Research on the Automatic Pattern Abstraction and Recognition Methodology for Large-scale Database System based on Natural Language Processing

Rong Wang[1]
[1] Hubei University of Science and Technology, Xianning, Hubei Province,437100 China

Cuizhen Jiao[1, *]
[1] Hubei University of Science and Technology, Xianning, Hubei Province,437100 China
*Corresponding Author：Cuizhen Jiao

Wenhua Dai[1]
[1] Hubei University of Science and Technology, Xianning, Hubei Province,437100 China

Abstract. In this research paper, we research on the automatic pattern abstraction and recognition method for large-scale database system based on natural language processing. In distributed database, through the network connection between nodes, data across different nodes and even regional distribution are well recognized. In order to reduce data redundancy and model design of the database will usually contain a lot of forms we combine the NLP theory to optimize the traditional method. The experimental analysis and simulation proves the correctness of our method.

Keywords: Pattern Abstraction and Recognition; Database System; Natural Language Processing.

Introduction

In distributed database system is usually based on data fragmentation and distribution strategy, the related distribution data is stored on multiple nodes. Each other through the network connection between nodes, when a database receives data request, the request will can send it to the corresponding node for processing, reducing the load of a single node. In addition, in the distributed database, in order to avoid the loss of data due to node failure or unable to access, usually for creating multiple copies of data are stored in different nodes. In this way, even if there is a node fails, will not affect the normal use of the database, and greatly improve the reliability of the database. A distributed database is to solve the data storage bottleneck of centralized database, and improve the reliability of the database. But in practice, it still has some problems. In distributed database, through the network connection between nodes, data across different nodes and even regional distribution are well recognized. If the application exists in the join operation, and involves all the table data are stored in different nodes, then execute this application available across nodes will cause data interaction. Traditional software testing is often overlooked to ensure the quality of the database system, and to test the database system is the database of an important means of quality assurance. A complex database at the time of design and implementation, often will have some changes and mistakes, in order to discover and to avoid these errors, we put forward a test method of the mode of the database.

The database schema matching algorithm based on fuzzy matching is based on the database schema conflict concept inspired by the fuzzy match. The method is put forward in the prophase research institute method on the basis of the improved. Database schema fuzzy matching algorithm ER model and logical model of the elements of information, data dictionary information extracted, the element information contain entity here, contact, attribute, the logical model is refers to in the table and table attribute information and then use data dictionary

information to generate a matcher, and by using the algorithm of ER model set and logic model elements are classified and calculate in each category, "entity relationship model" elements of the semantic matching degree as well as to the "entity relationship model" to each pair of the attributes of the candidate matching structure similarity calculation. In practice, data design usually follow paradigm. However, the strict paradigm design there is no guarantee that the database in any case can have the best performance. In order to reduce data redundancy, and model design of the database will usually contain a lot of forms, therefore, the query might need to connect more need all of the data acquired after table and too much join operation is the main factor database performance decline, especially in the case of large amount of data, a greater impact. Therefore, by reducing the number of tables, at the same time increase the data redundancy, decrease as far as the possible in the query join operation, thus improve the database performance, this is known as the paradigm of the database schema design. Especially in the business of the data warehouse, because of the relatively large amount of data, query request is also more complex, the pattern design is a good way to improve database performance.

Natural language database allows users to use natural language for the contents of the database of various operation requirements, and then by the system automatically converts it to the operation of the database language, thus in the database query to the right information, provide to the user. Pattern matching is one of the most simple natural language analysis technologies and it is the input of language as a whole. Mode variable can match any words in sequence and the interpretation of the input of language expression is through the word model with the input expression of matching. Patterns and interpretation are the concept of recursion, associated with each pattern is a kind of explanation, the interpretation can be used to construct a higher level of interpretation which is also can explain directly as an output.

In this paper, we conduct research on the automatic pattern abstraction and recognition method for large-scale database system based on natural language processing. Depending on the business query requires fields, we will form to split into multiple child table. By avoiding table joins redundant data, improve the query efficiency. Unlike the connection method, the method of redundant data is in the logical design, will need to connect more than originally form to obtain the properties of the stored in a table. Each business need to access their child table only can obtain the required data, reducing the need to search the amount of data. Split table structure generally includes vertical separation and horizontal split, vertical resolution is the attribute of the source table separated according to different business needs, in the different new tables, every new table redundancy primary key attribute of the original table, to ensure the record uniqueness. All attributes and horizontal resolution is retained the source table, split according to the scope of the primary key, and break up after each data to save people in the new table, the new table structure in line with the source table. In the following sections, we will discuss the algorithm in detail with theoretical analysis.

