Semantic Search using Thematic Similarity in Digital Documents

By
Madiha Butt
NUST201260809MSEECS61312F

Supervisor
Dr. Sharifullah Khan
Department of Computing

A thesis submitted in partial fulfillment of the requirements for the degree of Masters of Science in Computer Science (MSCS)

In
School of Electrical Engineering and Computer Science,
National University of Sciences and Technology (NUST),
Islamabad, Pakistan.

(December 2015)

Approval

It is certified that the contents and form of the thesis entitled **"Semantic Search using Thematic Similarity in Digital Documents"** submitted by **Madiha Butt** have been found satisfactory for the requirement of the degree.

Advisor: **Dr. Sharifullah Khan**

Signature: _____
\
Date: _____

Committee Member 1: **Dr. Khalid Latif**

Signature: _____

Date: _____

Committee Member 2: **Ms. Hirra Anwer**

Signature: _____

Date: _____

Committee Member 3: **Dr. Muhammad Muneeb Ullah**

Signature: _____

Date: _____

Abstract

Typical semantic-based search systems resolve semantic heterogeneity by augmenting keywords through domain ontology. They consider individual keywords i.e. either concepts or relationships of ontology, but ignore the semantic relationships that exist between keywords. Therefore, to answer complex queries accurately is not possible even augmenting the query's keyword with different semantic relationships. To find the right document is only possible, if a system knows the meanings of the concepts and relationships that exist among the concepts. The proposed system takes concepts as well as the relationship that exists among them for considering the context. The system performed searching by matching RDF triples rather than individual keywords. The documents are ranked according to their relevance score of triples.

To validate the proposed semantic similarity measure, a prototype system has been implemented. The proposed semantic similarity measure uses both the structure of ontology and statistical information content to compute the semantic similarity. By combining a taxonomic structure with empirical probability estimates, it provides a way of adapting a static knowledge structure to multiple contexts. Through RDF triple matching, we have computed context based information retrieval. The proposed system has been evaluated by repeating Charles and Miller experiment and by comparing the proposed measure with several other similarity measures. Experimental results demonstrate better performance over up-to-date similarity measures. We have also evaluated our measure using Pilot Short Text Semantic Similarity Benchmark Data Set (STASIS) and we have obtained 85% correlation with STASIS. In future, we intend to consider the most appropriate sense of a concept to further improve its accuracy.

Certificate of Originality

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any degree or diploma at NUST SEECS or at any other educational institute, except where due acknowledgement has been made in the thesis. Any contribution made to the research by others, with whom I have worked at NUST SEECS or elsewhere, is explicitly acknowledged in the thesis.

I also declare that the intellectual content of this thesis is the product of my own work, except for the assistance from others in the project's design and conception or in style, presentation and linguistics which has been acknowledged.

Author Name: **Madiha Butt**

Signature:———————————

I would like to dedicate this thesis to my loving parents and teachers for their unconditional support and encouragement . . .

Acknowledgements

First and foremost, I would like to thank ALLAH ALMIGHTY for giving me courage and motivation during this thesis to cater all the difficulties and problems in an amicable manner.

I offer sincere gratitude to my supervisor Dr. Sharifullah Khan who has put his great effort throughout the thesis phase with his knowledge, expertise and valuable suggestions. He has provided full support, mentorship and continuous assistance that enabled me to learn new concepts of related domain and developed an understanding of how to perform research. I wish to thank my committee members Dr. Khalid Latif, Ms. Hirra Anwer and Dr. Muhammad Muneeb Ullah for their kind support that helped me to refine this work.

I would extend my appreciation towards DELSA lab administrators who provided me With useful resources in order to perform my work smoothly. In the end I thank to all my friends and everyone who supported me in any manner for the completion of my thesis.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter introduces the research work that has been carried out in this thesis. It includes motivation, problem definition followed by a discussion of objectives.

1.1 Motivation

The growth and use of the World Wide Web has been increasing tremendously. Therefore it has been becoming a challenge to extract the required information from this huge amount of repository. One of the most important reasons of this anomaly is semantic heterogeneity, means that the same information is represented differently at various places, such as the word bus can be associated to an automobile or it can be a computer terminology and the word computer and PC (personal computer) are represented to same concept. And another important factor is absence of context. Context is needed to accurately answer the question. As an example if I ask you to distinguish the two concepts which are the more related between the two pairs (Monkey, Phone) and (Monkey, Banana), people will most of the time agree that the two concepts Monkey and Banana are more related. But keyword based search systems cannot deal with these problems and extract only few amount of information or huge amount of irrelevant information [1]. We can reduce this problem by considering context and meaning of concepts.

The current representation of web resources is more appropriate for humans rather than for machines or software agents. To make it more understandable for machines, semantic web introduced a new frame work known as Resource Description Framework (RDF). Triple is a basic building block of RDF which consists of three parts named as a subject i.e. madiha, an attribute or a value also known as property i.e. teach, and an object i.e. programming. It is also known as subject predicate object (SPO) representation [2, 3, 4, and 5].

Efficient information retrieval is very important for the success of web. Due to the limitations of Conventional information retrieval systems which do not involve the semantics of concepts, semantic based search systems are motivated [6].

Existing semantic based search systems resolve different semantic heterogeneity issues such as synonymy by augmenting keywords through domain ontology. They consider the semantics of individual concepts but ignore the relationships that exists among the concepts [7, 8]. A pattern consists of at least two concepts and relationships among these concepts. A pattern can signify the context in which a concept needs to be considered. RDF represents this pattern in the form of triples. Therefore we have focused on matching triples instead of keywords.

This motivated us to propose a semantic similarity measure to compute the semantic similarity among RDF triples. The proposed semantic similarity measure uses both the structure of ontology and statistical information content. We have evaluated our system by repeating charles and miller experiment [9] and by comparing our measure with several other similarity measures. Experimental results have shown the highest correlation to charles and miller experiment. We have also evaluated our measure using "pilot short text semantic similarity benchmark data set (STASIS)" [10] and we have obtained 85% correlation with STASIS.

1.2 Problem Statement

Context is very important to retrieve accurate information. Keyword based search systems do not consider the context they just focus on the semantics of individual keywords. Context can be considered by focusing on the relationships that exist among keywords. In [12] Jibran, et. al. has focused on the matching of a pattern. A pattern can represent the context that is the circumstances in which something happens or to be considered. He has used distance based approach to compute the semantic similarity of the pattern. However, many new and efficient techniques have been emerged to find the semantic similarity of the concepts. Like in [13] they have measured the semantic relatedness of concepts by finding basic expansion terms using density measure (DM), betweenness measure and semantic similarity measures. After that they use ontology alignment to find the new expansion terms. At the end the expanded terms are weighted by combination of different semantic measures. In [14] class match measure, centrality measure, density measure and semantic similarity measures have been used to find the semantic relatedness of the concepts. All above mentioned techniques are concept matching techniques based on edge counting. The drawback of edge based approach is that it assumes that all the taxonomy links have uniform distances [31]. But it is not easy to maintain and control. Another way of finding ontology based semantic similarity is to compute information content. IC is greatly affected by shallow annotations, but it is not affected by variated link distances. The combination of taxonomic structure and empirical probability estimates provides a way of using static knowledge structure to multiple contexts [44]. We have extended the Jibran's methodology by proposing new semantic similarity measurement for triple matching so that we can consider the context of the keywords. Which will ultimately improve the precision and recall of information retrieval.

1.3 Research Goal

The overall goal of our research is to design and develop a system that can compute the semantic similarity between RDF triples. By giving a query and a document, the system will tell us how much the document is semantically related with query.
The objective of our methodology is to

- Compute the semantic similarity by considering the context during search
- Rank the documents based on their semantic relatedness to query
- To improve the precision of information retrieval

1.4 Proposed System

We have proposed a semantic similarity measure to compute the similarity between triples of RDF. Our measure comprises two distinct features named as depth of overlap in path (DOP) and information content of most informative common ancestors of two concepts. The combination of

taxonomic structure and empirical probability estimates provides a way of using static knowledge structure to multiple contexts [44]. Through RDF triple matching, we have computed context based information retrieval. Proposed measure is evaluated on two data sets i.e. Miller and Charles (MC) and "Pilot Short Text Semantic Similarity Benchmark Data Set (STASIS)" on two different levels. We have compared our similarity measure with several existing similarity measures using Miller & Charles experiment. Our measure is showing highest correlation to MC which is 84%. We have also constructed triple pairs and compare them with STASIS dataset and got 85% correlation.

1.5 Thesis Distribution

This document is organized as follows: Chapter 2 presents a background to semantic web, its standards and important terminologies of RDF schema. Chapter 3 discusses various techniques and similarity measures to compute the similarity between concepts. In chapter 4, we have given a detailed description of the proposed System methodology, explaining in detail the process of similarity computation between RDF triple pairs. Chapter 5 gives a complete overview of implementation details and describes the experimental results and a comparison with the existing systems. Concluding remarks and future work are presented in chapter 6.

Chapter 2

Background

This chapter briefly explains some terms that are used throughout this thesis. Being inter disciplinary in nature, our study uses terminology from different domains such as information retrieval, query processing and Semantic web. So here we will briefly touch upon these domains, in particular the following fundamental areas:

- ✓ Evolution of web
- ✓ Semantic Web
- ✓ Resource Description Framework (RDF)
- ✓ RDF Schema
- ✓ RDF Layers VS RDF Schema Layer
- ✓ The Data Model-RDF Graph
- ✓ Ontologies as a specification mechanism
- ✓ Lexical Database WordNet

The organization of above areas is in accordance with the way they lay the foundation of the research study. Firstly, as the problem is about information retrieval, it is important to have knowledge about the evolution of web, Next, we will see the details of semantic web and its structure. In the details these areas are explained below:

2.1 Evolution of Web

Web 1.0 is known as readable web where only limited interaction is possible among users and websites. Users were only capable to receive information they could not post reviews or feedback. Whereas web 2.0 is recognized as writeable web. Users were allowed to interact with each other and provide feedback. Common examples of web 2.0 are Wiki, Flickr and Facebook. While web 3.0 is known as executable web that provides dynamic applications and machine to machine interaction. Machines can interpret and generate useful information like human. A digital video recorder known as Tivo is a famous example of web 3.0. Its recording program can search the web based on your preferences [15].

2.2 Semantic Web

“Semantic web is known as an extension of web”. It promotes common data format. The fundamental standard of semantic web is resource description framework (RDF)."The semantic web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries"[17]. In semantic web, data can be processed by machines [18]. In 2001 Berners-Lee shows the evolution of web to semantic web. In 2015, semantic web markup was presented in more than four million websites. W3C is taking care for the Standardization of Semantic Web [19].

2.3 Components of Semantic Web

Semantic web comprises multiple formats and technologies that enable it [17]. Linked data and its structuring is supported by the technologies that provide the description of concept and relationship between concepts within a given domain. “These technologies comprise the following:

- Resource Description Framework (RDF), a general method for describing information
- RDF Schema (RDFS)
- Simple Knowledge Organization System (SKOS)
- SPARQL, an RDF query language
- Notation3 (N3), designed with human-readability in mind
- N-Triples, a format for storing and transmitting data
- Turtle (Terse RDF Triple Language)
- Web Ontology Language (OWL), a family of knowledge representation languages
- Rule Interchange Format (RIF), a framework of web rule language dialects supporting rule interchange on the Web

The architecture of semantic web is presented Figure 2.1 by **Semantic Web Stack**, also known as **Semantic Web Cake** or **Semantic Web Layer Cake**”.

