DESIGN OF TECHNIQUES TO IDENTIFY SIMILARITY BETWEEN SEMANTIC WEB DOCUMENTS

THESIS

submitted in fulfillment of the requirement of the degree of

DOCTOR OF PHILOSOPHY

to

YMCA UNIVERSITY OF SCIENCE & TECHNOLOGY

by

POONAM CHAHAL

Registration No: YMCAUST/Ph57/2K11

Under the

Supervision of

Dr. Manjeet Singh
Professor,
YMCAUST, Faridabad

Co-supervision of

Dr. Suresh Kumar
Professor,
MRIU, Faridabad

Department of Computer Engineering
Faculty of Engineering and Technology
YMCA University of Science &Technology
Sector-6, Mathura Road, Faridabad, Haryana, INDIA

MARCH, 2017

DECLARATION

I hereby declare that this thesis entitled "**DESIGN OF TECHNIQUES TO IDENTIFY SIMILARITY BETWEEN SEMANTIC WEB DOCUMENTS"** by **POONAM CHAHAL**, being submitted in fulfillment of the requirements for the Degree of Doctor of Philosophy in the Department of Computer Engineering under Faculty of Engineering and Technology of YMCA University of Science and Technology, Faridabad, during the academic year March 2012 to March 2017, is a bonafide record of my original work carried out under the guidance and supervision of **DR. MANJEET SINGH, PROFESSOR, DEPARTMENT OF COMPUTER ENGINEERING, YMCA UNIVERSITY OF SCIENCE AND TECHNOLOGY** and co-supervision of **DR. SURESH KUMAR, PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, MANAV RACHNA INTERNATIONAL UNIVERSITY** and has not been presented elsewhere.

I further declare that the thesis does not contain any part of any work which has been submitted for the award of any degree either in this University or in any other University.

(POONAM CHAHAL)

Registration No:

YMCAUST/Ph57/2K11

CERTIFICATE

This is to certify that this thesis entitled **“DESIGN OF TECHNIQUES TO IDENTIFY SIMILARITY BETWEEN SEMANTIC WEB DOCUMENTS”** by **POONAM CHAHAL,** submitted in fulfillment of the requirement for the Degree of Doctor of Philosophy in Department of Computer Engineering, under Faculty of Engineering and Technology of YMCA University of Science and Technology Faridabad, during the academic year March 2012 to March 2017, is a bonafide record of work carried out under our guidance and supervision.

We further declare that to the best of our knowledge, the thesis does not contain any part of any work which has been submitted for the award of any degree either in this University or in any other University.

DR. MANJEET SINGH
Professor

Department of Computer Engineering,

Faculty of Engineering and Technology,

YMCA University of Science and Technology,
Faridabad

DR. SURESH KUMAR
Professor

Department of Computer Science and Engineering,

Faculty of Engineering and Technology,

Manav Rachna International University, Faridabad

Dated:

ACKNOWLEDGEMENT

After an intensive period of continuous learning, writing this note of thanks is the finishing touch on my thesis. It has been a period of intense learning for me, not only in the scientific field, but also on a personal level. Writing this thesis has had a big impact on me.

First and foremost I offer my sincerest gratitude to my thesis supervisor **Dr. Manjeet Singh**, for sharing his pearls of wisdom, patience, and motivation during the course of this research. I would like to thank him for the many valuable discussions that helped me understand my research area better, bestowing immense knowledge and diligence on me, for carefully reading and commenting on countless revisions of my work. Without his warm encouragement, I would not have been able to accomplish this thesis. One simply could not wish for a methodical and scrupulous supervisor.

My co-supervisor, **Dr. Suresh kumar,** has been always there to listen and give advice. I am deeply grateful to him for the long discussions that helped me sort out the technical details of my work. I am also thankful to him for his aspiring guidance, invaluably constructive criticism and friendly advice during the research work.

I would like to thank my husband for his wise counsel and sympathetic ear, for supporting me spiritually throughout writing this thesis and my life in general. You are always there for me.

My Parents, who offered their encouragement through phone calls, almost every day – despite my own limited devotion to correspondence.

I also like to thank my wonderful children: **Ranveer & Yashveer**, for always making me smile and for understanding on those weekend mornings when I was studying instead of playing with them.

Last but not the least; I would also like to thank my in-laws, friends for their humble support.

Again, thank you everyone for continuous support !

(POONAM CHAHAL)

ABSTRACT

The World Wide Web is the source of information in which information is present in the form of interlinked web pages. A search engine is an information retrieval tool that searches the information stored on WWW according to the specified query given to it by an individual. The basic architecture of a search engine consists of a crawler which fetches the documents as much as possible, an indexer which interprets these documents and creates an index based on the information available in each document, and a ranker which provides the ranked result-set as per the query given by the user? Due to huge amount of information available on the web, the users of web find difficult to retrieve the relevant information as per their requirement. The reason for inefficient retrieval of information from web is its representation in natural language. Thus, the result-set produced by the search engine are not up to the user expectation as the result-set contains many undesirable web pages which are not of user interest. This is due to the fact that the information retrieval tools such as Google search engine, Yahoo search engine etc. has several limitations. First, the commercial or traditional search engines do not lemmatize or part-of-speech tag. For instance, to identify the frequencies for the object-verb pairs there is a need of framing diverse queries and further it is desirable that the single query search should be used to do the similar thing. The issues will increase if the need is dealing with a language which is having more inflection and variability. Second limitation is that the search syntax is inadequate. Third, the restriction is on the count of queries and number of hits per query. Fourth, the number of hits is for the searched web pages rather than the instances. Thus, it means that the search engine does not use the highly structured searching techniques which are the basic requirement of Natural Language Processing applications. However, the algorithms used for ranking of web documents by the search engine to give user a ranked result-set as per user query depends on various factors like page authority, novelty of the web page content, organization of the web page, refresh rate of web page. The major issue is related to the understanding of the web page content by syntactic analysis along with semantic analysis to extract the meaningful information as desired by the web users.

The subsequent generation of semantic search engines deals with the issues of traditional search engine in the form of layered architecture of semantic web. Tim

Berners-Lee visualization of semantic web is basically a collection of resources along with the resource description. This resource description helps in interpreting the data/description of the web page content which is further efficiently processed by the machines. In recent times, several semantic web search engines developed like Ontolook, Swoogle, etc helps in searching and retrieving the meaningful information from the web content presented on semantic web. Similarity Computation is an essential concept which can be applied in many fields like Natural Language Processing, Artificial Intelligence, Machine Learning, Cognitive Science etc. The similarity computation between any given texts gives the base of analyzing, learning, specialization, generalization, and recognition. Basically, similarity measure between two texts can be classified in two kinds, one is the attributional similarity and the other is the relational similarity. When two entities are compared on the basis of attributes then their association is called attributional similarity. However, when the two entities are compared on the basis of semantic relationships between each pair of words then their association is termed as relational similarity. For example (car, automobile) word pair shows high degree of association between their attributes. On the other hand, (lion, cat) word pair have an implicit relationship that lion is a large cat. The semantic relationship "is a large" which is defined in an implicit way between the word pair which makes the words in the given pair relationally similar. Concept of similarity gives the measure of association between two documents, but if these documents are compared on the basis of keywords only, then the lexical similarity may not provide true results. The reason for this is that the author may use the synonyms of the words in a text and the keyword based approaches do not consider synonyms when the two texts are compared. To resolve such issues there is a need to detect the similarity on the basis of semantic analysis. Semantic analysis considers both attributional similarity and the relational similarity to measure the degree of association between any given texts.

In our research work, we refer text as an input data written in the natural language which is given to the machine for processing. The text size is defined as the combination of words and the relationships used by an author of the text to connect these words. This input text can be annotated with the semantic information by using schemes like Resource Description Framework (RDF) to make the text in a format which is easily processed by a machine. Generally, the text is considered of three types: Free Text, Structured Text, and Semi Structured Text. In Free Text, the elements are organized in a preset sequence of the words and relationships between the words which are written in Natural Language which follow the rules of grammar. For example, in research papers, e-books, news headlines etc. the method of doing any modification is significant as per the grammatical rules. This is due to the fact that the free text is processed into division like heading, sentence, paragraph, and document. Next, the Structured Text refers to the information which is accumulated in a file or database in an organized predefined format. The data/information management of the file or database can be easily accessed, updated, and dealt with the help of several computations techniques. The Semi-Structured Text is the form of text which lies between the structured text and unstructured text. In general, the semi-structured do not follow the particular format, but various kind of structuring is present in the text for example, web page written using HTML or XML.

Despite of the various favorable existing approaches and the challenges faced for similarity computation between the text/documents, there also exist various unique challenges which are required to overcome. First, is the recognition and extraction of probable set of concepts representing each word of a document written in natural language. Next, is to consider the relationships between these concepts so that the intention of author of the document can be captured. The intention of author means the idea, view, concept, description or information related to an event or thing which the author desires to communicate through the document. While analyzing various existing semantic similarity techniques, it is observed that the Natural Language Processing (NLP) and Ontology have significant roles to understand the text. Consequently, in our research work we have developed approaches for similarity detection and ranking scheme using NLP techniques and structured knowledge like Ontology further considering the issues like synonyms as discussed above.

In this thesis, we have given a few techniques for computing the semantic similarity between semantic web documents. In one proposed technique, we have considered the concepts available in a web document and the relationships between these concepts to compute the semantic similarity between web documents. In this relation based proposed technique, we have constructed the Vector Space Model for lexical matching and the Relation Space Model for relationship matching. The final similarity score between the documents is given by considering both lexical and relation matching. In second proposed technique, we are using the Genetic Algorithm to obtain the optimal ranked result-set of web documents with respect to user query. In this technique we are analyzing a document at two different levels i.e. Conceptual level and Descriptive level to extract the explicit and implicit information. The Conceptual level is related to the concepts i.e. explicit information available in the document and the Descriptive level is related to the implicit semantic information. The optimum values of weights to each level are assigned by using the Genetic Algorithm to compute the similarity between two documents.

The other three more proposed techniques, is related to identification of words/concepts and further forming the chains of such related words/concepts to construct document ontology. This document ontology will present the semantic information that is available in the content of the document. The extension of document ontology is further done by using current words being used in contemporary web called recent trends available related to a domain to uncover all the implicit related concepts. Finally, in all these three techniques the semantic similarity between the web documents is computed by comparing the constructed document ontology's. Further, two more techniques are proposed to provide ranking of web documents by computing the similarity between query and web document. In one ranking technique, the weighted relationships between the concepts of web documents and the user query are considered to provide user the relevant result-set as per their necessity. On the other hand, in second ranking technique, the relational probability of user query with respect to web document is computed which gives the relevance of web page with respect to the user query. Similarly, the relational probability of web page with respect to the base ontology is computed which gives the relevance of the web page with respect to the domain. Finally, the joint relational probability computation is done to rank the set of documents with respect to the user query.

All the proposed similarity detection techniques can be applied in various applications of information retrieval like Crawling, Indexing, and Ranking Etc. In general, the intend of the research is to focus on extensive analysis of the web documents for the purpose of finding similarities by exploring various NLP techniques and Ontology. This can be achieved by following the basic objectives for all the proposed techniques which consists of identifying the concepts and relationships among the concepts from a specific domain, representation of these identified concepts and relationships using a suitable formalism like ontology, development of a processing module which will identify certain form of semantic structure from a given document by using above said ontological structures and using NLP techniques, and computation of the similarity between the documents by using the semantic structures. The proposed approaches given in this thesis have been empirically evaluated on set of documents related to a domain showing their superiority as compared to existing similarity techniques.

This thesis is organized as follows: Chapter 1 gives the introduction of semantic similarity computation between web documents. Chapter 2 discusses the work done related to the field of detection of similarity between web documents into categories i.e. techniques based on lexical matching approaches and methods of semantic similarity detection using the knowledge structure ontology. The work related to the field of semantic similarity detection is analyzed deeply and the issues are considered while designing the new semantic similarity computation techniques. Chapter 3 gives the proposed semantic similarity techniques which makes the use of concept and the relationships between the concepts. Another technique which is used to compute similarity between query and web document and thus obtaining optimal ranked result-set using Genetic Algorithm is also given. In Chapter 4, we have given the techniques for semantic similarity computation which make the use of ontology and additionally construct the web document ontology by connecting the chains of concepts and the connected concepts. The novel techniques of semantic similarity detection based on the probability methods are also given in Chapter 5. The performance of all the developed techniques are analyzed deeply on the set of the web documents collected for testing the results corresponding to each techniques. In Chapter 6, we conclude our thesis with description of potential future work in the area of development of semantic similarity computation techniques between web documents.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure **Page No**

LIST OF ABBREVATIONS

WORLD WIDE WEB	WWW
INFORMATION RETRIEVAL	IR
RESOURCE DESCRIPTION FRAMEWORK	RDF
HYPER TEXT MARKUP LANGUAGE	HTML
EXTENSIBLE MARKUP LANGUAGE	XML
NATURAL LANGUAGE PROCESSING	NLP
ONTOLOGY WEB LANGUAGE	OWL
BAG OF WORDS	BOW
BAG OF CONCEPTS	BOC
LINK GRAMMAR PARSER	LGP
UNIFORM RESOURCE IDENTIFIER	URI
SEMANTIC SIMILARITY RETREIVAL MODEL	SSRM
ONTOLOGY STRUCTURE BASED SIMILARITY	OSS
LATENT SEMANTIC ANALYSIS	LSA
LATENT RELATIONAL ANALYSIS	LRA
SINGULAR VECTOR DECOMPOSITION	SVD
VECTOR SPACE MODEL	VSM
RELATION SPACE MODEL	RSM
LEXICAL MATCHING	LM
GENETIC ALGORITHM	GA
EUCLIDEAN DISTANCE METHOD	EUC
FUZZY LOGIC	FL
DOCUMENT ONTOLOGY	DO
EXTENDED DOCUMENT ONTOLOGY	EDO
NOUN PHRASE	NP
VERB PHRASE	VP
ADJECTIVE PHRASE	ADJP
GRAPHICAL USER INTERFACE	GUI

CHAPTER I

INTRODUCTION

1.1. WORLD WIDE WEB (WWW)

In the last many years, the Web has become a precious resource of information for almost each probable domain of knowledge [2]. The web is considered as applicable repository for tasks like information retrieval, knowledge acquisition etc. The tools like Google, Yahoo, etc. are being used by the users efficiently for information retrieval from WWW. But the information on the web is heterogeneous in nature and mainly written in natural language which is difficult for a machine to understand and hence it is difficult to give relevant response. An information retrieval process is mainly consisting of crawling, indexing and ranking of information. Therefore, it requires the comparison or understanding of texts/documents in order to detect the degree of similarity between the texts for either crawling, indexing, or ranking of documents. However, the similarity between numerical data can be compared by means of classical mathematical operators but the natural language similarity or relevant information retrieval is mainly done by semantic analysis techniques.

A search engine is a tool that helps in retrieving information stored on WWW. The search engine works by using a spider, robot or crawler to fetch the documents as much as possible. Another program, known as indexer, then examines these documents and generates an index based on the information contained in each document. The architecture of typical search engine is given in Figure 1.1. Each search engine makes use of a proprietary algorithm to generate its indices such that only meaningful results are returned for each end user query [1]. But, the outcomes of retrieval of information produced by various search engines are not up to the requirements of user. The reason is that there is a wide gap between the techniques required for automatic processing of information and the techniques presently used. This is due to inherent structure of plain web where web documents are written mainly according to human readability. To overcome the limitation, the next generation of search engines is required to deal with this problem in a layered architecture of web.

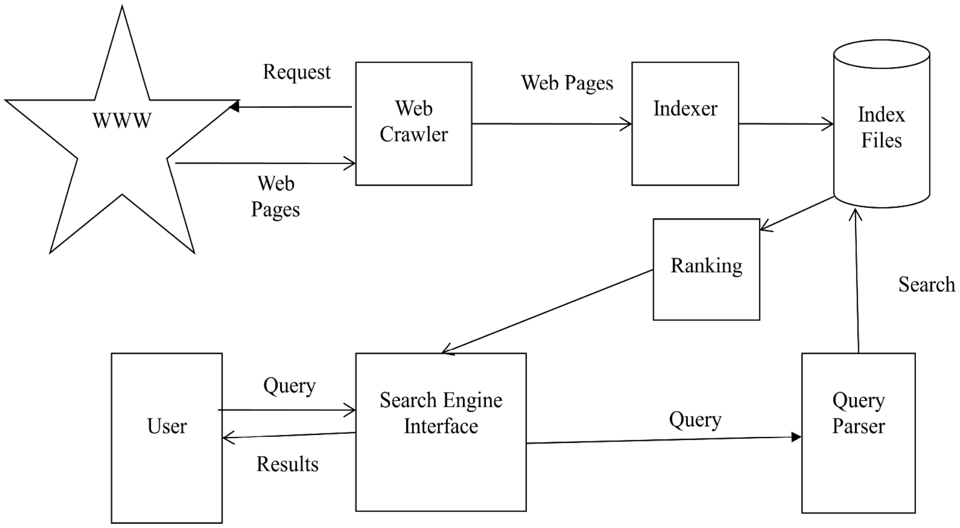


Figure 1.1: Web Search Engine Architecture [116]

The semantic web, given by Tim Berners-Lee [2] is a collection of resources and their description. In a semantic web, resource may be collection of web pages, service, product, application etc. The semantic web, thereby support machines to understand data/description in order to sustain/arrange the resources for information which is to be processed by a computer program or by any service/application later. In general, a Semantic Web presents a universal framework that permits mutual sharing of data and reprocess across relevance, project, and community boundaries. Computers, on the other hand, can only achieve inadequate understanding unless more explicit data is presented. The growth of Semantic Web typically involves dealing with descriptions of the data which is represented by using ontology's. According to Nicola Guarino et. al.[43]. Ontology is a specification of a conceptualization. In Semantic Web, ontology can be considered as a glossary used to describe a world model of a real domain. Specifically, ontology acts as a knowledge base which contains the representation or description of the classes/concepts and relationship names along with large number of entities that presents the instance population of the ontology.

