SAIS 2019 Proceedings

Southern (SAIS)

3-22-2019

Chitchat on the Nile about Using Similarity Measures to Evaluate Biomedical Ontology Records

Osama Rabie
King Abdulaziz University, obrabie@kau.edu.sa

CHITCHAT ON THE NILE ABOUT USING SIMILARITY MEASURES TO EVALUATE BIOMEDICAL ONTOLOGY RECORDS

Osama Bassam J. Rabie
King Abdulaziz University
obrabie@kau.edu.sa

ABSTRACT

This paper compares five binary similarity and distance measures that researches use in biomedical research. The measures are applied on seven ontology records to evaluate the usage accuracy of relationships *part_of* and *located_in*. Their use and definition in comparison to the definition found on the online medical dictionary provided by the U.S. National Library of Medicine. The practical contribution includes finding which similarity measures works better with biomedical ontologies among similarity measures that researchers use in biomedical field. Future work can use the similarity measure to fix the definition and use of *part_of* and *located_in* relationships. The premise is that aligning the definition and use of biomedical terminologies with accurate definition and use will improve the efficiency of the biomedical ontologies. This paper focuses on the use of the relationships *part_of* and *located_in*.

Keywords

Biomedical Ontology, Similarity Measure, Operations Management

INTRODUCTION

The main motive behind this work is to contribute in resolving the issues facing ontologies. The amount of data exchanged on the internet is huge and without a mechanism to process and automate, e.g., using ontologies, humans will not be able to process that amount of data. The data on the internet can supply humans with useful information. Web 3.0 is enabling the web to process itself and provide us with useful information by using ontologies (Berners-Lee et al., 2001).

Measuring the inaccuracy of relationships used in biomedical ontologies contributes into achieving the semantic web vision. Several scholars show it is impossible to manage something without measuring it. This study evaluates the similarity (or distance) levels between two relationships in the biomedical ontologies: *part_of* and *located_in*. The evaluation uses MedlinePlus, which is an online medical dictionary offered by the U.S. National Library of Medicine, National Institutes of Health and Merriam-Webster, Incorporated.

The importance of biomedical ontologies is encouraging to draft this paper. Ontologies help solving the challenge of big data automation. They enhance semantic information processing and knowledge exchange between machines. The biomedical ontologies are ontologies used in the field of bioinformatics (Maojo et al., 2006). As ontologies, the biomedical ontologies also deal with heterogeneous clinical data (Sugumaran and Storey, 2002). Two factors suggest the usage of biomedical ontologies; first, the biomedical ontologies are considered the most successful and practical innovation in the field of biomedicine (Maojo et al., 2006), second, the biomedical ontologies considered the most essential and advanced application of the semantic web (Domingue et al., 2011).

The work of this paper is with the relationships part_of and *located_in*. The relationships part_of and its inverse *has_part* are the second most important structural relationships after the relationship *is_a* (Schulz et al., 2006). However, the relationship *part_of* can be confused with the relationship *located_in* (Schulz et al., 2006). In this paper, five similarity and distance measures evaluate the usage accuracy of the relationship *part-of* at seven biomedical ontology records. The study contributes in proving the inconsistency of used biomedical ontology by evaluating the similarity of the used relationship and the dictionary definition.

Ontology

“An ontology is a "domain-specific dictionary." It captures the semantic meaning and relationship of terms which allows for further usage of the term’s concept” (Rabie and Norcio, 2013, pp01). In addition, it has terms and their defined properties that can be executed by a computer. Ontologies are created by anyone with suitable tools. Ontology is a formal conceptual representation of domain concepts. "Ontology is a representation of universals; it describes what is general reality, not what is particular" (Maojo et al., 2011). Databases are entities instances where ontologies represent entities classes. "Ontology provides engineers with the semantics of data which can be used with problem-solving methods along with reasoning services to produce a great system with fewer resources (Schulz et al., 2006)" (Rabie and Norcio, 2013). Based on the graph structure of the ontology, having one inaccurate instance means the rest of the information obtained is going to be inaccurate. "Semantics enable machines to understand context and the type of data they utilize" (Rabie and Norcio, 2013). Semantics give machines the ability to become closer to human understanding level (Benjamins et al., 2005). Machines should be able to manage knowledge independently by using semantics (Cayzer, 2004).

