The Analysis of RDF Semantic Data Storage Optimization in Large Data Era

Dandan He, Lijuan Wang, Can Wang

Dalian Institute of science and technology Zip code: 116052

Abstract: With the continuous development of information technology and network technology in China, the Internet has also ushered in the era of large data. In order to obtain the effective acquisition of information in the era of large data, it is necessary to optimize the existing RDF semantic data storage and realize the effective query of various data. This paper discusses the storage optimization of RDF semantic data under large data.

1. Introduction

The computer wants to understand the data, first of all it needs to achieve a variety of data effective query by solving the semantic data description, semantic data storage and semantic data query. The Semantic Web, as an intelligent network which is capable of understanding human language, it can identify various data in the network and pass on useful information to people after analysis and processing. It can be said that the idea of the Semantic Web and large data requirements have a certain degree of consistency, and by means of the concept of semantic web it can also provide some ideas which are related to the settlement of large data.

2. the Semantic Data Analysis

Semantic Web concept is a great theoretical framework, and contains a series of theoretical methods, and semantic data are given the meaning and the relevant data. The semantic data in these semantic web can be used to describe the complete semantics of the data through the Resource Description Framework Standard (RDF), and it can directly show the correlation between the data. Therefore, the Semantic Web is the best description for this stage semantic data pattern. In the RDF statement data description process, generally contains three tuples, RDF / XML and RDF Graph these three forms of description.

2.1. Triples

In the description of the triplet, a RDF file is made up of multiple statements and it can be directly divided into three parts: subject, attribute, and object. Where the subject represents the resource being described, the attribute indicates the characteristics and the relationship of the resource, and the object mainly refers to the value of the attribute.

2.2. RDF / XML

RDF can be implemented by means of a scalable markup PRI, which is a markup language used to mark a file to make it structurally structured. It is a source language that allows the user to define their own markup language. Through the XML the data mark and the definition of data types and other work can be achieved.

2.3. RDF Chart

In general, the RDF triplet can be used as the edge of the label, and the subject and object as a node part, make its attributes as the edge. RDF data conforms to the graph model structure and it has a direction. The semantic level of the RDF data can be effectively maintained by means of the semantic level of the RDF model, and the semantic information query can be carried out [1]. The RDF diagram is shown in Figure 1.

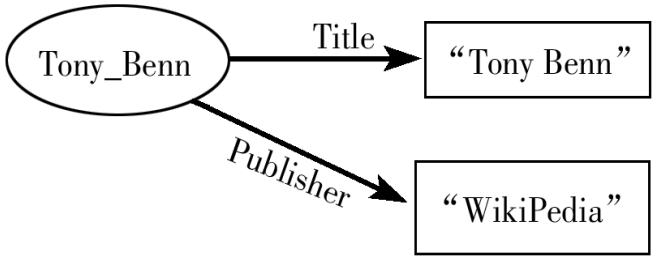

Figure 1: RDF representation

3. the Semantic Data Storage

The RDF resource description framework can solve the problem of the description of semantic data effectively. However, in the process of solving the RDF semantic data storage problem, it is necessary to realize that the storage problem is not simple to write data into the database and storage medium , but it will directly affect the entire data organization, read, maintain and query performance. And the difference of storage mode will directly affect the query efficiency and efficiency of RDF semantic data. In the current RDF semantic data storage process, it is mainly through the relational database, Big-Table, Key-Value, RDF and XML documents these five ways to carry out this article which mainly on the query performance optimization point of view Semantic data storage scheme was analyzed and studied.

3.1. Relational Database

Relational database as the most mature and most widely used database at this stage, in many studies we also try to directly stored the RDF semantic data in the relational database, to give full play to relational database existing application advantages. When the semantic data are stored through a relational database, it can be carried out in the following five forms.

