Research on Similarity Search Techniques for Time Series Subseries

Mengru Zhang

Abstract— **The representation of time series and complex similarity measurement are the basis of time series similarity research, and play a vital role in completing the task of time series similarity search. The similarity retrieval of time series subsequences is at the core of data mining technology. It is more and more widely studied in different application fields (including neuroscience, finance, meteorology, human health detection, data retrieval and so on). However, due to the massive and high dimensionality of data sequences, the difficulty of data mining is significantly improved. Therefore, after we obtain the time series, domain experts are faced with the problem of data analysis and processing. The commonly used method is to represent the feature of the data sequence, so as to effectively reduce the dimension of the data through the feature representation, and then use the distance measure to distinguish the similarity. Therefore, we use the combination of time series representation and similarity measurement to realize the similarity retrieval of time series subsequences.**

Index Terms— **Time series; Subsequence query; Time series representation method; Measurement method of time series; Similarity retrieval**

1. INTRODUCTION

In recent decades, with the vigorous development of computer technology such as artificial intelligence and big data, obtaining information with the help of data mining technology has become an essential part of people's daily life. Data series (also known as time series) is one of the most common data types. They exist in almost every scientific and social field, such as meteorology, medicine, neuroscience, finance, agriculture, entomology, sociology, voice detection[1], operational health monitoring, information retrieval, etc[2]. This makes the data sequence a specific important data type[3]. Therefore, mining similar time series is very important and common.Time Series

The similarity query of time series can be divided into full sequence query[4]and sub sequence query[5]. Full sequence query is that the length of our existing data sequence is the same as the given sequence to be queried. The subsequence query is that the given query sequence is relatively short. We want to find similar fragment information on the existing long data sequence. However, in the research process of time series subsequence similarity query, we will face the problem of low efficiency of the query process. Generally speaking, the subsequence query of time series is divided into two parts: time series representation[6] and time series similarity measurement[7].They interact and complement each other. Therefore, in the work of data mining, researchers need to first obtain the main features of the sequence by using feature representation and other methods, so that we can get the dimensionality reduced data. Then, we can calculate the similarity measurement of sub sequences for the dimensionality reduced data, find similar fragments and obtain the sequence information we want, Query completed.

The common retrieval technology is to query the similarity of time series through index establishment. In recent years, many experts and scholars are more and more interested in this research. The similarity retrieval of time series occupies an important position in data mining. This problem needs us to explore and innovate slowly. Similarly, similarity measurement, as the basis of data sequence similarity query, also occupies an important position in the field of data mining. Using index technology to achieve the purpose of query, can reduce the complexity of query, improve search efficiency and reduce the waste of resources, so as to achieve the purpose of query, which has also become an important link of sequence retrieval.

To sum up, we have great research value for data containing a lot of potential information such as time series. In order to reduce the difficulty of similarity measurement, we need to represent the time series first to reduce the dimension of the series. In order to improve the efficiency of similarity query, we build an index to achieve a more concise and efficient search. We propose a new analysis method, which is of great significance for the similarity query of time series variable length subsequences.

2. BACKGROUND AND RELATED WORK

In recent decades, data mining of time series has developed rapidly. In view of the relatively large scale of time series database, most studies are committed to the fast search process. At present, the most effective method is the technology mentioned above, which first reduces the dimension of the data, and then uses the spatial access method to index the data in the transformed space. The work related to time series data mining based on time series similarity query technology was initially proposed by foreign experts and scholars Agrawal et al[8] (1993), and extended and optimized by faloutsos et al. (1994), rafiei and mendelzon (1997), Keogh and pazzani (1999). The teams of Agrawal and pzzanai of well-known foreign IBM companies have been committed to the task of data mining of time series. Of course, the most prominent research on data mining technology also belongs to Professor Eamon Keogh of the University of California and the research team led by him. The UCR suite and sax representation for retrieval optimization proposed by him have provided great help to subsequent researchers.

Manuscript received March 11, 2022
Mengru Zhang, School of Computer Science and Technology, Tiangong University, Tianjin, 300387, CHINA

3. Time Series Representation and Similarity Measures

According to the transformation method of time series representation, we divide them into three categories[9], namely model-based representation, data adaptive representation and non data adaptive representation. The three representation methods are arranged according to the hierarchy, as shown in Figure 1 below.