The Proposed Methodology

The Principles of Natural Language Processing. From computer science, especially from the point of view of artificial intelligence, natural language understanding of the task is to build a computer model and the computer model is able to like people to understand the result of natural language. To make the machine understanding of human natural language, on the basis of long-term research formed the two basic methods: the method based on rules and based on the statistical method. The method based on rules is deductive in nature, based on statistical method is essentially inductive, two methods are lack of

more flexible and effective association and variable, not by adding knowledge to get better results. Actually, itself is a kind of knowledge representation, natural language of the extension model of natural language form as the formalization description of problems and extension transformation and extension reasoning process and methods to solve the problem of contradiction is applied in among them, the extension of natural language model base established, again through the extension of the model transformation for the knowledge representation and the improvement of machine intelligence has the very vital significance [1-2].

Natural language extension spatial visualization description: knowledge expressed by the sentence or a paragraph of text information, can be represented by establishing extension model, this description includes both the literal language information, include by the extension reasoning model of extension and transformation of knowledge useful information. Natural language is a hierarchical structure, generally can be divided into lexical analysis, syntactic analysis and semantic analysis and so on three levels. Influence each other and restrict each other between these levels, and eventually as a whole to solve the problem of natural language processing. From the perspective of the specific composition of natural language, a sentence consists of morphemes, words, phrases of which each level are constrained by the rules of grammar, and the realization of the hierarchy is directly reflected in the composition of the natural language sentence. As a result, the computer to deal with natural language should also be a hierarchical process. And, according to the rules of language structure, in the implementation process of natural language communication between man and computer, the computer in addition to the need to understand the given natural language text and also must be able to express in the form of natural language text processing results [3-5].

The Large-scale Database System Pattern. Paradigm design of database, can be reasonably organize data, makes the database each table has the very high between independence, greatly reduce the data redundancy. In the operations such as update, delete, owing to the high data independence, users often only need a list of data make a corresponding operation, reduce the abnormal because of the various operations caused by data redundancy. The system structure is shown in the figure one.

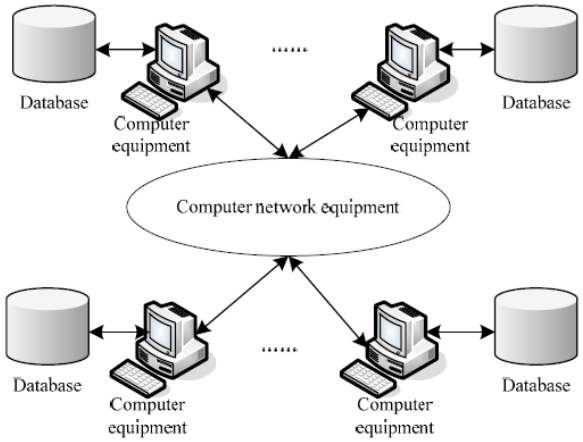

Figure 1. The System Structure of the Large-scale Database System

The relational database system for keyword query mainly has five parts, the first to analyze the input keywords, there are several key words; Then calls the full-text index, to examine the keyword belongs to, is the name of the table, attribute names or attribute value; Next query the database model, obtained several possible tree tuples connection. Finally connect the corresponding tuple tree into the relational database SQL queries, generate the query results, displayed in a two-dimensional tabular form. Keywords retrieval technology mainly, through the analysis of the key word belongs to type of user input to determine the tuple tree, which is converted into the corresponding SQL statement to query the relational database. If the user input the key word is the name of the table, will be a few natural connection table can be output after; If the user to enter the keywords that have the name of the table, attribute, then the attribute columns to the table in the output is the content

of the user's retrieval. User input keywords in the attribute value, the attribute value corresponding with the table or attribute column connection, according to the property value corresponding to a tuple to display the query results. Thus, for the same keyword, if it is more than one belongs to value, it will correspond to different SQL statements. In a relational database, the key word is connected through the main foreign key, therefore the relational database's data model, which is based on the modeling of patterning.

In order to reduce the query complexity, improve the efficiency of query, OLAP operations, the paradigm in database model design phase design is a kind of important method. The so-called design paradigm, it is to point to by reducing the traditional data standardization degree, in order to improve query performance process. Through the paradigm can significantly reduce the number of tables, and to reduce the number of join query request. At present, there have been many design method of application and research about the paradigm. Tupper door put forward two methods, respectively in the different stages of database design considering the paradigm, one way is in the design phase, just as much as possible to reduce the number of logical entity in view of the specific application, the method of using the paradigm increase the attribute table or table, fetching data redundancy and through the trigger to ensure redundancy data update consistency. On to adapt to changes in the data of the paradigm design, designers need to understand the business scenario, fully considering the data update frequently, if the data is updated frequently, so for these data will need to be careful with the paradigm. Although the pattern design has faults, but reasonable design can reduce the negative influence of shortcomings give play to the advantages of the paradigm can improve the performance.

