Information Retrieval/Integration Architecture for Heterogeneous Data with Situation Awareness

Jongwook Woo
Computer Information Systems Department
California State University Los Angeles
Email: jwoo5@calstatela.edu

ABSTRACT
It is not easy to retrieve and integrate heterogeneous data which is non- or semi-structured. Besides, the data integration systems need to be alerted when the data is changed. The paper proposes the model with which the buyer/user searches and aggregates organizational entity from multiple heterogeneous data sources in order to have a contract with the proper entity for the new project. The model supports a service that alerts the update of the entities to the user.

Keywords: information integration, heterogeneous data sources, situation awareness, information retrieval, search

1. INTRODUCTION
In order to implement IT solutions, people has manually researched and aggregated the profiles of the possible small and large business contractors. The profile includes the relevant business licenses, organizations’ structures, and data related to CMMI, Air Force SEP, and ISO 9000 etc. However, the current method involves manual search, manual aggregation, and the data must be constantly refreshed, which is labor intensive and time consuming processes. Therefore, we need a web application that can continuously search for information sources, aggregate and rank the results in a persistent database, and alert the users the change of the interesting entities at SA situation to provide an up-to-date profile of organizational entities, which will reduce the labor intense and avoid government over-commitment and abandonment of processes during an emergency, quick-response situation.

The paper illustrates the possible design and approach to implement this web application that searches, aggregates, manages data and profile qualifications for IT solutions.

In order to measure the qualification of an entity to have a contract, we can use CMMI (Capability Maturity Model Integration), the Air Force System Engineering Process (SEP), ISO 9000. However, it is labor intensive to search and aggregate data from heterogeneous data sources to asses an organizational entity - the heterogeneous data sources can be both structured (Database, XML and JSON files) and non-structured (plain files and web). It becomes worse when the user monitors the status of the interesting entities continuously over time, especially at the moment of situation awareness. Therefore, we need a web application that can continuously search for information sources, and aggregate and rank the results in a persistent database to provide an up-to-date profile of organizational entities.

The paper is composed of the following sections. Section 2 illustrates related works. Section 3 presents background such as system engineering, information integration and retrieval. Section 4 shows the proposed model. And, finally, section 5 is the conclusion

2. RELATED WORK
Naveen et al [31] implement Ebox system that provides integrated access to multiple heterogeneous data sources relevant to providing situational awareness during emergency response situations. The user of the system can receive the useful information from the multiple sources at the emergency situation.

Woo et al [16-18] design an information integration model on N-tier architecture with a global XML schema for a specific domain, which is a format that each heterogeneous data source uses to generate XML data to be migrated to a global data source. Woo et al [25, 26] also illustrates and presents a possible e-Business architecture integrated with an enterprise search engine.

Mashup is an application that provides a unique service by integrating data from multiple web sources. It is mostly web applications based on Web Service or SOA (Service Oriented Architecture). There are many commercial mashup tools that support user friendly GUI such as Yahoo Pipe, Microsoft Popfly, and Google Mashup editor etc [11-14]. You can use Yahoo search engine to search at youtube.com by using the RSS feed data for the top 10 songs from apple.com [14].

3. BACKGROUND
Situation Awareness (SA) is defined at Wikipedia as the perception of environmental elements in time and space, the comprehension of their meaning, and the projection of their status in the future [30]. In this paper, SA is defined as the situation that the user is alerted when the profiles of the interesting organizational entities are changed. The background of the processes can be described in the following subsections as follows. First, system engineering approaches - CMMI, SEP, and ISO 9000 - are presented. Second, data migration and integration concepts are explained. Finally, search technologies for structured and non-structured data are illustrated.