"Currently, vast amounts of data require transmission and manipulation, and without sharing it, the data may become useless" (Rabie and Norcio, 2013). Huge amount of data prevents people from being able to manually do transmission, interpretation, and manipulation. Semantic web (web 3.0) is meant to address those challenges (Berners-Lee et al., 2001; Lozano-Tello and Gómez-Pérez, 2004). On the other hand, handling a large amount of heterogeneous data can be an integration challenging. "Ontology can solve this problem by providing computers with an understanding of information and using it for reasoning without interference from humans" (Rabie and Norcio, 2013). However, using ontologies can be challenging given that ontologies languages are still under development (Lozano-Tello and Gómez-Pérez, 2004). In addition, ontologies including ontologies in use are corrupted, and finding an ontology to produce inference may be a challenge. Ontologies description is loos compared with databases (Baader et al., 2003) despite the effectiveness of using ontologies to process and exchange knowledge compared to the databases.

Ontology are important part of artificial intelligence implementation (Rabie and Norcio, 2013). Software agents understand the conceptual representation of a specific domain knowledge embedded in ontologies. "Ontology provides information management systems with a way to handle unstructured contents, which may be impossible for computers to handle without ontologies" (Rabie and Norcio, 2013, pp02). They supply knowledge description, natural language processing, and a reference for standardizing language modeling. Given the nature of ontology development, ontologies may co-exist with other ontologies for the same domain. Ontology creators do not have to check with domain experts, which is a reason for having imprecise knowledge represented by those ontologies.

PAPER MAP

The binary similarity and distance measures used in the experiment are next. A brief description of the ontology used in the experiment at the section after. This paper includes data analysis. Finally, the recommendations and the ranking of the measures used.

SIMILARITY AND DISTANCE MEASURES USED

Binary Similarity Measures and Binary Distance Measures are measures used to determine the level of similarity between two (i.e. binary) patterns, or the level of dissimilarity or distance between the patterns (Cassisi et al., 2012; Choi et al., 2010). To apply the measures data should be converted into pairs of time series. In each pair, one time series stands for the item to be compared where the other time series will represent the other item (Cassisi et al., 2012). In addition, both time series should have the same length (Cassisi et al., 2012). There are applications for these measures, enhancing clustering (Basu et al., 2003), enhancing image-processing (Willett, 2003a), and enhancing the detection of tumor (Fei et al., 2003). Moreover, fields like ecology (Jackson et al., 1989b), biometrics (Willett, 2003b), and handwriting recognition (Cha et al., 2003), are using the measures to help them deciding on the level of similarity between two patterns.

In this paper, the measures evaluate the accuracy of part_of usage in seven randomly selected biomedical ontology records. The rank of the binary similarity and distance measures is based on how close their similarity score to the average score of the ontology against medical dictionary (see table 1). Although the data can be represented in a variety of ways (Clifford and Stephenson, 1975), in this study the binary data is the type used to be able to process more records in future publications.

The measures picked based on their potential importance in earlier biomedical related studies. Being used in earlier studies makes this study the first one to study the biomedical ontology by using measures used in biomedical studies. The fact that those measures were used in the field makes it more justifiable to use in studying the biomedical ontology. The following are the measures used and a brief justification for each measure of its candidacy to be considered in the field of biomedical and this study:

Jaccard (Jaccard, 1901)

This similarity measure is considered the first to be used in a biomedical related field, ecology (Jaccard, 1901), and started the idea of using similarity measures in different fields (Choi et al., 2010). Jaccard is used in several ecological species classification (Cayzer, 2004; Jaccard, 1901; Jackson et al., 1989b) and in several biomedical ontology evaluations (Sánchez and Batet, 2011). In addition, no large data is needed for it to work properly (Noor Aznimah Abdul et al., 2010). Therefore, Jaccard is included in our experiment.

