Contents List available at RAZI Publishing
Acta Electronica Malaysia (AEM)
Journal Homepage: http://www.actaelectronicamalaysia.com
https://doi.org/10.26480/aem.02.2017.01.04

ISSN: XXXX (Print)
ISSN: 2590-4043 (online)

COMPARISON SEMANTIC SIMILARITY APPROACH USING BIOMEDICAL DOMAIN DATASET

Shahreen Kasim[1], Nurul Aswa Omar[1], Nurul Suhaida Mohammad Akbar[1], Rohayanti Hassan[1], Marzanah A. Jabar[2]

[1]Department Web Technology, [2]Soft Computing And Data Mining Centre, Faculty Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Parit Raja, Batu Pahat, Johor Malaysia.
[2]Department Software Engineering and Information Systems, Faculty of Computer Science & Information Technology, Universiti Putra Malaysia.
**Corresponding author email: nurulaswa@uthm.edu.my, shahreen@uthm.edu.my, nuraishahkamari93@gmail.com*

ARTICLE DETAILS

Article History:

Received 3 July 2017
Accepted 3 October 2017
Available online 1 November 2017

Keywords:

Ontology, efficiency, scalability, MeSH datasets, Semantic.

ABSTRACT

Semantic similarity is defined as the closeness of two concepts, based on the likeliness of their meaning. Most semantic similarity is applied uses ontology. This research are uses ontology as a case study due to their efficiency, scalability, lack of constraints and the availability of large ontologies. Ontology-based semantic similarity is used in two situations. The semantic similarity in a single ontology and when multiple ontologies are involved. In this research focus on single ontology where use MeSH datasets as a case study. Semantic similarity single ontology means similarities are compared from the same ontology. The importance of information in the biomedical field, semantic similarity measures have been of great interest.

1. INTRODUCTION

The study of semantic and ontology finds its origins as early as 1989 [1]. There is a growing number of various domain ontologies that organise concepts into hierarchies and semantic networks. Many researchers believe that the use of ontology can be translated as knowledge that is commonly understandable [2]. Semantic similarity, semantic relatedness and semantic distance can use the ontology as a case study in both single or multiple ontologies. Research about semantic in a single ontology is more likely to use the WordNet ontology as a case study. Table 1 shows that a large number of researchers use WordNet as a case study of single ontology in almost all methods of semantic similarity, semantic relatedness and semantic distance. All methods in the use of a single ontology use two benchmark data from Miller and Charles (MC) and Rubenstain and Goodenough (RG) [3, 4].

However, the use of the biomedical domain as a case study is also used in the single ontology method. In fact, most researchers nowadays prefer the biomedical domain as a case study, especially in semantic similarity for multiple ontologies. Rada similarity devised a semantic distance measure based on semantic networks. They used MeSH as a case study, which consists of biomedical terms organised in a hierarchy [5]. Besides that, some researcher also used MeSH as a case study [6]. This researcher proposed the FaITH method and used MeSH to evaluate the investigation of how FaITH performs with domain related ontologies. Schickel & Faltings, used two types of ontologies to evaluate their proposed method through general purpose ontologies (WordNet) and specific domain ontologies (Gene ontology) [7].

The use of the biomedical domain as a case study is not widespread in single ontology, but has increased in multiple ontologies. Several approaches for determining semantic similarity have been proposed. Ontology-based semantic similarity can be classified into structure-based approach, information content-based approach, feature-based approach and hybrid-approach.

