Chapter 50
Semantic Integration Framework Based on Domain Ontology Construction

Jike Ge, Xianqiu Xu, Yongwen Huang and Mingying You

Abstract In this paper, we proposed a semantic integration framework of heterogeneous data based on domain ontology construction. The mechanism of domain ontology construction and mapping for heterogeneous data are studied, the purposes of which are to enhance the capability of dynamic adaptability and optimization of domain ontology construction, to resolve the problems of data heterogeneity in the processing of semantic information integration, and to promote the flexibility of the semantic integration process. Next, as to the semantic query, the theory and methods of specification or dynamic expansion of semantic query based on social annotation and ontology, and the duplicate removal and aggregating optimization of semantic query results are explored, the purpose of which is to realize the usability of semantic integration system and the credibility of the query results. Finally, an experimental prototype system of semantic integration of oil and gas exploration data based on domain ontology construction is constructed, the purpose of which is to verify the feasibility and correctness of the proposed theories and methods.

Keywords Semantic integration · Ontology construction · Domain ontology

J. Ge (✉) · X. Xu · Y. Huang · M. You
School of Electrical and Information Engineering, Chongqing University of Science
and Technology, Chongqing, China
e-mail: gjkweb@126.com; gejike@gmail.com

X. Xu
e-mail: xuxianqiu2011@126.com

Y. Huang
e-mail: lanf@tom.com

M. You
e-mail: youmingying@163.com

Z. Zhong (ed.), *Proceedings of the International Conference on Information Engineering and Applications (IEA) 2012*, Lecture Notes in Electrical Engineering 219,
DOI: 10.1007/978-1-4471-4853-1_50, © Springer-Verlag London 2013

50.1 Introduction

Due to the wide applications of different information systems, a large amount of information and knowledge is accumulated. The information and knowledge is in different formats, e.g., electronic documents, databases, and hardcopy documents, scattered in various systems such as product lifecycle management (PLM), enterprise resource management (ERP), and office automation (OA) systems. Integrating different information sources semantically is a growing research area within different application domains. However, reaching semantic integration is not an easy task. While the real world is assumed to be unique, its representation depends on the intended purpose: every representation of reality is user-specific. Thus, different applications that share interest in the same real-word phenomena may have different perceptions and therefore require different representations. Differences may arise in all facets that make up a representation: what amount of information is kept, how it is described, how it is organized, how it is coded, what constraints, processes, and rules apply, how it is presented, etc., [1, 2]. Thus, the problem of data integration emerges as a new research challenge. Data integration is becoming even more necessary given the increasing availability of data from distributed and heterogeneous sources, as experienced in the development of the internet and semantic web. Such characteristics make it difficult to search for desired information since queries might be inappropriately answered or may have incomplete results if each data source is analyzed in isolation.

Together with the concept of data integration, the term federated databases emerged during the 1990s to characterize techniques for providing integrated data access to a set of distributed, heterogeneous, and autonomous databases [3]. The work reported in Busse et al. [4] defines the classical layered architecture of federated systems based on Sheth and Larson [5], which is widely referred to by many researches. The federated layer is one of the main components currently under analysis and study. Its importance comes from its responsibility to solve problems related to semantic heterogeneity. Different approaches have been used to model this layer. They are as much diverse as complementary in some cases, and can involve different perspectives, such as the use of ontologies [6] or the use of metadata [7].

Ontologies, as considered by the computer science community, comprise elements such as classes, individuals, properties, and relationships [8], which can be used to model the semantics of the domain related to integrated data sources. Most frequently, work on ontologies aims at developing single-world ontology, i.e., an ontology that represents a given conceptualization of the real world from a given perspective. In our work, we are particularly interested in those approaches using domain ontology because they are introduced to facilitate knowledge sharing and reuse among various agents (software and humans) [9].

Following this premise, we propose a semantic integration framework of multisource heterogeneous data based on domain ontology construction, based on two main processes: semantic integration and query. Semantic integration is carried out integrating multisource heterogeneous data schema to domain ontology schema in order to improve the understandability of data. We propose a semantic querying methodology based on social annotation, in which the use of a set of matching functions and inferences over the ontologies allowing users to find more suitable mappings according to the users' social annotation and domain ontology.

