H K . N C C P
International Journal of Intelligent Information and Management Science
ISSN: 2307-0692 *Volume 3, Issue 6, December 2014*

Keywords Retrieval in Relational Databases Based on Index Structure

Yan ZHAN, Hao CHEN
College of Mathematics and Computer Science, Hebei University, Baoding, CHINA

Abstract: At present, the keyword search based on relational database has some representative research results. This paper puts forward a query mapping index method different of the full-text index method, which creates a mapping table, and match index table and the corresponding key words. So it can improve the query speed. In the query mapping index method, setting as the threshold, it can use PSO algorithm to learn keywords number. According to the number of occurrences to each of the query keywords, it can determine the new keywords. So it can retrieve the log to mining, and achieve the goal of retrieving through the query the user's usage and form.

Keywords: relational database; keyword search; query mapping table; index structure; PSO

1. Introduction

Keyword search is the most popular information discovery method because the user does not need to know either a query language or the underlying structure of the data. The search engines available today provide keyword search on top of sets of documents. When a set of keywords is provided by the user, the search engine returns all documents that are associated with these keywords. Typically, two keywords and a document are associated when the keywords are contained in the document and their degree of associativity is often their distance from each other. In addition to documents, a huge amount of information is stored in relational databases, but information discovery on relational databases is not well supported. The user of a relational database needs to know the schema of the database, SQL or some QBE-like interface, and the roles of the various entities and terms used in the query. The user of DISCOVER does not need knowledge of any of the above. Instead, DISCOVER enables information discovery by providing a straightforward keyword search interface to the database.
At present, the keyword search based on relational database has some representative research results, such as DBXplorer [1], DISCOVER [2] and BANKS [3], and so on. The basic research on these methods to solve the problem is the same core idea, namely, these methods are based on the concept of graph theory and the reduced subtree. Because the nodes and edges in the graph associated text content, therefore, graph was a better representation of a relational database, semi-structured data and the Web data. The kernel problem of keyword search is how to efficiently find best result tree from the data graph.
Keyword search has been well studied for document databases ([4]). For example [5] presents the Google search engine. [6] offers an overview of current Web search engine design. It also introduces a generic search engine architecture and covers crawling and indexing issues. In [7], algorithms, data structures, and software are presented that approach the speed of keyword-based document search engines for queries on structural databases like parse trees, molecular diagrams and XML documents. [8] tackles the keyword search problem in XML databases. They propose an extension to XML query languages that enables keyword search at the granularity of XML elements, which helps novice users formulate queries, but do not consider keyword proximity search. [9] proposed a new approach to translating a practical lass of xpath queries over (possibly recursive) dtds to sql queries with a simple lfp operator found in many commercial rdbms. It introduces the SQL Tuning Advisor to help you tune SQL statements. It can then use the recommendations of this advisor to rewrite poorly performing SQL code.
Much of the keyword search methods in the relational databases are on the existing static database. For example, the database is a snapshot at some time in the method based on the data graph, which will not have updates. However, in practical applications, the database will be updated frequently, and users for a particular theme will show the lasting interest. [10] studied how to use multiple tuples units to answer queries, and designed a new index results to quickly find the query related tuples. It also adopted a new sorting technology and algorithm, realized the gradual formation of top-k query results. [11] puts forward a cascade top-k keyword query algorithm for the larger data graph which couldn't store the memo-

R. B. G. thanks the youth natural science Research Plan foundation of Hebei University(Grant no. 2010Q023) and the Soft Science Research Program of Hebei Prov-ince (Grant no. 14450318D)

ry. It generated "super node" needing less memory space which was different with other existing methods, so it could save the query response time. In [12] an efficient method is proposed to solve continuous top-k keyword query in relational databases. It integrated the grading sorting mechanism based on the existing relational data flow oriented keyword query mechanism. It could quickly calculate the static top-k results in the database, and when database were updated, it could update query results.

