Journal of Computational Information Systems 8: 21 (2012) 8819–8826
Available at http://www.Jofcis.com

A Flexible Data Integration Framework Using Ontology Similarity [⋆]

Juebo WU[*], Chenchieh FENG, Chihyuan CHEN

Department of Geography, National University of Singapore, 1 Arts Link, Singapore 117570

Abstract

In order to support for collaborative work between different individuals, teams, or organizations, data and their structures need to be frequently exchanged and integrated from one system to the other system. In this context, it is very significant and valuable to find out a fast way for achieving data integration. In this paper, we present a flexible approach for the framework of rapid data integration based on ontology similarity with language-agnostic. This is done first by extracting ontology from database, constructing the elementary semantic entities of ontology. Then, ontology similarity calculation is carried out to judge whether two ontologies are equivalent or not. We exploit the improved edit distance algorithm as the basic function for ontology similarity, and detail the intelligent framework of rapid data integration by web service. Finally, a system is developed to realize rapid data integration for two teams in a real collaborative project and the feasibility analysis is processed. It shows the high efficiency and flexibility of our approach.

Keywords: Rapid Data Integration; Ontology Integration; Ontology Similarity; Ontology Mapping

1 Introduction

Data integration provides a way to achieve comprehensive data sharing from different sources, formats, and characteristics in logical or physical aspects. Data sharing can enable more people to make full use of existing data resources, and reduce duplication of data collection, data acquisition and the corresponding costs [1]. In recent years, there have been many contributions made on data integration, mainly concentrated on 1) Schema mapping [2], 2) Middleware [3], 3) XML matching [4] and 4) Ontology mapping [5]. Ontology mapping was proposed to solve the issues that the former three technologies have such as its hard to extension to other fields.

Like the former three technologies, the contributions of ontology mapping are mostly concentrated on specific fields, and most of them require that the ontology or data should be described only in one language. Moreover, the current researches on ontology mapping are mainly about one aspect of the whole process of data integration, and it is a challenge task for users to utilize

[⋆] Project supported by the Center for Environmental Sensing and Modeling (CENSAM).

[*] Corresponding author.
Email address: wujuebo@gmail.com (Juebo WU).

1553–9105 / Copyright © 2012 Binary Information Press
November 1, 2012

such methods into their collaborative work systems. Therefore, its essential to establish a total framework based on ontology mapping which can carry out rapid data integration from original data in different systems. To fulfill this goal, the key issues that should be addressed in this paper are as follows. (1) How to convert data from database to ontology? (2) How to compare the similarity of two ontologies, and this way should be language-agnostic or language-independent? (3) In combination with the proposed approaches, how to build a whole process of rapid data integration concerning reliability and uncertainty? Regarding (1), we define four mapping relationships in accordance with OWL DL standard to extract ontology from database. For (2), we exploit the improved edit distance algorithm as the basic function for comparing two ontology entities. Because multiple features are participated in computing similarity, language-agnostic can be achieved. With respect to (3), the whole framework of rapid data integration is proposed based on web service. About reliability and uncertainty, we will discuss them in section 4. The remainder of the paper is organized as follows. Section 2 outlines our approach and improves edit distance algorithm to fit our framework. The way how to carry out ontology similarity calculation is given in Section 3. Section 4 describes the details the comprehensive approach and framework of rapid data integration. A case study is demonstrated in Section 5. The final section draws a conclusion for this paper.

2 Our Approach and Improvement

2.1 Motivation and our approach

According to the changing frequency of data or their structure between integration participants in collaborative work, we define the higher one as rapid data integration while the lower one as non-rapid data integration. Thus, the approaches that can be used in rapid data integration are also suitable for the situation of non-rapid data integration. Under the circumstance of rapid data integration, some systems have to change or re-design their data structure in order to meet the new demands for data integration. The underlying problem is that the collaborative systems cannot understand the meaning and data structure from each other, even though the data are the same.

In terms of abstract level, the term data can be divided from lower level to higher level as data, schema, xml and ontology. As the key issue (1), it requires extracting ontology from database, that is, we should extract ontology from the lowest level firstly. For this purpose, we choose OWL DL as ontology language for such data from database, and define several mapping relationships for ontology extraction. As OWL DL, several grammars and contents are used in this paper such as class, property, and individual.

