Development of an Integrated Retrieval System on Distributed KRISTAL-2002 Systems with Metadata Information

Young-Kwang Nam[1], Gui-Ja Choi[1], Byoung-Dai Lee[2]

Department of Computer Science, Yonsei University, Korea[1]
{yknam, gjchoi}@dragon.yonsei.ac.kr

Digital Home Solution Lab, Samsung Electronics, Co. LTD., Korea[2]
byoungdai.lee@samsung.com

Abstract. In this paper, we propose an integrated information retrieval system for multiple KRISTAL-2002 systems for different areas or systems for the same area with the different schemas by using the metadata information so that the users can get the answers by once from the whole systems. Our approach utilizes integrated metadata and mapping information between the metadata and the actual database schema information of participating systems. Therefore, we do not require modification of the participants for integration. We have implemented and deployed the proposed system that integrates six different databases distributed across multiple sites.

1. Introduction

Since the early 70's, many organizations in diverse areas have been developing and deploying information retrieval systems. Due to the availability of Internet, user demand for integrated search through the information retrieval systems that have been deployed has increased while the speed with which these systems are integrated does not catch up with the user demand. Integrating the information retrieval systems in each special area requires storing millions of bibliographic information either into different tables through generalization as in relational database systems or into a single table. The former approach may suffer from poor performance caused by JOIN operations. The latter approach may have low space utilization because there is significant duplicate information.

From service provider's point of view, it is not appropriate to modify existing information retrieval systems to build an integrated information retrieval system because they may need to stop servicing. However, from user's point of view, it is tedious to visit, register and search each system manually whenever they need information. In this paper, we propose an integrated information retrieval system (, which we call ***Integrated Server***) that does not require modification of participant systems (, which we call ***Source Server***). In the proposed system, mapping information between integrated metadata and actual databases or tables of each source

server is maintained and is utilized for integrated search. For integrated metadata, we follow ISO/IEC 11179 metadata registry procedure.

In order to make the integration process simple, the source server administrator is able to map the database schema of the source server to the meta-fields that will be used for integrated search. In this approach, users are able to search distributed information sources without considering their locations and the kinds of information each server maintains. The user query, however, must be re-generated for each source server based on the mapping information. This relationship is shown on Fig. 1.

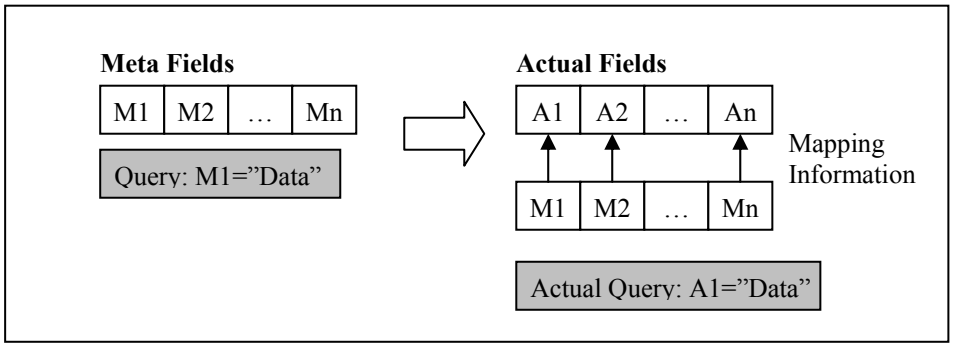

Fig. 1. Relationship between meta fields and actual fields.

We have implemented and deployed the proposed system using KRISTAL-2002 system, which will be explained in later section. We include six sources servers in different areas that are currently servicing independently: Science Literature DB, Scientific Technical Trend DB, Scientific Technical Report DB, Scientific Technical Analysis DB, Patent DB and Human Resource DB.

The paper is organized as follow: Section 2 gives related work and Section 3 describes KRISTRAL-2002 system. Section 4 presents the proposed integrated information retrieval system in details. Section 5 presents the prototype implementation and finally, we conclude in section 5.

2. Related Work

There have been considerable efforts in developing algorithms and protocols for integrating heterogeneous data distributed across multiple sources. However, schema integration and schema mapping techniques dealing with discrepancies of schema among distributed data sources require more research to address semantic interoperability.

