Research on Big Data Ad Hoc Queries Technology Based on Social Network Information Cognitive Model

Yang-bo Wu[1], Xi-liu Zhou[1(✉)], Ying Xu[1], Ming-xiu Wan[1], and Lin Ying[2]

[1] College of Mathematics and Computer, Xinyu University, Xinyu 338000, China
[2] College of Foreign Languages, Xinyu University, Xinyu 338000, China

Abstract. In order to improve the efficiency of big data ad hoc query, a big data ad hoc query technology based on social network information cognitive model is designed. Firstly, the hierarchical structure of data query is established, then the dimension measurement and query conditions are determined, and the result analysis function is proposed. On this basis, the linkage update model of distributed multi database information resources is constructed, and finally the query data is clustered and stored, so as to realize on-the-spot big data query. The experimental results show that the research of big data ad hoc query technology effectively improves the query efficiency, and improves the accuracy of data query.

Keywords: Social network information cognitive model · Ad hoc query · Big data

1 Introduction

In recent years, with the continuous development of society and the continuous improvement of business model, the business data in the enterprise has achieved explosive growth. The traditional management software of each business department has been unable to meet the growing data needs. It is urgent to seek a software that can not only deal with these huge data, but also provide decision support, provide reliable and accurate comprehensive decision support for enterprise management, and realize the integrated business intelligence of the whole enterprise. Business intelligence has built-in complete data analysis function. Without other auxiliary tools and rich it knowledge, users can realize data analysis work such as data specification, semantic mapping, real-time data query, multi-dimensional data analysis, Chinese style report compilation, visual graphic design, etc., improve the utilization rate of data, and significantly save users' investment. Similarly, business intelligence can also be closely combined with the existing business systems of enterprises to accurately analyze the massive data generated by the existing resource management system (ERP), customer relationship system (CRM), business process system (BPM), financial performance management system (FPSM), business analysis system and other core systems of enterprises. Generate clear and reliable data analysis results and embed them, so as to provide reliable data analysis means

S. Liu and X. Ma (Eds.): ADHIP 2021, LNICST 417, pp. 183–195, 2022.
https://doi.org/10.1007/978-3-030-94554-1_16

for users at different levels of government, finance, electric power, communication and other enterprises and institutions in various business scenarios, and significantly improve customers' existing system data analysis ability.

At present, there have been some good research results on big data query. Reference [1] presents a low latency publishing protocol for large data query process confidential data based on Manhattan metrics, classifies confidential data matrices by information gain method, extracts confidential data features from classified matrices, divides them into different data clusters according to feature differences, uses Manhattan metrics algorithm to measure the change of data cluster centers before and after, determines the low latency value according to the relative distance of cluster center changes, and judges the low latency situation of confidential data by comparing the low latency value with the selected low latency threshold size. Reference [2] presents a sampling method based on acceleration ratio and potential distribution, which supports various sampling algorithms, realizes the assurance of randomicity, performance assurance and nearness evaluation of sampling and query in a distributed environment, and is compatible with accurate query. This method can be quickly applied to the columns with a large amount of data, and has good scalability and maintainability. However, the two or more traditional methods cannot realize the impromptu query of large data and cannot be applied to social networks.

Based on this background, this study designs a big data ad hoc query technology based on social network information cognitive model. Ad-hoc Queries is a concept in the field of data warehousing that refers to queries that a user customizes to his or her current needs when using the system. Ad hoc statistics is an aggregation operation based on ad hoc query, which is a kind of statistical function of user's temporary definition of statistical indicators. Compared with fixed statistics, ad hoc statistics has the advantages of flexible use and personalized definition. This paper establishes the hierarchical structure of social network data query, determines the measurement of dimension and query conditions, and gives cognitive analysis function of social network information. This paper constructs a link updating model of distributed multi-database information resources, clusters and stores the query data, and realizes the extempore query of big data. This technology can greatly improve the efficiency of social network information processing.