50.2 The Proposed Data Integration Framework

The architecture of the proposed semantic integration framework of multisource heterogeneous data based on domain ontology construction system is shown in Fig. 50.1. In the framework, domain ontology is constructed by integrating multisource heterogeneous data. A user's interest ontology is generated by analyzing the user's demographic characteristics, personal preferences, and user's social annotation. An automatic retrieval specification and expansion method is utilized to categorize the information queried by a user. In the proposed retrieval specification and expansion method, the terms are determined by the domain ontology and user interest ontology base.

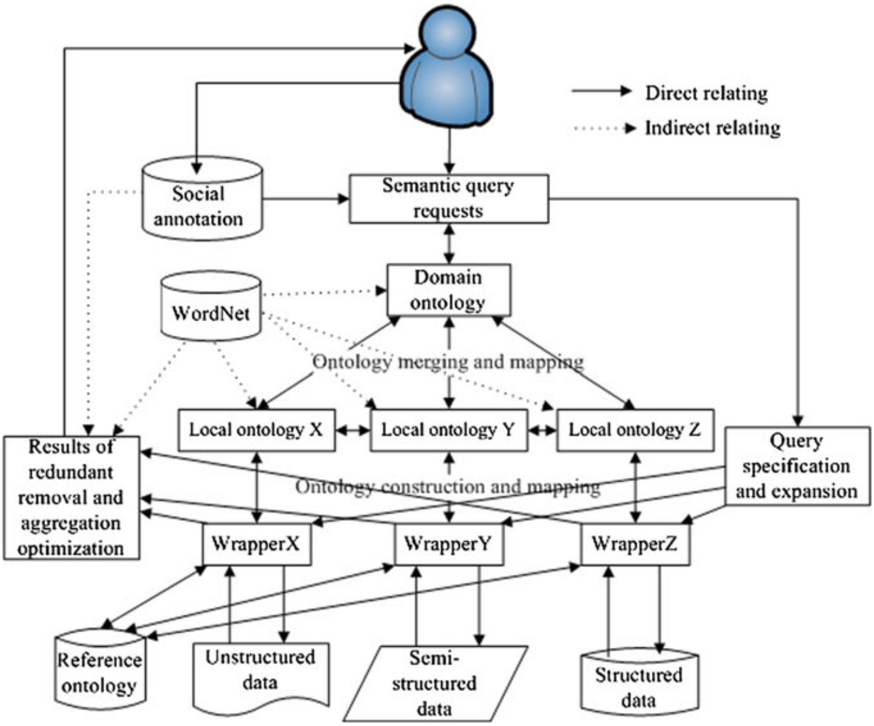

Fig. 50.1 The semantic integration framework based on domain ontology construction

50.2.1 Using Domain Ontology for Data Integration

Data integration systems (DIS) deal with two main problems: combining data located in different heterogeneous sources and providing the user with a unified view of gathered results [2]. By providing a unique, transparent and homogeneous view of heterogeneous data sources, it is possible to retrieve richer information, since different sources can have complementary data. In order to build an integrated view of data source s, some conflicts must be addressed, such as schematic and structural, and semantic ones [10].

To handle some of these conflicts, ontology provides a feasible methodology for semantic integration of the multisource heterogeneous information within the DIS. Specific domain ontology plays a key role in the data integration processing semantically. In this sense, domain ontology is constructed by integrating multisource heterogeneous information. The process of domain ontology construction includes two steps, namely, local ontology construction and merging local ontologies to domain ontology. The process of local ontology construction includes three aspects, namely from unstructured text documents, from structured relational data sources and from semi-structured data sources in XML files [11]. For structured relational data sources, the RDB2OWL mappings [12] can be regarded as documentation of the database-to-ontology relation. The RDB2OWL language reuses the OWL ontology structure as a backbone for mapping specification by placing the database link information into the annotations for ontology classes and properties. It features reuse of database table key information, user defined and table functions, as well as multiclass conceptualization that is essential for keeping the mapping compact in case of creating a conceptual partitioning of large database tables. For semi-structured data sources, XML2OWL [13] is a script to transform standard XML documents into neat OWL. The purpose of merging local ontologies to domain ontology is that building global domain ontology from some heterogeneous local ontology by using semantic distance-based ontology matching method [14].

