Journal of Southeast University (English Edition) Vol. 23, No. 3, pp. 381 – 384 Sept. 2007 ISSN 1003—7985

More efficient ontology matching algorithm for integrating heterogeneous web resources

Liu Chen[1,2] Han Yanbo[1] Chen Wanghu[1,2] Ding Weilong[3]

([1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China)

([2]Graduate University, Chinese Academy of Sciences, Beijing 100039, China)

([3]College of Information Science and Technology, Shandong University of Science and Technology, Qingdao 266510, China)

Abstract: To improve the performance of the ontology matching process, a more efficient ontology matching algorithm, which can effectively eliminate unnecessary operations of matching entities, is proposed. By the theoretical analysis and proof, a set of matching rules are summarized for depicting inherent relations among matching results of entities. Based on these rules, the proposed algorithm can reuse the matching results of two entities to directly determine the matching results of their adjacent entities. Thereby, redundant operations of matching adjacent entities can be avoided, which can improve the performance of the whole matching process. The experimental results show that, compared with related algorithms, the proposed algorithm has high matching accuracy and can remarkably reduce the consuming time of the whole matching process. So, the proposed algorithm is more competent for the large-scale ontology matching which often occurs in the practical heterogeneous web resources integration project.

Key words: ontology matching; matching performance; matching rule

Today, when integrating large-scale web resources, performance is becoming a key factor for the practical use of a matching algorithm to handle the semantic heterogeneities. As an example, in undergoing a practical project, a science and technology department plans to establish a centric portal for integrating heterogeneous web resources distributed over the local offices. Fig. 1 shows fragments of news ontologies provided by Science and Technology Department and Shandong Province Office. To integrate their resources, mappings must be first established among their heterogeneous ontologies. However, as increasing local offices are involved as well as richer and richer resources are supported, requirements of the performance of matching algorithm are becoming more and more exigent.

To provide higher matching accuracy, nowadays most of matching systems[1–6] depend on a domain specific thesaurus such as WordNet[7] to determine the relation holding in an entity pair (formed by two entities which come from the first and the second matching ontology respectively). However, matching entities by consulting a thesaurus are time consuming operations.

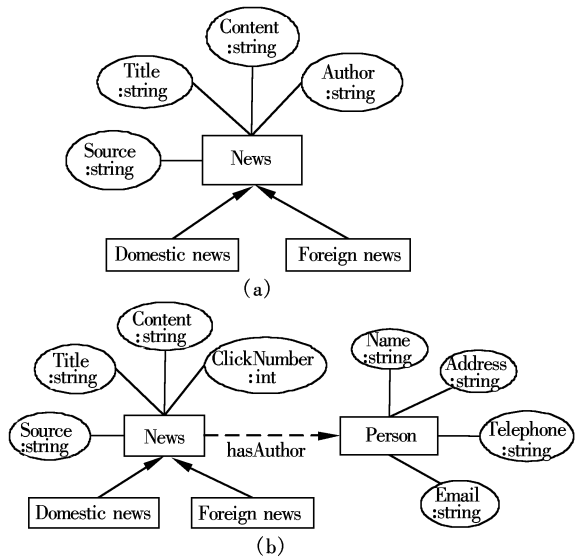

Fig. 1 Fragments of news ontology. (a) Ontology of Science and Technology Department; (b) Ontology of Shandong Province Office

We believe that the less such matching operations are executed, the higher will be the performance of an algorithm. Through experiments, we discover that the relation of two entities can be reused to determine the relations among their adjacent entities. As shown in Fig. 1, after finding that the entity pair 〈News, Person〉 have not any relations, without consulting the thesaurus again, we can directly draw the conclusion that there are also not any relations in the entity pairs 〈Domestic_News, Person〉 and 〈Foreign_News, Person〉. This

Received 2007-05-18.

Foundation items: R & D Infrastructure and Facility Development (No. 2005DKA64201), the National High Technology Research and Development Program of China (863 Program) (No. 2006AA12Z202).

Biographies: Liu Chen (1980—), male, graduate; Han Yanbo (corresponding author), male, doctor, professor, yhan@ict.ac.cn.

is because concepts "Domestic_News" and "Foreign_News" are firstly a concept "News". Therefore, two unnecessary operations of having to consult the thesaurus are passed over.

