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 Universe 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 likeness 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

TABLE OF CONTENTS

DECLARATION ii
DEDICATION iii
ACKNOWLEDGEMENT iv
ABSTRACT v
ABSTRAK vi
TABLE OF CONTENTS vii
LIST OF TABLES xii
LIST OF FIGURES xvi
LIST OF ALGORITHMS xix
LIST OF SYMBOLS AND ABBREVIATIONS xx
LIST OF APPENDICES xxv
LIST OF PUBLICATIONS xxvi

CHAPTER 1 INTRODUCTION 1
1.1 Overview 1
1.2 Problem statement 3
1.3 Research objectives 5
1.4 Research questions 5
1.5 Scope and research significance 6
1.6 Organization of the thesis 9
CHAPTER 2 LITERATURE REVIEW

2.1 Introduction 10
2.2 Similarity 11
2.3 Words similarity 12
2.4 Semantic similarity 16
2.5 Knowledge-based 19
  2.5.1 Ontology-based 19
  2.5.2 General purpose ontologies used with semantic similarity approaches 24
  2.5.3 Domain specific ontologies used with semantic similarity approaches 27
2.6 Classification of semantic similarity approaches according to ontology 33
2.7 Single ontology 37
  2.7.1 Structure-based approach 37
  2.7.2 Information content-based approach 40
  2.7.3 Feature-based approach 42
  2.7.4 Hybrid-based approach 43
2.8 Multiple ontologies 43
  2.8.1 Matching method 44
  2.8.2 Ontological feature matching for multiple ontologies 45
  2.8.3 Semantic similarity measure for multiple ontologies 47
2.9 Biomedical domain as a case study 56
2.10 Discussion of similarity 59
2.11 Chapter summary 61

CHAPTER 3 RESEARCH METHODOLOGY

3.1 Introduction 63
3.2 Structure of semantic similarity 63
3.3 Research framework 64
3.4 Data sources and preparation 68
  3.4.1 WordNet 71
CHAPTER 4 THE ONTOLOGICAL FEATURES
ALGORITHM IN SEMANTIC SIMILARITY FOR
MULTIPLE ONTOLOGIES 95
4.1 Introduction 95
4.2 Ontological features algorithm 97
  4.2.1 Matching matrix semantic relationship 97
  4.2.2 Semantic overlapping between subsumers 99
  4.2.3 Selecting the LCS 101
4.3 Result and discussion 103
  4.3.1 Ontological features evaluation result 103
  4.3.2 Matching method results 106
  4.3.3 Correlation between similarity measurement and method for multiple ontologies 109
4.4 Chapter summary 112

CHAPTER 5 Hyb-TvX: A HYBRID SEMANTIC SIMILARITY FEATURE-BASED MEASUREMENT FOR
MULTIPLE ONTOLOGIES 113
5.1 Introduction 113
5.2 Hyb-TvX: A hybrid semantic similarity feature-based measurement 116
  5.2.1 TvX-1: similarity measurement level 1 122
  5.2.2 TvX-2: similarity measurement level 2 125
5.3 Result and discussion 128
CHAPTER 6 CONCLUSION AND FUTURE WORKS

6.1 Introduction

6.2 Research summary

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

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

6.4 To propose and evaluate an ontological features algorithm in semantic similarity for multiple ontologies in terms of features matching accuracy

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

6.4.2 Does the proposed ontological features algorithm increase the accuracy of feature matching?

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

6.5 To develop and evaluate feature-based mechanism to measure semantic similarity of multiple ontologies to increase similarity accuracy
6.5.1 How does the proposed parameters support in improving the accuracy of similarity for multiple ontologies? 141

