International Journal of Database Theory and Application
Vol.8, No.3 (2015), pp.155-166
http://dx.doi.org/10.14257/ijdta.2015.8.3.13

A Metadata-based Method for Sharing Multiply Heterogeneous Information

Xiaotao Li[1], Xiaohui Hu[1], Weina Lu[1, 2] and Xi Liu[1]

[1] *Department of Automation and Electrical Engineering, Beijing University of Aeronautics and Astronautics, Beijing, 100191, China*
[2]*Department of Mechanics and Electronics, Hebei Normal University of Science and Technology, Changli, Hebei, 066600, China*
E-mail: taosmall@163.com

Abstract

As users' requirements for information integration enhance increasingly, how to integrate multiply heterogeneous data in a global sharing system has especially been a challenge for its large scale and diverse formats. To address the above problem, this paper proposes an information sharing approach for multiply heterogeneous data based on a two-layer metadata. Firstly, the architecture of the two-layer metadata is introduced. Secondly, the synchronization between different users for distributed heterogeneous data is realized by sharing table structures. Finally, Lucene search engine combined with the element GM-description of the two-layer metadata is presented to retrieve metadata, which reduces the response time compared to other retrieval methods. The experiment results illustrate the effectiveness of our approach and the conclusion is given.

Keywords: *Metadata, Information integration, Synchronization, GM-description*

1. Introduction

The information era has two characters, including information explosion and transmission. Both of them are based on the data storage. Along with information expanding and network popularization, the way that accessing single database doesn't satisfy the demand. The requirement to access distributed database is increased. The data can be distributed in any node of the network. How to enable users to find and use the data they required in the ocean of information becomes extremely urgent. Metadata is "data about data". Metadata is traditionally in the card catalogs of libraries. As information has become increasingly digitized, metadata are also used to describe digital data using metadata standards specific to a particular discipline. By describing the contents and context of data files, the usefulness of the original data is greatly increased. The main purpose of metadata is to facilitate in the discovery of relevant information, more often classified as resource discovery. Metadata also helps organize electronic resources, provide digital identification, and helps support archiving and preservation of the resource. Metadata assists in resource discovery by "allowing resources to be found by relevant criteria, identifying resources, bringing similar resources together, distinguishing dissimilar resources, and giving location information [1].

In this study, we present a metadata-based information sharing method. Our aim is to build an information sharing system (ISS) to share multiply heterogeneous data and realize the transmission between databases of different users. The first layer of metadata provides a uniform description format for one type of data, and the second layer of metadata manage the first layer of metadata to shield the difference among multiple metadata formats. The first layer of metadata is named IM (Individual Metadata), and the second layer of metadata is named GM (Global Metadata). Metadata extracted from

ISSN: 2005-4270 IJDTA

data resources are registered to ISS. Users could get data resource information through accessing metadata in ISS. ISS only stores metadata records which have a minimal size compared to original data. So it is possible to integrate massive distributed data resource. With the data synchronization, users are able to transmit the data resource from publishers' databases to local database. In this way, the information will not be affected when the server breaks down or the publishers delete the data released. It benefits the preservation and further application of data. Compared to other data sharing methods, the information sharing approach based on two-layer metadata has the following advantages:

(1) Could share multiple heterogeneous data resources distributed stored in a uniform platform;
(2) Has a global metadata retrieval interface for all categories of data;
(3) Has a faster response times for retrieve;
(4) Could synchronize data streams form publishers to subscribers.

The remaining paper is organized as follows: Section 2 reviews related work. Section 3 describes the structure of the two-layer metadata. Section 4 presents the synchronization approach between publishers and subscribers. Section 5 proposed the metadata retrieve method combining Lucene search engine with the metadata element GM-description. Section 6 presents experiments and Section 7 concludes the paper.

2. Related Work

Metadata as an important technology to share data resources has been applied in many fields. So far, there have been a lot of significant research results. In [2], metadata is obtained by cataloguing resources such as books, periodicals, DVDs, web pages or e-books. These data are stored in the integrated library management system, ILMS, using the MARC metadata standard. In [3, 4], uniform platforms for sharing education resource based on metadata are built to meet the demands of scholars (e.g. learners, teachers, etc.) for e-Learning. The teachers published their courses descriptions information. The students could choose interesting course to study according the course metadata information. In [5], a metadata approach for managing similarities and differences in clinical datasets in a standardized way that uses Common Data Elements (CDEs) is proposed to annotate heterogeneous clinical information, integrate and query it. As in [6, 7, 8], directory systems are set up to manage the scientific experiment data, which provide support to experimental scientists to access their raw data, facility managers to account for facility usage and other scientists who wish to re-use raw experimental data. As digital data becomes increasingly pervasive, metadata is an essential tool to share and retrieve media files, such as music, videos and images. Flickr [9], a large-scale, popular photo-sharing and archiving system owned by Yahoo, offers insight into the collection description and collection building practices of users with the help of collection metadata schemas and other information organization tools. In [10], the authors added new metadata to existing metadata objects of Flickr terms to improve indexing quality. In [11], the author offered a case study of the instrumental information technologies for digital music on computers and placed digital metadata within the broader history of recorded music specifically and digital objects more generally. In [12], the authors proposed a framework for querying a distributed database of video surveillance data in order to retrieve a set of likely paths of a person moving in the area under surveillance. In this framework, each camera of the surveillance system locally processes the data and stores video sequences in a storage unit and the metadata for each detected person in the distributed database.

