2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery

Semantic Similarity-Based Ontology Alignment for Enterprise Ontologies

Chong Xie
Shenyang Ligong University
Shenyang, Liaoning 110168, China
chongxie@126.com

Abstract

Ontology, as a representation of shared conceptualization of specific domain, is a key technology for the Semantic Web. In recent years, with the rapid development of the Semantic Web, many large enterprises have established their own ontologies. In order for enterprises to facilitate communication and interoperability across multiple ontologies, we need a flexible mechanism to align ontologies. In this paper we present an approach to ontology alignment based on ontology similarity measures. Particularly, we introduce semantic similarity-based measures, i.e., attributional, relational, and extensional similarity measures. Furthermore, we give evaluation metrics on our approach with reference to ontology alignment quality.

1. Introduction

Ontologies are increasingly seen as a key technology for enabling semantics-driven information processing [1]. Many applications benefit from semantically enriched information, including enterprise management and e-business. Organizations establish their own ontologies to provide a framework for sharing a precise meaning of symbols exchanged during communication. These ontologies cover unrelated or overlapping domains at different levels of detail and granularity. As a result, a critical issue related to working with ontologies in real-world enterprise applications is how to establish interoperability between agents or services using different individual ontologies [2]-[4]. Semantic interoperability can be grounded in ontology reconciliation. The underlying problem is called ontology alignment problem. Ontology alignment is a crucial and essential issue to resolve when building an integrated company-wide knowledge web or even a world-wide Semantic Web.

Generally, ontology alignment process can be manual, semiautomatic, or full automatic. To date, most tools are semiautomatic and vary according to the particular nature of the domain being treated. Full automation process of the ontology alignment is rarely achieved. Often, domain experts and knowledge engineers are needed to assist with decision making. Through alignment, correspondences between the objects in the ontologies can be identified and we are able to translate requests and map data between different ontologies to perform integration or even perform a complete merge of ontologies. As one can easily imagine, this cannot be done manually beyond a certain complexity, size, or number of ontologies any longer. Automatic or at least semiautomatic techniques have to be developed to reduce the burden of manual creation and maintenance of alignments.

There is already some related work on techniques for ontology alignment in recent years. For example, Noy *et al.* have developed and implemented PROMPT, an algorithm that provides a semiautomatic approach to ontology merging and alignment [5]-[7]. Ehrig *et al.* propose FOAM that is a framework for ontology alignment and mapping [8]-[9]. In addition, Lambrix *et al.* exploit a system (SAMBO) for alignment and merging biomedical ontologies [10]. Nevertheless, the methods for alignment of ontology of real-world enterprise-wide are still in the early stages. Therefore, many new methods will be presented on the way.

In this paper, we propose an approach to ontology alignment based on ontology similarity measures. Besides common lexical similarity measure, we introduce semantic similarity measures. Ontology alignment focuses on the aspect of how to organize the classes while the semantics are agreed to be identical, e.g., they are semantically close by equivalence or subsumption.

The rest of the paper is organized as follows. In the next section we introduce ontology similarity measures. In section 3 we propose our ontology alignment approach. In section 4 we introduce evaluation metrics on our ontology alignment approach. Finally we give our conclusions and present some topics for further

978-0-7695-3735-1/09 $25.00 © 2009 IEEE
DOI 10.1109/FSKD.2009.679

research in section 5.

2. Ontology similarity measures

In this section we first give ontology definition before further discussing ontology similarity measures.

2.1. Ontology definition

Definition 1 (Ontology). An ontology is defined as the tuple O = (C, P^C, R, I, A) where:

C is the set of classes (or concepts). The set is assumed a lattice, in which every pair of classes has a unique supremum and a unique infimum. This assumption ensures mathematical soundness of the structure.

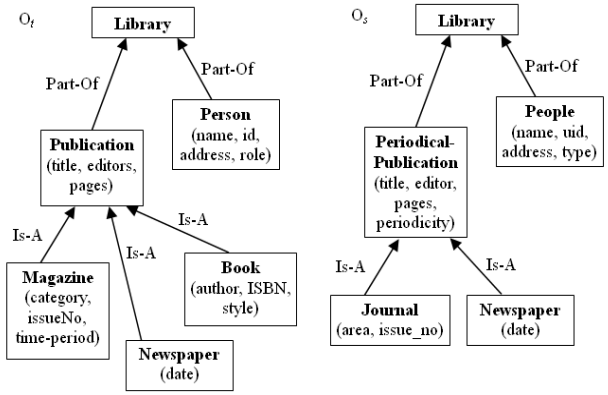
P^C is the set of properties. Each class C can be described by a set of properties denoted by $P^C(C)$.