① Triad table storage program: the storage program is mainly to build a ternary table, and then all the RDF triad are placed in the form. The program has a good versatility, but in the specific data query process there is still a problem with the performance of the query. ② Level storage program: The program is mainly to attributes all the as a column, and then form a large table. Its advantage is that the design is simple, and it can promote a single subject attribute value of the query work. The disadvantage is that the number of columns and the null values too much. ③ Attribute table: This is an optimized storage mode for horizontal storage mode, which can be divided into several sub-tables by means of the pattern of data attributes. But the query cannot take into account both the null value and query efficiency of these two aspects, once needs to query the number of tables involved, it will easily lead to a certain degree of decline in query efficiency. ④ Binary storage: The storage method is also an optimization model of the ternary table, which requires the triple table directly rewritten as n contains two columns of the table, and then the data above the table corresponding to the attributes as Table Name. The first column above all the properties on the attribute with the value of the main body, the second column is the subject above the property above the specific value. Binary storage This storage method has the advantage of small storage space and a null value, but it will directly increase the number of operands connected to the table, and directly affect the efficiency of data storage and query. ⑤ Full Traction Strategy: With the help of a full traction strategy, we can enumerate the possibility of all permutations and combinations in the hash table, and build the B +- tree on the basis

of this, which can compensate for the shortcomings of some simple vertical storage. Increase the storage space, so that by means of relational database it cannot fully meet the actual storage requirements of RDF semantic data.

3.2. Big-Table

Big-Table as a commonly storage program which is used in Google, the essence of a column storage. The data model is a sparse, distributed, persistent multi-dimensional sort mapping mode, the specific mapping of the index value is the line keywords, column keywords and time points. Each value in the concrete mapping process is an unparsed byte array. It can be said that the data model is a kind of map, and it can achieve the effective storage of web content through the table model, but it's difficult to meet the RDF semantic data storage needs.

3.3. Key-Value

The Key-Value data model mainly contains a specific Key value and a Value pointer, and directly point to a particular data. For example, Amazon's DYnamo storage platform is the use of Key-Value mode for data storage, and have a good scalability and application results. With the dynamic hash table model it can achieve the distribution of data storage and query work, but in the high readability on the basis of the data model this will lead to the loss of other performance, but can only take into the distributed and Large-scale data account. Key-Value data model can not be considered for the relevance of the data, and can only be distributed to the database to solve the problem, so that Key-Value is not suitable for RDF semantic data storage.

3.4. RDF File System

RDF file system is also known as the text storage, the basic idea is to text packaging data, and make the semantic Web base framework Jena as a tool for analysis. The document is a form of key-value pairs whose key ID can only be identified for a particular document and it has the advantage of being able to store large amounts of data and a model-free form. But the text storage mode can not directly reflect the relationship between the data, so that the RDF file system is not suitable for RDF semantic data storage.

3.5. RDF map storage

The above four kinds of database models are difficult to meet the RDF semantic data storage requirements, so many researchers are also beginning to consider the application form to RDF semantic data storage work, its application advantages are the following: ① by means of the map to describe RDF and it has a good intuition; ② to maximize the RDF data which is contained in the semantic information contained; ③ RDF model which can be a direct mapping, and in order to avoid the RDF data conversion link; ④ effectively avoid the RDF data reconstruction processing; ⑤ it can directly learn from some mature graph algorithm and graph database to RDF storage scheme and query algorithm design [2].

However, in the use of RDF map storage mode, it also has a large storage space and the complexity of the query algorithm is relatively high, and leads to the map storage model which can not be widely used [3]. In recent years, with the increasing capacity of storage media, making the unit capacity of the price which has become cheaper, which can directly ignore the need for storage space that is too large as the shortcoming, so the relevant researchers need to focus on research placed in the design of the algorithm above, and to reduce the data real-time query when the time complexity.

4. Neo4j Map Database

4.1.Neo4j Characteristics Analysis

Neo4j as a graphics database which uses Jave implementation, and fully compatible with the ACID, that can store the data through a graphical network for the optimal format directly to the disk above,

and it have the following three typical data characteristics: ① Data structure is not Is necessary, and it can change the pattern and delay the migration of data to play a certain role in simplification. ② For the common complex domain data sets, which can easily model, for example, CMS can be seen in the panic of the visit to some fine-grained access control table; ③ It can be based on analysis, social network to achieve data modeling work, and can be applied to a number of areas [4].

4.2. Neo4j System Implementation
After the installation of the Neo4j Server, you can use the geological: http: //127.0.0.0: 7474 / access, and it can monitor the operation of Neo4j Server, data browsing and maintenance, database interaction and browsing Yuan Data management object and many other functions.

