http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 13 (2): 397-400, 2014
ISSN 1812-5638 / DOI: 10.3923/itj.2014.397.400

An Improved Method for Ontology Instances Similarity Computation

[1,2]Gang Lv and [1]Cheng Zheng
[1]Key Laboratory of Intelligent Computing and Signal Processing, Ministry of Education, Anhui University, Hefei, 230039, China
[2]Key Laboratory of Network and Intelligent Information Processing, Hefei University, Hefei, 230601, China

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, Instances tend to better reflect the similarity between two ontologies, so instances similarity can be used to optimize the mapping. In the study, similarity calculation based on Bayesian theory was proposed, two concepts of the set similarity and the probability similarity were defined and model was derived, get a complete solution to an instance of mapping programs. It has been proved that the evaluation of similarity between ontologies is more accurate by considering both semantic similarity and semantic relativity.

Key words: Ontology mapping, instance similarity, bayesian probability

INTRODUCTION

Ontology mapping is a key problem in the fields such as information integration, semantic Web and knowledge management. The way that based on the similarity computing is the most frequently used method when conducting ontology mapping (Xia *et al*., 2010). In some occasions only the use of instance mapping can find the right pair of ontology mapping, because it is instances that truly reflect the semantic meaning of concept node (Ehrig and Sur, 2004). Thus could take advantage of the instance similarity to refine the mapping effect.

The core link of the semantic Web is retrieval and similarity computing is the core technology of achieving semantic retrieval. The quality of calculating ways defines the quality and reliability of instance retrieval (Li *et al*., 2010). Hence similarity computing plays an extremely important role when constructing and managing the system that based on the instance-based reasoning. Many instance similarity computing rely on the attribute characteristics of instances, employing similarity function to calculate the similarity between different instances (Kuster *et al*., 2007). Some are measuring by calculating the common parts within the two concepts of the proportion of whole instance set. Lv and Zheng (2011) puts forward the concepts of the richness and the equipoise to make the calculation of instance similarity come true. However, the instance mapping mentioned above can be easily affected by the content of instances which can lead to the instability of retrieval. This study based on the Bayes method presents a quite effective way of calculating ontology instance similarity.

METHODS

Actually, an instance is a sort of text set of concepts. To some extent, the common elements of two instance sets reflect the similarity. Meanwhile, for that on many occasions one element is composed of several phrases it must be pretreated (Tang *et al*., 2006). Bayesian decision makes use of subjective probability to estimate the partially unknown state under the incomplete information phenomenon and then Bayes formula is used to revise the probability of happening and finally make the most beneficial decision by expected value and revised probability. Through this process precision ratio and recall ratio could be considerably improved.

Instance pretreatment: The basic equation of Bayesian theory is:

$$P(A|B) = P(B|A) * P(A)/P(B)$$

the concept probability is used to represent similarity which means that the similarity between A and B equals the probability of similarity between A and B.

Corresponding Author: Gang Lv, Key Laboratory of Intelligent Computing and Signal Processing, Ministry of Education, Anhui University, Hefei, 230039, China Tel: 13855126906

Definition 1: Instance text set: A set that is composed of certain concept corresponding several related instances, recorded as Ic.

Every concept contained in the ontology does have its own instance text set. And the first step of instance pretreatment is to construct instance text set for the ontology. At first, need to compose text set respectively for the element value which each concept node was corresponded in every instance of the text trees, to get the instance text set of the ontology Next, need to separate the multiple-word in the set so as to get a new instance text set composed with multiple-word collection.

The basic theory of calculating instance similarity, mainly based on two assumptions:

- The more common instances the two concepts contain, the more similarities they have
- The more frequently the elements of certain instance set appear in another instance set, the more similarities they have

As for the assumption (1), can regard these two instance sets as two simple sets, calculate the ratio of the intersection and union as similarity; As for the assumption (2), can get the similarity based on Bayesian Theory via calculating the frequency of occurrence of the elements of certain instance text set appearing in another instance text set based on the Bayesian Decision Theory (Euzenat *et al*., 2006). At last, can combine these two results of calculations.