To summarize, the above researches proposed particular metadata standards to share the data resources in their research fields. The metadata model is the foundation of data discovery and uniform access of distributed resource data. However these standards above have respective emphasis and few of them will be still appropriate if the data category is changed. Even more regrettably, the present standards will be helpless when multiclass data resourced need to be shared in the same platform. Besides, the above researches focus on the centralized management of data but neglect the data transmission between distributed databases, which is an important complement for the preservation of shared data resources.

3. Architecture of Information Sharing

3.1. Process of Information Sharing

As is illustrated Figure 1, there are three classes of users in ISS: publisher, subscriber and manager. The publisher and subscriber have different roles in the process of information sharing, yet are both able to register and subscribe metadata.

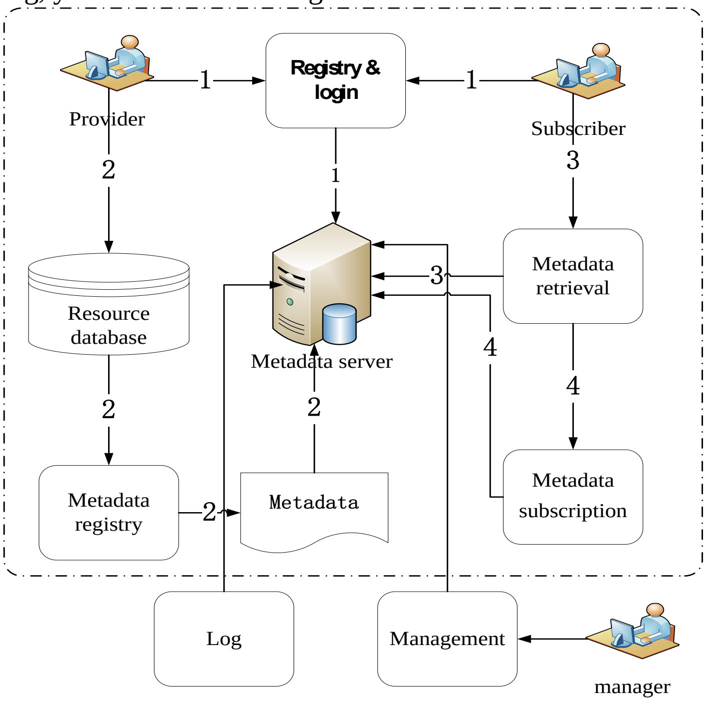


Figure 1. Process of Information Sharing

The process of information sharing based on metadata includes four steps: (1) the publisher and subscriber register an account and login ISS; (2) the publisher selects resource data and register it as metadata to the metadata server; (3) the subscriber retrieves the required metadata by keywords; (4) the subscriber subscribes the metadata, then ISS synchronizes the data mapped by the metadata from the database of the publisher to the database of the subscriber.

The manager monitors the operations of users to ensure the system runs normally. Every operation of the publishers and subscribers is recorded by log system. There are different log permissions between the manager and other users. The manager is able to view all logs of users, however, publishers and subscribers can only view their own logs. In this way, not only is users' privacy protected, but the monitoring capability of manger is also guaranteed.

3.2. Metadata Standard

Considering that ISS in this article mainly concerns data in databases and it needs to satisfy the requirement of users for data synchronization, a two-layer metadata standard is proposed. The characteristics of the two-layer metadata mainly include: (1) the metadata is able to describe the category and content of data resource; (2) the metadata is mapped to the data in databases; (3) permissions for different categories of metadata are different; (4) the metadata should be able to be expanded.

3.2.1. Structure of two-layer metadata: Each kind of data has its own IM standard, which is associated with the data category in ISS. So, there is a uniform description format for the same category of data distributed in different databases and systems. However, there is a distinction between every two IM standards. In order to shield this distinction, GM metadata is built on the basis of IM standards. Every piece of data is able to be described with the same GM standard. The structure of two-layer metadata is shown in Figure 2.

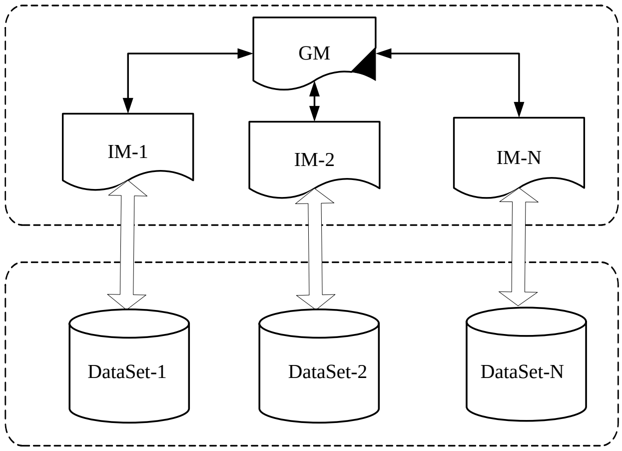


Figure 2. Hierarchical Structure of Two-layer Metadata

3.2.2. Elements of two-layer metadata: The IM elements are composed of required elements and optional elements. The required elements include: index, publisher, IP, database, table, identifier, date and type. The optional elements, which are related with the data category and could be expanded, may have different specific names. Considering the description completeness, there are at least two optional elements in IM.

