# Multi-strategy Based Matching Technique for Ontology Integration

S. Kumar and V. Singh

**Abstract** In modern era, ontology integration is most prominent and challenging problem in various domains of data mining. It helps greatly on defining interoperability in information processing systems. Ontology integration institutes interoperability by deriving semantic similarity and correspondences between the entities of ontologies. Ontology matching plays an integral role in whole integration process by identifying the similarity measure between source and target ontologies using various matching techniques, e.g. Element-level techniques and Structure level techniques. The scope of this paper is to discuss existing ontology matching techniques and propose a multi-strategy ontology matching approach. In Proposed algorithm is a multi-strategy matching approach, where multiple similarity measure are combined and finally an ontology tree based on the binary tree is created. Ontology mapping is used for creating ontology binary tree. The hybrid approach of Ontology integration performs significantly better than its predecessor with single similarity measures.

**Keywords** Ontology • String-based technique • Structure-based technique • Levenshtein distance • Matching • Similarity • Ontology integration

# **1** Introduction

In modern era, entire aspect of data integration is motivated by the information technology and social media, data cannot be stored on a single computer or with single semantics. "Semantic" means the meaning of data, the data stored at various geographical locations is linked based on their semantic similarity. The semantic

V. Singh e-mail: viks@nitkkr.ac.in

© Springer India 2015 L.C. Jain et al. (eds.), *Computational Intelligence in Data Mining - Volume 3*, Smart Innovation, Systems and Technologies 33, DOI 10.1007/978-81-322-2202-6\_12

S. Kumar  $(\boxtimes) \cdot V$ . Singh

Computer Engineering Department, NIT Kurukshetra, Kurukshetra, Haryana, India e-mail: er.satishkumar149@gmail.com

values of data can be a good source to define the relation with user query. The researchers presented the technology to help manage these data called semantic web. Semantic web has been used in various fields such as Information Systems, Search Engine etc. [1, 2]. Very large database systems to handle this with advent of data mining, because the large database analyze by finding patterns or relationships of data is an advantage of data mining. Research in area of semantic web integration is not vet widely explored, since there is a management tool for data mining of semantic web is less and data from the semantic web is stored in a format that cannot be used directly in data mining [3]. Semantic integration creates lots of interest of research in various disciplines, such as databases, information-integration, and ontologies. In the area of knowledge sharing and reuse, ontology integration is an important area of open research in domain of ontology engineering/management. Last few years, ontology integration catches the eyes of researcher in the field of ontology based systems development. There are few identified meanings of ontology integration. First, create ontology in a domain by reusing other existing ontologies in the same domain. Second, ontology is build by merging other existing ontologies into single ontology. Finally, when an application has built by using one or more ontologies, ontology integration is achieved through ontology matching and ontology alignment. Ontology matching concept is used to find contact between ontologies by finding the semantic similarities between the ontologies. Semantic similarity measures play a consequential role in text cognate area and application in ontology matching. Semantic similarities of data are being used in various fields e.g. Information retrieval, query answering etc. [4]. Identification of related attribute between the ontologies is the primary stage in ontology matching. In traditional applications ontology matching is a paramount operation e.g., Ontology evolution, Ontology integration, Data integration and Data warehouses [5]. An Ontology Integration algorithm is proposed, which begin with ontology identification followed by ontology matching and finally ontology mapping. Matching process is prime focus of this paper. The matching is base on some similarity measure. The similarity on words can be two ways, lexically and semantically. If the sequence of character is similar then they are lexically similar and if the two words are opposite of each other having similar meaning then they are semantically similar [6]. There are various types of similarity matching techniques to measure the similarity, the element-level techniques between two ontologies. An element-Level technique discusses the stringpredicated techniques, Language-predicated techniques and Constraint-Based techniques etc. On the other hand Structure-Level techniques are used to measure the similarity based on the structure. Structure-Level techniques e.g. Graph-predicated technique, Taxonomy-predicated technique and Constraint-predicated techniques are discussed briefly in next section in paper. Traditional approach uses single technique for measuring similarity, between source and target ontology. The proposed algorithm for multi-strategy based matching based on the multiple similarity measures, the similarities between the ontologies and then ontology alignment by combining the string- predicated similarity and structure-predicated similarity. In string based similarity measure, primarily levenshtein distance is used to measure the similarity between the ontologies; while for structure based similarity measure WordNet based semantic similarity is used [3]. In algorithm similarity measure is calculated by combining the similarity measure of both previous approaches. Experimentation performed for the result verification and validation of algorithm, which shows that multi-strategy algorithm perform significant better than some of single strategy based algorithm.

