Ontology Matching Techniques:
a 3-Tier Classification Framework

Nelson K. Y. Leung
RMIT International Universtiy, Ho Chi Minh City, Vietnam
nelson.leung@rmit.edu.vn

Seung Hwan Kang
Payap University, Chiang Mai, Thailand
seung_h@payap.ac.th

Sim Kim Lau
University of Wollongong, NSW, Australia
simlau@uow.edu.au

Joshua Fan
University of Wollongong, NSW, Australia
joshua@uow.edu.au

Abstract

Ontology matching can be defined as the process of discovering similarities between two ontologies, and it can be processed by exploiting a number of different techniques. To provide a common conceptual basis of ontology matching for the Semantic Web, researchers have started to develop classifications to distinguish ontologies. The most significant one is the classification proposed by Shvaiko and Euzenat. Their approach is to compare different existing ontology mediation systems as well as to design a schema-based matching system.

In our findings, the above classifications contain some improper identifications and vague categories. There are three insufficient elementary matching techniques out of ten matching techniques: the language-based matching, the repository of structures, and the upper level formal ontology techniques. The language-based matching technique is normally performed prior to string-based technique and has no direct engagement in the actual similarity computation between two ontologies. The repository of structures technique is a dynamic approach used to compare fragments of two ontologies and to eliminate the dissimilar portions. It may be regarded as the follow-up step. The upper level formal ontologies technique is an approach that uses external source of common knowledge in the form of ontology. There is insufficient evidence that specifies the input and design guidelines.

This paper therefore proposes a design and input-specific classification framework of ontology matching techniques to address the above problems based on the findings of the literature survey. The proposed framework consists of the layers: executive approach layer, basic technique layer and input layer. The executive approach layer identifies heuristic, probabilistic reasoning and semantic reasoning to execute the identified seven elementary ontology matching techniques. To provide a guideline between the executive and input layers, the basic technique layer consists of string-based, linguistic resources, constraint-based, alignment reuse, graph-based, taxonomy-based and model-based matching techniques. The input layer is classified into elementary input and structural input to sum up the *characteristics of the actual inputs.*

The 3-tier classification framework provides an effective way to design a new mediation tool that not only identifies the type of matching technique, as well as a practical executive approach that incorporates input of mediation system with the input layer in the proposed framework.

Keywords– ontology matching, 3-tier classification framework, ontology, the semantic web

I. INTRODUCTION

Ontology is a common conceptual basis to specify vocabulary used in the Semantic Web [10]. As the number of ontologies grow, sharing and reuse of ontologies are common in practice. Ontology matching takes an important role in the process of ontology mapping and merging with the purpose of establishing semantic relationships between two ontologies. In general, ontology matching can be defined as the process of discovering similarities between two ontologies [17]. It determines the relationships holding two sets of entities that belong to two discrete ontologies [19]. In other words, it is the process of finding a corresponding entity in the second ontology for each entity (for example, concept, relation, attribute, and so on) in the first ontology that has the same or the closest intended meaning. This can be achieved by analysing the similarity of the entities in the compared ontologies in accordance with a particular metric [7], [11]. The correspondence can either be expressed by one to one function or one-to-many function. One-to-one function denotes an entity in an ontology can only have one similar entity in another ontology whereas one-to-many function addresses the fact that an entity may have more than one similar entities in another ontology [3].

This paper investigates the classification and application of matching techniques as well as some of the most significant ontology mediation systems to develop a 3-tier classification framework of ontology matching techniques. The aim of this paper is to provide a guideline to identify the type of matching technique as well as its related executive approach and for designing new mediation tool. This paper is organized as follows. Section 2 describes the background of ontology matching. Section 3 presents a literature survey of ontology matching. Section discusses a proposed 3-tier classification framework. Finally, conclusion is given at Section 5.

2. ONTOLOGY MATCHING

Ontology matching can be processed by exploiting a number of different techniques. To provide a common conceptual basis, researchers have started to identify different types of ontology matching techniques and propose classifications to distinguish them. For example, Abels et al. [1] propose a classification that consists of nine matching techniques based on existing literature studies. Another example is the classification developed by Shvaiko and Euzenat [20]. Building on the foundation of Rahm and Bernstein's [18] schema matching techniques classification, Shvaiko and Euzenat develop a meticulous classification to categorize elementary ontology and schema matching techniques. Their classification focuses on techniques that exploit ontology-level information excluding instance data. There are two synthetic classifications that can be viewed in top-down and bottom-up manner. The top-down view is called "granularity/input interpretation layer" which is based on granularity of match and then on how input information is interpreted. The bottom-up view is called "kind of input layer" and it is based on the kind of input requires in the matching process. "Granularity/input interpretation layer" and "kind of input layer" are further divided into one common layer called "basic techniques layer". Ten different types of elementary matching techniques are identified in this layer: string-based, language-based, constraint-based, linguistic resource, alignment reuse, upper level formal ontologies, graph-based, taxonomy-based, repository of structures and model-based.