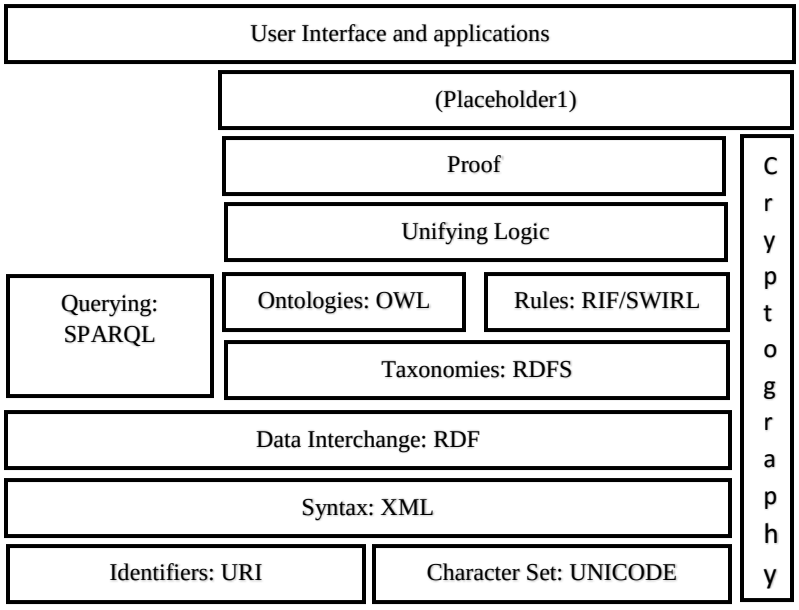


Figure 2.1 Semantic Web Stack

2.4 Resource Description Framework (RDF)

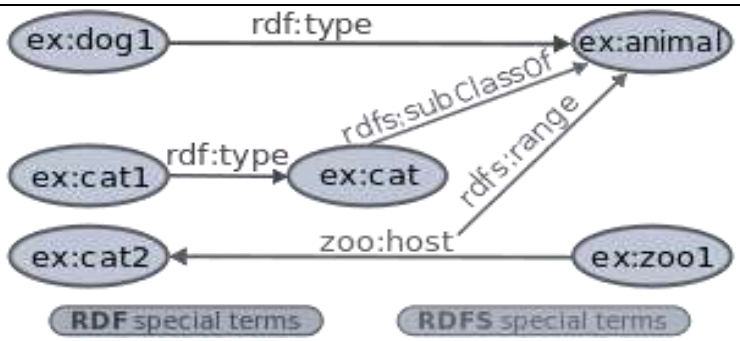
"It is a standard model for data interchange on web". In spite of different schemas, RDF supports data merging. It does not require all the data consumers to be reformed for the evolution of schemas [20]. In RDF, linking structure consists of URIs which define the relationship between two concepts and makes an RDF triple. RDF makes a mixture structure and semi structure data and share it to various applications. RDF defines a directed label graph. Graph nodes are represented to resources and edges are represented to relationships between these resources.

This *graph view* is the easiest model of RDF and is commonly used for visual explanations. Following is an example of a document and its RDF graph and it's RDF in Turtle format

Table 2.1 English Document & its RDF Representation and RDF Graph

<table>
<tr><th>In English</th><th>The graph</th></tr>
<tr><td>

Dog1 is an animal
Cat1 is a cat
Cats are animals
Zoos host animals
Zoo1 hosts the Cat2

</td><td> 
</td></tr>
<tr><th colspan="2">RDF/turtle</th></tr>
<tr><td colspan="2">@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix ex: <http://example.org/>.
@prefix zoo: <http://example.org/zoo/>.
ex:dog1 rdf:type ex:animal .
ex:cat1 rdf:type ex:cat .
ex:cat rdfs:subClassOf ex:animal .
zoo:host rdfs:range ex:animal .
ex:zoo1 zoo:host ex:cat2 .</td></tr>
</table>

2.5 RDF Schema

"RDF is known as a universal language to define resources in vocabularies. Semantics or not defined or assumed by RDF [22, 23].
By using RDF user can define:

- Classes and Properties
- Class Hierarchies and Inheritance
- Property Hierarchies

We must differentiate:

- Concrete things i.e. individual objects in the domain such as Operating System, Madiha Ilyas etc., and
- Sets of individuals sharing properties called Classes such as Assistant Professors, Programs etc.
- Individual objects of a class are known as instances of that class.
- The relationship among classes and their instances is defined by rdf:type property.

Classes: Groups are made by resources which makes a class. The members of a class are defined as *instances* of the class. A Classes is also categorize as a resource. Classes are described using RDF properties and mostly identified by IRIs .To make a resource as an instance of a class, the rdf:type property is used.

- A class may be an instance of itself or a member of its own class extension.
- Group of resources make a class called rdfs:Class.
- The rdfs:subClassOf is a property used to declare a class as a subclass of another class. If a class A is a subclass of another class B then all instances of A will also become the instances of B. Inverse of a subclass is called superclass [21].

Core Classes of RDF: Following are the core classes of RDF

- rdfs:Resource: It is a class of all resources
- rdfs:Class: It is a class of all classes
- rdfs:Literal: It is a class of all literals (e.g. strings)
- rdf:Property: It is a class of all properties.
- rdf:Statement:It is a class of all reified statements
- rdfs: Resource: In RDF, all described things are called *resources.* Every resource becomes the instance of the class rdfs:Resource. It is known as class of everything. All classes are called subclasses of rdfs:Resource. It is an instance of rdfs:Class [21].

rdfs:Class: It is a class of RDF resources. Group of resources define an RDF class. It is an instance of itself [21].

rdfs:Literal: String, integer and property values are known as literal values. The class of all literals are known as rdfs:Literal. Class of literal is an instance of rdfs:Class. It is a sub class of RDF resource class.

rdf:Property: It is an instance of rdfs:Class. It is the class of all properties [21]. rdfs:range and rdfs:domain are the instances of rdf:Property. These things define that the values of a property are instances of one or more classes.

rdf:Statement: A token of RDf triple makes a RDF statement. Rdf:statement is an instance of RDf class. It is a class of all statements [21]. rdf:subject: It is an instance of RDF property. Subject of a statement is defined by this property. A triple of the form: {A rdf:subject S} States that A is an instance of rdf:Statement and that the subject of A is S. RDF domain of subject is RDF statement and the RDF range of subject is RDFs resource. rdf:predicate is an instance of RDF property .Predicate of a statement is defined by this property. A triple of the form: {A rdf:predicate P} states that the predicate of A is P. rdf:object is an instance of rdf:Property. Object of a statement is declared by using it. A triple of the form: {A rdf:object O} states that the object of A is O. Domain and range of RDF predicate and object hold the same as subject of a statement."

2.6 RDF Layers VS RDF Schema Layer

Following diagram is showing the RDF representation and RDF Schema of the statement "Advance Databases is taught by Atif Kamal", the schema is itself written in a formal language. [22, 23]

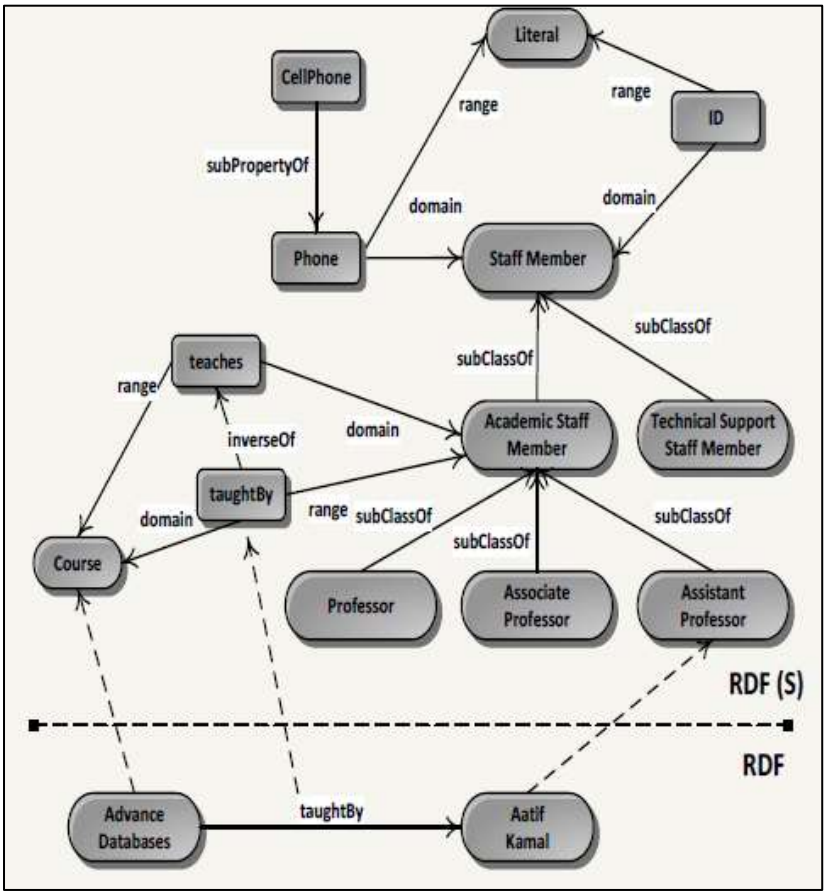


Figure 2.2RDF vs RDF (S)

2.6.1 The Data Model-RDF Graph

The data model that is considered in the proposed technique is a graph Gr={V,E,R} where V is the set of URIs which refers to specific elements (concepts, instances, Predicates), E the set of directed edges linking pairs of URIs according to a specific type of relationships (predicates) and R the set of predicates which can be used. Each edge is a triplet of URIs which is oriented, and unique. As an example the graph model adopted by us can be used to represent the simple (RDF) graph represented in Fig. 1 as each node can be identified by a unique identifier and directed edges are labeled by a particular semantic relationship.

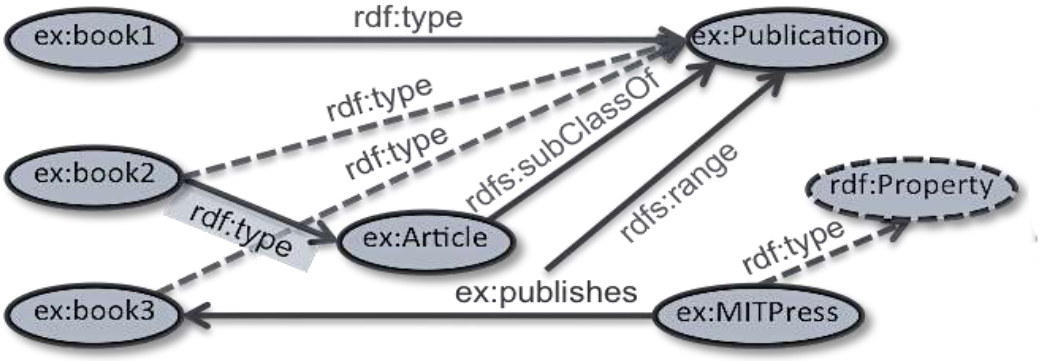


Figure 2.3 RDF Graph

In Figure 4.1 set of vertices comprise:

V= {ex:book1, ex:book2, ex:book3, ex:Article, ex:Publication, ex:MITPress, rdf:Property},

set of edges comprise:

E={rdf:type, rdf:type, rdf:type, rdf:type, rdfs:subClassOf, rdfs:range, ex:publishes, rdf:type}

and set of predicates comprise:

{rdf:type, rdfs:subClassOf, rdfs:range, ex:publishes}

2.7 Ontologies as a specification mechanism

Specification of a conceptualization defines an ontology [24]. A body of properly represented knowledge depends upon the concepts, objects, and all other things that are supposed to be in a domain and the relationships hold among these concepts. This phenomena is known as conceptualization. It is a simplified and abstract view of the domain that we want to use for any particular objective. Each knowledge base implicitly or explicitly is devoted to conceptualization at some extent [24, 25]. "**Ontology** is an explicit specification of a conceptualization. In philosophy, an ontology is a systematic view of existence". In AI "exist" means anything that can be represented.

All the objects that can be defined in a domain make the universe of discourse. Object set and relationships among these concepts make a representational vocabulary. A knowledge-based system use this representational vocabulary. Thus in AI domain, all representational concepts makes the ontology of a system. In a formal way it can be stated that an ontology is a representation of logical theory [24, 25].