A. System Engineering Approaches
There is a standard to measure the qualification of the businesses who can implement IT solutions. The system engineering approaches that are used to profile the
Capability Maturity Model Integration (CMMI) is a process improvement approach that provides organizations with the essential elements for effective process improvement in software engineering and organizational development. Thus, it guides process improvement across a project, a division, or an entire organization. An organization can be appraised not certified in CMMI so that it can be awarded a maturity level rating (1-5) or a capability level maturity profile. By the appraisal, the external customers and suppliers can be informed how well the organization’s processes to CMMI best practices. Besides, customers can use the appraisal for the contractual requirements [1-3]. We can refer to the web site of Software Engineering Institute at Carnegie Mellon University, which lists the appraisal results of many organizations [4].

System Engineering Process (SEP) is a process that applies system engineering techniques - technical management and problem-solving processes - to develop all kinds of systems. SEP has a meta model that is composed of four processes: Agreement, Project, Technical and Evaluation. Agreement process is to establish an agreement with the customer, in order to build a new system. Project process is to plan the project and be modified during technical process. Technical process designs and develops the project. Finally, evaluation process is to repeatedly check out whether the requirements are met, valid, and consistent [5, 6]. DoD system engineering is composed of 8 Technical Management Processes and 8 Management Processes. Technical Management Processes are equivalent to the Systems Analysis and Control portion to support and control the application of the Technical Processes during system development in order to meet program or project objectives [6].

International Standard Organization (ISO) 9000 addresses Quality Management families that consist of standards and guidelines relating to quality management systems and related supporting standards. The families fulfill the customer's quality requirements and applicable regulatory requirements while meeting customer satisfaction and achieving continual improvement of its performance in pursuit of these objectives. ISO 9001 is one of the families, which illustrates the requirements of quality management system. Proper quality management improves business, create a more efficient, effective operation, enhance marketing, and increase customer satisfaction. ISO 9001 sometimes makes people concern because of the amount of money, time and paperwork required for its registration [7, 8].

B. Data Migration and Integration

In order to integrate data from multiple heterogeneous data sources, we need two processes: data migration and data integration. Data migration is to retrieve and collect data from the data sources and store them into the third data source in a format specified. There are many issues to resolve for it. First, we must know the data sources. If data sources are at the same organization, we should connect data sources with data source connection information by a security authentication and authorization. If data sources are at the different organizations, it will become more complicated. Thus, people contract and share data as feed data such as in RDF and RSS format, which are XML data. Or, we need to search the web sites of related data, collect the relevant data, and transform them to the desired format [17-19].

Once we have data collected, we need to integrate data into our DB by converting them to the proper format of the DB. Data integration processes are 1) to check out if it exists in our DB and then update data and 2) to remove or combine duplicated data from the heterogeneous data [17, 18]. SA situation occurs when the data is updated or deleted. It can be simply implemented with email and Short Message Service (SMS) that sends text message to phone.

C. Search Engines

There are two kinds of search engines to search and retrieve data depending on the target data sources. One is enterprise search engine that searches local DBs, for example, Solr and Fast. After searching the data sources, the engine indexes data specified and stores indexed data into the repository of the engine. Another is web search engine that searches web sites.

Lucene is one of open source apache software projects supported by the Apache Software Foundation. It is the text search engine library in Java [19]. Solr is an open source enterprise search engine on the Lucene APIs. Solr runs on servlet engine and has many features such as hit highlighting, faceted search, caching, replication, and a web administration interface with XML/HTTP and JSON APIs [20]. FAST [23] is an enterprise search engine and a product of Microsoft. It supports geographical search function as well as all features of Solr. LocalLucene is to support geographical search capabilities to Lucene. LocalSolr provides the geographical search capabilities by using LocalLucene APIs [21, 22].

The major web search providers are Google and Microsoft/Yahoo’s Bing. The search providers support their search APIs to users so that the users can implement their web search application with the APIs. Besides, Apache has open source Nutch web search engine that can be integrated into Solr [24].

4. PROPOSED MODEL

A. Objectives

The paper proposes the model that searches and aggregates organizational entity from multiple heterogeneous data sources. Besides, when the profiles of interesting entities are updated, the user should be alerted. The model needs several processes: 1) selecting the data sources that have the profiles of entities, 2) migrating and integrating data from the data sources, 3) searching for the proper entities over the integrated data source, and 4) the user can keep the interesting entities and be alerted by email and SMS when the entities are updated. We assume that we have the information of the data and data sources that we can access. Thus, the selecting process 1) is based on the algorithm and use cases that follow the way to
research and aggregate data manually to assess an organizational entity.

