Research on an improved ontology mapping method

Kai Liu[a], Shanhong Zheng[b*], Wanlong Li[c], Kai Li[c]

Changchun University of Technology,Changchun 130012,China

[a]791568510@qq.com,[b]zhengshanhong@ccut.edu.cn(Corresponding author)

Key words: ontology;ontology mapping;conceptual similarity;multi-strategy

Abstract. In order to improve the accuracy of ontology mapping,an improved ontology mapping method is proposed.Firstly,simplify the process of calculating the conceptual similarity by adopting the candidate set.Then,determine the weight by referencing the rough set condition information entropy aimed at improving the quality of mapping.Last,the test set also called benchmark provided by OAEI is introduced to test the performance of our mapping algorithm.The experimental results display that the mapping efficiency and generality is maintained,the recall and precision of ontology mapping is significantly promoted as well.

1 Introduction

As it is known to us all,the generation of semantic web is due to the lack of semantic in the web.The concept of semantic web is proposed by Berners-lee[1],which adopts a new method that can be understood by the computers in order to promote the intelligent interaction between computers.Ontology as the important foundation of Semantic Web,it can be adopted to describe the semantic information in the process of dealing with the computer data,and it is also widely used in the field of the knowledge engineering,artificial intelligence,semantic web,information retrieval and the other related fields.Due to the the problem of ontology heterogeneity,the application of ontology in semantic web has been restricted up to a point.

In view of the above-mentioned facts,the research on an improved ontology mapping method is proposed in this paper,which is benefit for reducing the computation by introducing the strategy of candidate mapping.Meanwhile,all kinds of the ontology information such as instance similarity and structure similarity are overall considered.In the end,a comprehensive similarity calculation result is obtained by using the merging strategy,which makes the mapping results more accurate.

2 Ontology and ontology mapping

2.1 Ontology. With the ontology is introduced to the field of computer,Many scholars begin to elucidate the definition of ontology,the most widely recognized definition is from Studer et al.They present that ontology is a formal specification of a shared conceptual model[2].In this paper,we adopt the ontology definition form proposed by Gruber[3],It can be denoted by a five tuple,for instance,$O=<C,I,R,F,A>$.Where C represents the concept,I represents an instance,R represents the relationship between concepts,R represents the function and A represents the axiom.

2.2 Ontology mapping. Ontology mapping means a process of the semantic analysis of ontology which is constructed in the same domain.Shvaiko et al propose that the ontology mapping can be defined as f=<id,e,e′,n,R>[4],Where R represents the relationship between the entity e and e′and n represents the confidence level of mapping.

3 Improved Ontology Mapping Method

3.1 Method design idea. In this paper,we first to filter the candidate concept set,which need to reference the correlation of the two concepts in order to narrow the scope of concept and reduce the computation of the concept similarity.Then,calculate the concept similarity based on the name

strategy,concept attribute and concept relation,meanwhile,all kinds of information are synthetically considered.Finally,carry on the similarity merging by adopting the certain weight set by the rough set condition information entropy.As a result,the mapping efficiency is promoted.

3.2 Selection of candidate concept set. The conceptual similarity in ontology indicates the degree of the common characteristics between two concepts,but the concept correlation[5]means that the two concepts are related to a particular relationship.In this paper,we put forward to simplify the input concept set by computing the correlation degree between the concepts,meanwhile the Hirst-St-Onge semantic correlation algorithm[6] is adopted to calculate the concept correlation.The calculation formula of the correlation degree is as follows.

$$RelHs(S_1, S_2) = c - len(S_1, S_2) - k * turns(S_1, S_2) \quad (1)$$

Among them,where c , k is a constant, $turns(S_1, S_2)$ indicates the path turning times from the word S1 to the word S2, $len(S_1, S_2)$ represents the total length of the path from the word S1 to the word S2,and the range of $RelHs(S_1, S_2)$ is[0,1].When in the calculation of the correlation degree,first to set an initial threshold φ .If the calculated correlation value is less than φ ,it is considered that these two concepts are not related,and the concept will be removed from the concept set to be mapped.