Practically, Agents exchange queries and statements by using a common ontology. Ontological commitments enable us to use the shared vocabulary in a reliable way. There is no need to share the knowledgebase for sharing a vocabulary, because one agent may know the things which other does not. To constrain an ontology could not answer all the queries that can be formed in a shared vocabulary. In short, ontological commitment is assurance of consistency, but not completeness, regarding queries and assertions". Following is an example of ontology:

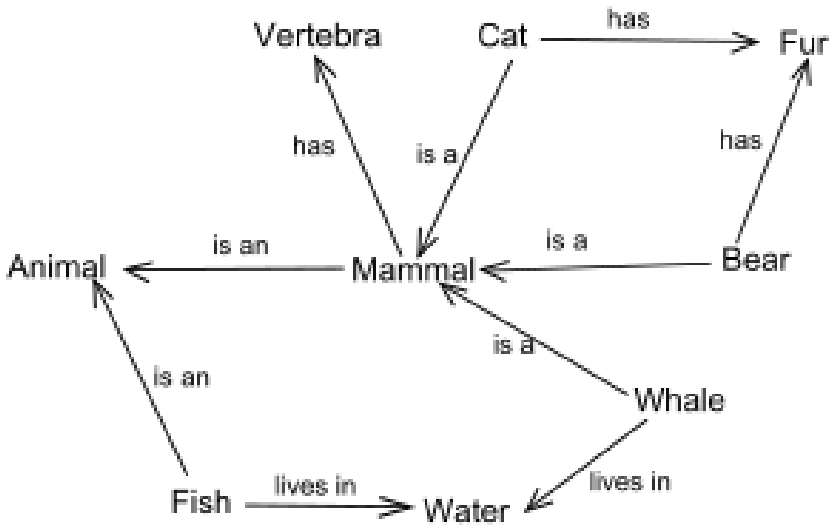


Figure 2.4 Example of Ontology

2.8 Domain ontology

A domain-specific ontology embodies concepts which belong to part of the world. Domain ontology provide the particular meaning of the concepts according to that domain. For example, the word *card* represents different meanings in different domains. In poker domain the ontology would model the term card as a playing card while in computer domain term card would be modeled as punched card or video card.

Domain ontologies represents terms/concepts in a very precise manner so ontologies are mostly incompatible. So for general representation ontologies need to be merged. In one domain there are many ontologies available due to different languages or different perception of the domain on the bases of cultural background or ideology. This gives a challenge to the ontology designers.

Ontologies that do not have a common foundation ontology are very difficult to merge. Two or more ontologies having common foundation ontology can be merge automatically. A lot of research work is being done in the domain of merging ontologies [26].

2.8.1 Word Net

"WordNet is a large lexical database of English. There are groups of nouns, verbs, adjectives and adverbs called synsets". Semantic and lexical relationships exist between these synsets. Its structure supports natural language processing [26]. By its sketch, it looks like a thesaurus. It makes groups of words according to their meanings. However, there exists some important

distinctions between WordNet and thesaurus. First, WordNet interlinks specific senses of words instead of just words. By doing so, words having close proximity become semantically disambiguated. Secondly, it mentions the relationships between concepts, whereas in a thesaurus the grouping of terms is made on the basis of meaning similarity instead of explicit pattern [26]. Following figure is showing the WordNet hierarchy:

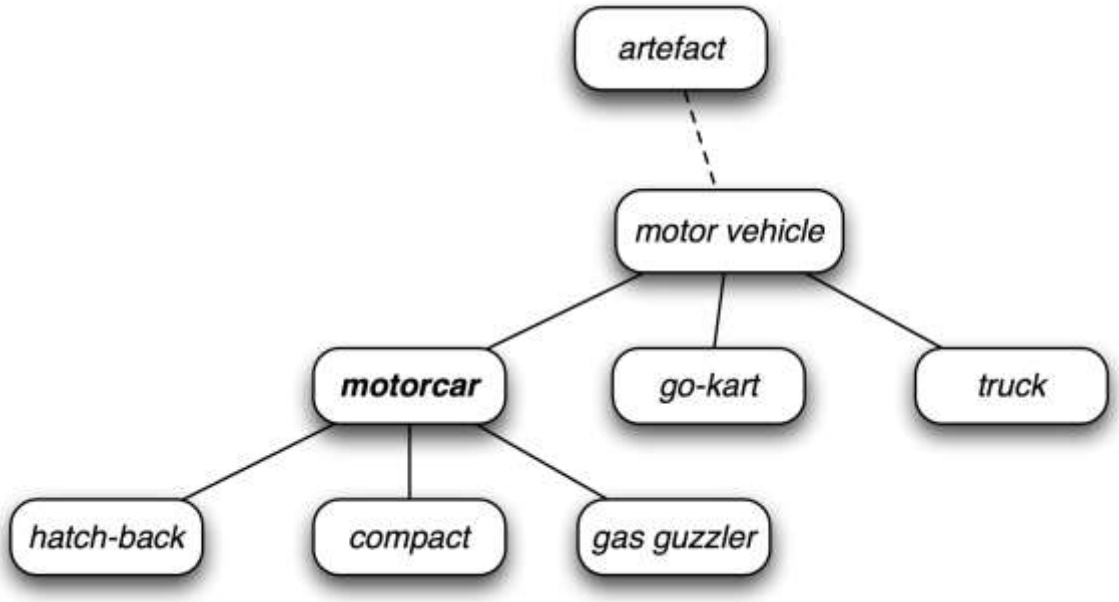


Figure 2.5 Fragment of WordNet Hierarchy

Structure: In WordNet the main relationship is synonymy [26], as between the terms Midday and noon or purchase and buy. Synonyms indicate the same concept and they are compatible in many contexts. Groups of synonyms called synsets. Each synset is linked with other synsets by using conceptual relations.

Relations: "Super-subordinate relation also called hyperonymy, hyponymy or ISA relation are the most widely used relations exist among sysets [26]. It establishes relation between general synsets such as furniture to more specific synsets like piece_of_furniture bed and bunkbed. So by WordNet the concept furniture includes bed, which includes bunkbed. In other way we can say that, bed and bunkbed concepts makes the concept furniture. All noun hierarchies eventually reach to the root. Hyponymy is a transitive relation means if a round table is a kind of table, and if a table is a kind of furniture, then round table is a kind of furniture. WordNet differentiates between common noun and specific persons, countries and geographic entities. Therefore, armchair is a type of chair and Nawaz Sharif is an instance of a president. Instances are represented by leaf nodes in their hierarchies."

Chapter 3

Literature Review

Over the past decade, many approaches have been proposed that exploits ontologies to calculate the semantic similarity between conceptual graph representations to improve the precision of information retrieval. Most of the proposed systems just based on conceptual similarity. The focus of this chapter is to present the overview of the existing techniques/factors that exploits semantics during information retrieval. The most advanced techniques are as follows:

- ✓ Ontology Development
- ✓ Linked data: Publishing Structured Data
- ✓ Ontology-based semantic similarity
- ✓ Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy
- ✓ Semantic Similarity Methods in WordNet
- ✓ Structure based measure
- ✓ Graph based measure
- ✓ Hybrid Semantic Similarity
- ✓ Information Retrieval with Conceptual Graph Matching
- ✓ Ranking of Web Documents using Semantic Similarity
- ✓ Semantic Concept Search based on Triple Matching Approach
- ✓ Learning Vector Representations for Similarity Measures

These approaches are discussed in the following sections.

3.1 Ontology Development

Researchers can share information in a domain by designing a common vocabulary i.e. ontology. It presents the definitions of the concepts and the relationships between them in such a way that machines can understand and interpret it [25].
There are many important reasons for ontology development:

- Develop common understanding about concepts among software agents and people.

- Make the domain knowledge reusable.
- Generates explicit domain assumptions.
- Put domain knowledge separate from operational knowledge.
- Make analysis of the domain.

The formal explicit description of the concepts of a particular domain define an ontology. The features and attributes of the concepts define the properties of the concept [25].
"In practical ontology development includes

- Deciding classes of the ontology,
- Arranging the classes in a taxonomic (subclass–superclass) hierarchy,
- Deciding slots and describing allowed values for these slots,
- Filling the values in slots for instances.
  Knowledge base is created by defining the instances of the classes. Deciding the slot values and additional slot restrictions."

We can then create a knowledge base by defining individual instances of these classes filling in specific slot value information and additional slot restrictions".

3.2 Linked Data

Linked data is a process to define the structure for publishing data. So that data become more useful for software agents and semantic queries can be resolved easily. It is based on standard web technologies like RDF, HTTP and URIs. It makes web pages automatically readable by machines. It makes a link between data at multiple sources [26].

Tim Berners-Lee defined following four principles of linked data in [27]

1. Things should be identified through URIs.
2. Things should be interpreted or dereferenced through HTTP URIs.
3. Provide useful information through open standards such as RDF, SPARQL, etc.
4. Refer to things using HTTP URI on the Web.

3.3 Ontology-based semantic similarity

[29] has presented a taxonomical feature based measure. Set of features is designed on the basis of subsumers of the concepts in a taxonomy hierarchy. Subsumers describe the concepts at different level of generality. This measure based on the taxonomic hierarchy instead of corpora dependency or parameter tuning. It is very efficient in computation as it explore on taxonomical branches and compute the mathematical properties for similarity computation like Everitt, et al., 2001; O'Sullivan, et al., 2005.
In [30] authors have discussed many approaches for finding similar concepts in and between ontology. This survey exploits various similarity method for query expansion to improve the information retrieval systems.The experiments provide better correlation values and gives direction of using them in ontology based information retrieval systems.

3.4 Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy

In [32] authors have proposed a lexical taxonomy structure and corpus statistical information for computing similarity among concepts. Computational evidence is derived from distributional analysis of domain data, which is used to compute the semantic distance between concepts. It

inherits edge counting scheme which is enhanced by information content based approach. Experiments have shown that the approach gives promising results. It gives 83% correlation on with the word pair data set.

3.5 Semantic Similarity Methods in WordNet

For improving the information retrieval process on web, this research has exploited information embedded in ontology and captured by using semantic similarity methods. By using this observation a mechanism known as semantic similarity retrieval model (SSRM) has proposed in [33]. This novel method incorporates conceptual similarity into retrieval mechanism. SSRM can work with any taxonomic ontology. Experiments have shown its excellent performance over many existing information retrieval methods.

3.6 Structure Based Measure

In [31] authors have proposed a semantic similarity measure (DOPCA) that uses the structure of ontology. It uses two features to compute similarity between terms. First one is degree of overlap in the path (DOP) and second one is depth of least common ancestor nodes. DOP uses both intersection and union of nodes and edges to compute the overlap in the path of two concepts. And second one is based on the assumption that concepts which lie deeper in the ontology would have more similarity with each other.

3.7 Graph Based Measure

To find the semantic similarity among two terms by using all senses of concepts, a lexical ontology based measure has proposed in [28]. This measure can consume the useful information available in the WordNet. The proposed measure is as follow:

$$\mathrm{sim}_{\mathrm{proposed}}(c_1, c_2) = \alpha \frac{\gamma}{\mathrm{dist}_{jc}(c_1, c_2) + \gamma} + (1 - \propto) \left(\frac{2 \times \log p(lcs(c_1, c_2))}{\log p(c_1) + \log p(c_2)} \right) \quad \textbf{(3.1)}$$

Four popular similarity methods (Resnik, Lch, Wup and Jiang & conrath) were implemented. An experiment had conducted with these Semantic Similarity Measures for English concepts. Based on comparison between experimental results and standard results the proposed semantic similarity measure was found performing well among four semantic similarity measures.

To compute the similarity between gene ontology terms, a simple measure and algorithm is proposed in [36]. Similarity among two terms is computed using length of shortest path and depth of lowest common ancestor in the ontology. This novel measure has eliminated the drawback of Nagar's method. This method neglects the depth feature. A small drawback of Wang J.Z.'s method has also been resolved by this method. However it is not assure that all the clusters are meaning full because gene ontology is not a standard ontology. And this measure is just based on GO.

3.8 Hybrid Measure

A hybrid measure known as *WNOntoSim* to compute semantic relatedness among ontologies using WordNet is proposed in [35]. It considers semantic information and ontology structure for computing semantic relatedness. Semantic similarity at elemental level and structural level is computed using WordNet as information source. Experimental results have shown that it performs

well in general ontology. But the drawback is that it cannot exactly compute the similarity among named entities due to the coverage problem of WordNet.

