A HYBRID SEMANTIC SIMILARITY FEATURE-BASED TO SUPPORT MULTIPLE ONTOLOGIES

NURUL ASWA BINTI OMAR

UNIVERSITI TUN HUSSEIN ONN MALAYSIA

A HYBRID SEMANTIC SIMILARITY FEATURE-BASED TO SUPPORT MULTIPLE ONTOLOGIES

NURUL ASWA BINTI OMAR

A thesis submitted in
fulfillment of the requirement for the award of the
Doctor of Philosophy

Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia

AUGUST, 2017

In the name of Allah, The Most Beneficent, The Most Merciful.

Thank you Allah for giving me such wonderful people.

Deep appreciation to my beloved husband,
Beni Widarman Bin Yus Kelana

My children,
Muhammad Faris Aiman
Nur Fatihah Aleeya

My parents and father-in-law,
Omar Othman; Shakinah Yaakob; Yus Kelana; Noraini

and my siblings (Nurul Atiqah, Muhammad Syafiq, Nurul Athirah, Muhammad Syamil, Muhammad Syarafi).

Thank you for your prayers, understanding, caring, compromising and everything.

ACKNOWLEDGEMENT

Firstly, I would like to express my thanks to Assoc. Prof. Dr. Shahreen Binti Kasim, and Assoc. Prof. Dr. Mohd Farhan Bin Md Fudzee for their willingness to accept me as a Ph.D. student. Their guidance, support, determination, encouragement, understanding and patience along this journey is greatly appreciated.

I am also grateful to the Ministry of Higher Education (MOHE) for sponsoring me under the Scheme of Academic Training for Public Higher Education Institution (SLAI) during my studies.

I am also greatly indebted to the Faculty of Computer Science and Information Technology (FSKTM) and the Centre for Graduate Studies (CGS) of Universiti Tun Hussein Onn Malaysia (UTHM) for providing good facilities and an inspiring environment for me to complete this study comfortably.

I am dedicating special thanks to my friends, especially the postgraduate members (Dr. Isredza, Dr. Norhanifah, Norhanani, Norlida, Mahfuzah, Yana Mazwin and Dr. Yusliza) for their help, motivation, encouragement and knowledge sharing throughout this journey. Many thanks to Dr. Arfian Ismail, Dr. Sazali Khalid, Dr. Riswan Effendi for helping in completing my thesis journey.

Last but not least, I would like to thank my husband, my son, my parents, my father-in-law and my siblings for their prayers, support, understanding, compromise and everything.

ABSTRACT

Semantic similarity between concepts, words, and terms is of great importance in many applications dealing with textual data, such as Natural Language Processing (NLP). 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. Coupled with the fact that only one ontology is exploited, this leads to insufficient knowledge, missing 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 was conducted for developing the taxonomy of semantic similarity that contributes to understanding the current approaches, issues and data involved. This research aims to propose and evaluate ontological features for semantic similarity with multiple ontologies. Additionally, 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. This research used two benchmark datasets of biomedical concepts from Perdesen and Hliaoutakis. Similarity value, correlation and *p*-value were also used in the evaluation of the relationship between the concept pair of multiple ontologies. The findings indicate that the use of a semantic relationship of concepts (hypernym, hyponym, sister term and meronym) can improve the baseline method up to 75%. Besides that, the Hyb-TvX mechanism produces the highest correlation value compared to the other two methods, that is 0.759 and the result correlation is significant. Finally, the ability to discover similarity concepts with multiple ontologies could be also exploited in other domains besides biomedicine as future research.

ABSTRAK

Persamaan semantik antara konsep, perkataan dan terma adalah penting dalam pelbagai aplikasi yang berkaitan dengan data teks seperti pemprosesan bahasa tabii. Persamaan semantik merupakan satu pendekatan bagi mengenalpasti persamaan konsep melalui perbandingan makna. Persamaan semantik lebih cenderung menggunakan ontologi berdasarkan kecekapan, berskala, kurang kekangan serta mempunyai ontologi yang besar. Walaubagaimanapun, penggunaan ontologi di dalam persamaan semantik masih dibatasi oleh kebergantungan ontologi yang terperinci dan hanya satu ontologi diterokai yang menyebabkan ketidakcukupan pengetahuan, kehilangan terma dan ketidaktepatan persamaan. Permasalahan ini boleh diatasi dengan mengeksplotasi kepelbagaian ontologi. Persamaan semantik dari pelbagai ontologi berpontensi meningkatkan ketepatan persamaan dengan pengiraan persamaan dalam situasi kehilangan terma serta gabungan pelbagai sumber ontologi. Kajian ini dijalankan untuk membangunkan taksonomi persamaan semantik dalam memahami pendekatan semasa, isu dan data yang terlibat. Kajian ini bertujuan mencadangkan dan menilai ciri-ciri ontologi untuk persamaan semantik dari pelbagai ontologi. Disamping itu, kajian ini juga bertujuan untuk membangunkan dan menilai mekanisme berasaskan ciri-ciri (Hyb-TvX) bagi pengukuran persamaan semantik pelbagai ontologi dalam meningkatkan nilai ketepatan persamaan. Kajian ini telah menggunakan dua penanda aras set data bioperubatan konsep yang terdiri daripada Perdesen dan Hliaoutakis. Nilai persamaan, kolerasi dan nilai p juga digunakan bagi melihat perhubungan dan kepentingan hubungan antara konsep dari pelbagai ontologi. Dapatan kajian menunjukkan penggunaan perhubungan semantik konsep (hypernym, hyponym, sister term dan meronym) telah meningkatkan kolerasi sebanyak 75%. Selain itu, kaedah pengukuran Hyb-TvX menghasilkan nilai kolerasi yang tinggi berbanding dua kaedah sebelum ini iaitu 0.759 dengan keputusan kolerasi signifikan. Akhir sekali, penyelidikan persamaan pelbagai ontologi boleh dieksploitasi dalam bidang selain bioperubatan di masa depan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

LIST OF SYMBOLS AND ABBREVIATIONS

$\forall$	-	For all, any, each, every of element.
$\in$	-	Is element of
σ	-	Selection of
$\geq$	-	Greater than or equal to
A	-	Set A
B	-	Set B
o	-	Objects
$\times$	-	Cross
R	-	Real number
$=$	-	Similar to
x	-	Concept x
y	-	Concept y
$\cap$	-	Intersection
*	-	Multiplication
C_1	-	First concept
C_2	-	Second concept
C_3	-	Third concept
N_1	-	N_1 are the number of *is-a* links from A
N_2	-	N_2 are the number of *is-a* links from B
N_3	-	N_3 are the number of *is-a* links from C
LCS	-	Lower common subsume
$W(c)$	-	$W(c)$ is the set of words (nouns) in the corpus
c / C	-	Concept (c) or concept (C)
$\sum$	-	Total

N	-	The total number of word (noun) tokens in the corpus
$S(a,b)$	-	Similarity between concept *a* and *b*
$Sim(C_1,C_2)$	-	Similarity concept 1 and concept 2
$Sim(a,b)$	-	Similarity concept *a* and concept *b*
O	-	Ontology
O_1	-	First ontology
O_2	-	Second ontology
R_i	-	Type (*i*) of semantic relationship for first ontology
R_j	-	Type (*j*) of semantic relationship for second ontology
X	-	*X* correspondences to sets of *a*
Y	-	*Y* correspondences to sets of *b*
$X \cap Y$ / $A \cap B$	-	Set *X* union set *Y* *or* Set *A* union set *B*
$\|X - Y\|$	-	The relative complement of *Y* in *X*
$\|Y - X\|$	-	The relative complement of *X* in *Y*
$A \cup B$	-	Set *A* union set *B*
$1-\alpha > 0$	-	$1-\alpha$ must be more the zero
α	-	Parameter for complement
$depth(a^p)$	-	Depth of ontology for concept *a*
$depth(b^q)$	-	Depth of ontology for concept *b*
$\alpha(a^p,b^q)$	-	Parameter for complement *a* and *b*
$w_w, w_u, w_n \geq 0$	-	Weighting parameter must be same or more than zero
(S_w)	-	Similarity word matching
(S_u)	-	Similarity feature matching
(S_n)/ $S_{neighborhod}$	-	Similarity neighborhood
$S_p(a^p,b^q)$	-	Similarity parts for concept *a* and *b*
$S_f(a^p,b^q)$	-	Similarity function for concept *a* and *b*
$S_u(a^p,b^q)$	-	Similarity attribute or concept *a* and *b*
S_{synset}	-	Similarity synset or synonym
$S_{description}$	-	Similarity description
$Total_{hypo}$	-	All hyponym

$total_{hypo}O_i(S_i)$	-	All hyponym in ontology *i* for subsumer *i*
(*r*)	-	Pearson correlation coefficient
n	-	Number of word pairs
$(\sum x_i y_i)$	-	Total multiplication of human judgments (x_i) and y_i is the corresponding *i*th element in the list of similarity value
$(\sum x_i)$	-	Total value human judgment or physician judgment
$(\sum y_i)$	-	Total value similarity based similarity measurement method
M_R	-	Matrix relationship
M_S	-	Matrix subsume
R_i (S_i)	-	The row represent the subsumer of relationship
R_j (S_j)	-	The column represents the subsumer of relationship
Hyb-TvX	-	Proposed method (A Hybrid Semantic Similarity Feature-based Measurement)
TvX-1	-	Similarity Measurement level 1
TvX-2	-	Similarity Measurement level 2
(*Int*)	-	Intersection
(*Un*)/*U*	-	Union
(*comp*)	-	Complement
(*max*)	-	Take the maximum value of the words
$x \in A$	-	*x* element of set *A*
$x \in B$	-	x element of set *B*
comp A/ A'	-	Complement for set *A*
comp B/ B'	-	Complement for set *B*
$x \in U$	-	*x* element of set Union
$x \notin A$	-	*x* not element of set *A*
$(Int \lvert C_1, C_2 \rvert)$	-	Intersection for concept 1 and concept 2
$(\max \lvert C_1, C_2 \rvert)$	-	Maximum for concept 1 and concept 2
$(Int \lvert A, B \rvert)$	-	Intersection set *A* and set *B*
$(Un \lvert A, B \rvert)$	-	Union set *A* and set *B*