2. Query Mapping Index structure

2.1. Full-text retrieval

With the development of science and technology and economic, the growth of electronic data in the libraries, press and publishing, the rapid development of the Internet, for people to choose information is rapidly expanding, traditional way of manual retrieval is increasingly difficult to meet the needs of the development. For retrieving full-text retrieval system powerful, easy to operate and more and more get the welcome of the masses of users. Foreign full-text retrieval software although have been applied earlier, but there are not applicable for Chinese users. Chinese full-text retrieval technology is consistent with es on the principle of full text retrieval, but the characteristics of Chinese characters itself makes the realization of the Chinese system is more complex than es system. The core of the technology of the full-text retrieval is how to index all the basic elements in the source document information record into the library. In the Chinese system, basic elements can be a single Chinese character, also can be a word. Therefore, there are two basic index structure, namely the library based on word table index and the index which is based on word library. Word table method, the source document records the position of every word to the index in the library. The library index for each different character is stored on a word table, and records all positions in the same word in the document. In corresponding with the above method, it using the a certain meaning word in the indexing library as the basic independent unit.

The integration of database system and full-text retrieval system implemented database-oriented full-text retrieval, played their respective advantages. Its application field is very broad, such as e-commerce, e-government, enterprise management information system, a business letter processing and so on. Its advantage is not only the recall ratio and high precision, fast response, and because the hypertext and multimedia information is stored in the database, security is a reliable guarantee.

Full-text retrieval views the text data as the main object, and achieves information retrieval according to the content of the data. "Wenheim needle" is image description of full text retrieval. From a technical perspective, the full-text search is main technology basis on finding information, filtering information analysis, information agent and information security control. At present, the research of Chinese automatic word segmentation technology has made great achievements, and the Chinese whole word retrieval ability had the very big enhancement. Especially with the combination of artificial intelligence technology such as knowledge base and reasoning machine, using the machine learning techniques, full-text retrieval has a powerful association reasoning function, and improve the efficiency of automatic retrieval, at the same time support logical expression and fuzzy retrieval in Chinese and in English.

2.2. Query Mapping Index

This paper illustrates the design method based on Chinese DBLP data set. In the dataset there are four tables: Paper, Scholar, Institution and ScholarPaper. As shown in Table 1 to Table 4, respectively, the field names in the table with "*" is the table's primary key field or fields.

Table 1. The structure of Paper table

Field name	Field meaning
*PaperId	Serial number of paper
Title	Chinese title
KeyWords	Chinese keywords
AuthorString	authors
WorkPlaceString	affiliation
Magzine	journal
PublishTime	Published time

Table 2. The structure of Scholar table

Field name	Field meaning
*ScholarId	Serial number of author
ScholarName	author name
ScholarDescription	Description of scholar

Table 3 The structure of Institution table

Field name	Field meaning
*InstitutionId	Serial number of institution
InstitutionName	Institution name
InstitutionDescription	Description of instituion

Table 4 The structure of ScholarPaper table

Field name	Field meaning
*PaperId	Serial number of paper
*ScholarId	Serial number of author
*InstitutionId	Serial number of institution
AppearOrder	The rankings of authors

Table 5 Query Mapping Index table（Key_Forkey）

Table name	Foreign key	Key of referenced table	referenced table
Paper			
Scholar			

Institution			
ScholarPa-per	PaperNum	PaperId	Paper
	ScholarNum	ScholarId	Scholar
	Institution-Num	InstitutionId	Institution

This paper puts forward a query mapping index method different of the full-text index method, which creates a mapping table, and match index table and the corresponding key words. So it can improve the query speed. Table 5 is the query mapping index table. ScholarPaper table's primary key consists of three attributes, i.e., PaperId, ScholarId and InstitutionId. At the same time, these three attributes are respectively foreign keys of ScholarPaper, that is , with reference to the primary key of the Paper table PaperId, Scholar table's primary key ScholarId and Institution table's primary key InstitutionId.

3. The number learning of occurrences with keywords

By setting keywords number as the threshold, it can set learning algorithm according to the number of occurrences to each of the query keywords, which can determine the new keywords. So it can retrieve the log to mining, and achieve the goal of retrieving through the query the user's usage and form.