1.2 TYPES OF RESOURCE DESCRIPTIONS

In this section, the basic terms and the idea on which the research work is based is given for the sake of reader convenience.

1.2.1 Text

In our research work, the text refers to the input data which is given to the machine in natural language form for processing. This text can also be annotated with the semantic associated with the content by using the RDF so that it can be made in a format which is easily processed by a machine. The text is basically considered of three types: Free Text, Structured Text, and Semi Structured Text [113].

1. Free Text

The free text means that the elements of a free text can be organized in a fixed sequence. This fixed sequence of the words and relationships is written in natural language which follows the rules of grammar. In free text like research papers, e-books, news headlines etc. the process of making any changes is relevant as per the grammatical rules, as free text is processed into parts like heading, sentence, paragraph, and document.

2. Structured Text

The information which is stored in a file or database is known as structured text as it is organized in a particular predefined format. The data/information stored in the file or database can be easily managed, accessed, and modify by performing various computations. There exist basic two types of databases i.e. traditional database and relational database. The traditional or conventional database are designed and developed to handle the organized form of data as they follow a predefined format. However, the relational database is the tabular representation of the data for accessing the stored data in several forms.

3. Semi-Structured Text

The semi-structured text, as the name suggests are the form of text which lies between the structured text and unstructured text. The semi-structured text generally do not follow the particular format, but some kind of structuring is there in the text like web page written using HTML or XML.

1.2.2 Document Set

A document or a web page text is basically considered as the content present in the web page which is in machine readable format. The content present in a web document may contain images, figures, tables etc. A document set is also known as local corpus which refers to the collection of documents that are interrelated with each other logically generally related to a domain. World Wide Web (WWW) is a collection of web documents which are one type of semi structured texts.

1.3 SEMANTIC SIMILARITY

In the field of semantic analysis, the computation of semantic similarity between two given texts plays a vital role for applications in the information retrieval task. In general, from the point of view of semantics, a text is basically the combination of words which are considered as labels representing set of concepts and relationships among these concepts. These set of concepts are widely used by many researchers in the era of relevant information retrieval as they help in depicting the semantic information present in a given text. It can be said that, Semantic Similarity, is, in particular, a discipline that intends to calculate the relatedness between words or concepts by determining, evaluating and exploiting their semantic information. There are mainly two types of similarity between words which are attributional similarity and relational similarity [3]. The attributional similarity is calculated by comparing the attributes of the words. And, the relational similarity is computed by comparing the semantic relations that are present between word pairs available in documents. However, the relational similarity between words has extensive relevance but it is a difficult task to execute because of many reasons. First, word pairs may contain more than one relation. Second, relations between the words can be represented by numerous ways. Third, relations between word pairs are dynamic in nature as they may vary with time. The objective of relational similarity is to capture the semantic information from a text.

1.4 SEMANTIC SIMILARITY MODELS

The work presented in this thesis deals with the measure of similarity detection in the field of information retrieval in domain of computer science. Similarity measure actually indicates or provides information regarding the degree of association/agreement between any two entities in the field of IR as suggested earlier also. In the sub sections, presented below, some of the major similarity computation is described briefly [109].

1.4.1 Similarity Computation Based on Distance

According to the widely accepted theoretical supposition, the similarity between two entities can be analyzed as the inverse association with the distance in several appropriate feature space which is considered to be metric space in many of the cases. Similarity score computation can be done by using the basic formula as sim=1-dis, where sim is the similarity score obtained for distance dis. The most common formulas for similarity computation [109] are given in the followed subsections.

1.4.1.1 Minkowski Distance

This Minkowski Distance measure defined as the distance Dij used for multidimensional data between any two parts i and j by using equation 1.1.

$$Dij = (\sum_{l=1}^{d} |\mathrm{x_{il}} - \mathrm{x_{jl}}\ \ |.^{1/\mathrm{n}})^{\mathrm{n}} \tag{1.1}$$

1.4.1.2 Manhattan Distance

The Manhattan is basically the Minkowski Distance defined at norm value of 1 computed by using equation 1.2. It gives the determination of absolute distinction between any two points.

$$\mathrm{Dij} = \sum_{l=1}^{d} |\ \mathrm{x_{il}} - \mathrm{x_{jl}}\ \ | \tag{1.2}$$

1.4.1.3 Euclidean Distance

The most commonly used similarity distance measure is Euclidean Distance which is defined as Minkowski distance at norm value of 2 and it is computed by using equation 1.3.

$$Dij = (\sum_{l=1}^{d} |\mathrm{x_{il}} - \mathrm{x_{jl}}\ \ |.^{\frac{1}{2}})^{2} \tag{1.3}$$

1.4.1.4 Cosine Similarity

The Cosine Similarity between any two vectors is computed by using the formula for Euclidean dot product as given below:

$$\mathrm{a.b} = \|\mathrm{a}\|\ \|\mathrm{b}\| \cos\theta$$

Depending upon the above Euclidean dot product formula, the cosine similarity represented by cosθ between two vectors having attributes A and B is computed using equation 1.4:

$$\text{Cosine Similarity} = \text{Cos}(\theta) = \frac{A.B}{||A||||B||} \qquad 1.4$$

1.4.1.5 Jaccard Similarity

The Jaccard index which is also known as Jaccard similarity measure is used for finding the similarity or dissimilarity between two sets. The Jaccard coefficient computation between any two finite set of texts is computed by given formula given in equation 1.5:

$$\text{J(A, B)} = \frac{|A \cap B|}{|AUB|} \qquad 1.5$$

To measure the dissimilarity using the Jaccard distance computation is obtained by complementing the Jaccard coefficient i.e. we need to subtract 1 from the Jaccard coefficient. The formula is given in equation 1.6.

$$\text{DJ(A, B)} = 1 - \frac{|A \cap B|}{|AUB|} = \frac{|AUB| - |A \cap B|}{|AUB|} \qquad 1.6$$

1.4.1.6 Dice Similarity

Similarly, the Dice Similarity formula was also used for similarity detection using original formula as given below in equation 1.7 which is applicable to the given data available in the two sets A and B for information retrieval.

$$\text{QS(A, B)} = \frac{2*|A \cap B|}{|A|+|B|} \qquad 1.7$$

Similar to the Jaccard similarity computation, the Dice computation can also be given in terms of operations on binary vectors like A and B which helps in calculating the common similarity metric over vectors as follows using equation 1.8.

$$\text{DS(A, B)} = \frac{2*|A.B|}{|A|2+|B|2} \qquad 1.8$$

1.4.1.7 Hamming Distance

The Hamming Distance is the most common measure of similarity for the binary attributes, thus it depends on the number of bits available in the binary attributes.

Therefore, it is described as the number of dissimilar bits in the two attributes between which similarity computation has to be done. For example, there are two strings as 10011001 and 10000101, the hamming distance between them is of 3 bits as the three bits needs to be altered to make them same. This distance method has a disadvantage that it can be applied only for the exact length comparison.

1.4.1.8 Levenshtein Distance

This distance method is the edited form of Hamming distance computation. Hamming distance gives the measure of the dissimilar bits between two strings, whereas the Levenshtein distance provides the means of edit operations like insertion, substitution, deletion etc. to make one string same as the other string.

Additionally, there are other distance similarity measures like Soundex Method, Matching Coefficient, Q-gram Distance, Overlap Coefficient etc.

1.4.2 Similarity Measures based on Features

The feature based similarity measures provide the information and computation related to the geometrical distance models. The most common feature based similarity measures is contrast model.

In the contrast model, the similarity computation is done by comparing the features of the two entities. If the two entities have more similar features than they are said to be closer and associated with each other. The formula for the same is given below in equation 1.9.

$$S(A,B) = \alpha g(A \cap B) - \beta g(A - B) - \gamma g(B - A) \qquad 1.9$$

Where, α, β, γ are the constants which are used to determine the respective weights of associated values.

g (A ∩ B) represents the common features in A and B,
g (A-B) represents distinctive features of A and,
g (B-A) of entity B.

1.4.3 Similarity Measures based on Probability

In the application like image processing, face recognition, multimedia database etc. where it is difficult to detect similarity by using exact features there is need of probability based similarity measure. The probability density functions in the probability similarity measure are used for certain features to determine the likelihoods between them. The probability measures have good performance in the applications like image processing, face recognition etc. but there is increase in the computational cost in terms of the complexity [101]. The following subsections will give the probability based similarity methods.

1.4.3.1 Maximum Likelihood Estimation (MLE)

The MLE method is based on R. A. Fisher's approach which organizes the parameters of probability model for experimental data, so that they can be made more similar.

1.4.3.2 Maximum a Posteriori (MAP) Estimation

MAP estimation of similarity computation is closely related to MLE estimation, based on the Bayesian approach where the distribution which is prior available is also used for similarity measure estimation. This method is very complex and also the priori sample of information is sometimes not available to the best information as required.
The fundamental probabilistic density models for data depiction significantly affect the correctness of similarity or likelihood calculations. Also the probabilistic similarity measures for image retrievals are used which depicts the relationships and use of Gaussian (Normal) model, Histogram model etc [104].

1.4.4 Additional Measures

Following are some more models based on recent computational methods.

1.4.4.1 Fuzzy Set Theory based Similarity Measure

A number of measures of similarity have been given based on fuzzy set theory. These fuzzy set similarity measures are basically based on union and intersection operations of fuzzy sets, maximum difference between fuzzy sets, and on the differences and summation of set membership values etc. [105]. Conventionally, the fuzzy similarity measure between two fuzzy numbers is given as

A= (a_1, a_2, a_3, a_4) B= (b_1, b_2, b3, b4)

Then the fuzzy similarity measure is computed using equation 1.10.

$$\mathrm{S(A,B)} = 1 - \frac{\sum_{i=1}^{4} a_i - b_i}{4} \qquad 1.10$$

Where S (A, B) ε [0, 1].

1.4.4.2 Graph Theory based Similarity Measure

Graph is a data structure used widely as graph matching is an effective technique to detect the similarity relationships between various parts of objects. The graph theory based similarity methods are used in several applications like content retrieval, computer vision [106] and structure analysis of document [107] etc. Practical implementation of strict graph matching is not common, thus the graph edit operations along with the cost function are used as given in [106] and also described below in equation 1.11.

$$D\ (g, g'{}_{)} = |g| + |g'| = 2|g''| \qquad 1.11$$

Although the information retrieval (IR) system as intact is accountable for storage, representation, organization, and access of data/information, the eventual goal is developing and designing similarity techniques for the efficient and relevant information retrieval process.

1.5 CHALLENGES

As large amount of information is available over World Wide Web (WWW), the Natural Language Processing (NLP) of the present information is a challenging task. The knowledge present in web is written by human beings and is constantly changing with the increasing amount of information. So, web is considered as valuable source of information for retrieval of relevant information by the user of a web. The retrieval of information can be done efficiently and effectively by applying various similarity measuring algorithms. The large amount of information available on web has been used successfully for retrieving the relevant information as per user's expectations by applying various similarity algorithms. Some researchers had already computed lexical matching of the text/documents present on web by using Jaccard similarity, Cosine Similarity, Dice Similarity, etc. [12]. Although the lexical matching provides the similarity score between the documents, but the result-set produced by the approach are not accurate as the lexical matching is purely keyword based approach which is not considering the synonyms, concepts and relationship between words/concepts. Next, the researchers had already made an attempt to compute the semantic similarity by considering synonyms of the words/concepts [44] [45] etc. The relationships between words have also been considered by various researchers providing the efficient algorithms for semantic similarity score computations [46] [47] [48] Etc.

Even though there are number of favorable approaches for semantic similarity computation between web documents, any processing approach/algorithm must overcome numerous distinctive challenges. First, the vast amount of information of web makes difficult for processing the entire content in each web document by using the given techniques of web similarity measure. Although, the storage systems like Google File System [97], and distributed computational models such as the Map Reduce [98] have already been developed which helps in data storage and processing. But, still from the computational cost point of view it is not easy to run or process a developed algorithm for similarity measure by considering the complete text on the web.

Secondly, the web page content which is written in natural language, the NLP systems also have to deal with the challenge of the superiority and the intensity of noise that exists in text of web. As large number of novel keywords which is called neologisms exists in the text of web and they are not registered in the manual created database like Word Net. The noise which creates interference in basic steps of document processing include part of-speech (POS) tagging, chunking of noun phrase (NP), syntactic or dependency parsing, or named-entity recognition (NER). Third challenge for NLP system is the reliability of information as there is high redundancy in text of web. It means that some web pages have same content of information which creates duplicate web pages and some have same content of information but using different set of keywords to represent the information. Some web pages also have content which gives contradictory information related to same topic, or also some web pages give false information.

The use of traditional search engines for processing of natural language has several limitations [99]. Firstly, the commercial or traditional search engines do not lemmatize or part-of-speech tag. For example, to detect the frequencies for the pairs of object-verb there is a requirement of framing different types of queries and it is desirable that the same thing should be done by the single query search. The issues will increase if the requirement is dealing with a language having more inflection, variability. Second limitation is that the syntax of search is limited. Third limitation is the constraints on count of queries and hit number per query. Fourth limitation is hits is not for instances but for searched pages. Thus, the ranking techniques used by the traditional search engine to rank the web pages according to the given query are not highly structured searching techniques which are the requirement of NLP applications. The major issue associated with traditional search engine is that it does not perform deep parsing of text by semantic analysis. However, the algorithms used for ranking of web documents by the search engine are dependent on several factors like page authority, novelty of the web page content, structure of the web page, page rate for refresh or update. Moreover, the exactly used ranking algorithm by a traditional/conventional search engine is also not available publicly which leads to the complex development of an NLP system which further cannot guarantee to find the relevant information web pages on the top of the searched and ranked result-set.

Many approaches have been given by various researchers to design and develop a search engine with capability and applicability for Natural Language Processing techniques [99]. Conversely, these several techniques still lack in achieving the high efficiency according to the scalability of the web information. The algorithms given in this thesis make use of linguistics approaches such as stemming, stop word removal and lexical patterns along with the semantic analysis of the information present in the web page. The web is predictable to develop continuously and thus the NLP algorithms for the dynamic and constantly increasing source of knowledge must be designed, developed and evaluated in a manner that any change in the web information would have a constructive effect on the performance of the algorithm.

Consequently, the development of techniques that does not deliberate the performance of a ranking algorithm even when the size of the web increases are desirable. In this observation, the utilization of web search engines as the interface to the enormous information existing on the web is attractive.

Despite the various favorable approaches presented above for finding the semantic similarity between the text/documents, there are several unique challenges which needs to be overcome. First, is the recognition and extraction of probable set of concepts representing each word of a document written in natural language. Next, is to consider the relationships between these concepts so that the intention of author of the document may be captured. The meaning of intention of author of the document is related to the idea, view, concept, description or information about an event or thing which the author wants to communicate through the document. For this, the ontology construction is to be done efficiently so that relevant relationships can be analyzed and extracted for a document. Finally, there is a need to consider the two ontology matching techniques so that semantic similarity score is computed to its true value which can meet user expectations.

1.6 MOTIVATION

In the section, as discussed above the main focus of the research work is to find similarity between documents by incorporating the semantic information using ontology's which are structured presentation of concepts used in a natural language text or sentence.

For ranking of web documents the semantic similarity is being computed between a query and stored documents by considering a user vision or expectations in mind i.e. by processing a query and extending it using ontology [5][4]. Automatic constructions of base ontology for similarity computation can also be done [6] [7].

Parsing of document to find the words and phrases from a document can be extended using WordNet [5] and then creation of a tree of each of the two documents between which similarity is to be calculated are merged using ontological information [8].

Many researchers have used the method of extracting keywords from a document and just considering and storing the noun, verb and adjective from the extracted keywords. Then the words retained are stored database and compared using ontology [1]. Although many approaches exist for similarity computation between texts but there is still a requirement of having more exhaustive techniques that are able to extract maximum semantic information from the content of web document. Based on the idea of similarity computation between text we will give the problem statement of our thesis in next section.

1.7 PROBLEM STATEMENT

In general, the major issue in information retrieval is the problem of representation and extraction of the semantic information in and from the content of a web document. The same issue has been considered for different proposed approaches for computation of the semantic similarity between documents. In survey, we have found that the Natural Language Processing (NLP) and Ontology plays important roles to understand text or to find similarity between documents. Therefore, in our research work we will develop approaches for similarity detection and ranking scheme based on NLP and structured knowledge like Ontology.

1.8 OBJECTIVES

The aim of this PhD research is to discover different techniques for finding the semantic similarity between semantic web documents which will definitely helps in various applications of information retrieval. In particular, the aim of the research is to focus on deep analysis of the web documents for the purpose of finding similarities by exploring various Natural Language Processing techniques.

Therefore, the major objectives of this thesis are:

- To identify the concepts and relationships among the concepts from a specific domain by analyzing a set of documents from the domain.
- To represent or encode these identified concepts and relationships using a suitable formalism like ontology.
- To develop a processing module which will identify certain form of semantic structure (the concepts and their relationships) from a given document by using above said ontological structures and using NLP techniques.
- Finally, computation of the similarity between the documents by using the semantic structures for ranking of the documents to provide the users results according to their necessity.