$S_c(C_1, C_2)$	-	Similarity concept for concept 1 and concept 2
$S_s(C_1, C_2)$	-	Similarity synonym for concept 1 and concept 2
$S_f(C_1, C_2)$	-	Similarity features for concept 1 and concept 2
(S_c)	-	Similarity concept
(S_s)	-	Similarity synonym
(S_f)	-	Similarity features
$\lfloor \max(S_c, S_s, S_f) \rfloor$	-	Identify value maximum between S_c, S_s, S_f
(w_a)	-	The proposed parameter for |*comp A*|
(w_b)	-	The proposed parameter for |*comp B*|
GB	-	Gigabyte
RAM	-	Random Access Memory
Ghz	-	Gigahertz
PHP	-	Hypertext preprocessor
GIS	-	Geographic information systems
STS	-	Semantic textual similarity
WordNet	-	Ontology WordNet
Snomed-CT	-	Systemized Nomenclature of Medicine Clinical Term
MeSH	-	Medical Subject Heading
GO	-	Gene ontology
IC	-	Information content
UMLS	-	Unified Medical Language System
ICD	-	International Clasification Disease
S	-	Synonym
SENSUS	-	Ontology SENSUS
Cyc KB	-	Cyc knowledge base
KB	-	Knowledge base
CPT	-	Current procedural terminology
ICD-10-CM	-	International Classification of Diseases, Tenth Revision, Clinical Modification
LOINC	-	Logical Observation Identifiers Names and Codes

NLM	-	National Library of Medicine
MH	-	Mesh Heading
STDS	-	Spatial Data Transfer Standard
LCA	-	Lower common ancestor
STS	-	Semantic textual similarity
RM	-	Root Matching
TM	-	Terminological Matching
TM	-	Terminological Subsumption
SS	-	Semantic Subsumption
LCS	-	Lower common subsume

LIST OF APPENDICES

LIST OF PUBLICATIONS

Journal:

(i) Nurul Aswa Omar, Shahreen Kasim, Mohd Farhan Md Fudzee (2016) (*in press*) "A Review of Semantic Similarity Approach for Multiple Ontologies." International Journal of Information and Decision Sciences. (Scopus Indexed)

(ii) Nurul Aswa Bt Omar, Shahreen Kasim, Mohd Farhan Md Fudzee (2016) (*in press*) "Extended TvX: A New Method Feature Based Semantic Similarity for Multiple Ontology." Journal of Telecommunication, Electronic and Computer Engineering (JTEC) Special Issue (Scopus Indexed).

CHAPTER 1

INTRODUCTION

1.1 Overview

The birth of the internet has led to an abundance of textual information in the World Wide Web. By using different language terminologies, people can access information from various sources in several formats (mostly text). However, resources are hard to manage due to the lack of textual understanding capabilities of computerized systems. The estimation of the semantic similarity between concepts, words, and terms is of great importance in many applications dealing with textual data, such as natural language processing (NLP) (Patwardhan & Pedersen, 2006), knowledge acquisition (Sánchez & Batet, 2011a), information retrieval (Al-Mubaid & Nguyen, 2006; Budanitsky & Hirst, 2006b; Mohameth-François *et al.*, 2012; Pedersen *et al.*, 2007), information integration (Petrakis *et al.*, 2006; Rodríguez & Egenhofer, 2003b) information extraction (IE) (Sánchez *et al.*, 2011; Vicient *et al.*, 2013) and knowledge-based systems (Al-Mubaid & Nguyen, 2009).

Similarity is derived from the word similar (adjective), similarity (noun), similitude (noun) and from the latin, similes, that is defined by like, resembling, or similar (Harper, 2016). In mathematics, the word may be encountered in several different contexts. In algebra, the terms that contain the same power of the variables involved are said to be like, or have similar terms. Terms that are not similar are called dissimilar. In geometry, two figures are said to be similar if they have the same shape, though not necessarily the same size (Schwartzman, 1996).

The most popular method used to compare two concepts is via similarity. In this method, similarity is used as a precondition to create interoperability between agents or words by using different sources (Ehrig & Staab, 2004). Similarity using semantics (semantic similarity) identifies similarity based on the likeliness of their meaning or their related information (Yuhua *et al.,* 2003; Martinez-Gil & Aldana-Montes, 2013). Semantic similarity improves the understanding of textual resources and increases the accuracy of knowledge-based applications (Jiang *et al.,* 2015).

Most items dealing with semantic similarity have been developed using taxonomies and ontologies (Batet *et al.*, 2013; Budanitsky & Hirst, 2006; Couto *et al.*,2007; Cross *et al.*, 2013; Liu *et al.*, 2012; Rodriguez & Egenhofer, 2003; Sánchez & Batet, 2013; Sánchez *et al.*,2010; Sánchez *et al.*, 2012). Ontology is of great interest to the semantic similarity research community as it offers a formal specification to a shared conceptualization. Several approaches to ontology-based semantic similarity have been proposed. The approaches can be classified into a structure-based approach, an information content-based approach, a feature-based approach and a hybrid-approach. They are used either in single ontology or multiple ontologies (Elavarasi *et al.*, 2014; Saruladha *et al.*, 2011a). Despite these approaches, most semantic similarities are used to compute the similarity between concepts within a single ontology and are rarely used in multiple ontologies (Rodríguez & Egenhofer, 2003b, Petrakis *et al.*, 2006, Batet *et al.*, 2013). The use of a single ontology does not ensure complete integration across a heterogeneous knowledge system. With an increasing problem in integrating heterogeneous knowledge sources, it is more dire that semantic similarity via multiple ontologies is studied. The exploitation of multiple ontologies would also provide additional knowledge that can improve similarity estimation and solve cases where terms are not represented in an individual ontology. This is especially interesting within domains of knowledge such as biomedicine where several big and detailed ontologies are available and offer overlapping and complementary knowledge to similar topics (Al-Mubaid & Nguyen, 2009).

The following section in this chapter discusses the problem background in semantic similarity. The research objectives and research questions for each objective are explained and outlined followed by the scope and significance of this research and the overview of the organization of this thesis.

1.2 Problem statement

Most semantic similarity methods are used to compute the semantic distance between concepts within a single ontology, but is rarely used for multiple ontologies. Semantic similarity in multiple ontologies is necessary for information integration and retrieval. However, in order to relate to this, several factors need to be considered. The first factor defines the scope of the matching problem (Sánchez *et al.*, 2012). Each ontology is built according to the expertise of the engineer's point of view which leads to the heterogeneity of the ontology. Ontological concepts rarely constitute a perfect fit because of this heterogeneity and lack of consensus. In order to integrate different knowledge on ontology, one has to consider that ontologies may represent the same knowledge within the ambiguity of language. Most related works rely on terminological matching of the concept label. However, this approach tends to underestimate the real similarity between concepts due to differentiating text labels. Moreover, since identifying the concept heavily relies on terminological matching, sometimes concepts with differing labels that have the exact same definitions and the potential to obtain equivalent concepts may be omitted because the commonality concept may not have been properly evaluated. Figure 1.1 shows the problem for this type of situation. For example, the concepts compared are *antibiotic* in WordNet and *anti-bacterial agent* in MeSH. The identical terminology for the paired concepts *antibiotic* and *anti-bacterial agent* is the concept of '*Drug*'. However, this terminology underestimates the real similarity concept. This similarity concept pair is closer to *bactericide* in WordNet and *anti-bacterial agent* in MeSH. The '*Drug*' concept is more of a general subsumer compared to both *bactericide* and *anti-bacterial agent*. The second factor in dealing with ontology integration is to find the similar equivalents of the concept that can act as the least common subsumer (LCS). Previous research has used the taxonomic ancestry to identify the LCS (Batet *et al.*, 2012). However, this approach's limitation is that it only uses the concept of ancestry to find the LCS.

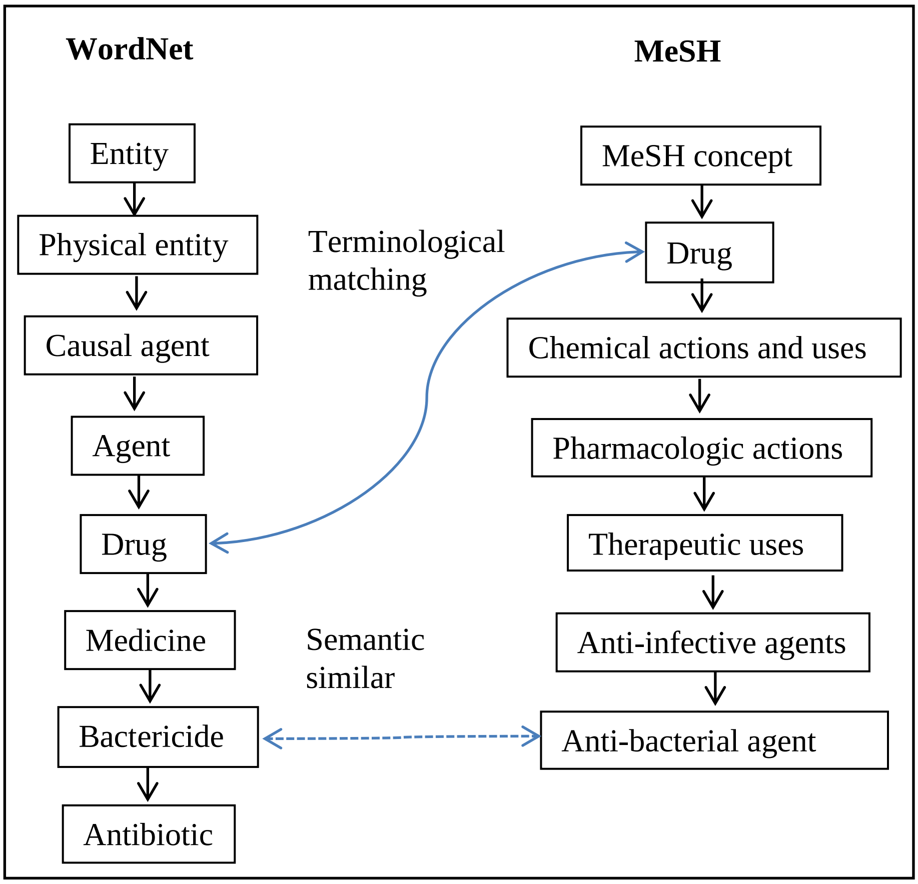


Figure 1.1: Example of terminological matching between subsumer of antibiotic (WordNet) and antibacterial agent (MeSH) (source: Solé-Ribalta *et al.*, 2014).

