Available online at www.sciencedirect.com

SciVerse ScienceDirect

Physics Procedia 25 (2012) 2167 – 2172

Physics Procedia

2012 International Conference on Solid State Devices and Materials Science

Combination of the Improved Method for Ontology Mapping

Rujuan Wang[1,2] , Lei Wang[1] and Lei Liu[2,*] Gang Chen[1] and Qiushuang Wang[3]

[1] College of Humanities & Sciences of Northeast Normal University
[2] College of Computer Science Jilin University Changchun 130012, China
[3]Department of agriculture JiLin University, Changchun,130062, China

Abstract

Most of current ontology mapping methods can not treat different mapping tasks in different ways referred to the features of the input ontology. And they combine different features of ontology without full consideration of the influences on mapping results caused mapping features. In view of the above questions, this paper proposes mapping method which can use entropy decision-making method to determine the combined weight of the different features of ontology. Experiments show that this method can maintain the stability and the commonality, and improve the recall ratio and the precision ratio at the same time.

Keywords: ontology mapping, Entropy, decision-making

1. Introduction

These years, ontology has become a hot topic in the field of artificial intelligence, knowledge representation, Semantic Web, data integration and information retrieval. But because the creators of ontology use different methods, there must be disparity between ontologies created by different domain experts. The goal of ontology mapping is to solve the knowledge sharing and reuse problems of different ontologies. Recognizing this, ontology mapping policies are exploited according to entities similarity computing of different ontologies, and these entities have infinite variety types of information (e.g. semantic information, structure information). All the information can be regarded as the features of the ontology, besides the unitary mapping methods can't get the whole information about entities of the ontology, so multi-strategy is widely used by present mapping methods[1-4]. But most of the methods combined features simply, they do not give due consideration to the features of ontology and the similar

* Corresponding author.

 doi:10.1016/j.phpro.2012.03.365

properties inside the mapping entity pair. This paper proposes an entropy decision-making method to determine the combined weight of the different similarities.

2. Definition

This section provides preliminary definitions [5] used throughout the paper.

A. 2.1 Ontology

Definition 1 Ontology

Ontology is a six-tuple of the form:

$$O = \left\langle C, P, H^C, H^P, A, I \right\rangle \tag{1}$$

consisting of a set of concepts C and a set of properties P, respectively arranged in the hierarchies H^C and H^P that associate each concept c_i with its sub-concepts $Sub(c_i)$ and each property p_i with its sub-properties $Sub(p_i)$, respectively A is a set of axioms. The set I contains instances of concepts and properties.

B. 2.2 Similar function

The ontology enteritis is described with different ontology information, so first every mapping strategy computes a similarity according to different information, and then combined these similarities to find the final similarity between source ontology and target ontology.

Definition 2 Similarity function

Given two ontologies entities $e_s \in O_s$, $e_t \in O_t$ their overall similarity is computed by the following function:

$$Sim(e_s, e_t) := F\left(\sum_{1}^{n} simX_i(e_s, e_t)\right) \tag{2}$$

where each $simX_i$ is the nth similarity function implemented by an individual similarity calculation method and F is a function that combines the different similarity scores.

3. Similarity calculations

Generally speaking, in the definition of ontology, three sources of information can be recognized: (i) *linguistic*, (ii) *structural* and (iii) *extensional*. Here we use the information associated with these methods to compute the similarity between ontology entities. All entities of the two ontologies are the input, the output are entity similarity vectors computed by each feature, finally returns a similarity matrix S.

$$S = \left\{Sim_{ij}(e_s, e_t)\right\}_{m \times n} \tag{3}$$

where $Sim_{ij}(e_s, e_t)$ is similarity of entity pair computed by each feature, m is the number of entity pairs, n is the number of similarity features.

C. 3.1 Lexical similarity for labels and Ids

Let $\sum$ denote a thesaurus, and *syn (l)* the set of synonyms and *ant (l)* the set of antonyms of label *l*; the lexical similarity measure between the labels of e_s and e_t, $S_l(e_s, e_t)$ is then given as follows[6]:

$$S_l(e_s, e_t) = \begin{cases} 1.0, \text{ if } l_s = l_t \\ 0.99, \text{ if } l_t \in syn(l_s) \\ 0.0, \text{ if } l_t \in ant(l_s) \\ \text{Lin}(l_s, l_t), \text{ if } l_s \in \sum \wedge l_t \in \sum \wedge l_t \notin syn(l_s) \\ \dfrac{\text{tok}(l_s) \cap \text{tok}(l_t)}{\max(|\text{tok}(l_s)|, |\text{tok}(l_t)|)}, \text{otherwise} \end{cases} \tag{4}$$

$\text{Lin}(l_s, l_t)$ denotes the information theoretic similarity proposed by Lin in [7]; it provides a good measure of closeness of meaning between concepts within a thesaurus. The tokenization function $\text{tok}(l)$ extracts a set of tokens from the label l, $S_{id}(e_s, e_t)$ is the ids similarity measure used the same way as with labels, except that the Lin function is not used.

