ISSN: 2521-0874 (Print)
ISSN: 2521-0505 (Online)

Contents List available at RAZI Publishing
Acta Informatica Malaysia(AIM)
Journal Homepage: http://www.razipublishing.com/journals/acta-informatica-malaysia
https://doi.org/10.26480/aim.01.2017.07.09

A Review on Feature based Approach in Semantic Similarity for Multiple Ontology

Nurul Aswa Omar, Shahreen Kasim, Mohd Farhan Md Fudzee, Azizul Azhar Ramli, Hairulnizam Mahdin, Seah Choon Sen

Faculty of Computer Science & Information Technology, Universiti Tun Hussein Onn Malaysia,
Beg Berkunci 101, 86400 Parit Raja, Batu Pahat, Johor Darul Takzim, Malaysia.
{nurulaswa, shahreen, farhan, azizulr, hairuln}@ uthm.edu.my, seanseah0702@gmail.com

ARTICLE DETAILS

Article history:

Received 22 January 2017
Accepted 03 February 2017
Available online 05 February 2017

Keywords:

Semantic similarity, features based, ontology, multiple ontology, cross ontology, heterogeneous sources.

ABSTRACT

Measuring semantic similarity between terms is an important step in information retrieval and information integration which requires semantic content matching. Semantic similarity has attracted great concern for a long time in artificial intelligence, psychology and cognitive science. This paper contains a review of the state of art approaches including structure based approach, information content based approach, and feature based approach and hybrid approach. We also discussed similarity according to their advantages, disadvantages and issues related to multiple ontology especially on method in features based approach.

1. Introduction

Similarity is the quality or condition of being similar, however different definitions of similarity have been discussed before. The difference of the definition of similarity refers to certain situation. Among them refers to [1] similarity can be defined based solely on the joint probability distribution of the concepts involved. However, in this study we believe that semantic similarity is define as the closeness of two concepts, based on the likeliness of their meaning. This refers to similarity between two concepts in a taxonomy or ontology [2] .

The history of semantic similarity has been used for years in psychology and cognitive science where different models have been proposed [3]. Besides that, semantic similarity has also been used in searching for similarities between images and visual [4]. However, by referring to [5] semantic similarity in recent years is widely used in obtaining the similarities between concept or between words where it is important to assist information extraction tasks [6] such as semantic annotation [7] and ontology learning [8], helping to discover semantically related terms.

Semantic similarity also is widely used in information retrieval tasks [9][10][11], to improve the performance of current search engines [12], information integration [11], ontology matching, to discover correspondences between entities belonging to different ontologies [11], [13], semantic query routing, to choose among the set of possible peers only those relevant and bioinformatics and to assess the similarity between proteins [14]. In addition, semantic similarity also can play an important role in both predicting and validating gene product interactions and interaction networks [15].

Many ontologies have been develop for various purposes and domains to represent an effective means of knowledge sharing within controlled and structured vocabulary. In recent years, information retrieval and data integration have emphasized the use of ontologies and semantic similarity as a mechanism for comparing objects that can be retrieved or integrated across heterogeneous repositories. However, semantic similarity have a few approach that have been develop in recent years and this paper will evaluate which approaches is suitable to use with multiple ontology.

2. Ontology Based Method

Ontology is defined as a formal, explicit specification of shared conceptualization. This definition means that ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents [16]. Ontology is an effective way to share knowledge within controlled and structured vocabulary [17]. Many ontologies have been developed for various purposes and domains [10][18][19]. According to [20] ontology is important in enabling interoperability across heterogeneous systems and semantic web applications. Ontology was develop with offer structured and unambiguous representation of knowledge in the form of conceptualizations which causes research on semantic similarity using ontology to increase [21]. There are a few ontologies used for semantic similarity such as; WordNet [19] is a lexical database for general English covering most of the general English concepts and supporting various purposes; and biomedical domain for example Unified Medical Language System (UMLS) includes many biomedical ontologies and terminologies (e.g., MeSH, Systemized Nomenclature of Medicine Clinical Term (SNOMED-CT), ICD family [10].

