ISSN 2278-3091
Volume 8, No.1.3, 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse3581.32019.pdf
https://doi.org/10.30534/ijatcse/2019/3581.32019

Hyb-Tvx: A Hybrid Semantic Similarity Feature-Based Measurement for Multiple Ontologies

Nurul Aswa Omar[1], Shahreen Kasim[2], Muhammad Azani Hasibuan[3], Mohd Farhan Md Fudzee[1],
[1] Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia, nurulaswa@uthm.edu.my
[2] Soft Computing And Data Mining Centre, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Johor, Malaysia
[3] School of Industrial Engineering, Universiti Telkom,Bandung West Java, Indonesia

ABSTRACT

Semantic similarity is defined as the closeness of two concepts, based on the likeliness of their meaning. It is also more ontology-based, due to their efficiency, scalability, lack of constraints and the availability of large ontologies. However, ontology-based semantic similarity is hampered by the fact that it depends on the overall scope and detail of the background ontology. This leads to insufficient knowledge, miss-ing terms and inaccuracy. This limitation can be overcome by exploiting multiple ontologies. Semantic similarity with multiple ontologies potentially leads to better accuracy because it is able to calculate the similarity of these missing terms from the combination of multiple knowledge sources. This research aims to develop and evaluate a feature-based mechanism (Hyb-TvX) to measure semantic similarity with multiple ontologies which can improve the accuracy of the similarity. Similarity value, correlation and p-value were also used in the evalua-tion of the relationship between the concept pair of multiple ontologies. Besides that, the Hyb-TvX mechanism produces the highest correla-tion value compared to the other two methods, that is 0.759 and the result correlation is significant..

Key words : Semantic similarity, similarity measurement, ontology, features-based.

1. INTRODUCTION

Semantic similarity can be defined as the closeness of two concepts, based on the likeliness of their meaning which means that both theories stated that the semantic similarity acts as a mechanism for comparing an object [1]. Multiple ontologies are a method to compare concepts from different ontologies. Nevertheless, most of these similarity approaches are not capable of measuring semantic similarity between concepts in multiple ontologies. This is due to differing backgrounds of ontology in allowing the integration of sources. The integration of multiple ontologies will affect the accuracy of the similarity concept. This is because each ontology has its own structure and features [2].

Previous research emphasized on the use of structure in the similarity measurement [3]-[7]. However, to find similarity between the concepts of multiple ontologies, the use of structure is not required. This is because every ontology has a different structure that cannot be directly compared [8].

Therefore, measurement in a feature-based approach tries to overcome the limitation of the structure-based approach [10]. A feature-based approach has higher potential to be used in similarity measurement of multiple ontologies as it exploits more semantic knowledge than the structure-based approach with the evaluation of commonalities and differences of compared concepts.

The Rodríguez & Egenhofer measurement [9] uses the depth of ontology (structure) as a source of relative importance of non-common features and depends on the weighting parameter that balances the contribution of each feature. X-similarity method [5] did not depend on the weighting parameters. However, this method omitted other features when a maximum value is used each time in the similarity measurements. This omitted feature has a high potential in similarity measurement. Besides that, when the X-similarity method assumes a similarity value of more than zero to one, an unreliable result will be obtained [10].

Based on the above situation, this research suggests improving the semantic similarity methods with the hybrid method X-similarity [10] and Tversky method [11]. The two combined methods above are proposed to be Hyb-TvX. The process of Hyb-TvX is illustrated in the block diagram, Figure 1. This method has two phases, TvX-1 and TvX-2. In the next section, the proposed method (Hyb-TvX) will be described

2. HYB-TVX: A HYBRID SEMANTIC SIMILARITY FEATURE-BASED MEASUREMENT

Medical Subject Heading (MeSH) is a controlled vocabulary and a thesaurus developed by the U.S. National Library of Medicine (NLM). WordNet is the lexical knowledge of a native speaker of English. This research used two benchmark datasets as proposed by [12] and [13]. The first dataset consists of 36 pairs of medical terms extracted from MeSH and WordNet. This dataset includes 1390 concepts from WordNet and 926 concepts from MeSH. The second benchmark consists of 30 concept pairs of medical terms extracted from MeSH and WordNet. This dataset includes 1186 concepts from WordNet and 712 concepts from MeSH. Hyb-TvX process is divided into two phases. The TvX-1 is a similarity measurement level 1 while the TvX-2 is a similarity measurement level 2.