### 2 Related Work

While many multi-strategy approaches has been proposed for ontology integration over the span of 10 years, among these approaches, recent approach of multi-strategy ontology integration using stack by Xia et al. [7]. Another similar effort made by Gang, L, Kunlun Wang and D. Liu, they proposed a multi-strategy ontology mapping. In algorithm characteristics of ontology structure and the instance of mapping on the ontology similarity are used for the ontology integration. Most recent work on the ontology integration is done by Fuqiang and Yongfu [8], as in proposed multi-strategy ontology integration approach, in which concept similarity between the ontologies is computed from the concept name, concept attribute and concept relationship.

# **3** Ontology Integration

Ontology provides the vocabulary of concepts that describe the domain of interest and a specific meaning of terms used in the vocabulary [5]. Ontology also provides the shared specification of conceptualization. Ontology reuse is one most adequate research issue in the ontology integration field. Ontology reuse, consist of two important processes, ontology merging and ontology integration. In ontology merging, ontology created in one domain by merging ontologies from two different domain ontologies. On the other hand in ontology integration, ontology is created by combining, assembling ontologies [9]. Therefore, ontology integration is major challenge and research issue. Main focus of this paper is on ontology integration by



Fig. 1 Ontology integration process

using ontology matching combined with ontology mapping because it is a crucial part in ontology alignment. Ontology integration's typically involves four activities, shown in Fig. 1.

# 4 Ontology Matching

The ontology matching process is utilized to measures the homogeneity between set of ontologies [9]. The matching process determines an alignment A', for two ontologies  $O_1$  and  $O_2$ , for this some parameters of the matching process are used like,

- (i) The utilization of an input alignment A
- (ii) Matching parameter like weights, threshold
- (iii) External resources utilized by the matching process like erudition-base and domain-categorical area.

**Definition 1** (*Matching Process*) The matching process considered as function f which emanates from the ontologies to match  $O_1$  and  $O_2$ , an input alignment 'A', a set of parameters p and a set of oracles and resources r, back to an alignment 'A' between these ontologies [4, 9].

$$\mathbf{A}' = \mathbf{f} \ (\mathbf{O}_1, \mathbf{O}_2, \mathbf{A}, \mathbf{p}, \mathbf{r})$$

Typical structure for ontology matching is shown in above Fig. 2. In the Ontology integration matching plays a crucial role and decider, how effectively ontologies are integrated. In ontology engineering, there are two ways of defining matching; first the matching is derived base on the element level e.g. String, Language, Constraint. Another ways matching is based on the structure used to store ontology i.e. Graph, Taxonomy, Instances etc. [9]. The classification is shown in below Fig. 3.







#### 4.1 Element-Level Technique

An element-Level matching technique, in the measure of homogeneous attribute on ontologies and their instances is isolation from their relation with the attribute and their instances [9]. This category of ontology matching techniques, consist of string-predicated, language-predicated and constraint-predicated methods. A string-predicated technique, is employ to define the match description of ontology entities, this technique draws matching between input ontologies by considering ontology as a sequence of letter in an alphabet [2, 9], it also consider various distance measures methods for assertion the distance between strings, some such are popular like Edit-distance, Hamming-distance etc. Alternative approach, which is based on string parity method, in which is string parity defines the homogeneous attribute between the strings. If the strings are identical, string parity returns 1, otherwise returns 0, which means, strings are not identical.

(S is set of strings) (s, t are Input strings) |s| = length of the string s, |t| = length of the string ts[i] for  $i \in [1, |s|]$  latter at position i of s.