1.9 ORGANISATION OF THESIS

The organization of thesis is as follows. Chapter 1 gives introduction related to semantic web and semantic similarity. Chapter 2 describes the related work carried out by other researchers in the domain of semantic similarity and semantic based ranking techniques. We discuss the various approaches of finding the semantic similarity using natural language processing techniques, ontology etc. Chapter 3 presents the proposed techniques for document similarity by using the concepts relationship and Genetic Algorithm. In chapter 4, the techniques of similarity detection between documents by constructing chains of concepts relationships and extending the chains of concept relationship by using the current trends are given. Chapter 5 discusses the proposed novel techniques for ranking of web pages corresponding to user query by considering the semantic information available in the user query, web page and base ontology. In Chapter 6, we conclude the research work discussed in all the chapters. Further the scope of the future work in this field is also given in this chapter.

CHAPTER II

RELATED WORK

2.1 INTRODUCTION

In this chapter, a literature survey is given in order to understand the requirements for processing of a document for information retrieval as well as to identify the problems with the existing work in the domain of information retrieval (IR). The field of IR is vast and crucial, thus there is the need to first understand the levels/phases at which information retrieval is done. The major concern in the research of IR is the detailed analysis and processing of a document which is having information/content stored and availability of the relevant information to the user of WWW by a search engine according to his/her necessity.

2.2 MODELS FOR INFORMATION RETRIEVAL

The main aim of information retrieval is to provide user the information which is relevant to them. Various major information retrieval models have been developed for exact matching and best matching of user query with stored documents or between any two documents [49]. The Boolean model and Statistical model are considered for exact matching by considering the vector space and the probabilistic retrieval model. The Linguistic and Knowledge-based models are considered for best matching as they conceptually analyze a document. The lexical level of matching considers the syntactic structure of a document, boolean retrieval extract words and relate with the thesaurus. While the statistical model considers phrases occuring in a document and also the clusters of phrases for retrieving information. The Linguistic and knowledge based models considers concepts and semantic relations between these concepts [19].

There are two measures which are primarily employed to compute the efficiency and relevancy of a retrieval method i.e. precision rate and recall rate. The *precision rate* is measure of the proportion of the retrieved documents that are actually relevant to a user according to the given query. Whereas, the *recall rate*, is measure of the proportion of all relevant documents that are actually retrieved from stored documents according to a given query.

The tool like search engine has been widely used for retrieving the required information from web by sending a query specifying the need regarding extraction of information related to a topic. But, queries given by a user to a search engine are generally not efficient to retrieve information in two respects: First, they may retrieve some irrelevant documents. Second, they may not retrieve all the relevant documents.

In fact, the procedure for retrieving considerable information with the assistance of a search engine is very vital. For relevant information retrieval one of the major requirements is assistance of semantic similarity. The semantic similarity working out between the documents has many applications [9] like:

- Detection of similar web pages on WWW during the process of Crawling, Indexing, and Ranking done by a search engine.
- Discovery of related web documents which represents analogous or same topic to know divergent versions of the documents.
- Identifying plagiarism, which is taking text written by other person and presenting it in one's own expression. This can have variety of structure like factual copying a section of text, copying the text structure, translating text, copying the idea, copying the text without quoting the source.
- Multi-document summarization.

For all the applications there is a need to develop and design the techniques which helps machine to process the web information as per the requirement of the field of IR by a user. Researchers have already exploited various techniques like keyword matching, NLP, Ontology based approaches etc. available for machine processing of information present on WWW. For processing of semantic web documents which are written using Resource Description Framework (RDF), Ontology Web Language (OWL) etc. the encrusted architecture has also been developed to handle semantic web.

2.3 EARLIER VIEWS ABOUT A DOCUMENT

In general, a corpus denotes a collection of digital text documents available on web. A document written by an author is defined as a chunk of text. In information retrieval process, a document may refer to a paragraph(s), sentence(s), phrase(s) or a chain of characters. In general, documents are termed as contexts or chunks. For application of semantic information retrieval, it is advantageous to analyze and accumulate documents as “semantically logical chunks of text”, where all the chunks convey a single idea or topic. To extract the meaningful information from WWW it has been found necessary to figure out what a person wants to convey from the usage of words. However, finding the statistical semantics similarity for efficient information retrieval has provided significant step sandstone towards more precise, computation-oriented instantiations, like the distance-based analysis of the bag-of-words representation of a document.

2.3.1 Bag-of Words (BOW) Representation of a Document

In mathematics, a bag, also named as multi set, which is a set with duplicates permissible. In general, a document is represented by the bag of words having its constituent tokens. Information Retrieval, the bag-of-words hypothesis for a document stipulates that the set of words may be used for the relevance of retrieving the information contained in a document. In other terms, it is said that the frequencies of individual words in BOW are adequately analytical of similarity association between any two documents, where one document may be a query given to a search engine by any user of web. On the other hand, it is noted that the bag-of-words hypothesis is completely immature from the linguistic point of view as it disregard order of words and any syntactic structure, which unavoidably acquire a severe loss of information. In observation documents are represented using a vector space model constructed with the help of bag of words obtained. So, similarity is computed by using different vector similarity measures like Cosine Similarity, Jaccard Similarity, and Dice Similarity etc also explained in Chapter 1. Although, there are formulas available for the computation of similarity which are classified into the categories like Set-Theoretic Models, Algebraic Models, and Probabilistic Models. The Set-Theoretic models are applied by using the Standard Boolean Models, Extended Boolean Models and Fuzzy Retrieval. This type of models considers the documents as bag of words or phrases. The Algebraic models are applicable by making use of vector space model, generalized vector space model, enhanced vector space model, latent semantic indexing/analysis, all of which consider the documents as tuples, vectors, or matrices. Similarly, Probabilistic models compute the similarity by finding the relevance of a document with respect to a query specified to a search engine by the user of WWW. These Probabilistic models are based on the theorems using probability for example Bayes' theorem. Usual models of probabilistic models are Binary Independence Models, Probabilistic Relevance Model, Uncertainty Based Models, and Latent Allocation Models, which consider or analyze a whole process of relevant retrieval of documents based on the inference achieved by using the probability. The category of Feature Based Models for retrieval of information analyzes the complete document as the vectors assessed on the values/score of the feature functions. These methods basically help in making ranking methods efficient to provide the user a relevant result-set depending on the feature functions of the document or query.

On the optimistic side, the conversion of the surface text to a Vector Space Model (VSM) is computationally simple and proficient. Changing the representation of text on web to the world of vectors and matrices also permits us to make use of prevailing techniques and algorithms available from the area of linear algebra. Possibly the most persuasive dispute in the above representation approach is the vast amount of text and flourishing applications based on this approach [10].

2.3.2 The Vector Space Model

Generally, document vectors structure the columns, while the elements of vector known as features structure the matrix rows. In more compound schemes, the weights of the integer event frequencies obtained from bag-of-words are assigned again depending upon the importance of the terms in context of semantic information associated with the word present in the document. In bag-of-words approach, there is an assumption that each vector dimension match to the frequency of a token. These structures of dimensions are called features of the data. Every document is construed as a dimension in a multidimensional feature space. Bag-of-word representation utilizes features with quantitative field. These features employ in other areas of machine learning etc.

An additional peculiarity of the bag-of-words method is the very elevated dimensionality of the feature space for every token. For every domain the bag-of-words approach take sparsity into consideration for efficiency of algorithms. Additional advanced methods used in practice make use of more composite vocabulary models, similar to similarity metric and additional vector transformations to retrieve more semantics from a document.

2.4 BASIC PHASES IN A DOCUMENT PROCESSING

Information retrieval on web is crucial task done in number of phases [25]. For text-based document, there are numerals of associated phases that must occur before any semantic dealing out takes place. The phases are given as follows:

- Tokenization: It is splitting of the text of a document into individual words.
- Token Normalization: It depends on the task which is to be performed on a document like it can be removing information about letter casing, morphology analysis, syntactic analysis etc.
- Spelling Correction: It is related to dealing with ambiguous spelling present in a document like won't vs. would not, limited vs. LTD. Etc. Depending on the application, the needed action may be performed either to use the form already available in the text, or normalize the words to a single canonical structure.
- Multi-Word Expressions: This is dealing with the more complex lexical components like dates, emoticons, special symbols etc. They are also crucial to handle in the intellect that errors at this basic level are very expensive to correct afterwards.

The pre-processing of a document is the primary requirement for text mining. There are various work done on pre-processing of text for information retrieval like classification of document by pre-processing based on Vector Space Model and Bayes' Rule [39], Efficient Pre-Processing Algorithm for IR [40] etc. In next sections, the related work regarding finding the similarity of documents/text in application to information retrieval using lexical approach, Natural language Processing techniques (NLP), Semantic analysis, Ontology Based Analysis is given.

2.5 LEXICAL MATCHING AND NLP TECHNIQUES

In keyword matching approach only keywords present in a document are taken into consideration. In this approach mostly researchers first parse the whole document using any parser like LGP Parser, Stanford Parser etc. to extract the set of keywords from the document. Then the vector space model of these set of keywords are constructed and matched to find the similarity between the documents using Cosine similarity, Jaccard similarity, Dice similarity etc [12]. The similarity computed using any of the similarity formula will be 100% only if the set of words extracted from the documents is same.

[11] Has given the concept of asymmetric similarity between any two documents. The authors discussed that the documents taken to compute similarity can be of equal size or of different size i.e. one document may be completely literally present in the other document. In this case if document A is contained in literal sense in document B then lexical similarity of A to B is 100% but B to A is not 100%.

Guenther Goerz and Martin Scholz [66] has discussed that using NLP techniques for processing of a document the selected informative words can be obtained and then analysis of set of documents is done by disambiguation of those words that have numerous meaning.

James W. cooper et. al. [13] detected similar documents by taking help of salient terms. The paper describes a system which rapidly determines similar documents among set of documents retrieved from information retrieval. The authors maintained a database having list of most important terms from each document which are ranked by using a rapid phrase recognizer system. Then, the document similarity is computed using database query. If the number of terms which is not present in both document is less than the predefined threshold as compared to the number of terms of the documents then these documents tends to be very similar. The authors also compared their approach with shingles approach which is a system described by Broder [114]. In their system each document region is named as "shingles" which are considered as a series of tokens and then summarized to a representation based on numerical analysis. These numerical representations are then converted to "fingerprints" by using a method given by Rabin [115]. In fact, the comparison of number of identical tokens can be evaluated and because of this similarity measure between documents could also be computed which shows the efficient retrieval of information.

Jan K. et. al. [9] presented a computer support system for determining similar documents using chunk based approach. In this approach, the authors split the document into chunks of text which is consecutive words selected from document itself. The two documents A and B similarity are computed as %age of chunks of A which are in B which is given below:

$$\begin{aligned}&Document\ similarity\ in\ percentage\\&\qquad = No.of\ chunks\ in\ A\,\&\,B * 100\backslash Total\ no.of\ chunks\ in\ B.\end{aligned}$$

Ziv bar-Yossef et. al. [118] has given the external global measuring functions like index freshness, corpus size, density of duplicate pages, density of spam etc. that are required over the set of documents which are indexed by a search engine. The authors also claim that these functions are also necessary for relevant retrieval of web pages according to a user query, as it requires accessing to the search engine query logs which are not publicly available. So, the authors developed a query log mining algorithms which computes index metric as per impression rank which is measure of visibility of a web page in the search engine.

Weifeng et. al. [119] has proposed a statistical based parsing query interface. The authors also discussed the classification of query interfaces based on rule based and learning based methods. In rule based, a predefined set of rules are used to parse the query interface whereas, in learning based methods a model is trained as per query interface and further that trained model is applied for query interface parsing. The authors statistical parsing is hybrid of both i.e. rule based and learning based methods.

A. Pisharody et. al. [1] proposed a method using relationships between keywords. The author used Link Grammar Parser (LGP) to parse a document which is having content containing noun, adjective, verb, determiner, preposition etc. From all the contents of a document the noun, adjective, and verb are accumulated in a database. The database constructed is then normalized to remove duplicate values and after removal process each remained word in database is communicated to WordNet to determine its relation sets. Now, the database is having words and its relatedness to other words. Now this is applied in ranking technique of a search engine, whenever a user gives query to a search engine, it is also parsed to retrieve its noun, adjective and verb. The retrieved word of a query is then sent to the database of a document for retrieval of all of its relations. If word is not available in database then the reverse Lookup algorithm is used for searching the relation part rather than query word. Thus, the authors tried to remove the disadvantage of keyword approach by building an intelligent database for documents having words and relations.

The work discussed above basically deals with the lexical analysis of a document which helps in providing information of a document and also helps in finding similarity between the documents but there are many features of similarity as discussed above. The aim of determination of similarity between the documents fulfills when we are able to analyze a document semantically. For semantic analysis, it is necessary to consider relationships between or among the words present in a document. Therefore, first concepts represented by words available in a document are extracted and then relationships between these concepts in document are found. On the other hand, it can be said that a document may be analyzed as set of concepts which is said to be Bag of Concepts (BOC) in contrast to Bag of Words (BOW).

2.6 SEMANTIC ANALYSIS PREREQUISITE

In order to discuss the various techniques based on semantic similarity and ontology, this section gives the introduction to basic terminology and technologies used in these techniques. For semantic analysis of a document by considering document as BOC, the most common structure called ontology has already been used by many researchers. Ontology in the field of computer science has been defined formally as "the specification of a conceptualization" by Tom Gruber [14]. Basically, ontology is described as set of collected entities along with the relationships that may exist between these entities. The representation of ontology can be done by using a graph having nodes representing the entities and edges representing the relationships between the entities.

The concept of ontology was initiated by the Greek philosopher named Aristotle. Wikipedia [15] defines ontology as "the philosophical study of the nature of being, existence or reality in general, as well as of the basic categories of being and their relations". In an ideal world, each one entity identified to man is symbolized by a URI (Uniform Resource Identifier) for exclusive recognition. Hence, all acknowledged relationships with other entities are stored for each entity. This would help in construction of all-encompassing ontology which is the eventual desire of any computer scientist. Hence, it can be said that ontologies are constructed in a way that they consist entities mainly from a particular domain. Thus, there is a basic requirement of domain expert for construction of this domain-specific ontology's. Instances of domain-specific ontologies comprise WordNet which is a glossary in the form of ontology.

The artificial intelligence area vision ontology's as prescribed logical theories whereby not only the consideration of significant terms and relationships is done, but also the context in which these term and relationships are applied. Well known Linguistic database like WordNet express numerous relationships like synonym, antonym, is-a, contains etc. between concepts but do not clearly describe the meaning of a concept formally. Therefore, there is major requirement of an ontology which defines a set of representative terms mainly called as concepts and the interrelationships between the concepts describe an intention world and also lexical database like WordNet [18] [20]. So, formally ontology can be constructed in two ways, domain dependent and generic. Like CYC [17] and Sensus [17] are instances of generic ontology's [16] [22] which helps in making a general framework for all the types encountered by human reality.

For general computation purposes, domain dependent ontologies are constructed which are generally much smaller as they provide concepts in a fine grain. The determined knowledge in domain dependent ontology's assists to disambiguate concepts available in ontology. In common, the approaches for building ontology can be done by using Dictionary, Text Clustering, Association Rule, Knowledge Base etc. [68].

Ontology construction involves six basic steps by identifying ontology scope, capture, encoding, integration, evaluation and documentation [67]. It is important consideration during the construction of ontology's that the constructed ontology should be:

- **Open and dynamic**: Ontology's should have the ability for growth and modification.
- **Scalable and inter-operable**: The constructed ontology should be easily scaled to a broader domain and also to adapt itself to novel requirements.

- **Easily maintained**: The structure of ontologies should be simple, clear and modular so that they can be inspected/analyzed easily. They should also be easy for humans to inspect.

There are numerous techniques like Resource Description Framework (RDF) available to serialize ontology. Other accepted language for defining ontology's is the Web Ontology Language or OWL which is used to define complex relationships and constraints on them that makes it much more communicative as compared to RDF [20] [69].

2.6.1 Benefits of Ontology

A high-level categorization on benefits of ontology has already been known. The classification distinguishes between three classes of importance as follows: -

- Communication among humans and systems
- Computational implication
- Reuse and association of Knowledge

It is to be noted that ontologies are used for Communication principle to: -

- Ascertain interoperability at the level of data and process among computer programs and humans.
- Disambiguate or exclusively identification of the meaning of a concept in a given domain or interest
- To facilitate knowledge, transfer by excluding unwanted interpretations through the usage of formal semantic.

Ontology's facilitate computational implication, which is further useful to

- Automatically derive implicit facts to enhance traditional browsing and retrieval technology.
- Helps in gaining to model domain knowledge independent of the implementation of the system and also facilitate the automatic creation of the code.
- Helps in indicating errors by finding logical inconsistencies.

Ontology's, are also means to organize and classify knowledge in reusable artifacts. There are other benefits of using ontology like

- Interoperability: This supports collaboration between different systems e.g., Generic medical ontology is shared in diagnostic and therapy-control medical systems
- Formal Community View: This formalizes a shared viewpoint over a definite universe of communication like conformity on how to model time.
- Model-based knowledge acquisition: This helps in modeling ontology to acquire knowledge related to a domain like medical ontology to obtain knowledge about medical guidelines in an ordered way.
- Knowledge-level validation and authentication like the medical guideline ontology can be checked by guidelines documents.

2.7 SEMANTIC SIMILARITY AND ONTOLOGY BASED APPROACHES

The approaches mentioned in section 2.5 considered keywords and its relatedness to compute the lexical matching between any two documents. Some researchers considered synonyms of words present in a document, concept of a word, relationships between concepts which can be represented by using graph theory, relational algebra etc, and [4]. In the graph construction of a document each node is represented by a concept and edges between the nodes represents the relationship that exists between the concepts. The similarity computation done by considering concepts and relationships between concepts provides the closer semantic relatedness of documents.

