Data Retrieval Method for Efficient Parallel Skyline Computing

Xiaofei Li
Jilin University of Architecture and Technology, Changchun 130012,China.

Abstract: Skyline retrieval is a broad data processing method, especially for multi-keyword retrieval. On the basis of analyzing the deficiencies of the existing Skyline algorithm, a data retrieval method for efficient parallel Skyline computing is proposed. A database index structure Par-Tree is constructed, and signature information is added to reduce the bit collision probability in the retrieval overshoot and filter out the retrieval area irrelevant to keywords, the irrelevant information points are pruned. Based on the Par-Tree index structure, a multi-keyword mining algorithm PSkyline algorithm is proposed. The experimental results show that the method improves the execution efficiency of data mining, and can effectively solve the multi-keyword Skyline retrieval problem.

Keywords: Data Mining; Skyline; Data Retrieval

Introduction

One of the key techniques for improving data mining performance is optimizing retrieval. Therefore, for data mining problems, finding an optimal mining plan has become an important content in data mining research[1-2].

Building a multi-connection search tree is the lowest cost of searching a database optimally. In order to solve the problem of optimizing the retrieval of large databases, many retrieval optimization algorithms have been proposed by scholars at home and abroad [3-4]. Traditional retrieval optimization algorithms use a full search algorithm. This type of algorithm is only suitable for when the number of connection relationships of objects in the database is small, and when the number is large, the retrieval speed and efficiency are very low. In the big data environment, the number of retrieval connections in the database is large. In order to solve this problem, relevant scholars proposed a dynamic programming algorithm for optimization, but the query efficiency is still low [5].

Based on the above problems, this paper improves on proposes a data retrieval method for efficient parallel Skyline computation, constructs the index structure according to the characteristics of database queries, and optimizes the Skyline algorithm. Experimental results show that this method improves the execution efficiency of data retrieval.

1. Index structure building

Aiming at the problem of low efficiency of multi-keyword matching in data mining, this paper constructs a keyword index structure Par-Tree.

Par-Tree is divided into two layers: (1) upper index: on the basis of the R-Tree index structure, the keyword is set for signature, and the signature information is added to the node, so as to find the relationship between the keyword information and the spatial region of the mining object; (2) Lower index: The structure of the inverted table is constructed, which can reflect the mapping relationship between keywords and mining object information.

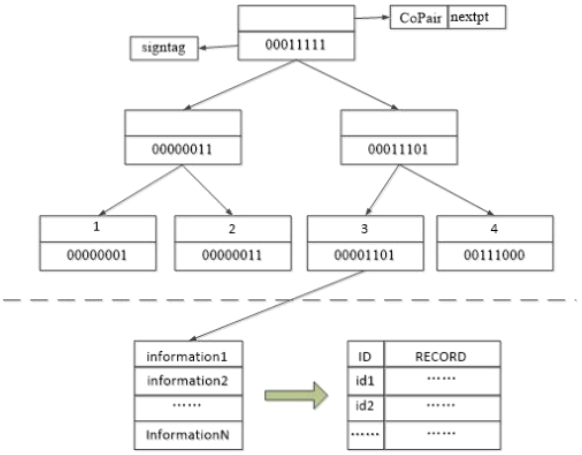

Figure 1 Par-Tree index structure

In the upper index, <CoPair, nextpt, signtag> are the storage structures of the node, where CoPair is the data information location coordinate pair, which represents the regional location information of the data; NextPT is a pointer to the next node; signtag is the signature information for the current location. The id in the leaf node represents the atomic number of the position. The signature information of the position is represented by an 8-bit binary code, and the keyword information is converted into a binary code through the Hash function. In the lower inverted index, keyword information and bit vector information are stored in each node. Each bit of the wherein bit vector corresponds to the internal number of the region.