String-predicated technique, which is contributory to quantify the similar attribute between the input strings, if we employ very similar string to represent the same concepts [10]. If we utilized the synonyms with dissimilar structure, then technique may be open-handed to the low kind attribute. The result of this technique is more supplementary, if we utilized the similar strings [11]. Language-predicated technique is also used to find the similarity between the words [9, 12]. This technique takes names as words/tokens, English language-predicated techniques utilized the various methods, which considered strings as sequence of characters. This method takes string as an input and fragment these string in words and these words may be identified sequence of words. Constraint-predicated technique [9, 12] is based on the internal structure of the ontology entities rather than comparing the designations or terms. This technique compares internal structure, because in this approach comparing the ontology entities with the other entities which are related.

# 4.2 Structure-Level Technique

Contrary to Element-level techniques, structure level measures the homogeneous attribute between the ontology and their instances to compare their relation with each other entities or their instances [9, 10]. In first among structure level techniques, Graph-predicated technique measures the homogeneous attribute between the pair of nodes of mapped graph of ontologies, as the ontology are predicated onto the graph with a specific positions in the graph [13]. Input for Graph-predicated technique is two ontologies labelled graphs. In the labelled graph, if two nodes are similar then their neighbours are additionally somehow homogeneous. Taxonomy-predicated technique is utilizing the graph-predicated approach but in this technique only the specialized relation are considered for the matching [9]. This type of technique is utilized as a comparison resource for matching classes. The perception after the taxonomy technique is that, specialization connect ontology/ tokens, those are already homogeneous (as a super set or subset of each other), therefore their neighbour are additionally somehow homogeneous. Instance-predicated technique is utilized for the comparison of sets of instances of classes [9]. On the substructure of the comparison it is decided that classes are match or not. This technique is relies on the set-theoretic reasoning and statistical techniques. The approach proposed in the paper, for the ontology matching, combine both string based similarity and structure based similarity to be used to define the ontology similarity measures. In String-based similarity, levenshtein distance technique which is a number of edit operation such as, insertion; deletion subtractions that are required to convert a string into other string, is used for measuring the similarity between entities of ontologies [10]. WordNet in case of Structured-based similarity, which is based on semantic similarity, is used to measure the similarity between ontologies [3].

# **5** Proposed Approach and Example

A hybrid ontology matching algorithm is proposed (Fig. 4) in which, matching ontologies of input documents or domains are mapped into a binary tree/Ontology tree. In traditional techniques, ontology matching technique prefer to base their matching approach on single similarity measure, while including more similarity measures to identify the measure is helpful to achieve good set of integrated ontology. Multiple similarity measures are combined for matching process in proposed approach. First, similarity measure, is based on string (grammar) based similarity measure and second is structure based similarity measure. In previous section basic details are discussed. For string based similarity measure, primarily levenshtein distance is used to measure the similarity between the ontologies. In Levenshtein distance method [9], the number of edit operation insertion, deletion and subtraction of characters to convert the one string into another. Another Proposed Algorithm: Input (D1, D2) Step 1: Ontology Identification/ creation from Documents. Step 2: Store Ontologies into Document vectors. Step 3: Begin Apply ontology matching technique (Set of Ontologies) using binary search tree /\* String based similarity measure method using Levenshtein distance \*/ End Step 4: Begin Apply another ontology technique (Set of Ontologies) using binary search tree /\* Structure based similarity measure method using WordNet \*/

End

Step 5: Combine the results of above two methods by adding the similarity matrices.

#### Output (Global Ontology)

Fig. 4 Proposed algorithm

technique is structure based similarity measure, where similarity measures between entities using WordNet based semantic similarity. WordNet is a lexical database of English. In proposed algorithm, both matching techniques have been used for measuring the similarities between entities of concepts using binary tree. Proposed algorithm iteratively constructs matching searches for the entities in both Source ontology S and target ontology T (see illustration in example as given in the next section). Matrix will be evaluated using matching techniques (string and structure based technique) and finally the hybrid approach based calculated matrix shows the better evaluation result to the user. Flowchart of proposed algorithm is shown in Fig. 6. Designed algorithm is useful as it provides a better ontology matching approach to match the entities of ontologies.

As shown in Fig. 5 two ontologies are given source ontology S and target ontology T with different entities  $(S_1, S_2, S_3)$  and  $(T_1, T_2, T_3)$  are concepts. Other entities are relations, properties etc.