The third factor is related to the accuracy of similarity. The measurement of similarity plays a crucial role in determining the similarities between concepts (Sánchez *et al.*, 2012). The similarities between concepts can create accurate information. The ontology structure has been widely employed in similarity measurement, especially on the semantic similarity single ontology, but not in multiple ontologies. The semantic similarity in multiple ontologies inappropriately uses the ontology structure in similarity measurements because the concept pair consists of two different ontologies that have different structures whereas the structure ontology cannot be compared with directly (Petrakis *et al.*, 2006). Besides that, the measurement structure-based ontology generally considers the shortest path between concept pairs. Consequentially, it is unsuitable for wide and detailed ontologies such as WordNet. As a result, several taxonomical paths are not taken into consideration. Besides that, with the use of structure ontology, other features are omitted as these features influence the semantic concept (e.g. common and non-common concepts). 'Feature-based' is an approach that overcomes the limitation of structure ontology. The method of this approach has more potential use in similarity measurements of multiple ontologies. This method considers measurements between sets of features (Sánchez *et al.*, 2012). However, some weaknesses were found in this approach as the measurements are very limited in their applicability ontology wherein this information is available. Another problem is that this approach depends on the weighting parameter that balances the contribution of each feature (Rodríguez & Egenhofer, 2003b). Only Petrakis *et al.*, (2006) did not depend on the weighting parameters. The maximum value is taken when the similarity synonym is more than zero (similarity synonym > 0 =1). Due to this, the contribution resulting from other features are omitted as sometimes, this feature has high potential in similarity measurement. Besides that, by assuming a similarity value of more than zero to one, this method will yield an unreliable result.

1.3 Research objectives

The goal of this research is to develop a feature-based semantic similarity in order to identify concepts that are similar from multiple ontologies. This research covers:

(i) To develop the taxonomy of semantic similarity that contributes understanding towards current approaches, issues and data related to the topic.

(ii) To propose and evaluate a ontological feature algorithm in semantic similarity for multiple ontologies in terms of feature-matching accuracy.

(iii) To develop and evaluate a feature-based mechanism to measure semantic similarity of multiple ontologies to increase the accuracy of similarity.

1.4 Research questions

Some research questions were developed based on the research objectives:

The research question for research objective 1:

(i) What is the appropriate approach for semantic similarity with multiple ontologies in order to understand the issues and data involved?

Research questions for research objective 2

(i) Does the feature-based measurement used is appropriate in measuring the ontological features algorithm proposed?

(ii) Does the proposed ontological features algorithm increase the accuracy of feature matching?

(iii) Is there any relationship between dependent variable (human scored) and independent variables (similarity values for each method)?

Research questions for research objective 3

(i) How do the proposed parameters support in improving the accuracy of similarity for multiple ontologies?

(ii) Does it a significant relationship between the Hyb-TvX methods with human scored?

1.5 Scope and research significance

This research, focused on the semantic similarity between multiple ontologies. The datasets used were WordNet and MeSH. The WordNet dataset was downloaded from https://wordnet.princeton.edu whereas the MeSH dataset was downloaded from http://www.nlm.nih.gov/mesh/MBrowser.html. This research will try to improve the ontological features algorithm. The ontological features algorithm consists of matching matrix semantic relationships where this component has a matching matrix of a subsumer. The matching matrix semantic relationships uses the semantic relationship matrix which is divided into four partitions; hypernym, hyponym, sister term and meronym/holonym. After the semantic relationship matrix is defined, the subsumer of each relationship in the phase matching matrix of subsumer is identified. Besides that, the ontological features algorithm also includes the semantic overlapping between subsumers where a matching matrix of subsumer computes to subsumers according to the number of hyponyms that they share. The last phase is selecting the most suitable and least common subsumer (LCS).

This research proposed the Hyb-TvX which is a hybrid of the X-similarity and Tversky methods. The proposed Hyb-TvX method proved its capability in dataets for different ontologies in the biomedical domain. The Hyb-TvX consists of two components: similarity measurement level 1 (TvX-1) and similarity measurement level 2 (TvX-2). The TvX-1, has two phases: (i) calculation of concept to find the similarity between concepts and (ii) calculation of synonyms to find the similarity synonym of each concept. The TvX-2 has two phases: (i) calculation of features to find a similarity concept through features and (ii) normalization of all calculations to find the similarity value for that concept. The evaluation measurement of Hyb-TvX encompasses the similarity of each concept pair as compared with the previous similarity method, the measurement evaluation (correlation) and the *p*-value. The focus of this research is to develop a better semantic similarity measurement method in order to obtain optimum accuracy.

This research could contribute to the "body of knowledge" in a feature-based semantic similarity to support multiple ontologies and the biomedical domain. This research also contributes to the improvement of statistical publications and citations for a number of articles in the areas of research. Through the web of science (12 January 2017), Figure 1.2 shows that the research semantic similarity for multiple ontologies and the feature-based semantic similarity is growing in importance and subsequently gaining more attention from other researchers around the world. However, research in the field of semantic similarity feature-based approach using multiple ontologies still needs to be explored by the researcher.

Furthermore, the results of this research could also act as a guide for expanded research in the field of semantic similarity with multiple ontologies using a feature-based approach that has, only previously received little attention by researchers. This research is also able to resolve the issues in the terminological matching method with the developed ontological features algorithm. An analysis of the similarity value and correlation indirectly provided a clear picture that the proposed ontological features algorithm can improve the similarity and correlation values that, in turn, can be used as empirical evidence. The issue of a similarity measurement that depends on the weighting parameter that balances the contribution of each feature (Rodríguez & Egenhofer, 2003b) and other features that are omitted can be addressed by developing the proposed method (Hyb-TvX). Alongside this, an analysis of the similarity, correlation and *p*-value will also indirectly confirm that the proposed method (Hyb-TvX) can increase the similarity, correlation and would be capable of showing a significant correlation. Overall, this analysis aims to obtain optimum accuracy of similarity between the concepts.

Additionally, the results of semantic similarity using the biomedical domain is important in assisting with the integration of heterogeneous clinical data such as clinical records with different formats. This research can improve the interoperability between medical sources, which are commonly dispersed and standalone. Therefore, it can assist in searching within the biomedical domain and ensuring that such searches are accurate.

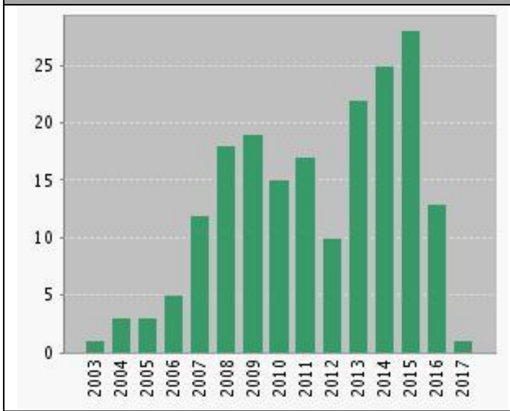
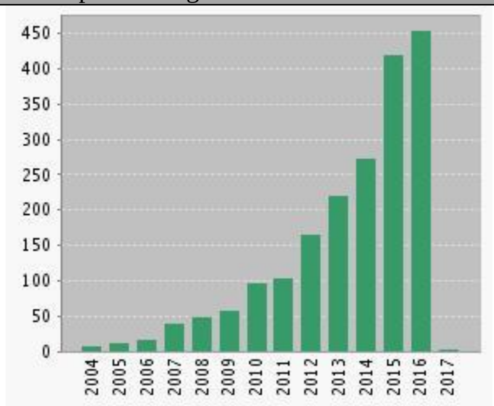
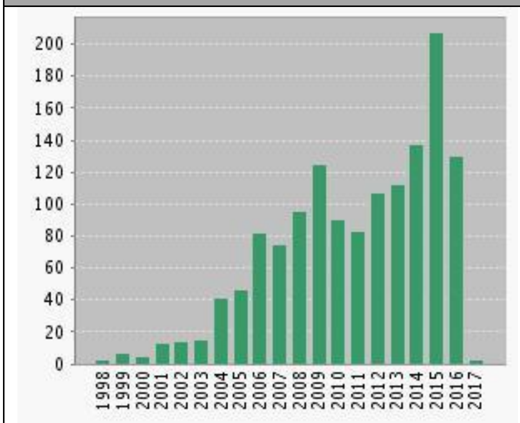
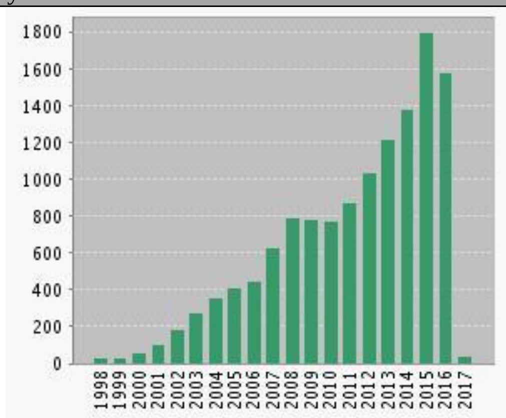
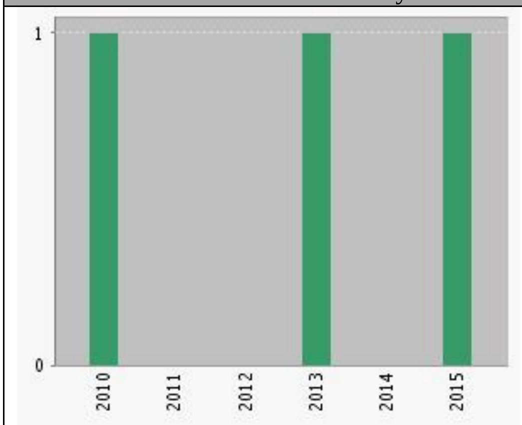
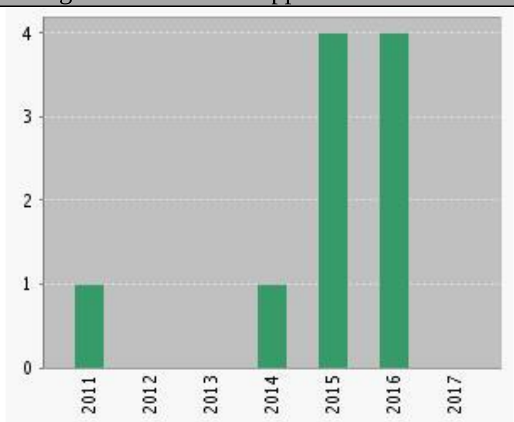
The number of articles published per year	The number of citations per year
Semantic similarity with multiple ontologies	
	
Semantic similarity feature-based	
	
Semantic similarity with multiple ontologies feature-based approach	
	

Figure 1.2: Comparison of the number of articles published every year and the number of citations per year (http://apps.webofknowledge.com)

1.6 Organization of the thesis

This thesis is organized in six chapters. Chapter 1 describes the introduction, problem statements, research objectives, research questions, scope and significance of this research. The rest of the thesis is organized into the following chapters. Chapter 2 reviews the main subjects used in the research which includes similarity, the approach of measurement analysis for semantic similarity and a feature-based semantic similarity approach for multiple ontologies. Chapter 3 describes the design of the semantic similarity method adopted to achieve the objectives of the research. This includes the research framework, data sources, instrumentation and result analysis. Chapter 4 highlights the development of the ontological features algorithm that aims to find the correspondence between compared concepts. This algorithm will describe the ontological features used in the similarity measurement method in the next chapter. Chapter 5 further elaborates the development of the feature-based Hyb-TvX method that utilizes a hybrid X-similarity and Tversky method. This method also includes ontological features and the use of a semantic relationship of concepts such as: hypernym, hyponym, sister term and meronym/holonym. This chapter also elaborates on the evaluation of Hyb-TvX with other methods in feature-based approaches. Finally, chapter 6 presents the general conclusions of the research, the contribution and proposed topics for future research.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter strives to give a better understanding of similarity while reviewing several works on semantic similarity. This chapter also explores the ontological features matching method for enabling semantic similarity from multiple ontologies. Furthermore, approaches to the measurement analysis of semantic similarity for multiple ontologies are discussed. The current research trends and directions are outlined before presenting the summary of this chapter. The content structure is represented in Figure 2.1.

