Extended TvX: A New Method Feature Based Semantic Similarity for Multiple Ontology

Nurul Aswa Bt Omar[1], Shahreen Kasim[2], Mohd Farhan Md Fudzee[3]
[1]*Department Web Technology,*
[2]*Soft Computing And Data Mining Center,*
[3]*Department Multimedia,*
Faculty Computer Sciences and Information Technology,
Universiti Tun Hussein Onn Malaysia, Malaysia.
nurulaswa@uthm.edu.my

***Abstract*—Semantic similarity between the terms is the main phase in information retrieval and information integration, which requires semantic content matching. Semantic similarity function is important in psychology, artificial intelligence and cognitive science. The problem of integrating various sources is the matching between ontological concepts. In this paper, we proposed to develop this method by analyzing the semantic similarity between the modeled taxonomical knowledge and features in different ontology. This paper contains a review on semantic similarity and multiple ontology that focuses on the feature-based approach. Besides that, we proposed a method, namely a semantic similarity that overcomes the limitation of different features of terms compared. As a result, we are able to develop a better method that improves the accuracy of the similarity measurement.**

***Index Terms*—Semantic Similarity; Feature Based; Ontology; Multiple Ontology; Cross Ontology; Heterogeneous Sources.**

I. INTRODUCTION

Semantic similarity can be defined solely based on the joint probability distribution of the concepts involved [1]. Besides that, semantic similarity also can be defined as the closeness of two concepts, based on the likeliness of their meaning, which means that both theory state that the semantic similarity acts as a mechanism for comparing an object.

By referring to Batet [2], semantic similarity in recent years has been widely used in obtaining the similarities between concepts or between terms, where it is important to support information extraction [3] such as semantic annotation [4] and ontology learning [5]. In addition, semantic similarity is also importantant in information retrieval [6-8] and information integration [6]. Information retrieval tasks improve the performance of current search engines [9] while information integration uses semantic similarity to discover concepts between entities belonging to different ontologies [6, 10].

Recently, the similarity approach is not limited to single ontology only. Currently, a similar approach also is used in multiple ontologies. Multiple ontology is a method to compare concepts from different ontologies, such as Wordnet and MeSH. Nevertheless, most of these similarity approaches could not measure the semantic similarity between concepts in multiple ontologies. This is due to different background of ontology in allowing the source of integration. Integration of multiple sources of the ontology in different ontology backgrounds will affect the accuracy similarity concept. This is because each ontology has its own structure and feature. Previous research has emphasized on the different structures of the ontology, but they do not give attention on the future when in a different background situation.

We proposed similarity approach that overcomes the limitation of different features of concept. This is due to the fact that each ontology has its own structure and feature. ExT-TvX is an extended method from Petrakis et. al [11] where this method has two phases TvX-1 and TvX-2. As a result, our method is better as it can improve the accurancy of the similarity. In our method, we have two contributions: Firstly, our method does not leave other features athough we use max value. Secondly, we use the way of "single ontology" approach to solve the problem of different features of each concept.

In the next section, we described in detail the works that have been done in semantic similarity.

II. RELATED WORK

Nowadays with the mushrooming of information sources on the web, there is a need to develop measurements that computes similarity among concepts in different ontologies [12, 6]. Multiple ontology similarity measurement will match the terms from different ontology. Multiple ontology often needs hybrid or feature based approach because the information content based and the structure based approach cannot be compared directly in different ontologies [12].

(i) Structure based approach
Path length approach is based on an ontology's structure, in which the ontological primary relationship is 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 [13] and [11]. Meanwhile, the depth relative approach considers the connecting edges of two concepts in structure ontology. It computes the depth from root to the target concept.

(ii) Feature based approach
This approach considers terms that are represented as collections of feature and the specific differentiating feature of each concept.