Table 1: Summary of datasets for single ontology used in previous work

Approach	Method	Semantic	Data source	Datasets
Structure-based	Rada *et al.*, (1989)	Distance	Ontology	MeSH
	Bulskov *et al.*, (2002)	Relatedness	Ontology	WordNet
	Sussna (1993)	Relatedness	Ontology	WordNet
	Palmer & Wu (1994)	Similarity	Ontology	WordNet
	Leacock & Chodorow (1998)	Similarity	Ontology	WordNet
Information content-based	Resnik (1995)	Similarity	Ontology + Corpus	WordNet
	Jiang & Conrath (1997)	Distance	Ontology + Corpus	WordNet
	Lin (1998a)	Similarity	Ontology + Corpus	WordNet
Feature-based	Tversky (1977)	Similarity	Ontology	WordNet
	Pirró & Euzenat (2010)	Relatedness	Ontology	MeSH
Hybrid-based	Li *et al.*, (2006)	Similarity	Ontology + Corpus	WordNet and Brown Corpus
	Schickel & Faltings (2007)	Similarity	Ontology	WordNet and Gene Ontology

There are several examples of biomedical domain ontologies available including: UMLS, MeSH, snomed-CT and gene ontology. The following section describes biomedical domain ontologies as follows:

1.1. UMLS

The Unified Medical Language System (UMLS) is a repository of biomedical vocabularies developed by the US National Library of Medicine. This ontology contains a very large, multi-purpose and multilingual meta-thesaurus containing information about biomedical and health related concepts. It is built from the electronic versions of a few different thesauri, code sets, classifications, and lists of controlled terms [8]. The UMLS contains information about over 1 million biomedical

concepts and 5 million concept names from more than 100 incorporated controlled vocabularies and classifications (some in multiple languages) systems. Vocabularies integrated in the UMLS meta-thesaurus include the National Center for Biotechnology Information (NCBI) taxonomy, Gene Ontology, the Medical Subject Headings (MeSH), Online Mendelian Inheritance in Man (OMIM) and the Digital Anatomist Symbolic Knowledge Base [9]. Each concept in the meta-thesaurus is assigned to at least one semantic type (a category). Certain semantic relationships can be identified between members of the various semantic types. The UMLS can browsed using https://www.nlm.nih.gov/researc h/umls/. Figure 1 shows a snapshot of the UMLS web page. The UMLS has three tools, which are known as the knowledge sources:

(i) Meta-thesaurus: Terms and codes from many vocabularies including *Current Procedural Terminology* (CPT), International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM), Logical Observation Identifiers Names and Codes (LOINC), MeSH and Snomed-CT

(ii) Semantic Network: Broad categories (semantic types) and their relationships (semantic relations)

(iii) Specialist Lexicon and Lexical Tools: Natural language processing tools

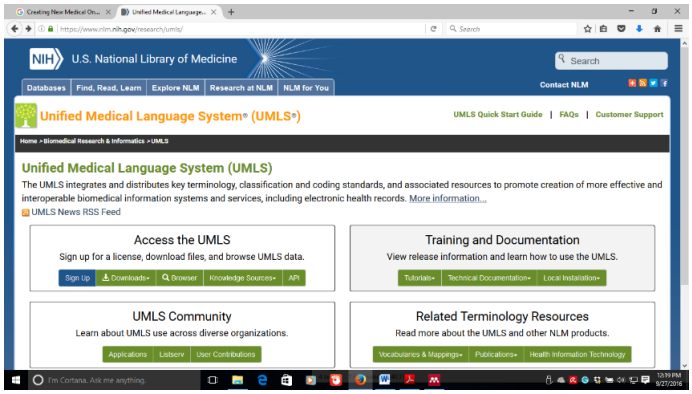

Figure 1: Snapshot of the UMLS web page.

1.2. MeSH

Medical Subject Heading (MeSH) is a controlled vocabulary and a thesaurus developed by the U.S. National Library of Medicine (NLM). This ontology focuses on indexing clinical documents through more than 22,000 medical concepts including 16 basic categories [10]. In MeSH, a concept may appear in more than one taxonomy. MeSH has several properties such as the MeSH heading (MH), scope note and entry term which is a synonym concept to MH. Besides that, MeSH's tree number is one important property that indicates the positions of the concept. This property can identify the hypernym for each concept in MeSH. Figure 2 shows an example of a MeSH web page and Figure 3 depicts an example of content in MeSH.