In this paper, a set of rules to reuse matching results has been summarized and strictly proved, which can be used to effectively cut down the unnecessary operations of matching entities. A more efficient algorithm (rule-match) based on these rules is developed.

1 Matching Algorithm

The result of a matching algorithm is called an alignment[8], which composes of a set of strongest relations[9] holding among entity pairs. The possible relations holding in an entity pair are equivalence ($=$), more general ($\subseteq$), disjointness ($\perp$) and overlapping ($\cap$). To compute an alignment, the technique for matching any two entities in a pair will be introduced in section 1.2. Then, in section 1.3, the algorithm is proposed to compute the set of possible relations for any two ontologies.

1.1 Ontology model

The currently implemented ontology language is based on RDF(S)[10-11], but contains some extensions such as cardinality. In the model, concepts, properties and instances are all called entities.

Definition 1 An ontology model is a tuple OntModel: $=(C, P, I, R)$, where

- C is the set of concepts. Concepts are collections of objects that have similar properties. Concepts constitute a subconcept—superconcept hierarchy with multiple inheritances.
- P is the set of properties. Properties can be divided into datatype properties and object properties. They are usually first-class objects. Besides, properties can be transitive, symmetric, or have inverses.
- I is the set of instances. Instances are individual members of concepts.
- R is the set of restrictions. Each property has a set of restrictions on its values, such as cardinality and range.

1.2 Entity matching technique

The determination of the relation holding in an entity pair is based on the exploitation of the entities' names. Referred to the global thesaurus WordNet[7], five terminological relations between two names t and t' are considered: ① Equal: t and t' are exactly the same with each other. ② Synonym: $t \neq t'$ but they are synonyms. ③ Hypernymy/Hyponymy: t has a more/less general meaning than t'. ④ Meronymy/Holonymy: t is part/whole of t'. ⑤ Unknown: The relation of terms cannot be recognized or supported. However, to form an alignment, these terminological relations need to be transformed into the above-mentioned entity relations. The transformation rules are demonstrated in Tab. 1.

Tab. 1 Terminological relations and their corresponding entity relations

Entity relation	Terminological relation	Examples
$=$	Equal $\cup$ Synonym	News, News
$\subseteq$	Hypernymy/Hyponymy	People, Person
$\cap$	Meronymy/Holonymy	Date, Time
$\perp$	Unknown	Title, Source

1.3 Rule-match algorithm

To improve the performance, the rule-match algorithm depends on the following matching rules to eliminate unnecessary matching operations. Here, we give the proofs of rule 1 and rule 3. The proofs of other rules are very similar.

Rule 1 (concept hierarchy rule) If concepts C_1 and C_2 have the relation "$\perp$", then any of their subconcepts also have the relation "$\perp$".

Proof C_1 and C_2 have the relation "$\perp$" means

$$C_1 \cap C_2 = \varnothing \tag{1}$$

Suppose that S_1 and S_2 are the subconcepts of C_1 and C_2, respectively. Then,

$$S_1 \subseteq C_1, \quad S_2 \subseteq C_2 \tag{2}$$

According to (1) and (2), $S_1 \cap S_2 = \varnothing$.

Therefore, S_1 and S_2 have the relation "$\perp$".

Rule 2 (property hierarchy rule) If property P_1 and P_2 have the relation "$\perp$", then any of their subproperties also have the relation "$\perp$".

Rule 3 (property domain rule) Suppose that concept C_1 and C_2 have the relation "$\perp$", if C_1 and C_2 are the domain of properties P_1 and P_2, respectively, then P_1 and P_2 have the relation "$\perp$".

Proof Suppose that R_1 and R_2 are the range of properties P_1 and P_2, respectively. $D_{11} \ldots D_{1m}$ and $D_{21} \ldots D_{2n}$ are the other domains of P_1 and P_2, respectively. Then,

$$\begin{aligned} P_1 &\subseteq (C_1 \cap D_{11} \cap \ldots \cap D_{1m}) R_1 \\ P_2 &\subseteq (C_2 \cap D_{21} \cap \ldots \cap D_{2n}) R_2 \end{aligned} \tag{3}$$

According to (1),

$$(C_1 \cap D_{11} \cap \ldots \cap D_{1m}) \cap (C_2 \cap D_{21} \cap \ldots \cap D_{2n}) = \varnothing \tag{4}$$

According to (1) and (4): $P_1 \cap P_2 = \varnothing$.

