Military Technical College
Kobry El-Kobbah,
Cairo, Egypt

10th International Conference
on Electrical Engineering
ICEENG 2016

Research on Task-focused Massive Multi-source Heterogeneous Information Sharing & Utilizing Method

By

WU Shan-shan *, Zong Shi-qiang

Abstract:

Nowadays the famous search engine companies are all providing the keyword web search capabilities. No one provides the high accurate & efficient user-requirements-oriented information Services. The task-focused massive multi-source heterogeneous information sharing & utilizing method and system is introduced in this paper. This paper presents the general metadata description model of heterogeneous information resource, the metadata extraction method of information resource, the service encapsulation method and registration publication mechanism of information resource, in order to standard the sharing process of large amounts of heterogeneous information resource. And then, the implement mechanism of information collection and index construction for huge amounts of information resource is given in the paper, by designing the optimized algorithm based on the map-reduce mechanism. Finally, the task-focused massive multi-source heterogeneous information precision searching mechanism is given in the paper in order to realize the information sharing and effective user of resources.

Keywords:

Task-focused, Massive Information, Multi-source Heterogeneous Information, Sharing & Utilizing Method

* Science and Technology on Information Systems Engineering Laboratory, Nanjing Research Institute of Electronics Engineering, Nanjing, China

1. Introduction:

In the information grid environment, enterprise information has the characteristics of heterogeneous structures, enormous quantity, dispersive locations and complex contents. Meanwhile the information resource, information processing node and information user requirement are all dynamic changing. In such complex conditions, the enterprise search engine companies are all providing the keyword web search capabilities. However, between the enterprises, the information is not only stored on web pages. And the uncertainty of information and the uncertainty of demand pose a devastating problem with information sharing and effective use. The simplicity keyword searching can't provide the high accurate & efficient user-requirements-oriented information Services. How to integrate entire network enterprises information resources, how to provide one-stop uniform information search mode, in order to improve the efficiency of information sharing and utilization, are all the problems , which will be solved in the paper.

2. The Design Of The System Framework:

The task-focused massive multi-source heterogeneous information sharing & utilizing system described in the paper is composed of the registration publication sharing module and the cloud search framework module. The registration publication sharing module is in the information source end. The cloud search framework module is in the information center. The system framework is as Figure 1.

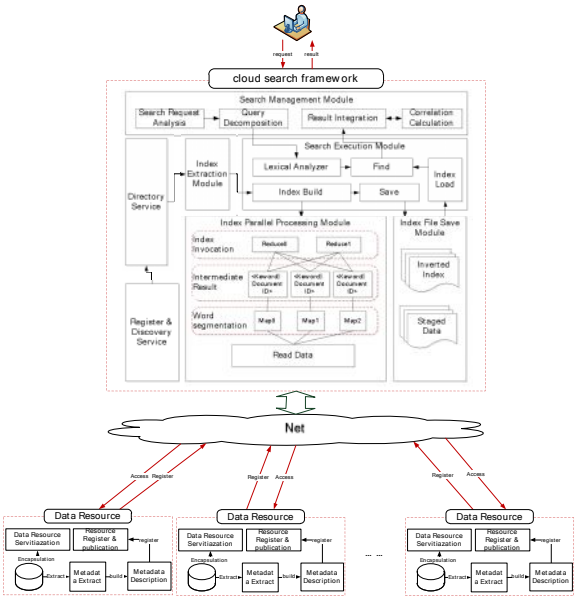

Figure (1): *The enterprise information one-stop search system Framework*

The registration publication sharing module ’which is in the information source end, is used to provide metadata automatically extracting module, general information resources metadata description module, the encapsulation of service access interface of information resources and information resources general registration and publish module. The information resources above-mentioned is composed by the all kinds of enterprise information resource, which are always in form of the database, maybe the Oracle database, SQL Server Database or DB2 database and so on.

The cloud search framework module, which is in the information center, is used to uniform handle huge amounts of information resources, construct the information resources index, providing information integration ability. At the same time, it is used to receive user searching requests, retrieve index in the index database, and return the search results to the user. It includes the register and found directory management module, the information resources index management module, and the searching management and execution module.

The register and found directory management module is used to realize all kinds of information resources registration and publication management and information resource directory management, in order to manage unified and publish the metadata of all kinds of information resources in the information grid.

The information resources index management module consists of three parts: the index extraction module, the index parallel processing module and the index file storage module. The index extraction module is used to fetch all kinds of information resource metadata information in the Register & Discovery Service, interact with the information resource encapsulation module, according to the metadata information of all kinds of information resource, and access the information content which can be shared in the information grid. The index parallel processing module is used to construct the index of all kinds of information resources in the distributed parallel processing approach of the cloud computing. And it provides organization management and integration ability of all kinds of information resources in the information grid. The index file storage module realizes the index storage in the distributed file storage way. The index, which is constructed by the index parallel processing module, includes the inverted index of all information resources and the information which is disposed.