Techniques for web search engines and directory-based portal services are the driving forces for advancing the information retrieval area. They must be able to search a large amount of data with short queries given by users. To address the problem, these techniques utilize ranking algorithms that use link information or structural information of the documents. This type of information is not included in

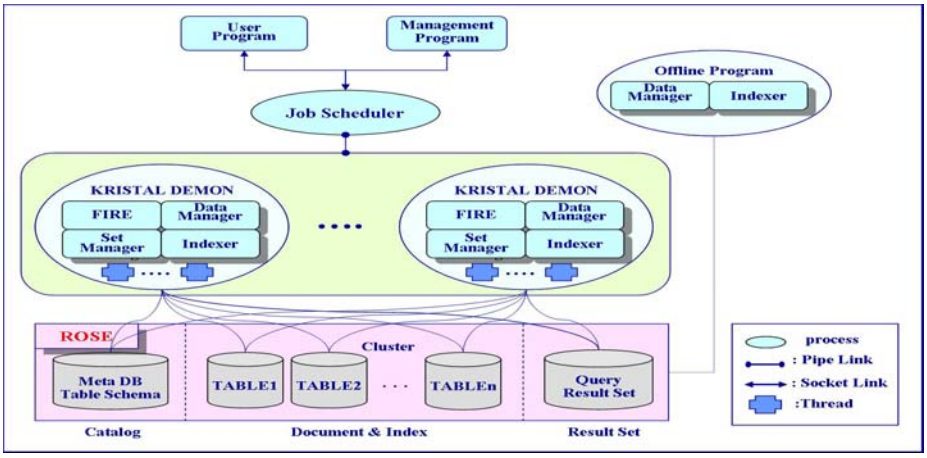

Fig. 2. The architecture of KRISTAL-2002.

the original documents. Crawling [18] is a new technique for web search engines. It collects and indexes documents distributed across multiple sites and the primary focus of crawling technique is how fast a given document can be included in the search target. To be more effective, it must address issues related to collecting, storing documents and serving user queries. Clustering [19] focuses on presenting accurate information to users by combining related search results. It, however, must improve performance and accuracy of clustering. Meta-searching technique ([20][21]) sends a user query to multiple web search engines and presents the results after combining the partial results obtained from them. It does not require web crawler or indexing a large amount of documents. However, it must be able to combine the results effectively that are found by different ranking algorithms of each source system.

Ontology ([8][9][10][11]) and similar researches address semantic interoperability between metadata in consideration of mapping the meaning and the presentation of source data into real-life objects and concepts. Examples of such research include RDF [4], Schema Integration [14], Intelligent Integration of Information [15] and Knowledge Sharing Effort [16], to name a few

3. KRISTAL-2002

KRISTAL-2002 is an information retrieval and management system developed by Korea Institute of Science & Technology. It runs on both Windows and Unix systems. KRISTAL-2002 consists mainly of five modules (Job Scheduler, FIRE, Data Manager, Set Manager and Indexer) and they communicate one another through sockets or pipes (Figure 2.). Job Scheduler distributes user requests to FIRE, which conducts actual searching. Set Manager stores and manages documents found by FIRE. Data Manager processes Document Update or Document Store requests from Job scheduler. Once it completes the requests, it sends acknowledgement to the Job Scheduler. If the operation is successful, it contacts Set Manager so that the Set

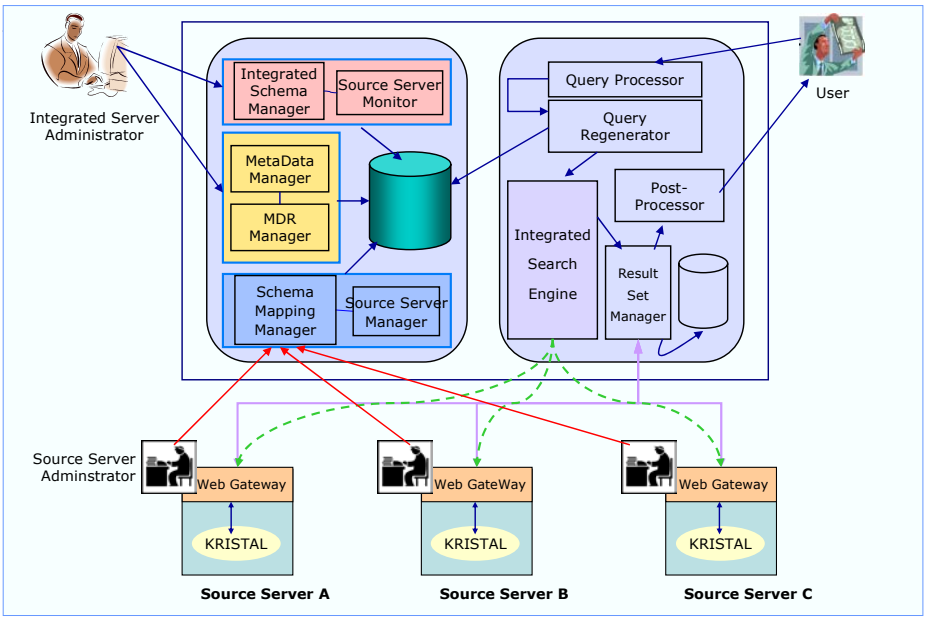

Fig. 3. System architecture of the integrated information retrieval system.