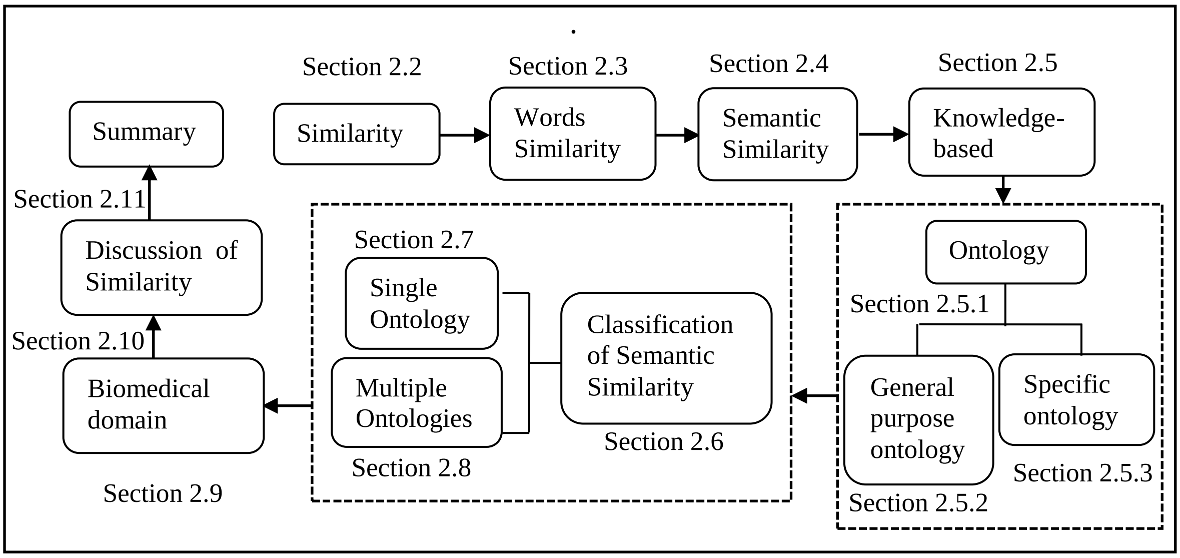


Figure 2.1: Content structure

2.2 Similarity

The first issue addresses the meaning of similarity. Generally, similarity is a quality or condition of being similar. However, many different definitions of similarity are possible, each being appropriate for specific and particular situations. This statement is also supported by Lin (1998a), who opines that the definition of similarity is normally according to the application and representation of knowledge. Similarity is also defined as a basis in making predictions, because similar things usually behave similarly (Quine, 1969). According to Lin (1998a), three formal definitions of the concept of similarity exist:

(i) Definition by Concept 1: Similarity between *x* and *y* relates to their commonality. The more commonality they share; they are more similar.

(ii) Definition by Concept 2: Similarity between *x* and *y* relates to the differences between them. The more differences they have, they are less similar.

(iii) Definition by Concept 3: The maximum similarity between *x* and *y* is reached when *x* and *y* are identical, no matter how much commonality they share.

Definition 2.1 (Similarity): A similarity $\sigma : o \times o \rightarrow R$ is a function from a pair of entities to real number (*R*) expressing the similarity between two objects (*o* and *o*). Table 2.1 shows the symbols that describe the meaning of similarity according to definition 2.1:

$$\forall x, y \in o, \sigma(x, y) \geq 0 \; (positiveness)$$

$$\forall x \in o, \forall y, z \in o, \sigma(x, x) \geq \sigma(y, z) \; (\max imality)$$

$$\forall x, y \in o, \sigma(x, y) = \sigma(y, x) \; (symmetry)$$

Table 2.1: Symbol and meaning in relation to similarity

Symbol	Meaning
$\forall$	for all, for any, for each, for every
$\in$	is an element of, is not an element of
σ	selection of
$\geq$	is less than or equal to, is greater than or equal to

2.3 Words similarity

In many fields of research on similarity, the use of the word has solved a lot of problems and has also given benefits in associated fields such as text categorization (Ko *et al.*, 2004), text summarization (Erkan & Radev, 2004; Lin & Hovy, 2003), word sense disambiguation (Lesk, 1986; Schütze, 1998), automatic evaluation of machine translation (Liu & Zong, 2004; Papineni *et al.*, 2002), evaluation of text coherence (Lapata & Barzilay, 2005; Wegrzyn-Wolska & Szczepaniak, 2005) and the classification of formatted documents (Wegrzyn-Wolska & Szczepaniak, 2005). Cohen (2000) stated that word similarity is vital in the retrieval of images from the web. This can improve the retrieval of images by utilising informative words.

Word similarity is used as the primary stage to assess the similarities between sentence, paragraph and document. This statement is supported by Lin (1998b) where similarity between two documents can be calculated by comparing the sets of concept in the documents or by comparing their stylistic parameter values, such as average word length, average sentence length, and average number of verbs per sentence. Several studies have assumed that words which are close in meaning will occur in similar pieces of text and context (Gomaa & Fahmy, 2013; Kolb, 2009; Landauer & Dumais, 1997; Lin, 1998b).

Word similarity in the context databases can be used in schema matching to solve semantic heterogeneity. The main problem regarding similarity in the context of a database is the data sharing system; whether it is a federated database, a data integration system, a message passing system, a web service, or a peer-to-peer data management system (Madhavan *et al.*, 2005). Word similarity also involves a joint operator as it joins two relations if their attributes are textually similar to each other.

Besides that, it also has a variety of application domains including integration and querying of data from heterogeneous resources, cleansing of data and mining of data (Cohen, 2000; Islam & Inkpen, 2008; Schallehn *et al.,* 2004). From the current studies, there are two approaches for word similarity. The first approach is similarity from a lexical perspective and the second approach is from the semantic approach.

The lexical approach is referred to as words that are similar if they have a similar character sequence. In this research, the lexical approach is introduced through the different string-based method. The string-based method works on string sequences and character structure (Bernstein & Rahm, 2001). This method typically finds the concepts of '*Book'* and '*Textbook'* to be similar, but not the concepts of *'Book'* and *'Volume'*. The string-based method is often used to match names and their description. This method assumes that the more similar the string, the more likely they are in representing the same concept (Bernstein & Rahm, 2001; Gomaa & Fahmy, 2013). There are many ways to compare the string depending on the way the string is viewed, for example as an exact sequence of letters, a set of letters, and a set of words (Euzenat & Shvaiko, 2007). There are two approaches of the string-based method identified in this research: character-based and term-based.

(i) The character-based approach considers distance as the difference between the characters. This is useful in the case of typographical errors. Among the works that have used this type include the Longest Common Substring (Allison & Dix, 1986), Damerau (1964), Jaro (1995), Winkler (1990), Needleman & Wunsch (1970), Smith & Waterman (1981) and N-gram (Kondrak, 2005).

(ii) The term-based approach is a similarity measure which incorporates the linguistic and semantic structures using syntactic dependencies. This type comes from information retrieval and considers a string as a multi set of words. These approaches usually work well on long texts (comprising of many words). This approach can also adapt to ontology concepts such as aggregating different sources of string, example identifiers, labels, comments and documentation. Besides that, this approach can also split the string into independent tokens. For example, '*renal failure*' becomes '*renal*' and '*failure*'. Examples of concepts that used this approach are Jaccard (1901) and Dice (1945).

The second approach is semantic. Semantic refers to words that can be similar if they have the same thing, used in the same way and in the same context (Gomaa & Fahmy, 2013). There are two types identified in this semantic: similarity and relatedness.

(i) Similarity or better known as semantic similarity is a comparison among entities/terms/concepts. This semantic similarity allows information retrieval and information integration to handle concepts that are semantically similar. Example of works that have used semantics are Resnik (1995), Jiang & Conrath (1997), Palmer & Wu (1994), Leacock & Chodorow (1998), Rodríguez & Egenhofer (2003a), Saruladha *et al.,* (2011b) and Sánchez *et al.*, (2012).

(ii) Relatedness, or otherwise known as semantic relatedness is a more general notion of relatedness, not specifically tied to the shape or form of the concept and they are not limited to considering *is-a* relations (Gomaa & Fahmy, 2013). Among works that have used this type are the *hso measure* (Hirst & St-Onge, 1998), *lesk* (Banerjee & Pedersen, 2002) and *vector pair* (Patwardhan & Pedersen, 2006).

Based on the number of reviews, the lexical approach is an easy method because it measures string sequences and character composition. However, a limitation on this approach exists, where it is considered as a simple traditional method in determining the similarity of concepts (Islam & Inkpen, 2008). This approach cannot identify the semantic similarity of concept. For instance, with the similarity between the concepts of *bactericide* and *anti-bacterial agent,* the current word similarity is unsuccessful in identifying any kind of connection between these words. The limitation of the lexical approach can be overcome by using the semantic approach. The semantic approach does not rely on the string sequence, but also measures similarity based on the likeness between words. This approach can identify the same meaning in different words. However, this approach needs an evaluation of the semantic evidence observed in knowledge sources such as ontologies or domain corpora. Most knowledge sources are presented in unprocessed and heterogeneous textual formats (Batet *et al.,* 2011). Table 2.2 displays the conclusion of words similarity approach as previously described.

