TELKOMNIKA, Vol.12, No.4, December 2014, pp. 1017~1022
ISSN: 1693-6930, accredited A by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v12i4.434

An Improved Entity Similarity Measurement Method

Gang lv*[1,2], Cheng Zheng[2], Sheng-bing Chen[3]
[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[1], chengzheng@ahu.edu.cn[2], chensb@hfuu.edu.cn[3]

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 $Com_f(c_1, c_2)$ 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 value(c,f) is defined to express Feature f's value of Class c, and 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 $Com_f(c_1, c_2)$ can be defined as follows:

$$Com_f(c_1, c_2) = \frac{|sem(f, c_1) \cap sem(f, c_2)|}{|sem(f, c_1) \cup sem(f, c_2)|} \quad (1)$$

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

$$Com_f(c_1, c_2) = \begin{cases} 1 & if\ value(c_1, f) = value(c_2, f) \\ 0 & otherwise \end{cases} \quad (2)$$

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:

$$com_f(c_1, c_2) = \frac{|value(c_1, f) \cap value(c_2, f)|}{|value(c_1, f) \cup value(c_2, f)|} \quad (3)$$

The value of conceptual feature $Com_f(c_1, c_2)$ 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 $Com_f(c_1, c_2)$ definition of a feature is as follows:

$$CM_f = \frac{\sum_{c_i, c_j \in C} com_f(c_i, c_j)}{n^2} \quad (4)$$

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:

$$W_f = 1 - CM_f \quad (5)$$

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_i^1 and e_j^2 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: $Sim_k\left(e_i^1, e_j^2\right)$ for the entity, and the similarity weight for e_j^1 and e_i^2 would be worked out after applying the iterative algorithm for k times and it would also be represented by $ISim_k\left(e_i^1, e_j^2\right)$.

$$ISim_k\left(e_i^1, e_j^2\right) = \frac{\left(Sim_k\left(e_i^1, e_j^2\right) \cap Sim_k\left(e_j^2, e_i^1\right)\right)}{2} \tag{6}$$

The entity e_i^1 which described by a set of feature collection can be defined as $F(e_i^1) =< f_1\left(e_i^1\right), f_2\left(e_i^1\right), \ldots f_l\left(e_i^1\right) >$, 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 $F(e_j^2) =< g_1\left(e_j^2\right), g_2\left(e_j^2\right), \ldots g_m\left(e_j^2\right) >$, among which $g_t \in F_2$, $t = 1\ldots, m$. After applying the iterative algorithm $ISim_k\left(e_i^1, e_j^2\right) = \max_m ISim_k\left(e_i^1, e_m^2\right)$ for k times, an adjusted result $A_k\left(e_i^1\right) = e_j^2$ can be worked out via calculation. In order to calculate $Sim_{k+1}\left(e_i^1, e_j^2\right)$, we defines $VF =< A_k(g_1), A_k(g_2), \ldots, A_k(g_m) >$ to adjust $F\left(e_j^2\right)$ to $F\left(e_i^1\right)$. And related Formula 7 is as follows:

$$Sim_{k+1}\left(e_i^1, e_j^2\right) = \frac{\sum_{f \in VF \cap F\left(e_i^1\right)} W_f \times Sim_k\left(f, A_k^{-1}(f)\right) \times SIM_k\left(value\left(e_i^1, f\right), value\left(e_j^2, A_k^{-1}(f)\right)\right)}{\sum_{f \in F\left(e_i^1\right)} W_f}$$

Among which the attribute value of SIM_k is based on its type:

(1) If X and Y are not in the same type, then $SIM_k(X,Y)=0$

(2) If X and Y are in the same type such as “character type” or “numeric type” and X=Y, the $SIM_k(X,Y)=1$, otherwise:

$$SIM_k(X, Y) = \frac{\left|sem(f, X) \cap sem(f, Y)\right|}{\left|sem(f, X) \cup sem(f, Y)\right|}$$

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

$$SIM_k(X, Y) = \frac{\sum_{e_1 \in X} \max_{e_2 \in Y} \left\{Sim_k(e_1, e_2)\right\}}{\max\left(|X|, |Y|\right)}$$

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:

$$Sim_{k+1}(Book_{01}, Book_{02}) = \frac{W_{r=\sup erclass} \times Sim_k(r, A_k^{-1}(r)) \times SIM_k\begin{pmatrix} value(Book_{01}, r) \\ value(Book_{02}, A_k^{-1}(r)) \end{pmatrix} + \cdots}{W_{\sup er\_class} + \cdots}$$

$$= \frac{Similarity\ OnSuperClass + Similarity\ OnLabel + Similarity\ OnSubClass}{W_{\sup er\_class} + W_{label} + W_{sub\_class}}$$

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.