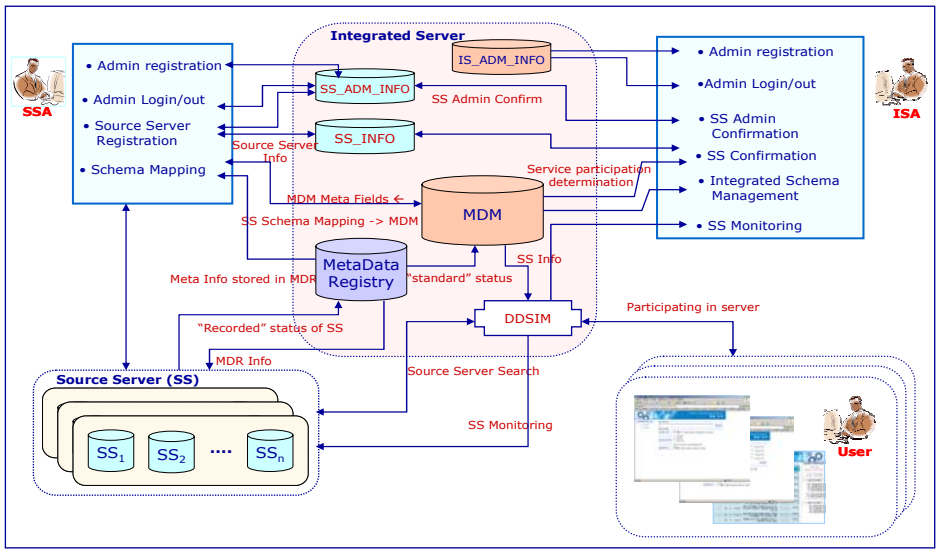

Fig. 4. Control and information flow between the source server and the integrated server.

Manager can update the documents it maintains. Indexer is able to analyze input documents and extract indexing words that can best describe the documents. Indexer supports analysis of Korean morpheme, English stamping, Chinese character conversion and user-defined dictionary, to name a few. FIRE, Data Manager, Set Manager and Indexer are integrated into a single daemon and it is able to manage multiple databases.

KRISTAL-2002 database supports processing simultaneous user queries and on/off-line data management and fast and safe backup. Searching and processing

BLOB (Binary Large OBject) is time-consuming. To address this, KRISTAL-2002 utilizes multiple threads for distributed query processing. The primary components of the databases are: Catalog, Document and Index, and Result Set. Catalog database maintains schema information such as table structures, indexing methods and primary keys and so on. Document and Index database is structured by a single or multiple clusters and each cluster is composed of tables that have the same schemas. Cluster is the basic unit of ranking. Each table in Document and Index database is composed of documents, primary keys, and index database and the structure of the table is defined by table schema stored in Catalog database. Result Set database maintains documents found so that it can respond to user requests quickly.

4. System Architecture of the Integrated Information Retrieval System

4.1 System Architecture

The proposed integrated information retrieval system is based on KRISTAL-2002 and it utilizes metadata information registered by participating source servers. To address the heterogeneity of data and schema information of each source server, metadata maintained in the integrated server is used for structural integration. Furthermore, standardized procedure proposed in ISO/IEC 11179 is employed for metadata registry procedure. Schema information of each source server must be mapped into metadata maintained in the integrated server. Therefore, schema that hasn't been mapped will not be considered for search process. Figure 3. shows the system architecture of the proposed system.

The source server administrator manages the source server and the structural information of the database maintained by the source server while MDR (MetaData Registry) manager standardizes data elements for metadata registry. Integrated server administrator manages and controls all the information of the whole system to support rapid and effective user query processing. Each component of the system will be explained in more details in later sub sections.

The control and information flow among the source servers and the integrated servers is depicted in Figure 4. Each source server administrator registers necessary information to the integrated server and the integrated server administrator authorizes the source server administrator so that he or she can register the source servers. The registered source servers must go through the same authorization process as with the source server administrator to participate in the whole system. Once the registered source servers are authorized by the integrated server, the source server administrators are entitled to do schema mapping. As with previous steps, schema mapping done by the source server must be confirmed by the integrated server. After these steps, the source server is included for integrated search.