The Numerical Analysis of the Method. Diagram will be generated in this paper is converted to a query, by executing the corresponding query and then get each relationship path corresponding to the query results. Because the diagram is carried out in accordance with the correlation between the returned, but this correlation is only the keyword mapping in the relationship between the levels and in order to make the query results more clear, in this paper, the results will refine, map the keywords to the relational level. In the following figure two, we simulate the method with result.

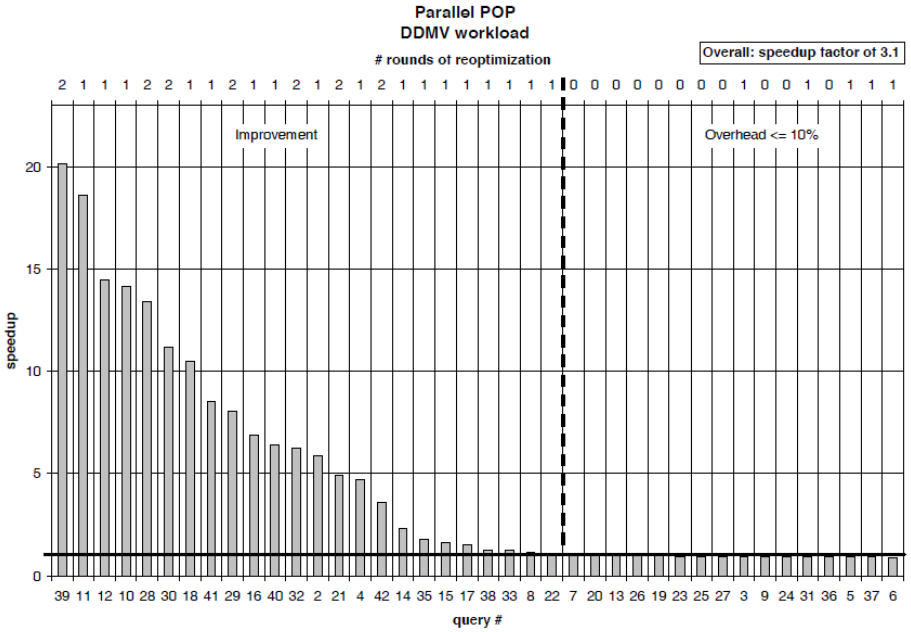

Figure 2. The Numerical Simulation of the Proposed Methodology

Conclusion

In this research paper, we research on the automatic pattern abstraction and recognition method for large-scale database system based on natural language processing. Natural language extension spatial visualization description: knowledge expressed by the sentence or a paragraph of text information which can be represented by establishing extension model. At present, there have been many design method of application and research about the paradigm. Our approach combines the NLP method with the existing method. The numerical simulation illustrates the effectiveness and feasibility of the method. Furthermore, we will conduct more literature review work in the near future.

Acknowledgement

（1）National Natural Science Foundation Project，61373108；
（2）Natural Science Foundation of Hubei Province,2015CFC778；
（3）Science and Technology Research Project of Department of Education of Hubei Province, 20082803；
（4）Humanities and Social Sciences Research Project of Department of Education of Hubei Province，2012D125。

References

[1] Simonovic S P, Li L. Methodology For Assessment Of Climate Change Impacts On Large-Scale Flood Protection System[J]. American Society of Civil Engineers, 2003, 129(5):361-371.

[2] Ma X, Eatherton M, Hajjar J, et al. Seismic Design and Behavior of Steel Frames with Controlled Rocking—Part II: Large Scale Shake Table Testing and System Collapse Analysis[C].

[3] Kyriakakis P, Chatzigeorgiou A. Maintenance Patterns of Large-Scale PHP Web Applications[C]// Software Maintenance and Evolution (ICSME), 2014 IEEE International Conference on. IEEE, 2014:381-390.

[4] Oubeidillah A A, Kao S C, Ashfaq M, et al. A large-scale, high-resolution hydrological model parameter data set for climate change impact assessment for the conterminous US[J].

[5] Weishar L L, Keon T, Stauble D K. Effects of Large Scale Morphological Changes to a Back-Bay System[J]. American Society of Civil Engineers, 2014.