The exclusive query language of the Neo4j graph database is the Cypher language, which serves as a graph to describe the query language, enabling efficient querying without traversing the graph structure, and enabling efficient updating of the graph storage. As an user-friendly query language, Cypher language can map database at any time query, and to be able to combine SQL and SparQL both query characteristics, the specific query language interface shown in Figure 2:

```
Neo4j-sh(?)S start n=node:node_auto_index(name= "Shana Willems" )
>match (m)-[r1:MOTHER_OF]->(n),(m)-[r2" MOTHER_OF]->(a)
>return n.name, r2, a.name;
==>+ -------------------------------------------------- +
==>| n.name          | r2                  | a.name    |
==>+ -------------------------------------------------- -+
-->| "Shana Willems" | :MOTHER_OF[19]{} | "Sharonda Peele" |
-->| "Shana Willems" | :MOTHER_OF[10]{} | "Melda Peza"     |
==>+ --------------------------------------------------+
==> 2 rows
==> 1 ms
```

Figure 2: Cypher query language interface

4.3 Query Advantage Analysis
In all databases , if you want to query the data work , you need to use the index which is applied to carry out , but the establishment of the index and update the work often needs to consume a lot of time . In the relational database through the index it can be a table in the corresponding record data to quickly find , but need to connect two tables in the process , you need to build a new and larger index which is based on all the data in the two tables . If you need to connect multiple tables , but also need to follow all the data onto these tables to carry out the refactoring of the index , as a fully recursive process , and need to spend a lot of resources . Semantic data has a very strong relevance , so in any meaningful data query process needs to be applied to the table connection , which also lead to the relational database of this storage mode query efficiency is low , and easy in the query request Fast-intensive circumstances led to the database because of the depletion of resources and the collapse of the situation [5] .

With the Neo4j graphics library for data query , it only needs to start the query process in the process of the index , you can on this basis of the starting point to traverse along the edge , and do not need to use the index case Can find the relevant data . So on the map database only needs to carry out

one or several effective index maintenance works , and the need for semantic data changes in the process of updating the index work , it can achieve good data storage and query results , and can save Large amounts to index creation and maintenance overhead . When the amount of semantic data appears to increase this situation , it will lead to a certain degree of increase in the index , the speed of its starting point will find a drop . But because the traversal of the network will not slow down , in the follow-up data query process will not slow down or the collapse of the database , etc . [6] Therefore , in the application of Neo4j RDF data query process , it also has a fast query which is not allowed to connect the cost of the table and overhead by the index update that is relatively small and stable performance and also have many other advantages, and play a certain role in optimizing to the existing RDF semantics Data storage technology.

5. Conclusion:

How to obtain useful data in massive data in large data environment is also one of the key issues that needs further study. In order to solve this problem, we need to start with the description, storage and query of semantic data, and to further improve the acquisition rate and accuracy of useful information. In this study, we compared the five main database models and found that the graph database is more suitable for the storage of semantic PDFs. The Neo4j is used to design the graphics database and the query is made using the Cypher query language to work, it will be able to achieve get the rapid access to useful goals for this purpose in a certain degree.

References:

[1] He Xiangwu. RDF semantic data storage optimization in large data [J]. Journal of Computer Applications and Software, 2015, (4): 38-41.
[2] Wang Yajun. RDF semantic data storage coding and query optimization [D] .Xi'an University of Electronic Science and Technology, 2015.
[3] Zhu Min. RDF data storage and query based on HBase [D]. Nanjing University, 2013.
[4] Gu Rong, Qiu Hongjian, Yang Wenjia, Hu Wei, Yuan Chunfeng, Huang Yihua. Goldfish: Large-scale RDF Data Storage and Query System Based on Matrix Decomposition [J / OL]. Journal of Computer Science, 2017,: 1-19.
[5] Yang Jian, Luo Jun. Overview of RDF Data Storage Strategy Based on Hadoop [J]. Information Security & Technology, 2015,6 (05): 46-48.
[6] Wang Linbin, Li Jianhui, Shen Zhihong. Overview of NoSQL-based RDF Data Storage and Query Technology [J / OL]. Journal of Computer Applications, 2015,32 (05): 1281-1286.