3.3 Concept similarity computation

Calculation method of concept similarity based on name. The theoretical basis of the concept name similarity strategy is that if the identifier of the two concept names are the same or similar,then their meanings are generally the same or similar.So,when carry on the concept similarity calculation,we need to consider from two aspects such as the character symbol and the semantics.The calculation method is shown as follows.

$$Sim_{name}(A, B) = \alpha * Sim_{sematic}(A, B) + \beta * Sim_{char}(A, B) \quad (2)$$

Among them,where A and B respectively represent the concept of ontology O1 and O2, $Sim_{sematic}(A, B)$ represents the similarity based semantic, $Sim_{char}(A, B)$ represents the similarity based on text symbols, α and β respectively represent the weight of the above two similarity,and the relation between α and β is that $\alpha + \beta = 1$.

1)Similarity based on semantic

The concepts of A and B are assumed to represent respectively the concepts extracted from the ontology O1 and O2,we need to carry on the pretreatment among the name of the concepts(such as word segmentation and stem extract etc.).Then,we will get the two word sets such as $\{A_i | i = 1,2,\cdots,m\}$ and $\{B_j \mid j = 1,2,\cdots,n\}$,which can be shorten and replaced by $\{A_i\}$ and $\{B_j\}$ .We can calculate the concept similarity according to the Wu-Palmer algorithm[7] as follows.

$$Sim(A_i, B_j) = \frac{2 * depth(sub(A_i, B_j))}{depth(A_i) + depth(B_j)} \quad (3)$$

Among them,where $sub(A_i, B_j)$ represents the common ancestor owned jointly by the concept A and B,and $depth(X)$ represents the depth of the concept X in the WordNet semantic tree.

For the concept B in each word B_j ,select the maximum value from $\{A_i\}$ as the similarity of the concept A and the word B_j ,which can be also expressed as $Sim(A, B_j)$ .Consequently,the concept name similarity can be defined as follows.

$$Sim_{semantic}(A, B) = \frac{\sum_{j-0}^{n} Sim(A_i, B_j)}{n} \quad (4)$$

2)Similarity based on the text symbols

By calculating,the minimum number of edits(for example insert,delete and replace)in the process of turning the name string S_1 derived form the concept A into the name string S_2 derived form the concept B is the desired edit distance also called $Edit(S_1, S_2)$ [8].Therefore,the calculation method based on the text symbols similarity is as follows.

$$Sim_{char}(A,B) = \max(0, 1 - \frac{2 * Edit(S_1, S_2)}{|S_1| + |S_2|}) \quad (5)$$

Calculation method of concept similarity based on attribute. In order to explain all aspects of attribute information,we can measure the similarity of concept attributes from three aspects,for example,the attribute name,data type and instance.We can denote the attribute of the concept A is a_i and the attribute of the concept B is b_j ,then the similarity between a_i and b_j is as follows.

$$Sim(a_i, b_j) = \alpha * Sim_{name}(a_i, b_j) + \beta * Sim_{data}(a_i, b_j) + \gamma * Sim_{instance}(a_i, b_j) \quad (6)$$

Where α represents the weight based on name attribute and β represents the weight based on the data type, γ represents the weight based on the instance,where there is accompanied by $\alpha + \beta + \gamma = 1$.

The attribute similarity corresponded with the concept A and B is as follows.

$$Sim_{attribute}(A,B) = \frac{\sum_{k=1}^{n} \omega_k * Sim(a_i, b_j)}{\sum_{k=1}^{n} \omega_k} \quad (7)$$

Calculation method of concept similarity based on instance. The concept instance is an important element in ontology,we can use the Jaccard coefficient to determine the similarity based on instance,as follows.

$$Sim_{instance}(A,B) = \frac{p(A \cap B)}{p(A \cup B)} = \frac{p(A, B)}{p(A, B) + p(A, \bar{B}) + p(\bar{A}, B)} \quad (8)$$

Calculation method of concept similarity based on structure. By overall consideration about the similarity of the parent concept,brother concept,and the sub concept,we can get the concept similarity based on structure,as follows.

$$Sim_{structure}(A,B) = \omega_1 * Sim_{parent}(A,B) + \omega_2 * Sim_{brother}(A,B) + \omega_3 * Sim_{son}(A,B) \quad (9)$$

The parameters can be set such as $\omega_1 \geq \omega_2 \geq \omega_3$ and $\omega_1 + \omega_2 + \omega_3 = 1$.we found that when ω_1 is set to 0.5, ω_2 is set to 0.3,and ω_3 is set to 0.2,the computing effect will be the best.