The GM elements include: GM-index, publisher, type, IM-index, date, GM-description, Metadata-permission, and expanded elements. The expanded elements in GM are mapped with the optional elements in IM. The components of two-layer metadata which contain two optional elements are shown in Figure 3.

Based on the above introduction, there are three important issues needed to be explained: (1) How does GM shield the difference of IM standards; (2) What is the relation between metadata and the data resource; (3) How to distinguish the permission of metadata.

As is illustrated in Figure 4, each category of IM standard has same required element titles. Consequently, a part of GM elements can be directly related to the IM elements. They express the same meaning, e.g., the publisher in GM is the same as the publisher in IM. However, each category in IM has particular optional element titles. Hence expanded elements in GM cannot be directly associated with optional elements in IM. To solve the problem, we combine the titles and content of IM optional elements to be a whole, and then associate it with the expanded element in GM. In this way, the difference of IM standards is shield by GM; what is more, IM information is not missed.

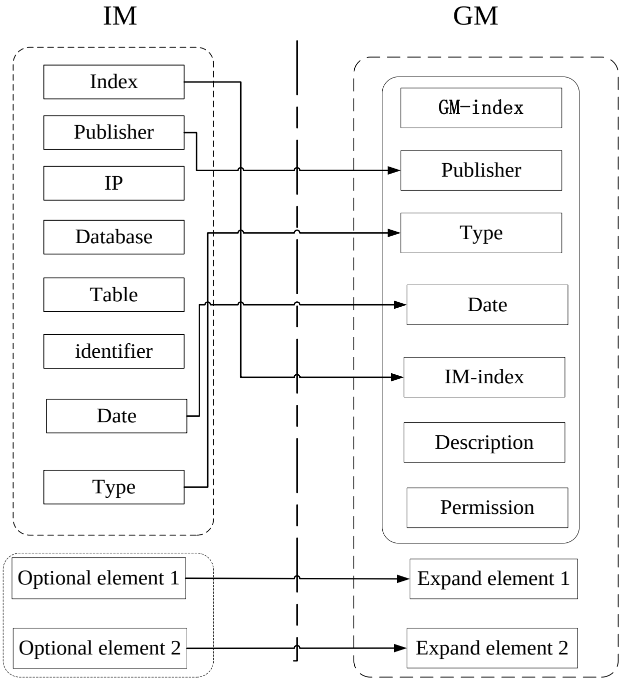


Figure 3. Components of Two-layer Metadata

The mapping relationship between metadata and data resource is performed by a couple of IM elements for locating data, which include IP, database, table and identifier. The four elements form a URI (Uniform Resource Identifier) of a piece of data in the database. What is important we would like to point is that the identifier must be able to precisely locate one data source in the database. Different types of IM may choose different identifiers, and the primary key of a data table is generally chosen as identifier. It is that the IM-index of GM and the index of IM which creates a link between GM and IM. Indeed the IM-index of GM and the index of IM have the same value. Through above mapping relationships, a GM is able to be uniquely mapped into a data resource.

In fact, various users and Metadata may never be part of equation. It is necessary to assign different permissions to users and metadata. The GM element Meta-permission is used to express the permission of the metadata, which is related with the category of GM. Similarity, the user permission is assigned when the user completes his registration. If the user permission is greater than the Meta-permission, the user would be able to view the metadata details. Conversely, users with low user permission cannot find metadata with high Meta-permission. In addition, the user permission is also a measurement of metadata reliability. When metadata is searched the results will be displayed in the order of highest to lowest user permission.

4. Information Resource Synchronization

After the two-layer metadata standard is formulated, how to enable users to synchronize data source from the database of publishers to local database is the primary issue needs to be solved. In the past researches on metadata sharing, the publishing work is completed after data resource is extracted and saved in the server according to metadata standards, which is obviously insufficient to achieve the synchronization. In this paper, there are two improvements performed on the original sharing process: (1) the structure of data resource table of the publisher is shared before metadata is registered; (2) the subscribe capability is extended, which bring about that data resource associated with the metadata is synchronized to the database of subscriber.

4.1. Resource Table Structure Sharing

Multiple database tables enable data synchronization on condition that their structures are uniform. If there is no database table for storing synchronized data, a new data table is needed to be created according to the data table structure of the publisher. Hence the publisher should publish the data table structure before registering metadata. The published data table structure information shown to subscriber as forms is stored in the StructTable (a database table) of the server. The structure of StructTable is illustrated in Table 1, and the column Structure-Info contains the structure information of the table named Table-name. The publisher need respectively to do a publish action for every category of data resource that will be published. If the publisher does not intend to be subscribed by other users, he can choose not to publish the data table structure. Thus, other users could only view the information of GM, IM and data resource, but not synchronize the data resource.

