Ontology-Based Query in Heterogeneous & Distributed Data Sources

NAJOOD AL-GHAMDI, MOSTAFA SALEH, FATHY EASSA

Faculty of Computing & Information Technology, King Abdulaziz University

najood@gmail.com , msherbini@kau.edu.sa, Fathy55@yahoo.com

ABSTRACT. Information sharing, exchanging and retrieving from heterogeneous data sources not only needs complete data accessibility, but also it needs solving data heterogeneity between these data sources. To solve this rising problems of heterogeneity, a lot of recent work has been done towards solving this issue. This research aims to develop a software system based on ontology to semantically integrate heterogeneous data sources such as XML and RDF to solve some conflicts that occur in these sources. We developed an agent framework based on ontology to retrieve data from distributed heterogeneous data sources. With this technique, user will be able to send a mobile agent with classical input query to a data source. Then the mobile agent will carry the needed module of the global ontology prepared by the user stationary agents and transport itself from the user site to a remote XML or RDF data source. At remote XML data source, stationary agents will transform the heterogeneous XML source into temporary local RDF ontology. The stationary agents in all sources perform a mapping between the local and global ontologies, convert the query to XML or RDF query, execute it and set the results in a suitable form. Finally, the mobile agent will return back with the results. A partial implementation of this framework has been carried out using some modules and libraries of Java, Aglet, Jena and AltovaXML.

KEYWORDS: Ontology, Ontology generation, Schema mapping, Multi-Agent System, Heterogeneous data sources, Mobile Agent.

.

1. Introduction

The Web contains abundant repositories of information that make selecting just the needed information for an application a great challenge since computer applications understand only the Web pages structure and layout and have no access to their intended meaning. To enable users get information from the Web by querying these heterogeneous data sources, the new trend is by using the Semantic Web technologies [11], [23]. The Semantic Web aims to enhance the existing Web with a layer of machine-interpretable metadata. The American Heritage Dictionary defines semantics as “the meaning or the interpretation of a word, sentence, or other language form” [11].

The emergence of the Semantic Web will simplify and improve knowledge reuse on the Web and will change the way people can access knowledge, agents will be a knowledge primary consumer. By combining knowledge about their user and his needs with information collected from the Semantic Web, agents can perform tasks via Web services [5] automatically. So agents can understand and reason about information and use it to meet user’s needs. They can provide assistance using ontologies, axioms, and languages such as DARPA Agent Markup Language which are cornerstones of the Semantic Web.

The layers approach for Semantic Web is presented in [24], where it mainly includes eXtensible Markup Language (XML) and Resource Description Framework (RDF). XML allows users to add arbitrary structure to their documents but does not state anything about what the structures mean. Meaning is expressed by RDF, which encodes it in sets of triples, each triple being rather like the subject, verb and object of an elementary sentence. The third basic component of the Semantic Web consists of collections of information is the ontologies. Ontology is the backbone of the Semantic Web.

The Semantic Web’s current focus is at the ontology and logic layers. However, work in the academic community on trust inference calculi across distributed information systems is ongoing and work on trust and the Semantic Web is beginning to appear [26].

The ontology consists of a specification of concepts to be used for expressing knowledge, including the types and classes of entities, attributes and properties, the relationships and functions, and constraints. It provides a share and a common understanding of a domain as a theory not a data structured container. Actually, there are different kinds of ontologies from lightweight to heavyweight for different purposes. Formal ontology supports data standardization to support interoperability which requires explicating all different representations and interoperations of the data in a particular subject domain. Therefore, it is difficult to create due to complexity and enormous details but once created it is easy to deploy.

Now, ontologies are a trendy research topic in various areas such as information integration, knowledge engineering, cooperative information systems, and natural language processing. In system integration, ontologies are playing an important role, mainly concerned with providing a set of mechanisms for resolving the semantic heterogeneity problems, resolving the queries, hiding the complexity of accessing data from different data sources, and describing the contents of all data source as concepts in a global ontology.

In queries resolving, the query is defined in terms of global ontology and a set of its pre-defined concepts such as relations and operations. At a data source, answering a query involves rewriting the query in terms of a local ontology and related concepts or new concepts must be added to the global concepts by merging mechanisms to offer a standard integrated structure.