2 Establishment of Data Query Hierarchy

Hierarchy is the collection of members in a dimension and the relative position between them. Hierarchy is the logical structure of organizing data, which enables us to organize and view data organically instead of facing a large number of detailed data directly. Generally speaking, drill down includes a behavior of traversing the hierarchy, that is, traversing from the high level of a one-dimensional hierarchy to the low level of detail, while roll up includes a process of generalizing the low-level detail data to the high-level data in a one-dimensional hierarchy. Hierarchy is an indispensable part of business, and it is also a basic problem in data warehouse and OLAP. There are two types of hierarchy: level based and parent child value based. This paper mainly discusses level based hierarchy, because it is more common and universal in practice. Hierarchical structure is also called hierarchical tree, which is a special type of "parent-child" connection. Children

represent low-level details or father's granularity. The columns in the dimension table represent a specific level in a hierarchy. Because there may be one or more relationships between multiple members in a dimension table, a dimension table can correspond to multiple hierarchical structures. There are several types of hierarchical structures: Normal hierarchy, skip hierarchy, ragged hierarchy, ragged with skip hierarchy. In addition, according to whether the data in the dimension of hierarchy can be summarized, it can be divided into aggregation hierarchy and non aggregation hierarchy.

Normal hierarchy: For example, there are three levels of country, province, and city in the regional dimension. The level of their composition is "country, one province, one city". Without considering municipalities, this level structure is a neat and balanced level structure.

Uneven hierarchy: It means that the process of traversing from high level to the lowest level of dimension will cross one or more levels. In the above-mentioned regional dimension, if the municipality directly under the central government is added, because there is no province at the upper level of Shanghai directly under the central government, but it is directly subordinate to the whole country, then the hierarchical structure will become uneven.

Unbalanced hierarchy: It means that there will be no data of the lowest level in the process of traversing from the high level to the lowest level of dimension. That is, the branches of the hierarchy are reduced to different levels.

Unbalanced and uneven mixed hierarchies: It refers to the hierarchical structure with both spanning branches and unbalanced branches in the process of traversing from high level to the lowest level of dimension.

Aggregation hierarchy and non aggregation hierarchy: In the data warehouse, data aggregation can be defined through hierarchical structure. For example, in commodity sales, from the perspective of time dimension analysis, a hierarchical structure of "year, quarter, and month" composed of level month, quarter, and year can be established, so that monthly commodity sales can be aggregated to quarterly commodity sales, and quarterly commodity sales can be aggregated to annual commodity sales. The hierarchy with such additive facts is called aggregation hierarchy. However, there are some factual values in the data warehouse, which do not show the additivity of commodity sales, such as the GDP growth rate of a province. If the annual growth rate is added up, it is meaningless. However, if the average value is taken, a meaningful value will be obtained. Such a measure is also called semi additivity. These semi additive or non additive facts are embodied in the hierarchy to form a non aggregated hierarchy. For example, in the national level of one province and one city, the GDP growth rate of each province can not be aggregated by the GDP growth rate of the cities below it, and the GDP growth rate of the whole country can not be aggregated by the GDP growth rate of each province. The provincial GDP growth rate and the national GDP growth rate must be stored as facts in the fact table of the data warehouse.

Through the predefined hierarchical structure in the dimension, the function of data summary and drilling can be realized. For the non aggregation level, it mainly uses the data drilling function of the level to view and display the data through the up and down analysis of the level. The hierarchical modeling of data warehouse determines the description of metadata, the design and implementation of ad hoc query software,

and the implementation of data drilling and aggregation. Hierarchical modeling also determines that data warehouse modeling adopts star model and snowflake model. At the same time, data aggregation also has a great impact on hierarchical modeling, which will be discussed below.

First, data aggregation [3]. As data is the normal level of aggregation, star model or snowflake model can be directly used for modeling. For the abnormal level of data aggregation, it is necessary to make the uneven branches and unbalanced branches in the level neat and balanced respectively in the process of modeling. If there are multiple levels between the child members and the direct parent, one level at a time will be processed by the same method, so as to realize the leveling of the non leveling level. Similar to the non-uniform hierarchy processing method, the non-uniform hierarchy can also be balanced by using placeholders, but the processing order is opposite to the non-uniform hierarchy. In the non-uniform hierarchy, the parent member occupies the position of the child member and iterates to the lowest level of the hierarchy, that is, the leaf node is reduced to the lowest level of the hierarchy tree.