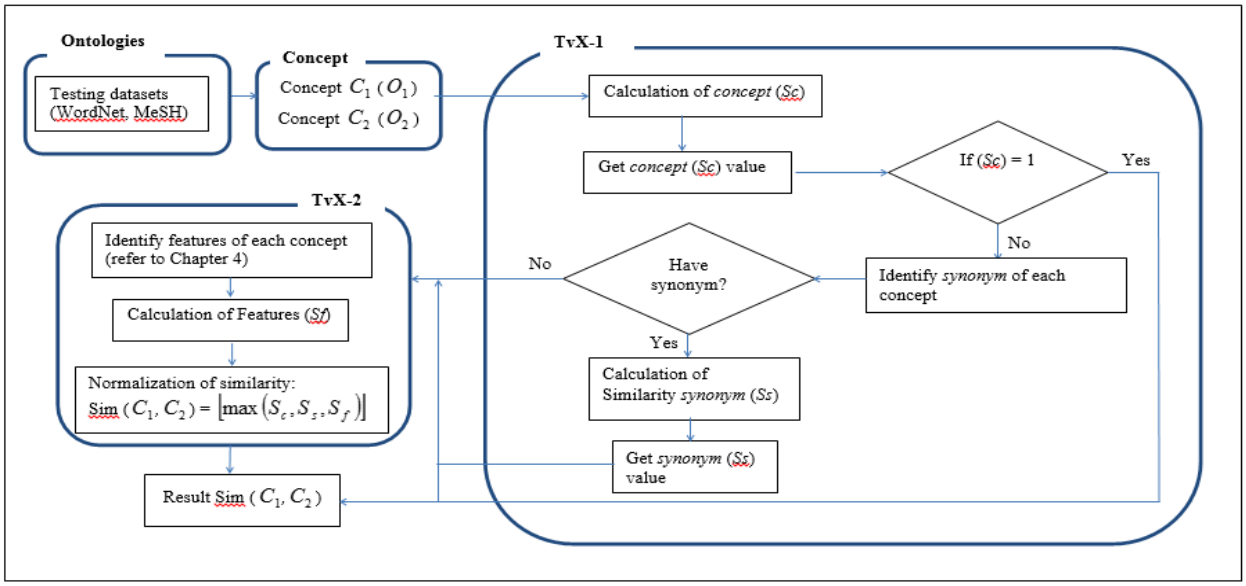

Figure 1: The flow process of Hyb-TvX

2.1. Tvx-1: Similarity Measurement Level 1

Two calculation steps are used for similarity in TvX-1. In the first step, the process begins by calculating the similarity concepts $(s_c(c_1, c_2))$ and the second step is the calculation of synonyms $(s_s(c_1, c_2))$.

2.1.1 Similarity Concept (s_c)

Similarity concept (s_c) is the calculation of similarity of concepts compared. Concepts compared are from different ontologies, the concept of the first ontology (o_1) is represented by the symbol (c_1) and the concept of second ontology (o_2) is represented by the symbol (c_2). In this section, the calculation of (s_c) can be derived by the implementation of $(Int|c_1, c_2|)$ and $(\max|c_1|, |c_2|)$. The two concepts compared are *renal failure* (c_1) and *kidney disease* (c_2), belonging respectively to ontology. The similarity concept (s_c) between the concepts of (c_1) and (c_2) is denoted in Equation (1):

$$s_c(c_1, c_2) = \frac{Int|c_1, c_2|}{\max(|c_1|, |c_2|)} \quad (1)$$

The example of concept and token are shown in Table 1:

Table 1: Examples of concepts

Concept		Token
(c_1)	renal failure	2
(c_2)	kidney disease	2

Calculation examples as follows:

$$s_c(c_1, c_2) = \frac{Int|c_1, c_2|}{\max(|c_1|, |c_2|)}$$

$$Int|c_1, c_2| = \{\ \}$$

$$\max(|c_1|, |c_2|) = \{2\}$$

$$s_c(c_1, c_2) = \frac{Int|c_1, c_2|}{\max(|c_1|, |c_2|)} = \frac{0}{2} = 0$$

Based on this calculation, the (s_c) of *renal failure* (c_1) and *kidney disease* (c_2) is equal to 0. There are two situations in this phase (i) If the value of $(s_c) = 1$, the value will declare a similarity value for (c_1, c_2) because the concepts compared have the same terminological concept, (ii) If the value of $s_c(c_1, c_2) < 1$, the second step of similarity synonym (s_s) in this phase will be continued.

