Multi-mode Retrieval Method for Big Data of Economic Time Series Based on Machine Learning Theory

Hai-ying Chen[1] and Lan-fang Gong[2(✉)]

[1] Wuhan Institute of Design and Sciences, Wuhan 430205, China
chenhaiying3223@163.com
[2] Guangdong Polytechnic of Water Resources and Electric Engineering, Guangzhou 510925, Guangdong, China
liuyoudan2018@163.com

Abstract. For traditional search methods affected by the index build time, resulting in poor search results, a multi-mode retrieval method for big data of economic time series based on machine learning theory is proposed. According to the good extensibility of big data, construct a retrieval model and use binary data conversion methods to match big data. The binary sequence is defined by the relationship between different data, the similarity of data features is calculated, and the candidate candidate sequence is filtered. Data with no similar features are filtered, and each sub-sequence set matching the pattern is given by similarity size. After the threshold is added, on the basis of slightly reducing the filtering amplitude, the calculation of the similarity matching in the big data retrieval process is greatly reduced, and combined with the fixed interval sampling matching method to determine the characteristics of big data, thereby realizing the machine learning theory. The multi-mode retrieval method for big data of economic time series based on machine learning theory retrieval. According to the experimental comparison results, the retrieval efficiency of the method can reach 95%, which provides effective help for large-scale retrieval of massive data.

Keywords: First machine learning · Second economic time series · Third big data · Forth retrieval

1 Introduction

As computer technology continues to innovate, data related to inter-industry trade can be collected and is growing at an unprecedented rate. Due to the booming and rapid spread of cloud computing, the Internet, and the Internet of Things, all kinds of information data are exploding, the big data era has come, The real-time analysis of big data is an important analytical tool that promotes in-depth research and discussion in the academia and industry. Econometrics is a discipline that studies the quantification of actual economic phenomena. Most economic data were time series and have a strong theoretical background [1]. Machine learning theory uses regular thinking to avoid overfitting the model. In linear regression, most of the performance was to add more

S. Liu and G. Yang (Eds.): ADHIP 2018, LNICST 279, pp. 118–125, 2019.
https://doi.org/10.1007/978-3-030-19086-6_13

and more independent variable parameters, which can easily lead to overfitting, therefore, in machine learning theory, it is necessary to add penalty items involving the complexity of the model to the parameter estimation objective optimization equation, predict and mine the interesting features in the data and describe them, according to a series of nonlinear data analysis methods, the economic time series of machine learning theory not only has a strong search effect on known data, but also has good prediction ability for unknown data [2].

2 Big Data Multi-pattern Retrieval Scheme Design

2.1 Search Model Construction

For the economic time series of machine learning theory, the design of multi-mode retrieval for big data is divided into six modules according to good scalability: data extraction conversion and loading access tools, data indexing/storage service, big data index Library, data repository, data retrieval service and Web query interface [3, 4]. The retrieval model design is shown in Fig. 1.

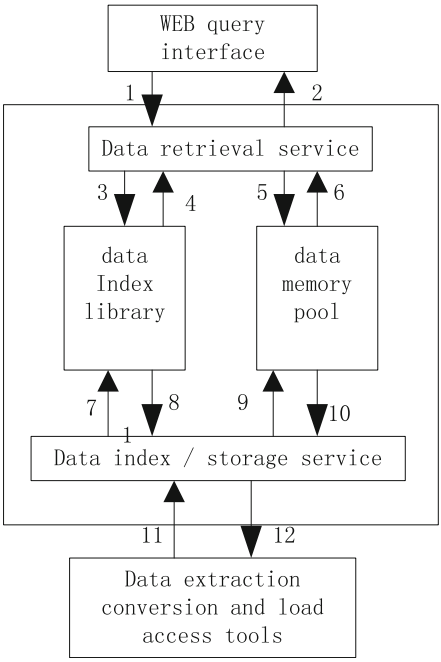

Fig. 1. Retrieval model

In Fig. 1, 1 indicates a sending request, 2 indicates a returned result, 3 indicates a search index library, 4 indicates a return, 5 indicates usage, 6 indicates return, 7

indicates index creation, 8 indicates return status, 9 indicates write storage, and 10 indicates return Status, 11 indicates write data, 12 indicates return status.