R is an n-ary relation on the domain of classes.

I is a set of instances of classes and relationships associated with the ontology.

A is the set of axioms. Each axiom in A is a constraint on the class, instances and relationships between class objects.

For example, we give two example ontologies that are handled by two different enterprises about management of publication and staffs in Figure 1.



(a) (b)

Figure 1. Two example ontology

The components of the ontology in Figure 1(a) are represented as follows:

C={Library, Publication, Person, Magazine, Newspaper, Book}

P^C(Person) ={name, id, address, roles}

P^C(Publication)={title, editors, pages}

P^C(Magazine)={category, issueNo, time-period}

P^C(Book)={author, ISBN, style}

P^C(Newspaper)={date}

R={(Publication, Library), (Person, Library), (Magazine, Publication), (Newspaper, Publication), (Book, Publication)}

I={Magazine(People Magazine, Star Magazine, PC Magazine), Newspaper(The Sun Newspaper, New York Times, Washington Post), Book(Handbook)}

Formally, this ontology is defined according to O = (C, P^C, R, I, A). Here, we omit axioms, e.g., A={$\forall x\ \exists y$ Book(x) $\Rightarrow$ Publication (y)}, which is a notation according to first-order logic. The example ontology does not contain all constructs. The sample data shown is for illustrative purposes and purposely very limited.

2.2. Computing ontology similarity

Ontology similarity refers to the comparison of whole ontologies or sub-elements thereof. Specially, ontology similarity denotes the correlation between two entities of different ontologies. The correlation returns a numerical value indicating whether the two entities have a high or low degree of similarity. This can be formally laid down through the following definition.

Definition 2 (Ontology similarity). A similarity function

sim: $2^E \times 2^E \rightarrow [0, 1]$

is a function that maps a pair of entity sets (expressed through the power set 2^E of entities) of their corresponding ontologies O (e.g., source ontology O_s and target ontology O_t) to a real number expressing the similarity between two sets such that:

$sim(e, f) \geq 0$ (positiveness)

$sim(e, e) \geq sim(f, g)$ (maximality)

$sim(e, f) = 1 \Leftrightarrow e = f$: two entity sets are identical.

$0 < sim(e, f) < 1$: two entity sets are similar/different to a certain degree.

$sim(e, f) = 0 \Leftrightarrow e \neq f$: two entity sets are different and have no common characteristics.

Entity e, f, or g ($\forall e, f, g \in 2^E$) may be sets of classes, relations, or instances. This can include subtrees or even whole ontologies. A set may also consist of only one entity, thus in the extreme case reducing the similarity to a similarity between two individual elements.

A number of similarity measurement techniques such as the cosine similarity measure [11], Dice's coefficient [12], and Jaccard's index [13] have been defined to compute this similarity. Using set theory, Tversky [14] defined a similarity measure in terms of a matching process. This measure produces a similarity value that is the result of common as well as different characteristics of objects.

In this paper, we introduce a computational model that assesses similarity by combining a feature process

with semantic similarity-based measures. There are at least three kinds of semantic similarity, i.e., attributional similarity, relational similarity, and extensional similarity. Attributional similarity is correspondence between attributes, relational similarity between relations, and extensional similarity between instances. Attributes are used to state properties of classes. Relations express relations between classes. Instances of each class are declared as its members.

The computational model uses the number of common and different features between two entity classes. The global similarity function $SIM(e_s, e_t)$, given in (1), is a weighted sum of the similarity values for lexical and semantic part, where $e_s \in O_s$ is referred to as the source entity set and $e_t \in O_t$ is referred to as the target entity set, $sim_l(e_s, e_t)$, $sim_a(e_s, e_t)$, $sim_r(e_s, e_t)$, and $sim_e(e_s, e_t)$ denote similarity value for lexical, attributional, relational, and extensional parts, respectively, and w_l, w_a, w_r, and w_e are weights of each similarity value. These weights define the relative importance of lexical, attributional, relational, and extensional similarity that may vary among different ontologies. Their respective weights add up to 1.0. Furthermore, only common specification components can be used in a similarity assessment.