Researchers have also tried to take the advantage of the ontology based similarity matching. The ontology can be constructed using tools like Protégé, Sweet, and WordNet etc [21]. In ontology based approach, concepts are extracted from a document and these can be extended using ontology with the hyponym (means more precise term or a subordinate grouping word or phrase), meronym (means fraction of a whole), synonym (means word or phrase that means precisely or almost the similar as another word or phrase in the identical language), hypernym (means a word with a wide meaning comprising a class into which words with more precise meanings lies) etc. In ontology, the parameters considered are measurement of shortest path, deepness of most precise common subsumer, density of concepts of the shortest path, density of the concepts from the root to the most precise common subsumer.

Giannis V. et. al. [5] proposed another method for computation of semantic similarity using WordNet for information retrieval from the web. In their proposed approach terms (concepts) are represented in the form of ontology and then analyzing their relationship from it. The author's method is accomplished with detection of semantic similarity between documents which are not lexicographically similar. In first part of the method, detection of semantically similar words is computed by using WordNet. Next, the author applied Semantic Similarity Retrieval Model (SSRM) method for final computation of semantic similarity. The steps in SSRM are as follows:

1. Queries and documents are analyzed syntactically and reduced to term (noun) vectors. Very frequent and infrequent words are eliminated to reduce noise.
2. Each term is represented by weight and it is computed by frequency occurrence in the document collection.
   Di=tfi*idfi where di is weight of term i in doc d, tfi is frequency of i in document
   and idfi is inverse frequency of i in whole document collection.
3. Term Reweighting: The weight of qi of each query term i is adjusted based on its relationship with other semantically similar terms j within same vector.
4. Term Expansion: Query is augmented by synonym. Then with hypernym, hyponym. Each query term is represented by tree then again weight is adjusted.
5. Document similarity: Similarity between an expanded and reweighting query q and document d is calculated.

In this approach only the query terms are expanded and reweighted. The document terms dj are computed as tf*idf it means they are neither expanded nor reweighted.

Sheetal A. et. al. [12] has also given method for measuring semantic similarity between Words by using web documents. The approach presented by the authors makes use of snippets for semantic processing of information returned by the Wikipedia or any encyclopedia such as Britannica Encyclopedia. The snippets retrieved are pre-processed for removal of stop words and stemming. Next, the significant words are extracted from the obtained pre-processed snippets. Semantic similarity measure proposed by the authors depends on the five diverse association measures in Information retrieval, namely simple matching, Dice, Jaccard, Overlap, Cosine coefficient.

B. Hajian et. al. [8] used a multi-tree model for measuring semantic similarity based on structure knowledge retrieved from ontology and taxonomy. The method described by the author's uses multi tree resemblance algorithm to determine likeness of two multi tree constructed from taxonomic relationships between dissimilar entities in ontology. The two multi-tree built are considered to obtain a final multi-tree for the set of documents which are compared. The semantic similarity is analyzed by finding the commonality of feature describing the properties of a concept. The final similarity between any two documents compared is considered as the score of similarity of root node. The author's explained the proposed approach by an example which multi tree transaction of d1 and d2 are shown in Figure 2.1 and Figure 2.2 and combined multi tree of d1 and d2 in Figure 2.3.

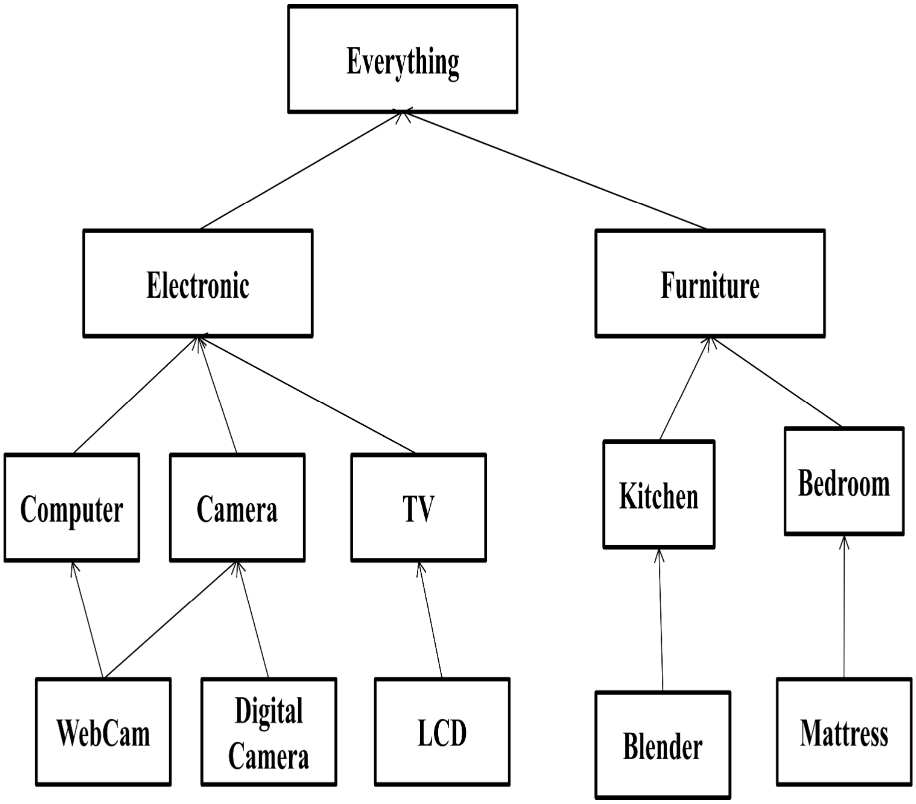


Figure 2.1: First Multi-Tree Representing Transaction D1 [8]

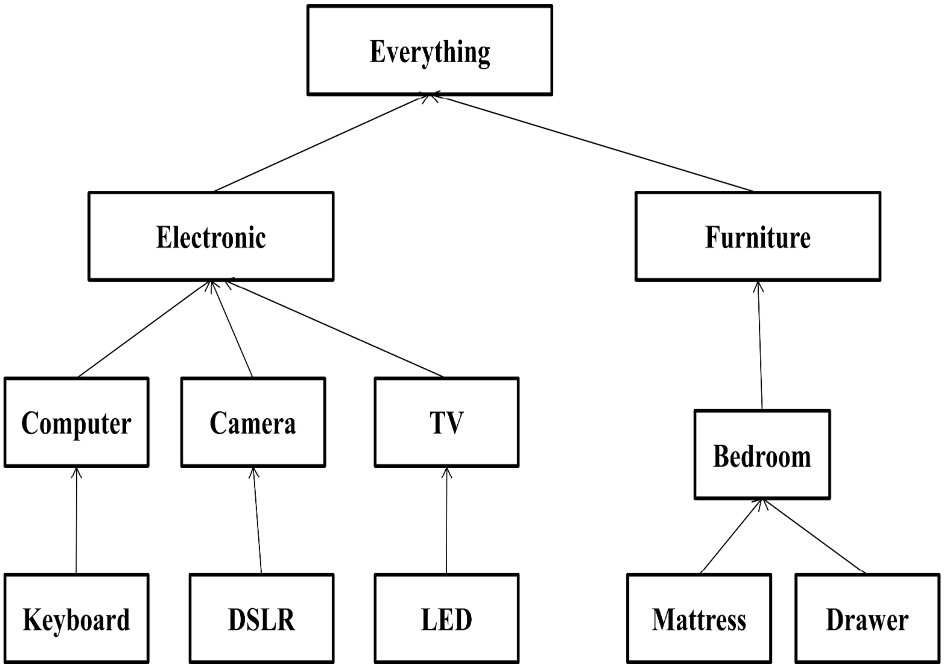


Figure 2.2: Second Multi-Tree Representing Transaction D2 [8]

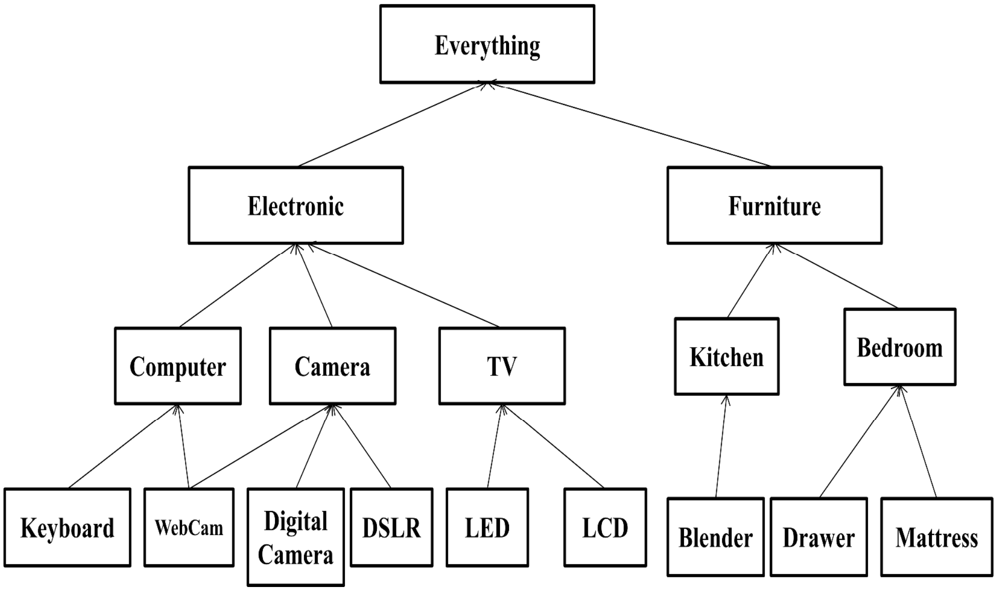


Figure 2.3: Multi-Tree Combined for Previous Multi-Tree [8]

Figure 2.3 gives the combined multi tree obtained from previous multi tree of d1 and d2 as shown in Figure 2.1 and Figure 2.2. The calculations of similarity score are as follows:

Calculating Similarity between the d1 and d2 using combined multi-tree.

W(Computer)=W(TV)=W(Camera)=(1-1/e)(0)+(1/e)=0.369

W (Bedroom) = (1/2)*(1-1/e) + (1/e) =0.684

W (Electronic) = (1/e)*(1-1/e^2) + (1/e^2) =0.457

W (Furniture) = ((0+0.684)/2)*(1-1/e^2) + (1/e^2) =0.431

W (Everything) =0.444

Y. Li et. al. [24] has given a semantic search engine named ONTOLOOK based on relationships that exists between concepts which can process related keywords with the support of architecture of semantic web. The method followed by the ONTOLOOK is first analyzing the input given by a user by determining keywords combinations. Then, the concepts pairs are accumulated to find the relationships between them which are defined in ontology. After retrieval of relationships a concept-relation graph is constructed based on information obtained. The sub graphs are obtained by cutting some unusual arcs, and the keywords and relations between them are fetched to find property-keyword candidate set used to get the relevant result set for a user.

Fabrizio L. et. al. [4] proposed an algorithm for ranking in semantic web search engine. The techniques proposed for ranking of semantic web search engine exploit the significance feedback and post methods result-set which analyze relations among keywords which are available in a web page. The proposed ranking technique is used in combination with the semantic web search engine as it is based on the information which is extracted from queries given by a user and on annotated web pages. The page significance is calculated by using probability, that a page is containing a relation whose existence was implicit by user at instant of query definition. The methodology of relation based algorithm starts from a page sub graph computation of an annotated web page and generation of all possible arrangement of edges except cycles of the sub graph. In this process, the authors constructed the graph for underlying ontology, query, page annotation and page sub graph to compute probability for a page by considering relations in all the graphs. To consider all the concepts which are of user interest even if any of them do not connect to other concept needs consideration of spanning trees.

Vladimir O. et. al. [26] has focused on ontology driven semantic comparison of documents. Generally, ontologies are considered as structured knowledge base which includes term along with properties and relations among the terms for efficient extraction of knowledge from an available text. The author's represented the ontology by using graph-model which is used for text analysis. In the approach proposed, author's compared enhanced documents by using ontology extraction algorithm and similarity is being computed between the two sub-ontology obtained.

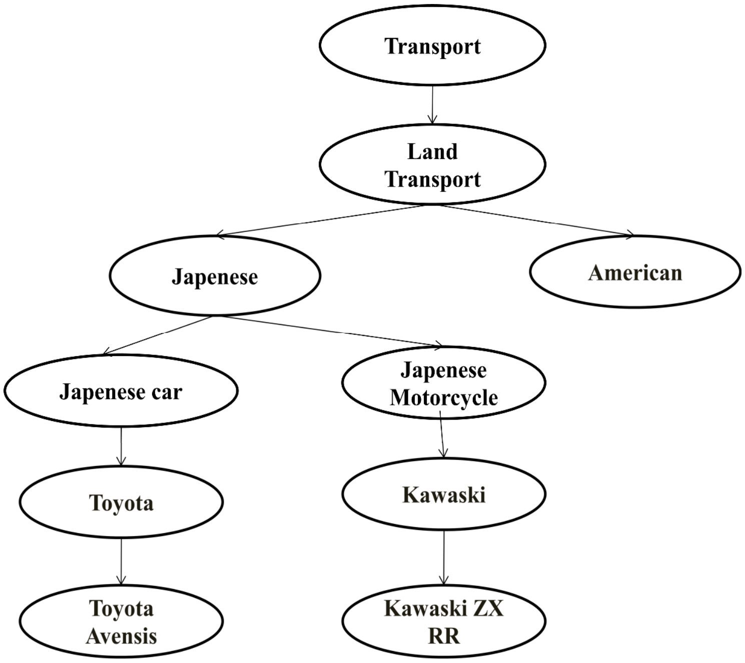


Figure 2.4: Ontology for Transportation [26]

As an example: the two text documents named t1 and t2 having contents as given below:

t1: Following the Toyota Avensis "best-ever" score in the EuroNCAP crash test, Toyota Manufacturing UK has collected a second prestigious safety accolade in recognition of its industry leading safe working environment. (From: Safety Success At The Double For Toyota).
t2: After a long winter of intensive testing on the Kawasaki ZX-RR, development is continuing at a rapid pace as Garry McCoy and Andrew Pitt prepare for the start of the 2003 MotoGP world championship on Sunday. (From: ROAD RACING - Kawasaki Hopes For Top 10 Posted By Paul Carruthers, Cycle News Online).

The main ontology is shown in Figure 2.4 and text ontology's O1 and O2 are in Figure 2.5 having comparison vector result=<1, 1, 1, 0, 0, 0>. The given texts t1 and t2 are found similar in consideration of main given ontology using the approach given by author's [26] in the logic that the both texts are related to Japanese land transportation.

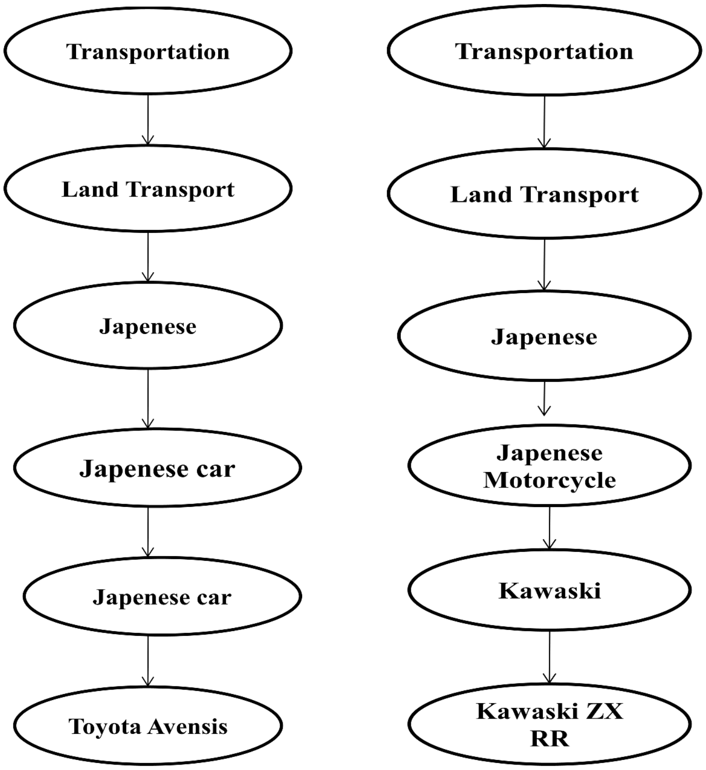


Figure 2.5: Ontology O1 and O2 for Text t1 and t2 [26]

R. Thiagarajan et. al [27] also focused on computation of similarity semantically using ontology's. In general, a web page is represented as set of words known as Bag of Words (BOW). The BOW approach considers only keywords which lead to lacking of intelligence. So, the author's considered a document as set of concepts known as Bag of Concepts (BOC) to represent a web page more semantically. The process of spreading is used to include more related term to a concept in BOC by taking help of ontology such as WordNet, Wikipedia. Spreading process used involve two schemes i.e. set spreading and semantic network as described by the authors.

Li Y. et. al. [28] proposed a method for measuring sentence similarity which application is given on conversational agents. The author's algorithm computes similarity between very short texts of sentence by considering two scores computed by semantic similarity and word order similarity. Firstly, semantic similarity is computed between two sentences resulting from information using an ordered lexical database and from corpus data. Secondly, computation of word order similarity is done from the location of word appearing in a sentence.

Yin G. et. al. [29] gives a method of computing similarity which is based on ontology by dividing the method into two i.e. concept similarity and description similarity. The concept similarity is computed by measuring the distance of concepts present in the ontology which helps in providing the shortest path length. The description similarity is further divided into two i.e. the similarity of relation and attribute. The relation similarity contributes to the similarity score emphasizing the relationship between the concepts in the ontology whereas the attribute similarity considers each attribute as a concept in the ontology.

