Topic Maps Matching Computation Based on Composite Matchers

Jungmin Kim[1] and Hyunsook Chung[2,*]

[1] School of Computer Engineering, Seoul National University, Korea
jmkim@idb.snu.ac.kr
[2] Department of Computer Engineering, Chosun University, Korea
hsch@chosun.ac.kr

Abstract. In this paper, we propose a multi-strategic matching approach to find correspondences between ontologies based on the syntactic or semantic characteristics and constraints of the Topic Maps. Our multi-strategic matching approach consists of a linguistic module and a Topic Map constraints-based module. A linguistic module computes similarities between concepts using morphological analysis, and language-dependent heuristics. A Topic Map constraints module takes advantage of several Topic Maps-dependent techniques such as a topic property-based matching, a hierarchy-based matching, and an association-based matching. It is not necessary to generate a cross-pair of all topics from the ontologies because unmatched pairs of topics can be removed by characteristics and constraints of the Topic Maps. Our experiments show that the automatically generated matching results conform to the outputs generated manually by domain experts, which is very promising for further work.

Keywords: Ontology matching, Topic Maps, multi-strategic matching process.

1 Introduction

In recent years, many approaches for ontology matching have been proposed. However, all of these earlier approaches for schema or ontology matching focused on providing various techniques for effective matching and merging of schemas or ontologies[1]. They were far from efficiency considerations and thus are not suitable for practical applications based on ontologies of real world domains[7]. Also, earlier approaches convert ontologies or schemas of relational database, object oriented database, and XML, to a graph model with only nodes and edges for supporting different applications and multiple schema types[2,3,11]. This conversion results in low efficiency because the characteristics of ontologies that are useful for similarity computation are overlooked. Another problem with the existing matching methods is that given two ontologies O_1 and O_2, for each entity in ontology O_1, they are compared with all entities in ontology O_2. This full scanning on ontology O_1 and O_2 also ends up with low efficiency.

* Corresponding author.

Y. Shi et al. (Eds.): ICCS 2007, Part II, LNCS 4488, pp. 696–703, 2007.

In this paper, we present an approach that considers features of Topic Maps to reduce the matching complexity and linguistic analysis to improve the matching performance. Our approach does not require ontologies to be converted into a generic graph model and the entities to be fully scanned into two ontologies. Furthermore, our approach is a composite combination of four matching techniques: name matching, internal structure matching, external structure matching, and association matching. This composite matching approach combines the results of four matching techniques that are independently processed to measure the unified similarity of each pair.

To evaluate the quality of our approach, we use the philosophy ontology[5] which is constructed from Korean philosophy learning domain, Wikipedia philosophy ontology which is constructed from philosophy-related contents of Wikipedia, and German literature ontology which is constructed from contents on German literature in the yahoo encyclopedia as experimental data.

We use three measurements such as precision, recall, and overall, which were derived from the Information retrieval field, to evaluate the quality of our approach. We then evaluated the approach by computing three measurements based on a set of manually determined matches and a set of automatically generated matches by matching operations. Based on the experimental results, we could conclude that automatically generated matches by our matching operation can cover most of the manually determined matches.

2 Related Work

With respect to matching and merging ontologies, there have been a few approaches, such as PROMPT[9], Anchor-PROMPT[10], Information flow[13], FCA-Merge[2], QOM[7], and so on.

According to Topic Maps Reference Model, two Topic Maps can be mapped and merged only if two topics have identical subject identity regardless of their name-based similarity. But it is not always the case that all topics, which represent the semantically same concept, have a standard subject identity. Furthermore, Topic Maps whose topic does not have a subject identity can be built.

To overcome this weakness, SIM(Subject Identity Measure)[6] was used to measure the similarity between topics based on their name similarity and occurrence similarity. In the SIM, the processes were only string comparison of the name of topics and resource data of occurrences. The hierarchical structure and association in Topic Maps are not considered.

Table 1 represents characteristics of the methods at a glance. Abbreviated column names mean that Language(L), Patterns(P), Experimental Data(D), Results(R), and Complexity(C). Patterns column indicates matching approaches, which terminological(T), internal structure(IS), external structure(ES), extensional(E), and instance(I). Our approach, which is named TM-MAP, is similar with QOM in terms of the use of features of a data model for an ontology to reduce the complexity of matching operation. The difference is that our approach treats the matching problem of distributed Topic Maps.