Ochiai (Ochiai, 1957a)

Also known as Ochiai-I (Choi et al., 2010) where few literatures consider it the same as the Cosine similarity (Sánchez and Batet, 2011; Willett, 2003b). However, we consider Ochiai-I and Cosine similarity to be two different measures. Ochiai is a similarity measure used in many aspects related to the biomedical field especially in the field of biology (Jackson et al., 1989a; Jongman, 1995; Ochiai, 1957b; Sánchez and Batet, 2011; Vavilova and William Jr, 1998; Willett, 2003b). Therefore, we included Ochiai-I in our experiment.

Forbes (Forbes, 1907)

Also known as ForbesI (Choi et al., 2010) was proposed to help in spices classification (Forbes, 1925; Forbes, 1907). Therefore, ForbesI is on our list.

Simpson (Simpson, 1960)

It was proposed to study classification of faunal (Simpson, 1960). In addition, the measurement was used to estimate the semantic similarity in the biomedical domain (Sánchez and Batet, 2011). It was used in studying the similarity of molecular fingerprints (Willett, 2003a). Therefore, Simpson is on the similarity measures list.

Yule (Yule, 1903)

Also known as YuleQ (Choi et al., 2010). Yule was used in species classification (Jackson et al., 1989a). In addition, it was used in studying the similarity of molecular fingerprints (Willett, 2003a). Therefore, it is on the list.

BIOMEDICAL ONTOLOGIES USED

The ontology used is called Foundation Model of Anatomy which is developed by Unified Medical Language System (UMLS). The ontology created by Unified Medical Language System (UMLS). The records picked from the Foundational Model of Anatomy (FMA) view of neuroanatomy ontology. The ontology version is 3.0 and the view upload date is Sep 18, 2009. The ontology category is anatomy.

There are three reasons for choosing this ontology. First, the projects using this ontology include:

• Electrophysiology Ontology (for The Johns Hopkins University): Per NIH, electrophysiology ontology supports the automation of representation and meta-analysis of electrophysiology data. The electrophysiology is a growing science and sciences need ontologies to support their research thus fixing their ontology is a priority;

• Neural ElectroMagnetic Ontologies (NEMO) (for University of Oregon and Georgia State University): Per the NIH, NEMO is EEG and MEG ontology that supports their representation, classification, and meta-analysis of brain electromagnetic data. The use of EEG and MEG is increasing including the use of their ontologies and ontology tools thus fixing their ontologies is a priority.

Second, Unified Medical Language System (UMLS) develop it. UMLS was made by U.S. National Library of Medicine (NLM) and still maintained by the NLM. According to NLM, NLM is “the world's largest medical library.” The UMLS project started beck in the 1986 by Donald A. B. Lindberg, M.D. According to NLM, UMLS is made to provide “health and biomedical vocabularies and standards” to be used and exchanged by computer systems.

Finally, the ontology is publicly available. From this ontology seven records were randomly picked by running through the ontology records and pick the ones with related relationships. The records were randomly picked from different classes in the ontology view.

RESULTS AND DISCUSSION

“In the “post-genomic era”, biomedical ontologies are becoming increasingly popular in the computational biology community as the focus of biology has started to shift from mapping genomes to analyzing the vast amount of information resulting from functional genomics research,” (Bodenreider et al., 2005, pp76). The use of ontologies is ever increasing, and biomedical ontologies should have more accuracy than less risky fields. The improvement of biomedical ontologies can save lives and resources via allowing better automation. The improvement of the biomedical ontology accuracy helps ontology users getting better and more accurate results. This paper exams which similarity measure among similarity measures researchers use in biomedical applications works better with biomedical ontologies. A follow up research will use the similarity measure to analyze the biomedical ontologies definition and use of the relationships part_of and located_in.