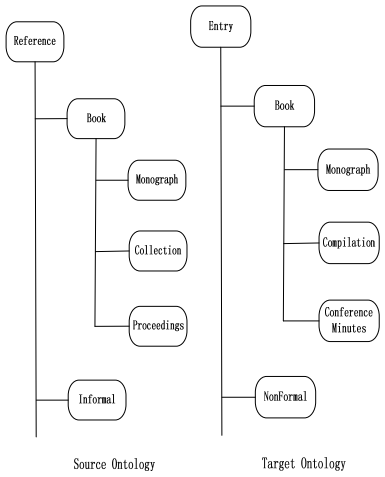

Figure 1. The Ontology Representation of a Sample

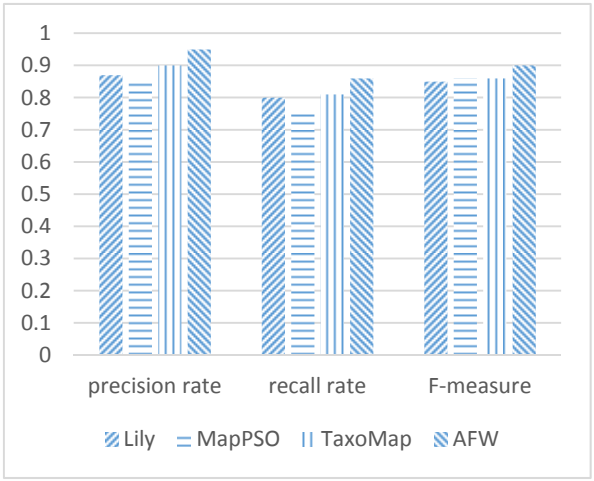

Figure 2. Comparison on Experimental

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 O1,O2
OUTPUT: Alignnment A
 BEGIN
   W1=ComputeWeight(O1)
   W2=ComputeWeight(O2 )
   A0=ComputeInitialAlignment(O1, O2)
     Sim0=ComputeInitialSimilarity (O1, O2,A0)
     k = 1
     WHILE k≠-1
       FOR ei in O1
         FOR ej in O2
          PUT(Simk, ComputerSimilarity(ei , ej , Ak-1))
         END_FOR
        END_FOR
        Ak= GetAlignment(Simk)
        IF Simk ≒ Simk-1 AND Ak ≒ Ak-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×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:

$$p = \frac{|extracted \cap standard|}{|extracted|}$$

$$r = \frac{|extracted \cap stanard|}{|standard|}$$

$$F = \frac{2pr}{p + r}$$

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

[1] Zhou Sheng-chen, Qu Wen-ting, Shi Ying-zi, Shi Xun-zhi, Sun Yun-chen. Overview on Sentiment Analysis of Chinese Microblogging. *Computer Applications and Software.* 2013; 30(3): 161-164.
[2] Gang Lv, Cheng Zheng. A novel framework for concept detection on large scale video database and feature pool. *Artificial Intelligence Review*. 2013; 40(4): 391-403
[3] AH. Doan, J. Madhavan, P. Domingos, A. Halevy. Ontology matching: A machine learning approach. *Handbook on Ontologies in Information System*s. 2003: 397–416.
[4] Xiong Fang, Huang Hong-bin, Huang Yu-cheng. An Approach of Information Semantic Clustering Based on Semantic Similarity. *Computer Engineering and Science*. 2012; 34(11): 175-179.

[5] Jiang Men-jin, Zhou Ya-qian, Huang Xuan-jing. Synonymous Entity Expansion Based Information De-duplication*. Journal of Chinese Information Processing*. 2012; 26(1): 42-50.
[6] Cui Xiao-Jun, Xiao Hong-yu, Ding Li-xin. Distance-Based Adaptive Record Matching for Web Databases*. Journal of Wuhan University (Natural Science Edition)*. 2012; 12(1): 1-9
[7] Zhao Hai-xia, Li Dao-shen, LIU Yong, et al. Research on entity extraction method of Deep Web data integration. *Computer Engineering and Applications*. 2012; 48(36): 160-163.
[8] Erlin E, Rahmiati R, Rio U. Two Text Classifiers in Online Discussion: Support Vector Machine vs Back-Propagation Neural Network. *TELKOMNIKA Telecommunication Computing Electronics and Control.* 2014; 12(1): 189-200.
[9] Qi Yu-dong, Yan Xiao-bin, Xie Xiao-fang. Conceptual models similarity computation based on LISA theory. *Computer Engineering and Applications*. 2012; 48(3): 40-42.
[10] Dong Deng-hui, Xiao Gang, Zhang Yuan-ming, Lu Jia-wei, Xu Jun. An SOA Reference Model Based on Multi-granularity Service Library and Its Application. *Computer Applications and Software*. 2012; 29(10): 152-155.
[11] Sun Ming, Lu Chun-sheng, Xu Xiu-xing. A Web Entity Information Extraction Method Based on SVM and AdaBoost. *Computer Applications and Software.* 2013; 30(4): 101-106.
[12] Albarda A, Supangkat S H, Kuspriyanto K, et al. Information Interchange Layer based on Classification of Information Use (IU). *TELKOMNIKA Telecommunication Computing Electronics and Control.* 2014; 12(2): 485-492
[13] Y.s R. Jean-Mary, EP. Shironoshita, et al. Ontology matching with semantic verification. *Journal of Web Semantics.* 2009; 21(4): 121-135.