2.1 Classification semantic similarity approaches according to ontology

Several approaches for determining semantic similarity have been proposed. In ontology, semantic similarity approach can be classified into single ontology and cross ontology [2], [11]. Four approaches that could be used to find the similarity between concepts are:

(i) The hierarchical structure based approach

The ontology based approach takes into an account the path length and depth of an ontology. This is also known as edge counting approach or structure based approach. This approach computes the similarity in terms of the shortest path between the concepts in the ontology. Path length approach is based on ontology structure, in which the ontological primary relationship are connected through is-a type relation. Thus this similarity calculates the shortest path while the degree of similarity is determined based on path length. There are various measurements for path length approach which have been used by [22] and [23]. Meanwhile, the depth relative approach considers the depth of the edges connecting two concepts in structure ontology. It computes the depth from root to the target concept. [24], [25] and [26] are examples of the similarity method mention in this type.

(ii) Information content based approach

Information content (IC) based measurement which is also known as corpus based determines the similarity between two concepts based on probabilities to each concept in ontology on word occurrences in a given corpus. However this approach is less commonly used due to ontology which causes the occurrence probability of a node to decrease when the layer of the node goes deeper, and hence the IC of the node increases. Therefore, the lower a node is in the hierarchy, the greater its IC [27]. The various information content based measurement are [28], [29] and [30].

(iii) Features based approach

Feature of terms based approach takes into account terms that are several

represented as collections of features and also the specific differentiating features of each concept. Various feature based measurement are [31][32] and [33].

(iv) Hybrid based approach

Hybrid based approach combines different sources of information to measure the score of similarity between concepts. These approach usually consider. features such as attribute similarity, ontology structure, information content and depth of LCA node [27]. The major advantage of these approaches is if the knowledge of an information source is inadequate then it may be derived from the alternate information sources [2]. Hence the quality of similarity measure would be improved. Some representative of this approach are [34] and [35].

Table 1 presents a summary of semantic similarity approach according to ontology. The use of semantic similarity has been used in two categories of ontology namely single and multiple ontology. Semantic similarity in single ontology is compared to terms from the same ontology, for instance from WordNet itself and multiple ontology or also known as cross ontology is compared to terms from different ontologies such as WordNet and MeSH.

Table 1. Classification method according to category of ontology

Approach	Category Ontology	Example Research	Advantage	Disadvantage
Information Content	Single Ontology	[28], [36], [30]	• Suitable to use in big ontologies such as WordNet and Biomedical ontology.	• Require big and fine grained taxonomies/ontologies with a detailed taxonomical structure
Feature	Single Ontology	[31], [3]	• Exploit more semantic knowledge than edge-counting approaches • Evaluate both commonalities and differences of compared concepts	• Depends on features such as glosses or synsets that can cause limitation of application to ontologies • Suitable for big ontologies • Depends on the weighting parameter
	Multiple ontology	[31], [32], [33],[37]		
Structure	Path Length approach: Single ontology	[22], [23]	• Depends on graph model which requires a low computational	• Edge-counting measures only consider the shortest path between concept pairs. • Several taxonomical paths are not taken into account when using wide ontologies. • The number and distribution of common and non- common taxonomical ancestors are neither considered • The structure of every ontology is different and cannot be compared directly
	Path Length approach: Multiple Ontology	[10]		
	Depth approach: Single Ontology	[24], [25], [26]		
Hybrid	Single Ontology	[38], [34]	• Improved performance in terms of increased accuracy	• Difficult to combine different approaches where the characteristics are different • Complex algorithm

3. Semantic Similarity Multiple Ontology/ Cross Ontology

Nowadays with mushrooming information sources on the web, there is a need to develop measurements that will compute similarity among concepts in different ontologies [2], [11]. Cross ontology measurement will match the words from different ontology. The cross ontology often needs hybrid or feature based approach. This is due to the structure and information content between diverse ontologies cannot be compared directly [2]. Similarity measurements between concepts in different ontology are classified as:

(i) Approach path length based

Information about this approach is as mentioned in section 2.1(i). An example measurement for this approach is [10].

(ii) Approach based on features based on terms

Information about this approach is as mentioned in section 2.1(iii). Various features based measurement are [31], [32] and [33] .

In this study, we concentrate on the use of approaches based on features of terms to measure the similarity concept between two ontologies. Feature based approach is more general and potentially used in multiple ontology because the concept of two different ontologies also have a different structure. This is due to the structure between diverse ontologies that cannot be compared directly [33][21] and [2].