3. LITERATURE SURVEY

This section presents a literature survey on some of the most significant mediation tools, frameworks and methods. Our focus is to examine their inherent matching process, in particular their similarity computation task at the matching stage of the process based on Shvaiko & Euzenat's classification [20]. In this way, a detailed description can be provided to demonstrate how these ten matching techniques are performed in the actual mediation environment.

Our finding consists of fifteen mediation tools, frameworks and methods with their inherent matching techniques. The most popular ontology matching techniques are string-based, taxonomy-based, constraint-based as well as linguistic resources techniques. Each of them is used by at least seven out of the fifteen mediation systems as shown in Table 1. In contrast, the least popular matching techniques are repository of structures technique and upper level formal ontologies. While the former technique is adopted by only one mediation system, the latter is not adopted by any system at all. Almost all systems in the survey incorporate a graph algorithm as their matching technique (either graph-based or taxonomy-based technique) with the exception of iPROMPT and Chimaera. For those who use graph algorithm as a matching technique except Glue and ITTalks, they include at least one additional matching technique in the system. Most of the mediation systems exploit multiple matching strategy which contains more than one matching technique. For instance, both COMA and COMA++ include six matching techniques in their inherent matching strategy. Thus leaving iPROMPT, Chimaera and Glue to engage with a single strategy in which only one matching technique is included in each system. In terms of execution approach, heuristic is widely implemented for carrying out string-based, language-based, constraint-based, linguistic resources, alignment reuses, graph-based, taxonomy-based and repository of structures matching techniques. Probabilistic reasoning approach, such as Bayesian network and machine learning, also play a part in the execution of taxonomy-based technique,

TABLE 1

SUMMARY OF MEDIATION SYSTEMS AND THEIR INHERENT MATCHING TECHNIQUES

	Automation	String-based	Language-based	Constraint-based	Linguistic Resources	Alignment Reuse	Upper Level Formal Ontologies	Graph-based	Taxonomy-based	Repository of Structures	Model-based
QOM	Full	Heuristic		Heuristic	Heuristic				Heuristic	Heuristic	
NOM	Full	Heuristic		Heuristic	Heuristic				Heuristic		
S-Match	Full	Heuristic	Heuristic		Heuristic				Heuristic		Semantic Reasoning
FCA-MERGE	Full		Heuristic		Heuristic						
IF-Map	Full	Heuristic							Heuristic		Semantic Reasoning
IPROMPT	Semi	Heuristic									
ANCHORPROMPT	Semi	Heuristic							Heuristic		
Chimaera	Semi	Heuristic									
COMA	Full / Semi	Heuristic	Heuristic	Heuristic	Heuristic	Heuristic		Heuristic			
COMA++	Full / Semi	Heuristic	Heuristic	Heuristic	Heuristic	Heuristic		Heuristic			
MAFRA	Full			Heuristic	Heuristic				Heuristic		
GLUE	Semi								Probabilistic Reasoning		
ITTalks	Semi								Heuristic or Probabilistic Reasoning		
OMEN	Full			Heuristic					Probabilistic Reasoning		
OLA	Full	Heuristic	Heuristic	Heuristic	Heuristic			Heuristic			

whereas semantic reasoning is the dedicated approach used to execute model-based technique. Out of the fifteen mediation systems, eight of them are capable of performing ontology matching automatically, five of them still rely on human intervention and the remaining two allow users to execute ontology matching either automatically or semi-automatically.

4. 3-TIER CLASSIFICATION FRAMEWORK

This paper proposes a design and input-specific classification framework of ontology matching techniques to address the shortcomings identified. As shown in Fig. 1, there are three main layers in the proposed framework, namely executive approach, basic technique and input layer. Language-based, upper-level formal ontologies and repository of structures matching techniques are excluded from the proposed framework to address the misidentification problem, thus leaving only string-based, linguistic resources, constraint-based, alignment reuse, graph-based, taxonomy-based and model-based matching techniques in the basic technique layer of the proposed framework.