Set similarity

Definition 2: Set similarity: The set similarity of two instance sets is the calculated ratio of the intersection and the union, regarding these two instance sets as sets. The greater the ratio is, the greater proportion of the common parts is in the two instance text sets which means the more similar the two instance text sets are. Vice versa counter. However, in a word-separated instance text set, a multi-word is separated into several lemmas. Therefore, there are two ways to calculate the set similarity:

- Regard all the lemmas as one set
- Express each multi-word as a set of lemmas and the instance text set is the collection of those sets of lemmas

The first design separates each multi-word so that each lemma is definitely alone which neglects the original meaning of multi-words. In this way, the result is single-faceted and lack of precision. Therefore, this study adopted the second solution. Hence, the two instance text sets are transfer into sets of sets. When looking for common elements, so long as a certain multi-words set contains a common lemma, this multi-words set is a common element of two instance text sets.

Assuming that these two instance text sets are, respectively s_i and s_j, there: $s_i = \{g_1, g_2, g_3, \ldots\ldots, g_m\}$, $s_j = \{g_z, g_2, g_3, \ldots\ldots, g_n\}$. While $g_i = \{w_1, w_2, w_3, \ldots\ldots, w_p\}$, here g_i is a multi-words set and $w_i(1 <= I <= p)$ is the lemma. As for the element which itself is a lemma, g_i is the self-formed set at the length of 1.

Then define $s_i * s_j$ and $s_i \cup s_j$: $g_k \in (s_i * s_j)$, if and only if $\wedge\exists \{w_p \in g_k \wedge g_k \wedge (\exists g_t \in sj \rightarrow w_p \in gt_j\}$ or $g_k s \in j_k, \wedge(\exists g_t \in s_i \rightarrow w_p \in g_t\}$, that is $s_i * s_j$ is a set constituted by the multi-words sets containing common lemmas in s_i and s_j.

Finally, the set similarity of instance text set is $p_i(s_i, s_j) = |s_i * s_j| / |s_i \cup s_j|$.

Probability similarity: Supposing that two instance text sets s_i and s_j, there $s_i = \{g_1, g_2, g_3, \ldots\ldots, g_m\}$, $s_j = \{g_1, g_2, g_3, \ldots\ldots, g_n\}$, $g_i = \{w_1, w_2, w_3, \ldots\ldots, w_p\}$, where g_i is a multi-words set, $w(1<=i<=p)$ is the lemma. The frequencies that the lemmas of s_i occur in s_j and the lemmas of s_j occur in s_i both reflect the similarity of two instance text sets to some extent. Therefore, the two frequencies can be used to calculate the similarity. This study brings about the following two concepts to express the two frequencies mentioned above:

- **Definition 3: Weight:** The weight of certain multi-words set is the proportion of its intersecting multi-words sets in the contained instance text set
- **Definition 4: External weight:** The proportion of multi-words sets which are intersected with another multi-words set in or outside its contained instance text set which indicates the weight of this multi-words set in the instance text set

Weight reflects the importance of multi-words sets in the contained instances text set. External weight refers to the importance of any multi-words set in an instances text set. Usually this multi-words set originates from another instance text set. When calculating the similarity of instances, this multi-words set originates from mapping another instance text set.

As for the two instances text sets s_i and s_j, there: The similarity degree that s_i maps to the s_j and the similarity degree that s_j maps to the s_i, are usually not the same, that is to say, this kind of similarity degree has a feature of asymmetry. This study puts forward a concept of "one-way probability similarity" which is used to describe the similarity degree that certain instance text set maps to another instance text set and it is directional. For example, can regard the one-way probability similarity of s_i to s_j as $p(s_i|s_j)$. It is decided by external weight that every

multi-words set of s_i in the s_j and weight that this multi-words set in the S_k. Based on Bayesian basic formula, can come to:

$$p(s_i \mid s_j) = \frac{\sum_{\xi=1} p(s_i \mid g_j) * p(g_j)}{\text{count}(s_j)} \quad (1)$$

$$p(s_i \mid g_j) = \frac{\underset{w \cap g_j \neq \varnothing}{\underset{w \in s_i}{\text{count}(w)}}}{\underset{w \in s_i}{\text{count}(w)}} \quad (2)$$

$$p(g_j) = \frac{\underset{w \cap g_j \neq \varnothing}{\underset{w \in s_j}{\text{count}(w)}}}{\underset{w \in s_j}{\text{count}(w)}} \quad (3)$$