The data extraction transformation and loading access tool is mainly responsible for extracting the required raw data from the database. After conversion and cleaning, the data information including the data characteristics, applicable environment, retrieval process, and judgment flow is formed. Then the API interface provided by the index/storage service is called to load the data information into the data index database and data repository. Data extraction conversion and loading access tools currently support the following types of data sources: Kafka, Mongodb, Mysql, HBase, etc., by simply configuring the data source, data conversion XML, and data cleansing XML, the data extraction transformation and loading configuration of the access tool can be completed [5]. The data information index/storage service masks the data information index database and the data information repository externally by providing a unified call interface, thereby simplifying the access operation of external data information. In addition, in order to avoid the situation that the data information is only stored in the index storehouse and the storehouse when the index is not stored or only the index is not stored, the transaction control logic of the write operation is added.

The data information index database and the data information database are used to store the indexes and data of the data information, respectively. The advantage is that the full text retrieval function based on the inverted index is provided, and the obtained data performance is very stable and efficient. Therefore, the model separates the index and repository. In the process of writing data information into the model, the data information is written into the data information storage and the corresponding RowKey information is returned. Then use the data information and RowKey information to create an index in the case information index database [6]. In the query, firstly, it searches in the data information index database, returns the RowKey information, and then obtains data according to Rowkey in the data information index database, and returns it to the data information querying party. The data information retrieval service externally masks the data information index database and the data information storage library by providing a unified search interface and simplifies external retrieval operations. The WEB query interface mainly provides the WEB interface to the user to implement the query and display of data information [7].

2.2 Big Data Pattern Matching

In the above retrieval model, big data is extracted, and according to machine theory, binary data conversion method is used to match big data. The specific matching process is as follows:

(1) Data conversion: A binary sequence conversion method of character sequences based on the trend of change amplitude. This method defines the binary sequence by the relationship between the adjacent three points, thereby accurately reflecting whether the three points are convex growth (Decrease) or concave growth (Reduce) relationship;
(2) Data reduction: In order to facilitate the calculation of similarity between candidate sequences and patterns, a data reduction method based on trend proportions

is proposed. Both candidate sequences and patterns are reduced to the interval [0, 1], and the candidate after reduction is reduced. The minimum value of the sequence and pattern is 0, the maximum value is 1;

(3) Similarity calculation and filtering: In order to distinguish between the amplitudes of the convex growth (decrease) or the concave growth (decrease) of different change amplitudes, similarity is calculated for the reduced sequence, and the similarity degree is finally filtered. Give each set of sub-sequences that match the pattern [8].

For the matching of big data patterns, it can effectively solve the problem of data oscillation amplitude out of control, and solve the problem that the data sequence and pattern sequence segmentation rule are not similar as a whole, greatly improving the string matching efficiency, and simultaneously matching multiple candidate strings with other patterns. Sorting the similarity between them provides a basis for further accurate data retrieval [9].

2.3 The Implementation of Retrieval Scheme

Based on the above results of big data matching, an economic time series retrieval method based on machine learning theory was designed. The large data set A to be searched for uses the economic time series to perform similarity filter matching in the reference database $\{A_1, A_2, A_3, \ldots, A_n, \ldots\}$, as follows:

(1) If the total feature quantity of the intermediate data of the large data set A is greater than the total feature quantity of the intermediate data of the reference set, then it is directly determined that the second audio A is not similar;

(2) Set thresholds K_1 and K_2. In the process of data retention and match, if the average distance from the first n data economic time series is less than K_1, then it is directly determined that data n is a possible result;

(3) The threshold S is set. In the above process, the distance X of the previous m data economic time series is accumulated. If X is larger than the threshold S, the search is directly determined to be dissimilar;

(4) The threshold W is set. When the original data is used for matching, the first t feature feature similarity Y_t is compared first. Only when Y_t is greater than the threshold, the similarity degree Y_V of the overall data feature can be calculated; otherwise, it is directly determined that the search is not similar;

(5) The threshold η is set. In the above process, the original data feature retrieval error Rε is accumulated for the first n times. If Rε is greater than the threshold η, the retrieval is directly determined to be dissimilar;

(6) The threshold P is set. Because the adjacent economic time series have extremely high similarity, when the similarity between the data is lower than the threshold P, several times of data can be extracted appropriately, similarity matching is performed, and the big data retrieval is completed [10].