Shahrul N. et. al. [31] proposes extraction and modeling of the semantic information content present in web documents to incorporate semantic document retrieval. The authors discussed the existing system extracting relevant information by mainly considering the extraction of important key phrases that represent the content of the documents using a domain based ontology and NLP techniques. The authors approach helps in constructing the semantic model for a document represented in XML. Finally, all the semantic model for each documents are integrated to construct global semantic model for obtaining global knowledge model of domains.

Jun F. et. al. [33] has given a novel method for document classification by using ontology reasoning and similarity computation measures. Firstly, the weighted set of terms is extracted from a document. Now, all the categories are represented using ontology's for representing the conceptualization of a category, then the lowest concept available in ontology is computed using available ontology reasoning techniques. The whole similarity score for a set of documents is computed by considering set of lowest concepts in ontology and due to small set of lowest concepts as the performance and accuracy at run-time would be better. The authors perform computation of similarity score by using Google Distance measure, to assign the documents to the categories.

Boanerges A. et. al. [32] has given the method for semantic ranking of documents by using ontological relationships. The authors aim in semantic document ranking is to consider semantic relationship that exists between the entities in the populated ontology. The key difference which author discussed in their approach is that the approach proposed does not require the interlinking of documents like in other link analysis algorithms i.e. Page Rank. The Page Rank algorithm relies on the hyperlinks for assigning the score based on number of references received by a page. The authors also introduced a measure of relevance that is based on traversal and the semantics of relationship that link entities in the ontology.

Jun F. et. al. [30] proposed automatic classification and ranking of web documents based on ontology. The authors proposed approach first extracts the weighted term set from a document to build ontology by using an effective ontology construction method which augments the existing ontology taken as per the requirements of authors. Next, the similarity score between documents and the ontology built is computed based on WordNet with the help of EMD i.e. earth mover distance method. Finally, the web documents based on similarity score are assigned to the categories and documents in the categories are also sorted using simple ranking method.

Fabio S. et. al. [34] has given a retrieval model of information for the semantic web. The authors in this paper find the information items with similar content which is present in the user query. The internal representation of information items is based on the user interest groups named semantic cases. The model proposed describes a similarity measure to order the results based on the semantic distance between semantic cases items. The model proposed is the quadruple (D, Q, F, R (d, q)), where D and Q respectively are internal representation of documents and queries, F is a framework for modeling document representation, queries, and their relationships and R(d,q) is a function to similarity measure between documents. In the model D and Q represent set of concepts. The framework of the proposed model is created using reasoning services and a semantic case-based strategy which defines how metadata are organized into the internal representation of documents. Finally, the model provides a matching process that uses the concepts to find related document and a semantic similarity function for the retrieval results ranking.

Danushka B. et. al. [35] has given an approach to measure semantic similarity using web search engine. The authors have given a novel algorithm for pattern extraction and pattern clustering to identify the various semantic relations that can exists between two given words. The optimal combination of lexical pattern clusters and page counts-based co-occurrence measures is learned using support vector machines.

Vincet S. et. al. [36] has proposed a function of semantic similarity based on hierarchical ontology's. The authors have given a novel approach that allows similarities to be asymmetric by using information contained in the structured ontology. The proposed approach is named as Ontology Structure Based Similarity (OSS) as it is based on ontology structure to compute similarity between any two concepts in three basic steps. First, the authors infer the score of the concept b from a. From the inferred score obtained the authors analyze how much has been conveyed between these two concepts. Finally, a distance function is applied which converts the transfer of score into a distance score.

Juhum K. et. al. [37] has given a method based on similarity graph computed for retrieving similarity for semantic web. The method given by authors using similarity graph mainly resolves the interoperability issue by providing mapping technique and similarity properties for computation of similarity. The main contribution of authors is to provide a core technique of computing similarity across ontologies of semantic web.

Peter D. et. al. [38] has proposed a method named LRA i.e. latent relational analysis for measuring semantic similarity which extends the VSM i.e. vector space model in three ways:

i) Automatic derivation of patterns from the corpus.
ii) Frequency data is smooth using SVD i.e. Singular value decomposition.
iii) Reformulation of word pairs using synonyms.

LRA process includes finding the alternates, filtering the extracted alternates, determining the phrases for the set filtered alternates. Next, the detection of the patterns for the phrases, mapping of the pairs to rows and mapping of the patterns to column is done. Then the sparse matrix is constructed, entropy is computed, SVD is applied, projection is done, and alternates are evaluated to compute final relational similarity.

Jun F. et. al. [33] has discussed the issue of classifier training and also not considering the semantic relations between words in traditional machine learning algorithm.

Generally, document classification is done in three stages. First, extraction of document characteristics and categories is done. Second, similarity is computed by using the extracted information between documents and the categories. Finally, classification of documents is done on the basis of similarity score measured. In the method proposed by the authors the issues are resolved by first extracting the weighted terms from a document and the categories extracted are represented by the ontology's. Next, by using Google distance measure the similarity between the documents and the ontology is computed. Finally, the assignment of web documents to the categories is done according to the similarity score.

Shahrul et. al. [31] proposed semantic document retrieval with the assistance of techniques of natural language analysis and a domain specific ontology. The authors extracted the set of candidate concepts by using heuristic rules. Next, for constructing the content of semantic the sentences having the concepts extracted are analyzed and evaluated with the document ontology. The representation of semantic document model which is extracted and constructed is done in XML. Finally, the creation of the global semantic model to give the global knowledge for some domains is done by integrating the semantic model.

Boanerges et. al. [32] has given a method for ranking of documents using semantic relationships independent of any specific structure of the documents or links between the documents by considering the one or two query from any user. Out of the two queries given by a user, first query is used to retrieve the documents that facilitate in matching query as part of annotation and the second query helps in retrieving the documents that match the keyword based searching. The ranking of documents is done by considering entity-matches from annotated query. The proposed method is basically based on traversal and the relationships semantics that link entities in an ontology.

Danushka B. et. al. [3] proposed the method of representing the various semantic relations that are available for linking the words by means of automatically extracted lexical patterns. Then the extracted lexical patterns are collected to identify different pattern that convey a precise semantic relation, and computation of the similarity between semantic relations is done with the assistance of a metric learning approach.

Pushpa et. al. [100] proposed pattern retrieval algorithm for computation of supervised semantic similarity between pair of words. The proposed algorithm makes use of the web snippet method and page count method. The authors submit query of word pair to the search engine to get the page counts. These page counts are used by them to compute the co-occurrence by using Web Dice, Web Jaccard, Web PMI, and Web Overlap methods. Then, the query is given to the search engine in the form A*****B to the search engine and retrieve the snippets. Finally, the patterns are retrieved and their frequency is computed by using the proposed pattern algorithm.

Eduardo et. al. [103] has proposed an approach based on semantic logic to compute the similarity between two given texts. The authors main contributions is derivation of logic form transformation (LFT) from semantic representation and thus further encoding knowledge at different levels. The proposed textual similarity approach is based on the derivation of semantic features from logic prover in combination with the machine learning approach. The prover gives the similarity score depending upon the features and LFTs and thus the final score of similarity is computed by combining all these scores.

Peipei et. al. [102] has given probabilistic approach for term similarity by using semantic network. The authors define the term in context of concepts performing the clustering on these concepts. The similarity is defined by the highest score obtained for the sense of one word in context with the available sense of the other word.

Ronald et. al. [121] has developed a framework for the construction of document spanner which maps an input string over the set of relationships that span over the input string. Georgina et. al. [117] has given Plagate, which is a novel tool for detection of plagiarism. This tool when integrated with the existing plagiarism tool improves the performance of detecting plagiarism by providing graphical evidences by using the well-known technique of information retrieval i.e. latent semantic analysis (LSA).

The work discussed above basically deals with the semantic analysis of a document by making the use of the domain knowledge base which is called ontology. The use of domain ontology further helps in providing the conceptual information of a document and thus finding relatedness between the documents. The comparison table to summarize the work of different researchers on the basis of concepts, relations and ontology used to extract semantic information from text is given in Table 2.1.

Table 2.1: Comparative Analysis of Various Approaches for Finding Similarity

Author	Concepts	Relations	Ontology
Cordi	√		√
Pisharody	√	√	
Thiagarajan	√	√	
Oleshchuk	√		√
Hajjan		√	√
Li		√	√
Peter D.		√	
Boanerges		√	√
Yin	√	√	
Lamberti		√	√

As discussed above and also seen from Table 2.1 it has been found that for semantic analysis it is essential to consider associations between the words/concepts available in a document. Thus, there is a need to design and develop the techniques for information processing to bridge the gap between the human understanding and the machine processing. In next section, we are giving our research problem in revised form after searching and understanding the techniques given by numerous researchers.

2.8 PROBLEM DEFINITION REVISED

The research problem in our thesis is related to the consideration and resolving of issues that are discussed above in the field of information retrieval. To extract the semantic information from a web document for relevant IR the semantic similarity computation techniques/methods given in our thesis consider the following:

- Generally, a semantic web document is written by using the schemes like Resource Description (RDF), Ontology Web Language (OWL) etc. But, in our proposed research work, we are assuming that a pre-processing to remove all language specific tags has already been done to get the plain text. Therefore, in all proposed schemes a document is defined as the collection of natural language constructs (plain text).
- The semantic analysis of a web document is done by processing each document in a way to extract the semantic information from it. For this processing, we need a lexical database and a base ontology like other researchers [1] [4] [8] [13] [26] [27] as already discussed in previous sections.
- In our research work, we will also construct a data structure which will be called as domain specific dictionary. This domain specific dictionary will be constructed by identifying the concepts related to a domain.
- Additionally, a base ontology will also be constructed by identifying the relationships between the concepts available in domain dictionary. This process of extraction of the relationships between concepts would help us to understand the semantic information related to the domain. This semantic information will further give the idea, view, concept, description or information about an event or thing implied in the content of each web document which the author wants to convey to the user/reader.
- Next, the above constructed domain dictionary and base ontology, will be used to for identification of concepts and relationships between these concepts from a web document. These identified concepts and relationship between the concepts from a web document will be represented in suitable formalism like ontology.
- Finally, the constructed ontologies for web documents will be used by various proposed approaches to compute the semantic similarity between any two web documents. The semantic score obtained from computation will further helps in ranking of the semantic web documents to provide the relevant result-set of web documents for a query given by the users of search engine according to their necessity.

2.9 SUMMARY AND DISCUSSION

In this chapter, we introduced the classical models in which documents are represented as set or vectors of words/terms. In the Boolean model, queries are represented as Boolean expressions of disjunction of conjunctive vectors. Each term in this representation has a weight associated which defines the term importance in the document or query. The Boolean system is flexible and easy to implement in search engine and information retrieval system as it allows evaluation of document and query by the use of hierarchical aggregation [41]. But still, there is a need for improvement in terms of scalability and fast analysis of terms as the drawbacks of systems is that it uses reasonably simple representations of semantics by implying search strategies on the terms or combination of terms.

One understandable extension to Boolean systems is embracement of additional knowledge in the structure of taxonomy of terms which will help in providing an evaluation method considering order of terms rather than just occurrence. This retrieval model is extending classical model using natural language processing techniques in combination with the knowledge contained in ontology constructed for domain. Using ontology, the similarity can also be computed based on the close principle which gives two related concepts that are in ontology. But still there is a challenge for improvement in the techniques available for processing of information by the machine which is readable by the user of information like representation of the ontology structure, organization of concepts and relationships between the concepts in a domain ontology, analysis of the stored concepts and relationships, retrieval of relevant related concepts etc.

The following chapters discusses the proposed work on the issue of extracting relevant concepts and relationships for a domain so that the more exhaustive and scalable approach for measuring semantic similarity score between any two web documents can be designed for efficient information retrieval for users of web.

CHAPTER III

DOCUMENT SIMILARITY BASED ON CONCEPT RELATIONSHIPS AND GENETIC ALGORITHM

3.1 INTRODUCTION

In previous chapters, various semantic similarity approaches have been discussed. These approaches have given an insight on the measure of relevance by analyzing the concepts/words and relationships between the concepts/words. Such semantic similarity techniques involve extraction of words/concepts and associations between them from a document. In this chapter, we are presenting novel techniques to exploit the extracted concepts and relationships to improve the semantic similarity computation between any texts. The proposed approaches make use of the conceptual knowledge available for a domain for extraction of concepts and relationships to compute semantic similarity.

3.2 RELATION BASED SIMILARITY COMPUTATION: A PROPOSED APPROACH

The implicit semantic relations have already been captured from semantic web by clustering extracted lexical patterns and then semantic similarity is measured by using a metric learning method [3]. Similarly, the advantages of online corpus and grammatical set of laws have already been utilized for improving the performance of similarity detection between texts [50]. The ontology as a knowledge base has also been considered by many researchers for detection of the connection between ontology terms/concepts [51]. The existing methods of similarity detection have been classified into categories considering: semantic distance based methods, information content, method of terms based properties, ontology based hierarchy, and hybrid methods [52]. The technique given in this chapter, also make use of the thesaurus like WordNet, knowledge base called ontology in form of graph having nodes as concepts/terms and edges as the relationships between the concepts/terms. It has also been assumed, for the proposed scheme, that the pre-processing to extract the plain text from the web document is already applied by using HTML parser.

The similarity computation between web documents by considering words and relationships is done by using a domain based constructed data structures named as domain specific dictionary and a base ontology graph. The collection of words from set of domain related documents and the consequent synonyms represented by each word are jointly stored in domain specific dictionary. This domain specific dictionary is constructed with the help of online available traditional dictionary i.e. WordNet. The base ontology is having the nodes representing concepts stored in constructed dictionary and the relationships that exist between each concept pair are represented as edges between them.

In first stage of semantic similarity computation, extraction of words from the documents is done by using Stanford Parser. Then, the visualization and disambiguation of these words is done using synonyms available in domain specific dictionary for each extracted word. Now, the document is represented as set of words and visualized interrelated words from constructed dictionary. Next, relationships between the identified words and interrelated words of a document are extracted by using base ontology constructed for a domain. This base ontology act as a knowledge base, for the extraction of relationships that exists between the known concepts of a document which further helps in computation of semantic similarity. One key element in the construction of base ontology is that each relationship between any two concepts stored in the ontology is assigned with weight by referring to the domain documents. The process of assigning weights to the concepts relationship is done only once during the construction of the ontology. The weights assigned to the relationships present in the ontology depend on many factors like type of relationships, class-instance relationships between the concepts. This process of assignment of weight to each relationship is done to construct the Relation Space Model (RSM) of each document by using ontology and domain specific dictionary. The constructed RSM is like the Vector Space Model (VSM) which consists of the words from the document along with the frequency of the word in the same document. In the similar manner, the RSM for a document will constitute relationships between concepts pairs with the frequency of each relationship in a document and the already assigned weights to the corresponding relationships.

The complete architecture for the approach considering concepts and relations is shown in Figure 3.1. The major components of the scheme are Ontology Processor to construct the ontology, Document Processor to analyze the document for concept retrieval, Semantic Score Computation module for final computation of similarity between any two documents. The Ontology Processor, is basically having the concept analyzer and relation analyzer for extraction of the concepts and their relationships for a document. The document processor is constructed with the help of syntactic analyzer and semantic analyzer for extraction of words by using Stanford Parser and consequently the extracted words are analyzed by using the domain specific dictionary. The architecture of the proposed technique as shown in Figure 3.1 works in two stages. First, the Document Processor extracts the keywords from the document and then analyzes and replaces them with the corresponding synonyms present in domain specific dictionary for retrieving the lexical patterns between concepts. In second stage, the Ontology Processor provides the relationship for construction of RSM. The results of calculation retrieved from the lexical matching and the RSM matching are given to comparator to combine the information and score obtained from lexical patterns and RSM. This comparator provides the complete information of both the stages of similarity to semantic score computation module for final calculation of semantic similarity between documents which is to be given to the user interface.

The proposed approach of computation of semantic based similarity by considering concept relationships can be formally explained as follows:

For set of two documents D_1 and D_2 for which similarity measure is to be computed, words extracted from these documents are represented by corresponding synonyms which we consider as concepts stored in dictionary. Each pair of concepts from both the documents is considered like for example (C_1, C_2) from document D_1 and (C_3, C_4) from document D_2 for similarity detection. The similarity computation is done in two stages as explained above. First, the common information retrieval tool i.e. search engine is used to extract the lexical patterns that may exists between each extracted concepts pair. The lexical patterns are retrieved by extracting snippet between each concept pair.