2.1.2 Similarity Concept (s_s)

The second step in this phase is to calculate the similarity synonym (s_s) of each concept. In this step, each concept contains two kinds of conditions. The first condition has a synonym concept while the second condition does not have a synonym. In the first condition, the calculation of similarity of synonym (s_s) is executed while the second condition will continue in the next phase (TvX-2).

Similarity synonym (s_s), is the calculation of similarity of the synonym for the concepts compared. The concepts compared are from different ontology, the concept of the first ontology

(O_1) is represented by the symbol (C_1) and the concept of (O_2) is represented by the symbol (C_2). In this calculation, $(Int|A, B|)$ terms are involved without separating them to single words as before, where set *A* denoted (O_1) and set *B* denoted (O_2). Besides that, this calculation also uses the union (Un) synonym for concepts (C_1) and (C_2) $(Un|A, B|)$.

The same example (concepts compared are *renal failure* (C_1) and *kidney disease* (C_2) in Table 1 , the synonyms for *kidney disease* are *renal failure* and *kidney failure* as stated in Table 2 below:

Table 2: Example for concepts and synonym

Concept		Synonym
(C_1)	Renal failure	kidney failure
(C_2)	Kidney disease	renal failure, kidney failure

The calculation of synonym is presented in Equation (2).

$$S_s(C_1, C_2) = \frac{Int\,|A, B|}{Un|A, B|} = \frac{1}{2} = 0. \qquad (2)$$

$Int|A, B|$: {*kidney failure*}

$Un|A, B|$: {*renal failure, kidney failure*}

Based on this measurement, (s_s) for *renal failure* (C_1) and *kidney disease* (C_2) is equal to 0.5. These measures are based on the measurement by X-similarity, where it is used to measure synonym as defined in the literature. This value will be brought to the next phase (TvX-2). Therefore, calculation TvX-2 will continue and value (s_s) will compare maximality.

2.2. Tvx-2: Similarity Measurement Level 2

The process in the second phase aims to discover the limitation of X-similarity method in which other features are omitted when the max value is taken. Due to this limitation, Hyb-TvX method has a second calculation (TvX-2). The second phase calculates the similarity for ontological features such as hypernym, hyponym, sister term and meronym/holonym or also known as similarity features (s_f).

This phase involves the calculation used by the Tversky method. The calculation involves the use of operation sets such as intersection (Int) and complement $(comp)$. In addition, this calculation also involves the use of parameters (w_a) and (w_b) to balance the non-common features involved.

In the (TvX-2) phase, the ontological features were computed using the Tversky method as the basis for calculation. This calculation uses the parameters of α and β *or* (*1- α*) in the Tversky method, following $\alpha + \beta = 1$ (for instance, if $\alpha = 0.2$, $\beta = 0.8$). The measurement of Rodríguez & Egenhofer method uses the ontology structure to gain parameters. They use the depths of ontology (C_1) and (C_2) to obtain the parameters.

In this method, the proposed parameters (w_a) and (w_b) where it depends on the value of |*comp B*| and |*comp A*|, which means that if |*comp B*| > |*comp A*|, the parameters must be (w_a) = 0.1 and (w_b) = 0.9 and if |*comp B*| < |*comp A*|, the parameters must be (w_a) = 0.9 and (w_b) = 0.1 to obtain the optimum value of similarity and balancing the non-common features while measurement occurs.

In this section, the calculation of (s_f) can be derived by the implementation of $(Int|A, B|)$, |*comp A*|, |*comp B*| and the proposed parameters (w_a), (w_b) The two concepts compared are *renal failure* (C_1) and *kidney disease* (C_2), belonging respectively to ontology. The similarity concept (s_f) between the concepts of (C_1) and (C_2) is shown in Equation (3):

The calculation features is demonstrated in Equation (3).

$$S_f(C_1, C_2) = \frac{Int|A, B|}{Int|A, B| + (w_a)|comp\ B| + (w_b)|comp\ A|} \qquad (3)$$

According to the concept in Table 3, ontological features that are related to a specific concept have been extracted.