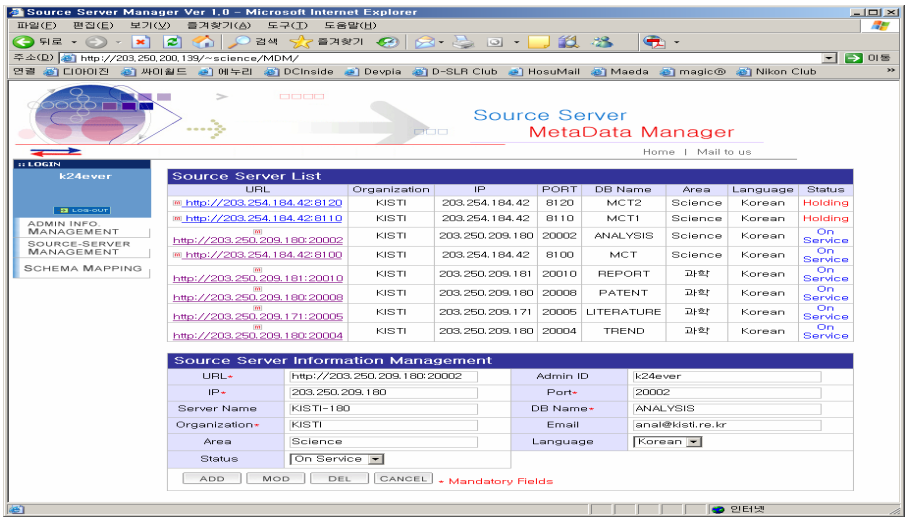

Fig. 5. Web interface for source server registration.

4.2 Source Server Manager

Each source server can run its own database systems and are responsible for searching and providing documents stored in the database when requested by the integrated server. The source server administrator is able to add/remove the source server to/from the set of source servers that will be included for integrated search, and to map the database schema of the source server to the database schema of the integrated server. All these operations are conducted through web interfaces.

4.2.1 Source Server Information Management

Figure 5. shows the web interface for source server registration. When URL column on Source Server List is clicked, the information of the selected server will be shown on Source Server Information Management and the source server administrator can change information of the server or remove from the list. Status column on Source Server List represents three different status information of the source servers: *Holding, On Service, Unavailable*. Although the source server has been registered, it is not included for the integrated search until it is confirmed by the integrated server. During this period, the status of the source server is Holding. Once confirmed, the status of the source server is On Service. If the source server cannot service temporarily due to some reasons, for example, updating operating system of the source server, the source server administrator can change the status of the source server to Unavailable. When the source server returns to the normal condition, the administrator can change the status of the server to “On Service”.

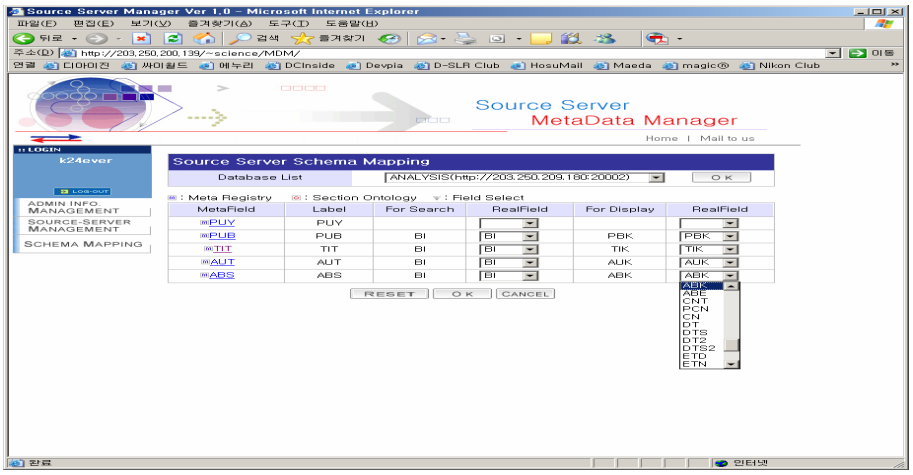

Fig. 6. Web interface for schema mapping.