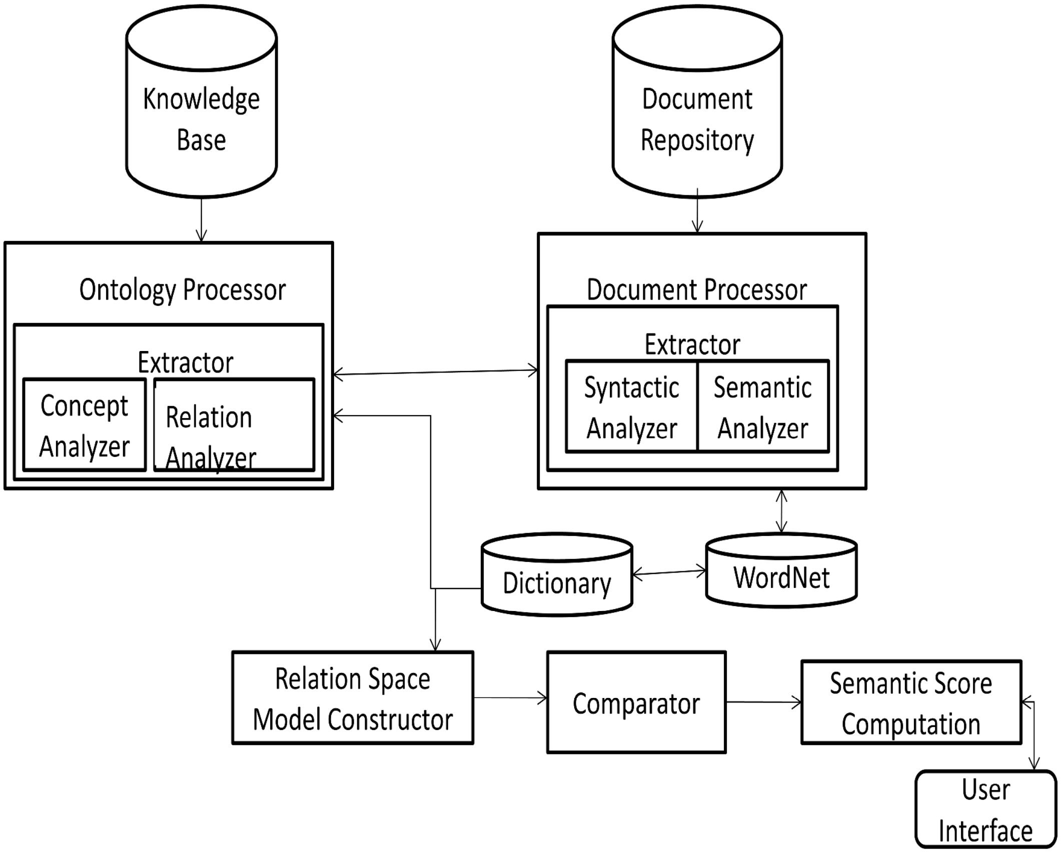


Figure 3.1: Structural Design of Proposed Semantic Similarity Model

The snippet is basically the text/phrases between concept pairs given by the search engine to provide context in which the two concepts relates with each other. The retrieval of snippet between two concepts C_1 and C_2 is done by giving seven types of queries C_1* C_2, C_2* C_1, C_1** C_2, C_2** C_1, C_1*** C_2, C_2*** C_1, and C_1 C_2 where * is a wildcard character which represents the extracted snippet. These extracted lexical patterns are used to compute the similarity using available Cosine similarity formula i.e.

$$\mathrm{Sim}_{\mathrm{cosine}}(\mathrm{e}_i, \mathrm{e}_j) = LMij = \frac{\vec{\mathrm{V}}(\mathrm{e_i}).\vec{\mathrm{V}}(\mathrm{e_j})}{|\vec{\mathrm{V}}(\mathrm{e_i})||\vec{\mathrm{V}}(\mathrm{e_j})|} \qquad 3.1$$

In second stage, these concepts pairs are analyzed to extract the relationships between them by using the base ontology graph O. The extracted relationships help in constructing Relation space model (RSM) for similarity detection as it has the information stored related to relationship type, relationship frequency in a document, relationship weight related to concepts present in a document etc. The RSM is then normalized to remove the duplicate relations and inconsistency retrieved for each concept pairs. Next, the RSM is sorted according to the frequency of each relation between concepts of documents as the frequency will give the importance of relation between concepts. The RSM constructed is used to compute the similarity using equation 3.2.

$$\mathrm{Sim}_{\mathrm{cosine}}(\mathrm{r_i}, \mathrm{r_j}) = RSMfij = \frac{\vec{\mathrm{V}}(\mathrm{r_i}).\vec{\mathrm{V}}(\mathrm{r_j})}{|\vec{\mathrm{V}}(\mathrm{r}_i)||\vec{\mathrm{V}}(\mathrm{r}_j)|} \qquad 3.2$$

Finally, the lexical matching score obtained in first stage of technique given and the RSM computation score obtained in second stage are used to detect the final semantic similarity between any document pair by using equation 3.3.

$$SSc = (RSMf_{ij} + LM_{ij})/2 \qquad 3.3$$

Where $\mathrm{RSM_{fij}}$ is the cumulative frequency of relationships weighted score for Document i and Document j.

$\mathrm{LM_{ij}}$ is the obtained similarity score between Document i and Document j from the lexical matching.

The detailed Concept Relationship Algorithm of our proposed approach based on relationships between concepts present in each document is as follows:

1. Construct a Domain Specific Dictionary D having keywords/terms along with the synonyms.
2. Construct a Domain specific weighted Ontology O.
3. For each document in domain related document set do
    - i. For each sentence in the document $\mathrm{D_i}$ extract keyword/term/concept $\mathrm{t_i}$.
    - ii. Search the term $\mathrm{t_i}$ in domain specific dictionary D to find the synonyms $\mathrm{c_i}$ which is also available in the base ontology O as a node of the graph.
    - iii. Consider $\mathrm{t_i}$ with $\mathrm{c_i}$.
    - iv. Extract the snippets $\mathrm{r_i}$ between each concept pair that exists in the document.
    - v. Compute Lexical Matching by using

$$\mathrm{sim}_{\mathrm{cosine}}\left(\mathrm{e}_{i},\mathrm{e}_{\mathrm{j}}\right)=LMij=\frac{\vec{\mathrm{V}}(\mathrm{e}_{\mathrm{i}}).\vec{\mathrm{V}}(\mathrm{e}_{\mathrm{j}})}{\left|\vec{\mathrm{V}}(\mathrm{e}_{\mathrm{i}})\right|\left|\vec{\mathrm{V}}(\mathrm{e}_{\mathrm{j}})\right|}$$

4. For any pair of two documents D_i, D_j do:
   i. Construct the document RSM having relationships that exists in document along with frequency of the relationship r_i by using O.
   ii. The RSM created is sorted according to the frequency of relationships available in the space model and normalized.
   iii. Compute Relation Matching by using

$$\mathrm{sim}_{\mathrm{cosine}}\left(\mathrm{r}_{i},\mathrm{r}_{\mathrm{j}}\right)=RSMfij=\frac{\vec{\mathrm{V}}(\mathrm{r}_{\mathrm{i}}).\vec{\mathrm{V}}(\mathrm{r}_{\mathrm{j}})}{\left|\vec{\mathrm{V}}(\mathrm{r}_{\mathrm{i}})\right|\left|\vec{\mathrm{V}}(\mathrm{r}_{\mathrm{j}})\right|}$$

5. Finally, calculate the semantic score among two documents by using

$$SSc=(RSMf_{ij}+LM_{ij})/2$$

3.2.1 Implementation and Explanation Using Example

As per the proposed scheme, the Stanford Parser is being used for the syntactic analysis of the sentences for set of documents given in Appendix I Table 1.1. The tree of each document is created by using the library Stanford-parser.jar and lexicalized parser class in our system which is implemented using Java. For example, the two documents D_1 and D_2 having the sample content related to domain artificial intelligence as follows:

D_1: Artificial intelligence is the intelligence of machine and robot and the branch of computer science that aims to create it.

D_2: Artificial intelligence textbook define that artificial intelligence is the intelligence of machine and robot, the field as study and design of intelligent agent where an intelligent agent is system that perceives its environment and takes action that maximizes its chance of success.

The document content is kept in the word file and the same is parsed by using the Stanford Parser to construct the tree of each document as discussed above. The structure of the parse trees of above sentences present in document D_1 and document D_2 are given below in Figure 3.2 and Figure 3.3.

```
[NLPParser] |--> ROOT [129.595]
[NLPParser]   |--> S [129.444]
[NLPParser]     |--> NP [22.512]
[NLPParser]       |--> JJ [9.647]
[NLPParser]         |--> artificial
[NLPParser]       |--> NN [7.993]
[NLPParser]         |--> intelligence
[NLPParser]     |--> VP [105.796]
[NLPParser]       |--> VBZ [0.147]
[NLPParser]         |--> is
[NLPParser]       |--> NP [101.027]
[NLPParser]         |--> NP [35.197]
[NLPParser]           |--> NP [10.386]
[NLPParser]             |--> DT [0.641]
[NLPParser]               |--> the
[NLPParser]             |--> NN [7.993]
[NLPParser]               |--> intelligence
[NLPParser]           |--> PP [24.153]
[NLPParser]             |--> IN [0.667]
[NLPParser]               |--> of
[NLPParser]             |--> NP [23.087]
[NLPParser]               |--> NN [7.253]
[NLPParser]                 |--> machine
[NLPParser]               |--> CC [0.165]
[NLPParser]                 |--> and
[NLPParser]               |--> NN [9.862]
[NLPParser]                 |--> robot
[NLPParser]         |--> CC [0.165]
[NLPParser]           |--> and
[NLPParser]         |--> NP [60.360]
[NLPParser]           |--> NP [10.557]
[NLPParser]             |--> DT [0.641]
[NLPParser]               |--> the
```

```
[NLPParser]                 |--> NN [8.164]
[NLPParser]                    |--> branch
[NLPParser]              |--> PP [20.325]
[NLPParser]                 |--> IN [0.667]
[NLPParser]                    |--> of
[NLPParser]                 |--> NP [19.259]
[NLPParser]                    |--> NN [6.016]
[NLPParser]                       |--> computer
[NLPParser]                    |--> NN [9.022]
[NLPParser]                       |--> science
[NLPParser]              |--> SBAR [25.543]
[NLPParser]                 |--> WHNP [1.447]
[NLPParser]                    |--> WDT [0.880]
[NLPParser]                       |--> that
[NLPParser]                 |--> S [23.646]
[NLPParser]                    |--> VP [23.369]
[NLPParser]                       |--> VBZ [6.812]
[NLPParser]                          |--> aims
[NLPParser]                       |--> S [12.009]
[NLPParser]                          |--> VP [11.745]
[NLPParser]                             |--> TO [0.010]
[NLPParser]                                |--> to
[NLPParser]                             |--> VP [11.716]
[NLPParser]                                |--> VB [5.716]
[NLPParser]                                   |--> create
[NLPParser]                                |--> NP [3.966]
[NLPParser]                                   |--> PRP [1.320]
[NLPParser]                                      |--> it
[NLPParser]        |--> . [0.004]
[NLPParser]           |--> .
```

Figure 3.2: Parse Tree for D_1

```
[NLPParser] |--> ROOT [332.527]
[NLPParser]     |--> S [332.376]
[NLPParser]         |--> NP [36.306]
[NLPParser]             |--> JJ [9.647]
[NLPParser]                 |--> artificial
[NLPParser]             |--> NN [7.993]
[NLPParser]                 |--> intelligence
[NLPParser]             |--> NN [11.935]
[NLPParser]                 |--> textbook
[NLPParser]         |--> VP [288.415]
[NLPParser]             |--> VB [9.009]
[NLPParser]                 |--> define
[NLPParser]             |--> SBAR [276.027]
[NLPParser]                 |--> IN [0.651]
[NLPParser]                     |--> that
[NLPParser]                 |--> S [275.048]
[NLPParser]                     |--> NP [22.512]
[NLPParser]                         |--> JJ [9.647]
[NLPParser]                             |--> artificial
[NLPParser]                         |--> NN [7.993]
[NLPParser]                             |--> intelligence
[NLPParser]                     |--> VP [252.207]
[NLPParser]                         |--> VBZ [0.147]
[NLPParser]                             |--> is
[NLPParser]                         |--> NP [101.326]
[NLPParser]                             |--> NP [35.197]
[NLPParser]                                 |--> NP [10.386]
[NLPParser]                                     |--> DT [0.641]
[NLPParser]                                         |--> the
[NLPParser]                                     |--> NN [7.993]
[NLPParser]                                         |--> intelligence
[NLPParser]                                 |--> PP [24.153]
[NLPParser]                                     |--> IN [0.667]
```

```
[NLPParser]                 |--> of
[NLPParser]              |--> NP [23.087]
[NLPParser]                 |--> NN [7.253]
[NLPParser]                    |--> machine
[NLPParser]                 |--> CC [0.165]
[NLPParser]                    |--> and
[NLPParser]                 |--> NN [9.862]
[NLPParser]                    |--> robot
[NLPParser]        |--> , [0.000]
[NLPParser]           |--> ,
[NLPParser]        |--> NP [24.607]
[NLPParser]           |--> NP [9.903]
[NLPParser]              |--> DT [0.641]
[NLPParser]                 |--> the
[NLPParser]              |--> NN [7.510]
[NLPParser]                 |--> field
[NLPParser]           |--> PP [14.046]
[NLPParser]              |--> IN [4.044]
[NLPParser]                 |--> as
[NLPParser]              |--> NP [9.604]
[NLPParser]                 |--> NN [7.263]
[NLPParser]                    |--> study
[NLPParser]        |--> CC [0.165]
[NLPParser]           |--> and
[NLPParser]        |--> NP [34.059]
[NLPParser]           |--> NP [10.428]
[NLPParser]              |--> NN [7.770]
[NLPParser]                 |--> design
[NLPParser]           |--> PP [22.973]
[NLPParser]              |--> IN [0.667]
[NLPParser]                 |--> of
[NLPParser]              |--> NP [21.907]
[NLPParser]                 |--> JJ [10.207]
```

```
[NLPParser]                    |--> intelligent
[NLPParser]                 |--> NN [8.112]
[NLPParser]                    |--> agent
[NLPParser]     |--> SBAR [142.029]
[NLPParser]        |--> WHADVP [1.965]
[NLPParser]           |--> WRB [1.896]
[NLPParser]              |--> where
[NLPParser]        |--> S [137.619]
[NLPParser]           |--> NP [25.453]
[NLPParser]              |--> DT [3.233]
[NLPParser]                 |--> an
[NLPParser]              |--> JJ [10.207]
[NLPParser]                 |--> intelligent
[NLPParser]              |--> NN [8.112]
[NLPParser]                 |--> agent
[NLPParser]           |--> VP [111.836]
[NLPParser]              |--> VBZ [0.147]
[NLPParser]                 |--> is
[NLPParser]              |--> NP [107.067]
 [NLPParser]                 |--> NP [8.685]
[NLPParser]                    |--> NN [6.028]
[NLPParser]                       |--> system
[NLPParser]                 |--> SBAR [96.080]
[NLPParser]                    |--> WHNP [1.447]
[NLPParser]                       |--> WDT [0.880]
[NLPParser]                          |--> that
[NLPParser]                    |--> S [94.183]
[NLPParser]                       |--> VP [93.906]
[NLPParser]                          |--> VP [25.719]
[NLPParser]                             |--> VBZ [9.462]
[NLPParser]                                |--> perceives
[NLPParser]                             |--> NP [12.179]
[NLPParser]                                |--> PRP$ [0.864]
```

```
[NLPParser]                                |--> its
[NLPParser]                              |--> NN [7.762]
[NLPParser]                                |--> environment
[NLPParser]                          |--> CC [0.109]
[NLPParser]                            |--> and
[NLPParser]                          |--> VP [63.903]
[NLPParser]                            |--> VBZ [4.855]
[NLPParser]                              |--> takes
[NLPParser]                            |--> NP [53.807]
[NLPParser]                              |--> NP [9.312]
[NLPParser]                                |--> NN [6.654]
[NLPParser]                                  |--> action
[NLPParser]                              |--> SBAR [42.193]
[NLPParser]                                |--> WHNP [1.447]
[NLPParser]                                  |--> WDT [0.880]
[NLPParser]                                    |--> that
[NLPParser]                                |--> S [40.297]
[NLPParser]                                  |--> VP [40.020]
[NLPParser]                                    |--> VBZ [11.195]
[NLPParser]                                      |--> maximizes
[NLPParser]                                    |--> NP [24.126]
[NLPParser]                                      |--> NP [12.782]
[NLPParser]                                        |--> PRP$ [0.864]
[NLPParser]                                          |--> its
[NLPParser]                                        |--> NN [7.449]
[NLPParser]                                          |--> chance
[NLPParser]                                      |--> PP [10.938]
[NLPParser]                                        |--> IN [0.667]
[NLPParser]                                          |--> of
[NLPParser]                                        |--> NP [9.872]
[NLPParser]                                          |--> NN [7.531]
[NLPParser]                                            |--> success
[NLPParser]    |--> . [0.004]
```

```
[NLPParser]           |--> .
```

Figure 3.3: Parse Tree for D_2

We have also given some of the terminologies related to the above constructed parse trees in Table 3.1. Next, the graph of each document parse tree D1 and D2 is constructed. The graph of each document as shown in Figure 3.4 and Figure 3.5 respectively is displayed by using library jgraphx.jar.