3.9 Information Retrieval with Conceptual Graph Matching

A system based on conceptual graph of a document and query is proposed for improving the precision of information retrieval in [37]. Proposed system uses a useful characteristic known as dice coefficient and many characteristics of conceptual graph known as conceptual similarity and relational similarity. Proposed measure is good for comparison of text especially short text. Currently the system has adopted is-a hierarchy and synonymy. The method is being used for text mining and document classification.

3.10 Semantic Concept Search based on Triple Matching Approach

For the given query, search is performed by matching Triples in spite of keywords [50]. User gives query in triples form. The query is expanded by different properties like synonyms and semantic neighborhood. Computation is based on distance based approach and domain ontology. Relevance is computed and relevant documents are find out. The identified documents are ranked by their relevance to user query. The relevance is computed by tf.Idf scheme for triples rather than for keywords.

Semantic concept search systems for searching of concepts from the knowledgebase has be proposed in [39]. The approach is able to automatically formulate natural language query to triple representation and uses the produced triple to make inference on what the answer will be based on triple matching. The approach has contributed by automating the semantic search systems. The system has proposed semantic concepts search that does not depend on the usual inference engines used by other researches.

ReWOrD measure relatedness among RDF predicates which are used to define statements about the terms being compared [38]. Informativeness is measured by the Predicate Frequency Inverse Triple Frequency (PF IT F). To refine relatedness it considers paths between concepts. It does not require preprocessing of data and can exploit any source of knowledge for which a SPARQL endpoint is available.

3.11 Learning Vector Representations for Similarity Measures

A novel approach that learns new vector representations for similarity measures is proposed in [42]. Abstractly, the system can be viewed as a two-layer Siamease neural network, where the first layer is the original term vector with term-level features as input and the second layer captures the relations among different terms to form the output vectors. The system is a simplification of several existing frameworks. When learning extended term vectors, the model parameters form a sparse Mahalanobis distance matrix. System also subsumes the recently proposed term weighting learning framework, TWEAK. In both settings, experiments have shown that the proposed system performs better than other methods for measuring document similarity, and is comparable to HDLR in learning concept vectors.

3.12 Entity Extraction: From Unstructured Text to DBpedia RDF Triples

An automatic triple extraction system is proposed in [41]. It can extract RDF triples from unstructured text. It incorporates text processing modules which consists of semantic parser and a coreference solver. All the actions and properties described by multiple sentences are grouped by

coreference chain and transformed into RDF triples. This system has been applied to over 114,000 Wikipedia articles and it could extract more than 1,000,000 triples.

3.13 Critical Analysis

Semantic closeness is known as semantic similarity [43]. For calculating relatedness among terms in an ontology, several methods have been presented in the past few years. In a broad sense, we can divide these methods into three types: structure based approach, information content based approach and hybrid method which utilizes multiple properties of an ontology. Structure based approach uses distance measure to compute the relatedness among two concepts of ontology. Few methods of this type are "Rada et al which considers the shortest path among two concepts [7], Wu and Palmer considers the distance from root node to lowest common ancestor node [8], Leacock and Chodorow considers the no. of nodes in the shortest path [9], Mubaid and Nguyen considers the commonality feature [10], Wang et al. [11] and Zhang et al. [12] combine multiple features to calculate the similarity among two nodes of an ontology". These methods are very useful for variety of problems, but still there exist some drawbacks. None of the above mentioned methods take into consideration the problem of multiple lowest common ancestor nodes. IC based methods compute relatedness by measuring IC of the deepest common ancestor of the two concepts. For example Resnik et al [14], Lin et al. [15] and Jiang and Conrath [16].But the problem with IC based methods is that these are affected by shallow annotation [11, 18]. Hybrid methods consider multiple features like ontology structure, IC, depth of LCA node etc. For example Yin and Sheng [20]. Hybrid methods are usually complex and do not give good performance. With this observation we have proposed a similarity measure called DOPCA, which comprises two distinct features: degree of overlap in path (DOP) and information content of most informative common ancestor, which overcome the drawback of varying link distances in the ontology.

Chapter 4

Proposed System Design

In this section we discussed the proposed semantic similarity measure to compute the relatedness between query and documents by considering the context. It employs distinct features, such as degree of overlap (DOP) and information content of most Informative Common Ancestors (ICCA) to find out the semantic relatedness between two RDF triples using domain ontology. Domain ontology provides the meaning of a concept with respect to that domain. We have used WordNet as a domain ontology. First of all concepts similarity is computed. Secondly the relationship (predicate) similarity is computed. Then in the last step, the similarity of RDF triples is computed to find out the relatedness between two RDF triples. The system is implemented by using open source Semantic Measure Library and Eclipse IDE and Apache-Jena API.

4.1 Proposed Methodology

Following is a diagram of proposed system methodology. RDF query and documents are provided as input to the proposed system. Then the system computes the subject predicate and object similarity by using the proposed similarity measure DOPIC. Finally the documents are ranked on the basis of their relatedness to query. In the following subsections we will discuss the detail of each step.

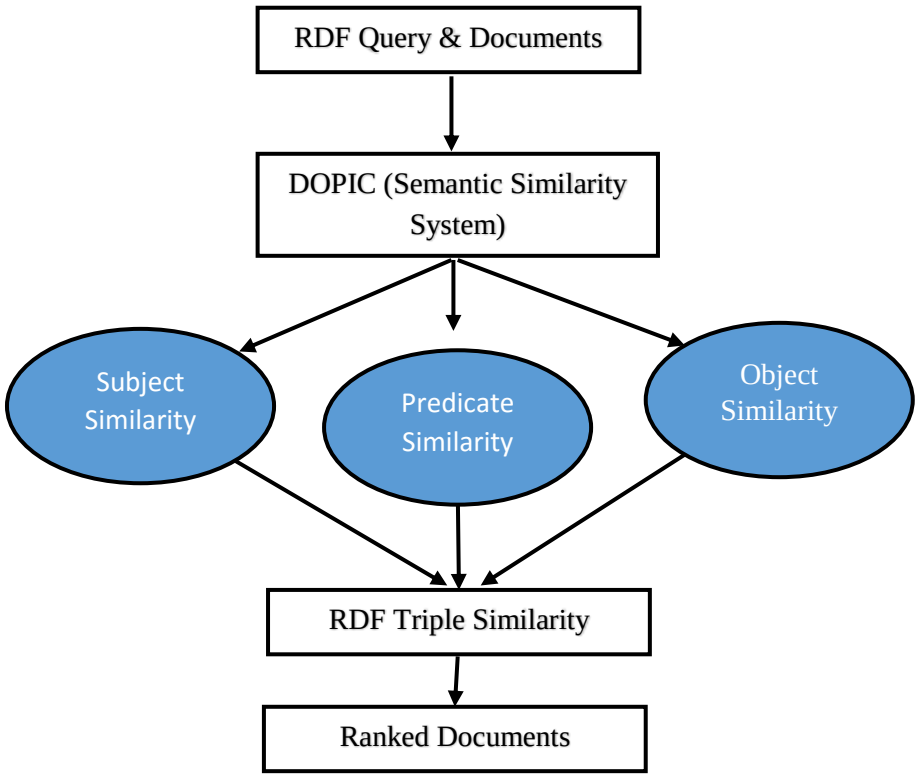


Figure 4.1 Proposed Methodology for measuring Semantic Similarity between RDF query and documents

4.2 Graph based proposed Semantic similarity Measure

We have designed a graph based semantic similarity measure to compute the relatedness between concepts and relationships of the ontology. Semantic similarity/relatedness refers to semantic closeness or proximity [43]. In a broad sense, we can divide similarity measures into three types: structure based approach, information content based approach and hybrid method which utilizes multiple properties of an ontology. Structure based approach uses distance measure to compute the relatedness among two concepts of ontology. These methods are very useful for variety of problems, but still there exist some drawbacks. None of these methods take into consideration the problem of multiple lowest common ancestor nodes. IC based methods compute relatedness by measuring IC of the lowest common ancestor node of the two terms. But the problem with IC based methods is that these are affected by shallow annotation [11, 18]. Hybrid methods consider multiple features like ontology structure, IC, depth of LCA node etc. Hybrid methods become complicated and do not perform well. With this observation we have proposed a method called DOPCA, which combines two features: degree of overlap in path (DOP) and information content of most informative common ancestor (ICCA), which overcome the problem of varying link distances in the ontology. The detail of both techniques is given in following section.

4.3 Concept Similarity

Concept similarity has been computed by the proposed semantic similarity measure which consists of two parts: The first part of the proposed semantic similarity measure computes the degree of overlap in the path (DOP) of two concepts of domain ontology by using the structure of ontology [31]. DOP is based on the following assumptions that Ontology G is a directed acyclic graph where set of vertices (V) of a concept comprising the term itself and all its ancestors, similarly the set of edges (E) of a concept will be a set of all the edges connecting the concepts in the set of vertices. Definition of Degree of overlap follows the Lin [45] semantic similarity measure which is the ratio of the information required to describe the commonality of two concepts and the information required to fully describe them [31]. A segment of WordNet is shown in fig.4.2 which would be used in following segments for computation.

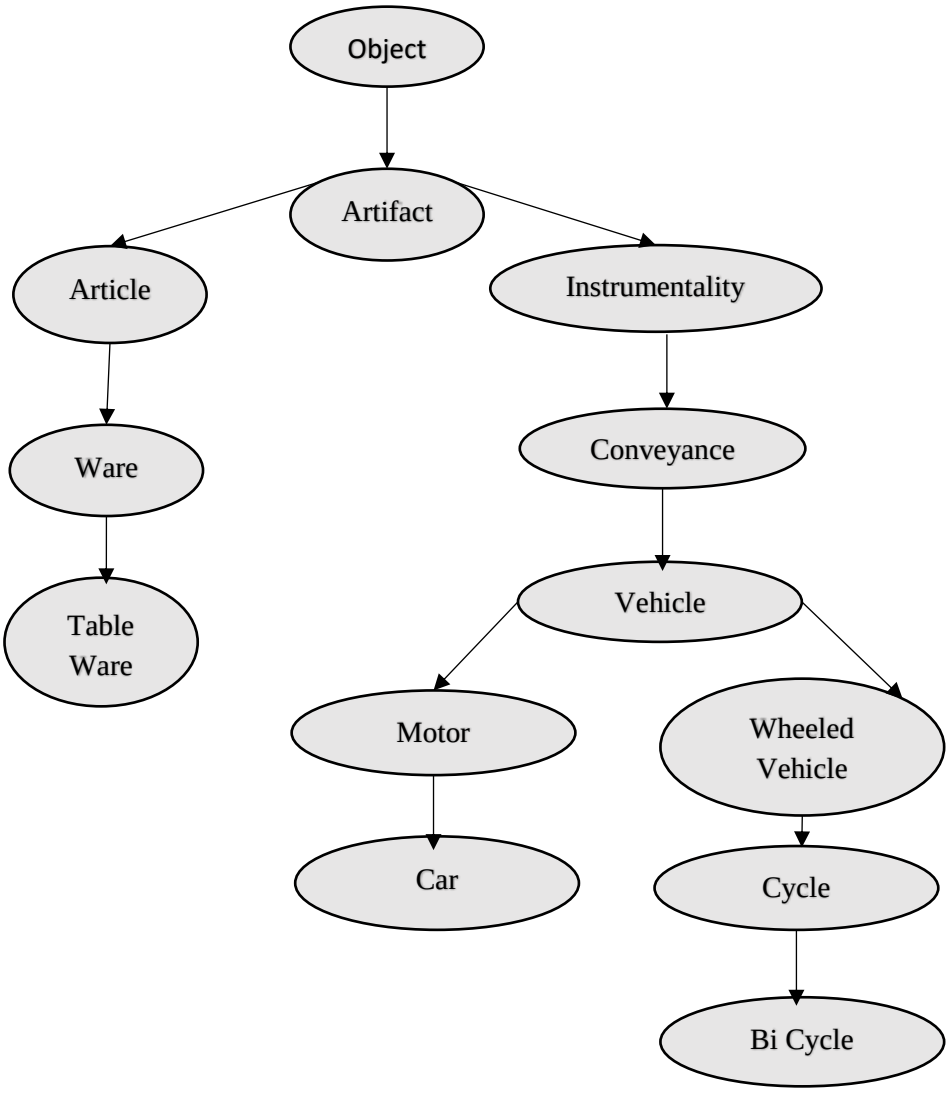


Figure 4.2 Ontology Segment concerning vehicles

4.4 Degree of Overlap (DOP)

The combination of taxonomic structure and empirical probability estimates provides a way of using static knowledge structure to multiple contexts [44].By considering this we have used DOP measure to explore the taxonomic structure. Degree of overlap in the path of two concepts A and B of the ontology can be computed by the following equation.

$$Sim_{DOP}(A,B) = \frac{|V_{A\cap B}| + r * |E_{A\cap B}|}{|V_{A\cup B}| + r * |E_{A\cup B}|} \quad (4.1)$$

Where V and E are represented to vertices and edges of the concepts respectively. And $r = W_e/W_v$ where W_e is the weight of edges and W_v is the weight of vertices. If two concepts would have more than one lowest common ancestor (LCA), the path from root to all LCA nodes will be computed which would solve the problem of multiple lowest common ancestors.