As shown above, ontologies/concepts have been compared from source ontology S with target ontology T, and based on the similarity other entities in S such as Name and instances can be matched with the corresponding entities such as Name in target ontology T. In proposed algorithm (Fig. 6), two similarity measures have been used to measure the similarity between source ontology S and target ontology T.

First, Similarity is string based similarity using (levenshtein distance), following Similarity matrix shows the similarity values of ontology matching,



Fig. 5 Source ontology S and target ontology T



Fig. 6 Flowchart of proposed algorithm

Multi-strategy Based Matching Technique for Ontology Integration

$$Mat_{(String-Based)} = \begin{bmatrix} S_1rT_1 & S_1rT_2 & S_1rT_3 & S_1rT_4 \\ S_2rT_1 & S_2rT_2 & S_2rT_3 & S_2rT_4 \\ S_3rT_1 & S_3rT_2 & S_3rT_3 & S_3rT_4 \\ S_4rT_1 & S_4rT_2 & S_4rT_3 & S_4rT_4 \end{bmatrix}$$
$$= \begin{bmatrix} 0.21 & 0.25 & 0.13 & 0.14 \\ 0.15 & 0.21 & 0.15 & 0.10 \\ 0.21 & 0.26 & 0.21 & 0.10 \\ 0.14 & 0.12 & 0.13 & 1.0 \end{bmatrix}$$

Similarity matrix for Structure based similarity using WordNet is shown below,

$$Mat_{(Structured-Based)} = \begin{bmatrix} S_1rT_1 & S_1rT_2 & S_1rT_3 & S_1rT_4\\ S_2rT_1 & S_2rT_2 & S_2rT_3 & S_2rT_4\\ S_3rT_1 & S_3rT_2 & S_3rT_3 & S_3rT_4\\ S_4rT_1 & S_4rT_2 & S_4rT_3 & S_4rT_4 \end{bmatrix}$$
$$= \begin{bmatrix} 0.0 & 0.0 & 0.0 & 0.11\\ 0.0 & 0.0 & 0.0 & 0.0\\ 0.0 & 0.0 & 0.0 & 0.0\\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}$$

Mat<sub>(String-Based)</sub> shows, the calculated string similarity matrix based on the levenshtein distance and Mat<sub>(Structured-Based)</sub> shows the calculated structured similarity matrix based on WordNet based semantic similarity. The matrix has been calculated on the basis of formal matching techniques. In proposed technique, for the similarity calculation, addition of both previous similarity matrixes is done, as shown below

$$Mat_{(Hybrid-Approach)} = (Mat_{(String-Based)} + Mat_{(Structured-Based)})/2.0$$

$$Mat_{(Hybrid-Approach)} = \begin{bmatrix} S_1rT_1 & S_1rT_2 & S_1rT_3 & S_1rT_4 \\ S_2rT_1 & S_2rT_2 & S_2rT_3 & S_2rT_4 \\ S_3rT_1 & S_3rT_2 & S_3rT_3 & S_3rT_4 \\ S_4rT_1 & S_4rT_2 & S_4rT_3 & S_4rT_4 \end{bmatrix}$$
$$= \begin{bmatrix} 0.10 & 0.12 & 0.06 & 0.12 \\ 0.07 & 0.10 & 0.07 & 0.05 \\ 0.10 & 0.13 & 0.10 & 0.05 \\ 0.07 & 0.06 & 0.06 & 1.0 \end{bmatrix}$$

In traditional techniques, prefer only either string based similarity method or structured based similarity. The performance of single similarity measure in matching is not significant as the approach with multiple similarity measure. In next section, hybrid approach of multiple similarity measures based approach is explained.

# 6 Experiment and Quality of Measures

In order to evaluate this method, we used two pairs of ontologies as shown in Table 1. Source ontology S and target ontology T. Source ontology contains the entities Faculty, Associate-Professor, Assistant-Professor and Name (S1, S2, S3, and S4). Target ontology contains the concepts Academic-Staff, Lecturer, Senior-Lecturer and Name (T1, T2, T3 and T4). First step, Sources and target ontologies are created.

