Research on Time Series Query Method Based on Linear Hash Index

Xi Lu and Xin-Ai Xu

Abstract: With the continuous development of China's information technology, many data will also be produced. In these data, time series data is an important data type. In this paper, we propose a new query processing method for time series, in order to reduce the index creation time and improve query efficiency. Experimental results show that the linear hash index in this method is reduced in the time of creation. In the query phase, the method of combining the K nearest neighbor and the lower bound distance is used to filter out the redundant results.

Keywords: time series; linear hashing; K- nearest neighbor; lower bound distance

In this paper, we propose a new method to deal with time series query based on linear hashing as an index technique. In the preprocessing stage of time series, a new normalization method is proposed, which preserves the original form of time series. The linear hash indexing mechanism is used to access the time series effectively and naturally. In the query phase of time series, a lower bound distance method is proposed. And combined with the K nearest neighbor method, the effect of query is improved, and the query needs of users for the time series are completed.

1. Overview

Time series refers to the same numerical index according to the time sequence. It has happened a series of it as a special form of temporal data in many fields, such as temperature and humidity of the financial stock price, medicine, meteorology, heart EEG trend of enterprise production. Said the time series is aimed at the structure of time series and will take complex time series
deformation technology; time series index is for how to conduct efficient storage and fast query technology, time series. Because of the large amount of data and complex structure of time series, it is difficult to represent and index. The existing search technology, processing large data often takes a lot of time. The similarity measure is not accurate enough. Researchers at home and abroad have provided a lot of similarity measurement techniques, but the integrity and accuracy of the query still needs to be improved.

2. Related work

2.1 Time series is a kind of complex data types which change with time. In the form of a time series, where the element is a sequence of points, which represent time, representing the value of the time series at the moment. The research on query processing can be divided into time series representation, indexing and similarity measurement. The method of time series representation includes discrete wavelet transform and discrete Fu Liye transform, singular value decomposition, piecewise linear representation, symbolic approximation method and so on.

2.2 Time series indexing technology based on space division index and based on data partition index, space partition based on K-D tree, four fork tree, grid file, data classification based on R-tree, iSAX-tree and ADS-tree.

2.3 Similarity measure of time series is a method to measure the relationship between time series. Similarity measurement is an important task in data mining. In general, each similarity measure of time series can represent one or more time series features.

3. Time series preprocessing

3.1 Time series normalization

The standardization of the so-called time series is to convert the original time series into the interval in the case of preserving the original general trend. Thus, the original time series has been standardized, compared with the previous, the standard time series, the trend of the original sequence is perfectly preserved, and the data distribution is relatively uniform.

3.2 Time series segmentation

Usually the size of a single time series is also larger. The definition of a time series, a collection of time series of the database, assuming that the length of the time series of Y n, the N unit of the elements into N segments. For simplicity, suppose that N is a factor of n. A length of N time series X, calculated by vector formula. Assuming that the temperature data from 1 to 16 points in a day, a time series with a factor of 16. The number of selected segments is 4, and the time series is divided into 4 regular boxes. The average value of each frame is calculated. Vector is a representation of time series.
segmentation. This transformation converts the original time series into a piecewise constant approximation.

3.3 Time series discretization

Most of the time series are consistent with the Gauss distribution, which can be discretized. A standard normal distribution time series is established, which decides the breakpoint to divide the target variable of time series into several regions.

4. Linear hash index

Linear hashing is a dynamic hashing technique, the basic idea is to use the hash function, the time series retrieval value is mapped to a fixed hash number, then you can find the time sequence of unknown origin. Through the normalization of time series, most of the time series are distributed in the standard normal distribution, which can solve the problem of the heterogeneity of the data distribution. Linear hash round robin splitting mechanism: defines a loop level that is used in one loop level and the two hash functions. After the beginning of the cycle split one by one, after the end of a cycle split began to split the next round, until the end of the cycle.

5. Time series query

5.1 Approximate query

The application of data mining requires approximate query, and the linear hash index can support fast approximate query, because the symbolic representations of two similar time series are often the same. The time series representation of the results from the linear hash index file is searched. Based on the exact query algorithm of linear hash index, the results obtained by approximate query (BSF) are used as input. Because the time series between the BSF results in the distance is relatively small, to the nearest neighbor query to create conditions in the approximate query stage will prune most of the search space, which improves the query precision, but also reduces the query time.

The core of the K nearest neighbor is to find the neighbor of the time series, which is to find the time series adjacent to the target sequence. Measurement of two time series is not a neighbor criterion, can be intuitively understood as the distance between the two time series, if the distance in the acceptable range, you can determine the two time series is a neighbor, because the distance between two time series in the feature space can reflect the degree of similarity between the two time series.

K nearest neighbor query will approximate query results as K nearest neighbors in the data set, when the input of new time series, in the time series data set and find the target time sequence nearest K neighbors, that this K time
series and time series of the most similar. When K takes 1, the query arrives at the exact query.

According to the distance measure method, we find out the nearest K time series in the time series data set (S) (BSF), covering the neighborhood of the K time series. The corresponding feature space partition is the corresponding K nearest neighbor model.

It is assumed that all points of the feature space form an approximate query result set BSF. Given a query example time series X0, through the X0 rough approximate query result set BSF, then from the existing BSF results as time series to eliminate and X0 may not close, get accurate result set. The distance between the two time series indicates the degree of similarity between them, and there are many methods to calculate the distance, usually using Euclidean distance, distance and Minkowski distance. Suppose Xi and Xj are two time series in the feature space

5.2 Closing discussion

The BSF result set is not directly ordered to traverse the result set because the time consumption is still relatively large, and the results are filtered out by adding a tight TLB. The so-called TLB is a very meaningful method for similarity measure. Among them, T and S are two time series. The advantage of TLB is that B achieves a complete free measure, which can effectively predict the validity of the index. If the value of TLB is 0, it is proved that the index needs to read out the time series from the disk, and the index is not efficient. If the value of TLB is 1, it is proved that a slight adjustment to the index can be used to retrieve the required time series, and to ensure that the real nearest neighbor.

6. Verification results

Selection of time series length, cardinality and segment size. The system takes complete testing, comparison and selection, mainly divided into two parts to create contrast index time and query time, the index creation time is early time series representation, sum index creation time. Either way, the index creation and preprocessing takes up most of the time spent in the whole time, and the query time is a small part of the whole time. The performance improvement of index creation is not obvious compared with iSAX; it is mainly to improve the query, so as to verify the validity of the lower bound distance.

7. Conclusion

In this paper, the time series based on the existing method, proposed a standardized time series method, and try to use the linear hash create time series index in time series similarity query is proposed and a new lower bound distance to measure the similarity between time series. After the validity of the
obtained time series standardization method, verified the linear hash in the index time series, index creation time is not greatly improved, but reduces the time sequence of query time. The lower bound distance method is discussed in this paper.

*Jiangxi Provincial Department of Education Science and technology research projects (GJJ151255); Project supported by Nanchang Teachers College (15KJZD39); Project supported by the data mining innovation team of Nanchang Teachers College.

8. References