Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Science and Engineering

Elixir Comp. Sci. & Engg. 45 (2012) 8032-8033

ISSN: 2229-712X

Basic technologies for integrating data and information from disparate source

Gholamreza Jandaghi and Hamid Haji Molla Mirzaee
Faculty of Management, University of Tehran, Qom College, Iran.

ARTICLE INFO

Article history:
Received: 27 February 2012;
Received in revised form:
10 April 2012;
Accepted: 27 April 2012;

Keywords
Technology,
Information extraction,
Data cleansing,
Disperse source.

ABSTRACT

This article discus a bout basic technologies which normally are used for different integration solutions and approaches to integrating data and information for enterprise information systems. In fact we are trying to find out the techniques that might be utilized in every solutions and approaches in order to integrate data and information from disparate sources this article discussing a bout basic and core technologies that work as hart of data integration solutions. This article contains eight sections and each section describes one technology. The eight basic technologies are included; Information Extraction, Data Cleansing, Extensible Markup Language (XML), Schema Matching, Schema Mapping, Schema Standards, Web Dynamic Technologies, Keyword Search.

© 2012 Elixir All rights reserved.

Introduction

Information extraction [1] is included a set of methods which create structured information from free form text. Interested concepts are extracted from document collections by using a set of annotators, that might either is custom code or particularly created extraction policies that are translated and compiled by an information extraction system. In some scenarios, when enough considered training information is accessible, machine learning techniques might also be used.

The essential tasks contain named entity identification for example to recognize people name, places, jobs and companies or affiliation extraction for example in the healthcare system identifies the patient's phone number or patient's address.

While a text fragment is identified as a concept then the fact will be recorded by enclosing it with XML tags which recognize the concept with inserting an entry in an index or by duplicating the values into a relational table.

The effect is better constructed and organized information which can more simply be merged with other data thus supporting integration.

Data Cleansing

The act of detecting and removing or correcting a dirty data in the databases is called data cleansing. The data that is incorrect, out-of-date, redundant, incomplete or formatted incorrectly often called dirty data. The main idea and main objective of data cleansing is not only the cleaning up the data in a database but bringing consistency to different sets of data that have been merged from separate databases is the goal of data cleansing.[2]

Extensible Markup Language (XML)

In all approaches mentioned before the integrated view of information from all sources should be generated. Each data element with the semi structured format will be tagged, Thus the entire element which their values are known necessary to be contained.

This capability to manage disparities of information content drives EII systems to test with XML.[3] the flexibility of XML causes an attractive format for integrating data through systems with different illustrating of data.

In some integration approaches it might not require to identify a frequent scheme and data will be combined into single self describing XML document from two resources Although it is necessary in some of the approaches such as warehousing to convert the original data into a well identified format. The XML is very flexible and very smart technique in the approaches that very common and applicable to transfer data and information between systems and also work as format to store data.[4]

Schema Matching

Schema matching is the process of identifying two objects that are semantically related for instant in two schemas DB1.Student (Name, SSN, Level, Major, Marks) and DB2.Grad-Student (Name, ID, Major, Grades); possible matches would be: DB1.Student ≈ DB2.Grad-Student; DB1.SSN = DB2.ID .

Actually large schemas have more than thousand elements, offering a major problem for a schema mapping tool. To map an element of Schema 1 into a possible match in Schema 2, the designer might have to scroll through dozens of screens. To avoid this boring process, the tool might suggest a schema matching algorithm,[5] which uses heuristic or machine learning techniques to discover possible matches based on whatever information which it has accessible for example similarity of name , data type, structure , an externally supplied glossary, or a library of previously matched schemas. Then the human user should authenticate the match. Schema matching algorithms do well at matching individual elements with somewhat similar names, for example " Salary_of_Employee" in the schema 1 and EmpSal in schema 2 or when it matches predefine synonyms, such as Salary and Wages. Some of the methods leverage data values.

But matching algorithms are unproductive when there are no hints to use.

Schema Mapping

Schema mapping is the process of transforming between two objects that are semantically related for example in the two schemas DB1.Student (Name, SSN, Level, Major, Marks) and

Tele:
E-mail addresses: jandaghi@ut.ac.ir

© 2012 Elixir All rights reserved

DB2.Grad Student (Name, ID, Major, Grades); possible transformations or mappings would be: DB1.Marks to DB2.Grades (100-90 A; 90-80 B..). A basic process for all data integration approaches is to identify how a source database schema relates to the target integrated schema. Schema mapping Tools usually present three vertical panes.