Table 1. Comparison of the matching and merging methods

Methods	L	P	D	R	C
PROMPT	Graph	T/ES	HPKB	Merge	$O(n^2)$
Ctx-Match	Graph	T/E	Toy	Matching	$O(n^2)$
IF-MAP	Graph	T/I	Toy	Matching	$O(n^2)$
FCA-Merge	Graph	T/I	Toy	Matching	$O(n^2)$
QOM	RDF	T/IS/ES/E	Real Onto.	Matching	O(nlogn)
TMRM	Topic Maps	T	-	Merge	$O(n^2)$
SIM	Topic Maps	T/IS	Toy	Matching	$O(n^2)$
TM-MAP	Topic Maps	T/IS/ES/E	Real Onto.	Merge	O(nlogn)

3 Problem Definition

3.1 Topic Maps Data Model

Topic Maps is a technology for encoding knowledge and connecting this encoded knowledge to relevant information resources. It is used as a formal syntax for representing and implementing ontologies[4,8]. Topic maps are organized around topics, which represent subjects of discourse; associations, which represent relationships between the subjects; and occurrences, which connect the subjects to pertinent information resources. These entities have different meaning and usage, and so we measure the similarity between same entity types rather than whole entitites.

Definition 1. We define a Topic Map model as following 7 tuples:

$$TM := (T_C, T_O, T_A, T_R, T_I, R_H, R_A)$$

- T_C *denotes a set of topic types*
- T_O *denotes a set of occurrence types*
- T_A *denotes a set of association types*
- T_R *denotes a set of role types*
- T_I *denotes a set of instance topics*
- R_H *denotes a set of subsumption hierarchy relations*
- R_A *denotes a set of associative relations*

3.2 Topic Maps Matching Process

Our ontology matching process is composed of following 6 steps.

1. **Initialization** step takes two serialized Topic Maps documents, so-called XTM (XML Topic Maps)[12], as input and interprets them to build Topic Maps in memory. During interpretation, PSI and TopicWord indexes are generated for each Topic Map.
2. **Topic pairs generation** step creates the reduced number of entity pairs rather than whole entity pairs of two Topic Maps.
3. **Similarity computation** step apply composite combination of matching techniques to measure similarity between topics based on the linguistic analysis. Our composite matching approach combines the results of independently executed four

matching algorithms: name matching operation, property matching operation, hierarchy matching operation, and association matching operation.

4. **Similarity aggregation** step aggregates similarity values of four matching operations to generate a combined similarity value for each topic pair.
5. **Match candidates selection** step automatically chooses match candidates for a topic by selecting the topics of the other Topic Map with the best similarity value exceeding a certain threshold.
6. **Post-processing** step manually corrects the errors of automatically generated match results by domain experts.

4 Similarity Computation

Definition 2. A matching function **map** is defined as following expression:

$$\begin{aligned}\textbf{map}(A, B, D) = \textbf{map}(A.T_C, B.T_C, D) \cup \textbf{map}(A.T_C, B.T_I, D) \cup \textbf{map}(A.T_I, B.T_C, D) \cup \\ \textbf{map}(A.T_O, B.T_O, D) \cup \textbf{map}(A.T_A, B.T_A, D) \cup \textbf{map}(A.T_R, B.T_R, D)\end{aligned}$$

A and B are source Topic Maps and D is domain-specific term dictionary. A matching function *map(A, B, D)* is processed by matching functions of different entity types. A matching function **map** is composed of following matching operations.

4.1 Name Matching Operation

Name matching operation compares strings of base names and variant names of topics. In the field terminology, a single term can refer to more than one concept and multiple terms can be related to a single concept. Name matching operation find multiple terms refer to a same concept by application of two main categories of methods for comparing terms: String-based methods and linguistic knowledge-based methods. Both x and y are tokens and c is the largest common substring of them. The similarity value between two strings based on the token and substring-based method is computed by following expression. In this expression x_i is the i-th token of string a and y_j is the j-th token of string b. In our morphological analysis these phrases or sentences are divided into a several stems and inflectional endings, which attached to stems and represent various inflections or derivations in Korean. Thus, in order to improve the quality of string matching results between words, we use word order and ending information, which classify corresponding ending groups according to their meaning and usage.

$$SIM_{token}(x, y) = 2|c| / |x| + |y|$$

$$TS\text{-}SIM_{string}(a, b) = \sum SIM_{token}(x_i, y_j) / | a \cup b |$$

4.2 Internal Structure Matching Operation

If two topics have m occurrences and n occurrences each other, internal structure-based strategy computes similarity values of m by n pairs of occurrences to measure the similarity between topics. An occurrence is defined by an occurrence type and an

occurrence value which is a textual description or URI address. For example, a topic, *Immanuel Kant*, has a occurrence which type is '*figure*' and value is '*http://www.encyphilosophy.net/kant/figure.jpg*'. Thus, the similarity values of occurrence types and occurrence values need to be combined to determine the internal structure-based similarity value of the paired topics.