Therefore, P_1 and P_2 have the relation "$\perp$".

Rule 4 (object property range rule) Suppose that concepts C_1 and C_2 have the relation "$\perp$", if C_1 and C_2 are the range of properties P_1 and P_2, respec-

tively, then P_1 and P_2 have the relation "$\perp$".

The rule-match algorithm is demonstrated as follows:

Algorithm 1 Rule-match(O_1, O_2)

Input: Ontology O_1 and O_2

Output: Alignment $A(O_1, O_2)$

1 Initialize an empty alignment

2 Get the entity set (E_1 and E_2) of each ontology

3 Sort each set by the level of an entity

4 Starting matching

for each pair $e_1 \in E_1$ and $e_2 \in E_2$

4.1 If the relation of e_1 and e_2 is determined, then continue;

4.2 Get the relation of e_1 and e_2 by looking up the WordNet;

4.3 Add the result of 4.2 to the alignment;

4.4 Applying the rules to predict the relation among adjacent entities.

endfor

5 Return to the alignment

Note that step 3 aims to ensure that the upper level concepts and their properties will take higher priorities to be firstly matched. Step 4.4 will eliminate the unnecessary matching operations based on the relation between e_1 and e_2.

2 Evaluation

As shown in Fig. 1, the above-mentioned practical project for integrating science and technology web resources provides the first group of experimental data for us. Besides, to ensure the fairness and accuracies, two other ontologies about the art history of Google and Yahoo concept hierarchies are developed as the second group of experimental data. They are shown in Fig. 2.

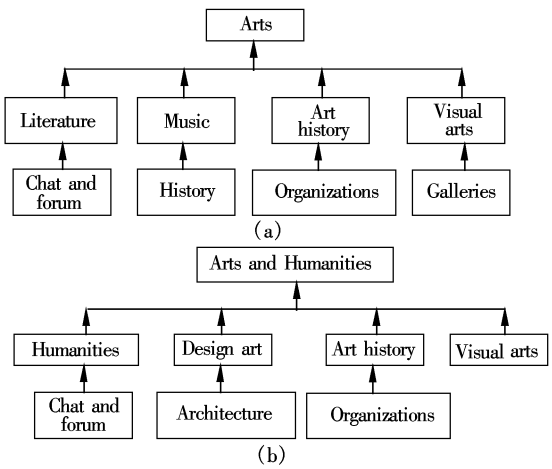

Fig. 2 Fragment of ontologies about art history of Google and Yahoo concept hierarchy. (a) Google; (b) Yahoo

The proposed algorithm is mainly compared with two other well-known algorithms (WN-Matcher in the CMS[1] and CtxMatch[2]). However, since the CtxMatch algorithm is designed to be only applied to the concept hierarchies, it does not attend the test of the first group of data. The following tests have been executed on an Intel Dothan machine at 1.60 GHz with 1 GB RAM and IDE disks.

The whole experimental process will be further divided into function tests and performance tests. The function test is aimed at finding whether the proposed algorithm can work as well as other algorithms. Referring to standard evaluation measures in information retrieval, recall and precision are adopted to evaluate the results. Here, recall is defined as the proportion of matched entities which are detected, while precision is defined as the proportion of detected entities that are actually similar. The results of this test are shown in Tab. 2. It shows that the proposed algorithm also has high accuracy and can work as well as other algorithms.

Tab. 2 Results of function test

Algorithm	Recall		Precision	
	Test 1	Test 2	Test 1	Test 2
CMS	0.83	0.75	0.83	0.50
CtxMatch		0.75		0.75
Rule-match	0.83	0.875	0.71	0.78

Note: CtxMatch does not attend the first test due to its limitations, so the corresponding cells in two tables are empty.