Figure 2: The snapshot of Medical Subject Headings (MeSH)

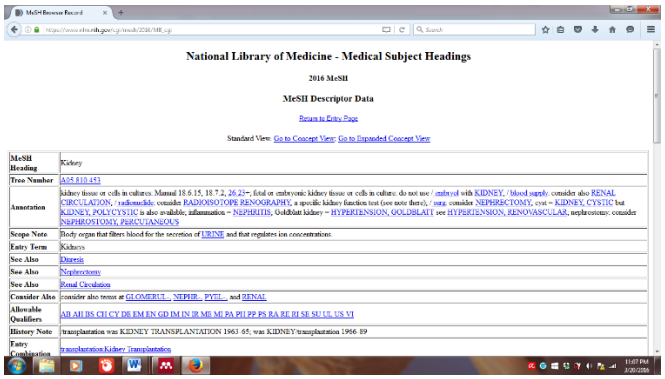

Figure 3: The snapshot of MeSH content for kidney disease

1.3. Snomed-CT

Snomed-CT stands for Systemized Nomenclature of Medicine Clinical Term, which was included in the UMLS in May 2004 [11-12]. Snomed-CT is a comprehensive clinical ontology maintained by the International Health Terminology Standards Development Organization (IHTSDO) [13]. These ontologies are some of the largest sources in the Unified Medical Language System (UMLS). Snomed-CT is used for indexing electronic medical records, ICU monitoring, clinical decision support, medical research studies, clinical trials, computerised physician order entry, disease surveillance, image indexing and consumer health information services.

This ontology contains more than 311,000 concepts with unique meanings. It also has formal logic-based definitions organized into hierarchies which include clinical findings, procedures, observable entities, body structures, organisms, substances, pharmaceutical products, specimens, physical forces, physical objects, events, geographical environments, social contexts, linkage concepts, qualifier values, special concepts, record artefacts, and staging and scales. The concepts of Snomed-CT link with 1.36 million relationships [14]. The Snomed-CT ontologies can be downloaded from https://www.nlm.nih.gov/research/umls/licensedcontent/snomedctfiles.html. This ontology can be downloaded via three types of Snomed-CT international release files. Figure 4 shows an example of the Snomed-CT web page.

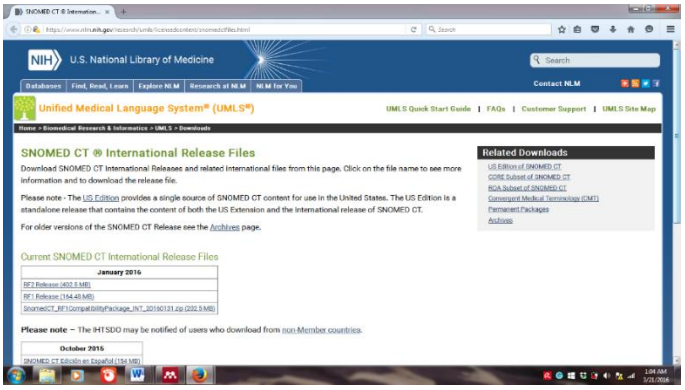

Figure 4: The snapshot of Snomed-CT web pages

1.4. Gene ontology (GO)

Gene ontology (GO) describes gene proteins and all concerns of organisms as a structured network of defined terms. The GO is developed based on a project utilising collaborative effort to address the need for consistent descriptions of gene products in different databases. The GO includes three function ontologies that describe gene products in terms of their cellular components, biological processes, and molecular functions in a species-independent manner [2]. The molecular function supplies information on the role played by a gene product. The biological process refers to a biological objective to which a gene product contributes. The cellular component represents the cellular localisation of the gene product, including cellular structures and complexes [15]. The GO ontology is structured as a directed acyclic graph where each term defines relationships to one or more other terms in the same domain, and sometimes to other domains. The GO vocabulary is designed to be species-

agnostic, and includes terms applicable to prokaryotes and eukaryotes, and single and multicellular organisms. The GO can be browsed at http://geneontology.org/index.html.