Second, the data is not aggregated, and the non aggregated hierarchy mainly uses the data drilling function of hierarchy to view and display the data through the analysis of roll up and drill down by hierarchy. Due to the non aggregation of data, the data of the column corresponding to dimension in fact table comes from all nodes in dimension hierarchy tree. Star model is difficult to realize this kind of hierarchical modeling, and snowflake model is generally used.

3 Dimensional Measurement and Query Condition Determination

Measurement: Columns in the database that can be used for aggregate calculation of sum, average and count.

Dimension: The angle from which users observe and analyze the data. For example, time, region.

The main contents include:

1) Find the data model by searching or in the model list. After users get the data model they want, they can view dimensions or measures.
2) Configure query criteria. According to the needs of enterprises or organizations, dimensions or measures in the model are selected.
3) Set filter conditions, such as time or some metric limits.
4) In order to browse query results more clearly, users can also sort dimensions and measures separately.
5) According to different query needs, users can select list or crosstab. A list is a linear table, and a crosstab is a report with groups in both row and column directions (Fig. 1).

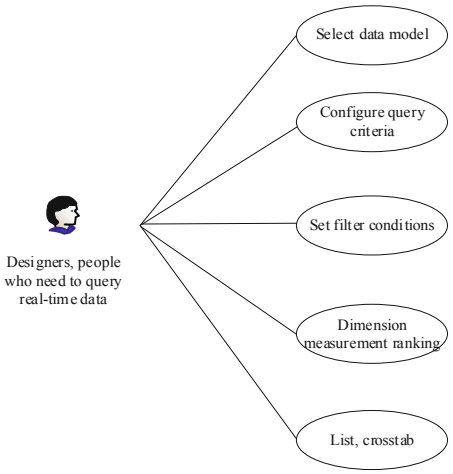

Fig. 1. Use case diagram of configuration table

4 Result Analysis Function Development

If the data is simply listed, it is impossible to find the key points and trends at a glance. This requires designers to use the idea of data visualization to show the data graphically and explore its in-depth value. The main functional steps involved are as follows:

1) Create a new calculation column [4] to calculate the function of the metric. Get the custom calculation result column.
2) Summarize and calculate the measurement. Merge multiple data into one cell, such as sum, average, count, etc.
3) Set an alert. According to some rules or opinions, highlight important data reports, prompt special situations, so as to avoid the ignorance of injury or lack of preparation conditions, so as to minimize the loss caused by danger.
4) For the data, the trend analysis chart can be drawn to facilitate observation and comparison.
5) Query results can generate graphics, which can help analysis through data visualization (Fig. 2).

Edit query data table function. As a complete and detailed report, in addition to accurate data and clear analysis results, designers also need to further improve the unit, font size, decimal point and so on, so as to complete a unambiguous and professional work result [5].

1) Designers can add units to the table data. One is to directly add currency symbols in front of the characters in the cell; the other is to add the unit name in front of the header.
2) The user can change the data alignment in the cell. Font, font size and other style settings.

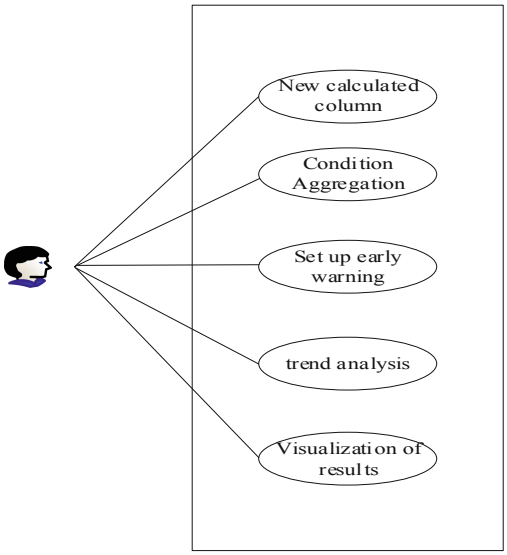

Fig. 2. Use case diagram of result analysis function

3) For the data in the table, the user can set the mantissa after the decimal point and the thousandth (Fig. 3).