Table 2.2: Words similarity approach

Approach	Techniques	Example Research	Advantages	Disadvantages
Lexical	Character-based	Allison & Dix (1986); Damerau (1964); Jaro (1995); Winkler (1990); Needleman & Wunsch (1970); Smith & Waterman (1981) and Kondrak (2005).	Easy method because this approach measures on the string sequences and character composition. It is useful if use very similar string to denote the same concepts. These approaches are most often used in order to detect very similar string used.	Different concepts with different structure characters are used, this will yield a low similarity. Concept pair with low similarity in turn yields many false positive.
	Term-based	Jaccard (1901); Dice (1945)		
Semantic	Similarity	Resnik (1995); Jiang & Conrath (1997); Palmer & Wu (1994); Leacock & Chodorow (1998); Rodríguez & Egenhofer (2003a); Saruladha *et al.*, (2011b) and Sánchez *et al.*, (2012)	This approach does not depend on the sequence of string. Semantic similarity computes the likeness between words, understood as the degree of taxonomical proximity.	Need evaluation of the semantic evidence observed in knowledge source. Knowledge source presented in unprocessed and heterogeneous textual formats
	Relatedness	Hirst & St-Onge (1998); Banerjee & Pedersen (2002) and Patwardhan & Pedersen (2006).		

2.4 Semantic similarity

Basically, semantic similarity is the quality or condition of being similar. The differences being purely dependant on certain situations. Doan *et al.,* (2004) mentioned that semantic similarity can be defined based on the joint probability distribution of the concepts involved. According to Elavarasi *et al.,* (2014) semantic similarity is defined as the closeness of two concepts based on the likeliness of their meaning, which refers to the similarity between two concepts in a taxonomy or ontology. Besides that, according to Jiang *et al.,* (2015) semantic similarity relates to computing the similarity between concepts, words, terms or short text expressions, where the concepts have the same meaning or have relatively matching information despite not being lexically similar (Martinez-Gil & Aldana-Montes, 2013; Yuhua *et al.,* 2003).

According to Schwering (2008), semantic similarity is central for the proper function of semantically enabled processing of geospatial data. It is used to measure the degree of potential semantic interoperability between data or geographic information systems (GIS). This semantic similarity is used to deal with vague data queries, vague concepts or natural language. This is also supported by Zhang *et al.,* (2015) whom indicated that semantic similarity is the degree of semantic equivalence between two linguistic items, where the items can be concepts, sentences or documents. However, semantic similarity between sentences or documents can also be known as semantic textual similarity (STS).

Semantic similarity has been used for years in psychology and cognitive science where different models have been proposed (Pirró & Euzenat, 2010). Besides that, semantic similarity has also been applied in searching for similarities between images and visual media (Deselaers & Ferrari, 2011). However, in recent years, semantic similarity is widely used in obtaining similarities between concepts or words where it assists in information extraction tasks (Sánchez & Isern, 2011) such as semantic annotation (Sánchez *et al.,* 2011), and ontology learning (Iannone *et al.,* 2007).

According to Schwering (2008), semantic similarity is also widely used in information retrieval tasks (Al-Mubaid & Nguyen, 2009; Budanitsky & Hirst, 2006b;

Saruladha *et al.,* 2011a) to improve the performance of current search engines (Hliaoutakis *et al.,* 2006), information integration (Saruladha *et al.,* 2011a), ontology matching (Pirrò *et al.,* 2009, Saruladha *et al.,* 2011a), semantic query routing, and bioinformatics to assess the similarity between proteins (Wang *et al.,* 2007). Additionally, semantic similarity can also play an important role in both predicting and validating gene product interactions and interaction networks (Pesquita *et al.,* 2009). Table 2.3 shows the field and application that utilises the semantic similarity approach.

Table 2.3: Fields and applications that use the semantic similarity approach

<table>
<tr><th>Approach</th><th>Field</th><th>Application</th><th>Reference</th></tr>
<tr><td rowspan="7">Semantic Similarity</td><td rowspan="2">Information extraction</td><td>Semantic annotation</td><td rowspan="2">Sánchez & Isern, (2011); Sánchez et al., (2011); Iannone et al., (2007)</td></tr>
<tr><td>Ontology learning</td></tr>
<tr><td rowspan="5">Information retrieval</td><td>Performance search engine</td><td rowspan="5">Al-Mubaid & Nguyen (2009); Budanitsky & Hirst (2006b); Saruladha et al., (2011a); Hliaoutakis et al., (2006); Pirrò et al., (2009); Wang et al., (2007)</td></tr>
<tr><td>Information integration</td></tr>
<tr><td>Ontology matching</td></tr>
<tr><td>Semantic query rating</td></tr>
<tr><td>Bioinformatic</td></tr>
</table>

Semantic similarity is generally based on certain background information (Maind *et al.,* 2012). Two background information are identified in semantic similarity: structured information and unstructured information (Abdelrahman & Kayed, 2015).

(i) Structured information is often in a hierarchical form that is known as knowledge-based such as the WordNet (Miller, 1995), MeSH (Hliaoutakis *et al.,* 2006), the Systemized Nomenclature of Medicine Clinical Term (Snomed-CT) (Garla & Brandt, 2012), Wikipedia (Gabrilovich & Markovitch, 2007) and Gene Ontology (GO) (Ashburner *et al.,* 2000). Similarity measurement from knowledge-based determines the degree of similarity between texts using information derived from semantic networks.

Examples of similarity using knowledge-based are found in Resnik (1995), Jiang & Conrath (1997), Palmer & Wu (1994) and Leacock & Chodorow (1998).

(ii) Unstructured information refers to corpus-based, which are a collection of texts. Some examples that use a corpus-based includes the Brown Corpus and Wall Street Journals (Zhang *et al.,* 2015). Similarity measurement from corpus-based defines the similarity between texts as dependant on the information gained from the large corpora. A corpus is a large collection of written or spoken texts that is used for language research. There are several studies using corpus-based; for example, Gabrilovich & Markovitch (2007), Landauer & Dumais (1997) and Lund *et al.,* (1995).

Based on a number of reviews, the corpus-based unstructured information represents the semantic of words by distribution in large multilingual corpora. These unstructured information rely on the assumption that related words exist in the same document (Aggarwal, 2012). However, similarity that uses corpus-based performs well in document similarity, but needs further improvement for short text or phrases. Knowledge-based is one potential issue in semantic similarity, especially for applications dealing with textual data (Batet *et al.,* 2014; Pirró & Euzenat, 2010) where semantic similarity measurement between words provides a valuable tool to the understanding of textual resources (Sánchez *et al.,* 2012). According to Batet & Sánchez (2014), semantic similarity measurement that uses knowledge-based is able to capture the semantics inherent to the knowledge modelled in ontology. Two types of knowledge are exploited; (i) explicit knowledge such as the structure of a taxonomy and (ii) implicit knowledge such as information distribution (Sánchez *et al.,* 2010). However, knowledge-based semantic similarity needs a proper understanding of the semantics concept. Encouraging an improved use and integration of heterogeneous sources as well as higher information retrieval accuracy (Saleena & Srivatsa, 2014) are some of the ways that will help improve understanding. Table 2.4 depicts the advantages and disadvantages based on the background information of the semantic similarity approach.

Table 2.4: Advantages and disadvantages based on the background information in the semantic similarity approach

Approach	Background Information	Advantage	Disadvantage
Semantic similarity	Structure Information (knowledge-based)	Determines the degree of similarity between texts using information derived from semantic networks. Capable to capture the semantics inherent to the knowledge model led in ontology.	Need proper understanding of concept semantics
	Unstructured Information (corpus-based)	Similarity used corpus-based perform well in document similarity.	Rely on the assumption that related words exist in the same document. Need to improve for short text or phrases.

2.5 Knowledge-based

Knowledge-based similarity determines the degree of similarity between concepts using information derived from semantic networks. WordNet (Fellbaum, 1998) is the most common semantic network in the area of measuring the knowledge-based similarity between concepts. Knowledge-based uses ontology when calculating similarity (Batet, *et al.*, 2013). The reason ontologies are so popular is due, in large part, to what they promise: a shared and a common understanding of some domains that can be communicated across people and computers (Studer *et al.*, 1998). Ontology is a type of knowledge-based. Ontology describes concepts through definitions that are sufficiently detailed to capture the semantics of a domain. Ontologies are widely used to enrich the semantics of the web (Alasoud, 2009).

2.5.1 Ontology-based

"An ontology is defined as a formal, explicit specification of a shared conceptualisation" which means that ontology is defined as a formal representation of concepts within a domain and the relationship between those concepts (Studer *et al.*, 1998). Ontology is an effective way to share knowledge within controlled and structured vocabulary (Spasic *et al.*, 2005). Many ontologies have been developed for various purposes and domains (Al-Mubaid & Nguyen, 2009; Hliaoutakis, 2005, Miller, 1995). Furthermore, in reference to Noy & McGuinness (2001) ontology is built for some reasons such as sharing a common understanding of the structure of information among people or software agents, enabling the reuse of domain knowledge, making explicit domain assumptions, separating the domain knowledge from the operational knowledge and analysing domain knowledge. Besides that, ontology is also crucial in enabling interoperability across heterogeneous systems and semantic web applications (Choi *et al.*, 2006).

Ontology contains concepts, the definitions of these concepts, and rich relationships among these concepts. Consider the computer ontology example shown in Figure 2.2.

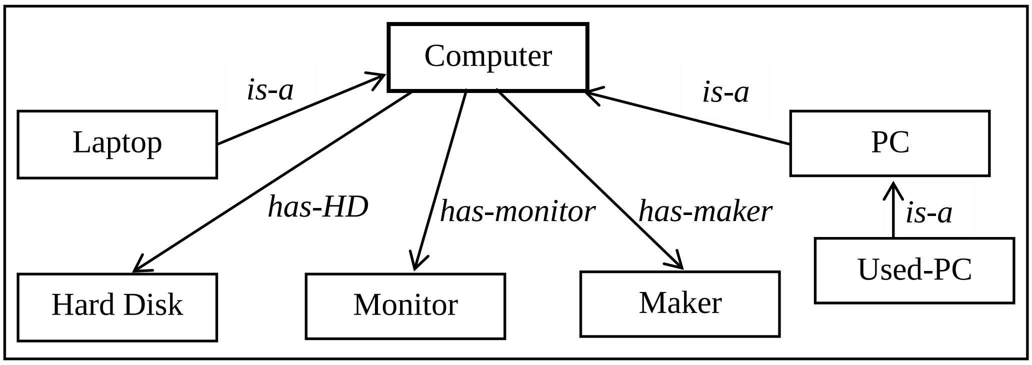


Figure 2.2: The computer ontology (Alasoud, 2009)

Three basic components of ontology are:

(i) Concept or classes

These are concepts of the domain or task, usually organised in taxonomies. In our ontology example, Computer, PC, Laptop, Hard Disk, etc. are examples of the concept.