3.1. Particle Swarm Optimization algorithm

Particle Swarm Optimization (Particle Swarm Optimization, PSO) algorithm is a kind of evolutionary computation methods which is put forward by American social psychologist J.Kennedy and electrical engineer R.E berhart [13] in 1995. PSO algorithm is originated in artificial life research, especially the behavior mechanism imitation of birds, fish and other groups. It referenced the biological group model by biologists F.Heppner, and at the same time integrated the idea of evolutionary computation.

As a bionic algorithm, there is no perfect mathematical theory in PSO algorithm, but as a new optimization algorithm it has showed a good application prospect in many fields. So this paper using the PSO algorithm to learn the number of occurrences to each of the query keywords.

In 1995, [13, 14] simulated the migration and the cluster in the process of the birds' flock foraging: When the birds were foraging, from one place to another in the process of migration, always there was one bird having good insight on the general direction of the food source. At the same time, when looking for the feed source, they conveyed information to each other at any time through a set of their unique way, especially the good information. Under the guidance of "good news", the flock "swarm" flew towards the food source, which achieved the food source cluster. Particle swarm optimization algorithm is used to gain enlightenment and is used to solve optimization problem. In particle swarm optimization algorithm, a bird is called a "particle", XieQun is equivalent to a flock, migration from one place to another is equivalent to the evolution of XieQun, the "good news" is the optimal solution in each generation evolution, food source is equivalent to the global optimal solution.

Assumption in a N d target search space, m particles are a community, of which the ith position of the particle in N dimensional space $Xi=(x_{i1},x_{i2},\cdots,x_{iN})^T$, $i= 1, 2,...m$; The speed $V_i=(v_{i1},v_{i2},\cdots,v_{iN})^T, i=1,2,\ldots,m$; Fitness value $fitness_i=f(X_i)$, and P_{besti} 和 $X_i^{Pbest}=(x_{i1}^{Pbest},x_{i2}^{Pbest},\cdots,x_{iN}^{Pbest})^T$, respectively, in order to adapt the *ith* a particle has the biggest value and its corresponding position. *gbest* is the best position for all particles in group, and its index number of is *g*. For every generation, its dth dimension will change according to the following equation:

$$v_{id} = wv_{id} + c_1r_1(x_{id}^{pbest} - x_{id}) + c_2r_2(x_{gd}^{gbest} - x_{id}) \quad (1)$$

$$x_{id} = x_{id} + v_{id} \quad (2)$$

v_{id} in (1) is the dth flight velocity vector component for a particle *i*; x_{id} is the dth position vector component for a particle *i*; r_1,r_2 are as a random number between [0, 1]; c_1,c_2 is the acceleration coefficient; *w* is for inertia weight. In (1), the first is the speed of the particles previously; $c_1r_1(x_{id}^{pbest} - x_{id})$ is as "Cognitive Term", which is as the associated with the cognitive experience of particles; $c_2r_2(x_{gd}^{gbest} - x_{id})$ is as "Social Term" , which stands for information sharing and cooperation between the particles. Equation (2) is new coordinates for particle *i* . They decide the position of particle *i* at next movement.

3.2. Keywords number learning with PSO

In the query mapping index method, setting as the threshold, it can use PSO algorithm to learn keywords number. According to the number of occurrences to each of the query keywords, it can determine the new keywords. So it can retrieve the log to mining, and achieve the goal of retrieving through the query the user's usage and form. The value of Keywords is the particle swarm algorithm to optimize object, encoding the coding of keywords value.

Keywords value parameter study is based on the idea of traditional particle swarm optimization algorithm, and it will modify appropriate for the parameter optimization problem. The specific algorithm is as follows:

- Initialize the particle swarm. A group of size m, within the scope of the allowed random set the initial position and velocity of particles, set the inertia weight;
- Evaluate each particle's fitness, namely compute the objective function value $fitness_i$ of the each particle;

- For all $i \in \{1, 2, \cdots, m\}$, if $fitness_i > Pbest_i$, then has made $fitness_i = Pbest_i, Xi^{Pbest} = X_i$, if $fitnessi > gbest$, then resetting the index number g of $gbest$;
- According to (1) and (2) to adjust each particle's position and speed;
- Check the termination conditions. If gets the maximum number of iterations gen_{max} or stagnation of best solution no longer changes, terminate the iteration, otherwise return step 2.