Example

Suppose we want to compute the degree of overlap between the car and bicycle terms of the ontology represented by *Figure 4.2*.

First of all we will compute the common vertices of both terms $(V_{A\cap B}) = 5$, secondly we will compute the common edges of both terms $(E_{A\cap B}) = 4$, thirdly we will compute the ratio $r = W_e/W_v$ since currently we are dealing with is-a edges only so we are considering the weight 1 for all edges hence r=0.8 similarly we will compute $V_{A\cup B}$ and $E_{A\cup B}$ to finally compute the DOP.

$$Sim_{DOP}(Car, Bicycle) = \frac{(5 + 0.8 * 4)}{(10 + 0.8 * 9)} = 0.4 \quad (4.2)$$

4.5 Information Content of Most Informative Common Ancestors (ICCA)

The second part of the proposed semantic similarity measure uses information content to compute the similarity of two concepts. IC is not affected by varying link distances in the ontology [44]. It computes the information content of the most informative common ancestor (ICCA) of the two concepts of ontology. The similarity of two concepts depends on the specific concept that subsumes them both in the ontology. Let C be a concept in ontology G and **P(C)** is the probability of concept C then information content of concept C can be defined as $-log\ P(C)$ [30].

$$sim_{ICCA}(A, B) = -\log P(C) \quad (4.3)$$

Where C is the most informative common ancestor of A and B.

Example

The most informative common ancestors of the terms ***Car*** and ***Bicycle*** is ***Vehicle*** shown in Fig. 2 so ICCA of the two mentioned concepts is 0.34 which is computed by using Semantic measure library [48].

$$sim_{ICCA}(Car, Bicycle) = 0.34 \quad (4.4)$$

4.6 DOPIC Similarity Measure

By combining the above mentioned degree of overlap in the path and information content of most informative common ancestors of two concepts, we have computed the similarity of two concepts of ontology as follow [31]:

$$\mathbf{sim_{DOPIC}(A, B)} = \boldsymbol{sim_{DOP}(A, B)} + \boldsymbol{sim_{ICCA}(A, B)} \quad (4.5)$$

To normalize the above equation between (0, 1) we have added weights to the above equation as follow:

$$\mathbf{sim_{DOPIC}(A, B)} = \boldsymbol{W_i} * \boldsymbol{sim_{DOP}(A, B)} + \boldsymbol{W_j} * \boldsymbol{sim_{ICCA}(A, B)} \quad (4.6)$$

The value of weights would be between (0, 1) as the sum of W_i and $\mathbf{W_j}$ equals to 1. It is recommended that $\mathbf{W_i = W_j = 0.5}$ so that we can equally consider both attributes of the proposed measure. System will not be affected by shallow annotations in that case DOP will play its role. The edge counting and information content based methods work by using the structure of ontology i.e. position of concepts in the hierarchy and IC of the concepts. These methods are best for comparing the concepts within an ontology as structure and IC of different ontologies are not directly comparable [12]. Therefore the proposed measure comprises both the best features to compare the concepts of same ontology.

4.7 Relationship Similarity

Similarity between two relationships will be computed the same as that of concepts. As for concepts we have used noun taxonomy of WordNet [46], but for relationship we have used verb taxonomy of WordNet.

4.8 RDF Triples similarity

Finally the similarity between two RDF triples will be computed as follow:

$$\mathbf{Sim_{DOPIC}(t1, t2)} = \{\mathbf{Sim_{DOPIC}(t1_{sub}, t2_{sub})} * \mathbf{Sim_{DOPIC}(t1_r, t2_r)} * \mathbf{Sim_{DOPIC}(t1_{obj}, t2_{obj})}\} \quad (4.7)$$

Where ***t1*** and ***t2*** are representing to triples and $\boldsymbol{t_{sub}}$, $\boldsymbol{t_r}$ and $\boldsymbol{t_{obj}}$ are representing to the subject, relation/predicate and object of the triple respectively. If we want to compute the similarity between a query and a source consisting of ***m*** RDF triples, then the above equation can be transformed into following:

$$\boldsymbol{Sim_{DOPIC}(q, s)} = \frac{\{\sum_{i=1}^{n} \prod_{j=0}^{m} \boldsymbol{Sim_{DOPIC}}(q_{sub}^i, s_{sub}^j) \boldsymbol{Sim_{DOPIC}}(q_r^i, s_r^j) \boldsymbol{Sim_{DOPIC}}(q_{obj}^i, s_{obj}^j)\}}{\boldsymbol{m*n}} \quad (4.8)$$

Where ***q*** and ***s*** are representing to query and source respectively. Where m*n is giving the total no of triples being compared. Finally the summation is being divided by total no of triples to get the normalized result of relatedness between (0-1).

4.9 Use Case Example

Suppose we want to compute the similarity between following two RDF graphs (Q, S) using WordNet as a domain ontology:

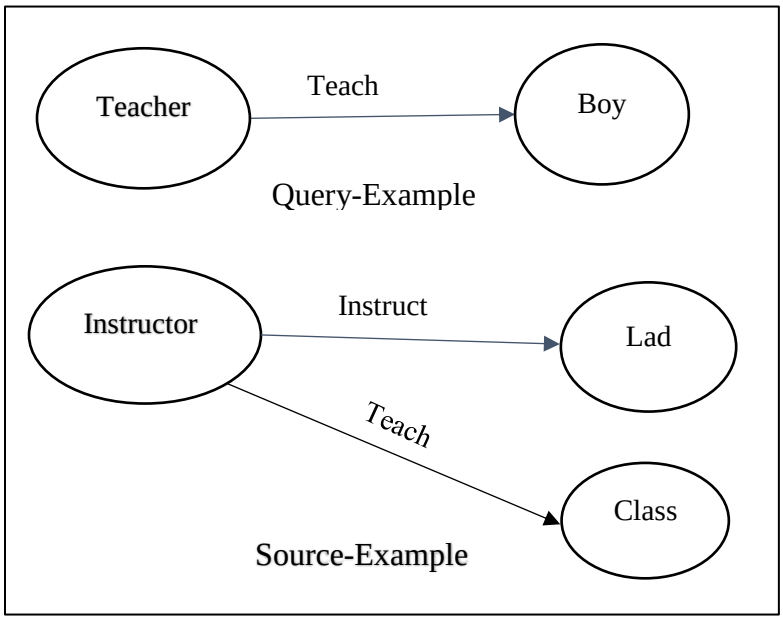


Figure 4.3 Query and Source Example

First of all, concept and relation similarity has been computed using equation No. 4.6. Table 4.1 shows the computed similarity.

Table 4.1 DOPIC Concept/Relation Similarity

Triple ID	Subject Similarity	Relation Similarity	Object Similarity
1	*Sim_{DOPIC} (teacher, instructor)=0.853*	*Sim_{DOPIC} (teach, Instruct)=0.857*	*Sim_{DOPIC} (boy, lad)=0.83*
2	*Sim_{DOPIC} (teacher, instructor)=0.853*	*Sim_{DOPIC} (teach, teach)=1.0*	*Sim_{DOPIC} (boy, class)=0.03*

After computing concept/relationships similarity, Similarity between RDF statements/Triples has been computed using *equation 4.7*

$$\boldsymbol{Sim_{DOPIC}(Q1, T1) = 0.853*0.857*0.83 = 0.60} \quad (4.9)$$

$$\boldsymbol{Sim_{DOPIC}(Q1, T2) = 0.853*1.0*0.03 = 0.03} \quad (4.10)$$

Hence the similarity between two RDF graphs is computed by using *equation 4.8 is as follow:*

$$\boldsymbol{Sim_{DOPIC}(Q, S) = (0.60+0.03)/2 = 0.3} \quad (4.11)$$

Summary

In this chapter a detailed discussion of proposed methodology has been presented. Our proposed semantic similarity measure consists of two distinct features such as DOP and ICCA has been discussed in detail. Finally the concepts and triple similarity has been computed on example concepts and triples respectively.

Chapter 5

Implementation and Evaluation

This chapter has two main parts. The first part will discuss the technical details about the system implementation and the second part will focus on the evaluation of the proposed system.

5.1 System Implementation

This section is further divided into three sub-sections; (i) System Specifications, (ii) Software Specifications, (iii) Sample output of each module, illustrated through a series of screen shots.

System Specifications: The system specifications used in the development of the system are shown in Table 5.1

Table 5.1 System Specifications

Processor	Pentium(R) Dual-Core CPU T4400 @2.20 GHz 2.20 GHz
RAM	4.00 GB
Operating System	Window 7 Professional

Software Specifications: The software specifications used in the development of the system are shown in Table 5.2

Table 5.2 Software Specifications

Development Language	JAVA
API	Apache-Jena
IDE	Eclipse Kepler

Library	Semantic Measure Library 0.7.3 (SML)
Domain Ontology	WordNet-2.0

SML [22] is an open source java library to measure semantic similarity between concepts using domain ontology.

5.2 Sample Output

As discussed in System Methodology Chapter, to compute the similarity between two RDF triples, the similarity between Subject, Predicate and object has been computed respectively. Next Figure 5.1 is showing the semantic similarity between two given Subject Concepts, Figure 5.2 shows the semantic similarity between two given Predicates and Figure 5.3 illustrates the semantic similarity between two given object concepts.