After ontology extraction, the task of the key issue (2) is to compute similarity between ontologies with a quantitative method, namely ontology mapping. From ontology point of view, there are many features to describe it, such as class, individual and property etc. So we can separate ontology mapping into two steps. At first, calculate similarity for ontologies. And then choose the highest value one as similar ontology. Building on the idea of simplicity and rapidity, we establish a similarity measures by introducing edit distance [6]. Because most of the names are not always the same from different integration participants, it needs to do some improvement for edit distance which will be discussed later.

2.2 Improved edit distance

Edit distance was proposed by Levenshtein [6] to compute the similarity for a source string S and a target string T. Here, we select it as the foundation block of forming equation for obtaining ontology similarity.

We define the similarity of two strings as:

$$S_{ij} = 1 - \frac{ED_{ij} + \alpha}{maxLength(S,T) + \alpha} \tag{1}$$

where

ED_{ij}: edit distance of source string S and target string T,

$maxLength(S,T)$: maximum length between S and T,

α: control parameter that uses to enlarge or reduce the similarity in various applications.

Its common that naming for the same data from different integration participants has various orders. Thus, we modify the algorithm by using substring for similarity comparison instead of single character, given as below.

$$ED_{ij} = \begin{cases} min \begin{cases} ED_{i-1,j} + 1 \\ ED_{i,j-1} + 1 \\ ED_{i-1,j-1} + (s_i == t_j ?0 : 1) \end{cases} & \\ 0 & i == 0 \&\& j == 0 \\ ED_{i-k-1,j-k-1} + 1 & if(s_{i-k}s_i == t_j t_{j+k} \&\& (i > 0 || j > 0)) \end{cases} \tag{2}$$

It can be seen from equation (2), when k=1, the swap operation becomes adjacent position swap operation, which is traditional edit distance. By using equation (2), the edit distance of the same example above is 1 and the similarity is 0.75. Therefore, the result of improved algorithm is more close to reality.

3 Ontology Similarity Generation

We design a quantitative representation for ontology similarity that emphasizes ontology extraction and similarity calculation, also including algorithm optimization.

3.1 Ontology extraction

In our presented approach, ontology extraction is the first and foremost process and ontologies are extracted from different databases or distributed network nodes. We define the following four mapping relationships to extract ontology from data schema to ontology entities, as OWL does. The detail mappings are described in [7].

3.2 Similarity calculation

Similarity calculation is carried out after ontology extraction. The target is to compare the similarity for each ontology entity. In our approach, several entities are involved to compute similarity for one ontology entity, aiming to ensure objectivity. As described in [7], there are more than four main similarities to be calculated, such as similarity between classes or subclasses, similarity of properties, similarity of individuals etc.

From the presented approach, it can be seen that each similarity of ontology entities is computed depending on more than one feature. For example, similarity between classes or subclasses is based on class, property and domain etc. Its different from the comparison only by two names or strings. Therefore, the similarity of ontology described in different languages can be also accepted, such as combining Chinese with English. Therefore, the framework of rapid data integration by using the presented approach for ontology similarity has the ability of language-agnostic or language independence.

3.3 Similarity optimization

Since the data come from different data sources, they usually carry names most familiar to database designers. It is critical to normalize synonyms so as to reduce computation. In terms of synonym, the same meaning with different words can be normalized into the uniform expression. Therefore, we carry out synonym normalization immediately after ontology extraction. WordNet is chosen as synonym normalization tool [8] because of its simplicity. It is easy to get a group of semantically equivalent data elements by WordNet Library.

4 Framework of Rapid Data Integration

Reliability and uncertainty are two essential aspects relating to ontology. To that end, we take the below strategies in our framework of rapid data integration by (1) authentication before integration and (2) interaction interface for user modification of ontology similarity.

4.1 Framework

In accordance the presented approach, the framework of the ontology similarity-based rapid data integration is shown in Fig. 1, which is composed of five layers:

(1) System interface. This layer gives a number of system interfaces and user operation interfaces, which are designed to support the user who can access and operate all kinds of functions.