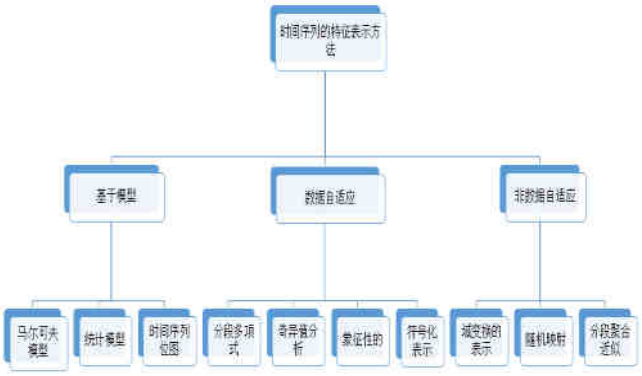

Figure 1 : three types of feature representation methods of time series

Piecewise aggregation approximation (PAA) reduces the dimension of time series by sliding windows. For the time series values within each window, the mean method is used for aggregation approximation, and the calculated mean value is used to represent the coefficient values within each window. PAA is calculated as follows: set the time series T=(x_1,……, x_n), there is m < n, and after dimensionality reduction, there is T'=($\bar{x}_1$,……$\bar{x}_m$).Expressed as:

$$\bar{x}_i=\frac{m}{n}\sum_{j=\left(\frac{n}{m}\right)(i-1)+1}^{\left(\frac{n}{m}\right)i} x_j$$

When some window sizes cannot be divided, you can choose to discard or average the remaining sequences again. The dimension reduction with PAA representation retains the characteristics of the original time series better, because the PAA dimension reduction method obtains the average value of the time series segments, and the dimension reduction speed is fast. Of course, this dimension reduction representation is suitable for the situation that the data series changes more smoothly, because in the case of severe change amplitude, PAA processing may lose some characteristic information of the original sequence.

Symbolic aggregation approximation (sax) representation is based on PAA. It needs to segment the time series first. It is to convert the time series into the form of string. It is represented by Sax (D, W, |alpha| - 1), where d represents the time series, w represents the dimension divided by the time series, and| alpha| - 1 represents the number of breaks.

The symbolization is completed by approximating the sequence values of breakpoint sequence B and PAA. According to the data, the Sax breakpoint diagram is shown in table 1:

Table 1: sax breakpoint diagram

α / β_i	2	3	4	5	6	7	8
β_1	0.00	-0.43	-0.67	-0.84	-0.97	-1.07	-1.15
β_2		0.43	0.00	-0.25	-0.43	-0.57	-0.67
β_3			0.67	0.25	0.00	-0.18	-0.32
β_4				0.84	0.43	0.18	0.00
β_5					0.97	0.57	0.32
β_6						1.07	0.67
β_7							1.15

Euclidean distance (ED), also known as Euclidean distance, calculates the distance between two sequences, and EUC(T_i^w, S_j^w) representsThe distance between T_i^w and $S_j^w$and w sequences, where W is the length of the sliding window, $1 \le i \le n$, $1 \le j \le n$. Measure the absolute distance between points in n-dimensional space[10]. Euclidean distance is used to calculate the similarity between two sequences. The operation is relatively simple. It represents the actual distance between two points in n-dimensional space. The Euclidean distance in two-dimensional and three-dimensional space, that is, the distance between two points, is used more frequently. Given $T(a_1, b_1)$ and $S(a_2, b_2)$, the Euclidean distance between T and s can be calculated by the following formula:

$$\rho = \sqrt{(a_1 - a_2)^2 + (b_1 - b_2)^2}$$

4. Overview of Index Tree Structure

4.1 Index construction

The index is constructed given the base cardinality b, word length w and threshold th. The index structure subdivides the Sax space hierarchically, resulting in the difference between time series entries until the number of entries in each subspace is less than. The index tree contains three types of nodes: root node, internal node and terminal node[11].

1. Root node: the root node represents the complete Sax space, and its function is similar to that of the internal node. The Sax word encountered corresponds to some internal nodes or terminal nodes, and the index function is guided accordingly.
2. Internal node: an internal node specifies a split in Sax space. Usually, an internal node is created when the number of time series contained in the terminal node exceeds a given th.

3. Terminal node: the terminal node is a leaf node, which contains a pointer to the index file with the original time series on the disk.

4.2 Algorithm flow

Firstly, we read in a period of time series ts, and then preprocess the data, which is the key step of time series analysis. In this paper, z-normalization is used for data standardization. After standardizing the data into data with standard deviation of 1 and mean value of 0, the data sequence is cut and dimensionality reduced. Firstly, the sequence is divided into equal segments, and the average value of each segment is calculated as the value of this segment to reduce the overall length of each segment. The dimension reduced data of PAA is used as input and converted into data expressed in characters for subsequent index construction and data query.

During index construction, we should pay attention to the difference between terminal nodes and internal nodes. The tree is constructed according to the information of Sax. The tree structure includes root nodes, internal nodes and terminal nodes. The general construction process is Firstly, sequence a is cut into subsequences with the same length as B, which are recorded as alist, then B is symbolized as BS, then alist is iterated, and each subsequence is symbolized as ATS. Then the Euclidean distance between ATS and BS is calculated through the character similarity comparison table, which is recorded as seqdis. After traversing, the minimum value of seqdis is obtained, and the corresponding subsequence is the subsequence with the highest similarity with B.

