An Improved Entity Similarity Measurement Method

Gang Lv*, Cheng Zheng, Sheng-bing Chen

1 Key of Intelligent Computing and Signal Processing, Ministry of Education, Anhui University, Hefei Anhui 230039, China
2 Key Laboratory of Network and Intelligent Information Processing, Hefei University, Hefei Anhui 230601, China
*Corresponding author, e-mail: lvgang@hfuu.edu.cn, chengzheng@ahu.edu.cn, chensb@hfuu.edu.cn

Abstract

To facilitate the integration of learning resources categorized under different ontology representations, the techniques of ontology mapping can be applied. Though many algorithms and systems have been proposed for ontology mapping, they do not have an automatic weighting strategy on class features to automate the ontology mapping process. A novel method of computing the feature weights is proposed. By feature semantic analysis, the different entities similarity calculation model and weight calculation model were defined. The results show that it makes the ontology mapping process more automatic while retaining satisfying accuracy. Improve ontology mapping effectiveness.

Keywords: semantics of features, ontology mapping, feature weight

1. Introduction

Being one of the best instrument of knowledge presentation and the basis of semantic web technologies, ontology is mainly described with RDF (Resource Description Framework) and OWL (Ontology Web Language) released by W3C besides CYCL, DOGMA, F-Logic and the like developed and used by other organizations. Currently, domain ontology has been applied in many fields such as artificial intelligence, software engineering, library science and semantic web[1],[2]. The resources represented by different ontologies in different fields would be integrated and classified via ontology mapping. As the key factor of ontology mapping, the entity similarity measurement can be generally divided into three methods with different bases, namely, terminology, structure and semantics. Besides, the process of mapping can also be classified into three types, namely, manual, semi-automatic and automatic [3],[4].

Influenced by factors such as classification scheme, representation language, and background knowledge, the ontology in a same field may appear quite different. Therefore, when studying the issue of ontology mapping, besides the researches on the class matching of different entities, the features (i.e. relations) between them also matters. Generally, the system of ontology mapping possesses two strategies, namely, single strategy and multi-strategy [5],[6]. When multi-strategy is adopted, different similarity measurements shall be combined into a single one properly. During the process, most weight distribution of resources is made based on the experiences or experiments of the experts nowadays, while this method remains time-consuming and unstable when used in Web resources represented by different ontologies[7],[8].

Ontology mapping is a kind of process in which the entity of the source ontology (including class and features) would be mapped and represented by a target ontology, and the similarity measurement also includes other related entities owing to certain relational features besides the entity itself. A concept of “universality” among classes in ontology representation is proposed in this thesis: if a feature possesses a high universality, the partition degree of a class would become low and the similarity would thus remain indistinguishable, namely, the larger a feature’s universality becomes, the smaller the weight will get. And the following comes the detailed explanations.

2. Semantics Features

Since the ontology possesses many feature types such as tags, annotations, attributes, relations (parent class and subclass) and examples, the distinctive feature among entities is called “uniqueness” [9],[10]. As a hypothesis, if the ontology of a feature is unique, meanwhile

Received August 24, 2014; Revised November 1, 2014; Accepted November 17, 2014
there is a class with same features in another ontology, then we consider the above ontologies equal to each other. Just as we can easily distinguish human beings when they were put into a group of animal by noticing the feature of “thought”, since they are the only species who possess the ability of thinking. On the contrary, since they maintain the same features, it is hard to distinguish them when in a crowd.

This thesis defines $\text{Com}_f(c_1, c_2)$ as two different semantic forms which represent two feature-based ontologies with different feature types in related semantics similarities. For instance, related semantics for the string type “tag” and “annotation” might be a set of synonyms, while the relational semantics of a related feature might be a set of classes which connect via certain relations. If $\text{value}(c, f)$ is defined to express Feature f’s value of Class c, and $\text{sem}(f, c)$ to express Feature f’s semantic associated value of Class c, the formula for the value of Feature f, ontology $c_1$ and $c_2$ and $\text{Com}_f(c_1, c_2)$ can be defined as follows:

$$\text{Com}_f(c_1, c_2) = \frac{\text{sem}(f, c_1) \cap \text{sem}(f, c_2)}{\text{sem}(f, c_1) \cup \text{sem}(f, c_2)}$$

(1)

Among which $\text{sem}(f, c_1)$ and $\text{sem}(f, c_2)$ are respectively the synonyms of $\text{value}(c_1, f)$ and $\text{value}(c_2, f)$. Besides, the similarities between binding property $f$, $c_1$ and $c_2$ can also be defined as follows:

$$\text{Com}_f(c_1, c_2) = \begin{cases} 1 & \text{if } \text{value}(c_1, f) = \text{value}(c_2, f) \\ 0 & \text{otherwise} \end{cases}$$