50.2.2 The Implementation Processing of Proposed System

In our proposed system, we employ domain ontology in order to integrate multisource heterogeneous data sources. In addition to data integration, we combine social annotation with domain ontology [15] to perform semantic query expansions, retrieving approximate results that are relevant to user requirements. The operational flow of proposed system as follows.

The starting point is the query formulation in user interface. User types the query in a text box, once the query is submitted to the data integration system; it is analyzed to check possible syntactic errors and homonyms in relation to the domain ontology and user's social annotation. After the necessary corrections to the query, if they are fit for querying ontology directly, it will implement ontology matching, if not, it will continue to the query specification and expansion module. In this module, the query is modified if both domain ontology constructs and user-defined expansion parameters indicate that semantic expansions are needed. For further information on semantic query expansion procedures.

In the next step, wrapper receives the query, creates a thread for each wrapper and distributes the query to them. Each wrapper is responsible for converting the user query into respective queries that are specific to the corresponding data source. Moreover, wrappers perform the translation of the data returned from sources to a common model, previously defined by the domain ontology. All wrappers keep two important types of XML documents containing: information on how to connect to its respective data source; and mappings for associating each ontology concept with a corresponding term in the data source. These mappings are an essential element to support query translation, since mapping rules are written depending on the schema and the query language considered by a data source. During the query translation, wrappers handle some heterogeneity problems, such as naming, lack of data, attribute, value, and identifier conflicts. After the query is translated, it is submitted to each respective data source. As soon as the results are returned from the data sources, wrappers check the mappings and the reference ontology to verify if returned answers contain terms related to user's query. In this step, specific functions offered by the reference ontology are called by the wrappers for filtering the results and discarding those coordinates that does not relate to the terms in the query.

Once all results have been received, they are redundant removal and aggregation optimization, checking for equalities and gathering them in a single tabular result. Subsequently, it can be properly formatted and presented to the user, concluding the query flow.

50.3 Case Study: Integrating Oil and Gas Exploration Data Sources

To evaluate the feasibility of proposed semantic integration framework based on domain ontology construction in a real environment, a case study was performed in the domain of oil and gas exploration.

We have constructed general prototype architecture of proposed data integration framework, it shows as Fig. 50.2. All modules, which have been developed using open source technologies, are organized in a traditional client–server structure. Users access the system through a web browser. The web server application has been built in Java Server Faces with the ICE faces extension which is an Ajax framework that allows the development of rich internet applications in Java. The concurrency of the system (as multiple users may be accessing the application at the same time) is addressed by the framework using a thread pool