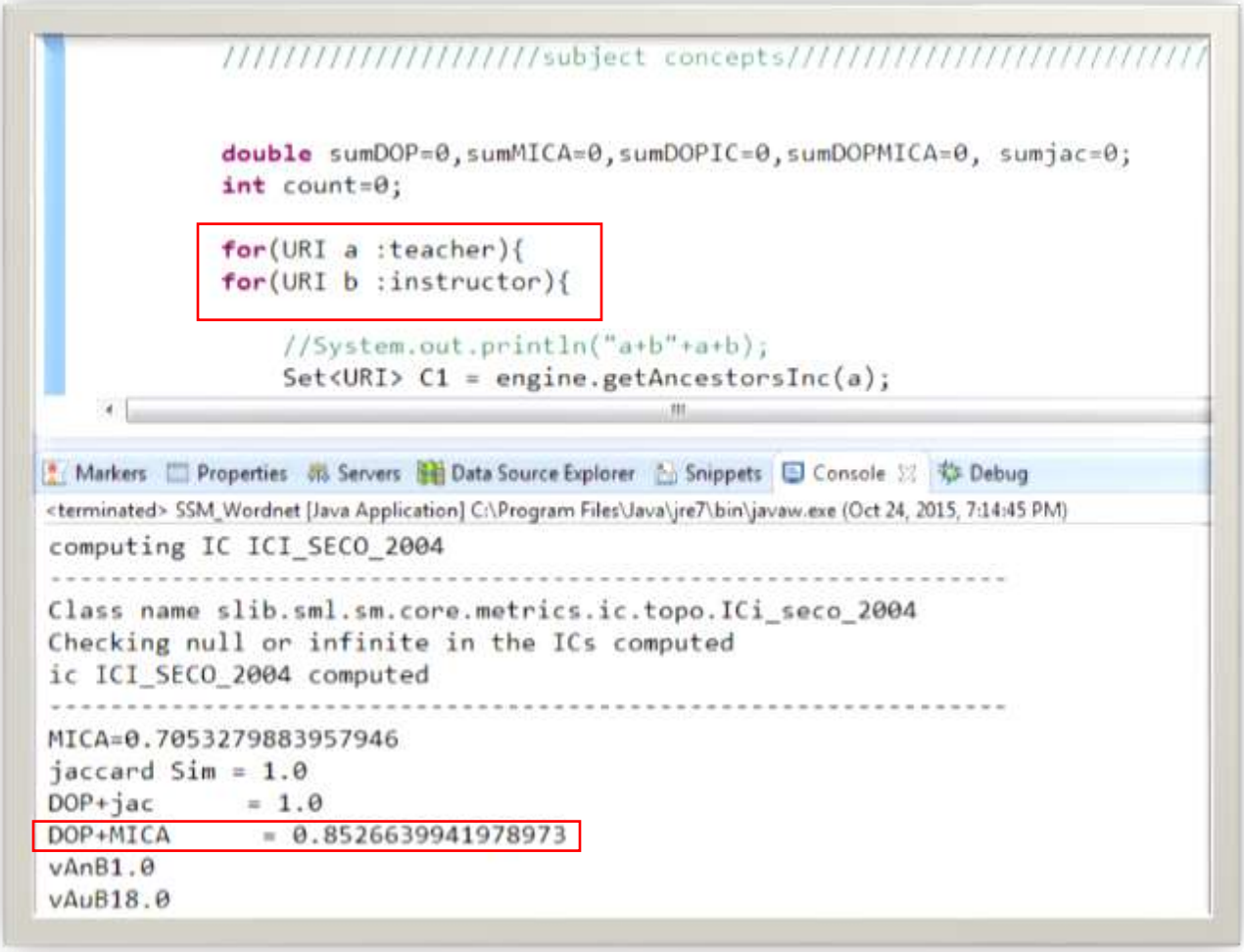


Figure 5.1 Illustration of Subject Semantic Similarity

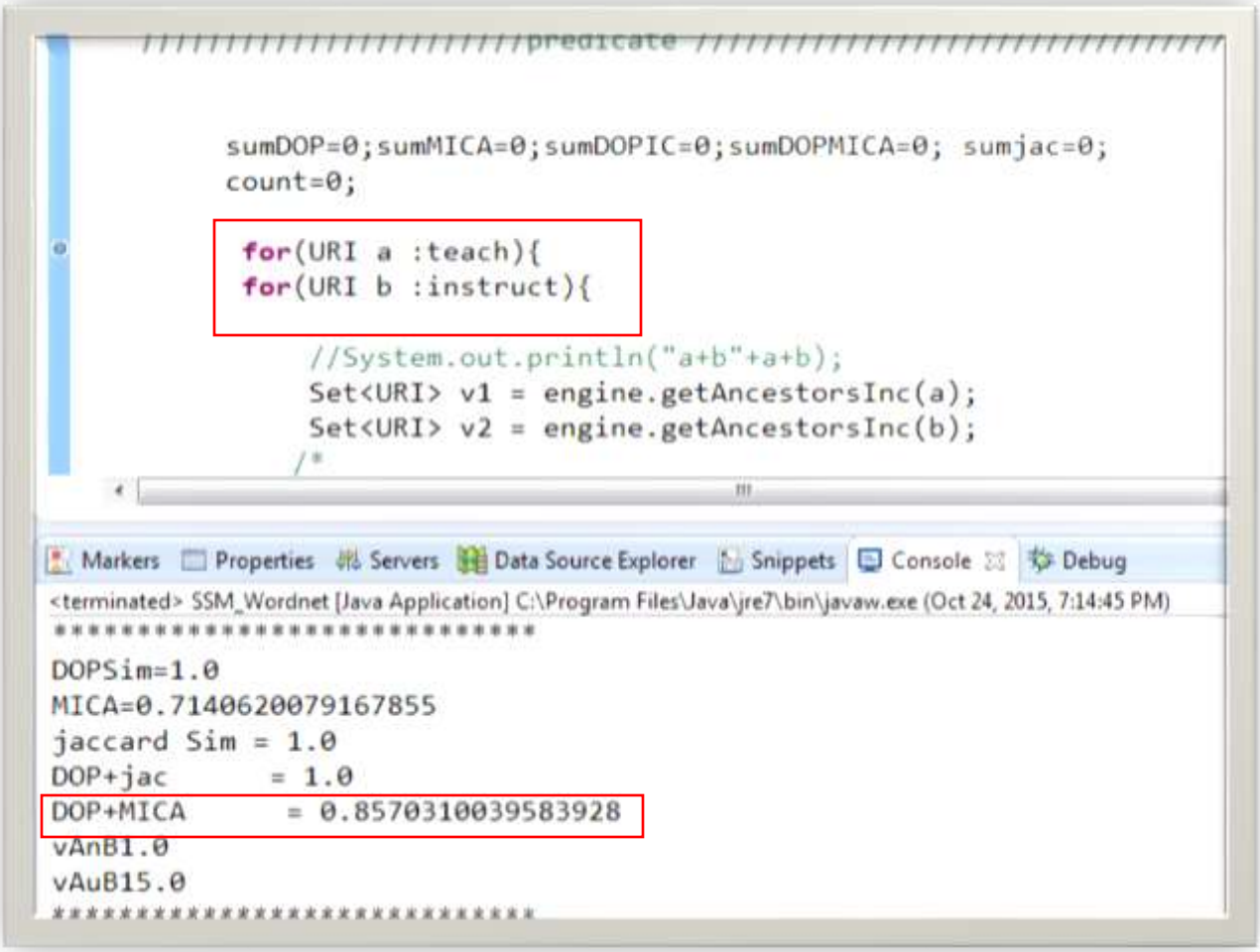


Figure 5.2 Illustration of Predicate Semantic Similarity

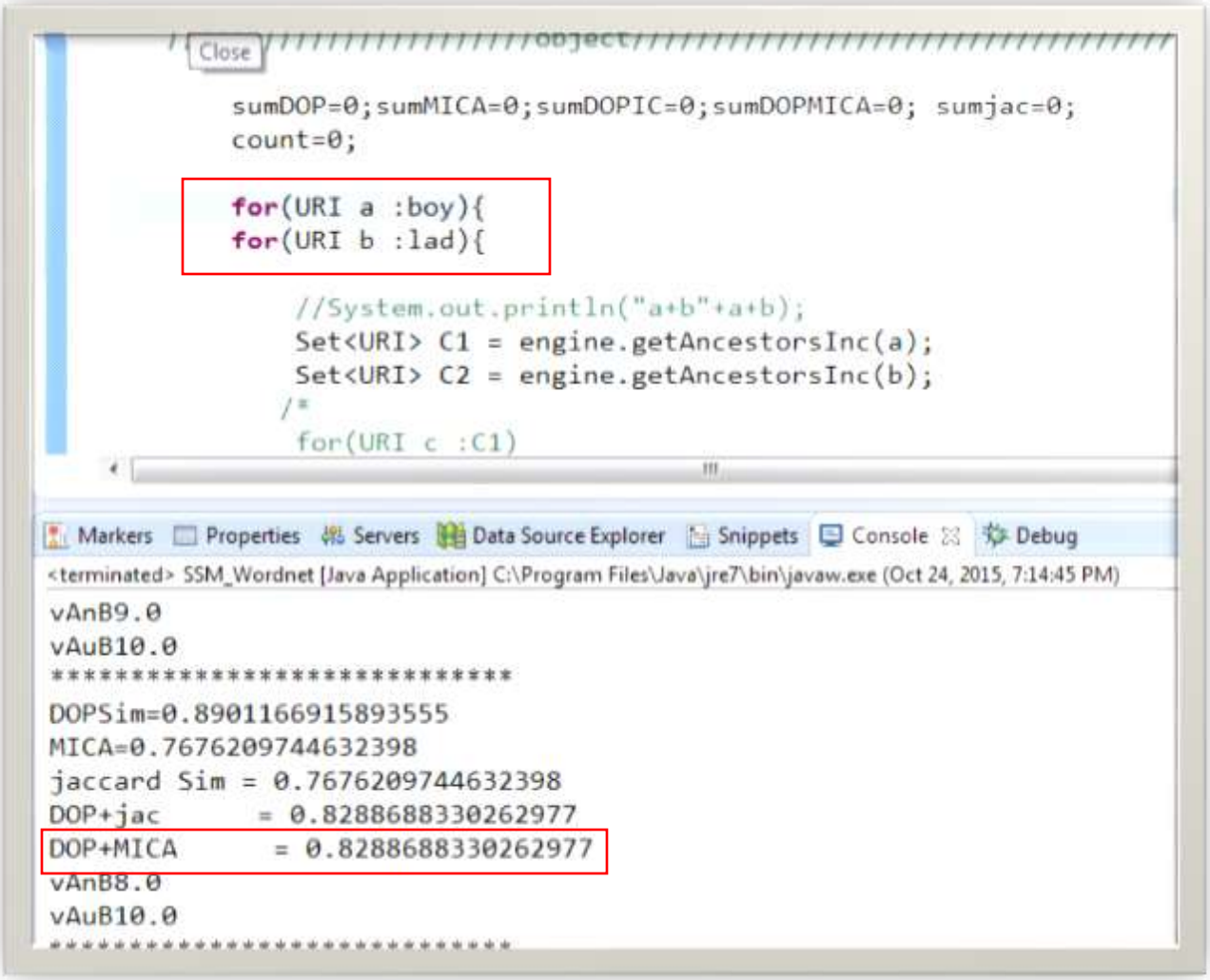


Figure 5.3 Illustration of Object Semantic Similarity

5.3 Evaluation Approaches

The proposed semantic similarity measure was evaluated at two levels: concepts and triples.

Concepts Based Evaluation: In this approach the proposed measure was evaluated on a pair of concepts. It has been used in [31].

Triple Based Evaluation: In this approach the proposed measure was evaluated on a pair of triples. There is no any standard triple pair dataset. We have designed a triple pair dataset and evaluated it with the human ratings given in STASIS [10].

5.4 Evaluation Metric

Two random variables or two data sets have some type of statistical relationship. Correlation is the best measure to find out the relation among two variables. Pearson product-moment correlation coefficient is most commonly used measure to find out the dependence [49]. If X and Y are two variables and their n series of measurements are written as x_i and y_i where $i = 1, 2\ldots n$. Then the sample correlation coefficient is written as follow:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \quad (5.1)$$

Equation 5.1 Pearson's correlation coefficient

Where x and y are representing the sample means of X and Y, and s_x and s_y are the sample standard deviations of X and Y.

Perfect Direct Correlation (Increasing): If two variables have perfect direct increasing linear relationship. Then the Pearson correlation has the value of +1.

Perfect Direct Correlation (Decreasing): If two variables have perfect decreasing (inverse) linear relationship also called anticorrelation. Then the Pearson correlation has the value of -1.

Linear Dependence: For all other cases the pearson correlation has value between −1 and 1. It shows the extent of linear dependence among two variables. If correlation coefficient approaches zero its mean that two variables are uncorrelated. And if the value of coefficient is close to either −1 or 1then the two variables are more correlated.

5.5 Dataset Specifications

To date there is no publically available standard dataset to measure the semantic similarity between RDF triples. We have evaluated our measure using two data sets. The specification of each dataset is described below:

Miller and Charles: This dataset is publically available. It consists of 30 pair of noun evaluated by 38 under graduation student. The similarity of each pair is evaluated on a scale from 0 which means not similar to 4 means perfect similar. This dataset has been used in [31].