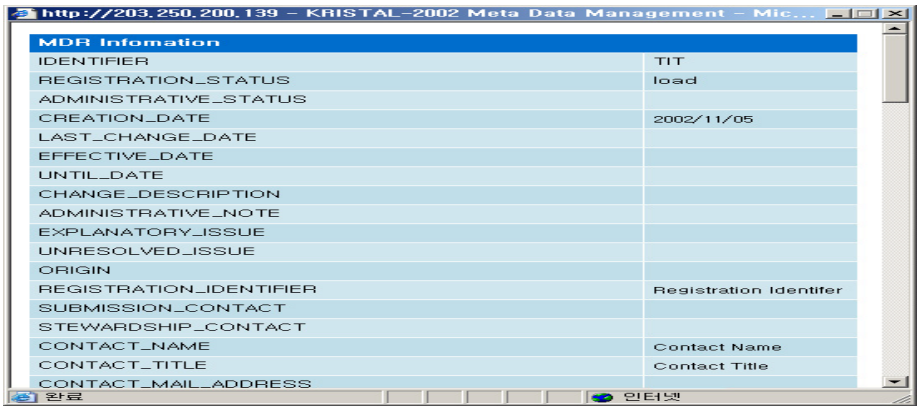

Fig. 7. Metadata information

4.2.2 Source Server Schema Mapping

Figure 6. shows the web interface for schema mapping for the source server. If multiple source servers are managed by an administrator, only those source servers whose statuses are On Service are listed on the web page. MetaField column represents integrated meta schemas loaded from metadata registry and Label column represents the names of the meta schemas. When values in MetaFields are clicked, the detailed information of the meta field is shown on a different window (Figure 7.). RealField represents database schema of the source server mapped into the integrated meta schema. The values listed in the combo box are loaded directly from the source server.

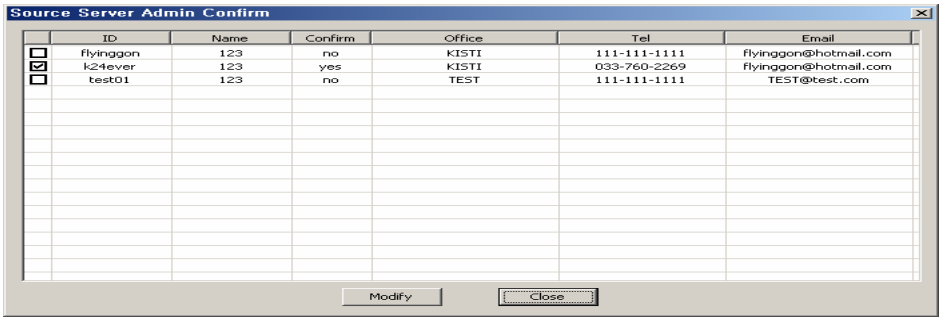

Fig. 8. Source server administrator confirmation.

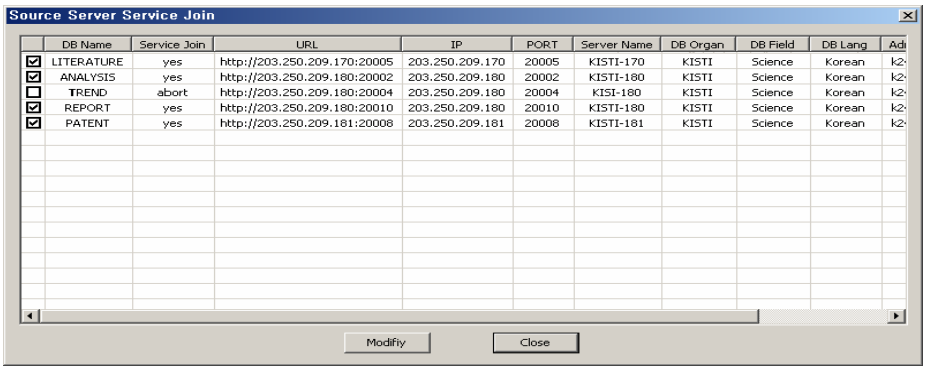

Fig. 9. Source server confirmation.

4.3 Integrated Server Manager

The role of the integrated server is to register metadata based on ISO/IEC 11179 and to maintain integrated meta schema and mapping information registered by the source servers. It also plays a role of entry point for the integrated search. Integrated server administrator determines which source servers will be included for the integrated search service and validity of schema mapping of the source servers. It monitors the status of the source servers included in the service. When it is detected that some source servers are not running correctly, the integrated server immediately exclude the servers from the source server list.

4.3.1 Source Server Confirmation

Integrated server administrator is able to confirm the requests from the source server administrators who want to participate in the integrated search service (Figure 8). Once confirmed, the source server administrators are entitled to add their source servers into the service (see 4.2.1.). For the source servers registered by the source server administrator, the integrated server administrator needs to decide whether or not to add them into the server list that will actually service end-users (Figure 9). Note