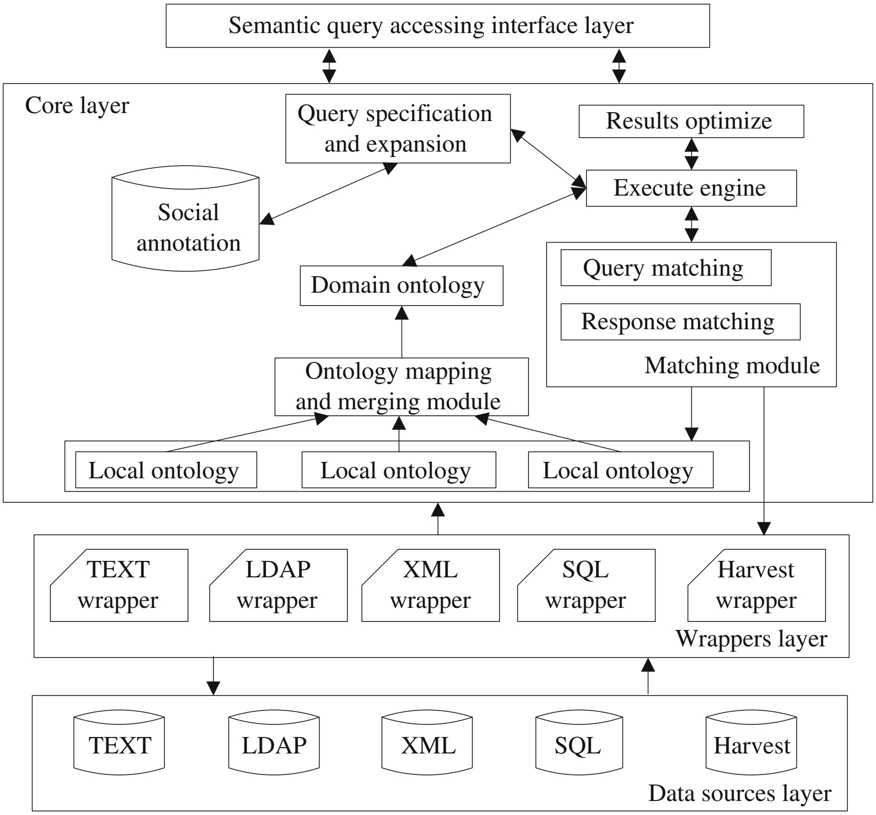

Fig. 50.2 The architecture of domain ontology construction based semantic integration

which provides bounded thread usage in large-scale applications. The core of the architecture is developed in Java, which handles the interaction between all the modules. Moreover, it manages the user's social annotation dynamically updating its state after each user action. This allows the system to take into account the behavior of the user and provide more accurate results. The Jena framework was also used to make inferences from the domain ontology written in web ontology Language.

Data sources containing the unstructured data, such as documents, the structured relational data sources (MySQL and SQL) and also semi-structured data sources in XML files.

The first step of system deployment is the construction of domain ontology, describing the semantics related to oil and gas exploration data sources. We have defined oil and gas exploration ontology with support of domain experts. The oil and gas exploration ontology was developed according to bottom-up approach and codified in OWL. Figure 50.3 shows part of this ontology.

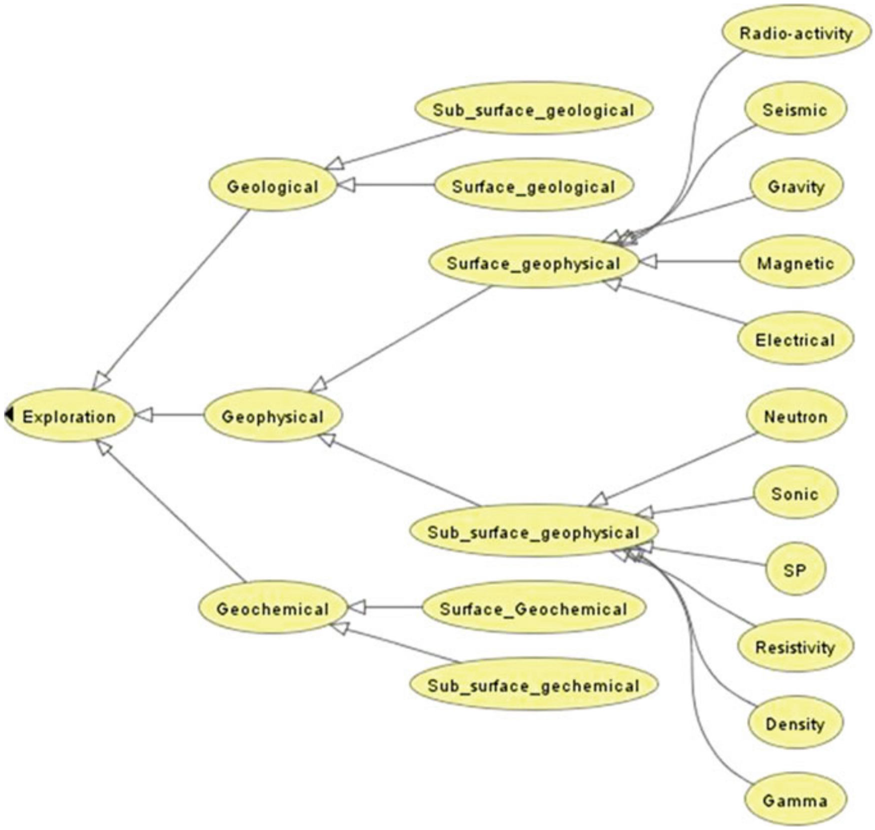

Fig. 50.3 The part of oil and gas exploration domain ontology

To illustrate the semantic query expansion and data integration processes involving oil and gas exploration data sources, consider query Q executed by researchers, who required information on the sub-surface geological and density in Sichuan basin:

$$\begin{aligned} Q:\ &\textit{sub-surface geological AND density AND exploration} \\ &= \textit{Sichuan basin} \end{aligned} \tag{50.1}$$