Table 1. StructTable

Column	Type	Length	Is-Key
ID	Number	4	Yes
publisher	Varchar	20	No
Table-name	Varchar	80	No
Structure-Info	Varchar	500	No

4.2. Synchronization between Publisher and Subscriber

It creates conditions for data transmission between multiple databases that publishers share the structures of data tables. The data synchronization of ISS is different from the common database synchronization, since the data synchronization of ISS should satisfy needs of users, while common database synchronization is a simply data copy of different tables. In ISS, the publisher could decide which data to be registered as metadata. Similarly, the subscriber is also able to choose which metadata to subscribe. The subscribed metadata may come from different publishers and have different categories. Indeed users hardly concern a specific data, but generally pay close attention to a category of data. So, the data flow between publisher and subscriber is composed of a category of data which come from the publisher. Therefore the minimum subscribed unit is a specific category of metadata of a publisher in this paper. After the subscriber completes subscription, the subscribe relationship is recorded by ISS. According to the subscribe relationship, ISS pushes the synchronous dataflow to the database of the subscriber. When the publisher registers new metadata which has the same category with the subscribe category or modifies the data resource related with the published metadata, ISS goes on pushing the variations to the subscriber. When the subscriber unsubscribe, the relationship between subscriber and publisher will be lifted, and then the primary subscriber will not be influenced by operations of the primary publisher. The synchronization flow is shown in the figure 4.

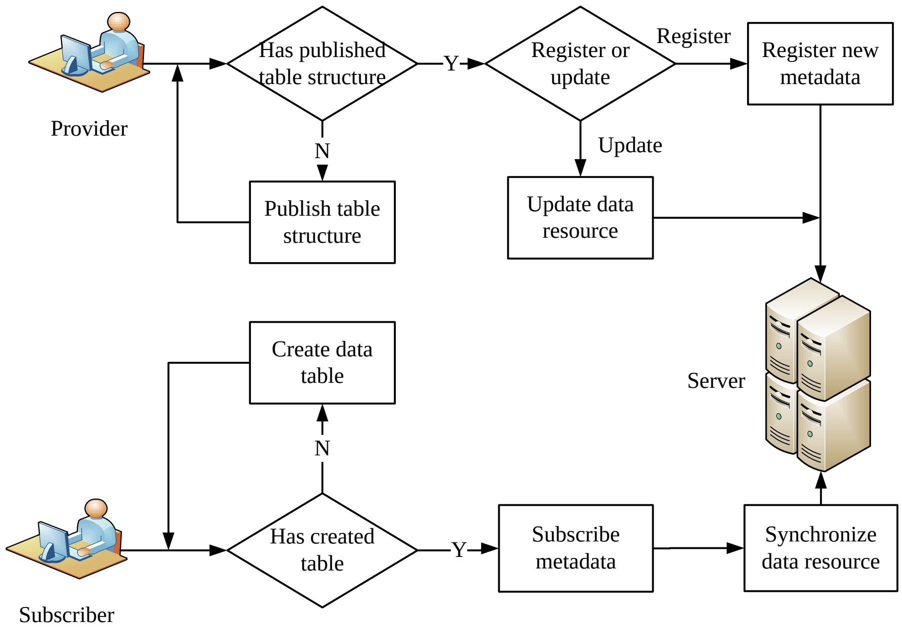


Figure 4. Synchronization Flow

5. Metadata Retrieval

In this paper, the metadata retrieval time is optimized by the GM element GM-description combined with Lucene search engine. Apache Lucene is an open source information retrieval software library, supported by the Apache Software Foundation. While suitable for any application which requires full text indexing and searching capability, Lucene has been widely recognized [13] for its utility in the implementation of Internet search engines and local, single-site searching. The core of Lucene's logical architecture is the idea of a document containing fields of text. This flexibility allows Lucene's API to be independent of the file format. Text from PDFs, HTML, Microsoft Word, and OpenDocument documents, can all be indexed as long as their textual information can be extracted [14]. Since the search operations of Lucene are performed in the indexed file, the metadata records, which are stored in relational database, should be converted to the indexed file in advance. Although Lucene has powerful retrieval capability, the conversion time between metadata records and indexed file still needs to be considered. The retrieve time rises as the size of indexed file increases. In this article, we extract the main information of a GM record by the GM element description to convert indexed file instead of full metadata record. The GM is described by GM-description with a format of RDF (Resource Description Framework) [15]. The element GM-description can be indicated as a statement includes subject, predicate and object. The subject means that the publisher registers which category of GM; the predicate denotes properties of GM; the object denotes property values. Here, an illustration is taken to show the content of a weather metadata GM-description: "user in 2014-8-17 publishers a weather metadata: region is Beijing, weather is clear, temperature is 25℃". In this way GM-description contains whole information of GM. In the traditional keyword searching method, all elements will be visited to determine whether keywords are contained or not. Through GM-description, the keyword retrieval for metadata can be focused on the GM-description. Therefore the indexed file just needs two fields, one is the primary key of metadata records, and the other one is GM-description. After a series of above-mentioned processing, the size and fields of indexed file decrease, so the retrieval time will reduce. Figure 5 shows the structure of GM-description.