4.3 Similarity search of variable length subsequences

After the index tree is constructed, the next step is to search. Given another small time series, the similarity search is carried out on the constructed index tree. When doing the similarity search, a priority queue needs to be created. When searching, a BSF needs to be calculated to judge the nodes in the priority queue whose distance is less than the BSF. If the distance is less than the BSF, the search judgment will be carried out first. Otherwise, the point and its child nodes are discarded. Finally, the similar sequence found will be output to nodes, and the sequence diagram will be output. Moreover, the length of the exact similar search is uncertain. The length of the search sequence of the similar search sequence is determined according to the given query sequence, that is to say, the search sequence is not necessarily the same length each time. In a word, the similarity search of variable length subsequences is carried out, and the index tree is built to speed up the search speed.

5. Experimental conclusion and analysis

In this experiment, the purpose of establishing the index is to realize more effective similarity search and find subsequence fragments similar to the query sequence more accurately in the subsequence of time series. Therefore, the index of experimental evaluation in this section is the comparison of the running time of the algorithm and the fragment pruning rate during the operation of the algorithm. In this section, two data sets are used and compared with the violence algorithm.

5.1 impact of different algorithms on running time

Firstly, the random walk data set is used to compare the violence algorithm, hash table construction algorithm and retrieval algorithm based on index tree structure. From the experimental results, it can be seen that our retrieval algorithm based on tree structure is significantly better than the first two retrieval algorithms in speed, and the clipping rate is more than 90%, indicating that the second algorithm (variable length sequence retrieval algorithm based on Sax index tree) can reduce computing resources to a great extent. The retrieval time of specific experimental data is shown in table 2 below:

Table 2: comparison of running time between two search algorithms

A/B length	1000000/60	2000000/80	3000000/100	4000000/120	5000000/140
Sequential scanning/s	7.226	10.779	13.414	14.764	15.896
Hash table structure/s	7.42	0.016	0.013	7.049	8.340
Index tree structure/ms	299	106	118	99.7	98.2

Conclusion

Through a large number of experiments, this paper verifies that our algorithm has certain advantages compared with the previous sequential scanning and its UCR suite. The high dimensionality of time series determines the challenge of data mining. Although in this paper, we put forward innovative ideas for subsequence query to realize query optimization, these concepts still need to be further discussed, and the shortcomings of the solution need to be further analyzed and processed. In view of the shortcomings of this work, our subsequent research can be explored from the following aspects. For the measurement tool of index tree method, we choose the measurement based on Euclidean distance, which limits the scalability of data query. We can consider combining index tree with DTW based measurement method to realize the similarity retrieval of variable length subsequences.

References

[1] Kashino K, Smith G, Murase H. Time-series active search for quick retrieval of audio and video[C]//1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258). IEEE, 1999, 6: 2993-2996.

[2] Palpanas T. Data series management: The road to big sequence analytics[J]. ACM SIGMOD Record, 2015, 44(2): 47-52.

[3] Linardi M, Palpanas T. Scalable, variable-length similarity search in data series: The ULISSE approach[J]. Proceedings of the VLDB Endowment, 2018, 11(13): 2236-2248.

[4] Dau H A, Begum N, Keogh E. Semi-supervision dramatically improves time series clustering under dynamic time warping[C]//Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016: 999-1008.

[5] Qu J F, Yuan L, Huang Y, et al. An efficient algorithm for attribute-based subsequence matching[J]. Information Sciences, 2016, 334: 323-337.

[6] Deng W, Wang G, Xu J. Piecewise two-dimensional normal cloud representation for time-series data mining[J]. Information Sciences, 2016, 374: 32-50.

[7] Hajihashemi Z, Popescu M. A multidimensional time-series similarity measure with applications to eldercare monitoring[J]. IEEE Journal of biomedical and health informatics, 2015, 20(3): 953-962.

[8] Zoumpatianos K, Palpanas T. Data series management: Fulfilling the need for big sequence analytics[C]//2018 IEEE 34th International Conference on Data Engineering (ICDE). IEEE, 2018: 1677-1678.

[9] Belhadi A, Djenouri Y, Nørvåg K, et al. Space–time series clustering: Algorithms, taxonomy, and case study on urban smart cities[J]. Engineering Applications of Artificial Intelligence, 2020, 95: 103857.

[10] Kadiyala S, Shiri N. A compact multi-resolution index for variable length queries in time series databases[J]. Knowledge and information systems, 2008, 15(2): 131-147.

[11] Lines J , Bagnall A . Time series classification with ensembles of elastic distance measures[J]. Data Mining & Knowledge Discovery, 2015, 29(3):565-592.