After the threshold is added, on the basis of slightly reducing the filtering amplitude, the similarity matching calculation amount in the big data retrieval process is greatly reduced, and the filtering speed and the retrieval speed can be effectively

improved. In combination with the fixed-interval sampling matching method, the big data feature is rapidly implemented. Similarity judgments can achieve more efficient retrieval results, thereby realizing economic time series big data multi-model retrieval based on machine learning theory.

3 Experiments

In order to verify the effectiveness of the economic time series big data multi-mode retrieval method based on machine learning theory, the following experiment was conducted. The experimental reference database uses 2,000 randomly collected data from the network and repeatedly loads the above data 20 times to a database with a scale of 40,000. The experimental data are all randomly selected from the database.

3.1 Experimental Results and Analysis

The data is divided by the size of 32 MB, 64 MB, 128 MB, 256 MB, 512 MB and the different index creation time is shown in Fig. 2.

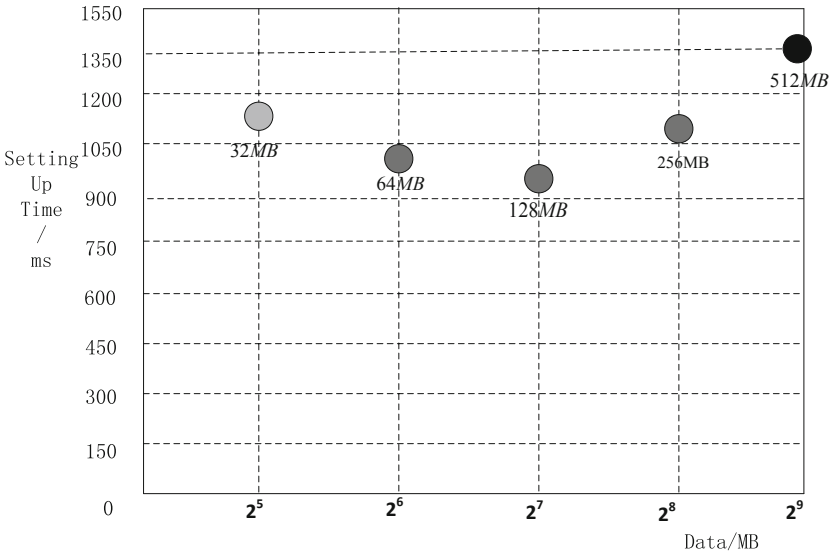

Fig. 2. Time for different index establishment

From Fig. 2 we can get the time for different index establishment. When the data is 32 MB, the index creation time is about 1100 ms. As the data grows larger, the index creation time is also reduced. Until the data size increases to 512 MB, the index creation time increases to 1350 ms. The index established using the traditional method is when the performance of 2 GB is the lowest, and it is lower than the normal time, this is because the set data size is the same, resulting in all the calculation results being

processed, which greatly wastes the computing power of the server, thereby reducing the performance of data retrieval. On the other hand, the large-data multi-mode retrieval method of economic time series using machine learning theory is not affected by this point, and has a good search effect.

In order to verify whether the different index creation time has an effect on retrieval results, the traditional method is compared with the economic time series method of machine learning theory. The results are shown in Fig. 3.