(2)

Moreover, as for the value of relational features such as “parent class”, “subclass” and “example”, it can be considered as a collection of ontologies which originate from a certain feature. And the relational features of $c_1$ and $c_2$ could thus be defined as the follows:

$$\text{com}_f(c_1, c_2) = \frac{\text{value}(c_1, f) \cap \text{value}(c_2, f)}{\text{value}(c_1, f) \cup \text{value}(c_2, f)}$$

(3)

The value of conceptual feature $\text{Com}_f(c_1, c_2)$ drawn from the above calculation can be used to calculate the value of weight of a feature. If we define O as the ontology, C for a set of entities which belongs to O, $F$ for a group of features of C, which include “tags”, “annotations”, “parent class”, “binds”, “relations”, “examples” and the like, the $\text{Com}_f(c_1, c_2)$ definition of a feature is as follows:

$$CM_f = \frac{\sum_{c_i, c_j \in C} \text{com}_f(c_i, c_j)}{n^2}$$

(4)

Among which $n$ represents the amount of classes in $C$; and $c_i$, $c_j$ are the classes of $C$, the weight of Feature $f$ could thus be defined as the follows:

$$W_f = 1 - CM_f$$

(5)

3. Similarity Measurement

If the weight of the entity features represented by the two ontologies is worked out, the similarity measurement of different classes can be calculated by integrating various feature weight, for the similarity measurement of class and feature inter influences each other during the process[11],[12]. Since a class is described by a set of features, the similarity of features
should be taken into account when doing similarity measurement. When make ontology mapping, iterative algorithm would be adopted in this thesis.

Definitions: original ontology \( O_1 = \{C_1, F_1\} \), target ontology \( O_2 = \{C_2, F_2\} \), \( c_1 \) and \( c_2 \) for class collection, \( F_1 \) and \( F_2 \) for feature collection. \( e^1_i \) and \( e^2_j \) for entities, and the classes and features also belongs to their ontologies \( O_1 \) and \( O_2 \). In order to benefit the expression of the aforementioned algorithm, related variables are defined as follows: \( \text{Sim}_k(e^1_i, e^2_j) \) for the entity, and the similarity weight for \( e^1_i \) and \( e^2_j \) would be worked out after applying the iterative algorithm for \( k \) times and it would also be represented by \( ISim_k(e^1_i, e^2_j) \).

\[
ISim_k(e^1_i, e^2_j) = \frac{\text{Sim}_k(e^1_i, e^2_j) \cap \text{Sim}_k(e^2_j, e^1_i)}{2}
\]

The entity \( e^1_i \) which described by a set of feature collection can be defined as \( F(e^1_i) = \{f_1(e^1_i), f_2(e^1_i), \ldots, f_l(e^1_i)\} \), among which \( f_t \in F_1, t = 1, \ldots, l \). Another entity which described by a set of feature collection can also be defined as \( F(e^2_j) = \{g_1(e^2_j), g_2(e^2_j), \ldots, g_m(e^2_j)\} \), among which \( g_t \in F_2, t = 1, \ldots, m \). After applying the iterative algorithm \( ISim_k(e^1_i, e^2_j) = \max_m ISim_k(e^1_i, e^2_m) \) for \( k \) times, an adjusted result \( A_k(e^1_i) = e^2_j \) can be worked out via calculation. In order to calculate \( ISim_k(e^1_i, e^2_j) \), we defines \( VF = A_k(g_1), A_k(g_2), \ldots, A_k(g_m) \) to adjust \( F(e^2_j) \) to \( F(e^1_i) \). And related Formula 7 is as follows:

\[
ISim_{k+1}(e^1_i, e^2_j) = \frac{\sum_{f \in F, f \in F_1} w_f \times \text{Sim}_k(f, A_k(f)) \times \text{SIM}_k(\text{value}(e^1_i, f), \text{value}(e^2_j, A_k(f)))}{\sum_{f \in F} w_f}
\]

Among which the attribute value of \( \text{SIM}_k \) is based on its type:

1. If \( X \) and \( Y \) are not in the same type, then \( \text{SIM}_k(X, Y) = 0 \)
2. If \( X \) and \( Y \) are in the same type such as “character type” or “numeric type” and \( X=Y \), the \( \text{SIM}_k(X, Y) = 1 \), otherwise:

\[
\text{SIM}_k(X, Y) = \frac{\text{sem}(f, X) \cap \text{sem}(f, Y)}{\text{sem}(f, X) \cup \text{sem}(f, Y)}
\]

3. If \( X \) and \( Y \) are both entity sets, then:

\[
\text{SIM}_k(X, Y) = \frac{\sum_{e \in X, f \in Y} \max_{e \in X} \{\text{Sim}_k(e_1, e_2)\} \delta_{X, Y}}{\max(|X|, |Y|)}
\]