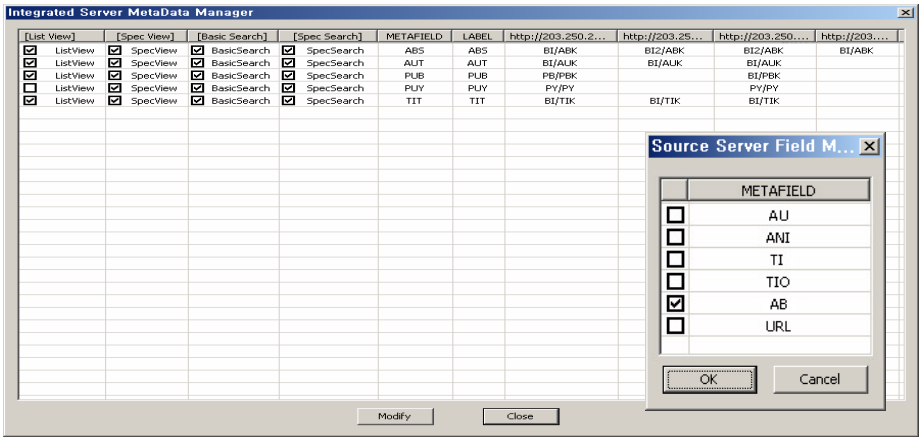

Fig. 10. Source server schema mapping manager.

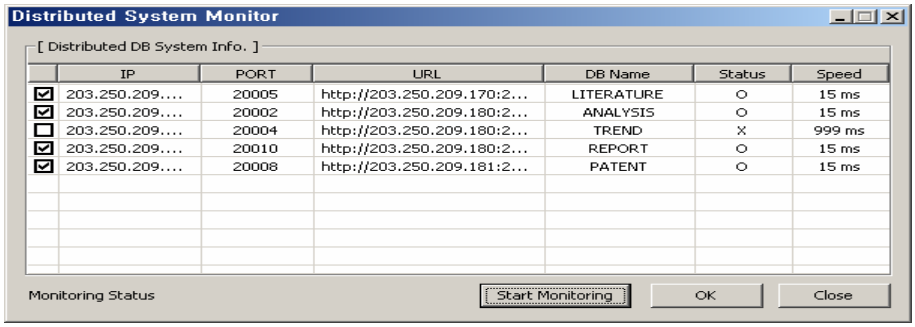

Fig. 11. Source server monitor.

that these two processes are different in that the former validates the source server administrators while the latter validates the source servers that the valid source server administrator registered.

4.3.2 Schema Mapping Management

Schema Mapping Manager of the integrated server determines integrated metadata schema and given schema mapping registered by the source server administrator (see 4.2.2), it checks the validity of the mapping (Figure 10.). Each row in Figure 10 provides schema mapping information of all the source servers participating in the integrated search service. ListView and SpecView columns determine which meta fields will be shown to users after search completes. Typically, those fields marked as ListView will give rough information of the selected documents while those fields marked as SpecView will be used to provide more detailed information of the documents. BasicSearch and SpecSearch columns determine which meta fields must be compared against the given user query. We support two different modes for search operation: Basic Search and Specific Search. In basic search, user query will be

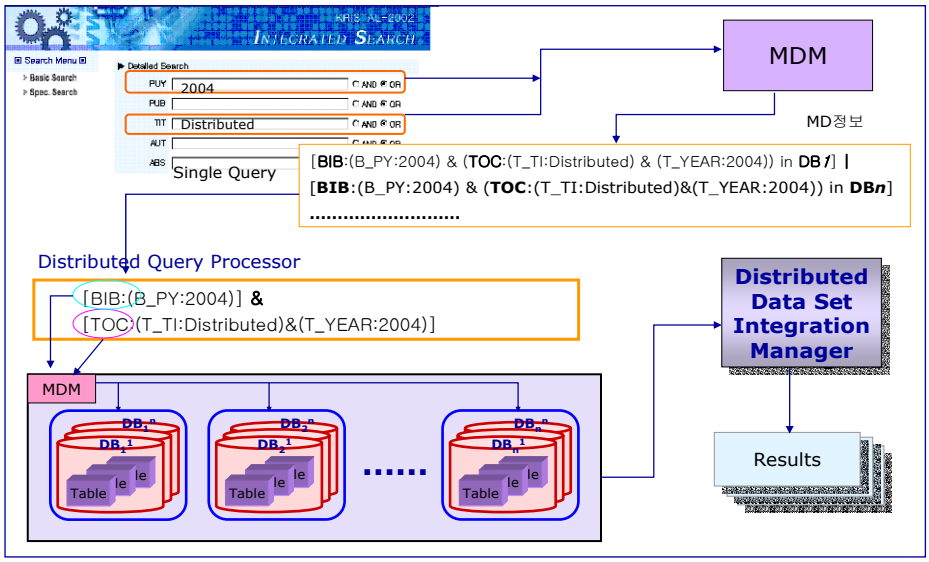

Fig. 12. Relationship between DQP and other modules in system.

compared to every target meta fields defined by the integrated server. However, in specific search, the user is able to determine which meta fields he/she wants to compare the query to.