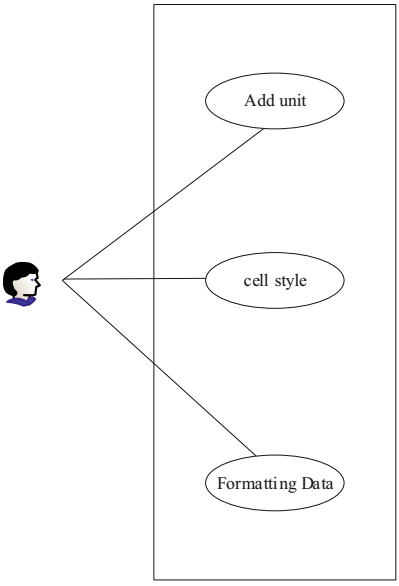

Fig. 3. Use case diagram for editing query data table

5 Construction of Distributed Multi Database Information Resources Linkage Update Model

Based on the completion of the above-mentioned basic work, a distributed multi-database information resource linkage update model is constructed. In order to better realize the linkage update of information resources, the fuzzy theory is used to process the incremental data packets in the space, extract the information characteristics collected by the sensors, and define the information collected by each sensor as a fuzzy input [6]. The fuzzy proposition judgment support degree of each information is given to obtain the spatial information optimization degree. The calculation formula for processing the observation information of the space sensor is:

$$J = \sum_{b} i \frac{g}{Tbv \times f} \tag{1}$$

In the formula (1), J represents the set of recognition sensors, $\sum_{b} i$ represents the rule inference coefficient in the data, and Tbv represents the fuzzy algorithm factor. This calculation does not perform directional analysis.

On the basis of the above calculation, the quality of information collected by different sensors in the database is evaluated, and the data with lower weights is eliminated. The calculation formula is:

$$E_b = \sqrt[w]{F} * \frac{p}{c \in v} \tag{2}$$

In formula (2), E_b represents the average of sensors in different spaces, $\sqrt[w]{F}$ represents the similarity matrix composed of sensors, and $c \in v$ represents the weight judgment factor. This calculation does not perform directional analysis.

Through the above calculation, the collected information features [7] are extracted, and the collected data is defined as the data volume of fuzzy proposition. On this basis, a distributed multi database information resource linkage update model is constructed. In order not to affect the efficiency of data update during work, the linkage update model is set to start at night, the incremental information extraction function is started within the planned time, the specified incremental information file is extracted, and the spatial data exchange format is automatically generated after extraction [8]. On this basis, the distance algorithm is used to measure the data in the distributed multi database, and the data in the database is updated by linkage. The update algorithm is introduced to synchronize the relevant data in the database to the database, and automatically update the new data, so as to complete the linkage update of distributed multi database information resources. The specific calculation is as follows:

$$K = \frac{B}{Q} \times G \times \sqrt{P_d} \tag{3}$$

In formula (3), K represents the unit value of the data, $\frac{B}{Q}$ represents the difference factor between the data, and $\sqrt{P_d}$ represents the linkage update factor. Directional analysis is not performed in this calculation.

Through the above calculation, the incremental data is imported into the database to complete the linkage update of distributed multi database information resources.

6 Query Big Data Clustering

On the basis of the above data update, the main steps of clustering the query data and information clustering are as follows:

Step 1: Users input query keywords, which is the first step in the operation of the query system in this study. Users input the keyword information of the title of the book, and the keyword information can be the title of the book, the author's name, the subject's name, the publishing house and so on [9];

Step 2: For keyword segmentation and query processing, because some users may not remember the complete information of the bibliography, so the association rules are defined for the bibliography. The keyword equivalence is defined as the smaller keyword which is most suitable for users' needs to query, so as to return multiple query results. The expression of association rules is as follows:

$$q = Q(m + s_j)/v \tag{4}$$

In formula (4), s_j represents the query record of the $_j$-th book, m represents the book query association rule, v represents the query support, and Q represents the frequent item set of the query information.

The recommendation process of association rules is as follows (Fig. 4):

Step 3: Content clustering is used to cluster the bibliographic information returned from the previous query. Firstly, the clustering center is initialized, and the clustering algorithm is used to determine the initial clustering center. Assuming that there are n objects, n objects are divided into different clusters. The purpose is to cluster the objects with high content similarity into the same cluster as much as possible. In the object set, the mean value of all objects in each query content cluster is taken as the new cluster center. The content clustering expression is as follows:

$$M = \frac{1}{n}\sum_{i=1} s/c \tag{5}$$

In formula (5), s represents the distance between the query object and the query object, and c represents the number of cluster objects.