2. Skyline data retrieval methods

2.1 Skyline data mining algorithms

In order to solve the problem of multi-keyword Skyline search efficiency, based on the Par-Tree index structure, the Skyline data retrieval algorithm-PTSL algorithm is proposed.

In the process of traversing the Par-Tree index, the algorithm filters the data information text collection by comparing the keyword position information with the query keyword information in the upper index.

In the lower index, the leaf nodes are traversed, and the data satisfying the search keywords are obtained through fast operations between bits, so as to obtain the candidate set of the relevant region. The Skyline data mining algorithm based on Par-Tree index is as follows:

PTSL algorithm

Enter: search point p, search keyword p.k, search range W, data information set S, Par-Tree index

Process:

```
1.  TempS ←{ };TS ←{ };
2.  While !Node.isEmpty() do
3.  NS ←Node.pop( )
4.  if NS.isInRange(p.k,W) then
// The search keyword p.k matched the search range W
5.  if NS.isLeaf() then
6.  TS ← getSet(p.k)
//  Obtain a collection TS that satisfies the search keywords
7.  for ts in TS
8.  TempS ← PSkyline(TempS,TS,p.k,W)
9.  Else
10. Node.push(NS.getChild());
```

In the PTSL algorithm, the unaccessed nodes in the upper index nodes of Par-Tree are first maintained in the form of stacks, and then the retrieval area is determined, and when the leaf nodes are retrieved, the inverted index is used to calculate the set TS that meets the search conditions. Finally, the method is called in a loop to dominate the judgment keyword and generate a new intermediate result set TempS.

2.2 Filtering Strategies

Due to the calculation of keyword dominance judgment between intermediate result set TempS and candidate set TS, the method is time-consuming and frequent. Therefore, this paper performs space optimization to improve the mining efficiency through filtering strategies.

Based on the above definition, this article adopts the following filtering strategy:

(1) Min filtering method: set a small top heap structure, and the top object tp is the object point closest to the retrieval point p-weighted distance in the intermediate result set TempS. The weighted distance of the candidate point TS is then judged, if less than TP, according to (2). According to this law, in the subsequent calculation, only the keyword dominance relationship needs to be determined, and the points that are not dominated by TS in the intermediate result set can be directly deleted.

(2) Sum filtering method: According to the attributes of numeric objects, the filter is determined based on the numerical sum of keywords.

3. Experimental results and analysis

This experiment mainly analyzes the change of execution time with the size of the dataset. As can be seen from Figure 2, the execution time of the PSkyline algorithm increases approximately exponentially as the tuples in the dataset increase, while the execution time of the INKS algorithm and the STD algorithm accounts for about 10% of the PSkyline algorithm. The initial execution time of the STD algorithm is similar to that of the INKS algorithm, and the execution time of the STD algorithm is gradually lower than that of the INKS algorithm with the increase of the size of the dataset. The size of the result set varies with the size of the dataset, and the experimental results are shown in Figure 2. The PSkyline algorithm produces a smaller number of tuples in the result set, which can reduce the number of extra tuples.