(ii) Roles or properties

These are the types of interaction between instances of concepts in the domain. For example, has-HD (has Hard Disk), has-monitor, and has-maker are roles which are shown in Figure 2.2.

(iii) Individual or Instances

Individuals or instances represent specific elements.

Ontology is a type of knowledge-based that describes concepts through definitions that are sufficiently detailed to capture the semantics of a domain. A few ontologies such as the WordNet (Miller, 1995) have been used for semantic similarity. The WordNet is a lexical database for general English covering most generic English concepts and supports various purposes. Besides that, other ontologies are also used for the same purpose as the Unified Medical Language System (UMLS), that includes many biomedical ontologies and terminologies (e.g., MeSH, Snomed-CT) (Saruladha *et al.,* 2011b), and the International Classification Disease (ICD) family (Al-Mubaid & Nguyen, 2009). These ontologies are specifically created for the biomedical domain that is different from WordNet.

Ontology-based semantic similarity is used in two situations. The semantic similarity in a single ontology and when multiple ontologies are involved.

(i) Single ontology means similarities are compared from the same ontology, an example is illustrated in Figure 2.3.

(ii) Multiple ontologies mean that the similarity concepts are compared from different ontologies, example is illustrated in Figure 2.4.

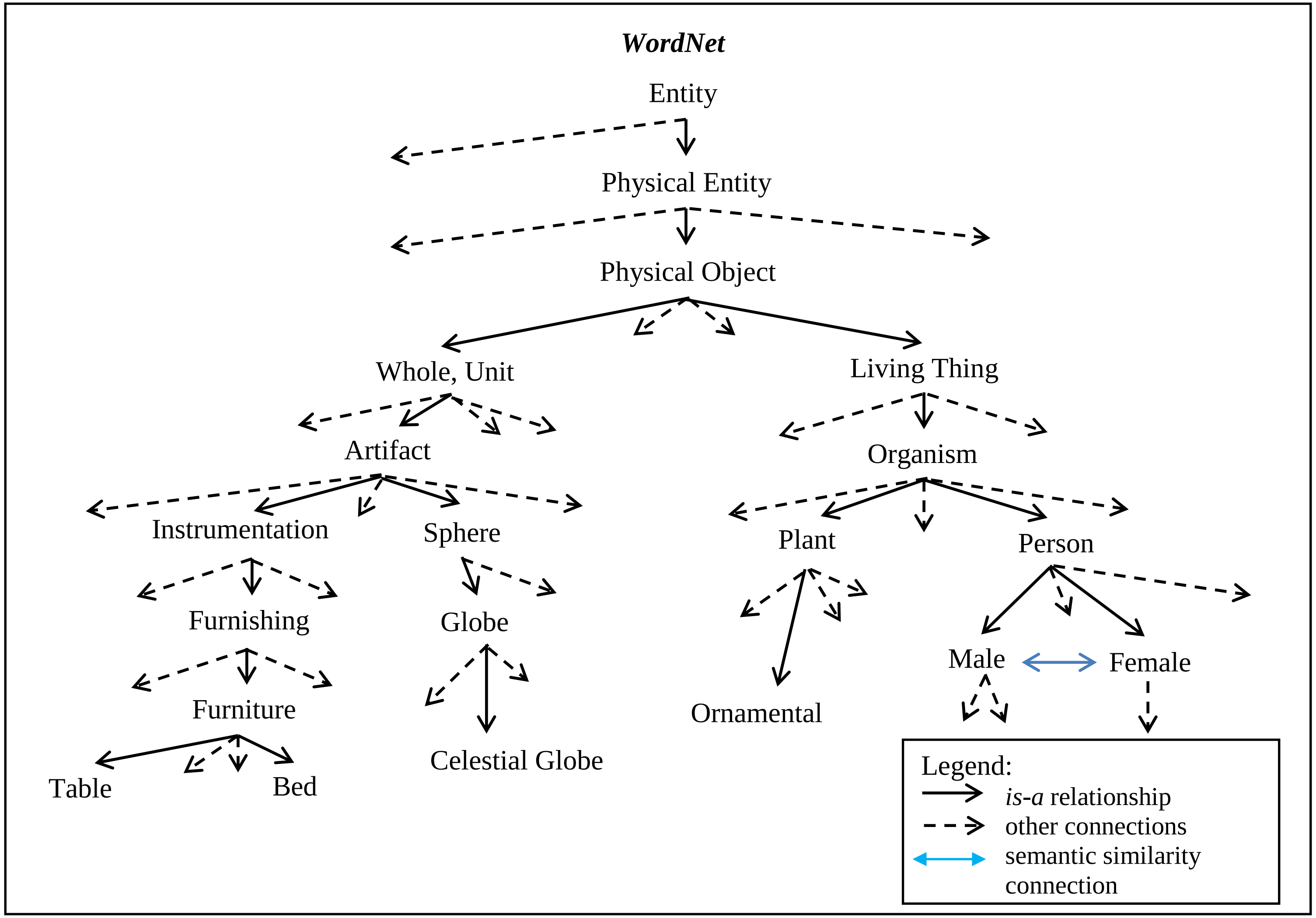


Figure 2.3: Semantic similarity in single ontology for male and female WordNet ontology (Yatskevich & Giunchiglia, 2007)

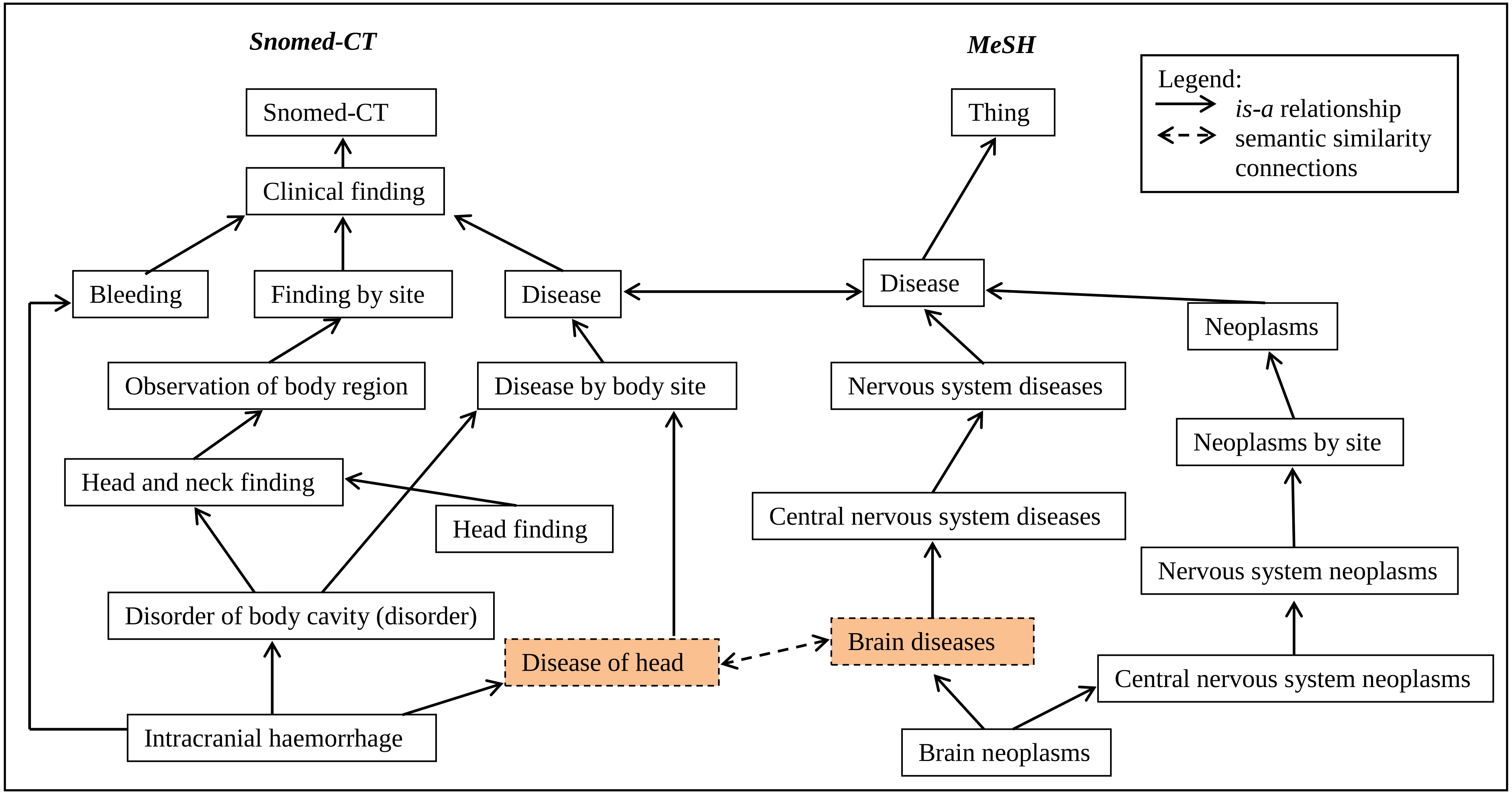


Figure 2.4: Semantic similarity in multiple ontologies for intracranial hemorrhage (in Snomed-CT) and brain neoplasms (in MeSH) (Batet *et al.*, 2014)

2.5.2 General purpose ontologies used with semantic similarity approaches

There are several examples of general purpose ontologies available including: WordNet, SENSUS, and the Cyc knowledge base. The following section describes the general purpose ontologies as follows:

2.5.2.1 WordNet

WordNet is the lexical knowledge of a native speaker of English. The latest version of WordNet is v3.1 which was released in June 2011. WordNet has 117,659 synsets and 206,941 general concepts of different domains (Slimani, 2013). These databases are semantically structured in ontological ways. It also contains nouns, verbs, adjectives and adverbs that are linked to synonym sets (synset), where each synset consists of a list of synonym word forms and semantic pointers that describe the relationships between the current synset and other synsets (Hliaoutakis *et al.*, 2006). Different types of relationships can be derived between the synsets or concepts (related to other synsets higher or lower in the hierarchy).

The hyponym/hypernym relationship (i.e., *is-a* relationship), and the meronym/holonym relationship (i.e., *part-of* relationship) are the most recognized relationships in WordNet. WordNet also introduces a larger amount of abstract concepts at the top of the taxonomic tree (Solé-Ribalta *et al.*, 2014). This is due to it being a general lexical database that does not merely focus on a singular domain. Figure 2.5 denotes the snapshot of WordNet web pages. The WordNet typically displays information such as synonym (S), direct hyponym (children of concept), direct hypernym (direct parents), full hyponym (all children), inherited hypernym (all parents), and sister term (shared direct parents). The WordNet contains a description of the concept in the form of tree structure as displayed in Figure 2.6.

