

Similarity Measures and their Aggregation in Ontology Matching

# S. Jan, I.A. Shah, I. Khan, F. Khan and M. Usman

Abstract—The growing number of ontologies on the web has emerged new research issues in the field of semantic web. It is now common to have more than one ontologies on the web for a specific domain. These ontologies may have various types of heterogeneities and therefore it is a challenging task for information retrieval systems to utilize the knowledge of all available ontologies. A number of ontology alignment systems have proposed by researchers in the last decade to bridge this semantic gap. During an alignment process, the entities from different ontologies are compared to find the semantically similar entities. For this purpose, various similarity measures are discussed according to the type of ontological heterogeneities. This paper investigates most common types of heterogeneities exist in ontologies and matchers which can deal with such heterogeneities. Furthermore, the role of aggregation methods in alignment systems have also been analyzed and discussed.

*Index Terms*— Semantic Gap, Ontology Alignment, Similarity Measures and OWL

## I. INTRODUCTION

**O**NTOLOGIES play an very imperative role in semantic interoperability because they define basic terms, relations of a domain concept and rules for relating these terms [1], thus, enabling machines to process information between heterogeneous platforms and applications. NF. Noy *et. al.* [2] stated the main reasons for developing ontology are given as: (*i*) to share common understanding of the structure of information among people or software agents, (*ii*) to enable the reuse of domain knowledge and to make domain assumptions explicit, and (*iii*) to separate domain knowledge from the operational knowledge and to analyze domain knowledge.

Ontologies are used in almost every field of information systems like businesses, information security, bio-information and knowledge management [3], [4], [5], [6], [7] and [8]. Several context-aware systems have been developed using Semantic Web technologies [9] such as ontologies, Resource Description Framework (RDF) [10] and Web Ontology Language (OWL) [11]. The semantic web technologies are not only used for World Wide Web (WWW) but also for other information retrieval systems and even for personal devices as suggested in some research directions presented in [12] and then the idea was enhanced by integrating semantic search support in [13]. RDF is a standard model for data interchange between applications and is widely used to share and communicate knowledge. It also offers common properties and syntax to describe data and information. Extensible Markup Language (XML) only addresses the document structure but RDF provides a data model which can be extended for addressing ontology representation and relevant techniques. The RDF does need translation as domain model can be presented to define objects and relations. RDF is also capable to exchange the knowledge between different metadata languages [14].

However, one of the major limitations of RDF is that it cannot define the cardinality constraints. Several ontology languages have been proposed by research community which includes Simple HTML Ontology Extensions (SHOE) [15] and OIL [16]. OWL was originally designed to be used by such applications which need to process the information contents and representing machine-interpretable contents on the web. Compared to RDF, OWL adds more vocabulary with a formal semantics and it allows rules which are more expressive. The main advantage of OWL over the RDF is that its ability to define cardinality constraints to define ontologies. OWL itself is an evolution of DAML + OIL [17] and it is divided into three sub-languages; OWL-Lite, which provides hierarchy of classification and constraints; OWL-DL have maximum expressiveness with computational completeness and OWL Full, which has maximum expressiveness without computational guarantee.

The ontology alignment systems make use of such ontological information to match entities from different ontologies. This paper thoroughly investigates and discusses the types of heterogeneities and techniques to utilize such

Sadaqat Jan is with the University of Engineering and Technology, Mardan Campus, Pakistan (e-mail: sadaqat@nwfpuet.edu.pk). *Ph:* +92-937-9230295 *Fax* :+92-937-9230296.

Ibrar Ali Shah is with the University of Engineering and Technology, Mardan Campus, Pakistan.

Imran Khan is with the University of Engineering and Technology, Mardan Campus, Pakistan (e-mail: imrankhan@nwfpuet.edu.pk).

Fahim Ullah Khan is with the University of Engineering and Technology, Mardan Campus, Pakistan (e-mail: fahimwazir@gmail.com).