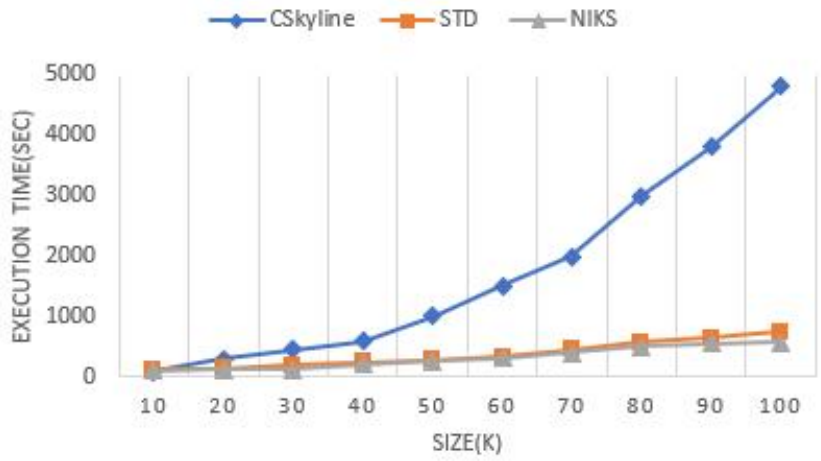

Figure 2 The influence of dataset size

In order to verify the influence of the number of keywords on the performance of the algorithm, the architectural design case database is partially extracted in the experiment, and the data dimension is 4. The change in the execution time of the algorithm increasing the number of keywords from 1 to 10 when the coordinates of the q region of the retrieval point are consistent, as shown in Figure 4. It can be seen that the PSkyline algorithm in this paper is significantly better than the other two algorithms when the keywords are high, and the signature information is used for multi-keyword matching and the hash function is used for mapping, which effectively improves the mining speed of multi-keywords.

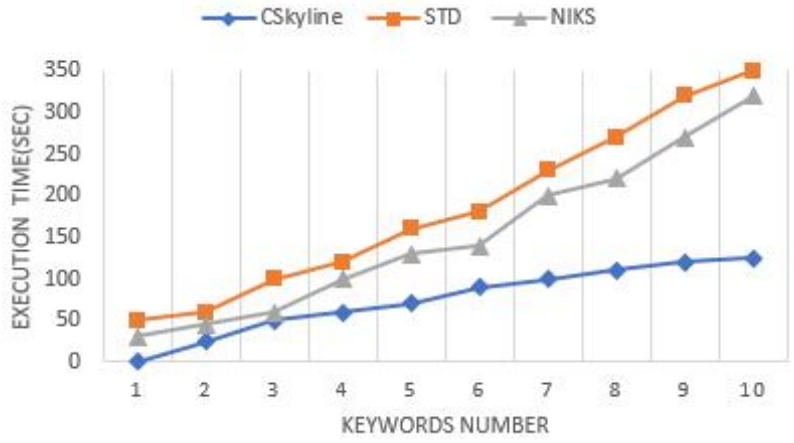

Figure 3 The influence of the number of keywords

4. Conclusions

This paper proposes an architectural design data mining method based on Skyline algorithm, constructs an index structure Par-Tree according to the characteristics of database query, adds signature information, reduces the probability of bit collision in the retrieval overshoot, filters the search area unrelated to keywords, and prunes irrelevant information points. Based on the index structure Par-Tree, the PSkyline algorithm of multi-keyword mining algorithm is proposed. Experimental results show that this method effectively improves the execution efficiency of architectural design data mining, and can effectively solve the multi-keyword Skyline retrieval problem in architectural design cases.

References

[1] Kalyvas, C.; Maragkoudakis, M. A skyline-based decision boundary estimation method for binominal classification in big data. In Proceedings of the 2020 5th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM), Corfu, Greece,25–27 September 2020

[2] Ouadah, A.; Hadjali, A.; Nader, F.; Benouaret, K. Sefap: An efficient approach for ranking skyline web services. J. Ambient Intell. Humaniz. Comput. 2019, 10, 709–725.

[3] Gil D, FerrÆndez A, Mora-Mora H, and Peral J, ‘‘Internet of Things:A review of surveys based on context aware intelligent services’’ Sensors,vol. 16, no. 7, p E1069, Jul. 2016.

[4] Song H, Rawat D, Jeschke S, and Brecher C, Cyber-Physical Systems:Foundations, Principles and Applications. Boston, MA, USA: Academic,2016, p. 514.

[5] Chen ZJ, Li SY, Liu WY. Range-constrained Top-k keyword query on road networks[J].Journal of Chinese Computer Systems, 2017, 38(12): 2707-2713.

Fund project: Jilin Institute of Architecture and Technology Research Project: A Data Partition Method for Efficient Parallel Skyline Computing. (NO.[2021]025ZQKJ)