$$SIM_{occ}(t_1, t_2) = \sum(SIM_{occtype}(t_1.occurrence_i, t_2.occurrence_j) \times SIM_{occvalue}(t_1.occurrence_i, t_2.occurrence_j) / |m| \times |n|, \text{ for } 1 \le i \le m \text{ and } 1 \le j \le n$$

4.3 External Structure Matching Operations

External structure matching measures the similarity between two class topics based on the combined similarity between their child topics. The following expression computes the similarity value between two topics based on the similarity of their hierarchical structure. In this expression, t_1 and t_2 are topics that have m and n parent topics and x and y child topics respectively. And $t_1.parent_i$ is i-th parent topic of t_1 and $t_2.parent_j$ is j-th parent topic of t_2. We average SIM_{name} and SIM_{occ} of $t_1.parent_i$ and $t_2.parent_j$ to determine a combined similarity value between parent topics of t_1 and t_2. Likewise, $t_1.child_k$ is k-th child topic of t_1 and $t_2.child_l$ is l-th child topic of t_2. We average SIM_{name}, SIM_{occ}, and SIM_H of $t_1.child_k$ and $t_2.child_l$ to produce the combined similarity value SIM between them. In the expression, w is a weight ranging from 0 to 1. We set a different value to w in order to emphasize the similarity of parent topics or child topics.

$$SIM_H(t_1, t_2) = (1-w)(\sum(SIM_{name+occ}(t_1.parent_i, t_2.parent_j))/|m| \times |n|) + w(\sum(SIM(t_1.child_k, t_2.child_l))/|x| \times |y|)$$

4.4 Association Matching Operation

Association matching operation determines the similarity between association types. An association type is composed of a set of members, which have their roles in the relation. Thus, the similarity between association types is determined by similarities between members of them. Following expression measures the similarity between association types. Given two association types, t_1 and t_2, for a set of pairs of members the similarity value between paired members is computed. M and N is the number of members of two association types each other. m_i is the i-th member of t_1 and m_j is the j-th member of t_2. r_i is role of m_i and r_j is role of m_j.

$$SIM_{assoc}(t_1,t_2)=\sum SIM(m_i,m_j)\cdot SIM(r_i,r_j)/|M| \times |N|, \text{ for } 1 \le i \le M,\ 1 \le j \le N$$

5 Experiment

We set up three kinds of data groups, which are group A, group B, and group C, for our experiment. Oriental philosophy ontology(T_1), modern western philosophy ontology(T_2), and contemporary western philosophy ontology(T_3) are grouped in group A, because these ontologies are philosophy domain's ontologies and created by the same philosophy experts. Group B includes Wikipedia philosophy ontology(T_4) which is constructed from philosophy-related contents of Wikipedia. Group C

includes German literature ontology(T_5) that was constructed from German literature encyclopedia provided by Yahoo Korea portal. Table 2 shows the characteristics of our experimental data.

Table 2. The statistics of experimental ontologies

Ontologies	Group A			Group B	Group C
	T_1	T_2	T_3	T_4	T_5
Max level	11	10	9	9	4
# of Topics	1826	983	1266	417	30
# of Topic types	1379	384	603	182	3
# of Occ. types	86	56	62	13	2
# of Ass. types	47	40	43	7	2
# of Role types	22	15	18	4	2
# of PSIs	653	328	345	0	3

In this work, we use performance measurement of information retrieval such as precision, recall, and overall, to measure performance of our ontology matching operations. To evaluate the quality of our matching operations, we need to know the *manually determined match set(R)* and the *automatically generated match set(P)* which can be obtained by matching the processes. By comparing these match results, we get *true-positive set(I)* which includes correctly identified matches. We can measure match quality of automatic matching processing by evaluating following expression. Figure 1 shows the experimental result that represents high recall and precision.

$$precision = \frac{|I|}{|P|} \qquad recall = \frac{|I|}{|R|} \qquad overall = recall * (2 - \frac{1}{precision})$$

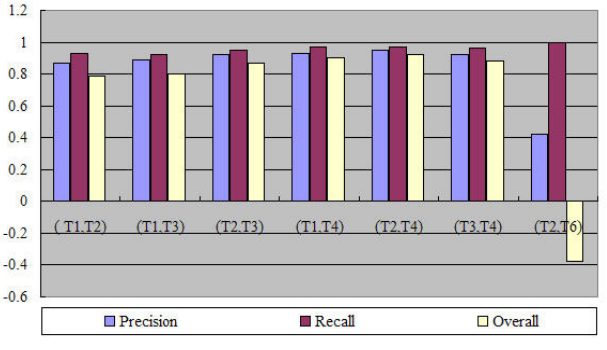

Fig. 1. Experiment results of pairs of Topic Maps

Pairs of ontologies in group A are matched based on the ontology schema layer because these ontologies are constructed from the same knowledge domain and a

group of experts. These ontologies share a common schema, known as the philosophy reference ontology, for standardizing and validating them. The pair (T_2,T_3) of group A has maximal matches because both ontologies are components of the philosophy ontology and have some relationships in terms of philosophers, texts, terms, doctrines, and so on.