As is shown in Figure 1 the ontology representation of synonyms, the formula of similarity measurement of “Book”, an entity in the source ontology and the one in target ontology is as follows:

\[
\text{Sim}_k(X, Y) = \frac{\sum_{e \in X, f \in Y} \max_{e \in X} \{\text{Sim}_k(e_1, e_2)\} \delta_{X, Y}}{\max(|X|, |Y|)}
\]
\[ Sim_{k+1}(Book_{01}, Book_{02}) = \frac{W_{\text{sup \ class}} \times Sim_{k}(r, A_{k}^{-1}(r)) \times SIM_{k}\left(\frac{\text{value}(Book_{01}, r)}{\text{value}(Book_{02}, A_{k}^{-1}(r))}\right) + \cdots}{W_{\text{sup \ class}} + \cdots} \]

\[ = \frac{\text{Similarity OnSuperClass}_{ss} + \text{Similarity OnLabel}_{ss} + \text{Similarity OnSubClass}_{ss}}{W_{\text{sup \ class}} + W_{\text{label}} + W_{\text{sub \ class}}} \]

Among which Similarity On Superclass, Similarity On Label and Similarity On Subclass are the corresponding similarity of features (Sim_k) and the feature weight (W) for features, super-class and sub-class by running the similarity measurement.

Figure 1. The Ontology Representation of a Sample

Figure 2. Comparison on Experimental Results

During the process of round-robin, if nearest adjustment function \(A_{k+1}\) and the similarity function \(Sim_{k+1}\) are in the same value with \(A_k\) and \(Sim_k\), then end the circulation and iteration. The adjustment algorithm is as follows:

**PROCEDURE:** Ontology Mapping

**INPUT:** Ontology \(O_1, O_2\)

**OUTPUT:** Alignment \(A\)

**BEGIN**

- \(W_1 = \text{ComputeWeight}(O_1)\)
- \(W_2 = \text{ComputeWeight}(O_2)\)
- \(A_0 = \text{ComputeInitialAlignment}(O_1, O_2)\)
- \(Sim_0 = \text{ComputeInitialSimilarity}(O_1, O_2, A_0)\)
- \(k = 1\)

**WHILE** \(k \neq -1\)

- **FOR** \(e_i \in O_1\)
- **FOR** \(e_j \in O_2\)
- **PUT**(\(Sim_k, \text{ComputerSimilarity}(e_i, e_j, A_{k-1})\))
- **END FOR**
- **END FOR**
- \(A_k = \text{GetAlignment}(Sim_k)\)

**IF** \(Sim_k \approx Sim_{k-1} \text{ AND } A_k \approx A_{k-1}\) **THEN**

- \(k = -1\)

**ELSE**

- \(k = k + 1\)

**END IF**

**END WHILE**

**OUTPUT**(\(A\))
END BEGIN
END_PROCEDURE

As is shown in the above algorithm, if the respective entity number for Ontology \( O_1 \) and \( O_2 \) is \( n \) and \( m \), the time complexity of the very algorithm would be \( O(n \times m) \).

4. Experimental results and Analyses

The test data of this thesis is OAEI 2009 Corpus (http://oaei.ontologymatching.org/), and the evaluation of the performance standard remain to be precision rate, recall rate and F-measure [13]. The respective definitions of the calculation formula are as follows:

\[
p = \frac{|\text{extracted} \cap \text{standard}|}{|\text{extracted}|}
\]

\[
r = \frac{|\text{extracted} \cap \text{standard}|}{|\text{standard}|}
\]

\[
F = \frac{2pr}{p + r}
\]

The test data includes 33 identified classes, 24 relations, 44 attributes, 56 examples and 20 examples with no attribute. The experiment has also compared the proposed method (AFW) with Lily, MapPSO, and TaxoMap, and as is displayed in Figure2 below, owing to the adoption of automatic feature weight calculation, the matching efficiency and the three performance standards have been improved significantly.

5. Conclusion

By emphasizing the importance to represent features via the method of weight and analyzing the semantics of features, this thesis has designed the computing model of entity weight and calculated the similarity weight among various relations. Due to the adoption of iteration method and automatic feature weight calculation, the Ontology-mapping efficiency has been improved in related experiments. Besides, it also possesses better characteristics in precision rate, recall rate and F-measure when comparing with other systems. Priorities would be given on the studies of improving the robustness and adjustable capability of the algorithm in the near future.

Acknowledgement

Project was supported by the Nature Science Foundation of AnHui (2013SQRL074ZD, 1408085MF135).Key Constructive Discipline of Hefei University, No. 2014xk08, Training Object for Academic Leader of Hefei University, No. 2014dtr08

References