Muhammad Usman is with the University of Engineering and Technology, Mardan Campus, Pakistan (e-mail: usman@nwfpuet.edu.pk).

information for ontology matching purposes. The accuracy of the alignments produced by any system not only depends on the similarity measures but the aggregation technique also plays an important role.

The rest of the paper is structured as follows. Section II presents and explains the ontology alignment process and fundamental stages. Section III gives details about the importance of similarity aggregation techniques and its role in the accuracy of final alignments and finally, Section IV concludes the paper, followed by the references.

## II. ONTOLOGY ALIGNMENT PROCESS

### A. Ontology Heterogeneity

Ontology heterogeneities have been categorized in many aspects in literature and analysed in different research studies [18],[19],[20] and [21]. However, there are two most important and common types of heterogeneities, *semantic* and *terminological. Semantic* heterogeneity; which occurs due to various reasons like using different axioms or due to disparity in modelling the same concept in different ontologies for the same domain. For example, the object-property "address" might has been used for the concept namely "organization" in one ontology and for "Publisher" in the second one.

*Terminological* heterogeneity; which emerges by using synonyms for the same entity in different ontologies. In the Figure 1, for example, the entity named as "*Publisher*" in one ontology which may have a different name like "*PublishedBy*" in the second but both represent the same concept. The semantic heterogeneity has been one of the most challenging tasks in a matching process because it derives from the difference in design or scope of ontology domains in the process of knowledge presentation.

## B. Ontology Matching

Ontology matching process is to find the semantic mapping between two ontologies. Entities of the different ontologies are compared to find correspondences between them, however, they do not necessarily have to be the same but they should have certain degree of semantic similarity. This degree of semantic similarity can be used as the alignment threshold in the ontology alignment process. It has been a challenging task to find the semantic similarity between the entities of two heterogeneous ontologies. For this purpose, some information should be available about the internal structure of entities in order to match them.

OWL is an emerging language to represent ontologies in the semantic web and is recommended by the World Wide Web (WWW). Its vocabulary is used to describe the semantics of ontology and can also be used to find some indications for matching entities during the ontology alignment process. In the Figure 2, a part of the OWL syntax is shown, which is used for the same fraction of ontology which is shown in the



Fig. 1. Ontology matching example



Fig. 2. A fragment of OWL ontology

Fig. 1. For example, *owl:Class rdf:ID="Institution"* is used to define a class named as Institution. Similarly, the syntax *rdfs:subClassOf* defines a class which is a sub-class of another defined class in other ontology. The *owl:ObjectProperty* and *owl:DatatypeProperty* are used to define the object and data properties. The properties can also have sub-properties which are define by the syntax *rdfs:subPropertyOf*. The *rdfs:domain* and *rdfs:range* syntax are used to classify the range and domain of properties, showing that a property is associated to which classes and what type of values a property may have.

Fig. 2, the syntax owl:ObjectProperty In the rdf:ID="school" indicates the object-property labelled as "school" while syntax "<rdfs:range rdf:resource="#School"/" shows that the property is associated with the class named as "School". This kind of information greatly helps in defining the internal structure of an ontology. There are a large number of matchers which are used to find terminological heterogeneity.

These types of matchers like string-based and linguisticbased, do not take into account the structural position of the entity and operates on element level while comparing two entities from different ontologies. Such matchers are mostly used in schema based matching systems. For example, the "Publisher" and "PublishedBy" can be compared by using string based matchers to find the similarity. External resources are always helpful in finding matches where some background knowledge is required about the entity names.

WordNet is an example of the widely used external resource and many ontology alignment systems have exploited its capability in different ways. For example, several mapping systems have translated the entity labels to their respective WordNet senses and then drawn the mapping from there [22], [23], [24], while *J. kwan et al.* [25] exhaustively used the relationships of synsets to measure the lexical similarity between the entities. LOM [26] is another example of alignment tool which make use of lexicon-based matching.