6.5.2 Does it a significant relationship between the Hyb-TvX methods with human scored? 143

6.6 Contribution of the research 143

6.7 Future research 144

REFERENCES 145

APPENDIX 157
LIST OF TABLES

2.1 Symbol and meaning in relation to similarity 12
2.2 Words similarity approach 15
2.3 Fields and applications that use the semantic similarity approach 17
2.4 Advantages and disadvantages based on the background information in the semantic similarity approach 19
2.5 Classification approach according to the categories of ontology 38
2.6 Advantages and disadvantages of the Tversky method 42
2.7 Types of ontological features in each semantic similarity method 45
2.8 Descriptions of two concepts from WordNet and MeSH 49
2.9 Method feature-based approach for multiple ontologies 56
2.10 Summary of datasets for single ontology used in previous work 58
2.11 Summary of datasets for multiple ontologies that used in previous work 58
3.1 Set of 36 medical term pairs with averaged experts similarity scores (Hliaoutakis et al., 2006) 68
3.2 Set of 30 medical term pairs with averaged physician and coder ratings scores (Pedersen et al., 2007) 70
3.3 The level of strength correlation coefficient

3.4 Example of results human similarity and proposed method

3.5 Calculate the correlation coefficient \((r)\) table

3.6 Result implementation parameters for situation 1

3.7 Result implementation parameters for situation 2

3.8 Result implementation parameters for situation 3

3.9 Result implementation parameters for situation 4

3.10 Result implementation parameters for situation 5

3.11 Result implementation parameters for situation 6

3.12 Result implementation parameters for situation 7

3.13 Result implementation parameters for situation 8

4.1 Comparison previous method and OntF

4.2 Similarity result of the proposed method (OntF) using four similarity feature-based measurements in Pedersen benchmark datasets based on physician ratings

4.3 Similarity result of the proposed method (OntF) using four similarity feature-based measurements in Pedersen benchmark datasets based on coder ratings

4.4 Similarity result of the proposed method (OntF) using four similarity feature-based measurements in Hlioutakis benchmark datasets based on expert ratings
4.5 Similarity results of multiple ontologies. The bold column represents the proposed method (OntF) 108

4.6 Correlation values based on four similarity feature-based measurements using different methods. The row in bold shows the results of the proposed method (OntF) 110

4.7 Improvements that the OntF achieved over the average of the three other methods using the Sánchez measure based on Physician, Coder and Expert ratings 111

4.8 Improvements that the OntF achieved over the average of the three other methods using the Jaccard measure based on Physician, Coder and Expert ratings 111

4.9 Improvements that the OntF achieved over the average of the three other methods using the Dice measure based on Physician, Coder and Expert ratings 112

4.10 Improvements that the OntF achieved over the average of the three other methods using the Ochiai measure based on Physician, Coder and Expert ratings 112

5.1 Example for concepts extracting a set of token 118

5.2 Example for concepts with tokenization of words 118

5.3 Example for concepts and synonym 119

5.4 Examples of concepts 122

5.5 Example for concepts and synonym 124

5.6 Example of concepts and features 126

5.7 Comparison similarity of the proposed method with the X-similarity and Rodriguez
and the Egenhofer method using Pedersen benchmarks datasets

5.8 Comparison similarity of the proposed method with the X-similarity and Rodriguez and Egenhofer method using Hlioutakis benchmarks datasets

5.9 Comparison similarity of proposed method with physician, coder and experts ratings (averaged)

5.10 Correlation of similarity method on feature-based approach for multiple ontologies according to the WordNet and MeSH dataset

5.11 $p$-value of similarity method on feature-based approach for multiple ontologies

6.1 Investigated the proposed parameters ($\alpha$) in eight situations involve condition $|comp B|$, $|comp A|$, differentiation complement and intersection value
LIST OF FIGURES

1.1 Example of terminological matching between subsumer of antibiotic (WordNet) and antibacterial agent (MeSH) 4
1.2 Comparison of the number of articles published every year and the number of citations per year (http://apps.webofknowledge.com) 8
2.1 Content structure 10
2.2 The computer ontology 20
2.3 Semantic similarity in single ontology for male and female WordNet ontology 22
2.4 Semantic similarity in multiple ontologies for intracranial hemorrhage (in Snomed-CT) and brain neoplasms (in MeSH) 23
2.5 The snapshot of lexical database for english (WordNet) 25
2.6 The snapshot of WordNet content for renal failure 25
2.7 The snapshot of the Cyc web page. 26
2.8 Snapshot of the UMLS web page. 28
2.9 The snapshot of Medical Subject Headings (MeSH) 29
2.10 The snapshot of MeSH content for kidney disease 29
2.11 The snapshot of Snomed-CT web pages 30
2.12 An example of a biological process tree for biological_process (GO: 0008150). It is the
parent while biological regulation (GO:0065007), localisation (GO:00051179), phosphorus utilisation (GO:0006794), and signaling (GO:00023052) are its children.