Other works in this similar method are Tversky [31], Rodriguez and Egenhofer [32] and X-similarity [33]. Tversky was developed to represent objects as a collection of features and similarity is described as a feature matching process. Equation (1) from Tversky which X and Y correspondences to sets of a and b where |X∩Y| is set function intersect and |X-Y| denotes the relative complement of Y in X. Further, α and β > 0 are parameters of the Tversky index. The Tversky method is as follows:

$$S(a,b) = \frac{|X \cap Y|}{|X \cap Y| + \propto |X-Y| + \beta |Y-X|}, S(a,b) \text{ is similarity between a and b} \quad (1)$$

In the meantime, Rodriguez and Egenhofer methods [32] also use features to obtain similarity measure. Their similarity function determines similar entity classes by using a matching process over synonym sets, semantic neighborhoods and distinguishing features that are classified into parts, function and attributes. To compute the synonym set, semantic neighborhood and feature matching, Equation (2) as shown below is used where ap and bq is the entity class of ontologies p and q, α is a function that defines the relative importance of the non-common characteristics:

$$S(a,b) = \frac{|X \cap Y|}{|X \cap Y| + \propto (a,b)|X-Y| + (1-\alpha(a,b))|Y-X|} \quad (2)$$

$$\text{Where } \alpha(a^p, b^q) = \frac{depth\ (a^p)}{depth\ (a^p) + depth\ (b^q)},\ if\ depth(a^p) \leq\ depth\ (b^q)$$

or

$$\alpha(a^p, b^q) = 1 - \frac{depth\ (a^p)}{depth\ (a^p) + depth\ (b^q)}, if\ depth(a^p) > depth\ (b^q)$$

In order to integrate the information obtained from similarity assessments of synonym sets, distinguishing feature and semantic neighborhoods, they propose a similarity function that is defined by the weighted sum of the similarity of each specification component as shown in Equation (3). The functions Sw, Su and Sn are the similarity between synonym sets, features and semantic neighborhoods between entity classes a of ontology p and b of ontology q and Ww, Wu and Wn are the respective weights of the similarity of each specification component.

$$S(a^p, b^q) = W_w.S_w(a^p, b^q) + W_u.S_u(a^p, b^q) + W_n.S_n(a^p, b^q) \quad for\ W_w, W_u, W_n \geq 0 \quad (3)$$

X-Similarity, a novel cross-ontology similarity method developed by Petrakis et. Al [33]. X-similarity relies on matching between synsets and term description sets. Rodriguez and Egenhofer [32] used α parameters to calculate the depth of the terms in the two ontologies while according to Petrakis et. al [33] cross ontology matching should not depend on ontology structure information. Due to this, Petrakis et. al [33] propose replacing Equation (2) to Equation (4) below with a plain set similarity. Where A and B denote synset or term description sets.

$$S(a,b) - \frac{|A \cap B|}{|A \cup B|} \quad (4)$$

They also proposed Equation (5) where the set similarities are computed per relationship type (e.g., is-A and part-Of) because they believe that not all terms in the neighborhood of a term are connected with the same relationship, where i denotes relationship type.

$$S_{neighborhood}(a.b) = max\ \frac{|A_i \cap B_i|}{|A_i \cup B_i|} \quad (5)$$

The above idea are combined into a single formula as shown in Equation (6)

$$Sim\ (a,b) - \begin{cases} 1, & if\ S_{synsets}\ (a,b) > 0 \\ max\{S_{neighborhoods}(a,b), S_{description}(a,b)\}, & if\ S_{synsets}\ (a,b) = 0 \end{cases} \quad (6)$$

Features based approach has tried to overcome the limitation of structure based approach regarding the fact that taxonomical links links in an ontology do not necessarily represent uniform distances [37]. However, this approach also has its disadvantages where it depends too much on the information provided. Table 2 below describes briefly the pros and cons

of each method in features based approach.

Table 2. Method features based for multiple ontology

Method	Advantage	Disadvantage	References
Tversky [31]	• Can generate a similarity value based on not only common but also distinct features of terms • Objects are represented as collections of features	• This feature model allows the representation of ordinal and cardinal features, but the similarity measure does not account for their ordering • Rely on information that is available in ontologies	[31],[27], [21]
Rodriguez and Egenhofer [32]	• Take into account semantic neighbourhoods in calculation of similarity	• Incomplete part for calculation will cause low accuracy • Parameter γ takes into account the depth of the terms in the two ontologies	[32],[33],[39],[21]
X-Similarity [33]	• Does not depend on weighting parameter • The maximum similarity provided by each feature alone is taken	• The contribution of other features is omitted if only the maximum value is taken at each time	[33],[21]