The search management and execution module is mainly used for receiving information search requests coming from the browser, and returning the search results to the user. It is composed of the search management module and the search execution module. The search management module is mainly used for receiving and analyzing user search

request, query decomposing query request, interacting with search execution module, receiving the search results returned from the search execution module, and doing correlation calculation and results integration. The search execution module is the core processor part of the whole index building and executing management. It is used to realize the lexical analysis of search requests submitted by the search management module, loading index information stored in the index file storage module, fetching index and returning the index result.

3. The Model And Method:

A. Enterprise Information Resource Metadata Description Model:

The enterprise information resource metadata description model is mainly used for specification and form a unified metadata description of the enterprise information resources in information grid. It provides the unified management mechanism for the registration of publishing the information resources. It includes Information the resource description metadata, the information resource access metadata and the information resource structure metadata. The enterprise information resource metadata description model is shown in Figure 2.

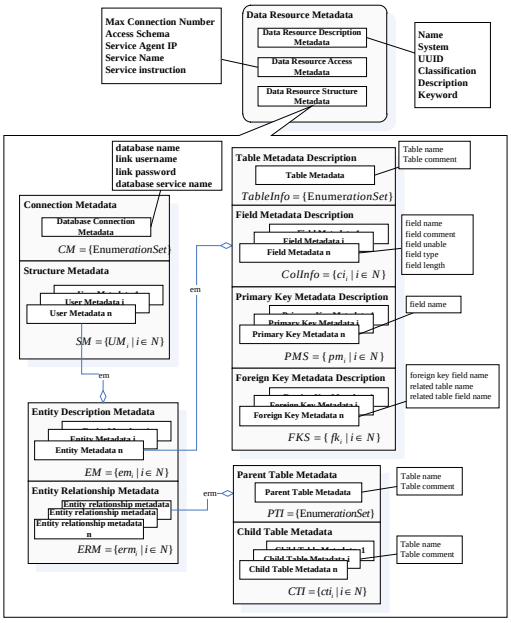

Figure (2): *Enterprise Information Resource Metadata Description Model*

The resource description metadata includes information resource name, belong system, UUID, classification, description and keyword; The information resource access

metadata includes max connection number, access schema, service agent IP, service name and service instruction; The information resource structure metadata is composed by the connection metadata and the structure metadata; The connection metadata (CM) is composed of the database name, the link username, the link password and the database service name, and these four items are the necessary information to establish a database connection; The structure metadata (SM) is composed of several user structure database metadata. And one user structure database metadata can be divided into entity description metadata (EM) and entity relationship metadata (ERM); Entity description metadata (EM) indicates the description of the database table under the database users. A database user always has a lot of database tables, so the user metadata entity description metadata is made up of a set of entity description metadata. The entity description metadata is covering all of the basic database table information. It is composed of the table metadata description, the field metadata description, the primary key metadata description and the foreign key metadata description four parts.

The table metadata description describes the table name and table comment information. The field metadata description describes all field information of the table. The field information includes field name, field comment, field unable, field type and field length. The primary key metadata description describes the primary key fields defined in the table. Usually, it is the set of field name in the table. The foreign key metadata description describes the relationship between the table and other tables. It uses the set of the foreign key field name, the related table name and the related table field name to describe.

Entity relationship metadata (ERM) is used to describe the constraint relationship between the entities. It also means the relationship between the parent table and the child among all database tables. It is described by an entity relationship set. One entity relationship metadata includes the parent table name, the parent table comment and name list and comment list of its children tables.

B. Metadata Automatic Extraction Method:

The metadata automatic extraction method provides the method how to automatic extract the metadata information of the information resources. The metadata automatic extraction tool is designed as the method said. The main procedure is as Figure 3.