### C. Ontology Alignment

The Ontology alignment process greatly varies and it depends on the techniques or algorithm used in the alignment system. The process may be varying in degree of mapping automation, the utilization of structural and lexical similarities and the degree of matching of such similarities. Mappings may be completed in one of the three modes, which includes manual, semi-automatic and automatic. In manual mapping, the user does the mapping by hand while in semi-automatic; the system suggests some mappings to the user for rejection or approval. Using automatic mapping, the system does all the process automatically without user's involvement. The manual mapping is the most time consuming but at the same time it gives more accurate matching results compared to the other two modes. The time and accuracy tradeoffs decision is made according to the application and its usage.

Alignment systems may also be different in the use of external resources in their matching processes such as web resources, external ontologies, dictionaries or lexical databases like WordNet<sup>1</sup> etc. Some of these systems use learning methods to improve mapping by using previous mapping results. Figure 3 shows a typical example of mapping two entities namely "*Publisher*" in source ontology and "*PublishedBy*" in the target ontology. Their structural similarity is exactly equal in terms of super classes while the string based similarity will not be equal by using any of the widely used string based matching techniques.

Semantically, the entities supposed to be matched by an alignment system, as it is suggested by the snippet of two ontologies given in the Figure 3; but, it totally depends on the algorithm used in the alignment system. Once the alignments are produced by alignment systems then these can be used in many ways for semantic interoperability between applications on the web or in other information retrieval systems for example [27] and [28], where the alignments are used for file retrieval.



Fig. 3. Mapping options of two entities from two different ontologies

### D. Similarity in Ontological Structures

Besides other similarities, the ontological structures can also be compared during the matching process. These comparisons are based on some of the intuitions that (i) If two classes from different ontology have similar upper classes in hierarchy, it is more likely that they represent the same concept in their respective ontologies (ii) If two classes from different ontology have similar sub classes in hierarchy, it is more likely that they represent the same concept in their respective ontologies (iii) If two classes from different ontology have similar properties, it is more likely that they represent the same concept in their respective ontologies (iv) if two entities have any combination of two or all the three mentioned similarities suggests more likelihood that they represent similar concept (v) if two entities have similar sibling classes, it is more likely that there is a degree of similarity between the entities. For example, as shown in the Figure 4, if an entity J have super classes A and B, and sub classes W, X and Y in an ontology O, and entity K have super classes A and B, and sub classes W, X and Y in ontology O', it is very likely that J and K represents the same concepts in O and O' respectively.

<sup>&</sup>lt;sup>1</sup> http://wordnet.princeton.edu/



Fig. 4. Comparison of super and sub classes of entities from different ontologies

Similarly, as shown in the Figure 5, if an entity J has a super class G, and sub classes W, X and Y in an ontology O, and entity K has super classes E and F, and sub classes W, X and Y in ontology O', up to some extent it is more likely that J and K represent the same concepts in O and O', respectively. The structural similarity of entities J and K have 50% of similarities in terms of super and sub classes in ontologies O and O'.



Fig. 5. Comparison of sub classes of entities from different ontologies

Similarly, as shown in figure 6, entities J and K have similar super classes *i.e.* A and B, but entity J have sub classes U and V in ontology O, while entity K have sub classes X and Y in ontology O'. This, again, suggests some likeliness that Jand K represent the same concepts in O and O', respectively. The structural similarity of entities J and K in the Figure 6 have 50% of similarities in terms of super and sub classes in the ontologies O and O'. Furthermore, sibling classes can also be compared between the entities from two different ontologies.



Fig. 6. Comparison of sub classes of entities from different ontologies

## III. SIMILARITY AGGREGATION

The fundamental similarity measures which includes string based similarity, semantic based similarity and structure based similarity are the most widely used measures in state of the art alignment systems. However, the real issue arises when these similarity measures give different results for the same entity during the matching process. Different techniques have been used to aggregate the results of different similarity matchers. For example, some of the alignment systems have used the average of all values returned by all similarity measures. The PROMPT [29] system was developed to support various ontology mediation techniques and it suggests the classes and properties for aligning. It uses linguistic and structural similarity measures to map two entities. PROMPT performs all the tasks automatically and resolves any found conflict by suggesting new mappings to the users.

