Mediating Databases and the Semantic Web: A methodology for building domain ontologies from databases and existing ontologies

S. Zhao
Digital Ecosystems and Business Intelligence Institute
Curtin University of Technology
Perth. WA, Australia

E. Chang
Digital Ecosystems and Business Intelligence Institute
Curtin University of Technology
Perth. WA, Australia

Abstract - *Ontologies play a vital role in the Semantic Web. However, ontologies are hard to construct due to a high cost in the manual development process and difficulty in reaching agreement. While there are two valuable resources can be utilized to tackle the problems: relational databases and existing ontologies. Thus this PhD research proposes a methodology for semi-automated ontology development using the two resources. Firstly, an initial ontology of the knowledge embedded in a relational database will be generated through reverse engineering, mapping techniques and data mining techniques. Secondly, the initial ontology will be improved by ontology merging and alignment techniques on existing ontologies and through semantic and linguistic analysis of thesauri and dictionaries in order to gain agreement. A mapping between the evolved final ontology and the original database will also be established. This is designed for allowing queries from the Semantic Web to be made on the database instances via the final ontology model.*

Keywords: Ontology development methodology, ontology learning, relational database reverse engineering, semantic web technology, knowledge mining in database

1 Introduction

The advent of ontology has brought a very promising means of knowledge sharing and reuse over the internet. In the last decade, ontology-based technologies have been gaining an increasingly important position in many research areas such as artificial intelligence, knowledge engineering, system integration and the Semantic Web [1]. The Semantic Web is a vision of the current web in the future with its data component - ontologies. However, ontology development is costly in terms of time, resources and skills. There are crucial issues on ontology development still remaining though of many significant progresses that have been made in this arena. This research will focus on three key issues, which are described as follows:

Issue 1 High cost of developing ontology from scratch impedes wide adoption of ontology-based technologies

Building an ontology from scratch requires a great effort in terms of time, resources and skills. Particularly, a considerable effort goes into building the conceptualization of the domain knowledge and into acquiring knowledge from various sources. Most existing ontologies are built mainly from scratch and developed manually such as Cyc [2], SENSUS [3], although with computer aided tools such as Protégé [4]. Additionally, the process of knowledge acquisition is also conducted on a manual basis. This process is slow and tedious and has become a bottleneck of ontology development. In such manual bases, an ontology development may need to take two or three years. Therefore, an effective means of acquiring knowledge from available resources and an automated engineering approach for the developing ontology is in high demand for making the best of ontology-based technologies.

Issue 2 Difficulties in reaching agreement

An ontology is a shared conceptualization of a domain knowledge. Therefore, the concept definitions, relationships among concepts, and the concepts hierarchy modeled in an ontology must be agreed upon by the domain of interest. However, one can easily find that there is a divergence in existing ontologies. This includes those ontologies that model the same domain knowledge and generic ontologies that model common knowledge across all domains. This happens because there naturally exists differences among ontology developers in their knowledge perception, background, and the requirements of the ontology under development. It is also infeasible, as argued by [5], to enforce a single standard ontology even within an organization, as the lengthy process of reaching agreement on unified terminologies and as one standard ontology impedes changes. The problem becomes severe in cross-organization and open environment ontology development. Hence, effective mechanisms for gaining agreement and incorporating different views during the process of ontology development are crucial.

Issue 3 The gap between the Semantic Web and heterogeneous databases remains unfilled

The Semantic Web [1] is a web of data. Its realization relies largely on the underlying data component i.e. ontologies. These ontologies model the knowledge of specific domains in an explicit and sharable form that can be processed automatically by intelligent software agents. However, current state of ontologies development for the Semantic Web is still in its infancy in many aspects. The Semantic Web demands much more ontologies, which represent domain knowledge, in order to function properly as in its initiative. On the other hand, there is a large amount of relational databases that store intensive business information and embed domain knowledge existing over the Internet. The owners of these databases need their data to be sharable and searchable across organizational and application boundaries at a certain level. Such kind of request is driven by the constantly increasing collaborations between organizations. For instance, patient data from individual health care systems needs to be integrated in the health care domain to enable provision of a better quality of health care services; travel information needs to be shared amongst travel service providers and agents in order to provide integrated services regarding transportation, accommodation and so forth. However, the forms of data held in the databases are application and enterprise dependent. Thus prevents these valuable databases from participating in the Semantic Web directly. This gap between the databases and the need for data in a sharable format in the Semantic Web remains unfilled.