There:

$$\underset{w \cap g_j \neq \varnothing}{\underset{w \in s_i}{\text{count}(w)}}$$

represents the number of the w contained in the s_i whose intersection with g_i is not null.

$$\underset{w \in s_i}{\text{count}(w)}$$

represents the number of the w contained in the s_i. g_j originates from s_j. $p(s_i \mid g_j)$ represents the external weight of g_j in the s_i. $p(g_j)$ represents the weight of g_j in the s_j. However, the probability similarity of s_i and s_j is made up of one-way probability similarity s_i to s_j and s_j to s_i:

$$p_k(s_i, s_j) = [p(s_i \mid s_j) + p(s_j \mid s_i)]/2$$

Based on the theory above, can calculate the similarity of two instance text sets via the set similarity and probability similarity:

$$p_k(s_i, s_j) = [p_1(s_i, s_j) + p_k(s_i, s_j)]/2$$

EXPERIMENTS AND ANALYSIS

Nowdays, some studies proposed instance similarity computings rely on the attribute characteristics of instances and employing similarity function to calculate the similarity between different instances. Some are measuring by calculating the common parts within the two concepts of the proportion of whole instance set. To improve the efficiency of ontology mapping with large scale multi-ontology, a method of multi-ontology mapping based on concept classification was proposed. Granular computing and semantic similarity computation are property of the classification trees; the process of concept classification was thronged by Quick Sort algorithm. Experiments show that the proposed method guarantees the accuracy while reducing the concept of mapping the number of comparisons, reducing the complexity and plan is feasible (Gang *et al*., 2011). This study has conducted experiments and proved that the optimization strategy proposed in the study can improve the effect of the basic strategies. Meanwhile, the source of the experimental data selected by this study is from Test Ontologies and Alignments which is based on the open-source resources-Frame Work for Ontology Alignment and Mapping (http://www.aifb.uni-karlsruhe.de/WBS/meh/foam/). This study provides 14 available ontologies and results of the different ontology mappings. And this study uses russia1.owl and russia2.owl as a data source. The tools used throughout the experiment are mainly Jena, VC++6.0 and protégé3.1. Russia1 and russia2 bring some instances on their own while the remaining instances are created manually in the protégé. In this experiment, each category creates 30 to 40 instances and the scale of the whole instances reaches around 3600. After calculation, the set similarity, the one-way probability similarity, the probability similarity of each node can be obtained and the instance similarity can be finally obtained as well. Table 1 shows a part of results. The probability similarity is obtained from the average value of the two one-way probability similarities. Meanwhile, in the table, the data filled in the "Probability

Table 1: Results of set similarity, one-way probability similarity, probability similarity of each node based on Instance text set s_1 and s_2 use MP_Bayes method

Instance text set s_1	Instance text set s_2	Set similarity	Probability similarity	Instance similarity
President	President	0.78	0.69 (0.71,0.67)	0.73
Food	Food	0.83	0.79 (0.79,0.79)	0.81
Plant	Plant	0.90	0.98 (0.98,0.98)	0.94
Traveller	Normal_traveler	0.82	0.74 (0.69,0.79)	0.78
Health_risk	Disease_type	0.73	0.71 (0.76,0.66)	0.72
Document	Document	0.36	0.60 (0.63,0.57)	0.48
President	Food	0.02	0.02 (0.02,0.02)	0.02
Plant	Food	0.81	0.71 (0.68,0.74)	0.76
Traveller	Document	0	0 (0,0)	0
Drink	Restaurant	0	0 (0,0)	0
Approval	Certificate	0.84	0.92 (0.91,0.93)	0.88
Unit	Unit	0.91	0.93 (0.92,0.94)	0.92

Table 2: Comprehensive data comparison between MP_Bayes system and other systems

	Rimon		SNAX_Map		MP_Bayes	
Test set	Recall ratio	Precision ratio	Recall ratio	Precision ratio	Recall ratio	Precision ratio
1	0.75	0.75	1.00	1.00	1.00	1.00
2	0.77	0.91	0.83	0.92	0.84	0.94
3	0.66	0.83	0.81	0.82	0.81	0.93