There are two different ways to study the proposed framework, either by middle view or bottom-up view. The middle view describes the relationships among elementary matching techniques, executive approach and input types. This view not only indicates the approach required to execute a particular matching technique (for example, heuristic approach can be used to execute string-based technique), it also provides a guideline for designing new mediation tool (for instance, to exploit model-based as a mediation system's matching technique). A tool designer must ensure the input type and executive approach are propositional formulas and semantic reasoning respectively. The bottom-up view provides an easier way to identify the type of ontology matching technique and its executive approach simply by comparing input of mediation system with the input types on the input layer. For example, the matcher is most likely to be string-based if it takes names and descriptions of entities as input.

In the executive approach layer, we identify heuristic, probabilistic reasoning and semantic reasoning as three major approaches to execute the above seven elementary ontology matching techniques. Heuristic approach exploits rules for comparing syntactic features, properties, linguistic and structural information of two or more different ontologies [3]. Alternatively, probabilistic reasoning approach, such as Bayesian network and machine learning, can also be used to execute the taxonomy-based technique [5], [13], [16]. Probabilistic reasoning uses probability measurement to represent similarity of two concepts from two different taxonomies that are similar or having the same instances [3]. When two independent taxonomies contain a pair of similar nodes, for example node A and B, it is possible to induce new set(s) of similar nodes from the taxonomies by considering the probabilistic similarity measured between node A and the neighbours of node B and between node B and the neighbours of node A. A semantic reasoning approach first requires to translate relationships of all possible matching candidates of two ontologies into some forms of propositional formula, such as axioms, or local logics [8], [9], [12]. Subsequently, the approach adopts sound deduction method to validate the matching between two ontologies in accordance with the semantic of propositional formulas. For instance, propositional satisfiability solver is used to check possible matching candidates by validating their propositional formulas [3]. Model-based technique deals with input based on its semantic interpretation using well grounded deductive methods such as propositional satisfiability and description logics.