(2) Integration layer. This layer is the key part in the framework. Its designed to publish the integration requirements for cooperation, which includes semantic conversion, ontology reasoning etc.

(3) Similarity calculation layer. This layer is responsible for calculating similarity between ontologies, including class similarity, individual similarity, and property similarity. By using the presented approach, ontology similarity can be produced automatically. User interface is also provided in this layer, in order to improve or modify the results by user.

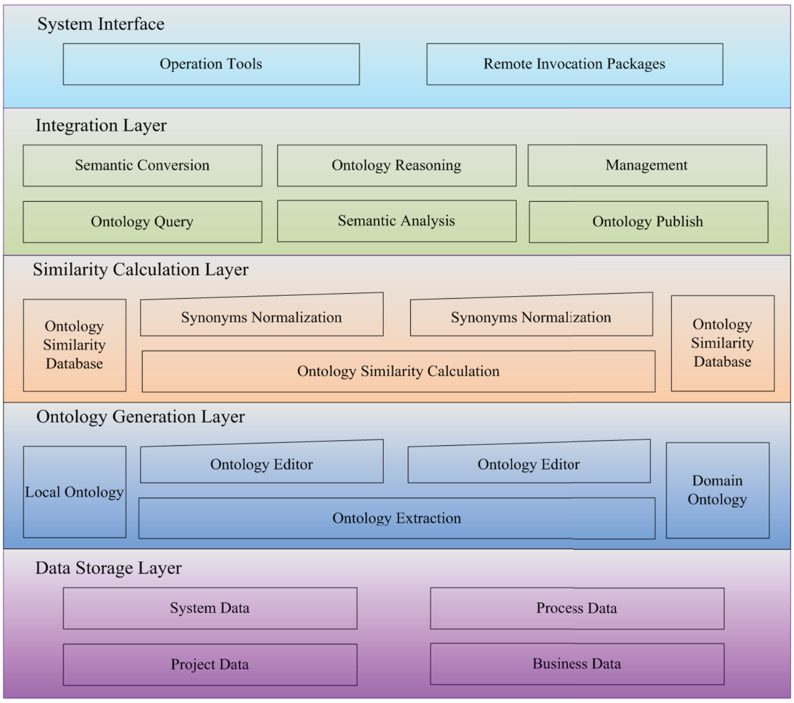

Fig. 1: Framework of ontology similarity-based rapid data integration

(4) Ontology generation layer. Ontology extraction and its editor is the bridge between database and ontology, and it adopts the presented mapping relationships to create ontology. A user interface is provided for improvement and modification of generated ontologies.

(5) Data storage layer. The data and the temporary files in processing steps are stored in this layer, including configuration information etc.

4.2 Web service-based platform

In order to achieve flexibility, web service is adopted as the foundation for rapid data integration platform (Fig.2).

In this platform, the related control parameters and data are wrapped by XML file mode to transmit, and the specific process is decided by the integration bus to deal with such data. Through user interface, many integrated interactive platforms are provided to clients, and the user can call the services in rapid integration platform. The browser can interpret the returned XML documents and provide the next step of operation graphical interface. The interaction between the users and the server is all based on graphical interface.

4.3 Framework reliability and uncertainty

General speaking, reliability requires the framework can perform and maintain its functions in routine circumstances, as well as unexpected circumstances [9]. To fulfill this goal, we exploit authentication mechanism to this framework based on our previous work [10].

For uncertainty, we concern three aspects, that is, data uncertainty, semantic uncertainty, and ontological uncertainty. We exploit human interaction method for improvement. The user can