To take advantage of standards based integration, many companies and institutes have been moving to XML but XML doesn't hold the meaning or the semantics of the data. A need to standardize business vocabularies of the increasing number of standard XML dialects, illustrate the need for a semantic transformation.

Semantics is possibly the most important vision in driving the Web to its next phase. Challenges in developing semantic techniques are applied in many fields such as AI, database, information system, data modeling, query and transaction processing, knowledge representation, etc. Those techniques propose new approaches for data integration. Semantic integration is considered to be the best framework to deal with the heterogeneity, enormous scale, and dynamic resources on the Web.

This paper presents a semantic framework to provide a query interface to users with a flexible accessing to remote heterogeneous data sources. An agent platform was selected as a framework for building an ontology-based search engine based on some available Semantic Web technologies. Using a global ontology attached with common vocabulary dictionary to solve the semantic heterogeneity, hide the complexity of retrieving information of heterogeneous data sources from the user and make the computing environment very flexible.

The remaining of this paper is organized as following: section 2 provides a background and a literature preview on important features of the paper framework. Section 3 presents an overview of our proposed framework and presents the system prototype module design. Section 4 discusses the framework implementation tools and problems raised during implementations. Also, presents a testing of the prototype by running examples and discussion. Finally, section 5 is directed to the conclusion and giving some suggestions for future work of this system.

2. Background and Literature Review

One of the difficulties in the Web information integration applications is the heterogeneity of the distributed data sources. These heterogeneities can be classified as syntactic, schematic, and semantic.

- **Syntactic.** Logical data models and their representations (relational, object-oriented) in the underlying DBMS [4]. For instance, XML and RDF provide two completely different paradigms for representing Web data. There are currently many attempts to use a conceptual-level schema (ontology) or a conceptual-level query language to integrate heterogeneous data sources [13].

- **Schematic.** Schematically the same real world phenomena can be represented with different abstractions in heterogeneous data sources. Within this heterogeneity range, different representations for equivalent data include:
    - entity vs. relation ;
    - entity vs. attribute ;
    - relation vs. attribute representations [4].
- **Semantic.** Semantics in linguistic is defined as the relationship between words and the things to that these words refer. Computer model's semantics is defined as the relationship among the computer representations and the corresponding real-world features within a certain context [4]. The semantic heterogeneity can by found in different ways, such the semantic problem with naming, when two terms represent same concept or one term represents two different concepts. Another problem is cognitive heterogeneity, when a fact or a real word object serves different purposes or have different views.

2.1. Ontology

An ontology is a conceptualization of an application domain in a human-understandable and machine-readable form, and typically comprises the classes of entities, relations between entities and the axioms which apply to the entities which exist in that domain [18]. A survey of Web tools [1] presented that extraction ontologies provide resiliently and scalability natively where in other approaches for information extraction, the problem of resiliently and scalability still remains.

One serious difficulty is creating the ontology manually is the need for a lot of time, effort and might contains errors. Also, requires a high degree of knowledge in both database theory and Perl regular-expression syntax. Professional groups are building metadata vocabularies or the ontologies. Large hand-built ontologies exist for example medical and geographic terminology. Researchers are rapidly working to built systems to automate extracting them from huge volumes of text.

The US Defense Advanced Research Projects Agency (DARPA) released a draft language known as the DARPA Agent Markup Language (DAML) [11]. The DAML language is an extension to XML and the Resource Description Framework (RDF). The language provides a rich set of constructs with which to create ontologies and to markup information so that it is machine readable and understandable. It leverages and extends the express-ability of RDF and RDF-Schema (RDFS) [18]. RDFS is a simple ontology language written in RDF that allows the creation of vocabularies with classes, properties, and subclass/super-class hierarchies [5]. DAML+OIL (DARPA Agent Markup Language + Ontology Inference Layer) is an extension of RDFS that allows finer-grained control of classes and properties with features such as cardinality constraints and inverses of properties [4].

2.2. Agent

An Agent is a computer program that acts autonomously on behalf of a person or organization. Agents come in several different types, usually depending on the nature of their environment. There are many ways to classify existing software agents e.g. by their mobility, by several attributes they have such as autonomy, learning and cooperation, by their roles, or hybrid agents.