B. Implementation

Figure 1 illustrates the proposed architecture. It is mainly composed of web search engine, data migration/integration service, enterprise search engine, and web application service. We implement the model in open source tools and APIs because the tools are free and proven in the market. Apache Solr and Nutch search engines are used in many companies such as AOL, CNET reviews, CitySearch.com etc.

B.a) Implementing Data Migration/Integration Function

Web Search Engine is needed to collect non-structured data, for example, html documents, from web sites. The search engine periodically crawls through, searches for, and indexes data of web pages at the web sites given. We can use Apache Nutch [24] engine for web searching by specifying the web sites given by buyers of data. And, in order to retrieve and store the data from the local data source from the repository of the search engine, we need to implement and provide Data Conversion and Parsing Service. After the periodic web data search and indexing, the service retrieves data of the repository by queries. And, the service parses and converts collected data to the desired data format for data integration and stored into local data source DS0. We combine Apache Solr [19-22] search engine and Nutch so that we can use Solr’s search APIs in order to collect, parse, and covert data, then finally store converted data into local data source DS0.

Data Migration and Integration service retrieves heterogeneous data periodically from multiple data sources. And, the service parses and converts the data to the specified format by using the dictionary and mapping service. Then, the data in the format needs to be merged into the Global Integrated DB.

For example, data of DS1 has data format \{local ID1, column 1, column 2, ..., column n\}. And, data of DS2 has data format \{local ID2, column 21, column 22, ..., column 2m\}. And, global data has data format \{global ID, list of local IDs, column 1, column 2, ..., column n\}. The columns of data DS1 can be mapped to the columns of the global data as follows: \{column 11 -> column 1, column 13 -> column 2, ..., column 1n -> column n\}.

Figure 3.2 illustrates how to integrate profiles’ data of an organization into DS1 from DS1 and DS2. IDs of Tables 1 and 3 are mapped to Local IDs of Table 2. Columns name, city, and state of Tables 1 and 3 are mapped to the correspondent columns of Table 2 respectively. CMMI Maturity level of Table 1 is mapped to Maturity column of Table 2. And, ISO 9001 certified column information of Table 3 is mapped to Certified column of Table 2.
B.b) Implementing Search Function

Once we have integrated data at the global integrated DB as shown in Figure 2, we need to index the columns for relevant search, which is related to machine learning algorithm. We use Solr [20] enterprise search engine to search for the proper organizational entities based on dynamic machine learning search algorithm. The developer (or administrator) of Solr needs to define indexable columns to index and specify the relevancy weight of the columns.

For example, if Table 3 has the data {00012123, {1111: DS1, 123:DS3}, “ABC Inc”, “Category: IT”, “310-340-xxxx”, “Santa Monica”, “CA”, ‘5’, ‘Y’}, we can specify indexable columns for category “Category: IT”, CMMI Maturity level ‘5’ and ISO 9001 certified ‘Y’. And, we could give weight 50% to the category, 75% to CMMI Maturity Level, and 75% to ISO 9001 certified columns respectively. Then, when the search engine display the search results, the highly relevant search results will be displayed first. Once we have more information how to retrieve and sort data from the buyer, we can define the relevancy as required. Besides, we can also implement dynamic relevancy – machine learning algorithm - that dynamically update the relevancy by giving the different weights to the columns from the information from the buyer.

As shown in Figure 1, Our Searchable Data Generator Service periodically retrieves data from the global DB and generates XML files. And, it posts XML files to the Solr search engine and the search engine indexes the XML elements as specified above. Then, the indexed data are stored at the repository of the search engine. XML file looks as shown in Figure 3. Besides, Data Alert Service is for SA, which keeps the user’s contact information and interesting data. When the profile data is updated, the service sends an alert to the user.