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.13</td>
<td>The snapshot web page of SDTS</td>
</tr>
<tr>
<td>2.14</td>
<td>Semantic similarity for single ontology</td>
</tr>
<tr>
<td>2.15</td>
<td>Semantic similarity for multiple ontologies</td>
</tr>
<tr>
<td>2.16</td>
<td>Single ontology</td>
</tr>
<tr>
<td>2.17</td>
<td>The concept similarity between $C_1$ and $C_2$</td>
</tr>
<tr>
<td>2.18</td>
<td>WordNet taxonomy, showing most-specific subsumers of nickel and dime and of nickel and credit card. Solid lines represent is-a links; dashed lines indicate that some intervening nodes have been omitted</td>
</tr>
<tr>
<td>2.19</td>
<td>Ontologies $O_1$ and $O_3$</td>
</tr>
<tr>
<td>2.20</td>
<td>The process to develop hybrid semantic similarity feature-based for multiple ontologies</td>
</tr>
<tr>
<td>3.1</td>
<td>Structure of semantic similarity method</td>
</tr>
<tr>
<td>3.2</td>
<td>Research framework</td>
</tr>
<tr>
<td>3.3</td>
<td>Research structure</td>
</tr>
<tr>
<td>3.4</td>
<td>Example of a taxonomical tree in WordNet</td>
</tr>
<tr>
<td>3.5</td>
<td>Example of a taxonomical tree MeSH</td>
</tr>
<tr>
<td>3.6</td>
<td>$</td>
</tr>
<tr>
<td>3.7</td>
<td>$</td>
</tr>
<tr>
<td>3.8</td>
<td>Results for condition 1 in situations 1, 2, 3 and 4</td>
</tr>
<tr>
<td>3.9</td>
<td>Results for condition 2 in situations 5, 6, 7 and 8</td>
</tr>
<tr>
<td>4.1</td>
<td>Ontological modelling of the concept lupus in MeSH</td>
</tr>
<tr>
<td>4.2</td>
<td>MeSH ontology for <em>myocardium</em> concept</td>
</tr>
<tr>
<td>5.1</td>
<td>The flow process of Hyb-TvX</td>
</tr>
</tbody>
</table>
5.2  Venn diagram for intersection  116
5.3  Venn diagram for union  117
5.4  Venn diagram for complement  117
5.5  Intersection of \( (C_1) \) and \( (C_2) \)  118
5.6  Maximum tokenization of \( (C_1) \) and \( (C_2) \)  119
5.7  Intersection synonym of \( (C_1) \) and \( (C_2) \)  120
5.8  Union synonym of \( (C_1) \) and \( (C_2) \)  120
5.9  Venn diagram for complement \( A \)  121
5.10 Venn diagram for complement \( B \)  121
# LIST OF ALGORITHMS