4. Conclusion

This paper describes the basics of semantic similarity measure and a brief introduction about the importance of the use of semantic similarity in various fields. Besides that, this paper also describes the classification of single and multiple ontology-based similarity measure. Advantages and disadvantages every approach are also described which may assist in the evaluation of the selection of the best approaches for multiple ontology. We also describe in more detail about method in the features based approach which is believed to be the most appropriate approach used to find the similarity between terms in multiple ontology. Feature based approach of potential in increasing efficiency and accuracy similarity between multiple ontology without using structure information. In future works, we want to study how different domain ontology are integrated using features based similarity approach as mechanism to comparing objects.

Acknowledgments. We are grateful to Dr Shahreen Kasim as our supervisor for her constructive comments, to Dr Mohd Farhan Md Fudzee as our second supervisor for his guidance throughout this research. We are also grateful to Gates IT Solution Sdn Bhd and Research Acculturation Grant Scheme (RAGS) vot no. R001 from Malaysian Ministry of Education for giving us the opportunity and confidence to do this project.

5. References

1. A. Doan, J. Madhavan, P. Domingos, and A. Halevy .: Ontology matching: A machine learning approach. In: Handbook on ontologies., pp. 1–20, (2004)
2. S. Elavarasi, J. Akilandeswari, and K. Menaga.: A Survey on Semantic Similarity Measure. In: International Journal of Research in Advent Technology, vol. 2, no. 3, pp. 389–398, (2014)
3. G. Pirró and J. Euzenat.: A feature and information theoretic framework for semantic similarity and relatedness. In: Semant. Web–ISWC 2010, (2010).
4. T. Deselaers and V. Ferrari.: Visual and semantic similarity in ImageNet. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1777–1784, (2011)
5. M. Batet, D. Sánchez, A. Valls, and K. Gibert.: Semantic similarity estimation from multiple ontologies. In: Applied Intelligence. vol. 38, no. 1, pp. 29–44, May (2012)
6. D. Sánchez and D. Isern.: Automatic extraction of acronym definitions from theWeb. In: Applied Intelligence., vol. 34, pp. 311–327, (2011)
7. D. Sánchez, D. Isern, and M. Millan.: Content annotation for the semantic web: An automatic web-based approach. In: Knowledge Information System, vol. 27, pp. 393–418, (2011)
8. L. Iannone, I. Palmisano, and N. Fanizzi.: An algorithm based on counterfactuals for concept learning in the Semantic Web. In: Applied Intelligence., vol. 26, pp. 139–159, (2007)
9. A. Budanitsky and G. Hirst.: Evaluating WordNet-based Measures of Lexical Semantic Relatedness. In: Computational Linguistics, vol. 32. pp. 13–47, (2006)
10. Al-Mubaid and H. Nguyen.: Measuring semantic similarity between biomedical concepts within multiple ontologies. In: IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews vol. 39, no. 4, pp. 389–398, (2009)
11. K. Saruladha, G. Aghila, and A. Bhuvaneswary.: COSS : Cross Ontology Semantic Similarity Measure-An Information Content Based Approach. pp. 485–490, (2011)
12. Hliaoutakis, G. Varelas, E. Voutsakis, E. G. M. Petrakis, and E. Milios.: Information Retrieval by Semantic Similarity. In: International Journal on Semantic Web and Information Systems., vol. 2, pp. 55–73, (2006)
13. G. Pirrò, M. Ruffolo, and D. Talia.: SECCO: On building semantic links in peer-to-peer networks. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 5480 LNCS, pp. 1–36, (2009)
14. J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C.-F. Chen.: A new method to measure the semantic similarity of GO terms. In: Bioinformatics, vol. 23, pp. 1274–1281, (2007)
15. C. Pesquita, D. Faria, A. O. Falcão, P. Lord, and F. M. Couto.: Semantic similarity in biomedical ontologies. In: PLoS Computational Biology, vol. 5, (2009)