2. METHODOLOGY

A methodology is a set of ideas or guidelines about how to proceed in gathering and validating knowledge of a subject matter. To ensure the effectiveness of the system in the future, all aspects should be emphasized. If inappropriate methodology is used or if appropriate methodology is used poorly, the result of study could be misleading. Methodology may include a few aspects which are publication research, interviews, surveys and other research technique. The method is use to achieve the objective of the project that will complete a perfect result. This research is focusing on Rodriguez and Egenhofer method [16]. This research was created by using MeSH dataset. This research has four phases. Firstly phase is focusing on data preparation. Secondly phase is similarity measure. Thirdly phase is feature based process and for the last phases is analysis result.

2.1. Data Preparation

The dataset that have been chosen in this research is MeSH as a sample for biomedical domain. The dataset of MeSH consist of MeSH Heading, Tree Number, Concept UI, Unique ID, Semantic Type and others. All of the names and data are the sources from MeSH Browser. Due to purpose of this research is to find similarity in single ontology, the term comparison that used in this research is come from MeSH dataset. The data used are twenty term to make a database and find the similar data for each.

Dataset of MeSH consisted of term, synonym, concepts and others. This research had used benchmark from MeSH and only use Rodriguez Egenhofer method to calculate similarity [16]. Figure 5 show MeSH dataset.

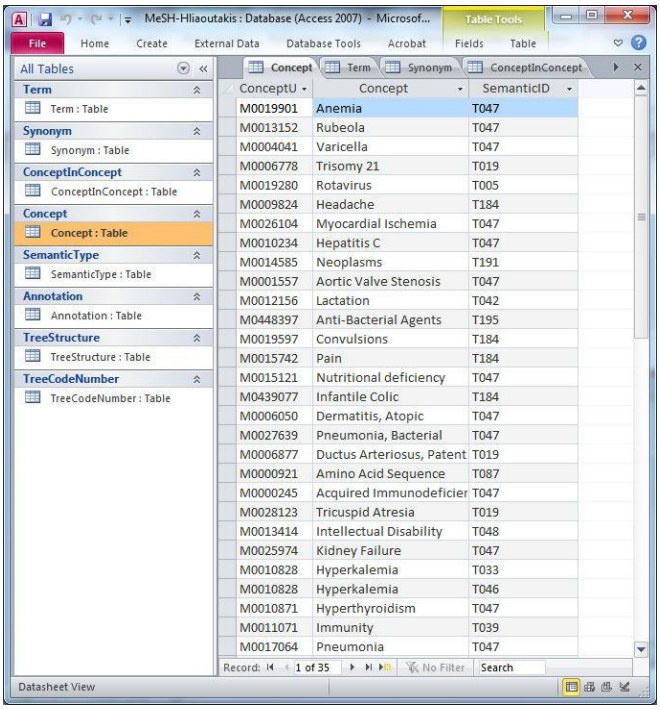

Figure 5: MeSH dataset

2.2. Similarity Measure

In this phase, measurement semantic similarity has been chosen. There are several approach is used in semantic a few method which is Tversky, similarity such as feature based, information content and structure. In this research feature based measure used as tool for this research. Feature based have Rodriguez and Egenofer and X- Similarity [17]. From both method, this research used Rodriguez and Egenhofer method [16] to measure.

2.3. Feature Based Process

This research had selected MeSH dataset, while in pre-processing; this research chooses thirty data from the dataset. Then we make comparison between the data in similarity process. .If the data similar, then we will make calculation based on the formula each method. Based on the phase 1, the method is picking and constructed based on existed algorithm. The dataset used Rodriguez and Egenhofer to measure [16].