D. 3.2 Lexical similarity for comments

The lexical similarity for comments is compute used the following equation:

$$S_c(e_s, e_t) = 1 - \frac{\text{op}(x_s, x_t)}{\max(|\text{tok}(x_s)|, |\text{tok}(x_t)|)} \tag{5}$$

$S_c(e_s, e_t)$ as a variation of Levenshtein distance but applied to tokens. Let x_s, x_t be the comments of e_s, e_t respectively, and let $\text{op}(x_s, x_t)$ denote the number of token operations needed, and $\text{tok}(x)$ denote the number of tokens in a comment.

E. 3.3 Entity-set similarity

In instance-based mapping semantic relations between concepts of two ontologies are determined based on the overlap of their instance sets. The well-known formula of similarity measure is Jaccard's coefficient does not take into account the number of instances the degree of difference, when the number of instances between concepts unevenly distributed, the mapping results is easy to distortion, so presented here richness(*r*) and equipoise (*eq*) the two key factors:

$$r = \min\left\{1 - \frac{1}{(P_A + \alpha)}, 1 - \frac{1}{(P_B + \alpha)}\right\} \tag{6}$$

$$eq = \frac{P_A}{P_B} \tag{7}$$

Where P_A, P_B represent the instances set of e_s, e_t respectively, richness value with the increase of the number of instances, but with the increasing number of instances, richness of growth should slow down; α is to make the richness value is not too small when the number of instance is one. The richer of instances of two concepts have, the more reliable of the strategy based on instance.

Equipoise reflects the difference of the richness between the instances of two concepts, the difference is larger, and the equipoise is smaller. When equipoise value is small, even

if $\left| P_A \cap P_B \right| = \left| P_A \right|$, that is the instance of e_s is completely matched, the final similarity of the instance may never reach the threshold. To avoid this, equipoise value is small, the denominator with $2 \cdot \min\left(\left|P_A\right|, \left|P_B\right|\right)$ to replace the $\left(\left|P_A\right| \cup \left|P_B\right|\right)$ Therefore, the similarity based on instance is calculated as follows:

$$S_i = \begin{cases} r \cdot JaccardSim(e_s, e_t) & eq \geq E \\ r \cdot \dfrac{\left|P_A \cap P_B\right|}{2 \cdot \min\left(\left|P_A\right|, \left|P_B\right|\right)} & eq < E \end{cases} \tag{8}$$

F. 3.4 Structural similarity

According to Tous et al. [8], built-in RDF(S) and OWL properties can be modeled by a vector space model in which each property is represented as a dimension of a k -dimensional space where k is the number of built-in properties considered. Similarities between entities are collected into a similarity matrix S_s whose values are calculated by iterating the following updating equation:

$$S_{k-1} = B \times S_k \times A^T + B^T \times S_k \times A, \quad k = 0, 1, ...n \tag{9}$$

where each element s_{ij} of S_k represent the similarity between a source entity $e_i \in O_s$ and a target entity $e_j \in O_t$ at iteration k. A and B are the adjacency matrices of O_s and O_t, respectively. The algorithm stops when a predefined difference between S_k and S_{k-1} is reached.

G. 3.5 Semantic similarity

Semantic similarity between names of ontology entities by considering their semantic meaning by relying on the knowledge defined in the WordNet [9] lexical ontology. By exploiting WordNet can discover similarity among apparently unrelated terms (e.g., automobile and car) defined in the ontologies to be mapped. In more detail, the following similarity function was implemented:

$$S_{sm}(e_s, e_t) = \frac{2 \times IC(sub(e_s, e_t))}{IC(e_s) + IC(e_t)} \rightarrow [0, 1] \tag{10}$$

Eq. (11) is adaptation of the distance measure defined in [11]. The function *IC* that returns the information content(*IC*) of a concept is defined as in [10].

4. Entropy decision-making to adjust weights

Entropy in information system is the measure of Information disorder, greater the entropy is, higher Information disorder is, the utility value of information is smaller; conversely, smaller the entropy is, lower Information disorder is, the utility value of information is greater.

Using the similarity matrix mentioned in the previous section *S*, *m* entity pairs as samples, *n* features as evaluating indicator, and this paper aims to evaluate the effectiveness value n features. Because the similarity matrix *S* is the same in dimension and quantity, here we pass over the standardization, and use y_{ij} instead of $Sim_{ij}(e_s, e_t)$ to facilitate the description.