PROMPT is a very useful alignment system where users are involved in the aligning processes. LILY [30] also uses linguistic and structural similarity measures to align the entities from different ontologies. It applies a propagation strategy to generate further alignments and then uses classic image threshold selection algorithm for the best suitable threshold. Finally, it extracts the final results based on the most stable marriage strategy. The QOM [31] ontology alignment system employs the RDF triples as features and it applies heuristic method for mapping the entities. It computes the similarities by using various functions and heuristics but avoids the complete pair-wise evaluation of ontology trees. OOM uses sigmoid function to aggregate the results of various similarity measures. The response time of QOM alignment system is faster than PROMPT. The alignment systems presented in [32], [33], [34] [35] and [36] uses different mapping approaches and aggregation techniques.



Fig. 7. A generic alignment process which uses more than one matcher in the alignment process

The existing alignment systems used diverse types of aggregation techniques which includes probabilistic, weighted sum and weighted product. Some of the alignment systems have used Fuzzy aggregation and Rough sets as presented in [37].

For more accurate alignment results, the representative alignment systems have used the basic matchers because these matchers provide crucial information of an entity in a given ontology. However, it is the aggregation technique which decides either to map an entity or not. Figure 7 shows a generic diagram of alignment system which uses more than matchers and then aggregates the results of all matchers.

#### IV. CONCLUSION

Various types of ontological heterogeneities have been discussed and analyzed which can be considered while developing an ontology alignment system. The existing matchers which compare similarities between two entities from different ontologies, have also analyzed thoroughly. The importance of aggregation technique in alignment systems has also been discussed and it has been concluded that the aggregation techniques necessitate more research efforts in order to get more accurate results from alignment systems.

#### REFERENCES

- R. Neches, R Fikes, T Finin, T Gruber, R Patil, T Senator, and W Swartout. "Enabling Technology for Knowledge Sharing", AI Magazine, 12(3):36–56, (1991).
- [2] N. F. Noy and D.L. McGuinness. "Ontology development 101: A guide to creating your first ontology". Technical Report SMI-2001-0880, Stanford Medical Informatics, (2001).
- [3] P Lord, S Bechhofer, MD Wilkinson, G Schiltz, D Gessler, D Hull, CA Goble, and L Stein. "Applying Semantic Web Services to Bioinformatics Experiences Gained, Lessons Learnt". In Proceedings of the 3rd International Semantic Web Conference, volume 3298 of LNCS, pages 350–364. Springer-Verlag, (2004).
- [4] A. Herzog, N. Shahmehri, and C. Duma. An ontology of information security. International Journal of Information Security and Privacy, 1(4):1--23, October-December (2007).
- [5] MC. Daconta, LJ Obrst, and KT Smith. The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management. Wiley, (2003).
- [6] D. Trastour, C Bartolini, and C Preist. Semantic Web Support for the Business-to-Business E-Commerce Lifecycle. In Proceedings of the 11th International Conference on World Wide Web, pages 89–98. ACM, (2002).
- [7] C. Pedrinaci, J. Domingue, and A. K. Alves de Medeiros. A Core Ontology for Business Process Analysis. In 5th European Semantic Web Conference, (2008).
- [8] T. Berners-Lee, James Hendler, Ora Lassila. The Semantic Web. Scientific American, May (2001).
- [9] Resource Description Framework (RDF), http://www.w3.org/RDF/, Last accessed May (2011).
- [10] M. Smith, C. Welty, and D. McGuiness," Web Ontology Language (OWL) Guide", August (2003).
- [11] S. Decker, Sergey Melnik, Frank Van Harmelen, Dieter Fensel, Michel Klein, Jeen Broekstra, Michael Erdmann, Ian Horrocks, "The Semantic Web: The Roles of XML and

RDF", IEEE Internet Computing, v.4 n.5, p.63-74, September (2000).