First step is analyzing Q, so that it can be rewritten in a common vocabulary language provided by the domain ontology. In this step, the domain ontology is analyzed to detect that density term can be associated with both density of gravity exploration and density of well logging concepts. In order to handle this ambiguity, the system interacts with the user so that one option is selected.

With regard to data integration, it is necessary to define mappings between ontology concepts and the data contained in the heterogeneous sources. The wrappers, specific to each data source, analyze those mappings in order to translate an ontology concept to the corresponding term in the data source. When a wrapper checks its mappings and they do not contain any of the terms of the query, it means that the database-specific query will not contain this term as well. In this case, the wrapper will not activate its respective database and will close the connection with local ontology.

Finally, results obtained from wrappers are sent back to results of redundant removal and aggregation optimization module, where result sets are merged. The integrated results are then presented to the user in a tabular format, including values for the original query as well as values for expanded terms.

50.4 Conclusions and Future Work

In this paper, we proposed a semantic integration framework of multisource heterogeneous data based on domain ontology construction, our framework supported integrated management of multisource information, providing a homogeneous view to access several heterogeneous data sources by constructing domain ontology. In addition, query expansions can also provide relevant retrieval results, it will support the researchers generating decision-making reports in less time. It will be our future works.

Acknowledgments The work described in this study is partially supported by the Natural Science Foundation Project of CQCSTC under Grant No. 2010BB2285, and the Research Foundation Project of Chongqing University of Science and Technology under Grant No. CK2010B06.

References

1. Bleiholder J, Naumann F (2007) Data fusion. ACM Comput Surv 41(1):1–41
2. Halevy A (2009) Information integration. In: Liu L, Özsu MT (eds) Encyclopedia of database systems, vol 5. Springer, US, pp 52–56
3. Hasselbring W (2000) Information system integration. Commun ACM 43(6):32–38
4. Busse S, Kutsche R, Leser U et al (1999) Federated information systems: concepts, terminology and architectures, vol 63. Technical report Nr. 99-9, Technical University of Berlin, Berlin, pp 25–27
5. Sheth AP, Larson JA (1990) Federated database systems for managing distributed, heterogeneous and autonomous databases. ACM Comput Surv 3(22):183–236
6. Buccella A, Cechich A (2007) Towards integration of geographic information systems. Electronic notes in theoretical computer science, vol 168, pp 45–59
7. Preece A, Hui K, Gray P et al (2002) Metadata integration assistant generator for heterogeneous distributed databases. In: Proceedings of international conference on ontologies, databases, and applications of semantics for large scale information systems
8. Uschold M, Grüninger M (1996) Ontologies: principles, methods and applications. Knowl Eng Rev 11:93–155

9. Fensel D (2003) Ontologies: silver bullet for knowledge management and electronic commerce, 2nd edn, vol 73. Springer-Verlag, Berlin, pp 15–27
10. Xue Y, Ghenniwa HH, Shen W (2012) Frame-based ontological view for semantic integration. J Network Comput Appl 35(1):121–131
11. Ge J, Li Z, Li T (2011) A context-based method for petroleum exploration domain ontology construction. Lecture notes in electrical engineering, vol 154, pp 683–690
12. Būmans G, Čerāns K (2011) Advanced RDB-to-RDF/OWL mapping facilities in RDB2OWL. In: Proceedings of perspectives in business informatics research, vol 90, pp 142–157
13. Lacoste D, Sawant KP (2011) An efficient XML to OWL converter. In: Proceedings of the 4th India software engineering conference
14. Ge J, Qiu Y (2008) Concept similarity matching based on semantic distance. In: Proceedings of 4th international conference on semantics, knowledge, and grid, vol 36, pp 26–27
15. Liu K, Fang B (2010) Ontology induction based on social annotations. Chin J Comput 33(10):1823–1834