Then cluster the bibliographic information and calculate the edit distance to optimize the clustering result. The expression is as follows:

$$D = \frac{r_k}{1 + d/a} \tag{6}$$

In formula (6), r_k represents the k-th attribute feature value, d represents the attribute weight of the query information, and a represents the comprehensive similarity of the query information of the two books.

Step 4: Sort the bibliographic information, sort the relevance of the query content according to the results of the content clustering, and view the title of the book closest to the cluster center, that is, the most relevant information the user queries, and it will be ranked at the top of the query results;

Step 5: The query result is displayed, the result of the query content is displayed to the user, and the data is accessed [10–12].

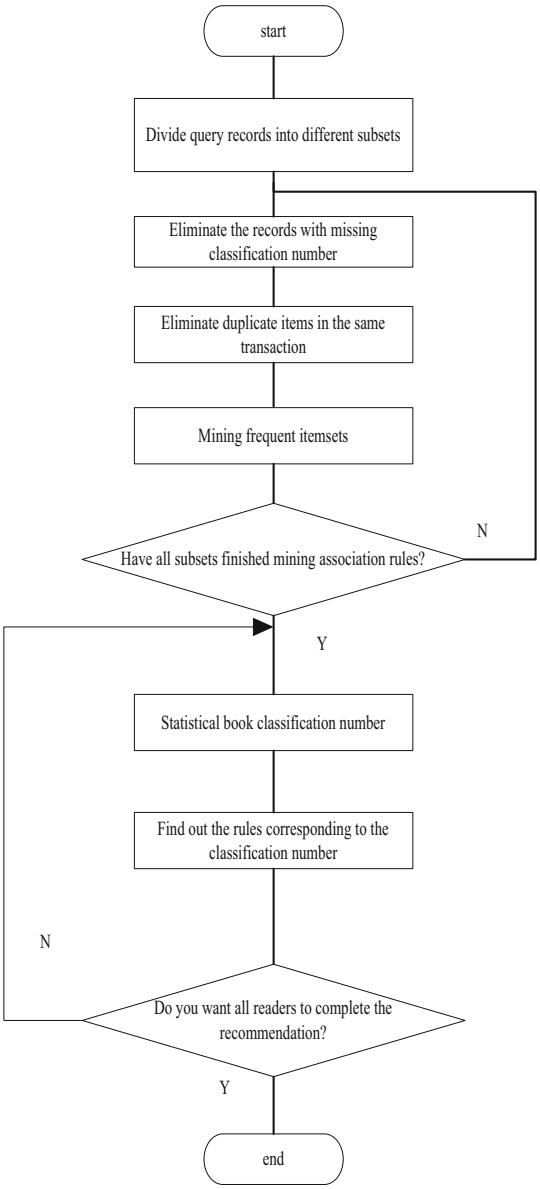

Fig. 4. Recommendation process of association rules

Step 6: Query data storage, OLAP is a kind of software technology that enables analysts, managers or executives to access information quickly, consistently and interactively from multiple perspectives, so as to obtain a deeper understanding of the data. The goal of OLAP is to meet the requirements of decision support or specific query and report in multi-dimensional environment. The core of OLAP technology is the concept

of "dimension". Dimension is the angle from which people observe the objective world, and it is a high-level classification. Dimension generally contains hierarchical relationship, which is sometimes quite complicated. By defining multiple important attributes of an entity as multiple dimensions, users can compare data on different dimensions. Therefore, OLAP is also a collection of multidimensional data analysis tools [13].

OLAP multidimensional data analysis operations, OLAP basic multidimensional analysis operations include roll up and drill down, slice, dice, pivot, drillcross, drill through and so on. Drilling is to change the level of dimension and the granularity of analysis. It includes roll up and drill down. Upward drilling is realized by climbing up the hierarchical structure of a dimension, that is, generalizing the low-level detail data to the high-level summary data in one dimension, or reducing the dimension by dimension specification [14]; On the contrary, drill down is implemented along the hierarchy of dimensions, that is, from summary data to detailed data for observation, or by introducing additional dimensions. Slicing selects one dimension of a given cube, resulting in a subcube. The slicing operation defines a subcube by performing a selection on two or more dimensions. Rotation is the direction of transforming dimensions, that is, reordering dimensions in a table.