2.4. Analysis the Result

Based on the phase 3, after testing the data, we have to compare the results. The value that will test is between 0 and 1 which means 0 is not similarity while 1 is exactly similar. If the value is more than 0, then the value will approximate to 1. If the value is not equal to 0, calculation with throughout with word matching, features, and neighbourhood can placed to get others data. Then the value with the highest or approximate correlation can be taken.

3. SIMULATION RESULTS AND ANALYSIS

This section also describes the system of the calculation which can be calculated and show the data. In this research, the model used a data created from the database which is access file that can be converting to SQL file. The designs of interface creating using bootstrap template. It also used Notepad++ as others tools help to create the interface design with Hypertext Preprocessor (PHP), and HyperText Markup Language (HTML) as a programming language. Figure 6 show the interface of medical system

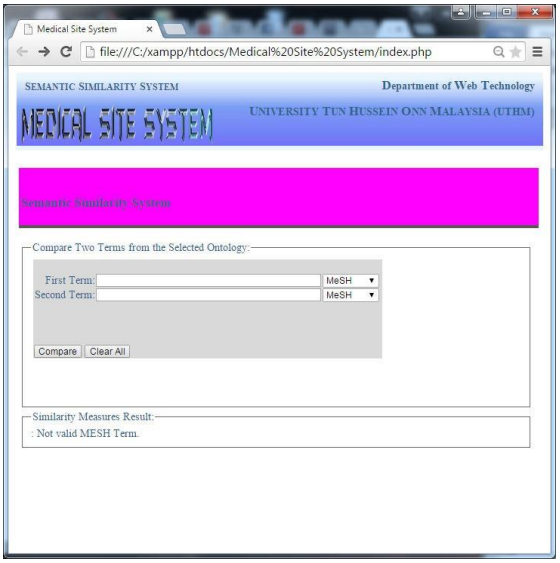

Figure 6: Interface of Medical System

The result is presented with an interface page which provides some simple output specifying success, with number of datasets as Figure 7. Similarity value is evaluate use correlation and the result correlation is show in Table 2.

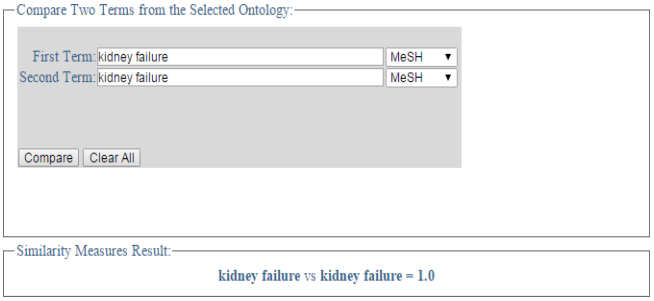

Figure 7: Status of result

Table 2: Correlation results

MeSH Term 1	MeSH Term 2	Correlations
Heart	Hearts	1.00
Miscarriage	Abortion applicant	0.00
Depression	Pain	0.78
Metastasis	Neoplasm	1.00
Mitral stenosis	Mitral valve stenosis	1.00
Diabetes mellitus	Rubeola	0.60
Syringe	Syringes	1.00
Myopia	Hyperkalemia	0.65
Kidney	Kidney failure	1.00

Allergy	Hypersensitivity	1.00
Osteoporosis	Tuberculosis	0.50
Delusion	Delusions	1.00
Alcoholic Cirrhosis	Headache	0.60
Anemia	Appendicitis	0.65
Pyelonephritis	Pneumonia	0.60

4. CONCLUSION

In conclusion, this research helps the user or programmers to gain the information that something new and can share their knowledge in term of semantic similarity. Apart of this, it discussed about how the semantic similarity works or flow the progress from the starting part until the end. On this research also explains about the method and the purpose why need to do this research. By the end of it, user can implement the feature based measure Rodriguez and Egehofer and MeSH as dataset.