Table 3.1: Terminologies Related to Parse Tree

S No	Representation	Explanation
1	S	Starting Node
2	NP	Noun Phrase
3	NN	Noun Singular
4	NNS	Noun Plural
5	NNP	Proper Noun, Singular
6	NNPS	Proper Noun, Plural
7	VB	Verb, base form
8	DT	Determiner
9	PP	Possessive Pronoun
10	ADJP	Adjective Phrase
11	ADVP	Adverb Phrase
12	SBAR	Subordinate Clause
13	CC	Coordinating Conjunction
14	JJ	Adjective
15	IN	Preposition
16	PDT	Pre Determiner
17	CD	Cardinal Number
18	JJR	Adjective Comparative
19	JJS	Adjective Superlative
20	VBN	Verb, Past Participle

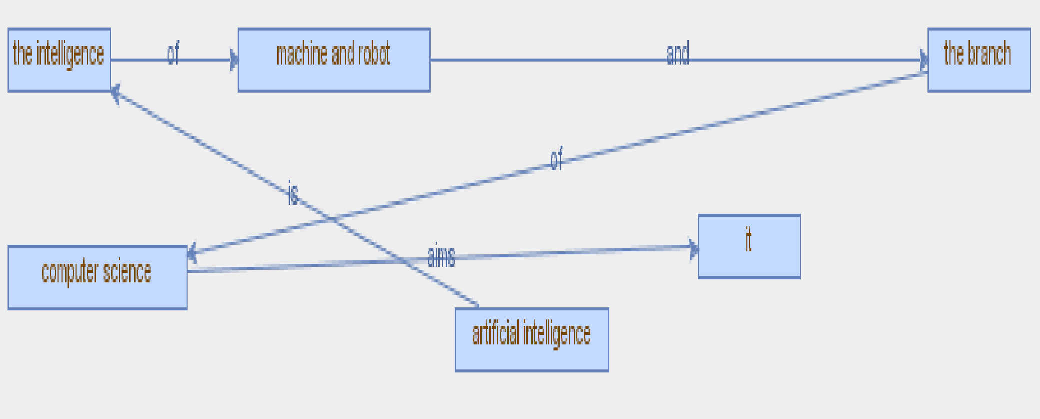


Figure 3.4: Original Document Graph for D1

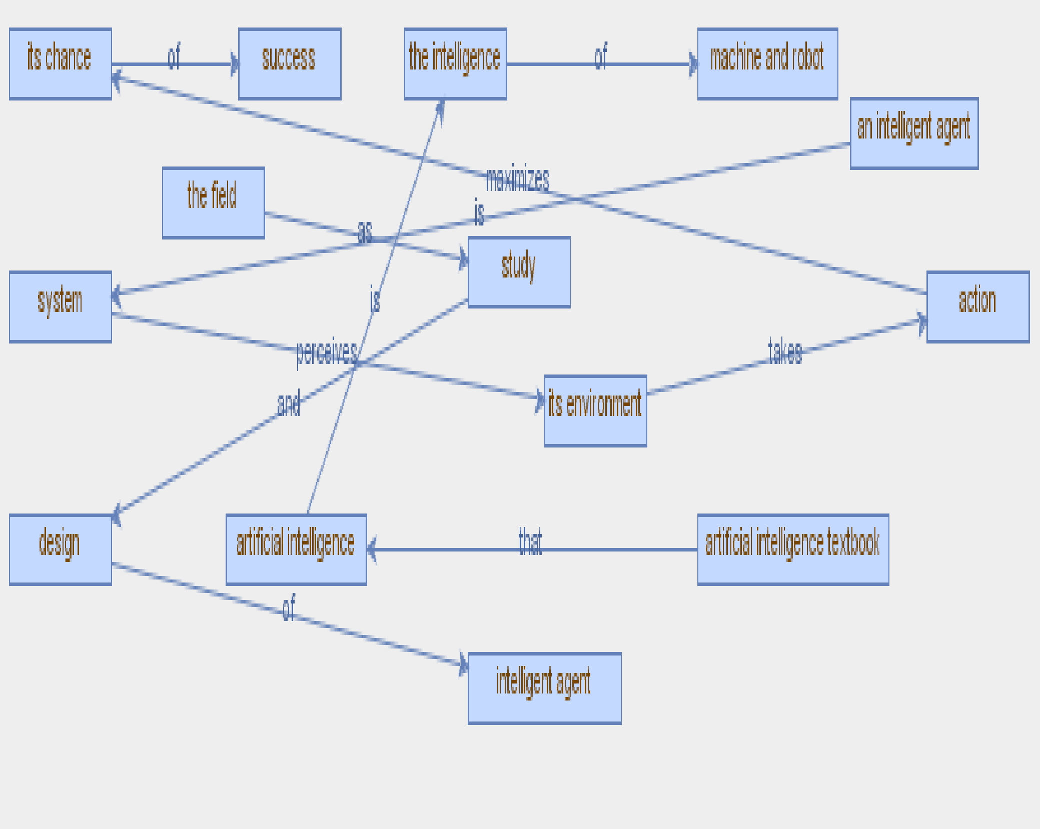


Figure 3.5: Original Document Graph for D2

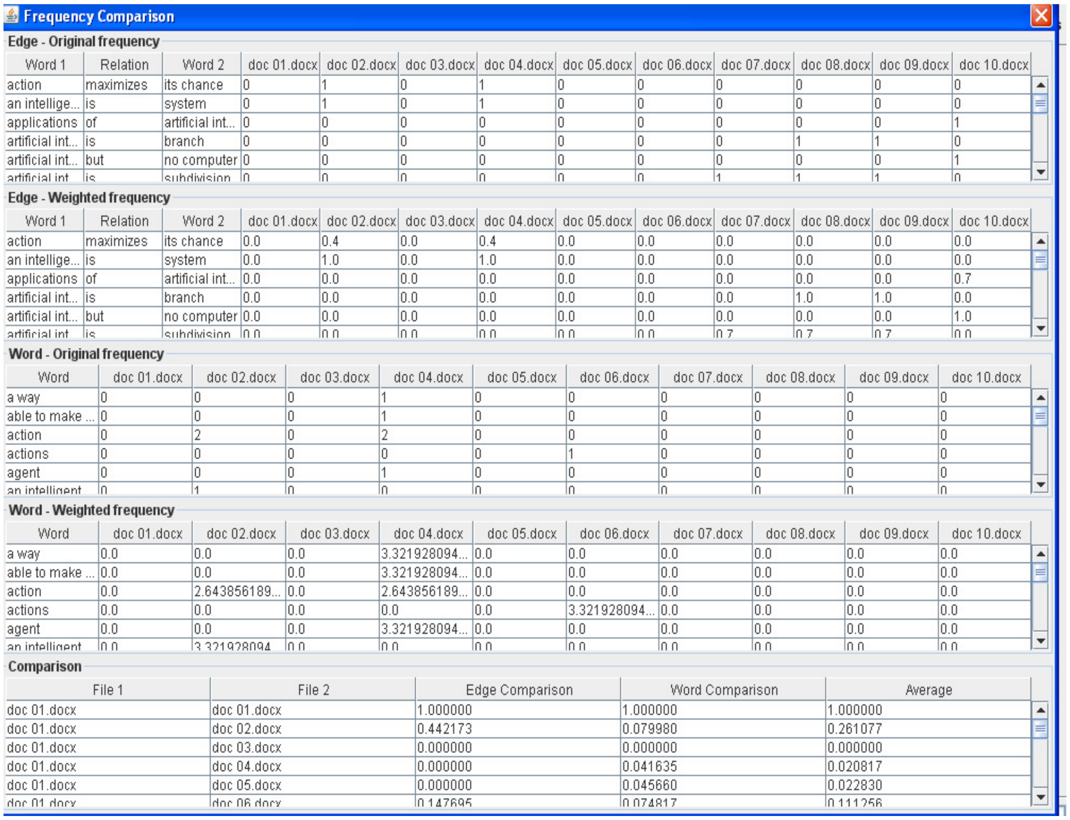
Now, the given two documents D1 and D2 are analyzed for similarity detection using the above explained relation based measuring technique. First, the words extracted from each document are considered to construct the set of semantically similar words also called concepts by using the domain dictionary which sample part is shown in Table 3.2. The complete dictionary is given in Appendix II Table 2.1.

Table 3.2: Domain Dictionary having Words and Concepts

S No	**Words**	**Synonyms as Concepts**
1	computer	machine, device, expert, calculator, estimator
2	study	survey, work, report, discipline, cogitation, examine, analyze, field
3	machine	device, product, mechanism, create, produce, make, shape
4	science	branch, discipline, field, power, ability, skill
5	intelligent	ability, knowledge, power
6	design	plan, blueprint, conception, innovation, contrive, pattern
7	agent	factor, broker
8	system	scheme, organization, arrangement
9	one	single, unity
10	expert	good, proficient, practiced
11	processing	treat, action, work
12	way	manner, mode, fashion, style
13	use	usage, role, purpose, apply
14	aid	assistance, assist, service, avail
15	computing	field, discipline, division
16	scheme	organization, arrangement
17	purpose	intent, objective, target, aspire
18	power	ability, information, knowledge
19	branch	discipline, field, subject, division
20	line	path, trend, row, track, flow
21	strong	stiff, substantial, firm, secure,
22	weak	light, unaccented, decrepit, feeble, infirm, frail

Next, the lexical patterns are retrieved between each concept pair of the two compared documents by using the Google search engine and the lexical similarity between snippets of concept pair is computed using equation 3.1. With the help of the words/concepts obtained above for each document the Vector Space Model is constructed having the words/concepts of each document along with the frequency (i.e. term frequency *tf*) of the same as shown in Word-Original frequency table in Figure 3.6. Now, the weight of each word/concept is computed by computing the *idf***tf*. The *idf* is the inverse document frequency which is calculated as $\log_2 (\frac{tf}{N})$ where N is the total number of documents as shown in Word-Weighted frequency table in Figure 3.6.

Similarly, the Relation Space Model of documents is constructed by finding relationships between each word/concept pair along with the corresponding frequency which is represented as Edge-Original frequency table in Figure 3.6. Next, these relation frequencies are multiplied with the weights of each corresponding word pair relationship which is represented as Edge-Weighted frequency table in Figure 3.6.



Edge - Original frequency

Word 1	Relation	Word 2	doc 01.docx	doc 02.docx	doc 03.docx	doc 04.docx	doc 05.docx	doc 06.docx	doc 07.docx	doc 08.docx	doc 09.docx	doc 10.docx
action	maximizes	its chance	0	1	0	1	0	0	0	0	0	0
an intellige...	is	system	0	1	0	1	0	0	0	0	0	0
applications	of	artificial int...	0	0	0	0	0	0	0	0	0	1
artificial int...	is	branch	0	0	0	0	0	0	0	1	1	0
artificial int...	but	no computer	0	0	0	0	0	0	0	0	0	1
artificial int	is	subdivision	0	0	0	0	0	0	1	1	1	0

Edge - Weighted frequency

Word 1	Relation	Word 2	doc 01.docx	doc 02.docx	doc 03.docx	doc 04.docx	doc 05.docx	doc 06.docx	doc 07.docx	doc 08.docx	doc 09.docx	doc 10.docx
action	maximizes	its chance	0.0	0.4	0.0	0.4	0.0	0.0	0.0	0.0	0.0	0.0
an intellige...	is	system	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
applications	of	artificial int...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.7
artificial int...	is	branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0
artificial int...	but	no computer	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
artificial int	is	subdivision	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.7	0.7	0.0

Word - Original frequency

Word	doc 01.docx	doc 02.docx	doc 03.docx	doc 04.docx	doc 05.docx	doc 06.docx	doc 07.docx	doc 08.docx	doc 09.docx	doc 10.docx
a way	0	0	0	1	0	0	0	0	0	0
able to make ...	0	0	0	1	0	0	0	0	0	0
action	0	2	0	2	0	0	0	0	0	0
actions	0	0	0	0	0	1	0	0	0	0
agent	0	0	0	1	0	0	0	0	0	0
an intelligent	0	1	0	0	0	0	0	0	0	0

Word - Weighted frequency

Word	doc 01.docx	doc 02.docx	doc 03.docx	doc 04.docx	doc 05.docx	doc 06.docx	doc 07.docx	doc 08.docx	doc 09.docx	doc 10.docx
a way	0.0	0.0	0.0	3.321928094...	0.0	0.0	0.0	0.0	0.0	0.0
able to make ...	0.0	0.0	0.0	3.321928094...	0.0	0.0	0.0	0.0	0.0	0.0
action	0.0	2.643856189...	0.0	2.643856189...	0.0	0.0	0.0	0.0	0.0	0.0
actions	0.0	0.0	0.0	0.0	0.0	3.321928094...	0.0	0.0	0.0	0.0
agent	0.0	0.0	0.0	3.321928094...	0.0	0.0	0.0	0.0	0.0	0.0
an intelligent	0.0	3.321928094	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Comparison

File 1	File 2	Edge Comparison	Word Comparison	Average
doc 01.docx	doc 01.docx	1.000000	1.000000	1.000000
doc 01.docx	doc 02.docx	0.442173	0.079980	0.261077
doc 01.docx	doc 03.docx	0.000000	0.000000	0.000000
doc 01.docx	doc 04.docx	0.000000	0.041635	0.020817
doc 01.docx	doc 05.docx	0.000000	0.045660	0.022830
doc 01.docx	doc 06.docx	0.147695	0.074817	0.111256

Figure 3.6: Similarity Computation Using Weighted RSM and VSM

These weights are already stored in the excel file for each relationship that exist between any word pair and for computation we have used the Apache POI libraries to read the data from excel file. The ontology based weights sample part is shown in Table 3.3 and the complete details are given in Appendix II Table 2.3. Next, the Relation Space Model is used for computation of relational similarity between two documents by using equation 3.2. Finally, the collective frequency weights are measured for computing semantic similarity score using equation 3.3. The empirical computation for set of documents related to domain Artificial Intelligence given in Appendix I Table 1.1 using above explained techniques is shown in Figure 3.6.

Table 3.3: Ontology Based Weights

Word/Concept	**Relationship**	**Word/Concept**	**Weight**
artificial intelligence	is	intelligence	1
intelligence	of	machine and robot	0.8
machine and robot	and	branch	0.8
the branch	of	computer science	0.6
computer science	aims	it	0.1
intelligent agent	is	system	1
system	perceives	environment	0.9
environment	takes	actions	0.3
action	maximizes	chance	0.4
help	of	fuzzy inference system	0.7
artificial intelligence	is	field	0.7
human intelligence	is	ability	0.8
sense	of	ambiguous message	0.7
expert	in	particular domain	0.8
applications	of	artificial intelligence	0.9

Word/Concept	Relationship	Word/Concept	Weight
artificial intelligence	are	expert system	1
expert system	is	program	1
program	as	expert	1
automatic programming	is	special programs	1
special programs	as	intelligent tools	0.9
complex behavior	of	individual or group	0.6
artificial intelligence	covers	key challenges	0.9
human knowledge	and	thought process	0.6

The relation based approach considers the concepts and weighted relationship among them from each document, and by analyzing these weights we get an idea that instead of assigning weight to each relationship of base ontology we can visualize each document at two levels to extract explicit and implicit information from a document. First, is at conceptual level which is related to the explicit information stored in the documents in the form of words/concepts, and second is the descriptive level which is related to the hidden/implicit semantic information in the document. In next section, we will propose a scheme of ranking where these two levels (conceptual and descriptive) will be used to retrieve the sounder results. Before giving the Genetic Algorithm based approach we will be giving some approaches that have already utilized the advantages of the same.

3.3 DOCUMENT SIMILARITY COMPUTATION USING GENETIC ALGORITHM (GA)

As we discussed in the previous section, that the document can be visualized at two levels to add the explicit and implicit information in a document. First, the conceptual level which is related to the explicit concepts available in the document, and second the descriptive level related to the implicit semantic information. The implicit semantic information is not present directly in the document but can be inferred from the existing concepts with the help of additional information present in dictionary like WordNet and knowledge structure like Ontology. Therefore, in this technique information at these two levels is extracted and used to calculate the similarity by giving them a fair weightage. The weightage to the conceptual and the descriptive information is decided by using Genetic Algorithm. The final values of weights to the conceptual and the descriptive information are calculated by taking average of the all values of weights to the conceptual and the descriptive information of all the documents under consideration. In coming sub-sections, we will discuss the early works using GA and the proposed technique of document similarity using GA in detail.

3.3.1 Early Works Using Genetic Algorithm

For efficient retrieval of information from WWW [2], Genetic Algorithm (GA) has been extensively used for dealing with the optimization problems [57]. A GA is a modification of stochastic beam exploration in which successor states are produced by merging two parent states, instead of transforming a single state. The Genetic Algorithm consists of four stages Initialization, Selection, Reproduction and Termination.

The relation based semantic similarity approach helps in capturing the lexical matching in combination with the consideration of relations along with the concepts of domain related documents. Although it helps in analyzing and processing of the documents but there is a need to analyze the document to the next level of understanding i.e. semantic level rather than syntactic structure level. Various approaches have been discussed imbedding the semantic in similarity detection techniques to provide the user a document of his/her interest while searching for particular information. In this era, of semantic similarity Genetic Algorithm has played a vital role.

[53] Has done the ontology evolution using semantic Genetic Algorithm to incorporate the concepts which are more relevant to a domain rather than irrelevant concepts. Similarly, Genetic Algorithm has also been used for searching the terms in Gene ontology [54] which helps in retrieving batch and deal with the large state space search. Wang Wei et. al. [55] has given an overview related to Semantic Search Systems which gives the survey on the traditional research trends in the semantic

search field. The analysis and findings based on which a generalized semantic search framework can be designed with the future scope for improvement in semantic search area is also given.

3.3.2 Semantic Similarity Using Genetic Algorithm for Ranking of Web Documents

Generally, various researchers analyze the text of a web page by extracting the keywords/concepts from the web page to find the relevance of the page with respect to the other document or a search engine query given by a user. To find the relevant semantic similarity of a web page with the topic/domain or to a query, the web page is analyzed by considering user view at conceptual level and as well as at descriptive level. The conceptual level is basically associated with the facts of the content available in the document with respect to the words or concepts which are physically present in each sentence of the document. Whereas, the descriptive level of analyzing the document considers the broad view of the content by identifying the relationships between words or concepts available in the document. Both these levels of viewing/analyzing a web page are significant but their relative importance may differ from a sentence to another sentence of the same document/web page. Therefore, the relative importance of these two levels is represented in terms of weights. These weights are determined prior to its usage by applying GA and by using the conceptual level and the descriptive level information present a given document in the sample set of documents. The final weights for a given set of documents are calculated by taking average of the individual final weights corresponding to the documents in the sample space. These final weights indicate how their relative importance is being used to write a document. These weights are then used to calculate the similarity between the query and the documents in the set of test documents.