When the actual fields of the source server is clicked, schema information of the source server is shown and the mapped fields for the given meta field are shown checked. By allowing multiple fields to be selected, we support ontology.

When an administrator clicks MetaField column, the detailed information of the meta field, is loaded from meta data registry, is shown on a different window. The information shown in Figure 6 is the subset of this information.

4.3.3 Source Server Monitor

Source Server Monitor monitors the status of the all the source servers participating in the integrated search service (Figure 11.). O mark in the status column indicates the source server is running normally. Otherwise, X mark will be shown. When the status of a source server is marked as X, by clicking checkbox on the leftmost column, the integrated server administrator can eliminate the selected server from the server list.

4.4 Integrated Query Processing

4.4.1 Distributed Query Processor (DQP)

Given the user query, DQP re-generates queries for source servers using schema mapping information and sends them to the corresponding source servers. DQP transforms user-input queries into Boolean or Vector queries. Figure 12. shows the relationship between DQP and other modules in the system. For the user-input entries,

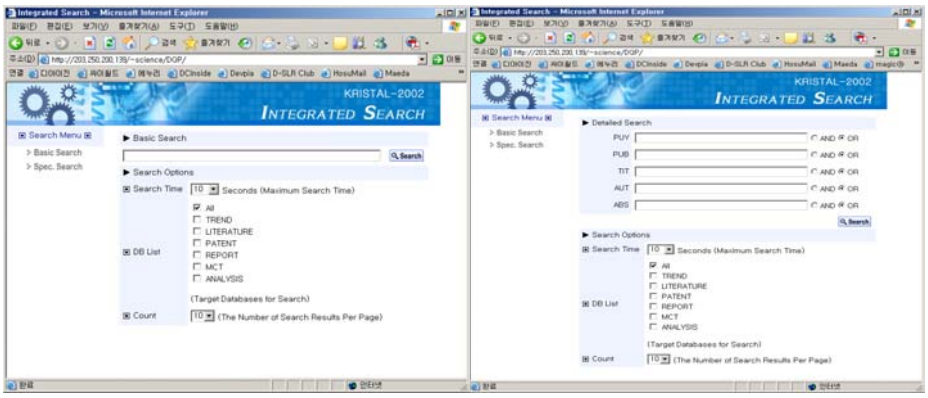

Fig. 13. Web interfaces for integrated search service.

DQP first read schema mapping information of each source server. If the source server has mapped the meta fields used for the integrated search into the actual fields of the database it manages, DQP extracts the table names and the actual fields that correspond to the user-input entries and regenerates new query as follows: *[table name: (actual field name: user-input entry)]*. If user-input entries are connected with AND or OR operations, then sub queries of the new query must be connected with the same Boolean operators. The following table shows an example.

1) AND operation [BIB: (B_PY:2004)] & [TOC:(T_TI:Distributed) & (T_YEAR:2004)]
2) OR operation [BIB: (B_PY:2004)] | [TOC:(T_TI:Distributed) & (T_YEAR:2004)]

4.4.2 Distributed Data Set Integrated Manager (DDSIM)

DDSIM is responsible for collecting and displaying the search results to users (Figure 12.). For the documents found by the source servers, it extracts those fields specified by schema mapping manager of the integrated server (See 4.3.2). Using the values, it generates HTML documents.

5. Prototype Implementation

We have implemented and deployed the proposed integrated information retrieval system. We have included six sources servers in different areas that are currently servicing independently: Science Literature DB, Scientific Technical Trend DB, Scientific Technical Report DB, Scientific Technical Analysis DB, Patent DB and Human Resource DB.

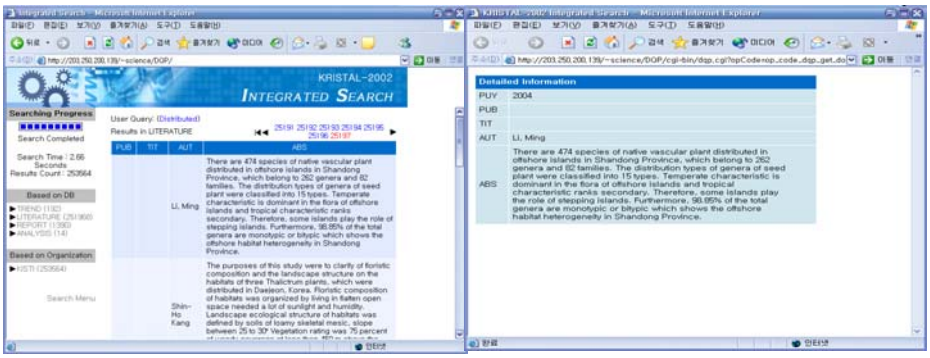

Fig. 14. Search results.