In this study, we concentrated on feature based approaches. Feature based approach is a more general approach. It is potentially used in multiple ontology because the concept of two different ontologies has a different structure. This is due to the fact that the structure between diverse ontologies cannot be compared directly [11, 12, 14].

Works in feature based approach are Rodriguez and Egenhofer [15] and X-similarity [11]. Rodriguez and Egenhofer [15] developed the method to represent terms as a collection of feature and their similarity as a feature matching process. Equation (1) from Rodriguez and Egenhofer uses X and Y that correspond to sets of a and b, where |X∩Y| is an intersect set function and |X-Y| denotes the relative complement of Y in X. They use similarity to determines similar entity by using matching process that are classified into parts of synonym sets (S_p), semantic neighborhoods (S_f) and attributes (S_a). To compute the synonym set, semantic neighborhood and feature matching, Equation (1) as shown below is used where ap and bq is the entity class of ontologies p and q:

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

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:

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

$$if\ depth(a^p) > depth\ (b^q)$$

To combine the information gained from similarity distinguishing feature, synonym sets and semantic neighborhoods, their similarity is defined by the amount sum of the weights of each component as shown in Equation (2). The functions synonym sets (Sw), feature (Su) and semantic neighborhoods (Sn) are similarity between entity classes a of ontology p and b of ontology q and Ww, Wu and Wn is the weight each specification component.

$$S(a^p, b^q) = W_w . S_w(a^p, b^q) + W_u . S_u(a^p, b^q) + W_n . S_n(a^p, b^q) \quad (2)$$

for $W_w, W_u, W_n \geq 0$

X-Similarity, developed by Petrakis et. al [11] is a novel multiple ontology similarity method. *X*-similarity depends on similarity between synsets (synonym) and description sets. Rodriguez and time according to Petrakis et. al [11] similarity multiple ontology should not consider ontology structure information. Due to this, Petrakis et al. [11] proposed replacing Equation (1) with Equation (3) below with a simple set of similarity, where *A* and *B* denote synset (synonym) or term description sets.

$$S(a.b) = \frac{|A \cap B|}{|A \cup B|} \quad (3)$$

They also proposed Equation (4), where the sets of similarities are computed per relationship type, such as is-A and part-Of , where *i* denotes relationship type.

$$S_{neighborhood}(a.b) = max\ \frac{|A_i \cap B_i|}{|A_i \cup B_i|} \quad (4)$$

The above idea is combined into a single formula as shown in Equation (5).

$$Sim\ (a,b) = \begin{cases} 1 & if\ S_{synsets}\ (a,b) > 0 \\ max\{S_n(a,b), S_d(a,b)\}, & if\ S_{synsets}(a,b) = 0 \end{cases} \quad (5)$$

Feature based approach has tried to solve the limitation of structure based approach concerning the fact that taxonomical links in an ontology does not necessarily represent uniform distances [16]. However, this approach also has its disadvantages, where it depends too much on the information provided. Table 1 below describes briefly the pros and cons of each method in feature based approach.

Table 1
Method feature based for multiple ontology

Methods	Advantage/s	Disadvantage/s	References
Rodriguez and Egenhofer [15]	Take into account semantic neighborhoods in the calculation of similarity.	Incomplete part for calculation will cause low accuracy. Parameter γ using the depth of the ontologies.	[11,15, 17, 14]
X-Similarity [11]	Does not be influenced by on weighting parameter. The maximum value is taken for every measurement feature	Omitted other feature is because the maximum value is taken at every time.	[11, 14]

III. Ext-TvX: A Matching Approach to Similarity Assessment

The process of extended TvX (Ext-TvX) is illustrated in the block diagram in Figure 1. This process is divided into two phases. TvX-1 is the calculation of similarity level 1, while TvX-2 is the calculation of similarity level 2.