REFERENCES

Abdelrahman, A. M. B., & Kayed, A. (2015). A Survey on Semantic Similarity Measures between Concepts in Health Domain. *American Journal of Computational Mathematics, 5*(June), 204–214. https://doi.org/10.4236/ajcm.2015.52017.

Aggarwal, N. (2012). Cross lingual semantic search by improving semantic similarity and relatedness measures. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 7650 LNCS*, 375–382. https://doi.org/10.1007/978-3-642-35173-0-26.

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. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5061528

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.

Alasoud, A. K. (2009). A Multi-Matching Technique for Combining Similarity Measures in Ontology Integration. Concordia University School of Graduate Studies. Concordia University. Thesis-PhD

Allison, L., & Dix, T. I. (1986). A bit-string longest-common-subsequence algorithm. *Information Processing Letters*. https://doi.org/10.1016/0020-0190(86)90091-8.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., & Al., A. P. D. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics, 25*(1), 25–29. https://doi.org/10.1038/75556.

Banerjee, S., & Pedersen, T. (2002). An Adapted Lesk Algorithm for Word Sense

Disambiguation Using WordNet. *International Conference on Intelligent Text Processing and Computational Linguistics*, *2276*, 136–145. https://doi.org/10.1007/3-540-45715-1_11.

Batet, M., Harispe, S., Ranwez, S., Sánchez, D., & Ranwez, V. (2014). An information theoretic approach to improve semantic similarity assessments across multiple ontologies. *Information Sciences*, *283*, 197–210. https://doi.org/10.1016/j.ins.2014.06.039.

Batet, M., & Sánchez, D. (2014). A Semantic Approach for Ontology Evaluation. *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*, 138–145. https://doi.org/10.1109/ICTAI.2014.30.

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.

Batet, M., Sánchez, D., Valls, A., & Gibert, K. (2010). Exploiting taxonomical knowledge to compute semantic similarity: 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.

Batet, M., Sánchez, D., Valls, A., & Gibert, K. (2013). Semantic similarity estimation from multiple ontologies. *Applied Intelligence*, *38*(1), 29–44. https://doi.org/10.1007/s10489-012-0355-y.

Bernstein, P. A., & Rahm, E. (2001). Generic Schema Matching , Ten Years Later, 695–701.

Blank, A. (2003). Words and concepts in time: Towards diachronic cognitive onomasiology. *Trends In Linguistics Studies And* Monographs, 6–25. Retrieved from http://www.metaphorik.de/01/blank.htm.

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

Budanitsky, A., & Hirst, G. (2006a). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*. https://doi.org/10.1162/ .2006.32.1.13.

Budanitsky, A., & Hirst, G. (2006b). Evaluating WordNet-based Measures of Semantic Distance. *Computational Linguistics*, *32*, 13–47.

https://doi.org/10.1162/coli.2006.32.1.13.

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.

Choi, N., Song, I.-Y., & Han, H. (2006). A Survey on Ontology Mapping, *College of Information Science and Technology Drexel University, Philadelphia, PA 19014* 35(3): 34–41.

Cohen, W. (2000). Data integration using similarity joins and a word-based information representation language. *ACM Transactions on Information Systems*. https://doi.org/10.1145/352595.352598.

Couto, F. M., Silva, M. J., & Coutinho, P. M. (2007). Measuring semantic similarity between Gene Ontology terms. In *Data and Knowledge Engineering*, *61*, 137–152. https://doi.org/10.1016/j.datak.2006.05.003

Cross, V., Yu, X., & Hu, X. (2013). Unifying ontological similarity measures: A theoretical and empirical investigation. In *International Journal of Approximate Reasoning* (Vol. 54, pp. 861–875). https://doi.org/10.1016/j.ijar.2013.03.003

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*. https://doi.org/10.1145/363958.363994.

Deselaers, T., & Ferrari, V. (2011). Visual and semantic similarity in ImageNet. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 1777–1784). https://doi.org/10.1109/CVPR.2 011.5995474.

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, *26*, 297–302. https://doi.org/10.2307/1932409.

Ding, L., Finin, T., Joshi, A., Pan, R., Cost, R. S., Peng, Y., Reddivari, P., Doshi, V., Sachs, J.(2004). Swoogle : A Semantic Web Search and Metadata Engine. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management* (pp. 652–659). https://doi.org/10.1145/1031171.1031289.

Doan, A., Madhavan, J., Domingos, P., & Halevy, A. (2004). Ontology matching: A machine learning approach. In *Handbook on Ontologies*, 1–20. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-24750-0_19.

Ehrig, M., & Staab, S. (2004). QOM–quick ontology mapping. *The Semantic Web–ISWC 2004*, 1–28. Retrieved from http://link.springer.com/chapter/10.1007/978-

3-540-30475-3_47.

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.

Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, *22*, 457–479. https://doi.org/10.1613/jair.1523.

Euzenat, J., & Shvaiko, P. (2007). Ontology Mapping. Springer Berlin Heidelberg.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. *British Journal Of Hospital Medicine London England 2005* (Vol. 71). https://doi.org/10.1139/h11-025.

Gabrilovich, E., & Markovitch, S. (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. *Proceedings of the 20th International Joint Conference on Artificial Intelligence* (pp. 1–6). https://doi.org/10.1145/2063576.2063865.

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.

Gomaa, W., & Fahmy, A. (2013). A survey of text similarity approaches. *International Journal of Computer Applications*, *68*(13), 13–18.

Grossman, J., Grossman, M., & Katz, R. (1980). The first system of weighted differential and integral calculus, ISBN 0-9771170-1-4

Harper, D. (2016). Online Etymology Dictionary. Access at June 25, 2016, from http:// http://www.etymonline.com/index.php?l=s&p=42

Hirst, G., & St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet - An Electronic Lexical Database*, 305–332. https://doi.org/citeulike-article-id:4893262.

Hliaoutakis, A. (2005). Semantic Similarity Measures in MeSH Ontology and their application to Information Retrieval on Medline. Technical University of Crete, Greek, Thesis PhD.

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.

Iannone, L., Palmisano, I., & Fanizzi, N. (2007). An algorithm based on counterfactuals for concept learning in the Semantic Web. *Applied Intelligence*, *26*, 139–159. https://doi.org/10.1007/s10489-006-0011-5.

Islam, A., & Inkpen, D. (2008). Semantic text similarity using corpus-based word similarity and string similarity. *ACM Transactions on Knowledge Discovery from Data*. https://doi.org/10.1145/1376815.1376819.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de La Société Vaudoise Des Sciences Naturelles*, *37*, 547–579. https://doi.org/citeulike-article-id:4118068.

Jaro, M. A. (1995). Probabilistic linkage of large public health data files. *Statistics in Medicine*, *14*, 491–498. https://doi.org/10.1002/sim.4780140510.

Jean-Mary, Y. R., Shironoshita, E. P., & Kabuka, M. R. (2009). Ontology Matching with Semantic Verification. *Journal of Web Semantics*, *7*(3), 235–251. https://doi.org/10.1016/j.websem.2009.04.001.

Jiang, J., & Conrath, D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv Preprint Cmp-lg/9709008*. Retrieved from http://arxiv.org/abs/cmp-lg/9709008.

Jiang, R.., Gan, M., & Doy, X., (2013). From ontology to semantic similarity: calculation of ontology-based semantic similarity. *The Scientific World Journal*, *2013*, 793091. https://doi.org/10.1155/2013/793091.

Jiang, Y., Zhang, X., Tang, Y., & Nie, R. (2015). Feature-based approaches to semantic similarity assessment of concepts using Wikipedia. *Information Processing and Management*, *51*(3), 215–234. https://doi.org/10.1016/j.ipm.2015.01.001.

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

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.

Ko, Y., Park, J., & Seo, J. (2004). Improving text categorization using the importance of sentences. *Information Processing and Management*, *40*, 65–79. https://doi.org/10.1016/S0306-4573(02)00056-0.

Kolb, P. (2009). Experiments on the difference between semantic similarity and relatedness. *Proceedings of Nodalida 2009* (pp. 81–88).

Kondrak, G. (2005). N -Gram Similarity and Distance. *Lecture Notes in Computer Science*, *3772*, 115–126. https://doi.org/10.1.1.69.2441.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*. https://doi.org/10.1037/0033-295X.104.2.211.

Lapata, M., & Barzilay, R. (2005). Automatic evaluation of text coherence: Models and representations. *IJCAI International Joint Conference on Artificial Intelligence* (pp. 1085–1090).

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.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries. *Proceedings of the 5th annual international conference on Systems documentation - SIGDOC '86* (pp. 24–26). https://doi.org/10.1145/318723.318728.

Li, H., Tian, Y., & Cai, Q. (2011). Improvement of semantic similarity algorithm based on WordNet. *Proceedings of the 2011 6th IEEE Conference on Industrial Electronics and Applications, ICIEA 2011* (pp. 564–567). https://doi.org/10.1109/ICIEA.2011.5975649.

Li, Y., McLean, D., Bandar, Z. A., O'Shea, J. D., & Crockett, K. (2006). Sentence similarity based on semantic nets and corpus statistics. *IEEE Transactions on Knowledge and Data Engineering*, *18*(8), 1138–1150. https://doi.org/10.1109/TKDE.2006.130.

Lin, C.-Y., & Hovy, E. (2003). Automatic evaluation of summaries using N-gram co-occurrence statistics. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, *2003*, 71–78. https://doi.org/10.3115/1073445.1073465.

Lin, D. (1998a). An Information-Theoretic Definition of Similarity. *Proceedings of ICML*, 296–304.

Lin, D. (1998b). Extracting Collocations from Text Corpora. *First Workshop on Computational Terminology* (pp. 57–63). https://doi.org/10.1.1.56.1687.

Liu, Y., & Zong, C. (2004). Example-based Chinese-english MT. *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics* (Vol. 7, pp. 6093–6096). https://doi.org/10.1109/ICSMC.2004.1401354.

Liu, H., Bao, H., & Xu, D. (2012). Concept vector for semantic similarity and relatedness based on WordNet structure. *Journal of Systems and Software* (Vol. 85, pp. 370–381). https://doi.org/10.1016/j.jss.2011.08.029

Lund, K., Burgess, C., & Atchley, R. A. (1995). Semantic and Associative Priming in High-Dimensional Semantic Space. In *Cognitive Science Proceedings, LEA* (pp. 660–665). https://doi.org/10.1016/j.jconhyd.2010.
08.009.