*Rimon and SNAX_MAP are two commonly used method name, *MP_Bayes is the name of the method proposed in this study

Similarity" column also lists one-way similarity probability. For example, in the first line, 0.69 (0.71, 0.67) shows that the one-way probability similarity of s_1 to s_2 is 0.71 and the one-way probability similarity of s_2 to s_1 is 0.67 while the probability similarity of the s_1 and s_2 is 0.69. Since this mapping algorithm can be applied to multi-strategy mapping system, so it is temporarily called MP_Bayes. In Table 2, 1-3 shows the centralized ontology number of the standard test data. The test results are respectively compared with the SNAX_Map methods and the test result of Rimon system which are proposed in reference 1. About the recall ratio and precision ratio, this study adopts the following definition:

Recall ratio: The ratio between the number of correct mapping pairs and the number of standard mapping pairs in the mapping results.

Precision ratio: In the mapping results, the ratio between the number of correct mapping pairs and the number of mapping pairs in the union set of the mapping results and standard results.

From result of the experiment, the only difference between MP_Bayes system and SNAX_Map system is the use of different methods of instance mapping. MP-Bayes system adopts the improved instance mapping method put forward in this study. Table 2 proved the highest recall ratio and precision ratio was from MP-Bayes system. That is to say, the improved multi-strategy instance mapping method not only guarantees the recall ratio but also improves the precision ratio obviously, so that the quality of final mapping result was also improved. At the same time, the performance of MP-Bayes system is more stable when dealing with different groups of testing data.

CONCLUSION

Concept similarity calculation is the essential base in many technologies including ontology mapping, service discovery and semantic retrieval. On the base of analyzing the existing functions of ontology instance mapping, this study proposed an improved ontology instance mapping method, making the most of Bayesian Theory, compensating the inherent defects of the instance mapping strategies, so that the deletion or repetition of some information would not affect the whole mapping system too much. The experiment indicates that the function of the system which applies the new calculation method improved a lot compared with that of the traditional multi-strategy method.

ACKNOWLEDGMENT

This work was carried out by Cheng Zheng and Ling-Chun Hu (Institute for Key Laboratory of Intelligent Computing and Signal Processing, Ministry of Education, Anhui University). We gratefully acknowledge their invaluable cooperation in preparing this application note.

REFERENCES

Ehrig, M. and Y. Sur, 2004. Ontology Mapping: An Integrated Approach. In: The Semantic Web: Research and Applications, Bussler, C.J., J. Davies, D. Fensel and R. Studer (Eds.). Springer, Berlin Heidelberg, Germany, pp: 76-91.

Euzenat, J., A. Ferrara, C. Meilicke, J. Pane and F. Scharffe *et al.*, 2006. First results of the ontology alignment evaluation initiative. Proceedings of the Workshop on Ontology Matching in the 5th International Semantic Web Conference, November 5, 2006, Berlin, pp: 73-78.

Gang, L.V. C. Zheng and C.L. Hu, 2011. Research on method of multi-ontology mapping based onconcept classification. Applic. Res. Comput., 28: 3335-3337.

Kuster, U., B. Konig-Ries, M. Stern and M. Klein, 2007. DIANE: An integrated approach to automated service discovery, matchmaking and composition. Proceedings of the 16th International Conference on World Wide Web, May 8-12, 2007, ACM Press, New York, pp: 1033-1042.

Li, J.J., J. Qi and J. Hu, 2010. Similarity measurement method based on membership function and it sapplication. Appl. Res. Comput., 27: 891-894.

Lv, G. and C. Zheng, 2011. A novel framework for concept detection on large scale video database and feature pool. Artificial Intell. Rev., 40: 391-403.

Tang, J., B.Y. Liang and J.Z. Li, 2006. Automatic ontology mapping in semantic web. Chin. J. Comput., 11: 1956-1976.

Xia, H.K., X.F. Zheng and X. Hu, 2010. Novel approach of ontology mapping extraction SME. Comput. Sci., 37: 233-236.