The left and right panes display the two schemas that to be mapped and the center pane is where the designer identifies the mapping, typically by drawing lines between the correct elements of the schemas and explaining the lines with the required transformations.[6], [7]

Once the mapping is identified, many tools able to create a program to transform data conforming to the source schema into data conforming to the target schema.[8] in the ETL engine, the tool may create a script in the Engine's scripting language. In the EII system, it may create a query in The query language, for example SQL. In the EAI system, it may transform XML documents from a source message format to target format. In the object to relational mapping system, it may create a view that transforms rows to objects.[9]

Schema Standards

One of the core technologies for integrating data and information from disparate sources is to use the standard schemas which means that integrating become very easy if different sources use same schema. So by using same schemas no need to reformat data before integration. And moreover it certifies that all sources have common meaning. Every source can be related together by using the common standards

Even though sources do not verify to a common scheme So for example two sources might be connected together by creating two maps that are related to the standard. This approach only enables integration of information that materializes in the standard, and because a standard is frequently a least frequent denominator, some information will be lost in the composition.

There are many schema standards technology. [10], [11], [12] some of them are oriented through general kinds of data, for example software engineering information or geographic information. The others that related to particular application domains for example medical billing, news stories, computer-aided design. While the schema standard is abstract and concentrates on making taxonomy of terms, it is normally named ontology. Ontology is frequently utilized as controlled and organized vocabulary.[13], [14].

Dynamic Web Technologies

When we apply portal for integrating data it typically requires to be dynamically created from databases which exist in backend servers.

The development of Web technologies is made to access data and information more easier. Particularly the dynamic web technologies includes web services and Really Simple Syndication (RSS) feeds with regarding that many sites along with many sites proposing XML for integrating their data.[15] Development technology also is evolved by quick improvement of languages, runtime libraries, and graphical development frameworks for dynamic creation of Web pages.

Mashup is one of the common approaches for integrating dynamic content that is a Web page which merges information and Web services.

For instance two functions can be proposed by service for presenting maps one to present a map and another to add a glyph which marks a labelled situation on the map. So it might be applied to build a mashup which presents a list of stores and their places on the map. To reduce the programming attempt of composing mashups the frameworks are now materializing to provide a layer of information integration analogous for EII systems, but which are adapted to the new environment that is named "Web 2.0" .[16]

Keyword Search

Key word search is a method of filing and locating information through the use of keywords that describe the content of records [17]. keyword search is one of the techniques for finding information from different sources in more enhance approaches the document that supposed to be search will be resided in several repositories like as digital libraries or content stores where it may not be crate a single index. Some times federated search may be applied to search every store separately one by one and combine the result [18] Keyword search can even be utilized for structured data to get a fast sense for what is available and put the step for more accurate integration.

Conclusion

In fact we have many approaches and solution that has multiple basic technologies that might be usable for all approaches but for applying the basic techniques we need to identify the problem and distinguish the condition then we are able to solve the problem and integrate data and information from disparate sources.

References

1. Andrew Mc Callum, Information extraction: Distilling structured data from unstructured text,2005.
2. http://www.webopedia.com/TERM/D/data_cleansing.html
3. Bertram Ludäscher, Yannis Papakonstantinou, Pavel Velikhov Navigation-Driven Evaluation of Virtual Mediated Views,2000.
4. http://www.informationweek.com/news/20900153
5. Erhard Rahm1, Philip A. Bernstein. A survey of approaches to automatic schema matching. 2001.
6. Renee J Miller, Laura M Haas and Maurico A Hernández. Schema mapping as query discovery2000.
7. Lucian Popa, Yannis Velegrakis, Renee J Miller, Mauricio A Hernández, and Ronald Fagin. Translating Web data 2002.
8. Laura M Haas, Mauricio A Hernández, Lucian Popa, and Mary Roth. From research prototype to industrial tool, 2005.
9. Sergey Melnik, Atul Adya, and Philip A Bernstein. Compiling mappings to bridge applications and databases, 2007.
10. Health Level Seven International, http://www.hl7.org.
11. Advancing Open Standards for the Information Society, www.oasis-open.org/specs.
12. OMG Standards; www.omg.org/technology/documents/modeling_spec_catalog.htm.
13. Foundational Model of Anatomy, http://sig.biostr.washington.edu/projects/fm/
14. the Gene Ontology, http://www.geneontology.org/.
15. Michael J Carey, Data delivery in a service-oriented world, 2006.
16. Mehmet Altinel, Paul Brown, Susan Cline, Rajesh Kartha, Eric Louie, Volker Markl, Louis Mau, Yip-Hing Ng, David Simmen, Ashutosh Singh. A data mashup fabric for intranet applications. 2007
17. http://www.answers.com/topic/keyword-search
18. Weiyi Meng, Clement Yu and King Lup Liu, Building efficient and effective meta search engines, 2002.