REFERENCES

[1] Pedersen, T., Pakhomov, S.V.S., Patwardhan, S., Chute, C. G. 2007. Measures of semantic similarity and relatedness in the biomedical domain. Journal of Biomedical Informatics, 40, 288–299. https://doi.org/10.1016/j.jbi.2006.06.004.

[2] Slimani, T. 2013. Description and Evaluation of Semantic similarity Measures Approaches. Journal of Computer Applicationsof Computer Applications, 80 (10), 1–10.

[3] Miller, G.A., Charles, W.G. 1991. Contextual correlates of semantic similarity. Language and Cognitive Processes, 6 (1), 1-28. https://doi.org/10.1080/01690969108406936

[4] Rubenstein, H., Goodenough, J.B. 1965. Contextual correlates of synonymy. Communications of the ACM, 8 (10), 627-633. https://doi.org/10.1145/365628.365657.

[5] Rada, R., Mili, H., Bicknell, E., Blettner, M. 1989. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics, 19 (1), 17–30. https://doi.org/10.1109/21.24528.

[6] Pirró, G., Euzenat, J. 2010. A feature and information theoretic framework for semantic similarity and relatedness. The Semantic Web–ISWC 2010, pp 615-630. Retrieved from http://link.springer.com/chapter/10.1007/978-3- 642-17746-0_39.

[7] Schickel-Zuber, V., Faltings, B. 2007. OSS: A semantic similarity function based on hierarchical ontologies. IJCAI International Joint Conference on Artificial Intelligence, pp. 551–556.

[8] Kleinsorge, R., Tilley, C., Willis, J. 2002. Unified Medical Language System (UMLS). Encyclopedia of Library and Information Science, 369–378. https://doi.org/10.1002/9781118479612.ch16.

[9] Bodenreider, O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research, 32 (1), D267–D270. https://doi.org/10.1093/nar/gkh061.

[10] Hliaoutakis, A, Varelas, G., Voutsakis, E., Petrakis, E. G. M., Milios, E. 2006. Information Retrieval by Semantic Similarity. International Journal on Semantic Web and Information Systems, 2, 55–73. https://doi.org/10.4018/jswis.2006070104.

[11] Al-Mubaid, H., Nguyen, H. A. 2006. A cluster-based approach for semantic similarity in the biomedical domain. In Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings (pp. 2713–2717). https://doi.org/10.1109/IEMBS.2006.259235.

[12] Batet, M., Sánchez, D., Valls, A. 2011. An ontology-based measure to compute semantic similarity in biomedicine. Journal of Biomedical Informatics, 44 (1), 118–125. https://doi.org/10.1016/j.jbi.2010.09.002.

[13] Garla, V.N., Brandt, C. 2012. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. BMC Bioinformatics, 13, 261. https://doi.org/10.1186/1471-2105-13-261.

[14] Batet, M., Sánchez, D., Valls, A., Gibert, K. 2010. Exploiting taxonomical knowledge to compute semanticsimilarity: An evaluation in the biomedical domain. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 6096 LNAI, 274–283. https://doi.org/10.1007/978- 3-642-13022-9_28.

[15] Kasim, S. 2011. Fuzzy C-Means Clustering By Incorporating Biological Knowledge And Multi-Stage Filtering To Improve Gene Function Prediction. University Teknologi Malaysia.Thesis-PhD.

[16] Rodríguez, M. A., Egenhofer, M.J. 2003b. Determining semantic similarity among entity classes from different ontologies. IEEE Transactions on Knowledge and Data Engineering, 15, 442–456. https://doi.org/10.1109/TKDE.2003.1185844.

[17] Petrakis, E., Varelas, G., Hliaoutakis, A., Raftopoulou, P. 2006. X-Similarity: Computing Semantic Similarity between Concepts from Different Ontologies. Journal of Digital Information Management, 4 (4), 233. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.114.3247.