Table 3: Example of concepts and features

Concept		Features
(C_1)	Renal failure	kidney failure, urologic diseases, kidney diseases
(C_2)	Kidney disease	kidney failure, renal failure, disease or syndrome, renal insufficiency, male urogenital diseases, urologic diseases, kidney diseases

The calculation of features is as follows:

$Int|A, B|$: {*kidney failure, urologic diseases and kidney diseases*}

|*comp B*|: {}

|*comp A*|: {*renal failure, disease or syndrome, renal insufficiency, male urogenital diseases*}

|*comp B*| < |*comp A*| = {(w_a) = 0.9 and (w_b) =0.1}

$$S_f(C_1, C_2) = \frac{Int|A, B|}{Int|A, B| + (w_a)|comp\ B| + (w_b)|comp\ A|}$$

$$s_f(C_1, C_2) = \frac{|3|}{|3| + (w_a)|0| + (w_b)|4|}$$

$$s_f(C_1, C_2) = \frac{|3|}{|3| + 0.9|0| + 0.1|4|} = \frac{3}{3+0+0.4} = 0.882$$

Then, (s_c), (s_s), (s_f) are calculated to get the maximum value between similarity concept, similarity synonym and similarity features $\left[\max(s_c, s_s, s_f)\right]$ as denoted by Equation (4). According to the concepts of *renal failure* (C_1) and *kidney disease* (C_2), the maximum value of similarity TvX-1 is equal to 0.882 which comes from (s_f).

$$s(C_1, C_2) = \left[\max(s_c, s_s, s_f)\right] \quad (4)$$

The final similarity $s(C_1, C_2)$ for *renal failure* (C_1) and *kidney disease* (C_2) is equal to 0.882. Using this similarity, similarity value for that concept have defined.

3. RESULT AND DISCUSSION

There are two approaches used to assess the accuracy of similarity values calculated by a given similarity measure. The first approach is to employ the similarity measure in applications that require similarity between words and the second approach is to compare the computed similarity values of the measure against the human similarity scores such as the correlation coefficient. This approach requires a datasets of term pairs scored for similarity by human. This research employed both approaches to evaluate Hyb-TvX method. Besides that, this research also used *p*-value to measure the "significance" of similarity result.

Table 4: Comparison similarity of proposed method with physician, coder and experts ratings (averaged).

WordNet	MeSH	Method	
		physician, coder and experts ratings (averaged)	Proposed (Hyb-TvX)
Renal failure	Kidney failure	1	1
Heart	Myocardium	0.875	1
Abortion	Miscarriage	0.787	1
Metastasis	Adenocarcinoma	0.562	0.588
Mitral stenosis	Atrial fibrillation	0.45	0.516
Pulmonary embolism	Myocardial infarction	0.375	0.327
Hypertension	Diabetes mellitus	0.250	0.491
Appendicitis	Anemia	0.031	0.144
Kidney failure	Hypertension	0.500	0.439
Hepatitis C	Hepatitis B	0.562	0.952
Aortic stenosis	Pulmonary stenosis	0.531	0.833
Convulsions	Seizures	0.843	0.714
Ache	Pain	0.875	1
Rubeola	Measles	0.906	1
Varicella	Chicken pox	0.968	1
Trisomy 21	Down syndrome	0.875	1
Anemia	Deficiency anemia	0.437	0.5
Anti-bacterial agents	Antibiotics	0.937	1
Malnutrition	Nutritional deficiency	0.875	1

The evaluation of similarity measures is usually performed by comparing (Hyb-TvX) the similarity values with those provided by human scored (physician, coder and experts ratings). Table 4 shows a comparison similarity of the proposed method (Hyb-TvX) with averaged physician, coder and experts ratings.

Based on Table 5, Hyb-TvX method produces the highest correlation value as compared to the other two methods, that is 0.759 and the correlation result is significant with its *p*-value <0.01 which is based on results shown in Table 6. Meanwhile, X-similarity method shows a correlation with human scored 0.554 also significant, as shown in Table 5 as the *p*-value is lower than 0.01. Next, the Rodriguez and Egenhofer method has recorded the lowest correlation compared to other methods in this research which is 0.429 with its *p*-value > 0.01 as in Table 5. This proves that the Hyb-TvX method is the best method to use for a feature-based approach.