$$SIM(e_s, e_t)=w_l \cdot sim_l(e_s, e_t) + w_a \cdot sim_a(e_s, e_t) + w_r \cdot sim_r(e_s, e_t) + w_e \cdot sim_e(e_s, e_t) \quad (1)$$

1) Lexical similarity measures. The lexical similarity measure is typically discovered by rules like Porter Stemming, Levenshtein, Prefix and Suffix, and Substring algorithm. Here we use the existing approach from [15] and [16] in (2).

$$sim_l(e_s, e_t) = \max(0, \frac{\min(|e_s|,|e_t|) - ed(e_s - e_t)}{\min(|e_s|,|e_t|)}) \quad (2)$$

The idea behind this measure is to take two strings to determine how many atomic actions are required to transform one string into the other one. Atomic actions would be addition, deletion, and replacement of characters, but also moving their position.

2) Semantic similarity measures. Further, for other types of distinguishing features we use a similarity function $sim_{tp}(e_s, e_t)$ given in (3) based on the ratio model of a feature-matching process [14]. Symbol *tp* can be *a*, *r*, or *e* denoting the type of features, i.e., attributional, relational, and extensional similarity in (1). Each entity *e* ($e_s \in O_s$ or $e_t \in O_t$) is characterized by a set of features *ftrs*(*e*), | | is the cardinality of a set, and α is a function that defines the relative importance of the non-common characteristics.

$$sim_{tp}(e_s, e_t) = \frac{|ftrs(e_s) \cap ftrs(e_t)|}{|ftrs(e_s) \cap ftrs(e_t)| + \alpha |ftrs(e_s) - ftrs(e_t)| + (1-\alpha) |ftrs(e_t) - ftrs(e_s)|} \quad (3)$$

To compute the value of function α we adopt the approach proposed by [17]. We assume connecting two independent ontologies being compared and having a mutual imaginary root of the two ontologies. Using the connected ontology, the function α of the matching model can be expressed in terms of the depth of the entity classes given in (4). The function *depth*() corresponds to the shortest path from the entity class to the imaginary root. This depth reflects the degree of granularity upon which the ontology was designed. For example, consider the ontologies in Figure 1. When the two classes Periodical-Publication$_s$ and Publication$_t$ are compared, *depth*(Periodical-Publication$_s$) is 2 and so the *depth*(Publication$_t$). So α(Periodical-Publication$_s$, Publication$_t$) is 0.5. With this definition of α, evaluation from deep to shallow ontologies usually result in greater value of similarity than evaluation from shallow to deep ontologies.

$$\alpha(e_s, e_t) = \begin{cases} \frac{depth(e_s)}{depth(e_s)+depth(e_t)} & depth(e_s) \le depth(e_t) \\ 1 - \frac{depth(e_s)}{depth(e_s)+depth(e_t)} & depth(e_s) > depth(e_t) \end{cases} \quad (4)$$

i) Attributional similarity measures. The amount of attributional similarity between two classes, A and B, depends on the degree of correspondence between the properties of A and B. A measure of attributional similarity is a function that maps two classes, A and B, to a real number from 0 to 1, i.e., $sim_a(A, B) \in [0, 1]$. The more correspondence there is between the properties of A and B, the greater their attributional similarity. For example, the classes Publication and Periodical-Publication have a relatively high degree of attributional similarity.

ii) Relational similarity measures. Concept analogies are often written in the form A:B::C:D, meaning A is to B as C is to D. e.g., traffic:street::water:riverbed. Traffic flows over a street; water flows over a riverbed. A street carries traffic; a riverbed carries water. There is a high degree of relational similarity between the concept pair traffic:street and the concept pair water:riverbed.

The amount of relational similarity between two pairs of classes, A:B and C:D, depends on the degree of correspondence between the relations between A and B and the relations between C and D. A measure of relation similarity is a function that maps two pairs, A:B and C:D, to a real number from 0 to 1, i.e., $sim_r(A:B, C:D) \in [0, 1]$. The more correspondence there is between the relations of A:B and C:D, the greater their relational similarity. For example, painter: painting and sculptor: sculpture have a relatively high degree of relational similarity.

