Sindh Univ. Res. Jour. (Sci. Ser.) Vol.45 (1):11-14 (2013)

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

An Overview of Concept Comparison Practices used in Ontology Alignment

S. JAN[++], I. KHAN*, I. A. SHAH**, G. AHMAD***, A. KHATTAK***, G. A. AL-SULTANY

Babylon University, Information Technology College, Networks Department, Iraq

Received 24th November 2012 and Revised 17th January 2013

Abstract: Concept comparisons play a vital role to find semantic similarity between entities from different ontologies. Semantically similar entities are used to exploit the knowledge of more than one ontology developed for the same or similar domain. In this paper, we have presented an overview of the most commonly used comparison techniques and their significance. Different combinations of comparison techniques were formulated in order to evaluate the alignment results in terms of precision, recall and f-measure. The results suggest that all these comparison techniques play an important role in finding the correspondences between the entities. It is almost impossible for a single comparison technique to handle all types of ontological heterogeneities.

Keywords: Semantic web, Ontology Matching, Alignment, Ontology heterogeneity.

1. INTRODUCTION

It is essential to bridge the heterogeneous resources on web in order to share the knowledge efficiently between users and applications. For this purpose, the concept of semantic web was presented by Berners *et al.*, 2001, where web of data can be processed by machines. The semantic web implementation has originated novel and diverse research directions. Ontologies were developed to explicitly define the concepts in order to enable the web users to access and share the knowledge based information. It is now common to find more than one ontology for a specific concept or domain which raised another issue of dealing with heterogeneities. This semantic heterogeneity is dealt with ontology alignment and the research community have proposed many state of the art alignment systems in the recent years. However, it still requires intensive investigations of the alignment techniques and approaches in order to align different ontologies perfectly. The mapped entities or alignments can be used for aligning, merging and/or transforming different ontologies.

During the ontology alignment process, entities from one ontology are compared and mapped, semantically, with the same entities from the second ontology. Ontology alignment systems use different types of comparison techniques to find the semantically equal entities. The most commonly used techniques include string-based, linguistic-based and structure-based comparisons. The results of these comparisons, referred as matchers, are than used to decide the final mapping of two entities. These alignments can also be used in other information retrieval systems. For example, a file retrieval framework proposed by Jan *et al.*, 2011a uses a generic ontology for the keywords used to annotate different files. The alignments were used in the extended search module presented in Jan *et al.*, 2011b for searching required files. We have analysed the significance of these three comparison techniques in ontology alignment and used the benchmark ontologies in our experiments in this paper. Four combinations of these matchers were formulated and the final aliments were decided based on the mean values of the results obtained.

The rest of the paper is structured as follows: Section 2 presents material and methods used in our experiments. An overview is presented about the fundamental three comparison techniques. Results and discussion are presented in Section 3 along with details of the evaluation process and benchmark ontologies. Finally, the conclusion is presented in Section 4 followed by references.

2. MATERIALS AND METHODS

String based Comparisons

The names or labels of ontological entities are considered as plain texts during the comparison process. Generally, the string comparison process starts after performing string normalization. The normalization process may include several steps which include the conversion of two strings to lower or upper case and elimination of special characters, dashes and blank spaces. The text normalization process itself plays a imperative role in string comparisons, for example the entities "*Published_by*", "*Published_by*", "*published_by*", "*PublishedBy*" and "*Published-By*" are normalized to "*publishedby*". To further evaluate string similarities, the following techniques are most widely

++Corresponding author: S. Jan, E-mail: sadaqat@brunel.ac.uk Phone: +44 (0)1895 274000
*School of Engineering and Design, Brunel University, London, UK
**Department of Telecommunication Engineering, University of Engineering and Technology, Peshawar (Mardan Campus), Pakistan
***Department of Electrical Engineering, University of Engineering and Technology, Peshawar, Pakistan

used in state of the art ontology alignment systems. Stoilos *et al.,* 2005 proposed a string based metric which can compute the string similarity based on their commonalities as well as their differences. This metric is calculated by subtracting the sum of differences and winkler similarity from the commonalities of both strings. Substring string-metric are used to calculate the commonalities. Using the n-gram technique Kondrak, 2005, the similarity is computed by taking the longest common subsequence to include the n-grams. Using Jaro, 2007 and Winkler, 1999 Jaro-Winkler technique, the number of matches are computed first and then the number of transpositions. The adjustment is the final rescoring of match score for initial characters of both strings in comparison.