Second step, apply the string based similarity measure using levenshtein distance. After apply this method a similarity matrix Mat (String-Based) between the entities is calculated on the basis of levenshtein distance.

$$Matrix_{(String-Based)} = \begin{bmatrix} S_1 rT_1 & S_1 rT_2 & S_1 rT_3 & S_1 rT_4 \\ S_2 rT_1 & S_2 rT_2 & S_2 rT_3 & S_2 rT_4 \\ S_3 rT_1 & S_3 rT_2 & S_3 rT_3 & S_3 rT_4 \\ S_4 rT_1 & S_4 rT_2 & S_4 rT_3 & S_4 rT_4 \end{bmatrix}$$
$$= \begin{bmatrix} 0.21 & 0.25 & 0.13 & 0.14 \\ 0.15 & 0.21 & 0.15 & 0.10 \\ 0.21 & 0.26 & 0.21 & 0.10 \\ 0.14 & 0.12 & 0.06 & 1.0 \end{bmatrix}$$

Table 1	Source	and	target	ontology
---------	--------	-----	--------	----------

Ontology name	Concepts
Ontology 1	Faculty, assistant-professor, associate-professor, name
Ontology 2	Academic-staff, lecturer, senior-lecturer, name



Fig. 7 Graph of string based similarity

As shown in Fig. 7 a tree or graph is constructed on the basis of calculated string-based matrix i.e.  $Matrix_{(String-Based)}$ . Third step, structure based similarity measure method has been used for calculating the similarity matrix Mat (Structured-Based) between concepts based on their structure. Structure based similarity measure used the WordNet based semantic similarity measure method. WordNet is a lexical database of English.

$$Mat_{(Structured-Based)} = \begin{bmatrix} S_1rT_1 & S_1rT_2 & S_1rT_3 & S_1rT_4 \\ S_2rT_1 & S_2rT_2 & S_2rT_3 & S_2rT_4 \\ S_3rT_1 & S_3rT_2 & S_3rT_3 & S_3rT_4 \\ S_4rT_1 & S_4rT_2 & S_4rT_3 & S_4rT_4 \end{bmatrix}$$
$$= \begin{bmatrix} 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 \end{bmatrix}$$

After the matrix calculation a tree or graph is constructed on the basis of similarity matrix as shown in Fig. 8. Fourth step, the similarity matrix of both methods is combined. Similarity matrixes are added to evaluate the purpose.

$$Matrix_{(Hybrid-Approach)} = (Mat_{(String-Based)} + Mat_{(Structured-Based)})/2.0$$

$$Matrix_{(Hybrid-Approach)} = \begin{bmatrix} S_1rT_1 & S_1rT_2 & S_1rT_3 & S_1rT_4 \\ S_2rT_1 & S_2rT_2 & S_2rT_3 & S_2rT_4 \\ S_3rT_1 & S_3rT_2 & S_3rT_3 & S_3rT_4 \\ S_4rT_1 & S_4rT_2 & S_4rT_3 & S_4rT_4 \end{bmatrix}$$
$$= \begin{bmatrix} 0.10 & 0.12 & 0.06 & 0.12 \\ 0.07 & 0.10 & 0.07 & 0.05 \\ 0.10 & 0.13 & 0.10 & 0.05 \\ 0.07 & 0.06 & 0.06 & 1.0 \end{bmatrix}$$



Fig. 8 Graph of structure based similarity



Fig. 9 Graph of Hybrid Based Similarity

After calculating similarity matrix measure for the individual, Matrix (Hybrid-Approach), based on the resulting matrix from methods, a tree or graph is constructed based on the combined matrix, and the resultant ontology is shown in Fig. 9. The given two ontologies are considered source ontology S and target ontology T, in example S contains 4 entities and T contain 4 concepts. Following mentioned parameters are used for comparative analysis of proposed algorithm in which quality of similarity measure use to compare the effectivity of matching techniques. These parameter are used in some traditional ontology matching techniques to justify the expediency of techniques also,

(1) **Precision** is a value defined in the range between 0 and 1; the higher the value, the fewer wrong mapping computed.

$$Precision = \frac{number of correct found alignments}{number of found alignments}$$

(2) **Recall**, is a value defined in the range between 0 and 1; the higher this value, the smaller the set of correct mappings which are not found.

$$Recall = \frac{number of correct found alignments}{number of found alignments}$$

(3) **F-measure** is a value defined in the range between 0 and 1, F-measure value also known as global measure of matching quality. F-Measure used the mean of precision and recall [3].