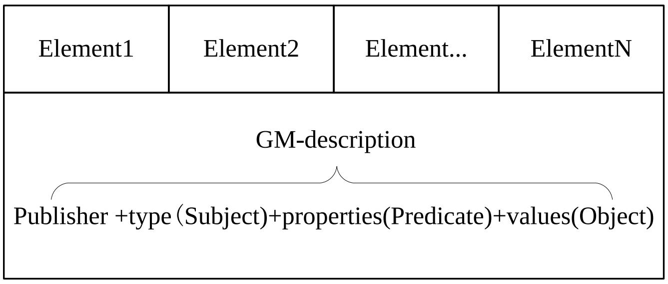


Figure 5. The Structure of GM-description

6. Experiment Results and Discussions

ISS in this paper adopts Brower/Server mode, which is developed by java server pages programming language. Experimental hardware environment: Intel dual-core P9600 processor clocked at 2.53GHz, 2G DDR3 1333Hz RAM, Windows XP Professional operating system. Experimental software environment: Oracle 11.0 database, MyEclipse 11.0 development platform. The experiment test data include 5 types: continental weather data, ocean weather data, air pollution data, typhoon warning data and thunder warning data. All of the test data resources are available in http://www.weather.com.cn.

6.1. Information Sharing Experiment

In this part of experiment, there are 5 client nodes and a server node. Each node has a unique username. The usernames are user1, user2, user3, user4 and user5, respectively. To reflect multiclass heterogeneous data sharing, we make every node publish a category of data resources. Figure 6 shows a part of registered metadata records of the ISS. The column names are corresponding to the GM element names. In this page, user1 published 2 continental weather data records, user2 published 2 ocean weather data records, user3 published 2 air pollution data records, user4 published 2 typhoon warning data records, and user5 published 2 thunder warning data records. These GM records have same structures, but the contents are different. For example, the optional elements of the continental weather metadata are “Area” and “Weather”, and the optional elements of ocean weather data are “Sea” and “Weather”. The metadata records are integrated in ISS and related with the distributed data resources in the 5 client nodes. Figure 7 shows the IM information of continental weather data, which is related with the first GM record registered by user1. The text marked in red is the content of the required elements of the IM, and the others denote the information of the optional elements. IM supports more detailed information, and the original data in the node of user1is able to be accessed through the IM required elements IP, Database, Table and Identifier. Therefore, we can see that our proposed method base on the two-layer metadata is able to shared multiply heterogeneous data at the same time.

