asnani, K., Khorjuvenkar, P. R., Desai, S., & Pawar, J. D. (2022). Konkani wordnet: corpus-based enhancement using crowdsourcing. *ACM transactions on Asian and low-resource language information processing, 21*(4), 1-18. https://doi.org/10.1145/3503156

[9] Burdzik, R. (2022). A comprehensive diagnostic system for vehicle suspensions based on a neural classifier and wavelet resonance estimators. *Measurement, 200*, 1-17. https://doi.org/10.1016/j.measurement.2022.111602

[10] Taha, A. A. & Malebary, S. J. (2022). A Hybrid Meta-Classifier of Fuzzy Clustering and Logistic Regression for Diabetes Prediction. *Computers, Materials, & Continua, 6*, 6089-6105. https://doi.org/10.32604/cmc.2022.023848

[11] Williams. B., Stokes, S. L., & Foster, J. (2022). Investigating record linkage for combining voluntary catch reports with a probability sample. *Fisheries Research, 251*(9), 1-9. https://doi.org/10.1016/j.fishres.2022.106301

[12] Liu, Z. & Valliant, R. (2021). Investigating an Alternative for Estimation from a Nonprobability Sample: Matching plus Calibration. *arXiv e-prints, 1*(2), 1-10.

[13] Lu, S., Cheng, G., Li, T., Xue, L., Liu, X., Huang, J., & Liu, G. (2022). Quantifying supply chain food loss in china with primary data: a large-scale, field-survey based analysis for staple food, vegetables, and fruits. *Resources, Conservation and Recycling, 177*, 1-8. https://doi.org/10.1016/j.resconrec.2021.106006

[14] Tontchev, N. T., Yankov, E. H., Gaydarov, V., & Hristov, N. (2022). Analysis and optimization of the properties of high-strength austenitic steels by approximation of a primary database. *Materials Science Forum, 1069*(95), 95-101. https://doi.org/10.4028/p-gs4b26

[15] Levis, J. M. & Zawadzki, Z. (2022). New directions in pronunciation research:previous research as primary data. *Journal of Second Language Pronunciation, 8*(3), 319-327. https://doi.org/10.1075/jslp.22049.lev

[16] An, Y. X. & Li, T. (2021). Tamper Proof Retrieval Method for Corpus Based on Distributed Cluster. *Computer Simulation, 38*(9):460-464.

Contact information:

Yun HE
College of Foreign Languages and International Education,
Quzhou University,
No. 78, Jiuhuabei Road, Kecheng District, Quzhou City, Zhejiang Province,
Quzhou, 324000, China
E-mail: 33043@qzc.edu.cn