Fusion of similarity. Based on the formula(2),(7)(8),(9),the comprehensive similarity of the concept A and concept B is equivalent to the weighted average come from the above similarity value.The computational method is shown in the formula(10).

$$Sim(A,B) = \omega_1 Sim_{name}(A,B) + \omega_2 Sim_{attribute}(A,B) + \omega_3 Sim_{instance}(A,B) + \omega_4 Sim_{structure}(A,B) \quad (10)$$

3.4 Weight determination based on conditional information entropy of rough sets. According to the rough set theory,where the global U represents the set of concept pair used for participating the similarity computation.Attribute set is $A = \{C, D\}$, the conditional attribute set is $C = \{Sim_{name}, Sim_{attribute}, Sim_{instance}, Sim_{structure}\}$,and decision attribute set is $D = \{d\}$ .Meanwhile the property value set is V .Where $f : U \times A$ represents the information function as well as the attribute value come from the global U .In addition,there is a decision table $S = \{U, C, D, V, f\}$.

Decision attribute $D(U/D = \{d_1, d_2\})$ is defined as the conditional information entropy of $C(U/C = \{c_1, c_2, c_3, c_4\})$ for each similarity attribute as follows.

$$I(D/C) = \sum_{i=1}^{4} \frac{|C_i|^2}{|U|^2} \sum_{j=1}^{2} \frac{|C_i \cap d_i|}{|C_i|} \bullet \left| 1 - \frac{|C_i \cap d_i|}{|C_i|} \right| \quad (11)$$

The importance of similarity attribute C is as follows.

$$Sig(c_i) = I(D/C - \{c_i\} - I(D/C) + I(D/C_i)) \quad (12)$$

The weight of each similarity attribute is as follows.

$$\omega(c_i) = \frac{Sig(c_i)}{\sum_{c \in C} Sig(c)} \quad (13)$$

The weights of each attribute are calculated by the method mentioned above.At the same time,the method can be applied in the mapping process.

4 Experimental results and analysis

4.1 Experimental data. The method name is tentatively called ICSCM.We select several sets of ontology in the standard test case set benchmark come from the OAEI as the experimental data.#1XX～#3XX reference ontology is used to test.The statistics such as the recall,precision and F-measure will be calculated before and after the process of mapping.Then the typical mapping system(such as Falcon[9]、RIMOM[10]、OntoDNA[11]、ASMOV[12])are selected to carry on the comparative analysis.The result indicates that when the correlation threshold is set to 0.1(When the correlation between the two concepts is less than 0.1,it is considered that they are not related.),then both the mapping quality and the mapping efficiency can be improved.

4.2 Experimental environment. In this paper,the algorithm is described by Jena language and Eclipse 8.6 is adopted as the development platform.Experimental tests are carried out in the computer with the Intel Core i7 quad core CPU,8G memory,and Windows 7 operating system.Furthermore,the TempLoadRunner tool is used to test the amount of calculation.

4.3 Evaluation measures. The precision(P),the recall(R)and the F-measure[13]are widely used in the area of information retrieval which are adopted as the evaluation standard.

$$P = \frac{A}{A+C} * 100\%$$

$$R = \frac{A}{A+B} * 100\%$$

$$F - measure = \frac{(a^2+1) * P * R}{a * P + R}$$

Among them,where A represents the number of mapping results correctly identified by the algorithm,B represents the number of mappings true but not identified by the algorithm,C represents the number of mapping results incorrectly identified by the algorithm.F-measure is a combination of two indicators which includes the recall and the precision.We can set a=1,which indicates that the recall and precision is of equal importance.