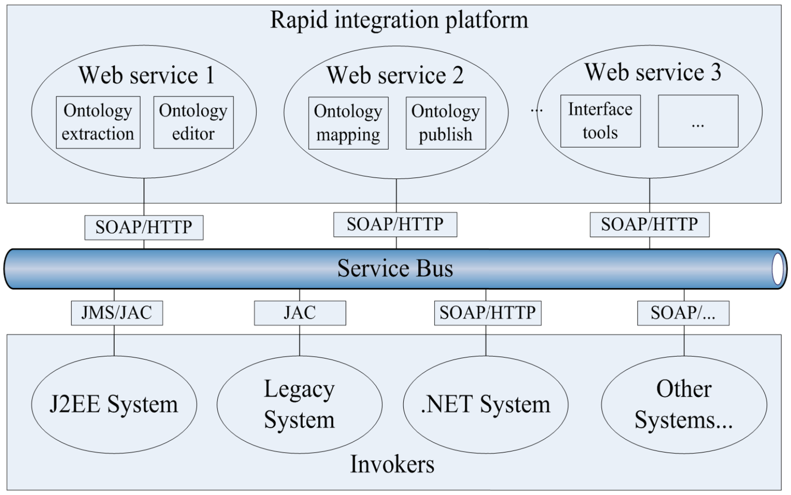

Fig. 2: Web service-based architecture of rapid data integration

adjust or modify the results of similarity calculation.

5 Case Studies

This section demonstrates the proposed approach to perform the whole framework for rapid data integration. The context is as follows. There are more than ten teams working for a geographical project, and they want to share and exchange data for each other frequently. Team A and team B are two parts in this project. There are some existing systems established already in Team A and Team B, which using different data structures. Although the function and design of these two systems are different, parts of the objective data are the same and all the two systems have stored such data. System A (developed by team A) can carry out spatial clustering while system B (developed by team B) can perform data normalization. These two systems want to exchange data for cooperation. System A carries out spatial clustering for the normalized data which come from system B. Since the data structure is different, the data cannot be shared and transferred without preprocessing. In order to realize data cooperation for the two systems without changing the existing system, the rapid integration platform is adopted for two systems cooperation as a bridge. JUDDI and AXIS2 are introduced as basic technical support so that web service can be rapidly developed and deployed.

5.1 Ontology extraction

The structures of data source chosen in this demonstration in system A and system B are different in parts. Parts of the table names are different, and parts of the column names are also different. The data records in system A and system B are from the same data source, and parts of contents in two systems are same while the rest are not. So, ontologies are extracted from these two tables. The related entities can be generated here including classes, properties etc.

5.2 Similarity calculation and optimization

System A sends the records to integration platform and it needs to figure out which records from system B are the same. And then, system B carries out data normalization and returns the results to system A. Ontology optimization is conducted before ontology calculation. Two classes are created from system A (i.e., cyclone) and system B (i.e., whirlwind) respectively.

5.3 Results analysis

By using the presented framework, we can achieve rapid data integration for two teams in the project. In combination with user participation, it can easily establish a reliable and complete ontology similarity database for data integration. Ontology construction can be finished by visiting information system database for only one time. The joint systems can send their data to integration platform without establishing a new data structure or updating their existing systems. These make data integration with rapid speed and more scalability.

For simplicity, we named the similarity methods from [11] as JE&PV and [12] as ME&YS by choosing their authors names for short. Likewise, the presented similarity method in this paper is short for JW&CF&CC. Fig.3 shows relationship between the number of fields in data table and computing accuracy.

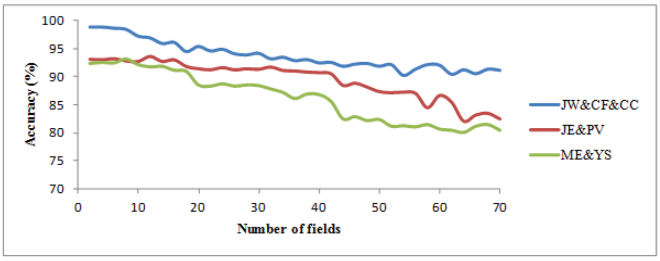

Fig. 3: Relationship between the number of fields in data table and computing accuracy

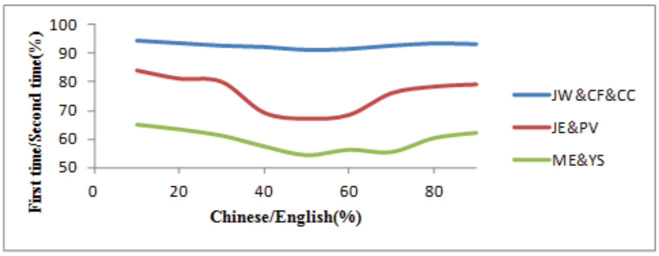

Fig. 4: Results of testing methods with language-agnostic

It shows that the results of average accuracy are respectively 93.94% for JW&CF&CC, 89.56% for JE&PV and 86.25% for ME&YS. That means our presented method can reach higher accuracy for ontology similarity. It also shows that the accuracy is going down when the number of fields in data table is increasing for all three ways. However, the accuracy of our approach is stable when the number is going up to a certain amount.