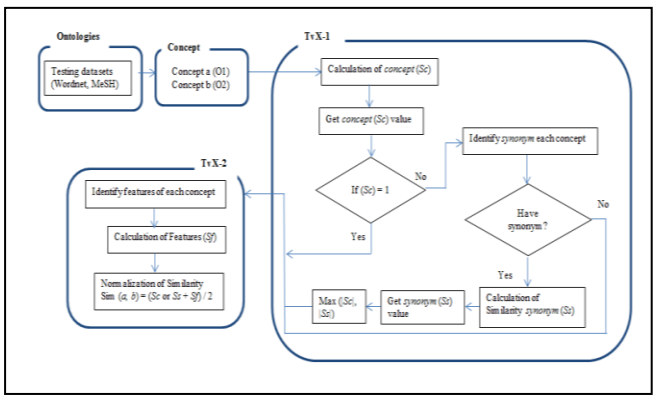

Figure 1: The process of Extended TvX

A. TvX-1: Similarity Calculation Level 1

In TvX-1, we have two calculation steps for similarity. In the first step, the process begins by calculating the similarity concepts ($S_c(a,b)$) and second step is the calculation of synonym ($S_s(a,b)$). The two concepts being compared are concept a (renal failure) and concept b (kidney disease), belonging respectively to ontology. The similarity concept ($S_c(a,b)$) between the concepts of a and b, is shown in Equation (6):

$$S_c(a,b) = \frac{Int|a,b|}{max(|a|,|b|)} \quad (6)$$

The concepts will be extracting a set of token, by dividing a string of punctuation and separation, blank spaces and uppercase changes. Similarity concept ($S_c(a,b)$) involves an intersection between concept a and b ($Int|a,b|$) and maximum tokenization ($\max(|a|,|b|)$). Examples of the calculation are as follows, which are based on Table 2.

Table 2
Example for concepts compared

	Concept	Token
Concept *a*	Renal failure	2
Concept *b*	Kidney disease	2

$$S_c(a,b) = \frac{Int|a,b|}{max(|a|,|b|)}$$

$$Int|a,b| = \{\emptyset\}$$

$$\max(|a|,|b|) = \{2\}$$

Based on this calculation:

$$S_c(a,b) = \frac{Int|a,b|}{max(|a|,|b|)} = \frac{0}{2} = 0$$

There are two situations: 1) If the value of *Sc(a,b) = 1*, the value will be brought to the next phase (TvX-2), 2) If the value of $S_c(a,b) < 1$, the second step in this phase will be continued. The second step in this phase is the calculated synonym ($S_s(a,b)$) 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 the similarity of synonym ($S_s(a,b)$) will be executed, while the second condition will be continued in the next phase (TvX-2).

Using the same example, the synonym for renal failure is kidney failure and synonyms for kidney disease are renal failure and kidney failure, as stated in Table 3 below:

Table 3
Example for concepts and synonym

	Concept	Synonym/s
Concept *a*	Renal failure	kidney failure
Concept *b*	Kidney disease	renal failure, kidney failure

Similarity synonym $S_s(a,b)$ involves intersection between concept a and b ($Int|a,b|$) and union ($Un|a,b|$). The calculation of synonym is as follows, as shown in Equation (7).

$$S_s(a,b) = \frac{Int|a,b|}{Un|a,b|} = \frac{1}{2} = 0.5 \quad (7)$$

if: $S_s(a,b) > 0 = 1$
$Int|a,b|$: {kidney failure}
$Un|a,b|$: {renal failure, kidney failure}

Get the maximum value between the similarity concept and the similarity of synonym (max {*Sc*,*Ss*}). According to the concepts of *a* (Renal failure) and concept *b* (Kidney Disease) the value of similarity TvX-1 is max{*Sc*,*Ss*} = 1. The value will be brought to the next phase (TvX-2).