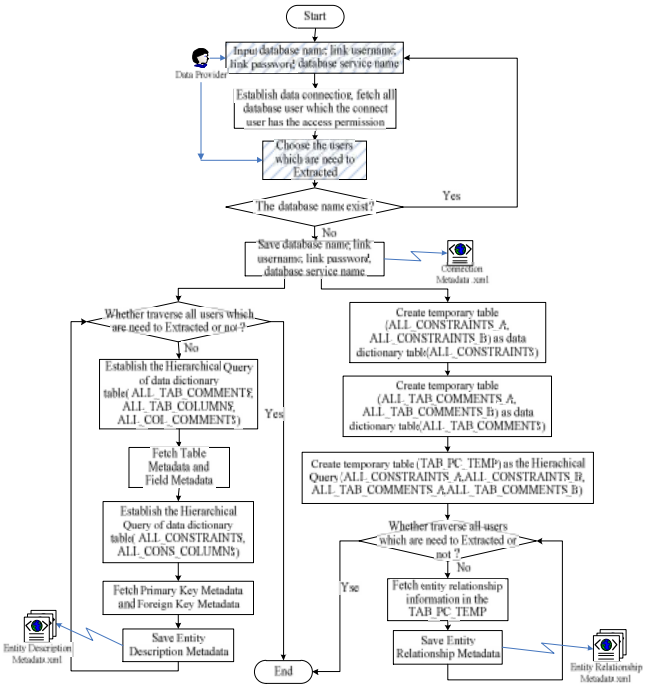

Figure (3): *Metadata Automatic Extraction Flow*

Firstly, Information provider use Input database name, link username, link password, database service name by the tool. The tool will establish information connection, fetch all database users which the connect user has the access permission. Information provider chooses the users, which are needed to extracted, from the fetched database users. Then the tool creates the new XML file (Connection Metadata .xml) for saving database name, link username, link password, database service name.

Secondly, the tool traverses all users, establishing the hierarchical Query of information dictionary table (ALL_TAB_COMMENTS, ALL_TAB_COLUMNS, ALL_COL_COMMENTS), fetching table metadata and field metadata. And establishing the hierarchical Query of information dictionary table (ALL_CONSTRAINTS, ALL_CONS_COLUMNS), fetching primary key metadata and foreign key metadata, then creating the new XML file (Entity Description Metadata.xml) for saving entity description metadata.

Thirdly, the tool automatically creates temporary table (ALL_CONSTRAINTS_A, ALL_CONSTRAINTS_B) as information dictionary table (ALL_CONSTRAINTS) and creates temporary table (ALL_TAB_COMMENTS_A, ALL_TAB_COMMENTS_B) as information dictionary table (ALL_TAB_COMMENTS). And then, creating temporary table (TAB_PC_TEMP) as the hierarchical query result (ALL_CONSTRAINTS_A, ALL_CONSTRAINTS_B, ALL_TAB_COMMENTS_A, ALL_TAB_COMMENTS_B). The

tool traverses all users which are needed to extract, fetching entity relationship information in the TAB_PC_TEMP and creating the new XML file (Entity Relationship Metadata.xml) for saving entity relationship metadata.

Finally, the tool compresses three XML file together and then commits to the center when registering the information resource.

C. The Design of Index Extraction Module:

The index extraction module will mainly interact with the register & discovery service, traverse all information resources registered in the register & discovery server, fetching the metadata of the information resource. The index extraction module execution flow diagram is shown in Figure 4.

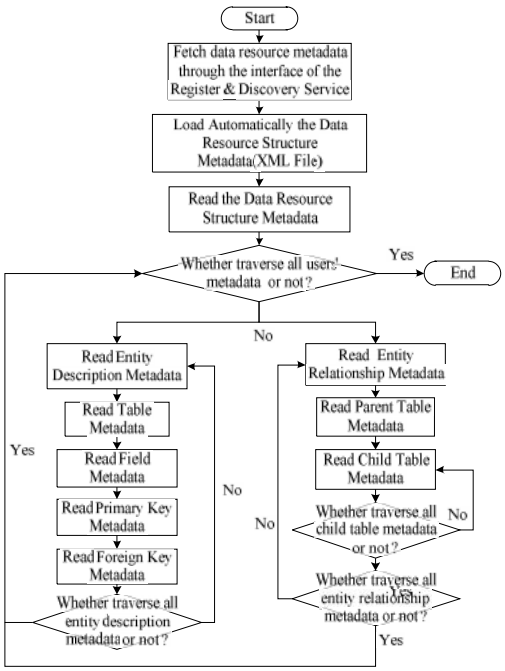

Figure (4): *Index Extraction Module Execution Flow*

Firstly, Fetch information resource metadata through the interface of the register & discovery server, load automatically the information resource structure metadata (XML File). And then, read entity structure metadata, traverse all user metadata, traverse read entity description metadata and entity relationship metadata respectively. Finally, generate entity array and entity relationship array.

D. The Design of Index Parallel Processing Module:

The index parallel processing module is composed by the single entity content parallel extraction module, the correlative entity content parallel extraction module and the index parallel building module.

The target of the single entity content parallel extraction module is to crawl the information content from the remote information resource, based on the entity array, which is generated by the index extraction module, and the service interface information descripted in the information resource access metadata. The whole process is parallel. Input: <server name, entity name>, use map mechanisms, map (server name, entity name) → output <key value1, content>, the entire process using multi-threading, crawling the information content from the remote information resource through the information resource encapsulation agent.