A technique proposed by (Levenshtcin, 1966) measures the difference between two strings through the number of single character edits which are required to change one string into other.

Structure based Comparisons

Hu *et al.*2005 proposed a structural matcher "GMO", which uses bipartite graphs for comparing the structural of two entities from different ontologies. Maedche *et al.,* 2002 used meta data based on definition of similarities for the structural comparison. The similarity distance technique which computes the relation amongst the ontological entities was proposed by Zhong, 2002. Each entity which represents a concept in ontological hierarchy is assigned a value while the distance between two concepts represents a path to a common parent concept.

The comparison of super and sub classes was used in an ontology alignment system by Jan, *et al.,* 2011 in order to find the similarity between the structures of two entities. Similarly, the internal structure can be compared to find the similarity between two entities for example, the comparison of properties belong to two entities can be used to find the semantic similarity. The Structural Topological Dissimilarity on Hierarchies technique presented by Valtchev, 1997 uses the dissimilarity of elements in a ontological hierarchy based on the distance from their nearest common parent concept.

Linguistic based Comparisons

The linguistic based matchers generally use external resources to find the semantic similarity between two entities. These resources include WordNet, Dublin Core, thesauri, language dictionaries or specific databases. Language dictionaries are used in linguistic matchers where ontologies are developed in languages other than English.

Similarly, the databases are useful for domain specific ontologies, for example, Library catalogue ontologies. There are several ontology alignment systems or linguistic matchers which have used the WordNet as an external resource, for example, Kwak, *et al.,* 2010 have used the features of WordNet such as hypernym, hyponym, meronym and holonym for lexical comparison of entities. LOM (Li, 2004) is lexicon based matcher for comparing the lexical similarities of entities. The aggregation of results obtained from individual similarity measures also play an important role when deciding the final alignments of entities from two ontologies.

Similarity measures used in Experiments

Before comparing the strings, we have normalized the strings to eliminate blank spaces and dashes. We used the *Smoa* to compute the string similarity between two entities as follows.
Let

- e_{01} be the entity from ontology-1
- e_{02} be the entity from ontology-2
- $String_Comparison$ be the string similarity between e_{01} and e_{02} .

Thus, $String_Comparison$ can be calculated using the Equation (1).

$$String_Comparison(e_{01}, e_{02}) = Smoa\ (e_{01}, e_{02}) \quad (1)$$

The WordNet is used to compute the linguistic similarity based on synonyms, hyponyms, hypernyms and antonyms.

Let

- w_{01} be the word from ontology-1
- w_{02} be the word from ontology-2
- $Linguistic_Comparison(\ w_{01},\ w_{02})$ be the linguistic similarity between w_{01} and w_{02}
- $syn(w_0)$ be the set of synonyms
- $hyp(w_0)$ be the set of hyponyms and hypernyms.
- $ant(w_{01})$ be the set of anonyms of w_{02}

thus,

$$Linguistic_Comparison(\ w_{01},\ w_{02}) = \begin{cases} 1 & if\ \ w_{02} \in syn(w_{01}) \\ 0.5 & if\ \ w_{02} \in hyp(w_{01}) \\ 0 & if\ \ w_{02} \in ant(w_{01}) \end{cases} \quad (2)$$

Using the structural similarity, we compare the super classes, subclasses and properties of the classes for class entities. The structural similarity is computed by $Structure_Comparison(e_{01}, e_{02})$, which is the average of three similarities.

3. RESULTS AND DISCUSSION

Evaluation Measures

The most widely used measures namely; *precision*, *recall* and *F-measure* are considered for all experiments.