<table>
<thead>
<tr>
<th></th>
<th>Algorithm Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.1</td>
<td>The proposed ontological features algorithm (OntF)</td>
<td>101</td>
</tr>
<tr>
<td>4.2</td>
<td>Features algorithm from Sánchez et al., (2012)</td>
<td>102</td>
</tr>
<tr>
<td>4.3</td>
<td>Features algorithm from Batet et al., (2013)</td>
<td>102</td>
</tr>
<tr>
<td>5.1</td>
<td>The process Hyb-TvX algorithm</td>
<td>128</td>
</tr>
</tbody>
</table>
**LIST OF SYMBOLS AND ABBREVIATIONS**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \forall )</td>
<td>For all, any, each, every of element.</td>
</tr>
<tr>
<td>( \in )</td>
<td>Is element of</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>Selection of</td>
</tr>
<tr>
<td>( \geq )</td>
<td>Greater than or equal to</td>
</tr>
<tr>
<td>( A )</td>
<td>Set ( A )</td>
</tr>
<tr>
<td>( B )</td>
<td>Set ( B )</td>
</tr>
<tr>
<td>( o )</td>
<td>Objects</td>
</tr>
<tr>
<td>( \times )</td>
<td>Cross</td>
</tr>
<tr>
<td>( R )</td>
<td>Real number</td>
</tr>
<tr>
<td>( = )</td>
<td>Similar to</td>
</tr>
<tr>
<td>( x )</td>
<td>Concept ( x )</td>
</tr>
<tr>
<td>( y )</td>
<td>Concept ( y )</td>
</tr>
<tr>
<td>( \cap )</td>
<td>Intersection</td>
</tr>
<tr>
<td>( * )</td>
<td>Multiplication</td>
</tr>
<tr>
<td>( C_1 )</td>
<td>First concept</td>
</tr>
<tr>
<td>( C_2 )</td>
<td>Second concept</td>
</tr>
<tr>
<td>( C_3 )</td>
<td>Third concept</td>
</tr>
<tr>
<td>( N_1 )</td>
<td>( N_1 ) are the number of is-a links from ( A )</td>
</tr>
<tr>
<td>( N_2 )</td>
<td>( N_2 ) are the number of is-a links from ( B )</td>
</tr>
<tr>
<td>( N_3 )</td>
<td>( N_3 ) are the number of is-a links from ( C )</td>
</tr>
<tr>
<td>LCS</td>
<td>Lower common subsume</td>
</tr>
<tr>
<td>( W(c) )</td>
<td>( W(c) ) is the set of words (nouns) in the corpus</td>
</tr>
<tr>
<td>( c / C )</td>
<td>Concept ( (c) ) or concept ( (C) )</td>
</tr>
<tr>
<td>( \Sigma )</td>
<td>Total</td>
</tr>
</tbody>
</table>
\begin{itemize}
  \item \(N\) - The total number of word (noun) tokens in the corpus
  \item \(S(a, b)\) - Similarity between concept \(a\) and \(b\)
  \item \(\text{Sim}(C_1, C_2)\) - Similarity concept 1 and concept 2
  \item \(\text{Sim}(a, b)\) - Similarity concept \(a\) and concept \(b\)
  \item \(O\) - Ontology
  \item \(O_1\) - First ontology
  \item \(O_2\) - Second ontology
  \item \(R_i\) - Type \((i)\) of semantic relationship for first ontology
  \item \(R_j\) - Type \((j)\) of semantic relationship for second ontology
  \item \(X\) - \(X\) correspondences to sets of \(a\)
  \item \(Y\) - \(Y\) correspondences to sets of \(b\)
  \item \(X \cap Y / A \cap B\) - Set \(X\) union set \(Y\) or Set \(A\) union set \(B\)
  \item \(|X - Y|\) - The relative complement of \(Y\) in \(X\)
  \item \(|Y - X|\) - The relative complement of \(X\) in \(Y\)
  \item \(A \cup B\) - Set \(A\) union set \(B\)
  \item \(1 - \alpha > 0\) - \(1 - \alpha\) must be more the zero
  \item \(\alpha\) - Parameter for complement