Pilot Short Text Semantic Similarity Benchmark (STASIS): This dataset is publically available. The data set contains the sentence pairs taken from Collins Cobuild Dictionary [3] with some minor modifications. The sentence pairs are rated by the mean and standard deviation provided by the ratings of 32 human participant.

Miscellaneous Dataset: To date there is not any publically available dataset to compute the similarity between RDF triples. We have constructed the RDF triples on the same set of sentence pairs given by STASIS by using WordNet as a background knowledge.

5.6 Performance Evaluation

Concept Based Evaluation: We have compared our method with several other similarity measures proposed in the literature on the same pair of concept pairs selected by Miller and Charles [31]. Experiment has shown the highest correlation (0.84) to Miller and Charles dataset. The result of correlation of different measures has been given in table 5.3.

Table 5.3 Correlation of similarity measures to CM

Method	Type	Correlation
Rada	Edge Counting	0.59
Wu	Edge Counting	0.74
Li et all	Edge Counting	0.82
Leacock	Edge Counting	0.82
Richardson	Edge Counting	0.63
Resnik	Information Content	0.79
Lin	Information Content	0.82
Lord	Information Content	0.79
Jiang	Information Content	0.83
Tversky	Feature	0.73
Rodriguez	Hybrid	0.71
X-Similarity	Hybrid	0.75
DOPIC	Hybrid	0.84

A graphical representation of correlation among semantic similarity measures and Miller and Charles dataset is given in Figure 5.4

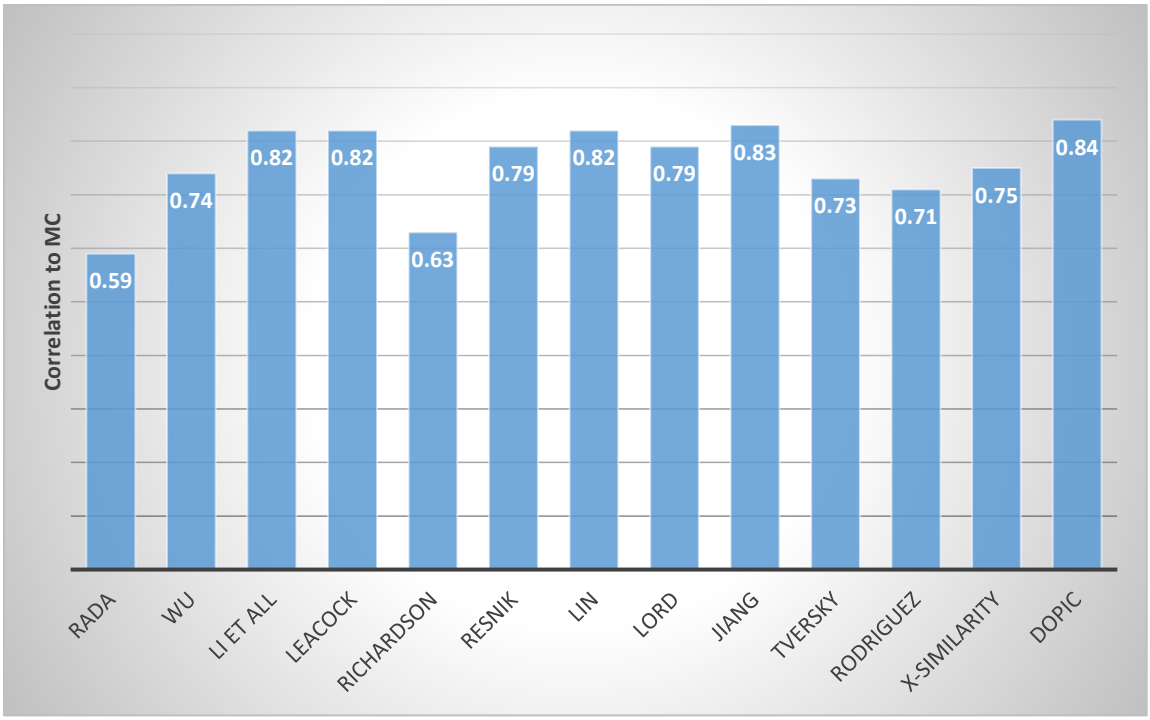


Figure 5.4 Correlation of similarity measures to Charles & Miller experiment

Triple based Evaluation: The detail of triple pairs and their semantic similarity computed by the proposed semantic similarity measure DOPIC has been given in Table 5.4. Experiments have shown 85% correlation to STASIS.

Table 5.4 DOPIC Similarities of Triple pairs

ID	Triple 1	Triple2	DOPIC	STASIS (NORMALIZED)
1	Cord is string	Smile is expression	0.00024	0.01
2	Rooster is chicken	Voyage is journey	0.0004	0.01
3	Noon is midday	String is rope	0.0009	0.01
4	Fruit grows on tree	Furnace produce heat	0.0108	0.05
5	I write autograph	I visit shore	0.0024	0.01
6	Wizard practices magic	Automobile stops road	0.00042	0.02
7	Mound is pile	Stove is equipment	0.068	0.01
8	Grin is smile	Implement is tool	0.0009	0.01
9	I visit asylum	I eat fruit	0.012	0.01
10	Asylum is hospital	Monk is person	0.024	0.04
11	I go to graveyard	I visit madhouse	0.032	0.02
12	Glass is substance	Magician is person	0.005	0.01
13	Boy is child	rooster is chicken	0.08	0.11
14	She purchase cushion	She buy jewel	0.06	0.05
15	Monk offer prayer	Slave provide service	0.16	0.053
16	Asylum is madhouse	Cemetery is graveyard	0.0072	0.04
17	She visit forest	She travel to coast	0.14	0.05
18	Grin is smile	Lad is boy	0.0004	0.01
19	Sharif visit shore	Sharif go to woodland	0.06	0.08
20	Madiha meet monk	Madiha know an oracle	0.21	0.11
21	Boy advise him	Sage advise him	0.4	0.04
22	He purchased automobile	He get cushion	0.24	0.02
23	I visit a shore	I make a mound	0.204	0.04
24	Lad impress audience	Wizard amused audience	0.0195	0.03
25	I visit forest	I go to graveyard	0.024	0.07
26	Boy eat food	Boy see rooster	0.003	0.06
27	People travel to cemetery	Folk travel to woodland	0.072	0.04
28	People visit shore	People enjoy voyage	0.009	0.02
29	He see a bird	He visit a woodland	0.1	0.01
30	They reach to a coast	They climb a hill	0.2295	0.1
31	Furnace is container	Implement is tool	0.198	0.05
32	Crane is machine	Rooster is chicken	0.22	0.02
33	We climb hill	We visit woodland	0.0056	0.15

ID	Triple 1	Triple2	DOPIC	STASIS (NORMALIZED)
34	We travel by car	We enjoy journey	0.0016	0.073
35	I visit cemetery	I make a mound	0.0042	0.058
36	I clean glass	I wash jewel	0.1185	0.108
37	Magician amuse people	Oracle impress people	0.02	0.13
38	Crane is machine	Implement is tool	0.16	0.185
39	I love brother	I like lad	0.28	0.128
40	wizard amuse him	Sage advise him	0.0336	0.153
41	Oracle astonish them	Sage impress them	0.27	0.2825
42	Bird fly in air	Crane move thing	0.018	0.035
43	I saw bird	I saw cock	0.15	0.1625
44	She eat fruit	She eat food	0.1	0.243
45	He meet brother	He meet monk	0.39	0.045
46	They visit asylum	They visit madhouse	0.66	0.215
47	Furnace provide heat	Stove provide heat	0.26	0.3475
48	Magician entertain people	Wizard amuse people	0.4131	0.355
49	I climb hill	I make a mound	0.35	0.293
50	She break cord	She break string	0.75	0.47
51	Glass	Tumbler	0.163	0.136
52	Grin is smile	Smile is expression	0.5335	0.485
53	I have a slave	I have a serf	0.49	0.483
54	I enjoy journey	I enjoy voyage	0.76	0.36
55	actor give autograph	Actor write signature	0.4368	0.405
56	I see coast	I visit shore	0.7695	0.588
57	We go to forest	We visit woodland	0.3382	0.626
58	He hold a tool	He have an implement	0.65	0.59
59	He have a cock	He have a rooster	0.97	0.863
60	she like boy	She love lad	0.574	0.58
61	He take cushion	He ask for pillow	0.7462	0.523
62	Hindu bury cemetery	Muslim bury graveyard	0.485	0.773
63	I have automobile	I buy car	0.4032	0.558
64	I sleep at midday	I sleep at noon	1	0.955
65	I have gem	I purchase jewels	0.49	0.65

A pictorial representation of correlation among DOPIC and STASIS has been shown in figure 5.5

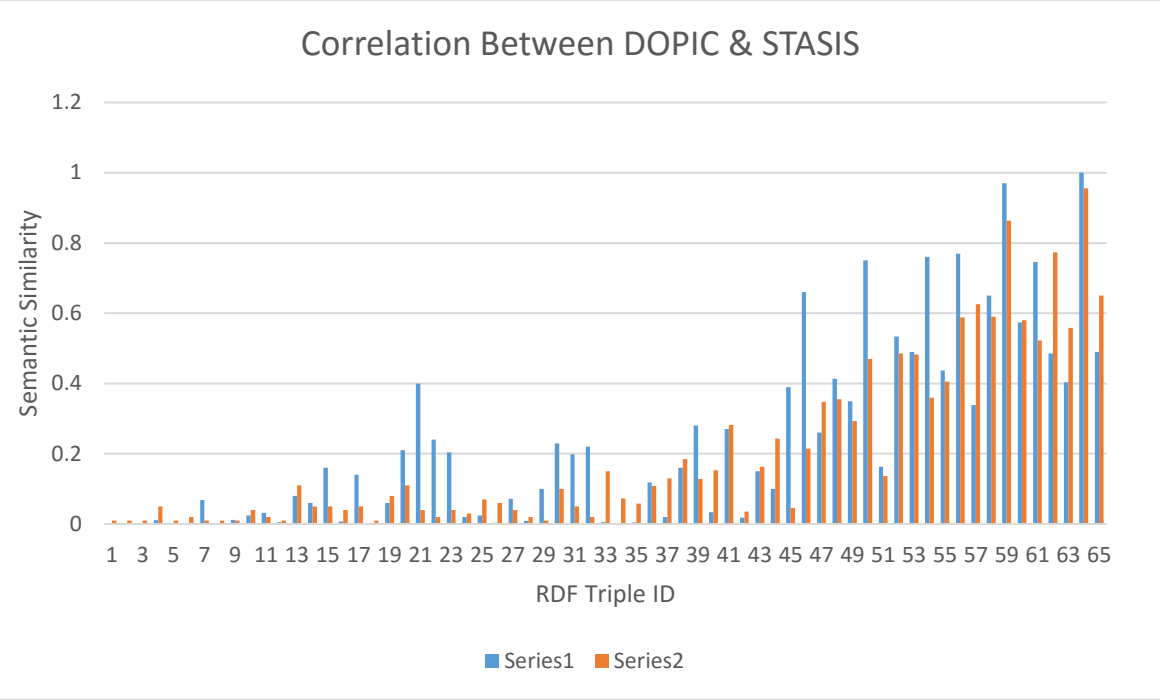


Figure 5.5 Correlation among DOPIC & STASIS

Summary

In this chapter we have discussed our system's implementation and evaluation. The hardware and software specifications were described and the output of system was illustrated with screen shots. The datasets used for the system evaluation and comparison were specified. The system evaluation against the Miller and Charles experiment and STASIS dataset was discussed, followed by a discussion on system's comparison with existing techniques based on methodology as well as Pearson product moment correlation coefficient.

Chapter 6

Conclusion and Future Direction

In this chapter we present a summary and the contributions of the research work documented in this thesis. Some of the fundamental limitations of our approach and an outlook of the future directions where this work can be extended are presented at the end of this chapter.