All the chosen records from the ontology can be considered to be ones of the part_of relation records. We wanted to make this study more comprehensive by not just indicating if the relationship is similar or not, but also measure its similarity. Therefore, if a relationship is accurate (i.e. part_of), its weight is 1, 100% similar. If the relationship can be part_of or located_in, its weight is 0.5, 50% similar. If the relationship should be located_in, the records weight is 0 to indicate it should be the other relationship. Finally, if the relationship cannot be part_of or located_in, the weight is -1. The weight is -1 to indicate it is a crucial error. Please remember that the relationships part_of and located_in are confused in many cases; therefore, it can be considered less of an error compared with not needing to have either of the relations.

The results from the medical dictionary are shown in Appendix B. The dictionary average is calculated as 0.5+(-1)+0+1+(-1)+(-1)+(-1) / 7 = -0.21429 = -21.429%. The -21.429% is considered the similarity resulted from the medical dictionary. Being from the U.S. National Library of Medicine, National Institutes of Health, and Merriam-Webster, Incorporated we consider the results from MedlinePlus to be the accurate relationship and will be used as the reference. On the other hand, all the records picked from the FMA view of neuroanatomy ontology are considered part_of; see appendix c. Therefore, the average is 1+1+1+1+1+1+1 / 7 = 1 = 100%.

The similarity is measured between the results from the medical dictionary (A) and ontology's records (B) by using the similarity measures discussed in section 2.

Jaccard = J (A, B) = |A⋂B| / |A⋃B | [23] = 0.666+0+0.333+1+0+0+0 / 14 = 0.14279 = 14.279%. In order to consider the weight of the relationships we weighted the relationships as follows:

The relationship of the value 1 = 1.

The relationship of the value 0.5 = 0.666.

The relationship of the value 0 = 0.333.

The relationship of the value -1 = 0.

Ochiai = A.B / √(Σ_(i=1)^n∑A_i .Σ_(i=1)^n∑B_i) [16] = (1X0.5)+(1X1)+(1X0)+(1X1)+(1X-1)+(1X-1)+(1X-1) / √(0.5+1+0+1+1+1+1) X √7= -2.5 / 6.205 = -0.4029 = -40.29%.

For Forbes, Simpson, and Yule measures:

a = Number (A⋃B). The value of a is calculated by adding all readings together, from the medical dictionary and the ontology, while considering the weight the reading. Therefore, a = 0.5+(-1)+0+1+(-1)+(-1)+(-1) + 1+1+1+1+1+1+1 = 4.5

b = Number (A-B). The value of b is calculated by adding the readings that exist in A, the reading does not equal 0, but does not exist in B. To consider the weight of the readings we calculated as follows:

b = 0+(-2)+0+0+(-2)+(-2)+(-2) = -8. The first record is A=0.5 and B=1 which indicates that the whole value of A exists in B so, it should be 0. Now, for the second, fifth, sixth, and seventh records, A=-1 and B=1, we are measuring the existence of -1 in 1. The distance between -1 and 1 is 2. Since we move from -1 we add a negative sign, -2. Finally, for the third (A=0 and B=1) and fourth (A=1 and B=1) records we check the existence in A and the inexistence in B in a straight forward manner

c = Number (B-A). The value of c is calculated by adding the readings that exist in B, the reading does not equal 0, but does not exist in A. To consider the weight of the readings we calculated as follows:

c = 0.5+2+1+0+2+2+2 = 9.5. The first record, A=0.5 and B=1, half of B existence in both A and B and the other half does exit in B but not in A. For the second, fifth, sixth, and seventh records, A=-1 and b=1, the distance between -1 and 1 is 2 and since we are traveling from positive, the result is positive, 2. Finally, the third (A=0 and B=1) and fourth (A=1 and B=1) records are calculated in a straight forward manner

d = Number (-A-B). The value of d is calculated by adding the number of elements that does not exist in either A or B. Since all elements exist in B, the results is d=0

n = a+b+c+d. Therefore, n=4.5+(-8)+9.5+0

Forbes = na / (a+b) (a+c) [8] = 6X4.5 / [4.5+(-8)](4.5+9.5) = 27 / -49 = -0.55102 = - 55.102%