B. TvX-2: Similarity Calculation Level 2

The process in the second phase starts with the calculation of the similarity (TvX-2) for features such as excessive hyponym, hypernym, meronym, holonym. Referring to the problems of multiple ontology in different backgrounds, different features of each concept will affect the accuracy. Therefore, this phase will use the concept of "single ontology" to solve this problem. As shown in Figure 2, the similarities between Rf of ontology 1 (O1) and Kf of ontology 2 (O2) can be seen by looking at the concept of Rf at ontology 2 (O2). Assuming that, Rf is similar to Kd, feature in Kd is compared to feature in Kf.

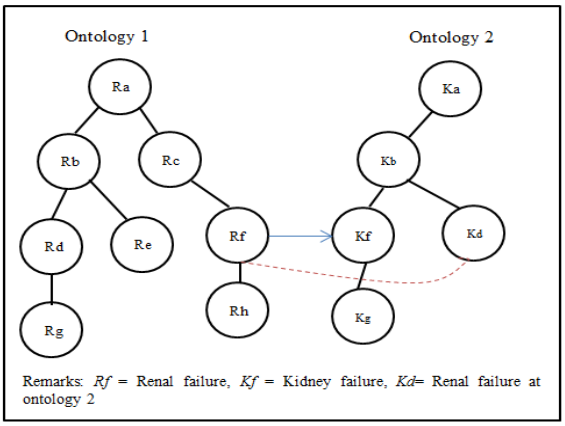

Figure 2: Illustration of connecting two ontologies

This feature is computed using Tversky method [18] as the basis for calculation. The function of α and β in Tversky method follow α + β = 1 for instance, if α = 0.2, β = 0.8. This will cause the similarity result to have more than one value. In this method, we use dynamic function W_a and W_b where it depends on the value of comp B and comp A, which means 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. The calculation features are as shown in Equation (8).

$$S_f(a,b) = \frac{Int|a,b|}{Int|a,b| + W_a|comp\ B| + W_b|comp\ A|} \quad (8)$$

According to the concept in Table 4, we extract all features that are related to the specific concept.

Table 4
Example of concepts and feature

	Concept	Features
Concept *a*	Renal failure	kidney failure, urologic diseases, kidney diseases
Concept *b*	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: {Ø}
comp A: {renal failure, disease or syndrome, renal insufficiency, male urogenital diseases}
comp B < *comp A* = {α= 0.9 and β=0.1}

$$S_f(a,b) = \frac{Int|a,b|}{Int|a,b| + W_a|comp\ B| + W_b|comp\ A|}$$

$$S_f(a,b) = \frac{|3|}{|3| + W_a|0| + W_b|4|}$$

$$S_f(a,b) = \frac{|3|}{|3| + 0.9|0| + 0.1|4|} = \frac{3}{3 + 0 + 0.4} = 0.882$$

To combine the information gained from the similarity calculation of concept, synonym and feature, we suggest the calculation of semantic similarity as shown in Equation (9).

$$S(a,b) = \frac{1}{2}\left[max(S_c, S_s) + S_f\right] \quad (9)$$

The final similarity $S(a,b)$ for concept a (renal failure) and concept b (kidney disease) is (1+0.882) / 2 = 0.941. Using this similarity we will define a value similarity for that concept.

IV. Experiment and Result

A. Dataset

Datasets used in these experiments are the domain of biomedical datasets. We used a set of 30 concept pairs. The dataset used in this evaluation are Wordnet [20] and Mesh [21] . Wordnet dataset describes more than 100,000 general concepts, which are structured of Wordnet in ontological form. The Medical Subject Headings (MeSH) [21] contains medical biological terms defined by US National Library of Medicine, which are structured in ontological way as well. In MeSH, there are 16 basic categories with more than 22,000 concepts.

We used Wordnet 2.0 as the first ontology, while Mesh as the secondary ontology. The Wordnet database was downloaded from http://wordnet.princeton.edu and the MeSH database was downloaded http://www.nlm.nih.gov/mesh/meshhome .html. This dataset has become synonymous in the study of semantic similarity, as previous researcher David Sanchez et al.[19] also used this dataset in their work.