This research views the issues discussed above as opportunities of turning the ontology development into an semi-automated engineering process. Therefore we aim to tackle these issues by introducing the methodology which will be described in detail in later sections. Section 2 outlines the objectives of this research. Then followed by related work in Section 3; in Section 4, the methodology is described in detail with a draft model of the methodology; and section 5 concludes the proposal and indicates future work for this research.

2 Objectives

Considering the issues mentioned above, this research aims at developing a methodology for semi-automated domain ontology development from existing databases, thesauri and existing ontologies. This methodology will serve four purposes which include:

1) Reduce the domain ontology development effort by semi-automated engineering processes;

2) Reduce the knowledge acquisition effort for ontology development by transforming knowledge embedded in existing databases to web ontologies;

3) Incorporate different views of the same domain knowledge from existing ontologies and enrich the ontology under development by using thesauri and dictionaries; and

4) Mediate databases and the Semantic Web towards data sharing and integration purposes. This means that by applying the proposed methodology, the data held in databases can be shared, queried and searched directly by external applications in the Semantic Web context.

The proposed methodology is tailored to an information system where a relational database is available. It will contain four parts. Firstly, a data model will be extracted from a relational database using a combination of available techniques including database reverse engineering, structure mining and data mining techniques. Secondly, the data model is transformed into an initial ontology in OWL DL [6] using mapping techniques. Thirdly, the initial ontology model will be enriched and enhanced in two ways: one is through ontology merging and alignment of existing ontologies including generic ontologies (i.e. top-level ontologies) and domain ontologies and the other way is through semantic and linguistic analysis on thesauri and dictionaries. By using these techniques for enrichment of the initial ontology model, this methodology aims to eliminate the bias of the ontology generated from one particular database, thus providing an effective means for gaining agreement on the resulting ontology. In the fourth part, the methodology will establish and maintain a mapping between the resulting ontology and the original database. The mapping will specify the counterparts between the final ontology model and the original database. This is necessary as the evolved ontology is designed to serve as a conceptual model of the domain knowledge without ontological instances. Searching and querying that come from the Semantic Web and other external applications to the actual database instances will be enabled through the ontology model then, via the mapping to the original database. In other words, *the ontology instances will be generated at run-time based on requests*. This seeks to keep the applications built on the original database unchanged or with minimum change, while enabling the updated data instances in the original database will be used through the final ontology model. More details of the proposed methodology will be described in section 4.

3 Related Work

Prominent research on ontology-based technologies has been carried out since the early 1990s. The definition of an ontology has evolved as 'a formal, explicit specification of a shared conceptualization' [7, 8, 9]. This definition indicates the power of ontology that unambiguously and precisely captures and represents the intended meaning of knowledge for software agents

processing without prior communication. Besides, the representation must be consensual, which means its concept definitions, relations and concept hierarchies must be agreed within a community of interest. Furthermore, ontologies intend to model static domain knowledge at knowledge level, which is independent of any particular application or enterprise. Therefore, the knowledge presented by an ontology is sharable and reusable across those boundaries.

The infrastructure of ontology-based technologies has almost been established. This is evidenced by available frameworks and methodologies for ontology development [10, 11], by existing tools and implementation techniques for ontology construction such as Protégé [4] and JENA [12], and by the rich literature in the ontology research community. However, there are many open issues as discussed in the previous section that prevent this technology from becoming fully mature. Many studies towards these issues including ontology learning and ontology merging have been carried out to resolve these issues and are still ongoing.

3.1 Ontology Learning from relational database

Research on '*ontology learning*' aims to address the issues associated with manual knowledge acquisition for ontology development. According to Gómez-Pérez et al. [13], '*Ontology Learning*' is defined as '*the set of methods and techniques used for building an ontology from scratch, enriching, or adapting an existing ontology in a semi-automatic fashion using several sources*'. Other terms such as ontology generation, ontology mining and ontology extraction are also used to refer to the semi-automatic construction of ontologies. The sources have been used for ontology learning in previous approaches including text, dictionary, thesauri, knowledge base, and semi-structure data such as XML Schema and Relational schema. Natural language analysis techniques, machine learning techniques, reverse engineering techniques and mapping techniques are commonly used in this area.