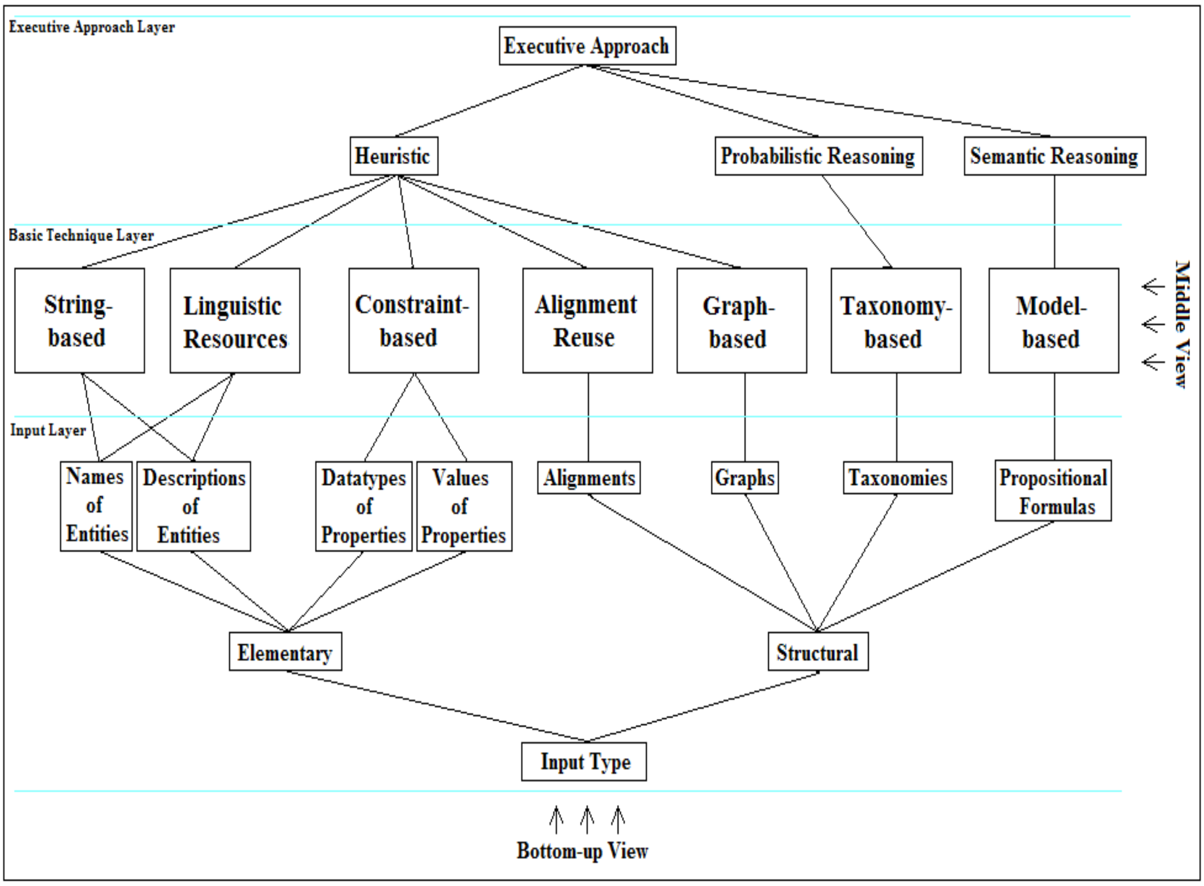

Fig. 1 A design and input-specific classification framework of ontology matching techniques

The basic technique layer includes string-based, constraint-based, linguistic resources, alignment reuse, graph-based and taxonomy-based matching techniques. In terms of string-based matching technique, heuristic approach establishes rules to determine the matching entities based on similarity computation of representational strings from two ontologies. While the string-based matching technique focuses only on calculating the string similarity of properties between two ontologies, the constraint similarity of the properties are taken care by the constraint-based technique. Here, heuristic approach applies rules to find matching properties based on the internal constraints that apply to each property. Linguistic resources matching technique uses a common knowledge or a domain specific thesaurus to derive meanings of entities in ontologies. By taking these meanings as input, heuristic rule is capable of determining the linguistic relations (such as synonyms, hyponyms and hypernyms) among the entities. For example, if a linguistic resources matcher derives from a common knowledge thesaurus that “Laptop” in Ontology A is a hyponym of “Computer” of Ontology B, then heuristic could determine “Laptop” in A is subsumed by “Computer” in B [2], [4]. Alignment reuse matching technique makes use of previously matching results at the level of ontology fragments or the entire ontologies to derive new matching results. Heuristic of the technique is built on transitive nature of the similarity relation between elements [2], [4]. This transitive nature means that if *x* is similar to *y* and *y* is similar to *z*, then *x* is very likely similar to z. In other words, it allows heuristic to reuse the available alignment information for matching analysis when Ontology B and C are required to match with each other, given that the matching results between A and B as well as that between A and C have been stored. Graph-based matching technique takes two ontologies in the form of labelled graphs as input and from which nodes from the ontologies are compared and analysed to derive the similarity of their neighbouring nodes. Similar to graph-based technique, taxonomy-based matching technique also takes graph as input. However, the graph intake here is more rigorous because neighbouring nodes on the graph are connected with is-a links to indicate they are superset/subset of each other. Heuristic can be applied to compare and identify similar nodes along the paths connected by is-a links [6], [7], [12], [14], [15].

In the input layer, inputs are divideded into two levels. The first level contains two keywords used to sum up the characteristics of the actual inputs on the second level: elementary and structural. Elementary input represents input that undergoes analysis in isolation during the matching process without the need of considering its relations with other entities. Names and descriptions of entities as well as data types and values of properties are classified as elementary input. In contrast, mediation systems analyse structural input in accordance with its relations with other entities in the process of ontology matching. Alignments, graphs, taxonomies and propositional formulas are categorized as structural input.

5. CONCLUSION

Based on the findings of the literature survey, existing ontology mediation system has the improper identification of matching techniques as well as the lack of an executive approach layer and a detailed input layer when matching between two ontologies. As a result, we propose a 3-tier classification framework of ontology matching techniques to address the above problems. The proposed framework consists of three layers, namely executive approach, basic technique and input layer. On one hand, the proposed framework provides a clear guideline on designing new mediation tool based on the middle view that describes the relationships among the three different layers. On the other hand, the bottom-up view provides an effective method to identify the type of the matching technique and its related executive approach simply by comparing input of mediation system with the input layer in the proposed framework. Further research will focus on implementation based on the proposed framework as a means of proof of concept.

REFERENCES

[1] Abels, S., Haak, L., Hahn, A. (2005). “Identification of Common Methods Used for Ontology Integration Tasks”. In: *1st International ACM Workshop on Interoperability of Heterogeneous Information Systems (IHIS)*, pp. 75-78. Bremen, Germany

[2] Aumueller, D., Do, H., Massmann, S., Rahm, E. (2005). “Schema and Ontology Matching with COMA++”. In: *ACM SIGMOD International Conference on Management of Data*, pp. 906-908. Maryland.