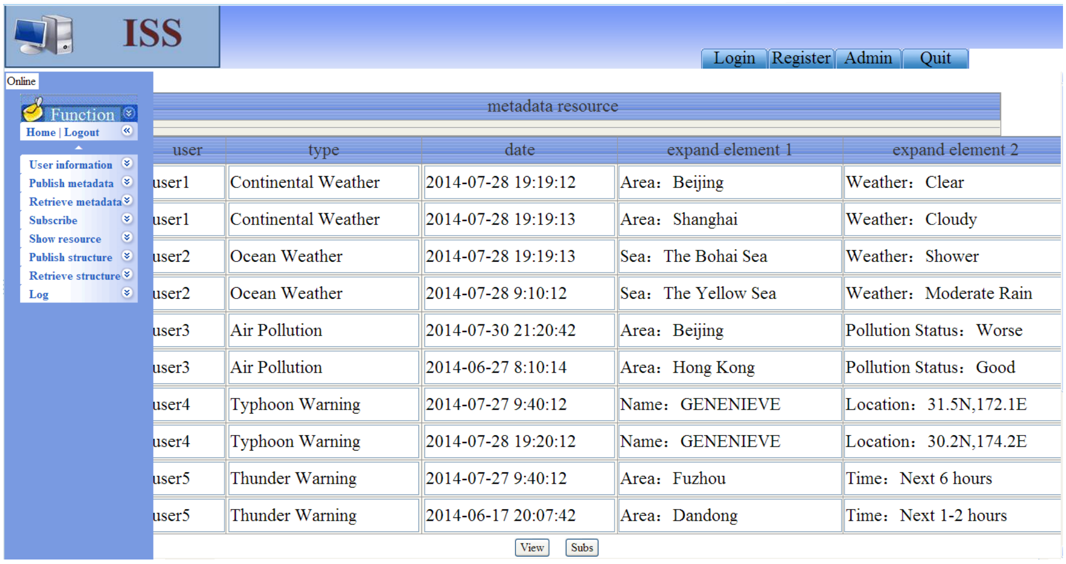


Figure 6. GM Records in ISS

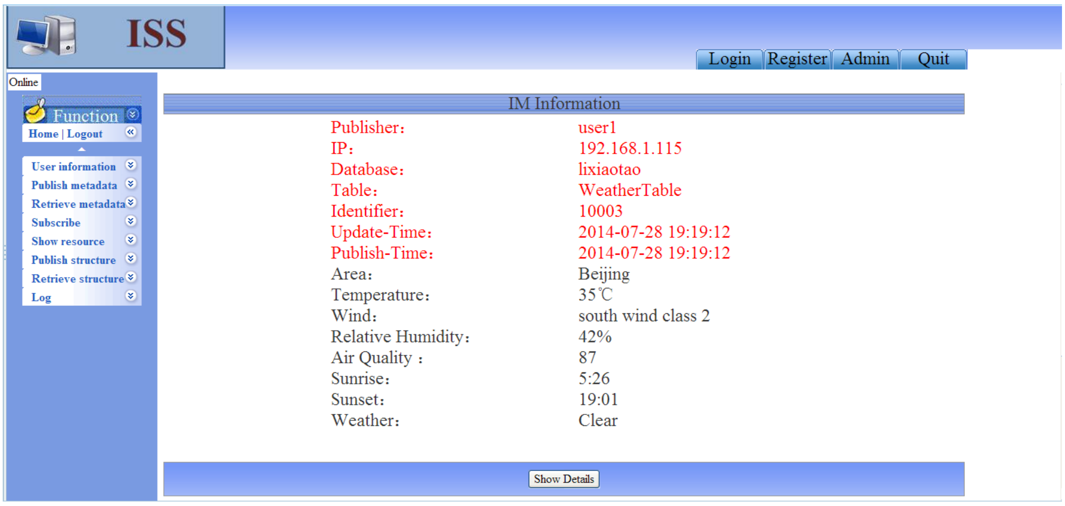
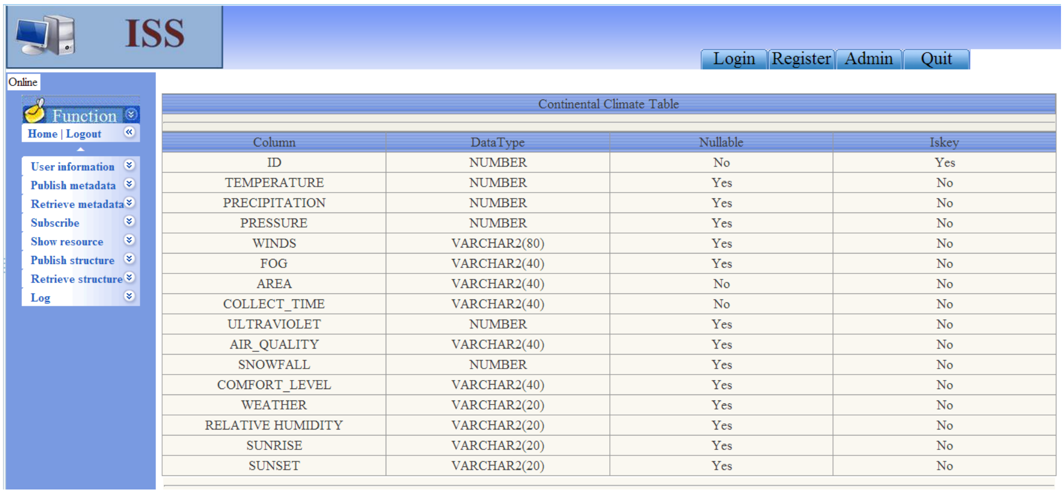


Figure 7. IM Information

6.2. Information Synchronization Experiment

In this part of experiment, we make the client node of user2 subscribe the continental weather metadata registered by user1. Before the subscription, user1 published the continental weather data table structure. Figure 8 shows the continental weather data table structure information, which contains the column name, data types, nullable, and the primary key information. This table structure information sets rules for a database table. The number, order and data types of columns must agree with the structure information. Besides, the primary key columns and nullable columns should be considered as well.

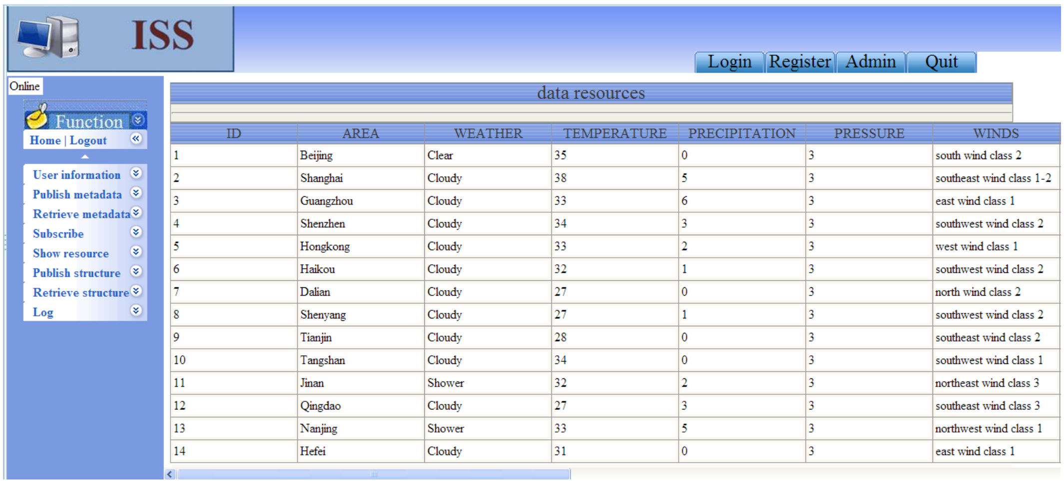
User2 created a new table to store synchronized continental weather data from the client node of user1 according to the data structure published by user1. After that the client node of user2 has a same continental weather data table with user1 except the data resources. Through metadata subscription, the continental weather data resources published by user1 are transmitted to the newly created table in the client node of user2. Figure 9 shows the synchronized result. The continental weather data records have been inserted to the newly created table.