A natural approach to compare the relational similarity between two entity classes is to determine the similarity of the immediate superclass and subclass of each entity class, respectively. Furthermore, different relation between superclass and subclass

including the Is-A relation, the Part-Of relation and other relations should be considered. Thus, we need assign weighs to different relations and normalize the relational similarity.

iii) Extensional similarity measures. For a class which just has some instances instead of subclass, we can apply extensional similarity measures where $ftrs(e)$ denoting the instances set of class e in (3).

3. Ontology alignment

The ontology alignment problem may be described as follows: given two arbitrary ontologies each describing a set of discrete entities (which can be classes, relations, or instances), aligning one ontology with another means that for each entity to find the correspondences, e.g., equivalence or subsumption holding between these entities. That is, for an entity in a source ontology O_s, we try to find a corresponding entity which has the same intended meaning in target ontology O_t. For example, we may find a subclass or equivalence relationships between two classes, a subPropertyOf or equivalentProperty relationships between two properties, and sameAs or differentFrom relationships that states two instances are semantically identical or different, respectively. This set of correspondences is called alignment between the pairs of entities of two ontologies. Furthermore, an entity can either be mapped to exactly one other entity or none.

Definition 3 (Ontology alignment). An ontology alignment function, *align*, based on the vocabulary, E, of all terms $e \in E$ and the set of possible ontologies, O, as a partial function:

$align: E \times O_s \times O_t \rightharpoonup E$,

$align_{rs}(e_s) = \{f_t \cup \perp \mid \text{for } \forall e \in E_{Os}, \exists f \in E_{Ot}\}$

Where "⊥" indicates that no valid alignment is found and "rs" denotes the relationships between the entity e_s and f_t. For example, $align_{subClassOf}$(Periodical-Publication$_s$)=Publication$_t$ means Periodical-Publication$_s$ is a subclass of Publication$_t$.

Once a (partial) alignment between two ontologies O_s and O_t is established, we say entity e_s is aligned with entity f_t iff $align_{rs}(e_s) = f_t$.

For example, for the ontologies depicted in Figure 1(a) and 1(b), the following set of correspondences gives a possible alignment:

$align_{subClassOf}$(Periodical-Publication$_s$)= Publication$_t$

$align_{equivalentClass}$(People$_s$)=Person$_t$

Whenever we consider a pair of classes or properties for alignment, we need to decide whether they should be in a subsumption or an equivalence relationship. We use the following decision procedure to determine the relationship. We first rank the attributional, relational, and extensional similarity and select the top one. Here, we suppose the top one similarity is attributional similarity and describe the decision process on attributional similarity.

Let $c_s, c_t \in C$ be two classes in the source ontology and target ontology, respectively. Let $ftrs(c_s, c_t) = \frac{|ftrs(c_s) - ftrs(c_t)|}{|ftrs(c_s)|}$ and $ftrs(c_t, c_s) = \frac{|ftrs(c_t) - ftrs(c_s)|}{|ftrs(c_t)|}$, where $ftrs(c)$ denotes the set of attributes in class c and $\beta \in [0, 1]$ be an arbitrary, but fixed threshold value.

If $ftrs(c_s, c_t) < \beta$ and $ftrs(c_t, c_s) \geq \beta$, then consider the alignment: $align_{subClassOf}(c_s)=c_t$.

If $ftrs(c_s, c_t) \geq \beta$ and $ftrs(c_t, c_s) < \beta$, then consider the alignment: $align_{subClassOf}(c_t)=c_s$.

Otherwise consider the alignment: $align_{equivalentClass}(c_s)=c_t$.

Specially, if c_s and c_t are in neither equivalence nor subsumption relationship then $ftrs(c_s, c_t) = ftrs(c_s)$ or $ftrs(c_t, c_s) = ftrs(c_t)$, which imply the two classes can not contain overlapping parts.

4. Evaluation metrics

Evaluation is an important and central issue when proposing new methods. Here, we just give evaluation metrics on our ontology alignment approach. We use the following terminology:

#*existing_alignmets*: the amount of all the alignments between source ontology and target ontology.

#*found_alignments*: the amount of found alignments.

#*correct_found_alignments*: the amount of actually correct found alignments.

$$Recall = \frac{\#correct\_found\_alignments}{\#existing\_alignmets}$$

$$Precision = \frac{\#correct\_found\_alignments}{\#found\_alignmets}$$

We use the metrics that are *Recall* and *Precision* to evaluate the quality of our ontology alignment approach. We consider a pair of alignment successful if both metrics have a high value: high recall indicates that many correct pairs are found, while high precision indicates the predominance of correct found pairs.