The keywords value parameter learning based on traditional particle swarm optimization algorithm mainly includes: (a) Particle flow in the whole problem space in the form of random and to evaluate their environment (calculate fitness). (b) Each particle can remember the best location itself and perception of neighboring particles have reached the best position. (c) At the time of changing speed it can consider the best location has reached and the best position of a nearby particle.

4. Experimental results

This paper does the experiments to illustrate the design method based on Chinese DBLP data set. In the dataset there are four tables: Paper, Scholar, Institution and ScholarPaper, which is shown in table 1 to table 4 respectively in the session 2. With tuples as the following Table 6:

Table 6 Tuples number of the four tables

Table name	Field name	Tuples
Paper	PaperId, Title, KeyWords, AuthorString, WorkPlaceString, Magzine, PublishTime	39484
Scholar	ScholarId, ScholarName, ScholarDescription	25693
Institution	InstitutionId, InstitutionName, InstitutionDescription	2362
ScholarPaper	PaperId, ScholarId ,InstitutionId, AppearOrder	92865

To evaluate the effectiveness of keywords retrieval system in relational databases, we use the standard recall, precision and F1 measure. Recall(R) is defined to be the ration of the number of the correctly assigned documents to the number of positive samples. Precision (P) is the ratio of the number of the correct documents in the positively assigned documents. The F1 measure combines recall and precision in the following way:

$$F1 = \frac{2 * R * P}{R + P} \quad (3)$$

In the experiments six data sets were selected and each data set was first partitioned into two sets of mutually exclusive randomly selected examples to form 90% of the databases as training data and 10% of the databases as testing data.

Experiment 1 (Exp.1) uses the keywords retrieval system in relational databases with the fixed keywords number. Experiment 2 (Exp.2) uses the keywords value parameter learning based on traditional particle swarm optimization algorithm. From Fig. 1-3 we can see experiment 2 improved the Precision, Recall and F1 value. Therefore, considering the keywords value parameter learning, it can get better retrieval performance compared to the conventional the results of the experiment. When considering the keywords value learning, it can find the more appropriate keywords number of each table, especially when adding some new data sets.

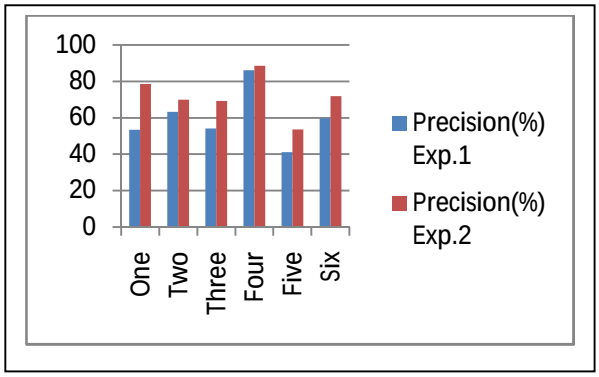

Figure 1. The Precision between Exp.1 and Exp. 2

Figure 2. The Recall between Exp.1 and Exp. 2

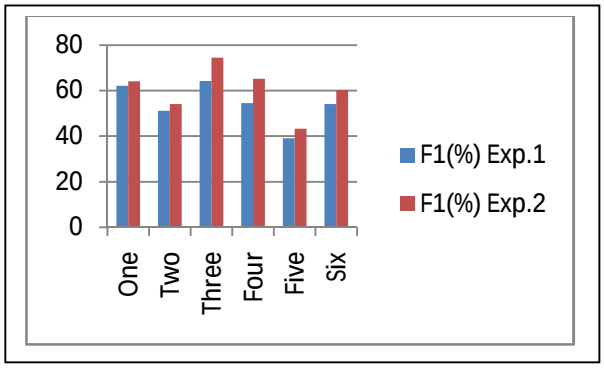

Figure 3. The F1 between Exp.1 and Exp. 2

5. Summary

Keyword search is the most popular information discovery method because the user does not need to know either a query language or the underlying structure of the data. This paper puts forward a query mapping index method, which creates a mapping table, and match index table and the corresponding key words. In the query mapping index method, setting as the threshold, it can use PSO algorithm to learn keywords number.