Simpson = a / Min(a+b, a+c) [8] = 4.5 / Min(4.5+0, 4.5+9.5) = 4.5 / 4.5 = 1 = 100%

Yule = ad-bc / ad+bc [8] = 4.5X0 – (-8) X 9.5 / 4.5X0 + (-8)9.5 = 76 / -76 = -1 = -100%

CONCLUDING REMARKS

In this paper, we are contributing by studying biomedical ontology inconsistency using similarity measures. The ontology records were taken from an ontology created by Unified Medical Language System (UMLS). The records picked from the Foundational Model of Anatomy (FMA) view of neuroanatomy ontology. The similarity between the results from the medical dictionary and the ontology records are calculated. In addition, the distance between the results from the measures and the results from the medical dictionary is computed, Table 1.

Measure	**Jaccard**	**Ochiai**	**Forbes**	**Simpson**	**Yule**
Similarity	14.27%	-40.29%	-55.102%	100%	-100%
Distance	21.429 + 14.279 = 35.708	40.29 – 21.429 = 18.861	55.102 – 21.429 = 33.673	100 + 21.429 = 121.429	100 – 21.429 = 78.571

Table 1. Similarity Measures, Ontology Average = -21.429%

Another contribution of the paper is ranking similarity measures the biomedical field uses; Ochiai, Forbes, Jaccard, Yule, and then Simpson.

ACKNOWLEDGMENTS

The author thanks Prof. Anthony F. Norcio, Professor of Information Systems, Department of Information Systems, University of Maryland Baltimore County (UMBC) for his support.

REFERENCES

1. Baader, F., Calvanese, D., McGuinness, D., Patel-Schneider, P., and Nardi, D. (2003) *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University press.
2. Basu, S., Bilenko, M., and Mooney, R. J. (2003) Comparing and Unifying Search-Based and Similarity-Based Approaches to Semi-Supervised Clustering, *Proceedings of the ICML-2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining*: Citeseer, pp. 42-49.
3. Benjamins, V. R., Casanovas, P., Breuker, J., and Gangemi, A. (2005) Law and the Semantic Web, an Introduction, in *Law and the Semantic Web*. Springer, pp. 1-17.
4. Berners-Lee, T., Hendler, J., and Lassila, O. (2001) The Semantic Web, *Scientific American* (284:5), pp. 28-37.
5. Bodenreider, O., Mitchell, J. A., and McCray, A. T. (2005) Biomedical Ontologies, *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*), pp. 76-78.
6. Cassisi, C., Montalto, P., Aliotta, M., Cannata, A., and Pulvirenti, A. (2012) *Similarity Measures and Dimensionality Reduction Techniques for Time Series Data Mining*. InTech.
7. Cayzer, S. (2004) Semantic Blogging and Decentralized Knowledge Management, *Communications of the ACM* (47:12), pp. 47-52.
8. Cha, S.-H., Tappert, C. C., and Srihari, S. N. (2003) Optimizing Binary Feature Vector Similarity Measure Using Genetic Algorithm and Handwritten Character Recognition, *ICDAR*: Citeseer, pp. 662-665.
9. Choi, S.-S., Cha, S.-H., and Tappert, C. C. (2010) A Survey of Binary Similarity and Distance Measures, *Journal of Systemics, Cybernetics and Informatics* (8:1), pp. 43-48.
10. Clifford, H. T., and Stephenson, W. (1975) *An Introduction to Numerical Classification*. Academic Press.
11. Domingue, J., Fensel, D., and Hendler, J. A. (2011) *Handbook of Semantic Web Technologies*. Springer.
12. Fei, B., Lee, Z., Duerk, J. L., and Wilson, D. L. (2003) Image Registration for Interventional Mri Guided Procedures: Interpolation Methods, Similarity Measurements, and Applications to the Prostate, *International Workshop on Biomedical Image Registration*: Springer, pp. 321-329.
13. Forbes, S. (1925) Method of Determining and Measuring the Associative Relations of Species, *Science* (61:1585), pp. 518-524.
14. Forbes, S. A. (1907) *On the Local Distribution of Certain Illinois Fishes: An Essay in Statistical Ecology*. Illinois State Laboratory of Natural History.
15. Jaccard, P. (1901) Étude Comparative De La Distribution Florale Dans Une Portion Des Alpes Et Des Jura, *Bull Soc Vaudoise Sci Nat* (37), pp. 547-579.
16. Jackson, D. A., Somers, K. M., and Harvey, H. H. (1989a) Similarity Coefficients: Measures of Co-Occurrence and Association or Simply Measures of Occurrence?, *The American Naturalist* (133:3), pp. 436-453.