Using the formula (11) to compute information entropy value of feature *j* based on information entropy

theory.

$$E_j = -k\sum_{i=1}^{m} y_{ij} \ln y_{ij} \tag{11}$$

where constant k related to system sample size m, a system with completely disordered information, the degree of order is zero, and its entropy is maximum, $E = 1$ .when m samples are in completely disordered distribution $y_{ij} = \frac{1}{m}$, calculated by the formula (12):

$$E = -k\sum_{i=1}^{m} \frac{1}{m} \ln \frac{1}{m} = k\sum_{i=1}^{m} \frac{1}{m} \ln m = k \ln m = 1 \tag{12}$$

$$k = \frac{1}{\ln m} \tag{13}$$

Because information entropy E_j is used to measure the utility value of feature j, when completely disordered distribution, $E_j = 1$. Here, the information of E_j (data of feature j target) utility value of overall evaluation is zero. So, the information utility value of an index is dependent on its difference value h_j of information entropy E_j and 1.

$$h_j = 1 - E_j \tag{14}$$

It is clear that using entropy method to estimate the weight of each feature, its essence is to compute with the cost coefficient of the feature information (similarity), if the cost coefficient is higher, the more important it means to the final result, so weight of feature j is:

$$w_j = \frac{h_j}{\sum_{j=1}^{n} h_j} \tag{15}$$

Entropy method can determine the weight of feature according to the embodied effectiveness of similarity different features, and it can avoid the results deviation caused by human intervention or weighting simply in traditional methods to ensure the mapping precision ratio.

5. Experimental results

This section discusses in detail the results of the method; the set of experiments was done using the 2008 benchmark series of tests created by the Ontology Alignment Evaluation Initiative (OAEI).We used the provided partial alignment as a gold standard, then ran the method using two different combinations of weights: the set is used for the OAEI contest, which had to be the same set as those used for the benchmark, The results of this further investigation are presented in Table I, the standard weights used were $w_l = 0.2$, $w_{id} = 0.1$, $w_c = 0.1$, $w_{instance} = 0.1$, $w_{sematic} = 0.3$, $w_{structure} = 0.2$.The entropy weights compute were $w_l = 0.32$, $w_{id} = 0.05$, $w_c = 0.11$, $w_{instance} = 0.124$, $w_{sematic} = 0.22$, $w_{structure} = 0.176$.

TABLE I Total number of correspondences found for partial alignment in Anatomy test.

	Standard weights	Entropy weights
Correct correspondences found	831	884
Correspondences found but not in gold standard	409	427
Correspondences in gold standard not found	138	85
Precision	0.670	0.674
Recall	0.858	0.912

6. Conclusions

This paper proposes a dynamic mapping policy which analyzes the similar information of the entities, which use entropy decision-making method to determine the combined weight of the feature similarity; the combination should especially consider the influence on mapping results caused by mapping feature itself. Overall, this method can maintain the stability and the commonality, and improve the recall ratio and the precision ratio at the same time.

This paper supported by the National Natural Science Foundation of China under Grant No.60873044.

References

[1] A. Doan, J. Madhavan, R. Dhamankar, P. Domingos, and A.Y. Halevy, 'Learning to match ontologies on the semantic Web', VLDB Journal 12 (4) (2003) 303-319. 2003.

[2] Juanzi Li, Jie Tang, Yi Li, and Qiong Luo, 'RiMOM: A Dynamic Multistrategy Ontology Alignment Framework', IEEE Trans. Knowl. Data Eng. 21(8): 1218-1232. 2009.

[3] Yves R. Jean-Mary, E. Patrick Shironoshita, and Mansur R. Kabukaa, 'Ontology matching with semantic verification', Web Semantics: Science, Services and Agents on the World Wide Web 7 (2009) 235-251. 2009.

[4] M. Ehrig and Y. Sure, 'FOAM-A framework for ontology alignment and mapping', Proceedings of ISWC, Demo, 2005.

[5] Giuseppe Pirró and Domenico Talia. 'UFOme: An ontology mapping system with strategy prediction capabilities', Data Knowl. Eng, doi:10.1016/j. datak. 2009.12. 002. 2010.

[6] Yves R. Jean-Mary, E. Patrick Shironoshita, and Mansur R. Kabukaa, 'Ontology matching with semantic verification', Web Semantics: Science, Services and Agents on the World Wide Web 7. 235-251. 2009.

[7] D. Lin, 'An information-theoretic definition of similarity', Proceedings of 15th International Conference of Machine Learning (ICML), pp. 296-304. 1998.

[8] R. Tous, and J. Delgado, 'A vector space model for semantic similarity calculation and OWL ontology alignment', Proceedings of DEXA, pp. 307-316. 2006.

[9] G. Miller, 'WordNet an on-line lexical database', International Journal of Lexicography 3 (4) (1990) 235 – 312.

[10] N. Seco, T. Veale, J. Hayes, 'An intrinsic information content metric for semantic similarity in WordNet', in: Proceedings of ECAI, 2004, pp. 1089 – 1090

[11] E. Rahm, P.A. Bernstein, 'A survey of approaches to automatic schema matching', VLDB Journal 10 (4) (2001) 334 – 350.