Table 5: Correlation of similarity method on feature-based approach according to the WordNet and MeSH dataset

Method	Correlation
Rodriguez and Egenhofer	0.429
X-similarity	0.554
Hyb-TvX	0.759

Table 6: *p*-value of similarity method on feature-based approach for multiple ontologies

Method	*p*- value	Result
Rodriguez and Egenhofer	0.01261	Not Significant
X-similarity	0.00080	Significant
Hyb-TvX	0.00001	Significant

4. CONCLUSIONS

The Hyb-TvX method is a hybrid of two previous methods, namely X-similarity and the Tversky method. The proposed Hyb-TvX method is a measurement of TvX-1 and TvX-2 where

TvX-1 is the original measurement of X-similarity. TvX-2 is an improved measurement that uses the Tversky method incorporated with ontological features. The proposed Hyb-TvX method overcomes the limitations of the X-similarity method wherein other features are omitted when when taking a maximum value of similarity synonym (similarity synonym > 0 = 1). Besides that, with the use of proposed parameters, the Hyb-TvX method has improved upon the Tversky method. The proposed method (Hyb-TvX) has three contributions: the first one is that the Hyb-TvX method does not leave out other features. The second contribution is that it uses two feature-based approaches to solve problems in accuracy. The last contribution is that the proposed parameter (w_a) and (w_b) are used to balance the non-common features during the similarity measurement process. This proposed Hyb-TvX method produces the highest correlation value as compared to the other two methods, that is 0.759 and the correlation result is significant with its *p*-value <0.01 which is based on the results shown in Table 6.

ACKNOWLEDGEMENT

We would like to thank Universiti Tun Hussein Onn Malaysia for supporting this research under the Contract Grant-Endowment (Vot number: A066), also, thanks to Gates IT Solution Sdn Bhd for the whole support.

REFERENCES

1. Elavarasi, S., Akilandeswari, J., & Menaga, K. (2014). **A Survey on Semantic Similarity Measure**. Ijrat.org, 2(3), 389–398. Retrieved from http://www.ijrat.org/downloads/march-2014/paper id-232014114.pdf.M. Davis, H. Putnam. A computing procedure for quantification theory. Journal of ACM, 7(3)(1960) 201-215. https://doi.org/10.1145/321033.321034
2. Sánchez, D., & Batet, M. (2012). **A new model to compute the information content of concepts from taxonomic knowledge**. International Journal on Semantic Web and Information Systems, 8, 34–50. https://doi.org/10.4018/jswis.2012040102.M. Davis, G. Logemann, D. Loveland, A machine program for theorem proving. Communications of the ACM, 5 (1962) 394-397. https://doi.org/10.4018/jswis.2012040102
3. 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.
4. Bulskov, H., Knappe, R., & Andreasen, T. (2002). **On measuring similarity for conceptual querying**. Flexible Query Answering Systems, 100–111. Retrieved from http://link.springer.com/chapter/10.1007/3-540-36109-X_8. https://doi.org/10.1007/3-540-36109-X_8
5. Al-Mubaid, H., & Nguyen, H. A (2009). **Measuring semantic similarity between biomedical concepts within multiple ontologies**. IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews, 39(4), 389–398. https://doi.org/10.1109/TSMCC.2009.2020689
6. Sánchez, D., Batet, M., Isern, D., & Valls, A. (2012). **Ontology-based semantic similarity: A new feature-based approach**. Expert Systems with Applications, 39(9), 7718–7728. https://doi.org/10.1016/j.eswa.2012.01.082.
7. Leacock, C., & Chodorow, M. (1998**). Combining Local Context and WordNet Similarity for Word Sense Identification**. WordNet: An electronic lexical database. (pp. 265–283). https://doi.org/citeulike-article-id:1259480.
8. Sánchez, D., & Batet, M. (2013**). A semantic similarity method based on information content exploiting multiple ontologies**. Expert Systems with Applications, 40(4), 1393–1399. https://doi.org/10.1016/j.eswa.2012.08.049.H Lin, JG Sun, YM Zhang. Theorem proving based on the extension rule, Journal of Automated Reasoning, 31 (2003) 11-21.
9. 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.
10. 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.
11. Tversky, A. (1977). Features of similarity. Psychological Review. https://doi.org/10.1037/0033-295X.84.4.3
12. 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.
13. 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.