On this basis, for data cube calculation, OLAP tools use data cube to model data and provide multi angle analysis. Data cube calculation is a basic task of OLAP. Full precomputation (full cube materialization) or partial precomputation (partial cube materialization) of data cube can greatly reduce the response time and improve the performance of OLAP [15]. But this kind of computing is a challenge, because it requires a lot of time and storage space. The core of multidimensional data analysis is to effectively calculate the aggregation of multiple dimensional sets. In SQL terms, these aggregations are called group by. Each group can be represented by a cube, and the set of groups forms the lattice of cubes that define the data cube. The figure below shows a cube with three dimensions (Fig. 5).

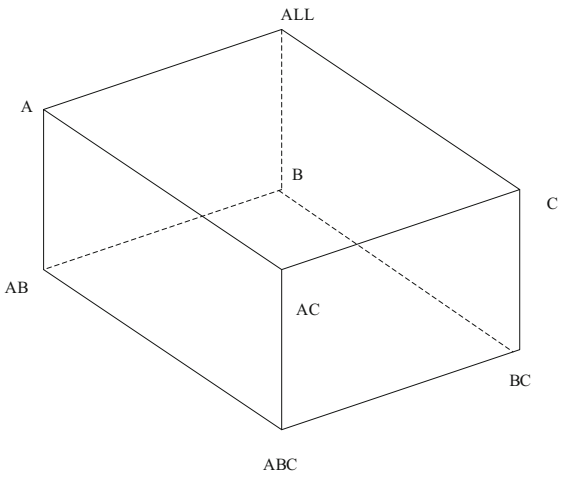

Fig. 5. Data cube

Based on the above process, it completes the big data ad hoc query and data storage.

7 Experiment

In order to verify the effectiveness of the big data ad hoc query technology based on the social network information cognitive model in this study, experiments are conducted to compare the query effects of the two methods.

7.1 Query Time Comparison

Comparing the data query time of the two methods, the comparison results are as follows:

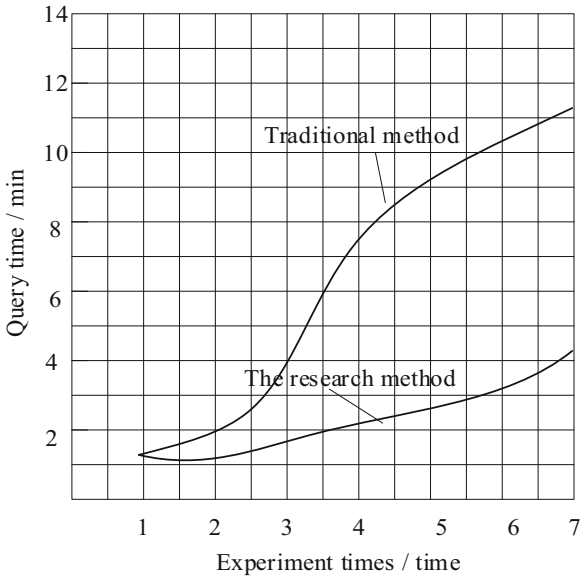

Fig. 6. Query time comparison

Through the analysis of Fig. 6, we can see that the query time of the two methods increases gradually in 7 iterations, but from the third experiment, the query time of the two methods is quite different. In the seventh experiment, the query time of the traditional method is about 11 min, in contrast, the query time of the research method is about 4 min, so we can see that the data extemporaneous query technology in this study has better application effect than the traditional query technology.

7.2 Comparison of Query Accuracy

Comparing the query accuracy of the two methods, the results are as follows:

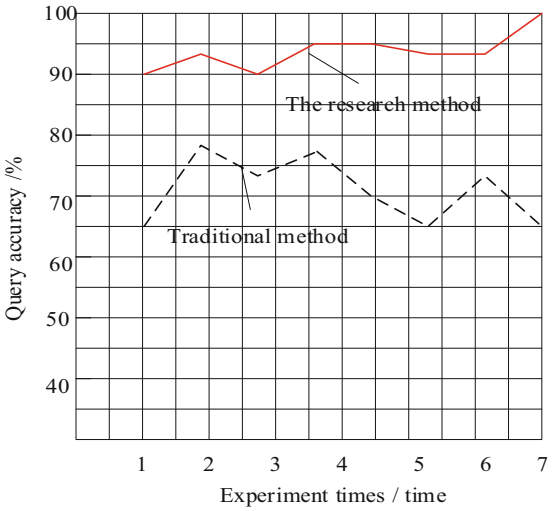

Fig. 7. Comparison of query accuracy

The results of the experiment in Fig. 7 show that the accuracy of the traditional method and the proposed method are stable throughout the experiment. In 7 iterative experiments, the range of the traditional methods is 65%–80%, and the range of the proposed methods is 90%–100%. Through the analysis of the figure above, we can find that the query technology in this study has higher accuracy and better application effect than traditional methods.