[3] Castano, S., Ferrara, A., Montanelli, S., Hess, G.N., Bruno, S. (2007), State of the Art on Ontology Coordination and Matching. BOEMIE Bootstrapping Ontology Evolution with Multimedia Information Extraction Project, FP6-027538 D4.4.

[4] Do, H., Rahm, E. (2002), “COMA – A System for the Flexible Combination of Schema Matching Approaches”. In: *28th International Conference on Very Large Databases (VLDB)*, pp. 610-621, Hong Kong.

[5] Doan, A., Madhavan, J., Domingos, P., Halevy, A. (2002), “Learning to Map between Ontologies on the Semantic Web”. In: *11th International Conference on World Wide Web*, pp. 662-673. Hawaii.

[6] Ehrig, M., Stabb, S. (2004), “QOM – Quick Ontology Mapping”. In: McIlraith, S.A., Plexousakis, D., van

Harmelen, F (eds.) *ISWC 2004*. LNCS, vol. 3298, pp. 683-697. Springer, Berlin/Heidelberg.

[7] Ehrig, M., Sure, Y. (2004), "Ontology Mapping – An Integrated Approach". In: Bussler, C., Davies, J., Fensel, D., Studer, R. (eds.) *ESWS 2004. LNCS*, vol. 3053, pp. 76-91. Springer, Berlin/Heidelberg.

[8] Giunchiglia, F., Shvaiko, P. (2003), "Semantic Matching". *The Knowledge Engineering Review Journal*, 18(3), 265-280.

[9] Giunchiglia, F., Shvaiko, P., Yatskevich, M. (2004), "S-Match: an Algorithm and an Implementation of Semantic Matching". In: Bussler, C., Davies, J., Fensel, D., Studer, R. (eds.) *ESWS 2004. LNCS*, vol. 3053, pp. 61-75. Springer, Berlin/Heidelberg.

[10] Gruber, T. (ed.) (1993), *Toward Principles for the Design for Ontologies used for Knowledge Sharing*, Kluwer Academic Pulishers, Padova, Italy.

[11] INTEROP. (2004), "Ontology Interoperability". State of the Art Report (SOA), WP8ST3 Deliverable, IST-508011.

[12] Kalfoglou, Y., Schorlemmer, M. (2003), "IF-Map: An Ontology-Mapping Method based on Information-Flow Theory". *Journal on Data Semantics* 1, pp. 98-127.

[13] Mitra, P., Noy, N., Jaiswal, A.: Ontology Mapping Discovery with Uncertainty. In: 4^{th} International Semantic Web Conference (ISWC). Galway, Ireland (2005).

[14] Noy, N., Musen, M. (2000), "PROMPT: Algorithm and tool for automated ontology merging and alignment". In: *17^{th} National Conference on Artificial Intelligence (AAAI)*. Texas.

[15] Noy, N., Musen, M. (2003), "The PROMPT Suite: Interactive Tools for Ontology Merging and Mapping". *International Journal of Human-Computer Studies*, 59(6), pp. 983-1024.

[16] Prasad, S., Peng, Y., Finin, T. (2002), "Using Explicit Information to Map between Two Ontologies". In: *1^{st} International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. Bologna, Italy.

[17] Predoiu, L., Feier, C., Scharffe, F., de Bruijn, J., Martin-Recuerda, F., Manov, D., Ehrig, M. (2006), State-of-the-art Survey on Ontology Merging and Aligning V2. EU-IST Integrated Project (IP) IST-2003-506826 SEKT: Semantically Enabled Knowledge Technologies, University of Innsbruck.

[18] Rahm, E., Bernstein, P. (2001), "2001, A Survey of Approaches to Automatic Schema Matching". *The International Journal on Very Large Data Bases (VLDB)*, 10(1), pp. 334-350.

[19] Shvaiko, P. (2004), "Classification of Schema-based Matching Approaches". In: *1^{st} International Semantic Web Conference (ISWC)*

[20] Shvaiko, P., Euzenat, J. (2005), "A Survey of Schema-based Matching Approaches". *Journal on Data Semantics*, IV, pp. 146-171

[21] Stumme, G. & Maedche, A. (2001), "2001 FCA-MERGE: Bottom-Up Merging of Ontologies". In: *7^{th} International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 225-230. Washington