Let SA_d denotes the set of output alignments, SA_t denotes the complete set of correct alignments. Then the precision, recall and *F-measures* can be defined using equations (4), (5) and (6) respectively.

$$Precision = \frac{|SA_d \cap SA_t|}{|SA_d|} \quad (4)$$

$$Recall = \frac{|SA_d \cap SA_t|}{|SA_t|} \quad (5)$$

$$F - measure = \frac{2 \times Prec \times Rec}{Prec + Rec} \quad (6)$$

For all experiments, the Ontology Alignment Evaluation Initiative (OAEI) 2010 benchmark data sets were used which includes a number of ontologies with different levels of ontological heterogeneities. The benchmark library and details are available at http://oaei.ontologymatching.org/2010/benchmarks/index.html

In OAEI data set, all the ontologies are based on a bibliography topic. The test ontology 101 is the base ontology and is used as the reference ontology in OAEI data sets. This ontology contains 33 named classes, 24 object properties and 40 data properties. In addition, it has 76 individuals in which 56 of them are nameless. The data set is divided into three groups which include simple, systematic and real life ontologies, as described in the **(Table-1)**. The simple group (1xx) has 4 ontologies and have very little differences amongst them. The systematic group (2xx) ontologies are the most difficult and challenging for alignment systems as these ontologies have diverse types of ontological deficiencies. These deficiencies may include;

- In this group, usually classes are replaced with other classes or prolonged.
- Some of the entity names are replaced with synonyms.
- Foreign language have been used to represent some of the comments at different level.
- Some of the properties are blocked or their restrictions on classes are ignored.
- Some instances are ignored at various levels.
- Hierarchies might have been expanded, suppressed or flattened.

Furthermore, ontologies in group 3xx are real world ontologies which are provided by different institutions. They are left unchanged in the benchmark data sets.

Table-1 Details of Benchmark ontologies sets

Sets	Detail of heterogeneities
101-104	Same structure but entity names may be different
201-210	The hierarchical structure is similar Different linguistics are used in some levels
221-247	Different structures and labels or names are the same
248-266	Both the structure and linguistics are different
301-304	Real world ontologies

To evaluate the efficacy of similarity measures, we have formulated 4 test scenarios namely *A, B, C* and *D.* Scenario *A* uses the mean value of two measures namely, $String_Comparison(e_{01}, e_{02})$ and $Linguistic_Comparison(w_{01}, w_{02})$, scenario *B* uses the mean value of $Linguistic_Comparison(w_{01}, w_{02})$ and $Structure_Comparison(e_{01}, e_{02})$, scenario *C* uses the mean value of $String_Comparison(e_{01}, e_{02})$ and $Structure_Comparison(e_{01}, e_{02})$ while scenario *D* uses the mean value of all three matchers $String_Comparison(e_{01}, e_{02})$, $Linguistic_Comparison(w_{01}, w_{02})$ and $Structure_Comparison(e_{01}, e_{02})$. All scenarios were implemented on group (1xx) ontologies and then 12 ontologies from group (2xx). From group (2xx), 12 ontologies were selected randomly for each round. Experiments were carried out five times for each group and results presented in this paper are the average values of corresponding groups.

Using group 3xx ontologies, the combination of string and structural matchers have outperformed the remaining three combinations as shown in **(Fig. 1)**. The main reason is that ontologies in group 3xx have more string heterogeneities as compare to linguistic and structural heterogeneities. The test scenario *B* does not use string matcher and it can be observed from **(Fig. 1)** that it has the lowest f-measure. Furthermore, the linguistic matcher did not perform well because some of the ontological entities have prefix texts and therefore all those combinations which include linguistic matcher have lower f-measures.

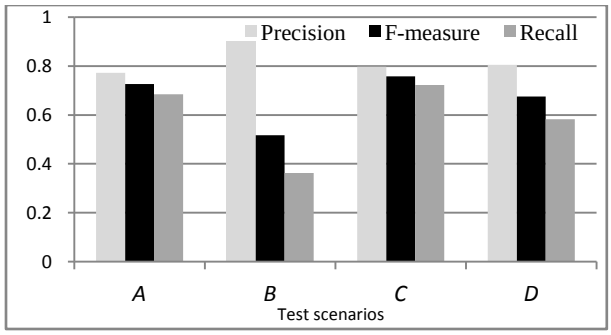


Fig. 1: Comparison of test scenarios using benchmark ontologies in group 3xx

Using group 2xx ontologies, the test scenario *D* has the best f-measure because it uses the combination of all three matchers as shown in **(Fig. 2)**. As mentioned in previous sections, group 2xx ontologies have most types of heterogeneities, therefore, using test scenario *D*, the combination of three matchers have showed their significance and tackled the corresponding heterogeneities. As predicted, the combination of different matchers performs well when the types of heterogeneities in ontologies are not known to an alignment system.