- [12] S. Jan, M. Li and H. Al-Raweshidy, "File Annotation and Retrieval on Mobile Devices", Personal and Ubiquitous Computing – Springer-Verlag London, pp volume 15, issue 7, pp. 771-779 (2011)
- [13] S. Jan, M. Li, G. Al-Sultany, H. Al-Raweshidy and I.A Shah "Semantic file annotation and retrieval on mobile devices", International Journal of Mobile Information System, vol. 7, no 2, pp. 107-122, (2011)
- [14] S. Decker, Sergey Melnik, Frank Van Harmelen, Dieter Fensel, Michel Klein, Jeen Broekstra, Michael Erdmann, Ian Horrocks, "The Semantic Web: The Roles of XML and RDF", IEEE Internet Computing, v.4 n.5, p.63-74, September (2000).
- [15] J. Heflin, J Hendler, and S Luke. "SHOE: A knowledge representation language for internet applications". Technical report, Dept. of Computer Science, University of Maryland at College Par, 1999. Technical Report CSTR- 4078 (UMIACS TR-99-71), (1999).
- [16] D. Fensel, FV Harmelen, I Horrocks, DL McGuinness, and PF Patel- Schneider. "OIL: An Ontology Infrastructure for the Semantic Web" IEEE Intelligent Systems, 16(2):38–45, (2001).
- [17] D. McGuinness, L., Fikes, R., Hendler, J. and Stein, L. A., "DAML+OIL: An Ontology Language for the Semantic Web". IEEE Intelligent Systems, 17(5): 72-80 (2002).
- [18] M. Benerecetti, P. Bouquet and C. Ghidini, Contextual reasoning distilled, Journal of Experimental and Theoretical Artificial Intelligence 12, pp. 279–305 (2000).
- [19] C. Ghidini, Giunchiglia F "A semantics for abstraction". In: M'antaras Rde, Saitta L (eds) Proceedings of ECAI'2004, including PAIS 2004, pp 343–347, (2004).
- [20] H. Wache, T. Voegele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Huebner. "Ontology-based integration of information a survey of existing approaches", In Proceedings of the workshop on Ontologies and Information Sharing at the International Joint Conference on Artificial Intelligence (IJCAI), pages 108–117, (2001).
- [21] J. Euzenat, Shvaiko, P., "Ontology matching", Springer, Heidelberg (2007)
- [22] P. Bouquet, L. Serafini, and S. Zanobini, "Peer-to-Peer Semantic Coordination", Journal of Web Semantics, 2(1), (2005).]
- [23] F. Giunchiglia, P. Shvaiko, and M. Yatskevich, "Semantic Schema Matching". in Proceedings of CoopIS'05, volume 3760 of LNCS, pages 347 – 360, (2005).
- [24] Giunchiglia, P. Shvaiko, and M. Yatskevich, "Discovering Missing Background Knowledge in Ontology Matching". in Proceedings of ECAI, (2006).
- [25] J. Kwan and Hwan-Sung Yong, "Ontology Matching based on hypernym, hyponym, holonym, and meronym sets in word net", International journal of Web & Semantic Technology (I West), Vol.1, No.2, April (2010).
- [26] J. Li, "LOM: A Lexicon-based Ontology Mapping Tool", Proceedings of the Performance Metrics for Intelligent Systems (PerMIS. '04), (2004).
- [27] S. Jan, M. Li, G. Al-Sultany and H. Al-Raweshidy, "Semantic based file retrieval on resource limited devices with ontology alignment support", Proc. of Eighth International Conference on FSDK. pp. 2699 – 2702 (2011).
- [28] S. Jan, M. Li, G. Al-Sultany and H. Al-Raweshidy, "File annotation and sharing on low-end mobile devices" 7<sup>th</sup> International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 2973-2977, (2010).