4.4 Results and analysis. Test comparison results are shown in the Figure 1.

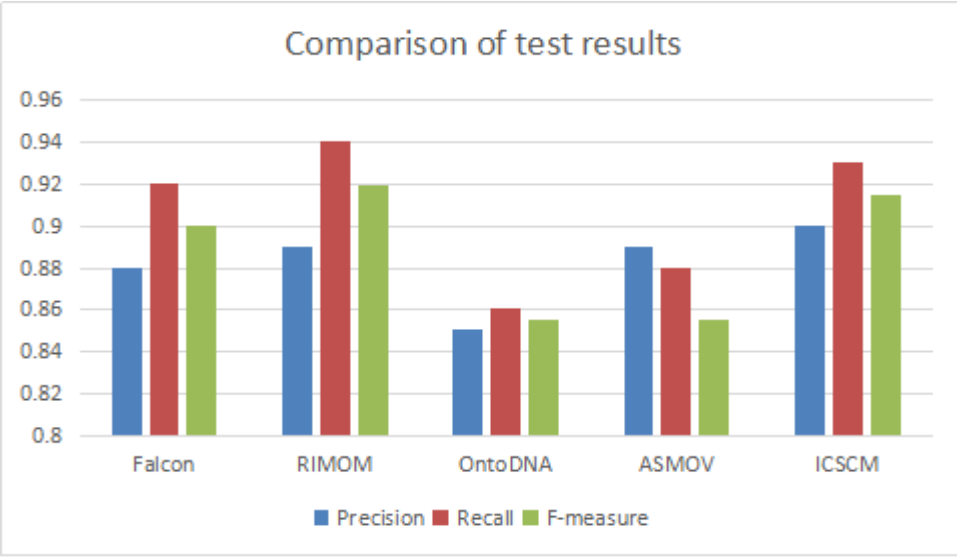

Fig.1 Comparison of test results

In this paper,the filter candidate concept set is introduced to reduce the calculation of concept similarity,and the strategy based on the name,concept attribute and concept relation is adopted.Furthermore,the rough set condition information entropy is introduced to determine the weight.The experimental results show that the ontology mapping is significantly improved.

5 Conclusions

In the paper, research on an improved ontology mapping method is proposed.In this method,firstly,the filter candidate concept set is selected to reduce the scope of the concept and the amount of computation.Then,make use of the strategy based on the name,concept attribute and concept relation.The rough set conditional information entropy is adopted to automatically generate the weight.As a result,the quality of the mapping is improved.The calculation of concept similarity is just one step.So,the next research need to focus on improving the accuracy of mapping and reducing the complexity by studying other steps of the ontology mapping.

Acknowledgment

This work is supported by Jilin province Natural Science Foundation support project (No.20130101060JC) and the Science & Technology Research Foundation of Jilin Department of Education during the Twelfth Five-year Plan support project (No.2014131;No.2014125), all support is gratefully acknowledged.

References

[1] BERNERS-LEE T,FISCHETTI M,DERTOUZOS M L.Weaving the Web:The Original Design and Ultimatc Destiny of the World Wide Web by Its Inventor[M].San Francisco:Harper.1999.

[2] STUDER R,BENJAMINS V R,FENSEL D.Knowledge Engineering,Principles and Methods [J].D-ata and Knowledge Engineering,1998,25(1-2):161-197.

[3] GRUBER T R.A Translation Approach to Portable Ontology Specifications[J].Knowledge Ac q-uisition,1993,5(2):199-200.

[4] SHVAIKIP,EUZENATJE.A Survey of Schema-Based Matching Approaches[J].Journal on Dat a SemanticsⅣ,2005,4(1):146-171.

[5] Liu Hong-zhe,Xu De.Ontology Based Semantic Similarity and Relatedness Measures Review[J].Computer Science,2012,39(2):8-13.

[6] Zhang Cheng-li,Chen Jian-bo,Qi Kai-yue.Improvement on the Arithmetic of Semantic Similarity Based on the Semantic Web[J].Computer Engineering and Applications,2006,42(17): 165-179.

[7] Wu Zhi-biao,PALMER M.Verb Semantics and Lexical Selection[C]//Procofthe 32nd Annual Meeting of the Association for Computational Linguistics.NewYork:ACM,1994:133-138.

[8] Li Rong,Yang Dong,Liu Lei.Research of Ontology-Based Conceptual Similarity Computation[J].Journal of Computer Research and Development,2011,48(Z2):312-317.

[9] Hu Wei,Qu Yu-zhong.Falcon-AO:A Practical Ontology Matching System[J].Web Semant-SciServ Agents WWW,2008,6(3):237-239.

[10] LI Juanzi,TANG Jie,LI Yi,et a1.RiMOM:A Dynamic Multistrategy Ontology Alignment Frame-work[J].IEEE Trans Knowl Data Eng,2009,21(8):1218-1232.

[11] KIUC C,LEECS.OntoDNA:Ontology Alignment Results for OAEI 2007[C]Proceedings of the ISWC'2007 Workshop on Ontology Matching.Heidelberg:Springer,2007:185-195.

[12] JEAN-MARYL Y R,KABUKA M R.ASMOV Results for OAEI 2007[C]//Proceedings of the ISWC' 2007 Workshop on Ontology Matching.Heidelberg:Springer,2007:216—224.

[13] EUZENAT J,RHONE-ALPES I.Semantic Precision and Recall for Ontology Alignment Evaluation[C]Proceedings of the 20th International Joint Conference on Artificial Intelligence.San Francisco:[s.n.],2007:248—253.