6.1 Conclusion

Context is very important to retrieve accurate information. Keyword based search systems do not consider the context they just focus on the semantics of individual keywords. Context can be considered by focusing on the relationships that exist among keywords. A pattern can represent the context that is the circumstances in which something happens or to be considered. We have propose a semantic similarity measure to compute the semantic similarity among RDF triples. The proposed semantic similarity measure uses both the structure of ontology and statistical information content. The combination of taxonomic structure and empirical probability estimates provides a way of using static knowledge structure to multiple contexts [44]. We have evaluated our system by repeating Charles and Miller experiment [9] and by comparing our measure with several other similarity measures. Experimental results have shown the highest correlation to Charles and Miller experiment. We have also evaluated our measure using Pilot Short Text Semantic Similarity Benchmark Data Set (STASIS) [10] and we have obtained 85% correlation with STASIS.

6.2 Contributions of the Research

Analysis by Software Agent: By considering semantics and RDF representation, critical analysis and useful information can be generated by software agents which a human being have to do manually in other case**.**

Improved Precision: By considering context, semantic heterogeneity can be eliminated which improves the precision of information retrieval.

Context based system: Semantic measure provides ability of making comparison of different units of a language on the basis of their meaning and context. Context is needed for accurate information retrieval. For example if we rank the two concept pairs (Monkey, Phone) and (Monkey, Banana) on the basis of relatedness. We will say that the concepts monkey and banana are more related. Without context we could not calculate the exact similarity in such situations.

6.3 Limitations and Future Work

WordNet has mentioned many senses of a concept. The proposed system have used competitive computation method i.e. each sense of a concept is being compared with each sense of other

concept. Future work involves to further improve the system by using the most appropriate sense of a concept instead of all sense.

Bibliography

1. Latifur Khan, Dennis McLeod, Eduard Hovy. “Retrieval effectiveness of an ontology-based model for information selection”. In: *The VLDB Journal*, (2004).13: 71–85.
2. Berners-Lee, Hendler, Lassila. “*The Semantic Web”*. The Scientific American, Feature Article, (2001).
3. Ivan Herman. “Introduction to the Semantic Web”. *Semantic Technology Conference*, (2009).
4. W3C.*Semantic Web-W3C Semantic Web Activity*.[online; accessed 6-December-2015].URL: http://www.w3.org/2001/sw/
5. Sharifullah Khan, Jibran Mustafa, Khalid Latif. **“**Intelligent Search in Digital Documents.” In: *Proceeding WI-IAT '08 Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (2008). Volume 01. Pages 558-561.
6. Miriam Fernández1, Iván Cantador2, Vanesa López1, David Vallet2, Pablo Castells2, Enrico Motta. “Semantically enhanced Information Retrieval: an ontology-based approach”. *Dissertation*. (2009).
7. Jonathan Poole and J. A. Campbell, “A Novel Algorithm for Matching Conceptual and Related Graphs”, *In the book of Conceptual Structures: Applications, Implementation and Theory, Santa Cruz CA, USA Springer-Verlag*, pp. 293-307 1995.
8. S. Khan and F. Marvon, “Identifying Relevant Sources in Query Reformulation”. In: *the proceeding of the 8th International Conference on Information Integration and Web-based Applications & Services (iiWAS2006),* Yogyakarta Indonesia (2006).
9. Jarmasz, Mario, and Stan Szpakowicz. "Roget's Thesaurus and Semantic Similarity." *arXiv preprint arXiv:1204.0245* (2012). [accessed 4-December 2015].URL: http://arxiv.org/ftp/arxiv/papers/1204/1204.0245.pdf
10. Pilot Short Text Semantic Similarity Benchmark Data Set: Full Listing and Description. [accessed 4-December 2015]

URL:https://www.researchgate.net/publication/238732681_Pilot_Short_Text_Semantic_Similarity_Benchmark_Data_Set_Full_Listing_and_Description

11. Semantic Web. [Online; accessed 4-December 2015]. URL: http://www.csie.ndhu.edu.tw/~showyang/SOC2008/05SemanticWeb.pdf
12. S. Khan and J. Mustafa. “Effective Semantic Search using Thematic Similarity”. In: *Journal of KSU Computer and Information Sciences*, ISSN: 1319-1578, Elsevier Publishing.
13. Neda Alipanah, Pallabi Parveen, Sheetal Menezes, Latifur Khan, Steven B. Seida_, Bhavani Thuraisingham: “Ontology-driven Query Expansion Methods to Facilitate Federated Queries”. *SOCA* 2010:1-8
14. Harith Alani, Christopher Brewster**: “**Ontology Ranking based on the Analysis of Concept Structures”. In: *Proceedings of Third International Conference on Knowledge Capture (K-Cap)*, Alberta, Canada. (2005). pp. 51-58.
15. WittyCookie. *Major differences among Web 1.0, 2.0 and 3.0*. [Online; accessed 4-December 2015]. URL: https://wittycookie.wordpress.com/2012/06/04/what-are-the-major-differences-among-web-1-0-2-0-and-3-0/
16. W3C. *XML and Semantic Web W3C Standards Timeline*. [Online; accessed 4-December 2015]. URL: http://www.w3.org/standards/timeline/
17. Herman, Ivan.*"W3C Semantic Web Activity". World Wide Web Consortium (W3C). November 7, 2011*.
18. Berners-Lee, Tim, James Hendler, and Ora Lassila. "The semantic web." *Scientific american* 284, no. 5 (2001): 28-37.
19. W3C Semantic Web. *Semantic Web standards*.[Online; accessed 4-December 2015].URL: http://www.w3.org/standards/
20. W3C Semantic Web. *Semantic Web* [Online; accessed 4-December 2015].URL: http://www.w3.org/RDF/
21. W3C Semantic Web. *RDF Schema*.[Online; accessed 4-December 2015].URL: http://www.w3.org/TR/rdf-schema/
22. Brickley, Dan, and Ramanathan V. Guha. "{RDF vocabulary description language 1.0: RDF schema}." (2004). [On-line; accessed 4-December 2015].URL: http://www.w3.org/TR/rdf--schema

23. Grigoris Antoniou, Frank van Harmelen. (2004). “Describing Web Resources in RDF” in A Semantic Web Primer. [On-line; accessed 4-December 2015].URL: https://www.dcc.fc.up.pt/~zp/aulas/1011/pde/geral/bibliografia/MIT.Press.A.Semantic.Web.Primer.eBook-TLFeBOOK.pdf
24. Thomas R. Gruber. “A translation approach to portable ontologies”. In: *Knowledge Acquisition*, 5(2):199-220, 1993.
25. Natalya F. Noy, Deborah L. McGuinness. “Ontology Development 101: A Guide to Creating Your First Ontology”. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001.
26. Wikipedia. *Linked data- Wikipedia, The Free Encyclopedia*. [Online; accessed 4-December 2015].URL: https://en.wikipedia.org/wiki/Linked_data
27. W3. *"Linked Data—Design Issues"*.[Online; accessed 4-December 2015].URL: http://www.w3.org/DesignIssues/LinkedData.html
28. Jagendra Singh1, Mayank Saini and Sifatullah Siddiqi.” Graph Based Computational Model for Computing Semantic Similarity”**.** In: *Proceedings of International Conference on Emerging Research in computing, ERCICA* 2013.
29. David Sánchez1, Montserrat Batet, David Isern, Aida Valls.” Ontology-based semantic similarity: a new feature-based approach”. In: *Expert Systems with Applications*, Volume 39, Issue 9 (2012), Pages 7718–7728
30. K.Saruladha, Dr.G.Aghila, Sajina Raj.” A Survey of Semantic Similarity Methods for Ontology based Information Retrieval”. In: *Machine Learning and Computing (ICMLC)*, 2010 Second International Conference (9-11 Feb. 2010). PP. 297 – 301.
31. Gan, Mingxin ; Dou, Xue ; Wang, Daoping ; Jiang, Rui. “DOPCA: A New Method for Calculating Ontology based Semantic Similarity”. In: *Computer and Information Science (ICIS)*, 2011 IEEE/ACIS 10th International Conference (2011). PP. 110 – 115.
32. Jay J. Jiang, David W. Conrath.” Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy”. In: *Proceedings of International Conference Research on Computational Linguistics (ROCLING X), (*1997), Taiwan. PP. 15.
33. Giannis Varelas, Epimenidis Voutsakis, Paraskevi Raftopoulou, Euripides G.M. Petrakis, Evangelos E. Milios.” Semantic Similarity Methods in WordNet and their Application to

Information Retrieval on the Web". In: *Proceeding WIDM '05 Proceedings of the 7th annual ACM international workshop on Web information and data management* (2005). PP. 10-16.

34. Sheng, Qiuyan , Ying, Guisheng." Measuring Semantic Similarity in Ontology and Its Application in Information Retrieval". In: *Image and Signal Processing*, 2008. CISP '08. Congress on (Volume: 2) (2008). PP.525 – 529.
35. Wei He, Xiaoping Yang ; Dupei Huang. "WNOntoSim: A Hybrid Approach for Measuring Semantic Similarity between Ontologies Based on WordNet". In: *Web Information Systems and Applications Conference (WISA)*, 2011 Eighth (2011). PP. 73 – 77.
36. Shaohua Zhang, Xuequn Shang ; Miao Wang ; Jingni Diao. "A New Measure Based on Gene Ontology for semantic Similarity of Genes". In: *Information Engineering (ICIE)*, 2010 WASE International Conference on (Volume:1)(2010). PP. 85 – 88.
37. Manuel Montes-y-Gómez, Aurelio López-López, Alexander Gelbukh. (2001-June-28) *Database and Expert Systems Applications Volume 1873 of the series Lecture Notes in Computer Science* pp 312-321.[Online; accessed 4-December 2015].URL: http://link.springer.com/chapter/10.1007/3-540-44469-6_29
38. Giuseppe Pirro. "REWOrD: Semantic Relatedness in the Web of Data". In: *Twenty-Sixth AAAI Conference on Artificial Intelligence* (2012-07-12).
39. Yauri, Aliyu Rufai; Kadir, Rabiah A; Azman, Azreen; Murad, Masrah A A. "Semantic Concept Search based on Triple Matching Approach". In: *Journal of Convergence Information Technology10.2* (2015). PP. 98-105.
40. Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre. "SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity". In: *Proceeding SemEval '12 Proceedings of the First Joint Conference on Lexical and Computational Semantics* - Volume 1: *Proceedings of the main conference and the shared task*, and Volume 2: *Proceedings of the Sixth International Workshop on Semantic Evaluation (*2012). PP. 385-393.
41. Peter Exner, Pierre Nugues." Entity Extraction: From Unstructured Text to DBpedia RDF Triples". In: *Proceedings of the Web of Linked Entities Workshop in conjuction with the 11th International Semantic Web Conference* (ISWC 2012). PP. 58-69.
42. Yih, Wen-tau, and Christopher Meek. Learning vector representations for similarity measures. *Technical Report MSR-TR-2010-139*, Microsoft Research, 2010.

43. Jose Paulo Leal CRACS & INESC-Porto LA.” Using proximity to compute semantic relatedness in RDF graphs”. In *the book of Computer Science and Information Systems* 2013 Volume 10, Issue 4, Pages: 1727-1746
44. Philip Resnik.”Using Information Content to Evaluate Semantic Similarity in a Taxonomy”. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence* (1995). PP. 6.
45. Lin, Dekang. "An information-theoretic definition of similarity." In *ICML (1998)*, vol. 98, pp. 296-304.
46. Miller, George A. "WordNet: a lexical database for English." *Communications of the ACM* 38, no. 11 (1995): 39-41.
47. Jarmasz, Mario, and Stan Szpakowicz. "Roget's Thesaurus and Semantic Similarity." *arXiv preprint arXiv:1204.0245* (2012).
48. Semantic measure library and toolkit. *Benchmark.* [Online; accessed 4-December 2015].URL: http://www.semantic-measures-library.org/sml/index.php?q=doc_benchmarks
49. Wikipedia. *Correlation and Dependence- Wikipedia, The Free Encyclopedia*. [Online; accessed 4-December 2015].URL: https://en.wikipedia.org/wiki/Correlation_and_dependence
50. Sharifullah Khan, and Jibran Mustafa. "Effective semantic search using thematic similarity." In: *Journal of King Saud University-Computer and Information Sciences* 26, no. 2 (2014): 161-169.