F-Measure =  $\frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$ 

On the basis of Precision, Recall and F-measure parameter values, as shown in above Fig. 10, various conclusions are pinched on the propped algorithm. The values used for comparison are various matching techniques e.g. Semantic based Techniques, String based matching, and Hybrid Based matching, Other hand values of three factors (Precision, Recall, F-measure) with respect to each of the techniques (Semantic based Techniques, String based matching, Hybrid Based matching). The Semantic based techniques are the poor among all techniques, as precision and recall is quite low. The String based matching techniques are better than the semantic based techniques, as the matching performed on the data structure on which the ontologies are stored/kept with some logical relation among them, due to this logical relation between the ontology the probability of correctness increased which result into improve values of Precision, Recall and F-measure. The proposed algorithm combine the similarity values of Semantic based techniques and String based techniques to derived the similarity measure and as shown in the above graph (Fig. 10). This algorithm outperformed the previous techniques in all three values, precision, recall, F-Measure.

#### 7 Conclusion

Ontology matching between the entities plays an important role in ontology integration. Ontology integration is an important process of ontology reuse, other approach is ontology merging. Ontology integration is strongly based on similarity measure between input ontology (source, target). Traditionally, two types of techniques are used to evaluate the matching among the ontology and each approach have certain degree of limitations. The proposed algorithm for ontology matching is based on the multi-strategy approach for the ontology matching. In the approach similarity measures of the two traditional approaches are combined and this combined (hybrid) value is used for defining mapping between the set of ontology. The levenshtein distance similarity measures in string based techniques and WordNet similarity measure in Semantic based techniques are used. The experimentation are performed on large amount of datasets (documents, ontologies) to validate the authenticity of proposed algorithm, which shows that the proposed algorithm outperformed the traditional algorithm in all three aspect (Precision, Recall, F-Measure), which are considered for the comparative study.

### References

- 1. Berners-Lee, T., Handler, J., Lassila, O.: The Semantic Web. Scientific American (2001)
- Huiping, J.: Information retrieval and the Semantic Web. In: IEEE International Conference on Educational and Information Technology (ICEIT), vol. 3, pp. 461–463. Chongqing, China (2010)
- 3. Perez-Lopez, A., Blace, R. Fisher, M., Hebeler, J.: Semantic Web Programming. Wiley Publishing, Inc., New York (2009)
- Huang, L., Hu, G., Yang, X.: Review of ontology mapping. In: 2nd International Conference on Consumer Electronics, Communications and Networks (CECNet), pp. 537–540. Yichang (2012)
- Shvaiko, P., Euzenat, J.: Ontology matching: state of the art and future challenges. IEEE Trans. Knowl. Data. Eng. 25(1), 158–176 (2013)
- Ye, B., Chai, H., He, W., Wang, X., Song, G.: Semantic similarity calculation method in ontology mapping. In: 2nd International Conference on Cloud Computing and Intelligent Systems (CCIS), vol. 3, pp. 1259–1262. Hangzhou (2012)
- Xia, H., Zheng, X., Hu, X., Tian, Y.: Multi-strategy ontology mapping based on stacking method. In: Sixth International Conference on Fuzzy Systems and Knowledge Discovery (2009)
- 8. Fuqiang, L., Yongfu, X.: The method of multi-strategy ontology mapping. In: ICCIS (2011)
- 9. Euzenat, E., Shvaiko, P.: Ontology Matching. Springer, Heidelberg (2013)
- Rodrfguez, A., Egenhofer, M.: Determining semantic similarity among entity classes from diferent ontologies. IEEE Trans. Knowl. Data. Eng. 15(2), 442–456 (2003)
- Li, C.M., Bo, S.J.: The research of ontology mapping method based on computing similarity. Sci. Technol. Inf. 1, 552–554 (2010)
- 12. Zhe, Y.: Semantic similarity match of ontology concept based on heuristic rules. Comput. Appl. 12 (2007)
- Doan, A.H., Jayant, M., Pedro, D. et al.: Learning to map between ontologies on the semantic web. In: Eleventh International World Wide Web Conference, Honolulu Hawaii, USA (2005)