- [29] N. F. Noy and M. A. Musen, "The PROMPT suite: interactive tools for ontology merging and mapping." International Journal of Human-Computer Studies 59(6): 983-1024, (2003).
- [30] P. Wang, Xu, B. LILY: The Results for the Ontology Alignment Contest OAEI 2007. In Proceedings of ISWC 2007 Ontology Matching Workshop. Busan, Korea, (2007).
- [31] M. Ehrig, and S. Staab, "QOM: Quick Ontology Mapping", Proceedings of the 3rd International Semantic Web Conference (ISWC), (2004).
- [32] P. Xu, Y. Wang, L. Cheng and T. Zang, "Alignment Results of SOBOM for OAEI 2010," Ontology Matching, pp. 203, (2010).
- [33] F. Isabel, Cruz, Flavio Palandri Antonelli, and Cosmin Stroe.,
  "AgreementMaker: Efficient Matching for Large Real-World Schemas and Ontologies", PVLDB, 2(2):1586–1589, (2009).
   [34] J. Noessner and M Niepert., "CODI: Combinatorial
- [34] J. Noessner and M Niepert., "CODI: Combinatorial Optimization for Data Integration–Results for OAEI 2010". In Proceedings of the 5th International Workshop on Ontology Matching, page 142, (2010).
- [35] F. Hamdi, Safar, B., Niraula, N., Reynaud, C. "TaxoMap in the OAEI 2009 alignment contest", In: ISWC Workshop on Ontology Matching, Chantilly (VA US), pp. 230–237, (2009).
- [36] M. Bhatt, Andrew Flahive, Carlo Wouters, Wenny Rahayu, David Taniar: "MOVE: A Distributed Framework for Materialized Ontology View Extraction", Algorithmica, 45(3): 457-481, (2006).
- [37] S. Jan, M. Li, G. Al-Sultany and H. Al-Raweshidy, "Ontology alignment using Rough Sets" Proc. of IEEE international Conference on Fuzzy Systems and knowledge Discovery 2011, pp. 2683 – 2686, doi: 10.1109/FSKD.2011.6020069, (2011)

**Dr. Sadaqat Jan** is working as Assistant Professor in Computer Software Engineering Department at University of Engineering and Technology, Mardan Campus, Pakistan. He received his PhD degree from Brunel University, UK and Master degree from NWFP University of Engineering and Technology Peshawar, Pakistan. His research interests include Grid Computing, Mobile Computing, Information retrieval, Distributed Systems, Semantic Web, Knowledge Engineering and Computer Networks. He is a member of IEEE.

**Ibrar Ali Shah** received his B.Sc. and M.Sc. degrees in Computer Systems Engineering from NWFP University of Engineering & Technology, Peshawar, Pakistan in 2004 and 2007 respectively. Currently, he is with Wireless Networks & Communications Centre, Brunel University, London UK working towards his PhD. His research mainly focuses on cooperative and non-cooperative behavior in Wireless Mesh Networks where he is working on routing and channel assignment problems by using game theoretic models for fairness, QoS and throughput optimization.

**Dr. Imran Khan** is working as Assistant Professor in Department of Telecommunication Engineering at University of Engineering and Technology, Mardan Campus, Pakistan since 2004. He received his PhD and Master degree from AIT, Thailand. He is student member of IEICE and IEEE. Mr. Khan's research interests include performance analysis of wireless communications systems, OFDM, OFDMA, MIMO, BICM-ID based systems and cooperative networks.

Fahim Ullah Khan is working as Lecturer in Department of Computer Software Engineering at University of Engineering and Technology, Mardan Campus, Pakistan since 2004. He received his Master degree from UET Peshawar. His research interests include Semantic Web, NLP, Probabilistic approaches and uncertainty analysis.

**Muhammad Usman** is working as Assistant Professor in the Department of Computer Software Engineering at University of Engineering and Technology Mardan Campus, Pakistan since 2006. He received his Master's Degree in Computer Engineering from Center of Advanced Studies in Engineering (CASE), Pakistan. Mr. Usman's research interests include Embedded Systems, Distributed Systems, Reconfigurable Computing, Semantic Web, Service Oriented architecture and Web Services.