Two types of search services in the current implementation are provided: Basic Search and Specific Search (Figure 13.). In basic search mode, users enter query words in a single text area of the web page. Users are able to select target source servers, the number of documents per page and the maximum search time. The user-input entries are compared against all of the meta fields marked for BasicSearch (see 4.3.2). When there are multiple meta fields for basic search, they are connected with OR operators. Unlike basic search mode, in specific search mode, users are allowed to select target meta fields that will be compared to the user-input entries. These target meta fields are those meta fields that have been marked for SpecSearch by the integrated server administrator (see 4.3.2). In addition, the Boolean connectors also can be determined by users.

Figure 14. shows the search results. Once the searching process completes, DDSIM shows only the summarized information of the selected documents. If the user wants more detailed information, he/she clicks the document and can obtained detailed information. The actual fields of source servers used for displaying documents are defined dynamically by the source server administrator (see 4.3.2)..

6. Conclusion

In this paper, we proposed an integrated information retrieval system that does not require modification of existing information sources. The proposed system consists of Source Server Manager, Integrated Server Manager, MetaData Registry Manager, Distributed Query Processor and Distributed Data Set Integrated Manager. We have implemented and deployed the proposed system and tested and verified it using six source servers that use different database schema.

Future work includes developing a system that does not assume that participating source servers are running the same database system.

References

[1] Z39.50 Gateway, http://www.loc.gov/z3950/gateway.html.
[2] ANSI/NISO Z39.50 Revision, January 2002. http://www.loc.gov/z3950/agency/revision/revision.html.
[3] http://giis.kisti.re.kr/download/k-protocol-03-02-14.pdf.
[4] D. Brickley, R. Guha, eds. Resource Description Framework Schema Specification. W3C Proposed Recommendation, March 1999.
[5] Metadata Registry, http://metadata-stds.org/11179.
[6] K. Beard, T. Smith. A Framework for Meta-Information in Digital Libraries, In Sheth A, Klas W (eds) Multimedia Data Management: Using Metadata to Integrate and Apply Digital Media. Mc Graw Hill: 341-365, 1998.
[7] Online Computer Library Center, Inc. 1997 Dublin Core Metadata Element Set: Reference Description. 1997. Office of Research and Special Projects, Dublin, Ohio. http://www.oclc.org:5046/research/dublin_core.
[8] Ontology, http://www.w3.org/2001/sw/WebOnt.
[9] D Calvanese, G. Giacomo, and M. Lenzerini. Ontology of Integration and Integration of Ontology, Proceedings of the 2001 Description Logic Workshop (DL 2001), 2001.
[10] D. Calvanese, G. Giacomo, M. Lenzerini, D Nardi, R. Rosati. Description Logic Framework for Information Integration, Proceedings of Principles of Knowledge Representation and Reasoning (KR'98), 1998.
[11] A. Farquhar, R. Fikes and J. Rice. The Ontiliqua Server: A Tool for Collaborative Ontology Construction. International Journal of Human-Computer Studies, 1997.
[12] D. Egnor and R. Lord. Structured Information Retrieval using XML, ACM SIGIR Workshop on XML and Information Retrieval, 2000.
[13] M. Kobayashi and K. Takeda, Information Retrieval on the Web, IBM Research Report, RT0347, 2000.
[14] Nishizawa Itaru, Takasu Atsuhiro, Adachi Jun. A Schema Integration and Query Processing Information Retrieval, IPSJ SIGNotes Database System Abstract, No.094009, 1993.
[15] Intelligent Integration of Information, http://mole.dc.isx.com/13.
[16] Knowledge Sharing Effort, http://www-ksl.stanford.edu/knowledge-sharing.
[17] Robert M. Losee and Lewis Church Jr., Information Retrieval with Distributed Databases: Analytic Models of Performance, IEEE Transactions on Parallel and Distributed Systems, Vol.14, No. 12, 2003.
[18] Soumen Chakrabarti, Martin van den Berg and Byron Dom, Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery, Elsevier Science, 1999.
[19] Oren Zamir and Oren Etzioni, Web Document Clustering: A Feasibility Demonstration, SIGIR, 1998.
[20] Hsinchun Chen, Haiyan Fan, Michael Chau, and Daniel Zeng, MetaSpider: Meta-Searching and Categorization on the Web, Journal of American Society for Information Science and Technology, Vol. 52, No. 13, 2001.
[21] Luis Gravano, Chen-Chuan K. Chang, Hector Garcia-Molina, STARTS: Stanford Proposal for Internet Meta-Searching, Proceedings of ACM SIGMOD, 1997.
[22] KRISTAL-2002 User's Manual V1.1, KISTI, 2002.