To test the influence of data table names designed by different language, we carry out several experiments by adjusting the proportion of Chinese to English. Fig.4 gives the results of the above experiments.

X-coordinate means the proportion of Chinese to English using for naming data in the table, while Y-coordinate denotes the accuracy proportion of the first time (only one language) to the second time (two languages). We can see from that our approach has a better compatibility when combining different languages between different systems. The accuracy proportion of the first time to the second time can reach 92.79% and the change is less. Therefore, we accomplish the goal that the presented framework is language-agnostic or weak language-related.

6 Conclusion

To support for collaborative work, we put forward a complete framework for rapid data integration based on ontology similarity. Three key issues in rapid data integration were solved. Unlike other current researches in ontology similarity, edit distance was introduced in our approach for similarity computation with high flexibility. The new approach has the advantage of extracting ontology from distributed systems or nodes on the Web and incorporating most of the descriptive features of ontology into similarity calculation. We considered the human interaction during the process by user-defined interface, and also including reliability and uncertainty from semantic aspect. Web service was also involved into increase flexibility. Moreover, the new framework can achieve language independence by concerning more entity features when doing similarity analysis. The case study proved the proposed framework is valuable and efficient in dealing with rapid data integration between two parties, with high scalability.

Acknowledgement

The research described in this paper was funded in whole or in part by the Singapore National Research Foundation (NRF) through the Singapore-MIT Alliance for Research and Technology (SMART) Center for Environmental Sensing and Modeling (CENSAM).

References

[1] H. Chueh and N. Lin, A data accessing model for dynamic clustering of object-oriented databases, Journal of Computational Information Systems, v 6, n 9, 2010, pp. 2787-2794.

[2] H. H. Do, Schema Matching and Mapping-based Data Integration, Verlag Dr. Mller(VDM), 2006.

[3] A. Halevy, A. Rajaraman, and J. Ordille, Data integration: the teenage years, In VLDB Conference, VLDB Endowment, 2006, pp. 916.

[4] M. Keulen, A. Keijzer, and W. Alink, A probabilistic XML approach to data integration, In Proc. ICDE-2005, IEEE Computer Society, 2005, pp. 459-470.

[5] P. Ceravolo, E. Damiani, A. Gusmini, and M. Leida, Using Ontologies to Map Concept Relations in a Data Integration System, in OTM Workshops (2), 2007, pp. 1285-1293.

[6] Y. Li. and B. Liu, A normalized levenshtein distance metric, Pattern Analysis and Machine Intelligence, IEEE Transactions on 29(6), 2007, pp. 1091-1095.

[7] J. Wu, C. Feng and C. Chen, Ontology Similarity Measurement Method in Rapid Data Integration, The International Conference on Data Technologies and Applications (DATA 2012), In press.

[8] V. D. Blondel, A. Gajardo, M. Heymans, P. Senellart, and P. V. Dooren, A measure of similarity between graph vertices: Applications to synonym extraction and Web searching, SIAM Rev., 2004, pp. 647-666.

[9] http://en.wikipedia.org/wiki/Reliability.

[10] C. Feng and L. Yu, A generic attribute-improved RBAC model by using context-aware reasoning, WorldComp'11, 2011, pp. 398-404.

[11] J. Euzenat and P. Valtchev, Similarity-based ontology alignment in OWL-lite, In Proc. 15th ECAI, 2004, pp. 333-337.

[12] M. Ehrig and Y. Sure, Ontology mapping-an integrated approach, In 1st European SemanticWeb Symposium, 2004, pp. 76-91.