Ontology learning from relational schemas aims to construct ontologies by (semi-)automated processes of acquiring knowledge from databases. Previous approaches to ontology learning from relational databases can be categorized into two groups. The first group is characterized as those that construct an ontology from a database schema then creates ontological instances from instant data in the database based on the previously constructed ontology. Approaches in this group include Kashyap [14], Meersman [15], Stojanovic, Stojanovic & Volz [16] and Astrova [17]. The techniques utilized in this group are reverse engineering from database schema to a logical or conceptual model and mapping technique that maps the extracted model into an ontology language. Key correlations are commonly used for identifying relationships between concepts. However, the key issue was that the knowledge embedded in data instance received no consideration.

The second group of approaches proposed mapping languages that directly map database instances into the Semantic Web data syntax such as RDF [18] and OWL [6]. This group includes approaches such as R2O [19] and D2R MAP [20]. The main drawbacks of this type of approaches are twofold: firstly, it ignored the fact that database instances will keep updating over time, as a result, constant synchronisation between the published ontological instances and the data in the original database is required; secondly, the meanings of the original data model such as relations, constraints and so forth are less reflected in the transformed ontological instances. However, these languages may be effectively used to populate ontological instances from databases when an ontology model of that database is available.

3.2 Ontology merging and alignment

According to Bruijn et al. [5] and Noy & Musen [21] '*Ontology Merging*' is defined as '*the processes of generating a new ontology from two or more ontologies*'. '*Ontology Alignment*' (ontology mapping or ontology relating, in the terms used by Bruijn et al. [5]) is defined as '*the process of establishing and specifying links or mappings between original ontologies, hence, the original ontologies are kept unchanged*'.

Based on Bruijn et al. [5], ontological re-engineering [22], ONIONS [23], and PROMPT [24], the generic phases of ontology merging and alignment can be summarized as follows:

- *Input of ontologies or other resources*. The number of ontologies to be merged or mapped can be two or more. An input source, other than ontologies, is used as a means of checking similarities between input ontologies. This includes documents that contain intensive keywords of domain concepts and dictionaries such as WordNet [25].

- *Transformation of heterogeneous inputs into a common format*. The input ontologies and heterogeneous sources can be in different formats but they need to be transformed or extracted into a common form or model in order to perform mapping or merging.

- *Identify similarities and differences*. The similarities among input ontologies are mainly of two types: the linguistic and semantic. Linguistic similarities are specified by the terms that are used to denote concepts and properties. They are commonly compared using special dictionaries such as WordNet [25]. Semantic similarities, on the other hand, are specified by the structure of the

ontologies - the concepts hierarchy. Different approaches have their own algorithms for defining the semantic similarities.

- *Operations in merging or mapping.* This process is performed based on the similarities and differences identified in the previous stage. Overlapping concepts may be merged into one; a new concept with or without attributes from one ontology may be copied into another; mappings that link concepts in one input ontology to those in another will be established.

- *User intervention.* The resulting ontology or mappings may be presented to the domain expert for refinement during or following merging or mapping processes.

As argued by Bruijn et al. [5] in a survey of ontology merging and aligning techniques, most previous approaches are not yet capable of being applied in real practice. However, all these previous works have contributed to the evolution of ontology development. Some of them will be adapted and extended in this research after further investigation and testing.

4 The proposed methodology for ontology development

This section describes the proposed methodology in detail. The methodology model is depicted in a draft model as shown in *Figure 1*. It consists of four components:

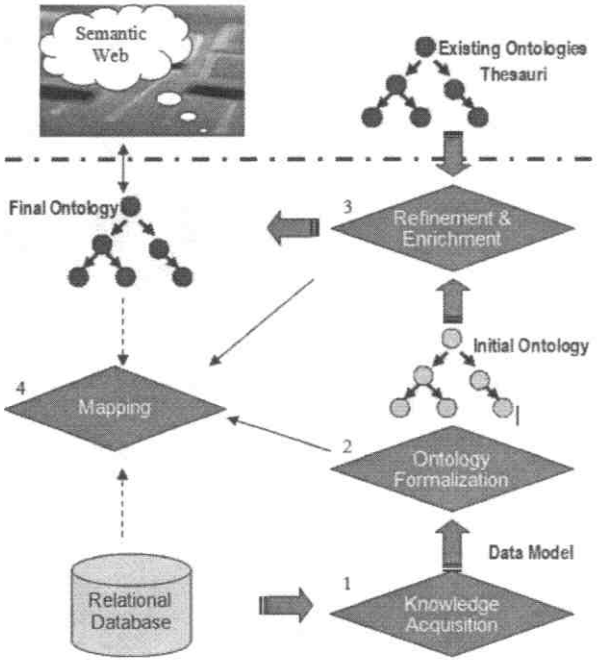

Figure 1 Draft model of the proposed methodology

4.1 Knowledge acquisition component

This component is designed to extract knowledge from a relational database schema and database instances. Techniques proposed in previous approaches such as Stojanovic, Stojanovic & Volz's [16], Astrova [17] and Meersman [15] may be combined and extended to extract a data model from the database schema. Unique to all previous approaches, data mining techniques will also be applied in this research. This is because both database schema and database instances embed valuable knowledge of the domain. It will need to incorporate both of these aspects to maximize the extraction of knowledge from the source database. Thus, existing data mining techniques on database instances will be adopted to reveal further knowledge from database instances and to refine the data model generated from database schema. Potential issues of existing databases including redundancy, poor naming and database vendor specific characteristics will also be tackled. This aims to address issues of knowledge acquisition from the database fully and in-depth.

4.2 Ontology formalization component

This component takes as input the data model produced from Knowledge Acquisition Component and transforms the data model into OWL DL by using mapping techniques. This will be done by defining and specifying correlations and counterparts between the constructs of the data model and each component of the OWL DL. The output ontology is called initial ontology.

4.3 Ontology refinement and enrichment component

The initial ontology model generated from one relational database represents domain knowledge perceived and viewed by one particular organization. It covers the common domain knowledge only at a certain level. It, more or less, is biased. To enhance its objectivity of the representation in order to eliminate the biases, the ontology model should be refined, enriched and adjusted by external data and knowledge sources. This can be done through the integration of different views of the knowledge which are represented by other existing ontologies and by publicly available knowledge in some thesauri or dictionaries. Existing approaches for ontology merging, and semantic and linguistic analysis may be adopted.

The existing ontologies can be generic ontologies and domain ontologies developed by other organizations. The candidates for generic ontologies to be used in this research include: Cyc [2] and Suggested Upper Merged Ontology (SUMO) [26]. On the other hand, domain ontologies are those that model the same domain but are developed by other organizations. They reflect different views of the same domain knowledge. The input of existing domain

ontologies will be at the choice of the users. However, this component will suggest tools such as the Semantic Web search engine 'Swoogle' [27] for the identification of existing ontologies. WordNet [25] and SENSUS [3] can be the candidates of thesauri or dictionaries. The ontology produced by this component is the final ontology.

4.4 Mapping Component

Mapping between the final ontology and the original database will be created and maintained during the previous processes. This is necessary in order to enable queries from external agents to be made into the database instances through the final ontology model, as the final ontology model has evolved. This component will also be in charge of dynamically generating ontological instances from the database instance based on the final ontology model upon the requests of external agents.

The structure of this mapping will be designed. An approach for populating mapping instances (e.g. class A in OWL maps to table A in the database), and reflecting the changes that occurred during the process from the initial model to the final model, will also be designed and developed.

5 Conclusions and future work

This proposal has outlined the methodology for developing ontology from relational database and exiting ontology and thesauri. This semi-automated methodology for developing ontology will turn the slow and manual ontology development work with a high cost into an effective engineering process. In the mean time, data and information contained in relational databases find an efficient way to become searchable to the Semantic Web without affecting exiting systems. Agreements on the resulting ontology will be much enhanced by incorporating a variety of views on the domain knowledge from thesauri and existing ontologies. The bias of the resulting ontology generated from one particular database will be limited. Hence, the resulting evolved ontology becomes more reusable and sharable.. Furthermore, the design of dynamic ontological instance generation from database instances not only eliminates the constant synchronisation between the ontological instances and database instances while providing timely updated, sharable data, but also suggests a means of enabling other applications built upon the original database keep unaffected. This ongoing research will take further two years to be consolidated and developed. A prototype of the methodology is also planned be developed for demonstration and testing purposes in this PhD research.

6 References

[1] W3C, "Semantic Web." Internet: http://www.w3.org/2001/sw/, [May. 30, 2006].

[2] CYC, "CYC." Internet: http://www.cyc.com/cyc/technology/whatiscyc, [April 25, 2006].

[3] B. Swartout, R. Patil, K. Knight, and T. Russ, "Toward Distributed Use of Large-Scale Ontologies," in *KAW96, the tenth Knowledge Acquisition Workshop*, Banff, Canada 1996, pp. 32.1-32.19.

[4] Stanford Medical Informatics, "Protégé." Internet: http://protege.stanford.edu, [June 9, 2006].