  \item \(\text{depth}(a^p)\) - Depth of ontology for concept \(a\)
  \item \(\text{depth}(b^q)\) - Depth of ontology for concept \(b\)
  \item \(\alpha(a^p, b^q)\) - Parameter for complement \(a\) and \(b\)
  \item \(w_u, w_u, w_u \geq 0\) - Weighting parameter must be same or more than zero
  \item \(S_u\) - Similarity word matching
  \item \(S_u\) - Similarity feature matching
  \item \(S_u / S_{\text{neighborhood}}\) - Similarity neighborhood
  \item \(S_p(a^p, b^q)\) - Similarity parts for concept \(a\) and \(b\)
  \item \(S_f(a^p, b^q)\) - Similarity function for concept \(a\) and \(b\)
  \item \(S_a(a^p, b^q)\) - Similarity attribute or concept \(a\) and \(b\)
  \item \(S_{\text{synset}}\) - Similarity synset or synonym
  \item \(S_{\text{description}}\) - Similarity description
  \item \(\text{Total}_{\text{hypo}}\) - All hyponym
\end{itemize}
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{total}_{\text{hypo}} O_i(S_i)$</td>
<td>All hyponym in ontology $i$ for subsumer $i$</td>
</tr>
<tr>
<td>$(r)$</td>
<td>Pearson correlation coefficient</td>
</tr>
<tr>
<td>$n$</td>
<td>Number of word pairs</td>
</tr>
<tr>
<td>$(\sum x_i y_i)$</td>
<td>Total multiplication of human judgments $(x_i)$ and $y_i$ is the corresponding $i$th element in the list of similarity value</td>
</tr>
<tr>
<td>$(\sum x_i)$</td>
<td>Total value human judgment or physician judgment</td>
</tr>
<tr>
<td>$(\sum y_i)$</td>
<td>Total value similarity based similarity measurement method</td>
</tr>
<tr>
<td>$M_R$</td>
<td>Matrix relationship</td>
</tr>
<tr>
<td>$M_S$</td>
<td>Matrix subsume</td>
</tr>
<tr>
<td>$R_i (S_i)$</td>
<td>The row represent the subsumer of relationship</td>
</tr>
<tr>
<td>$R_j (S_j)$</td>
<td>The column represents the subsumer of relationship</td>
</tr>
<tr>
<td>Hyb-TvX</td>
<td>Proposed method (A Hybrid Semantic Similarity Feature-based Measurement)</td>
</tr>
<tr>
<td>TvX-1</td>
<td>Similarity Measurement level 1</td>
</tr>
<tr>
<td>TvX-2</td>
<td>Similarity Measurement level 2</td>
</tr>
<tr>
<td>(Int)</td>
<td>Intersection</td>
</tr>
<tr>
<td>(Un)/U</td>
<td>Union</td>
</tr>
<tr>
<td>(comp)</td>
<td>Complement</td>
</tr>
<tr>
<td>(max)</td>
<td>Take the maximum value of the words</td>
</tr>
<tr>
<td>$x \in A$</td>
<td>$x$ element of set $A$</td>
</tr>
<tr>
<td>$x \in B$</td>
<td>$x$ element of set $B$</td>
</tr>
<tr>
<td>$\text{comp} A/\text{A'}$</td>
<td>Complement for set $A$</td>
</tr>
<tr>
<td>$\text{comp} B/\text{B'}$</td>
<td>Complement for set $B$</td>
</tr>
<tr>
<td>$x \in U$</td>
<td>$x$ element of set Union</td>
</tr>
<tr>
<td>$x \notin A$</td>
<td>$x$ not element of set $A$</td>
</tr>
<tr>
<td>$(\text{Int}</td>
<td>C_1, C_2]$</td>
</tr>
<tr>
<td>$(\text{max}</td>
<td>C_1, C_2]$</td>
</tr>
<tr>
<td>$(\text{Int}</td>
<td>A, B]$</td>
</tr>
<tr>
<td>$(\text{Un}</td>
<td>A, B]$</td>
</tr>
</tbody>
</table>
\[ S_c(C_1, C_2) \] - 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