8 Conclusion and Prospect

This paper designs a big data ad hoc query technology based on social network information cognitive model, and verifies the effectiveness of the research method through experiments. Through this research, the query technology can effectively improve the efficiency and accuracy of ad hoc query, but due to the limitation of research time, the design method still has some shortcomings, and further research is needed in the follow-up research.

In the future, the research on unified query statistics of structured data and unstructured data will be further carried out, as well as predictive data mining and analysis research using historical and real-time data, so as to more fully mine the value of social network information big data and provide new technical means for the processing of network big data.

Fund Projects. Science and technology project of Jiangxi Provincial Education Department in 2018: Research on big data ad hoc query and analysis technology (Fund No. GJJ181029).

References

1. Su, X.G., Xue, J.M., Xuan, Z.Y.: Big data query process confidential data low-latency release protocol simulation. Comput. Simul. **36**(7), 363–366 (2019)

2. Qi, W., Bao, Y.B., Song, J.: Column-oriented store based sampling query process on big data. Comput. Sci. **046**(012), 13–19 (2019)
3. Liu, S., Bai, W., Zeng, N., et al.: A fast fractal based compression for MRI images. IEEE Access **7**, 62412–62420 (2019)
4. Liu, W., Zhang, T., Liu, J.: Window-based multiple continuous query algorithm for data streams. J. Supercomput. **75**(9), 5782–5807 (2019)
5. Fu, W., Liu, S., Srivastava, G.: Optimization of big data scheduling in social networks. Entropy **21**(9), 902 (2019)
6. Chen, I.C., Hsu, I.C.: Open Taiwan Government data recommendation platform using DBpedia and Semantic Web based on cloud computing. Int. J. Web Inf. Syst. **15**(2), 236–254 (2019)
7. Safari, L., Patrick, J.D.: An enhancement on Clinical Data Analytics Language (CliniDAL) by integration of free text concept search. J. Intell. Inf. Syst. **52**(1), 33–55 (2019)
8. Liu, S., Lu, M., Li, H., et al.: Prediction of gene expression patterns with generalized linear regression model. Front. Genet. **10**, 120 (2019)
9. Khan, A.W., Bangash, J.I., Ahmed, A., et al.: QDVGDD: Query-Driven Virtual Grid based Data Dissemination for wireless sensor networks using single mobile sink. Wirel. Netw. **25**(1), 241–253 (2019)
10. Watanabe, Y., Sato, K., Takada, H.: DynamicMap 2.0: a traffic data management platform leveraging clouds, edges and embedded systems. Int. J. Intell. Transp. Syst. Res. **18**(1), 77–89 (2020)
11. Cao, Y., Chen, S.W.: Extended query model for MOOC education resource metadata based on big data. Int. J. Contin. Eng. Educ. Life-Long Learn. **29**(4), 374–387 (2019)
12. Cisneros-Cabrera, S., Michailidou, A.V., Sampaio, S., et al.: Experimenting with big data computing for scaling data quality-aware query processing. Expert Syst. Appl. **178**(1), 114858 (2021)
13. Wang, H., Mu, L., et al.: Management and instant query of distributed oil and gas production dynamic data. Pet. Explor. Dev. **46**(05), 169–176 (2019)
14. Sam, T., Mauricio, S., Jonathan, B., et al.: Internet search query data improve forecasts of daily emergency department volume. J. Am. Med. Inform. Assoc. **12**, 12 (2019)
15. Jia, B., Meng, B., Zhang, W., et al.: Query rewriting and semantic annotation in semantic-based image retrieval under heterogeneous ontologies of big data. Traitement du Signal **37**(1), 101–105 (2020)