The second test is to compare the performance of each algorithm, which is aimed at finding whether the proposed algorithm will greatly improve the performance of the matching process. The result of this test is shown in Tab. 3. It shows that the performance of the proposed algorithm is much higher than that of other algorithms. Thus, in the practical applications, it can be more adequate to match large-scale ontologies than other algorithms.

Tab. 3 Results of performance test ms

Test data	CMS	CtxMatch	Rule-match
Data 1	24 859		10 250
Data 2	18 312	11 500	7 859

3 Conclusion

This paper proposes a more efficient ontology matching algorithm for integrating heterogeneous web resources. Through theoretical analysis and proof, a set of matching rules is summarized and strictly proved. Depending on these rules, the proposed algorithm can greatly improve the performance of the matching process by decreasing unnecessary matching operations. It means that the proposed algorithm is quite competent for the large-scale ontology matching which often occurs in the practical web resources integration project. The practical application and experiment have

validated this conclusion.

References

[1] Kalfoglou Y, Hu B, Reynolds D, et al. Capturing, representing and operationalising semantic integration (CROSI) project-final report [R]. Department of Electronics and Computer Science of University of Southampton, 2005.

[2] Bouquet P, Magnini B, Serafini L, et al. A SAT-based algorithm for context matching [C]//*Proc of the 4th International and Interdisciplinary Conference on Modeling and Using Context*. Springer-Verlag, 2003: 66 - 79.

[3] Castano S, Ferrara A, Montanelli S. Matching ontologies in open networked systems: techniques and applications [J]. *Journal of Data Semantics*, 2006, **5**: 25 - 63.

[4] Giunchiglia F, Shvaiko P, Yatskevich M. S-match: an algorithm and an implementation of semantic matching [C]// *Proc of the European Semantic Web Symposium*. Springer-Verlag, 2004: 61 - 75.

[5] Euzenat J, Valtchev P. Similarity-based ontology alignment in owl-lite [C]//*Proc of the European Conference on Artificial Intelligence*. Amsterdam: IOS Press, 2004: 333 - 337.

[6] Ehrig M, Sure Y. FOAM—framework for ontology alignment and mapping; results of the ontology alignment initiative [C]//*Proc of the Workshop on Integrating Ontologies*. Banff, Alberta, Canada, 2005: 72 - 76.

[7] Miller A G. WordNet: a lexical database for English [J]. *Communications of the ACM*, 1995, **38**(11): 39 - 41.

[8] Euzenat J. An Api for ontology alignment [C]//*Proc of the International Semantic Web Conference*. Springer-Verlag, 2004: 698 - 712.

[9] Shvaiko P, Euzenat J. A survey of schema-based matching approaches [J]. *Journal of Data Semantics*, 2005, **4**: 146 - 171.

[10] Beckett D, McBride B. Rdf/xml syntax specification (revised) world wide web consortium [EB/OL]. (2004-02-10) [2007-05-10]. http://www.w3.org/tr/rdf-syntax-grammar/.

[11] Brickley D, Guha R V. Rdf vocabulary description language 1.0: Rdf schema. World wide web consortium [EB/OL]. (2004-02-10) [2007-05-10]. http://www.w3.org/tr/rdf-schema/.

一种使能异构 web 资源集成的高效本体匹配算法

刘 晨[1,2] 韩燕波[1] 陈旺虎[1,2] 丁维龙[3]

([1] 中国科学院计算技术研究所,北京 100080)

([2] 中国科学院研究生院,北京 100039)

([3] 山东科技大学信息科学与技术学院,青岛 266510)

摘要:为了提高本体匹配过程的性能,提出了一种能够有效减少冗余实体匹配操作的高效本体匹配算法 rule-match. 通过理论分析和证明,总结出了一组刻画实体匹配结果内在联系的匹配规则. 基于这组规则,匹配算法可以在匹配过程中重用 2 个本体实体的匹配结果来直接确定其邻接实体的匹配结果,避免了对邻接实体所进行的不必要的匹配操作,从而提升整个匹配过程的性能. 实验结果表明,相对于其他匹配算法,该算法不仅具有较高的准确率,而且能有效降低整个匹配过程所消耗的时间. 该算法适用于解决实际异构 web 信息资源集成项目中所常见的大规模本体匹配问题.

关键词:本体匹配;匹配性能;匹配规则

中图分类号:TP311