Our web site has search functions that are built with Solr search APIs. Thus, when the user searches for organizational entities by queries, the APIs request for search results. And, the search engine looks for the entities’ information at the repository then it sends the response back to the user with the relevant search results that are displayed on the user’s web browser. For example, Figure 4 shows the example GUI view for search. The user chooses “>=3” for CMMI maturity level that is greater than 3, “Yes” for ISO 9001 certified, and “IT” for category in order to search for entities that are good fit to the queries.

B.c) Use Cases

```xml
<docs>
<doc>
<field name = "id">00012123</field>
<field name = "name">ABC Inc</field>
<field name = "zip">90030</field>
<field name = "certified">"Y"</field>
</doc>
<doc>
<field name = "id">123</field>
<field name = "name">ABC Inc</field>
<field name = "zip">90040</field>
<field name = "certified">Y</field>
</doc>
...
</docs>
```

Figure 3: XML Example for Solr

In order to build the systems shown in Figure 1, we need several use cases that direct what to implement. The use cases with the services are illustrated as follows:

**Goal:** A user searches for and assesses the profiles of entities at the web site to find the proper contractors for IT solutions, which are collected from the multiple heterogeneous data sources.
Assumption:
(1) The buyers give us the steps and processes to search and retrieve the profiles of the organizational entities.
(2) Data sources include web sites that need to be searched for by web search engine.

Use Cases for Web Search Engine:
(1) Web Search Engine periodically crawls through, searches for, and collects the entities’ web sites given by buyers.
(2) Data Conversion/Parsing Service converts the collected data to the proper format and stores them to the data source.

Use Cases for Data Integration/Migration:
(1) Data Integration/Migration Service periodically retrieve data from the multiple data sources – including data source at the web search engine.
(2) The service combines by following the mapping information and stores the data to the global integrated DB.

Use Cases for Data Indexing:
(1) Searchable Data Generator Service periodically generates XML files from the global integrated DB.
(2) The service posts the XML files to the enterprise search engine.

Use Cases for Situation Awareness:
(1) Data Alert Service keeps the user’s contact information for the user’s interesting data.
(2) Data Alert Service sends an alert to the user when the user’s interesting data is updated.

Use Cases for User:
(1) The user opens the web site and logs in
(2) The user chooses the queries from the drop down menus given and clicks on submit button.
   a. Or/and the user types in the queries at the text field and clicks on submit button.
(3) The site moves to new page that displays the search results that show the profiles of the entities by relevant order.

5. CONCLUSION
The paper presents the model that integrates heterogeneous data sources and retrieves information from the integrated database in order to find the proper IT businesses that satisfy the system engineering standards. Besides, the profile data of the businesses are updated periodically so that the model has the function to update the data on SA. It is to assist buyers to easily search for the proper vendors who are qualified for the buyers’ projects.

The model is built in open source Solr/Nutch search engines to collect web and local data and J2EE. And, the data is collected in XML and integrated into local DB. The model is implemented on N-tier web architecture so that the buyers can conveniently find the proper vendors.

REFERENCES
[14] “Yahoo Pipe RSS Feed of Youtube and Apple Songs Top 10”, Jongwook Woo,
http://pipes.yahoo.com/pipes/pipe.run?_id=4bd90f50cbbfd0a0c6494b549c5a80e00&_render=rss


Jongwook woo is currently an Associate Professor at Computer Information Systems Department of California State University, Los Angeles. He received the BS and the MS degree, both in Electronic Engineering from Yonsei University in 1989 and 1991, respectively. He obtained his second MS degree in Computer Science and received the PhD degree in Computer Engineering, both from University of Southern California in 1998 and 2001, respectively. His research interests are Information Retrieval/Integration/Sharing, Map/Reduce algorithm on Hadoop Parallel/Distributed/Cloud Computing, and n-Tier Architecture application in e-Business, smartphone, social networking and bioinformatics applications. He has published more than 40 peer reviewed conference and journal papers. He also has consulted many entertainment companies in Hollywood.