The correlative entity content parallel extraction module's goal is to generate information content set of double or several entities, through crawling the hierarchical information content form the correlative entity, according to the entity relationship array. the whole map-reduce parallel processing as follows: Input:<server name , entity name>, using the map mechanisms, Map (server name, entity name) → output <key value2, content>, the whole process also uses multi-threading and crawl the information content from the remote information resource through the information resource encapsulation agent.

The goal of the index parallel building module is to build index in the map-reduce parallel process way, according to the information content crawled above. The processing is as follows. Input: <key value, Content> document, the key value above including key value 1 and key value 2, Map (): split different key value, performing segmentation of the information content according to the custom index granularity and segmentation methods. Reduce () call Lucence index plug-in to generate the inverted index files, output: inverted index.

E. The Design of the task-focused precision searching mechanism:

The task-focused precision searching mechanism is due to receive the user search request. And then it will parse the search request. Firstly, it will have to do the lexical analysis of the search request, generating the key-value pairs. And according to the key-value pairs, it will generate the query composition through constructing logical operation (and-or-not among) the key-value pairs. Secondly, it has to replace respectively the key and the value by their thesaurus in order to generate new key-value pairs, which can improve the accuracy of the search. Thirdly, it will do the index search

according to the all key-value pairs and generate index search result set. And then, it will do similarity computation with the index search result set and generate the result sorted list. Finally, it will integrate and filter search results according to the user type and the task type.

In order to improve the quality of its results, the task-focused filter search method is introduced in the paper. The method is based on the principal component analysis and logistic regression algorithms. In order to realize the task-focused information filtering, firstly is to identify information relevance between information and the user demand. The mainstream similarity calculation method is to extract information content characteristic keyword, establish characteristic vector, and compute the vector angle to the vector composed of the user demand keywords. However, when characteristic keywords are large, the accuracy and efficiency of this algorithm will fall. In order to resolve this problem, this paper designed the new information filtering method based on the principal component analysis and logistic regression algorithms.
The index extraction module execution flow diagram is shown in Figure 5.

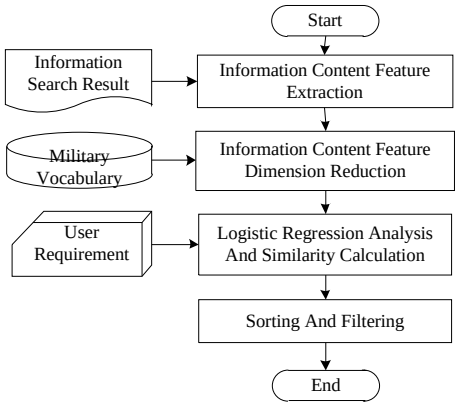

Figure (5): *Index Extraction Module Execution Flow*

It uses the principal component analysis method, extracting the information main characteristics, based on the military vocabulary, and reducing the information content feature vector, in order to get the information low content dimensions. And then, it analyses the low dimensions characteristics by using the logistic regression algorithms. The result is the similarity calculation module. It can calculate information content similarity. Finally, sorting And filtering the result by the similarity. This method can improve the quality of its results.

3. Conclusions:

All kinds of models or methods mentioned in this paper, which could effectively handle the problem of information sharing and utilization brought by heterogeneous structures, enormous quantity, dispersive locations and complex contents of enterprise information under the environment of grid, are mainly used to build up unified cloud search framework of enterprise information. It can provide timely and effective mass level information query access ability by using the cloud computing mechanism to build enterprise information index and the task-focused precision searching mechanism to improve search ability. It provides the main framework for building the enterprise information center in order to do integration and sharing among all enterprise information resources in the information grid.

References:

[1] You Chuan-chuan and Zhang Gui-gang, "A Kind of Efficient Search Method Based on Big Information" Computer Science, vol. 40, No. 2, March 2013, pp. 265-269.

[2] Zhu Ming-dong, Guo Zhi-long and Zhang Sheng, "Research on Information Sharing Service System Based on Information Center" Command Information System & Technology, vol. 1, No. 3, June 2010, pp. 18-22.

[3] Zhou Xiao-lei, Zhang Yan-qin and Sun Jin-hai, "Information Sharing Scheme for Network Centric Command Information System" Command Information System & Technology, vol. 2, No. 3, June. 2011, pp. 14-18.

[4] Cao Ju and Yin Zhe, "Clouds Search Optimization" Computer Engineering & Science, Vol.33, No.10, 2011, pp.120−125.

[5] Tang Yu Wang Ying-jie and Fan Ai-hua, "mDHT:A Search Algorithm to Extra-large Volume of Information Based on Open HDFS Platform and Multi-level Indexing" Computer Science, vol. 40, No.2, Feb. 2013 pp. 195–199.

[6] Wu Guang-jun, Wang Shu-peng and Chen Ming, "Massive Structured Information Oriented Storage and Retrieve System" Journal of Computer Research and Development, vol. 49, No. 1, Sep. 2012, pp. 1-5.