Madhavan, J., Bernstein, P. A., Doan, A., & Halevy, A. (2005). Corpus-based schema matching. In *Proceedings - International Conference on Data Engineering* (pp. 57–68). https://doi.org/10.1109/ICDE.2005.39.

Maind, A., Deorankar, A., & Chatur, P. (2012). Measurement Of Semantic Similarity Between Words: A Survey. *International Journal of Computer Science, Engineering & Informa;*, *2*(6), p51.

Martinez-Gil, J., & Aldana-Montes, J. F. (2013). Semantic similarity measurement using historical google search patterns. *Information Systems Frontiers*, *15*(3), 399–410. https://doi.org/10.1007/s10796-012-9404-7.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*, 39–41. https://doi.org/10.1145/219717.219748.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*. https://doi.org/10.1080/0169096910840
6936.

Mohameth-François., Ranwez, S., Montmain, J., Regnault, A., Crampes, M., & Ranwez, V. (2012). User centered and ontology based information retrieval system for life sciences. *BMC Bioinformatics*, *13 Suppl 1*, S4. https://doi.org/10.1186/1471-2105-13-S1-S4.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*, 443–453. https://doi.org/10.1016/0022-2836(70)900574.

Neter, J., Kutner, MH., Nachtsheim,CJ., Wasserman, W., (1996). Applied Linear Statistical Modelslrwin, Chicago.

Noy, N., & McGuinness, D. (2001). Ontology development 101: A guide to creating your first ontology. *Development*, *32*, 1–25. https://doi.org/10.1016/j.artmed.2004.01.014.

Ochiai A. (1957). Zoogeographical studies on the soleoid fishes found Japan and its neighboring regions. *Japan Society Fish Sciences, 22:526-30*.

Palmer, M., & Wu, Z. (1994). Verb Semantics And Lexical. *Proceeding ACL '94 Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, 133–138.

Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting on association for computational linguistics*, 311–318. https://doi.org/10.3115/1073083.1073135

Patwardhan, S., & Pedersen, T. (2006). Using WordNet-based Context Vectors to Estimate the Semantic Relatedness of Concepts. *11th Conference of the European Chapter of the Association for Computational Linguistics*, *1501*, 1–8. https://doi.org/citeulike-article-id:1574418.

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.

Pesquita, C., Faria, D., Falcão, A. O., Lord, P., & Couto, F. M. (2009). Semantic similarity in biomedical ontologies. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1000443.

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.

Pirró, G. (2009). A semantic similarity metric combining features and intrinsic information content. *Data and Knowledge Engineering*, *68*, 1289–1308. https://doi.org/10.1016/j.datak.2009.06.008.

Pirró, G., & Euzenat, J. (2010). A feature and information theoretic framework for semantic similarity and relatedness. *The Semantic Web–ISWC 2010*. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-17746-0_39.

Pirrò, G., Ruffolo, M., & Talia, D. (2009). SECCO: On building semantic links in peer-to-peer networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*

(Vol. 5480 LNCS, pp. 1–36). https://doi.org/10.1007/978-3-642-00685-2_1.

Pirró, G., & Seco, N. (2008). Design , Implementation and Evaluation of a New Semantic Similarity Metric Combining Features and Intrinsic Information Content. *Lecture Notes in Computer Science Volume 5332* (Vol. 5332, pp. 1271–1288). https://doi.org/10.1007/978-3-540-88873-4.

Quine, W. V. O. (1969). Ontological Relativity & Other Essays. *Ontological Relativity & Other Essays*, 1–25. https://doi.org/10.2307/2024305.

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.

Resnik, P. (1995). Using information content to evaluate seantic similarity in a taxonomy. *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI).*

Richardson, R., Smeaton, A., & Murphy, J. (1994). Using WordNet as a knowledge base for measuring semantic similarity between words. Retrieved from http://ducra.dcu.ie/wpapers/1994/1294.pdf.

Rodríguez, M. A., & Egenhofer, M. J. (2003a). Comparing Geospatial Entity Classes : An Asymmetric and Context-Dependent Similarity Measure. *International Journal of Geographical Information Science*, vol. 18, no. 3, pp. 229–256, 2004

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.

Rubenstein, H., & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*. https://doi.org/10.1145/365628.365657.

Saleena, B., & Srivatsa, S. K. (2014). Using concept similarity in cross ontology for adaptive e-Learning systems. *Journal of King Saud University - Computer and Information Sciences*, *27*(1), 1–12. https://doi.org/10.1016/j.jksuci.2014.03.007.

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.

Sánchez, D., & Batet, M. (2011). Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. *Journal of*

Biomedical Informatics, *44*, 749–759. https://doi.org/10.1016/j.jbi.2011.03.013.

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.

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.

Sánchez, D., Batet, M., Valls, A., & Gibert, K. (2010). Ontology-driven web-based semantic similarity. *Journal of Intelligent Information Systems*, *35*, 383–413. https://doi.org/10.1007/s10844-009-0103-x.

Sánchez, D., & Isern, D. (2011). Automatic extraction of acronym definitions from the Web. *Applied Intelligence*, *34*, 311–327. https://doi.org/10.1007/s10489-009-0197-4.

Sánchez, D., Isern, D., & Millan, M. (2011). Content annotation for the semantic web: An automatic web-based approach. *Knowledge and Information Systems*, *27*, 393–418. https://doi.org/10.1007/s10115-010-0302-3.

Sánchez, D., Solé-Ribalta, A., Batet, M., & Serratosa, F. (2012). Enabling semantic similarity estimation across multiple ontologies: an evaluation in the biomedical domain. *Journal of Biomedical Informatics*, *45*(1), 141–55. https://doi.org/10.1016/j.jbi.2011.10.005.

Saruladha, K., Aghila, G., & Bhuvaneswary, A. (2011a). COSS: Cross Ontology Semantic Similarity measure-An information content based approach. *2011 International Conference on Recent Trends in Information Technology (ICRTIT)*, 485–490. https://doi.org/10.1109/ICRTIT.2011.5972360.

Saruladha, K., Aghila, G., & Bhuvaneswary, A. (2011b). Information content based semantic similarity approaches for multiple biomedical ontologies. *Communications in Computer and Information Science* (Vol. 191 CCIS, pp. 327–336). https://doi.org/10.1007/978-3-642-22714-1_34.

Schallehn, E., Sattler, K. U., & Saake, G. (2004). Efficient similarity-based operations for data integration. *Data and Knowledge Engineering*, *48*, 361–387. https://doi.org/10.1016/j.datak.2003.08.004.

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).

Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, *24*, 97–123.

Schwartzman, S. (1996). The Words of Mathematics: An Etymological Dictionary of Mathematical Terms Used in English (Spectrum). *The Mathematical Association of America* (September 1996).

Schwering, A. (2008). Approaches to semantic similarity measurement for geo-spatial data: A survey. *Transactions in GIS*. https://doi.org/10.1111/j.1467-9671.2008.01084.x.

Singh, P. (2004). Web Ontology to Facilitate Semantic Web. *2nd International CALIBER-2004, New Delhi, 11-13 February, 2004* (pp. 11–13).

Slimani, T. (2013). Description and Evaluation of Semantic similarity Measures Approaches. *Journal of Computer Applicationsof Computer Applications*, *80–no 10*, 1–10.

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, *147*, 195–197. https://doi.org/10.1016/0022-2836(81)90087-5.

Solé-Ribalta, A., Sánchez, D., Batet, M., & Serratosa, F. (2014). Towards the estimation of feature-based semantic similarity using multiple ontologies. *Knowledge-Based Systems*, *55*, 101–113. https://doi.org/10.1016/j.knosys. 2013.10.015.

Spasic, I., Ananiadou, S., McNaught, J., & Kumar, A. (2005). Text mining and ontologies in biomedicine: Making sense of raw text. *Briefings in Bioinformatics*, *6*, 239–251. https://doi.org/10.1093/bib/6.3.239.

Studer, R., Benjamins, V. R., & Fensel, D. (1998). Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, *25*(1–2), 161–197. https://doi.org/10.1016/S0169-023X(97)00056-6.

Sussna, M. (1993). Word sense disambiguation using a massive of computer for free-text semantic indexing network. *CIKM '93 Proceedings of the Second International Conference on Information and Knowledge Management*, 67–74. https://doi.org/10.1145/170088.170106.

Thiagarajan, R., Manjunath, G., & Stumptner, M. (2008). Computing Semantic Similarity Using Ontologies. *Computing*, Germany. Retrieved from http://www.hpl.hp.com/techreports/2008/HPL-2008-87.pdf

Tversky, A. (1977). Features of similarity. *Psychological Review*. https://doi.org/10.1037/0033-295X.84.4.327.

Vicient, C., Sánchez, D., & Moreno, A. (2013). An automatic approach for ontology-based feature extraction from heterogeneous textual resources. *Engineering Applications of Artificial Intelligence*, *26*, 1092–1106. https://doi.org/10.1016/j.engappai.2012.08.002.

Wache, H., Vögele, T., Visser, U., Stuckenschmidt, H., Schuster, G., Neumann, H., & Hübner, S. (2001). Ontology-Based Integration of Information - A Survey of Existing Approaches. *IJCAI-01 Workshop: Ontologies and Information Sharing* (pp. 108–117). https://doi.org/10.1016/j.ijhcs.2008.07.007.

Wang, L., & Chen, M. (2013). A New Model to Compute Semantic Similarity from Multi-ontology. *China Semantic Web Symposium and Web Science Conference. Communications in Computer and Information Science book series* (CCIS, vol: 406) 1–10.

Wang, J. Z., Du, Z., Payattakool, R., Yu, P. S., & Chen, C.-F. (2007). A new method to measure the semantic similarity of GO terms. *Bioinformatics (Oxford, England)*, *23*, 1274–1281. https://doi.org/10.1093/bioinformatics/btm087.

Wegrzyn-Wolska, K., & Szczepaniak, P. S. (2005). Classification of RSS-Formatted Documents Using Full Text Similarity Measures. *Lecture Notes in Computer Science*, *3579*, 400–405.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research* (pp. 354–359).

Yatskevich, M., & Giunchiglia, F. (2007). Element level semantic matching using WordNet. *CiteSeerX*.

Yuhua Li, Zuhair A. Bandar, and D. M. (2003). An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on Knowledge and Data Engineering*, *15*(4), 871–882. https://doi.org/10.1109/TKDE.2003.1209005.

Zhang, S., Zheng, X., & Hu, C. (2015). A Survey of Semantic Similarity and its Application to Social Network Analysis. *IEEE International Conference on Big Data (Big Data)* (pp. 1–3).