\[ \max(S_c, S_s, S_f) \] - Identify value maximum between \( S_c, S_s, S_f \)

\[ (w_a) \] - The proposed parameter for |\text{comp } A|

\[ (w_b) \] - The proposed parameter for |\text{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
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLM</td>
<td>National Library of Medicine</td>
</tr>
<tr>
<td>MH</td>
<td>Mesh Heading</td>
</tr>
<tr>
<td>STDS</td>
<td>Spatial Data Transfer Standard</td>
</tr>
<tr>
<td>LCA</td>
<td>Lower common ancestor</td>
</tr>
<tr>
<td>STS</td>
<td>Semantic textual similarity</td>
</tr>
<tr>
<td>RM</td>
<td>Root Matching</td>
</tr>
<tr>
<td>TM</td>
<td>Terminological Matching</td>
</tr>
<tr>
<td>TM</td>
<td>Terminological Subsumption</td>
</tr>
<tr>
<td>SS</td>
<td>Semantic Subsumption</td>
</tr>
<tr>
<td>LCS</td>
<td>Lower common subsume</td>
</tr>
</tbody>
</table>
# LIST OF APPENDICES

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A2</td>
<td>List of total concept datasets MeSH for Hliaoutakis <em>et al.</em>, (2006) benchmark</td>
<td>159</td>
</tr>
<tr>
<td>A3</td>
<td>List of total concept datasets WordNet for Pedersen <em>et al.</em>, (2007) benchmark</td>
<td>161</td>
</tr>
<tr>
<td>A4</td>
<td>List of total concept datasets MeSH for Pedersen <em>et al.</em>, (2007) benchmark</td>
<td>162</td>
</tr>
</tbody>
</table>
LIST OF PUBLICATIONS

Journal:


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; Vicent 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 datasets 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.

<table>
<thead>
<tr>
<th>The number of articles published per year</th>
<th>The number of citations per year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic similarity with multiple ontologies</td>
<td>Semantic similarity feature-based</td>
</tr>
</tbody>
</table>

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 $\sigma : o \times o \to \mathbb{R}$ is a function from a pair of entities to real number ($\mathbb{R}$) 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:

\begin{align*}
\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) \ (\text{max imality}) \\
\forall x, y \in o, \sigma(x, y) &= \sigma(y, x) \ (symmetry)
\end{align*}
Table 2.1: Symbol and meaning in relation to similarity

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>\forall</td>
<td>for all, for any, for each, for every</td>
</tr>
<tr>
<td>\in</td>
<td>is an element of, is not an element of</td>
</tr>
<tr>
<td>\sigma</td>
<td>selection of</td>
</tr>
<tr>
<td>\geq</td>
<td>is less than or equal to, is greater than or equal to</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Approach</th>
<th>Techniques</th>
<th>Example Research</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical</td>
<td>Character-based</td>
<td>Allison &amp; Dix (1986); Damerau (1964); Jaro (1995); Winkler (1990); Needleman &amp; Wunsch (1970); Smith &amp; Waterman (1981) and Kondrak (2005).</td>
<td>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.</td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td>Term-based</td>
<td>Jaccard (1901); Dice (1945)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semantic</td>
<td>Similarity</td>
<td>Resnik (1995); Jiang &amp; Conrath (1997); Palmer &amp; Wu (1994); Leacock &amp; Chodorow (1998); Rodríguez &amp; Egenhofer (2003a); Saruladha et al., (2011b) and Sánchez et al., (2012)</td>
<td>This approach does not depend on the sequence of string. Semantic similarity computes the likeness between words, understood as the degree of taxonomical proximity.</td>
<td>Need evaluation of the semantic evidence observed in knowledge source. Knowledge source presented in unprocessed and heterogeneous textual formats</td>
</tr>
</tbody>
</table>
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>
<thead>
<tr>
<th>Approach</th>
<th>Field</th>
<th>Application</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Information extraction</td>
<td>Semantic annotation</td>
<td>Sánchez &amp; Isern, (2011); Sánchez et al., (2011); Iannone et al., (2007)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ontology learning</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Information integration</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ontology matching</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Semantic query routing</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bioinformatic</td>
<td></td>
</tr>
</tbody>
</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.

(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

<table>
<thead>
<tr>
<th>Approach</th>
<th>Background Information</th>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semantic similarity</td>
<td>Structure Information (knowledge-based)</td>
<td>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.</td>
<td>Need proper understanding of concept semantics</td>
</tr>
<tr>
<td></td>
<td>Unstructured Information (corpus-based)</td>
<td>Similarity used corpus-based perform well in document similarity.</td>
<td>Rely on the assumption that related words exist in the same document. Need to improve for short text or phrases.</td>
</tr>
</tbody>
</table>

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.

![Computer Ontology Diagram](image)

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.


Kolb, P. (2009). Experiments on the difference between semantic similarity and