Continental Climate Table

Column	DataType	Nullable	Iskey
ID	NUMBER	No	Yes
TEMPERATURE	NUMBER	Yes	No
PRECIPITATION	NUMBER	Yes	No
PRESSURE	NUMBER	Yes	No
WINDS	VARCHAR2(80)	Yes	No
FOG	VARCHAR2(40)	Yes	No
AREA	VARCHAR2(40)	No	No
COLLECT_TIME	VARCHAR2(40)	No	No
ULTRAVIOLET	NUMBER	Yes	No
AIR_QUALITY	VARCHAR2(40)	Yes	No
SNOWFALL	NUMBER	Yes	No
COMFORT_LEVEL	VARCHAR2(40)	Yes	No
WEATHER	VARCHAR2(20)	Yes	No
RELATIVE HUMIDITY	VARCHAR2(20)	Yes	No
SUNRISE	VARCHAR2(20)	Yes	No
SUNSET	VARCHAR2(20)	Yes	No

Figure 8. Table Structure of Continental Weather Data



data resources

ID	AREA	WEATHER	TEMPERATURE	PRECIPITATION	PRESSURE	WINDS
1	Beijing	Clear	35	0	3	south wind class 2
2	Shanghai	Cloudy	38	5	3	southeast wind class 1-2
3	Guangzhou	Cloudy	33	6	3	east wind class 1
4	Shenzhen	Cloudy	34	3	3	southwest wind class 2
5	Hongkong	Cloudy	33	2	3	west wind class 1
6	Haikou	Cloudy	32	1	3	southwest wind class 2
7	Dalian	Cloudy	27	0	3	north wind class 2
8	Shenyang	Cloudy	27	1	3	southwest wind class 2
9	Tianjin	Cloudy	28	0	3	southeast wind class 2
10	Tangshan	Cloudy	34	0	3	southwest wind class 1
11	Jinan	Shower	32	2	3	northeast wind class 3
12	Qingdao	Cloudy	27	3	3	southeast wind class 3
13	Nanjing	Shower	33	5	3	northwest wind class 1
14	Hefei	Cloudy	31	0	3	east wind class 1

Figure 9. Synchronized Continental Weather Data

6.3. Comparing Retrieval Time

At the same time, retrieve time experiments are performed to show the comparison of our searching method with the database "select like" keyword search method and single Lucene search method. The retrieval time is acquired from the log system of ISS, which subtracts the time of returning result from the time of clicking searching button. In the First experiment, the number of IM optional elements is set to 10, and the number of GM increases from 400 to 4000. The retrieval time results for the same keyword of our method and the other two methods are shown respectively in Figure 10. In the second experiment, the number of metadata is set to 1000, and the number of IM optional elements increases from 2 to 20. The contrast retrieval time results of our method and the other two methods are shown in Figure 10. It can be obviously seen from Figure 10 and Figure 11 that our method and Lucene search method both have an large advantage than "select like" search method under the conditions given by these two experiments. That is because Lucene search engine has more powerful search ability than "select like" search, especially when the number of records is massive. As the Number of metadata records and optional elements increase, our method gets faster response time. The reason is that our method combines Lucene search engine with the GM element GM-description. The indexed file mapped by GM-description has a smaller volume than the indexed file mapped by the whole GM records. In conclusion, our retrieve method has the best retrieval performance than the other two methods.