[5] J. d. Bruijn, F. Mart´ın-Recuerda, D. Manov, and M. Ehrig, "The state-of-the-art survey on ontology merging and aligning," Tech. Rep. Digital Enterprise Research Institute, University of Innsbruck EU-IST Integrated Project (IP) IST-2003-506826 SEKT: Semantically Enabled Knowledge Technologies, 2004.

[6] W3C, "Ontology Web Language," Internet: http://www.w3.org/2004/OWL/, [Sept. 5, 2006].

[7] T. R. Gruber, "A Translation Approach to Portable Ontology Specifications," *Knowledge Acquisition*, vol. 5, pp. 199-220, April 1993.

[8] W. Borst, "Construction of Engineering Ontologies," University of Twente, Enschede, 1997.

[9] R. Studer, V. R. Benjamins, and D. Fensel, "Knowledge engineering: Principles and methods," *Data & Knowledge Engineering*, vol. 25, pp. 161-197, 1998.

[10] M. Uschold and M. Gruninger, "Ontologies: Principles, Methods and Applications," *Knowledge Engineering Review*, vol. 11, pp. 93-155, Feb. 1996.

[11] M. Fernandez, A. Gomez-Perez, and N. Juristo, "METHONTOLOGY: from Ontological Art towards Ontological Engineering," in *AAAI97 Spring Symposium Series on Ontological Engineering*, Stanford, USA, 1997, pp. 33-40.

[12] Open Source Technology Group (OSTG), "Jena – A Semantic Web Framework for Java." Internet:

http://jena.sourceforge.net/index.html, [Sept. 20, 2006].

[13] A. Gómez-Pérez, D. Manzano-Macho, E. Alfonseca, R. Núñez, I. Blacoe, S. Staab, O. Corcho, Y. Ding, J. Paralic, and R. Troncy, "A survey of ontology learning methods and techniques," Tech. Rep. OntoWeb Consortium 2003.

[14] V. Kashyap, "Design and Creation of Ontologies for Environmental Information Retrieval," in *Twelfth Workshop on Knowledge Acquisition, Modelling and Management*, Voyager Inn, Banff, Alberta, Canada, 1999.

[15] R. Meersman, "Ontologies and Databases: More than a Fleeting Resemblance," in *OES/SEO Workshop Rome*, Rome, 2001.

[16] L. Stojanovic, N. Stojanovic, and R. Volz, "Migrating data-intensive Web Sites into the Semantic Web," in *the 17th ACM symposium on applied computing (SAC).*, SAC, 2002, pp. 1100-1107.

[17] I. Astrova, "Reverse engineering of relational database to ontologies," in *First european Semantic Web symposium, ESWS*, Heraklion, Crete, Greece, 2004, pp. 327-341.

[18] W3C, "RDF." Internet : http://www.w3.org/RDF/, [June 5, 2006].

[19] J. Barrasa, Ó. Corcho, and A. Gómez-Pérez, "R_2O, an Extensible and Semantically Based Database-to-ontology Mapping Language," in *Semantic Web and Databases, Second International Workshop, SWDB 2004*, Toronto, Canada, 2004.

[20] C. Bizer, "D2R MAP – A Database to RDF Mapping Language," in *the 12 thInternational World Wide Web*, Budapest, Hungary, 2003.

[21] N. F. Noy and M. A. Musen, "SMART: Automated Support for Ontology Merging and Alignment," in *the Twelfth Workshop on Knowledge Acquisition, Modeling and Management*, Banff, Canada, 1999.

[22] A. Gomez-Perez, M. Fernandez-Lopez, and O. Corcho, *Ontological engineering: with examples from the areas of knowledge management, e-Commerce and the Semantic Web*. London: Springer-Verlag, 2004.

[23] G. Steve, A. Gangemi, and D. M. Pisanelli, "Integrating Medical Terminologies with ONIONS Methodology," Internet : http://www.loa-cnr.it/Papers/onions97.pdf, 1997.

[24] N. F. Noy and M. A. Musen, "PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment," in *AAAI-2000*, Austin, Texas, 2000.

[25] WordNet, "WordNet: A Lexical Database for English." Internet : http://wordnet.princeton.edu, [July 30, 2006].

[26] SUOWG, "Suggested Upper Merged Ontology." IEEE Standard Upper Ontology working group, Internet: http://suo.ieee.org/, 2003 [3 October 2006]

[27] Swoogle, "The Semantic Web search engine," Internet: http://swoogle.umbc.edu, [Aug. 17, 2006].