17. Jackson, D. A., Somers, K. M., and Harvey, H. H. (1989b) Similarity Coefficients: Measures of Co-Occurrence and Association or Simply Measures of Occurrence?, *American Naturalist*), pp. 436-453.
18. Jongman, E. (1995) *Data Analysis in Community and Landscape Ecology*. Cambridge university press.
19. Lozano-Tello, A., and Gómez-Pérez, A. (2004) Ontometric: A Method to Choose the Appropriate Ontology, *Journal of Database Management* (2:15), pp. 1-18.
20. Maojo, V., Crespo, J., García-Remesal, M., De la Iglesia, D., Perez-Rey, D., and Kulikowski, C. (2011) Biomedical Ontologies: Toward Scientific Debate, *Methods of information in medicine* (50:03), pp. 203-216.
21. Maojo, V., García-Remesal, M., Billhardt, H., Alonso-Calvo, R., Pérez-Rey, D., and Martín-Sánchez, F. (2006) Designing New Methodologies for Integrating Biomedical Information in Clinical Trials, *Methods of information in medicine* (45:2).
22. Noor Aznimah Abdul, A., Siti Salwa, S., Mohamad, D., and Omar, M. (2010) Investigating Jaccard Distance Similarity Measurement Constriction on Handwritten Pen-Based Input Digit, *International Conference on Science and Social Research (CSSR)*: IEEE, pp. 1181 - 1185.
23. Ochiai, A. (1957a) Zoogeographic Studies on the Soleoid Fishes Found in Japan and Its Neighbouring Regions, *Bull. Jpn. Soc. Sci. Fish* (22:9), pp. 526-530.
24. Ochiai, A. (1957b) Zoogeographic Studies on the Soleoid Fishes Found in Japan and Its Neighbouring Regions, *Bulletin of Japanese Society of Scientific Fisheries* (22), pp. 526-530.
25. Rabie, O., and Norcio, A. F. (2013) Discussion of Some Challenges Concerning Biomedical Ontologies, in: *Human-Computer Interaction. Applications and Services*. Las Vegas, USA: Springer, pp. 173-180.
26. Schulz, S., Kumar, A., and Bittner, T. (2006) Biomedical Ontologies: What Part-of Is and Isn't, *Journal of Biomedical Informatics* (39:3), pp. 350-361.
27. Simpson, G. G. (1960) Notes on the Measurement of Faunal Resemblance, *American Journal of Science* (258:2), pp. 300-311.
28. Sugumaran, V., and Storey, V. C. (2002) Ontologies for Conceptual Modeling: Their Creation, Use, and Management, *Data & knowledge engineering* (42:3), pp. 251-271.
29. Sánchez, D., and Batet, M. (2011) Semantic Similarity Estimation in the Biomedical Domain: An Ontology-Based Information-Theoretic Perspective, *Journal of biomedical informatics* (44:5), pp. 749-759.
30. Vavilova, V. V., and William Jr, M. (1998) Temporal and Altitudinal Variations in the Attached Algae of Mountain Streams in Colorado, *Hydrobiologia* (390:1-3), pp. 99-106.
31. Willett, P. (2003a) Similarity-Based Approaches to Virtual Screening. Portland Press Limited.
32. Willett, P. (2003b) Structural Biology in Drug Metabolism and Drug Discovery, *Biochemical Society* (31), pp. 603-606.
33. Yule, G. U. (1903) Notes on the Theory of Association of Attributes in Statistics, *Biometrika* (2:2), pp. 121-134.