16. R. Studer, V. R. Benjamins, and D. Fensel.: Knowledge engineering: Principles and methods. In: Data & Knowledge Engineering, vol. 25. pp. 161–197, (1998)
17. I. Spasic, S. Ananiadou, J. McNaught, and A. Kumar.: Text mining and ontologies in biomedicine: Making sense of raw text. In: Brief. Bioinformation., vol. 6, pp. 239–251, (2005)
18. A. Hliaoutakis.: Semantic Similarity Measures in MeSH Ontology and their application to Information Retrieval on Medline. In: Interface. pp. 1–79, (2005)
19. G. A. Miller.: WordNet: a lexical database for English. In: Communications of the ACM, vol. 38. pp. 39–41, (1995)
20. N. Choi, I.-Y. Song, and H. Han.: A Survey on Ontology Mapping, vol. 35, no. 3, pp. 34–41, (2006)
21. D. Sánchez, M. Batet, D. Isern, and A. Valls.: Ontology-based semantic similarity: A new feature-based approach. In: Expert Systems with Applications., vol. 39, no. 9, pp. 7718–7728, Jul. (2012)
22. R. Rada, H. Mili, E. Bicknell, and M. Blettner.: Development and application of a metric on semantic nets. In: IEEE Transactions on Systems, Man, and Cybernetics., vol. 19, no. 1, pp. 17–30, (1989)
23. H. Bulskov, R. Knappe, and T. Andreasen.: On measuring similarity for conceptual querying. In: Flex. Query Answering Syst., pp. 100–111, (2002)
24. M. Palmer and Zhibiao Wu.: VERB SEMANTICS AND LEXICAL. In: Proceeding ACL '94 Proc. 32nd Annu. Meet. Assoc. Comput. Linguist., pp. 133–138, (1994)
25. M. Sussna.: Word sense disambiguation using a massive of computer for free-text semantic indexing network. In: CIKM '93 Proc. Second Int. Conf. Inf. Knowl. Manag., pp. 67–74, (1993)
26. C. Leacock and M. Chodorow.: Combining Local Context and WordNet Similarity for Word Sense Identification. In: WordNet: An electronic lexical database., pp. 265–283, (1998)
27. R. Jiang.: From ontology to semantic similarity: calculation of ontology-based semantic similarity. In: Sci. World J., vol. 2013, p. 793091, Jan. (2013)
28. P. Resnik.: Using information content to evaluate seantic similarity in a taxonomy. In : Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), (1995)
29. D. Lin.: An Information-Theoretic Definition of Similarity. In: Proceedings of ICML, pp. 296–304, (1998)
30. J. Jiang and D. Conrath.: Semantic similarity based on corpus statistics and lexical taxonomy. In: arXiv Prepr. C., no. Rocling X, (1997)
31. A. Tversky.: Features of similarity. In: Psychological Review, vol. 84. pp. 327–352, (1977)
32. M. Rodríguez and M. Egenhofer.: Determining semantic similarity among entity classes from different ontologies. In: IEEE Transactions on Knowledge and Data Engineering vol. 15, no. 2, pp. 442–456, (2003)
33. E. Petrakis, G. Varelas, A. Hliaoutakis, and P. Raftopoulou.: X-Similarity: Computing Semantic Similarity between Concepts from Different Ontologies. In: J. Digit. Inf. Manag., vol. 4, no. 4, p. 233, (2006)
34. V. Schickel-Zuber and B. Faltings.: OSS: A semantic similarity function based on hierarchical ontologies. In: IJCAI International Joint Conference on Artificial Intelligence, pp. 551–556, (2007)
35. Y. Li, D. McLean, Z. A. Bandar, J. D. O'Shea, and K. Crockett.: Sentence similarity based on semantic nets and corpus statistics. In: IEEE Trans. Knowl. Data Eng., vol. 18, pp. 1138–1150, (2006)
36. D. Lin.: An Information-Theoretic Definition of Similarity. In: Proc. ICML, pp. 296–304, (1998)
37. D. Sanchez and M. Batet.: A semantic similarity method based on information content exploiting multiple ontologies. In: Expert Syst. Appl., vol. 40, no. 4, pp. 1393–1399, Mar. (2013)
38. D. M. Yuhua Li and Zuhair A. Bandar.: An approach for measuring semantic similarity between words using multiple information sources. In: IEEE Trans. Knowl. Data Eng., vol. 15, pp. 871–882, (2003)
39. H. Li, Y. Tian, and Q. Cai.: Improvement of semantic similarity algorithm based on WordNet. In : Proceedings of the 2011 6th IEEE Conference on Industrial Electronics and Applications, ICIEA 2011, pp. 564–567, (2011)