B. Experimental Results

In this research, we used 30 concept pairs of biomedical terms to evaluate our proposed method, X-similarity method [11] and Rodriguez and Egenhofer method [15]. Unfortunately, according to Petrakis [11] standard evaluation benchmarks for multiple ontology method have not been proposed. Works that have been done before this are substantially different ontologies, such as Wordnet and MeSH. Results are compared according to similarity ratings provided by human experts.

Table 5
The comparison of similarity accurancy for Ext-TvX, X-similarity and Rodriguez and Egenhofer method

Wordnet	MeSH	Method			Wordnet	MeSH	Method		
		Proposed (Ext-TvX)	X-similarity	Rodriguez and Egenhofer			Proposed (Ext-TvX)	X-similarity	Rodriguez and Egenhofer
Renal failure	Kidney failure	1	1	0	Headache	Migraine	0.37	0.042	0
Myorcardium	Heart	0.5	0.183	0	Myocardial infarction	Myocardial ishemia	0.75	0.47	0
Hyperkalemia	Hyperlipidemia	0.5	0.182	0	Hepatitis B	Hepatitis C	0.65	0.42	0.016
Pneumonia	Asthma	0.294	0.07	0.0119	Carcinoma	Neoplasm	0.357	0.17	0.04
Diabetes mellitus	Diabetic nephropathy	0.3	0.205	0.018	Breast feeding	Lactation	0.084	0	0
Lactose intolerance	Irritable bowel syndrome	0.44	0.047	0.005	Measles	Rubeola	1	1	0.245
Urinary tract infection	Pyelonephritis	0.25	0.03	0.01	Malnutrition	Nutritional deficiency	1	1	0.143
Iron deficiency anemia	Sickle cell anemia	0.629	0.14	0.011	Varicella	Chicken pox	1	1	0.247
Psychology	Cognitive science	0.4	0.25	0.008	Down syndrome	Trisomy 21	1	1	0.146
Adenovirus	Rotavirus	0.25	0.16	0.018					

Analyzing the results shown in Table 5, there is a slight increase of accuracy similarity compared to the previous method. From these 30 concept pairs, 12 concept pairs showed an increase in accuracy similarity in comparison to the X-similarity method, while 5 concept pairs maintained the same result. Our proposed method achieves 23% better than X-similarity method. This is due to the use of dynamic function W_a, W_b and the calculation features in TvX-2: The calculation of Similarity level 2. Ext-TvX does not leave other features, although the value of similarity concepts and synonym are equal to 1. This is important due to the factor of a second calculation is needed to ensure that the concept is similar.

Based on Table 6, the correlation result for our method has improved in 0.5% from X-similarity and improved 14% correction compared with Rodriguez methods. This shows that our proposed method succeeded in increasing the accuracy of similarity.

Table 6
Correlation of similarity approach on feature-based method for multiple ontology

Method	Method Type	Correlation
Rodriguez	Feature-based	0.552
X-similarity	Feature-based	0.687
Ext-TvX	Feature-based	0.692

V. CONCLUSION AND FUTURE WORK

This paper described the basic of semantic similarity measure and a brief introduction about the importance of the use of semantic similarity in various fields. We also described in more detail about the method in the feature based approach, which is believed to be the most appropriate approach used to find the similarity between terms in multiple ontology. The feature based approach has the potential in increasing efficiency and accuracy similarity between multiple ontology without using structural information. Besides, we also described the proposed calculation(Ext-TvX), where this proposed multi-tier calculation of similarity is to ensure similarity of concepts. Results showed that our proposed method have improved than previous method. We used correlation (Pearson) coefficient to evaluate the improvement. In the future, we will make the prediction for the data with no features by using the method from msf-CluFA [22].

ACKNOWLEDGMENT

This works funded by the Universiti Tun Hussein Onn Malaysia(UTHM) under grant Research Acculturation Collaborative vote no. 1447. Many thanks to GATES IT Solution Sdn Bhd for ideas and collaboration.