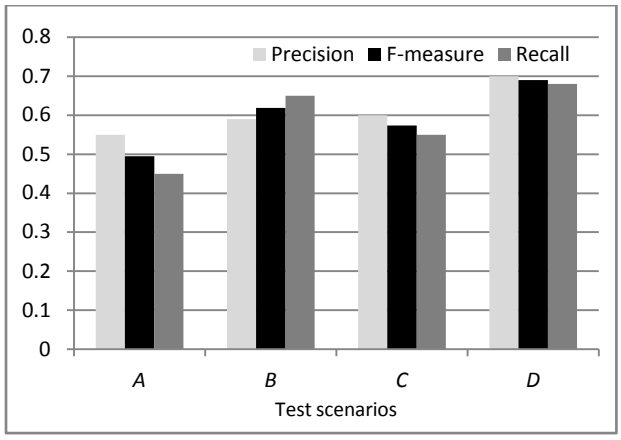


Fig. 2: Comparison of test scenarios using benchmark ontologies in group 2xx

4. CONCLUSION

In this paper, an overview of most commonly used similarity measuring techniques and their significance in over all ontology alignment is presented. For this purpose, four test scenarios were formulated and benchmark ontologies from OAEI were used to evaluate the performance of different combinations. The results show that no single similarity measure is adequate enough to give accurate alignments. Specifically, when the ontologies have diverse types of heterogeneities, the combination of various similarity measures can perform well. However, the aggregation of such measures still need to be investigated for more accurate alignments.

REFERENCES:

Berners L. T., H. James and O. Lassila (2001) The semantic web Journal of Scientific American 284, **(5):** 34-43.

Hu W., N. Jian, Y. Qu and Y. W. Gmo, (2005) A graph matching for ontologies in Proc. of K-CAP Workshop on Integrating Ontologies: 41-48. CiteSeerX, Banff, Canada.

Jan S., M. Li, G. Al-Sultany, H. Al-Raweshidy and I. A. Shah (2011) Semantic file annotation and retrieval on mobile devices Mobile Information Systems, Vol. **(7):** Issue 2:107-122.

Jan S., M. Li, G. Al-Sultany and H. Al-Raweshidy (2011) Semantic based file retrieval on resource limited devices with ontology alignment support) IEEE Eighth International Conference on Fuzzy Systems and Knowledge Discovery - (FSKD): 2699-2702.

Jaro M. A. (2007) Probabilistic linkage of large public health data files Statistics medicine Vol. **(14)**: 491-498.

Kondrak, G. (2005) N-gram similarity and distance Lecture Notes in Computer Science Vol. **(3772):** 115-126.

Kwak J. A. and H. S. Yong (2010) Ontology Matching based on hypernym, hyponym, holonym, and meronym sets in word net International journal of Web and Semantic Technology (I West) Vol. **(1):** No.2.44-48.

Lcvenshtcin V. I. (1966) Binary Coors Capable Or 'Correcting Deletions, Insertions, And Reversals In Soviet Physics- Doklady Vol. **(10):** No. 8. 78-86.

Li J. L. (2004) A Lexicon-based Ontology Mapping Tool IEEE Proceedings of the Performance Metrics for Intelligent Systems, Gaithersburg, Md, Washington:114-119.

Maedche A. And V. Zacharias (2002)Clustering ontology-based metadata in the semantic web Principles of Data Mining and Knowledge Discovery: 383-408.

Stoilos G., G. Stamou, (2005). A string metric for ontology alignment The Semantic Web–ISWC: 624-637.

Valtchev P. and J. Euzenat (1997) Dissimilarity measure for collections of objects and values Advances in Intelligent Data Analysis Reasoning about Data: 259-272.

Zhong J., H. Zhu and, Y. Yu (2002) Conceptual graph matching for semantic search In: Proceedings of Conceptual Structures: Integration and Interfaces: 92-106, Springer-Verlag London, UK.