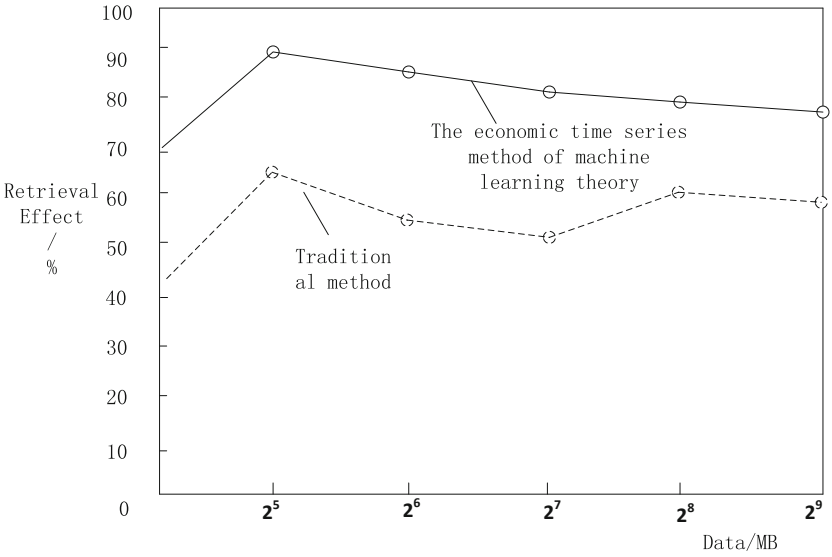

Fig. 3. Comparison of search results by different methods

From Fig. 3, we can see that the initial search effect of the traditional method can reach 78%, and the initial retrieval effect of the economic time series method of machine learning theory can reach 95%. As the amount of data increases, the retrieval efficiency gradually decreases. When the data is 32 MB, the retrieval effect by the traditional method is 65%, and the initial retrieval effect of the economic time series method of machine learning theory is 89%; When the data is 64 MB, the retrieval effect by the traditional method is 69%, and the initial retrieval effect of the economic time series method of machine learning theory is 85%; When the data is 128 MB, the retrieval effect is 55% using the traditional method, and the initial retrieval effect of the economic time series method of the machine learning theory is 82%; When the data is 512 MB, the retrieval effect is 59% using the traditional method, and the initial retrieval effect of the machine time learning method is 78%. It can be seen that using traditional methods will be affected by different index creation time, leading to poor retrieval results, while the economic time series method using machine learning theory will not be affected by different index setup time and have good retrieval results.

3.2 Experimental Conclusion

Based on the above experimental results, it can be concluded that when the data is 32 MB, the economic time series method of machine learning theory is 24% more efficient than the traditional method. When the data is 64 MB, the economic time series method of machine learning theory is 16% more efficient than the traditional method. When the data is 128 MB, the economic time series method of machine learning theory is 27% more efficient than the traditional method; when the data is 256 MB, the economic time series method of machine learning theory is 20% more efficient than the traditional method; when the data is 512 MB At that time, the economic time series method of machine learning theory was 19% more efficient than traditional method retrieval.

To sum up, the multi-mode retrieval method of big data based on machine learning theory is effective.

4 Conclusion

In the big data environment, many disciplines begin to merge, which brings the challenge of traditional search methods. Many emerging fields have attracted the attention of more and more researchers with the help of big data. Machine learning is one of the hot research directions. Machine learning pays more attention to the practical application value of research and relaxes the classical assumptions that are harsh on the model, focus on predicting and mining interesting data in the data and describe it. The experimental verification results show that, the use of machine learning theory in the economic time series big data multi-mode retrieval method has the retrieval effect of 95%, and has a good application prospect.

Fund Project. The 2018 annual scientific research project of Hubei Provincial Department of education. Based on modern statistical theory and machine learning theory, economic time series analysis is carried out (B2018371).

References

1. Su, T.F., Liu, Q.M., Su, X.C.: Research on crop remote sensing classification based on multiple vegetation index time series and machine learning. Jiangsu Agric. Sci. **45**(16), 219–224 (2017)
2. Dong, F., Liu, Y.F., Zhou, Y.: Abstracts based on LDA-SVM abstracts multi-classification emerging technologies prediction. J. Inf. **36**(7), 40–45 (2017)
3. Zhu, X., et al.: Research on network purchase behavior prediction based on machine learning fusion algorithm. Stat. Inf. Forum **25**(12), 94–100 (2017)
4. Sun, C.Y., Gong, L.T.: Research on interest rate pricing under the big data thinking: an empirical analysis based on machine learning. Fin. Theory Pract. **18**(7), 1–5 (2017)
5. Li, L., et al.: Parallel learning-a new theoretical framework of machine learning. Acta Automatica Sinica **43**(1), 1–8 (2017)

6. Jiao, J.Y., et al.: Review of typical machine learning platform under big data. J. Comput. Appl. **37**(11), 3039–3047 (2017)
7. Wu, Y.L., et al.: Construction and prediction of prospecting model based on big data intelligence. China Mining Mag. **26**(9), 79–84 (2017)
8. Xia, J.M., et al.: Physiological parameter monitoring system based on K-means and MTLS-SVM algorithm. Telecommun. Sci. **16**(10), 43–49 (2017)
9. Xing, X., et al.: Analysis of characteristics of multi-state traffic flow combined with viewable time series. Acta Physica Sinica **66**(23), 51–59 (2017)
10. Mei, Y.: Simulation of resource target information extraction in big data environment. Comput. Simul. **35**(03), 337–340 (2018)