REFERENCES

[1] Doan A., Madhavan J., Domingos P., and Halevy A., 2004. Ontology matching: A Machine Learning Approach. Handbook On Ontologies. 1–20.
[2] Batet M., Sánchez D., Valls A., and Gibert K., 2013. Semantic similarity Estimation From Multiple Ontologies, Applied Intelligence. 38(1):29–44.
[3] Sánchez D.and Isern D., 2011. Automatic Extraction Of Acronym Definitions From Theweb. Applied Intelligence. 34: 311–327.
[4] Sánchez D., Isern D., and Millan M., 2011. Content Annotation For The Semantic Web: An Automatic Web-Based Approach. Knowledge and Information Systems. 27:393–418.
[5] Iannone L., Palmisano I., and Fanizzi N., 2007. An Algorithm Based On Counterfactuals For Concept Learning In The Semantic Web. Applied Intelligence. 26:139–159.
[6] Saruladha K., Aghila G., and Bhuvaneswary A., 2011. COSS: Cross Ontology Semantic Similarity measure — An Information Content Based Approach. 2011 International Conference on Recent Trends in Information Technology (ICRTIT). 485–490.
[7] Al-Mubaid H.and Nguyen H., 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,
[8] Budanitsky A. andHirst G., 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. Computational Linguistics. 32:13–47.
[9] a Hliaoutakis, Varelas G., Voutsakis E., Petrakis E. G. M., and Milios E., 2006. Information Retrieval by Semantic Similarity. International Journal on Semantic Web and Information Systems. 2:55–73.
[10] Pirrò G., Ruffolo M., and Talia D., 2009. 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). 5480 LNCS. 1–36.
[11] Petrakis E., Varelas G., Hliaoutakis A., and Raftopoulou P., X-Similarity: Computing Semantic Similarity between Concepts from Different Ontologies. Journal of Digital Information Management. 4(4):233, 2006.
[12] Elavarasi S., Akilandeswari J., and Menaga K., 2014. A Survey on Semantic Similarity Measure. ijrat.org. 2 (3):389–398.
[13] Rada R., Mili H., Bicknell E., and Blettner M., 1989. Development And Application Of A Metric On Semantic Nets. IEEE Transactions on Systems, Man, and Cybernetics. 19(1):17–30.
[14] Sánchez D., Batet M., Isern D., and Valls A., 2012. Ontology-Based Semantic Similarity: A New Feature-Based Approach, Expert Systems with Applications. 39:7718–7728,
[15] Rodríguez M. and Egenhofer M., 2003. Determining Semantic Similarity Among Entity Classes From Different Ontologies, Knowledge and Data. 15(2)442–456.
[16] Sánchez D. and Batet M., 2013. A Semantic Similarity Method Based On Information Content Exploiting Multiple Ontologies, Expert Systems with Applications. 40(4): 1393–1399, Mar.
[17] Li H., Tian Y., and Cai Q., 2011. Improvement Of Semantic Similarity Algorithm Based on WordNet. in Proceedings of the 2011 6th IEEE Conference on Industrial Electronics and Applications, ICIEA 2011. 564–567.
[18] Tversky A., 1977. Features of Similarity. Psychological Review. 84:327–352,
[19] Sánchez D., Solé-Ribalta A., Batet M., and Serratosa F., 2012. Enabling Semantic Similarity Estimation Across Multiple Ontologies: An Evaluation In The Biomedical Domain., Journal Of Biomedical Informatics. 45(1):141–55, Feb.
[20] https://wordnet.princeton.edu
[21] https://www.nlm.nih.gov/mesh
[22] Kasim S., Deris S., Othman R. M. 2013. Multi-Stage Filtering For Improving Confidence Level And Determining Dominant Clusters In Clustering Algorithms Of Gene Expression Data. Computers In Biology And Medicine. 43(9):1120-1133.