5. Conclusion and outlook

To support the interoperability between enterprises using different individual ontologies, we need to identify the correspondences between the objects in the ontologies. We explore ontology alignment techniques to complete the task. In this paper, we present an approach to ontology alignment based on ontology similarity measures. The new approach has the advantage in that it uses a semantic similarity-based ontology alignment approach, i.e., attributional,

relational, and extensional similarity measures instead of just using common lexical similarity measures. In addition, evaluation metrics related to ontology alignment quality are presented.

In the scope of the ontology project, our future goal is to evaluate our approach's practicality and elicit the technical requirements that are crucial for successfully applying ontology alignment in enterprise resources management. Further, we will explore the combination of techniques from the areas of database, artificial intelligence and distributed systems technology.

References

[1] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web", The Scientific American, vol. 284, no. 5, pp. 33-43, May 2001.

[2] J. Lee and R. Goodwin, "Ontology management for large-scale enterprise systems", Electronic Commerce Research and Applications, vol. 5, pp. 2-15, 2006.

[3] Z. Cui, V. A. M. Tamma, and F. Bellifemine, "Ontology management in enterprises", BT Technol J, vol. 17, no. 4, pp. 98-107, Oct. 1999.

[4] A. Maedche, B. Motik, L. Stojanovic, R. Studer, and R. Volz, "Ontologies for enterprise knowledge management", IEEE Intelligent Systems, vol. 18, no. 2, pp. 26-33, 2003.

[5] F. Noy and M. A. Musen, "PROMPT: Algorithm and tool for automated ontology merging and alignment," in Proc. 17th National Conf. on Artificial Intelligence, Austin, 2000, pp. 450-455.

[6] F. Noy, and M. A. Musen, "Anchor-PROMPT: Using nonlocal context for semantic matching", in Workshop on Ontologies and Information Sharing at the 17th Int. Joint Conf. on Artificial Intelligence, Seattle, 2001, pp. 63-70.

[7] F. Noy and M. A. Musen, "The PROMPT suite: Interactive tools for ontology merging and mapping", International Journal of Human-Computer Studies, vol. 59, no. 6, pp. 983-1024, Dec. 2003.

[8] Ehrig and Y. Sure, "FOAM-Framework for ontology alignment and mapping results of the ontology alignment evaluation initiative", in K-Cap Workshop on Integrating Ontologies, Banff, 2005, vol. 156, pp. 72-76.

[9] Ehrig, S. Staab, and Y. Sure, "Bootstrapping ontology alignment methods with APFEL", in Proc. 4th Int. Semantic Web Conference, Galway, 2005, LNCS 3729, pp. 186-200.

[10] P. Lambrix and H. Tan, "SAMBO-A system for aligning and merging biomedical ontologies", Journal of Web Semantics: Science, Services and Agents on the World Wide Web, vol. 4, no. 3, pp. 196-206, Sep. 2006.

[11] D. Dhyani, W. Keong, and N. Bhowmick, "A Survey of Web Metrics", ACM Computing Surveys, vol. 34, no. 4, pp. 469-503, Dec. 2002.

[12] C. J. van Rijsbergen, Information Retrieval. London: Butterworths, 1979, pp. 25-26.

[13] L. Hamers, Y. Hemeryck, G. Herweyers, M. Janssen, H. Keters, R. Rousseau, et al., "Similarity measures in scientometric research: The jaccard index versus salton's cosine formula", Information Processing and Management, vol. 25, no. 3, pp. 315-318, Aug. 1989.

[14] A. Tversky, "Features of similarity", Psychological Review, vol. 84, no. 4, pp. 327-352, July, 1977.

[15] I. Foster, C. Kesselman, J. M. Nick, and S. Tuecke, "The physiology of the grid: An open grid services architecture for distributed systems integration", Technical report, Global Grid Forum, 2002.

[16] A. Maedche and S. Staab, "Measuring similarity between ontologies", in Proc. European Conf. on Knowledge Acquisition and Management, Madrid, 2002, LNAI, vol. 2473, pp. 251-263.

[17] M. A. Rodriguez and M. J. Egenhofer, "Determining semantic similarity among entity classes from different ontologies", IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 2, pp. 442-456, Mar./Apr., 2003.