In (T_1,T_4), (T_2,T_4), and (T_3,T_4) of group A and B, most of all matched topics result from topic name-based matching operation because paired Topic Maps have topics describing same philosophers, i.e. *Kant*, *Hume*, and *Marx*, same texts of philosophy, i.e. *Philosophy of Right*, *Critique of Pure Reason*, and *Discourse on the Method*, and same terms of philosophy, i.e. *reason*, *free will*, *ideology*, and *moral*. The recall of a pair of modern western philosophy and German literature, (T_2, T_6), is 1 because the number of matches between different domain's ontologies are very low and matching operations easily find matches based on topic names, such as *Nietzsche*, *Philosophy of Right*, and so on. This pair has poor overall, -0.38 in contrast to recall. This means that domain experts must make more efforts to adopt automatically generated matches than to determine matches in manual. In other words, it seems useless to match ontologies between different knowledge domains.

6 Conclusion

In this paper, we propose a multi-strategic matching approach to determine semantic correspondences between Topic Maps. Our multi-strategic matching approach takes advantage of the combination of linguistic module and Topic Maps constraints including name matching, internal structure matching, external structure matching, and association matching. By doing this, the system achieves higher match accuracy than the one of a single match technique.

The experiment results shows that precision of automatically generated match set is more than 87%, but the recall of the set is more than 90%. This means that automatically generated match sets include a large portion of all manually determined matches.

Matched topics are merged into a new topic or connected by a semantic relationship to enable ontology-based systems to provide knowledge-related services on multiple Topic Maps. However, merging or alignment of Topic Maps is not easy work although we found matches between Topic Maps. Ontology merging approaches concerning merging issues, such as conflict resolution, ontology evolution, and versioning will be investigated in the near future.

References

1. Erhard Rahm and Philip A. Bernstein. 2001. *A survey of approaches to automatic schema matchng*, VLDB Journal, 10(4):334-350.
2. Gerd Stumme and Alexander Maedche. 2001. *FCA-Merge: Bottom-up Merging of Ontologies*, In Proceedings of 17th Internaltional Joint Conference on Artificial Intelligence(IJCAI):225-234.
3. Hong Hai Do and Erhard Rahm. 2002. *COMA - a system for flexible combination of schema matching approaches*, In Proceedings of VLDB:610-621.

4. ISO/IEC JTC1/SC34. 2003. *Topic Maps - Reference Model*,URL:http://www.isotopicmaps.org/TMRM/TMRM-latest-clean.html, 2003.
5. JungMin Kim, ByoungIl Choi, and HyoungJoo Kim. 2005. *Building a Philosophy Ontology based on Contents of Philosophical Texts*, Journal of Korea Information Science Society, 11(3):275-283.
6. Lutz Maicher. 2004. *Merging of Distributed Topic Maps based on the Subject Identity Measure(SIM) Approach*, In Proceedings of Berliner XML Tags:301-307.
7. Marc Ehrig and Steffen Staab. 2004. *QOM: Quick ontology matching*, In Proceedings of ISWC:683-697.
8. Michel Biezunski, Michael Bryan and Steven R. Newcomb. 2002. *ISO/IEC 13250 TopicMaps*, URL:http://y12web2.y12.doe.gov/sgml/sc34/document/0322.htm.
9. Natalya F. Noy and Mark A. Musen. 2000. *PROMPT: Algorithm and Tool for Automated Ontology Merging and Alignment*, In Proceedings of the National Conference on Artificial Intelligence(AAAI):450-455.
10. Natalya F. Noy and Mark A. Musen. 2001. *Anchor-PROMPT: Using Non-Local Context for Semantic Matching*, In Proceedings of the Workshop on Ontologies and Information Sharing at the International Joint Conference on Artificial Intelligence(IJCAI):63-70.
11. Paolo Bouquet, Luciano Serafini, and Stefano Zanobini. 2003. *Semantic coordination: A new approach and an application*, In Proceedings of ISWC:130-145.
12. Steve Pepper and Graham Moore. 2001. *XML Topic Maps(XTM) 1.0*, TopicMaps.Org, URL:http://www.topicmaps.org/xtm/1.0.
13. Yannis Kalfoglou and W. Marco Schorlemmer. 2003. *IF-Map: An Ontology-Mapping Method Based on Information-Flow Theory*, Journal on Data Semantics I: 98-127.