In the later work, it can consider to graphical knowledge representation of the query results, which displays the query results in the more vivid and appropriate form, rather than a simple two-dimensional data table. When the user enters the keywords it can query search results to satisfy user needs. The results can be displayed with knowledge topology graphics, tables, or graphics. It should continue to study other forms of knowledge representation, then create unique expression characteristics.

6. Acknowledgment

R. B. G. thanks the youth natural science Research Plan foundation of Hebei University(Grant no. 2010Q023) and the Soft Science Research Program of Hebei Province (Grant no. 14450318D)

References

[1] Agrawal S, Chaudhuri S, Das G. DBXplorer: A system for keyword-based search over relational databases. In: Proc. of the 18th Int'l Conf. on Data Engineering (ICDE 2002). San Jose: IEEE Computer Society Press, 2002, pp.5-16. [doi: 10.1109/ICDE.2002. 994693]

[2] Hristidis V, Papakonstantinou Y. DISCOVER: Keyword search in relational databases. In: Proc. of the 28th Int'l Conf. on Very Large Data Bases (VLDB 2002). Hong Kong: Morgan Kaufmann Publishers, 2002, pp.670-681. http://www.informatik.uni-trier.de/~ley/db/conf/vldb/vldb2002.html

[3] Bhalotia G, Hulgeri A, Nakhe C, Chakrabarti S, Sudarshan S. Keyword searching and browsing in databases using BANKS. In: Proc. of the 18th Int'l Conf. on Data Engineering (ICDE 2002). San Jose: IEEE Computer Society Press, 2002, pp.431-440. [doi: 10.1109/ICDE.2002.994756]

[4] Gerard Salton. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. Addison Wesley, 1989.

[5] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. WWW Conference, 1998.

[6] Arvind Arasu, Junghoo Cho, Hector GarciaMolina, Andreas Paepcke, and Sriram Raghavan. Searching the web. Transactions on Internet Technology, 2001.

[7] Jason T., L. Wang, Xiong Wang, Dennis Shasha,Bruce A. Shapiro, Kaizhong Zhang, Qicheng Ma,and Zasha Weinberg. An approximate search engine for structural databases. SIGMOD, 2000.

[8] Daniela Florescu, Donald Kossmann, and Ioana Manolescu. Integrating Keyword Search into XML Query Processing. WWW9 Conference, 1999.

[9] Dandan Li, Lu Han, Yi Ding. SQL Query Optimization Methods of Relational Database System. 2010 Second International Conference on Computer Engineering and Applications. pp.557-560

[10] Jianhua Feng, Guoliang Li, Jianyong Wang. Finding Top-k Answers in Keyword Search over Ralational Databases Using Tuple Units. IEEE Trans. Knowl. Data Eng. (TKDE), 2011, 23(12), pp.1781-1794.

[11] Ziqiang Yu, Xiaohui Yu, Yang Liu. Cascading Top-k Keyword Search over Relational Databases. DOLAP 2011, pp. 95-100.

[12] Yangwei Xu. Scalabel Continual Top-k Keyword Search in Relational Databases. CoRR abs/1108.4516 (2011).

[13] Kennedy J, Eberhart R C. Particle swarm optimization[C]. Proc. IEEE Int. Conf.on Neural Networks, Perth, WA, Australia, 1995, pp.1942-1948

[14] Eberhart R C, Kennedy J A. A new optimizer using particle swarm theory[C]. Proc. The Sixth Int. Symposium on Micro Machine and Human Science, Nagoya, Japan, 1995, pp. 39-43