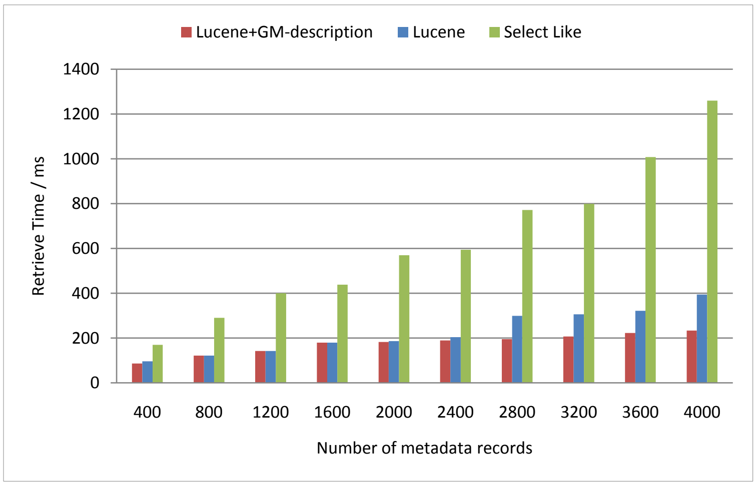


Figure 10. Retrieval Time Comparison with the Same Number of Metadata Items

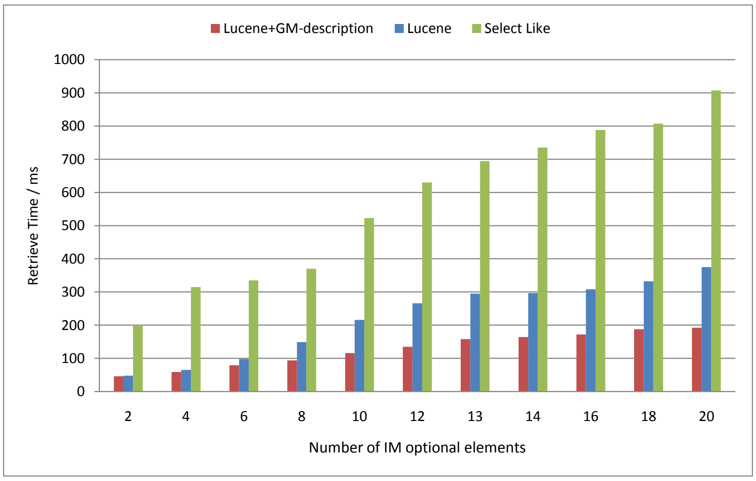


Figure 11. Retrieval Time Comparison with the Same Number of Optional IM Elements

6. Conclusion

This paper proposed a two-layer metadata standard to accomplish the sharing and synchronization of multiple heterogeneous data. The metadata standard is composed of GM standard and IM standard. IM standard is a uniform format of one category of data, and GM standard is a global format of all categories of IM standards. The data synchronization between publishers and subscribers, which is a characteristic work in this paper, is realized through sharing table structures. The retrieval time is optimized by combining Lucene search engine and the GM element GM-description. The experimented results proved that the method based on two-layer metadata is suitable for heterogeneous data sharing and the retrieval method decrease the response time of key words retrieval.

Acknowledgements

This research was supported by National Natural Science Foundation of China (No. 61273350).

Statement

This paper is a revised and expanded version of a paper entitled "Research on Metadata-based Multiclass Information Sharing Technology" presented at the 2014 2nd International Conference on Computer Science and Information Processing, Taiyuan, China, 24-25, May, 2014.

References

[1] R. Guenther and J. Radebaugh. "Understanding Metadata", National Information Standards Organization, Bethesda **(2004)**.
[2] M. Khoo, C. Hall, "Managing metadata: networks of practice, technological frames, and metadata work in a digital library", Information and Organization, vol. 23 , no. 2, **(2013)**, pp. 81-106.
[3] B. Zhao, M. K. Liu, F. Xiao and X. R. He, "On metadata of curriculum resource based on ethnic minority culture in Information technology teaching in the ethnic minorities' area", Proc. The 7th International Conference on Computer Science & Education, **(2012)**, pp. 1932-1935.
[4] W. Hu, S. Liu, H. Y. Ye and C. J. Su, "Metadata-based management for educational resources", Proc. The 6th International Conference on Information Management, Innovation Management and Industrial Engineering, **(2013)**, pp. 253-255.
[5] I. Papatheodorou, C. Crichton, L. Morris, P. Maccallum, J. Davies, J. D. Brenton and C. Caldas, "A metadata approach for clinical data management intranslational genomics studies in breast cancer", BMC Medical Genomics, vol. 66, no. 2, (**2009**), pp. 1-10.
[6] E. Yang, B. Matthews and M. Wilson, "Enhancing the core scientific metadata model to incorporate derived data Future Generation Computer Systems", vol. 29, no. 2, (**2013**), pp. 612-623.
[7] J. Qin, A. Ball and J. Greenberg, "Functional and architectural requirements for metadata: supporting discovery and management of scientific Data", Proc. Dublin Core and Metadata Applications, **(2012)**, pp. 62-71.
[8] C. Willis, J. Greenberg and H. White. "Analysis and synthesis of metadata goals for scientific data", American Society for Information Science and Technology, vol.63, no. 8, **(2012)**, pp.1505-1520.
[9] B. Stvilia and C. Jörgensen. "User-generaed collection level metadata in an online photo-sharing system", Library & Information Science Research, vol.31, no. 1, **(2009),** pp, 54-65.
[10] B. Stvilia, C. Jorgensen and S. H. Wu, "Establishing the value of socially-created metadata to image indexing," Library & Information Science Research, vol. 34, no. 2, **(2012)**, pp. 99-109.
[11] J. W. Morris, "Making music behaves: metadata and the digital music commodity," New Media & Society, vol. 14, no.5, **(2012)**, pp. 850-866.
[12] L. L. Presti, S. Sclaroff and M. L. Cascia, "Path Modeling and Retrieval in Distributed Video Surveillance Databases," IEEE Transactions on Multimedia, vol .14, no. 2, **(2012)**, pp. 346-360.
[13] M. McCandless, E. Hatcher and O. Gospodnetic. Lucene in Action, Second Edition: Covers Apache Lucene 3.0, Manning Publishers **(2010)**.
[14] P. Wittek and S. Daranyi. " Accelerating Text Mining Workloads in a MapReduce-based Distributed GPU Environment", Journal of Parallel and Distributed Computing, vol. 73, no .2, **(2013)**, pp. 198-206.
[15] F. Du, Y. G. Chen and X. Y. Du. "Survey of RDF query processing techniques", Journal of Software, vol.24, no.6, **(2013)**, pp. 1222-1242.