At conceptual level, the words are extracted from a document and query to construct the vector space model. The similarity value at conceptual level is computed using the cosine similarity function as

$$\mathrm{Sim}_{\mathrm{conceptual}}(\mathrm{D},\mathrm{Q}) = \frac{\vec{\mathrm{V}}(\mathrm{D}).\vec{\mathrm{V}}(\mathrm{Q})}{|\vec{\mathrm{V}}(D)||\vec{\mathrm{V}}(\mathrm{Q})|} \qquad 3.4$$

At descriptive level, the assignment of weights is done according to the description i.e. the relationships between the concepts available in the document by using base ontology and domain specific dictionary. In domain specific dictionary we are storing the words along with the concepts and in base ontology is having all the domain related concepts along with the relationships between the concepts that exist. Normally, the description of any document can be given in numerous ways, but primarily the description is associated with the number and type of relationships that exists between the concepts present in the document. The final similarity of a document with respect to the query is calculated by using the formula as:

$$\text{Sim (D,Q)}=w_{1f}*Sim_{Conceptual}(D, Q)+w_{2f}*\ \text{Sim}_{\text{Descriptive}}\ (D,Q)$$

where w_{1f} and w_{2f} are the weight constant used at the conceptual and the descriptive level respectively.

So, the fundamental approach of the proposed scheme is to use these two types of weighted information related to the conceptual and descriptive level with their respective weights (w_{1f} and w_{2f}). These values of weight constants w_{1f} (conceptual level) and w_{2f} range between interval [0, 1] excluding 0 and 1. It may be noted that, these values are average of the final values of w_1 and w_2 of all documents in the sample space. The values of w_1 and w_2 of a document in sample space are determined as a result of adjustment in these w_1 and w_2 by using the Genetic Algorithm (GA).

3.3.2.1 Applying Genetic Algorithm

Now, let discuss how we are calculating the w_1 and w_2 for a given document using GA. Initially a set of these weights containing *n* pairs are taken as initial population required by GA by using random values between 0 and 1. The second step is to select the most promising k numbers of pairs from this available population on the basis of a fitness function described next. As it is evident, from the nature of the problem, that a formal fitness function is not possible in this case. Therefore, an approximate fitness function is designed using final similarity values obtained for each document in the sample space on the basis of human analysis. This analysis is done by providing the documents in the sample space to individual expert in the domain. The probable pair of w_1 and w_2 is used to give the final value of similarity for a given document by considering, off course, using the information at the conceptual and the descriptive levels. The process of optimizing the values of w_1 and w_2 will be terminated if the desired level of optimization has been achieved. Otherwise, n-k of pairs of w_1 and w_2 are generated by using crossover and mutation operations. These newly generated n-k numbers of pairs will be added to the k number of parent pairs (chromosomes) to get once again n numbers of pairs and control will be given to next iteration for the further optimization. The overall steps of initial population, selection, evaluation, generation of new population (new pairs of weights w_1 and w_2) is depicted diagrammatically in figure 3.7.

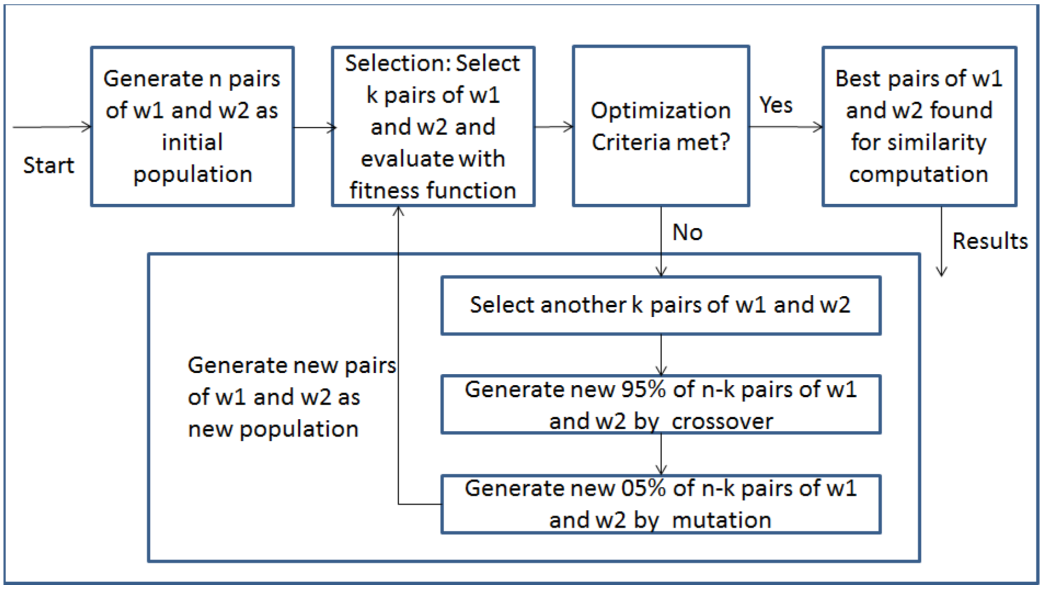

Start
Generate n pairs of w1 and w2 as initial population
Selection: Select k pairs of w1 and w2 and evaluate with fitness function
Optimization Criteria met?
Yes
Best pairs of w1 and w2 found for similarity computation
No
Results
Select another k pairs of w1 and w2
Generate new pairs of w1 and w2 as new population
Generate new 95% of n-k pairs of w1 and w2 by crossover
Generate new 05% of n-k pairs of w1 and w2 by mutation

Figure 3.7: Basic structure for generating promising pairs of w1 and w2 using GA

The similarity using w_1 and w_2 is computed as, for example, suppose the similarity score of a document with respect to query obtained from human analysis is 0.77. Now, the final similarity of the document with respect to query will be calculated by using the proposed approach. Let us consider the conceptual score ($Sim_{Conceptual}(D,Q)$) and descriptive score ($Sim_{Descriptive}(D,Q)$) of the document is 0.5 and 0.6 respectively and current values of w1 and w2 are 0.56 and 0.72 respectively. The final similarity is calculated by using the formula as:

$$Sim\ (D, Q)=w1*Sim_{Conceptual}(D,Q)+w2*\ Sim_{Descriptive}\ (D,Q)$$

$$Sim\ (D, Q)=0.56*0.5+ 0.72*\ 0.6$$

$$=0.712$$

This value of final similarity will be compared to human analyzed value i.e 0.77. If the difference between the calculated similarity and human analysis based similarity more than 0.02, the w1 and w2 once again will be modified using GA. In general, in order to see how much promising is the similarity value, we are comparing the computed similarity score with human analysis based score. The difference in the value indicates whether the pair of w_1 and w_2 are promising or not. The closer the calculated value/score with human analysis score indicates better the pair of w_1 and w_2 used in computation of similarity of a document with respect to the query. The computation of pairs of desired w_1 and w_2 for a given document is given in algorithm 3.1.

Algorithm 3.1: Computation of pairs of w_1 and w_2 using GA

Input: *n* random pairs of w_1 and w_2 for a document

Output: Optimized pairs of w_1 and w_2 for a document

1. Initial Population: Consider the n pairs of w_1 and w_2 generated by using random function.
2. Selection: k pairs of w_1 and w_2 from these n pairs are extracted based on the fitness function which provides the similarity score of a document based on human analysis.
    - i. For all n pairs of w_1 and w_2 compute document similarity

      $\text{Sim}(D,Q)_i = w_1 * Sim_{Conceptual}(e_i, e_j) + w_2 * \text{Sim}_{\text{Descriptive}}(D,Q)$.

    - ii. Select the k numbers of pairs of w_1 and w_2 which are closest to $\text{Sim}_{\text{Human}}$.
3. For all k $\text{Sim}(D,Q)_i$, check whether the difference between any of the $\text{Sim}(D,Q)_i$ and $\text{Sim}_{\text{Human}}$ less or equal to $\eta(0.02)$. Make the pairs of w_1 and w_2 corresponding to that $\text{Sim}(D,Q)_i$ as the final value for the document and terminate the process of optimization. Otherwise, go step 4.
4. Generate n-k number of new population of pairs of (w_1, w_2)
    - i. Generate 95% of n-k new pairs of w_1 and w_2 using crossover operation.
    - ii. Generate 05% of n-k new pairs by using mutation operation.

iii. Add these newly generated pairs of w_1 and w_2 in the k number of parent pairs.

iv. Shift the control to Step 2.

5. End while

Once the final values of pairs of w_1 and w_2 for all documents in the sample space containing m number of documents are determined, the final values of these weights across the sample space is calculated by taking average of all the w_1 and w_2 corresponding to the individual document in the sample space. Compute final w_{1f}` and w_{2f} for a document as:

$$w_{1f}\text{=avg }(w_{1'}, w_{1''}, w_{1'''}............ w_1^m)$$
$$w_{2f}\text{=avg }(w_{2'}, w_{2''}, w_{2'''}...........w_2^m)$$

3.3.2.2 Chromosome Representation

In order to apply the GA, the representation of w_1 and w_2 are done by using their floating point values between 0 and 1 upto two decimal. The floating point numbers are in turn represented as 32-bit single precision (IEEE 754).

For example, from the initial set of weights pairs, let take two values of w1, say them w_{11} and w_{12} represented in binary as follows:

w_{11}=0.5, Most accurate representation = 5.0E-1

Binary representation: 0x3F000000 = 00111111 00000000 00000000 00000000

Sign	Exponent								Mantissa																						
0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

w_{12}=0.6, Most accurate representation = 6.00000023841857910156250E-1

Binary representation: 0x3F19999A = 00111111 00011001 10011001 10011010

Sign	Exponent								Mantissa																						
0	0	1	1	1	1	1	1	0	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	1	0

Crossover Operation

Next, the crossover is performed on the pair for weight (w_{11}, w_{12}). The crossover computation is done by using single point crossover in which we are selecting the middle of the binary string as the crossover point. Then, the 16-bit binary string from start of chromosome to the selected crossover point is copied in the new chromosome from first parent chromosome and the remaining 16 bits are taken from second half of the second parent chromosome. For example, the weights w_{11} and w_{12} crossover is shown below:

00111111 00000000 00000000 00000000 X 00111111 00011001 **10011001 10011010**

= 00111111 00000000 10011001 10011010

Mutation Operation

Next, the mutation is performed by flipping the bits randomly of the binary representation of the remaining weights and generate new (w_{1i}, w_{1j}). The random bit is selected by generating a random number between 1 and 32. Using the new pairs of weights obtained from crossover and mutation the similarity of a document with respect to query is again computed by using the conceptual and descriptive scores. This is done iteratively, as discussed in the algorithm above, until we get desire values of w_1 and w_2 for a given document.

Once the final weight w_1 and w_2 for all documents in sample space is determined, these values will be used to calculate the final weights (w_{1f} and w_{2f}) which in term will be used to calculate the final similarity between the query and the test document in a given set of test documents related to the query. The final similarity calculated this way for all the documents in the given set are in turn used for ranking among the documents.

3.3.2.3 Final Similarity Score Calculation

In order to find the conceptual and descriptive similarity score, the document is first analyzed at sentence level, and then the sentences are combined to view the document as paragraphs which are further combined to view the document as a whole. The conceptual similarity is computed by extracting the words from a document and constructing the vector apace model of these extracted words. The vector space model is used to compute the similarity by using the cosine similarity formula. Next, the descriptive level similarity is computed by extracting all the concepts and relationships among the concepts using the domain specific dictionary and the base ontology. The weight of each relation that exists between the pair of concepts that are present in the document is computed by using the formula as:

$Wt(r_{ij})=Dis(C_i , C_j)/DP(C_i)+ DP(C_j)$

where Dis $(C_{i,} C_{j)}$ is the shortest path between the concepts in the ontology and, DP (C_i), DP (C_j) are depth of the concepts in the ontology.
This is done for each sentence of the document to obtain the sentence level descriptive score. The paragraph level score is obtained by using the statistical analysis measure for combining the sentence level score of the document. Similarly, paragraphs level scores are combined to compute the descriptive similarity score of the document. The overall algorithm for Similarity Detection is as follows:

Algorithm 3.2: Similarity Detection

Input: Set of documents (S), Query (Q) & Ontology (O).
Output: Sim (D, Q) // where D is the Document present in set S and Q is the user query.
Ranked set of documents D=(D1, D2, D3……..Dn)

```
Begin Process
1. For each sentence $S_i$ in document $D_i$
   a. Extract all the available concepts and the relations between these concepts
      with the help of ontology O.
   b. Compute $Wt(r_{ij})=Dis(C_i , C_j)/DP(C_i)+ DP(C_j)$.
// Where Dis $(C_{i,} C_{j)}$ is the shortest path between the concepts in the ontology and,
DP $(C_i)$, DP $(C_j)$ are depth of the concepts in the ontology.
   c. Compute $Sim_{Descriptive}(S_{i,}Q)=\sum Wt(r_{ij})$.
      // where $Sim_{Descriptive}(S_{i,}Q)$ is the descriptive level weight of a sentence of
      a document according to query.
2. For each document $D_i$
      a. Compute $Sim_{Descriptive}(D,Q)=\sum Sim_{Descriptive}(S_{i,}Q)/n$
      // where n is total number of sentences in the document.
```

b. Compute $\mathrm{Sim}_{\mathrm{Conceptual}}(\mathrm{D}, \mathrm{Q}) = \frac{\vec{\mathrm{V}}(\mathrm{D}).\vec{\mathrm{V}}(\mathrm{Q})}{|\vec{\mathrm{V}}(D)||\vec{\mathrm{V}}(\mathrm{Q})|}$

3. Computation of final similarity score using the computed weight sets: $\mathrm{Sim(D,Q)}=w_{1f}*Sim_{Conceptual}(\mathrm{D}, \mathrm{Q})+w_{2f}*\ \mathrm{Sim}_{\mathrm{Descriptive}}(\mathrm{D,Q})$.

   //w_{1f} and w_{2f} obtained using Algorithm 3.1 by taking average of all the optimized pairs of w_1 and $w_{2.}$

4. Ranked set of Documents:
   D=asc (D1, D2, D3……Dn) // Depending upon the similarity score obtained in step 3.

3.3.3 Explanation using Example

Further, the detailed working of our model for finding semantic similarity using Genetic Algorithm is explained with the help of examples. The set of seven documents related to the domain education are considered and they are ranked according to the calculation of semantic similarity with the query "what is education" given to the search engine. These set of documents were first processed at conceptual level to calculate the score of document with respect to the other document which is having the content of query by using lexical matching. Next, these set of documents were processed at descriptive level to calculate the score by considering the extracted relationship between concepts using ontology with respect to the query document. The scores computed at both the levels i.e. conceptual level and descriptive level were modified by using the weight constants w_1 and w_2 by using Genetic Algorithm for considering the scores which provides the relevant ranked set of documents with respect to the query document. The weights are then modified by changing the values of w_1 and w_2 in range as defined between [0, 1] using Genetic Algorithm to perform the number of iterations. Similar computations are done for the rest of the documents to get the final semantic score corresponding to each document. The complete process will provide the optimal ranked set of documents which is near to human analysis.

3.3.4 Performance Analysis of Relation based and Genetic based Semantic Similarity

Performance of the given approaches for detecting the semantic similarity between the web documents certainly rely on how the concepts corresponding to a word and the relations between these concepts are extracted from the document. The construction of such set of related concepts depends on the method of spreading used to create the ontology graph for a document which will again be diverse from one domain to another domain further relying on the formulation of domain related concepts. We have analyzed the performance of given Relation based Semantic Similarity technique by taking the set of 50 documents related to domain education. The education domain dictionary constructed is having words with the synonyms are from the education domain specific documents.

The evaluation of the performance of given relation based technique is done with the traditional Vector Space Model (VSM) and Euclidean Approach (EUC) [3] of lexical matching. The results of the comparison of these approaches are shown in Table 3.4.

Table 3.4: Results of Similarity VSM, EUC and Relation Based Semantic Similarity

S No.	**Set of Documents**	**VSM**	**EUC**	**Relation Based Semantic Similarity**
1.	D_1, D_2	.5	.6	.7
2.	D3, D_1	.4	.49	.8
3.	D4, D_5	.49	.6	.6
4.	D2, D3	.55	.57	.71
5.	D3, D4	.32	.36	.36
6.	D1, D4	.44	.47	.55

Note: The sample parts of documents are as follows:

D_1: Artificial intelligence is the area of computer science focusing on creating machine that can engage on behavior that human consider intelligent.

D_2: Artificial intelligence track focuses on fundamental mechanism that enable the construction of intelligent system that can operate autonomously, learn from experience, plan their actions and solve complex problems.

D_3: Knowledge representation and knowledge engineering are central to artificial intelligence research. Many of the problems machines are expected to solve will require extensive knowledge about the world.

D_4: Intelligent agent must be able to set goal and achieve them. They need a way to visualize future and be able to make choices that maximizes the utility of available courses.
D_5: Machine learning is central to artificial intelligence research. It is study of computer algorithm that improves automatically through experience.

The results produced in Table 3.4 shows the difference in similarity score by considering lexical analysis and on the other hand considering the related concepts to understand the information present in the document by a machine processing technique.

Similarly, the performance of our proposed technique using Genetic Algorithm for detection of semantic similarity is analyzed on the set of documents related to a domain education and its application was shown in ranking of the set of document for a user query related to domain education. Using this approach, the semantic similarity score is improvised by analyzing and processing the document at description and conceptual level for better relevance. Genetic Algorithm helped in computing the score for the conceptual and descriptive level by processing number of iterations retaining the optimal solution to the problem. The results shown in Table 3.5(a) provide the details of some iteration performed to calculate the w_1 and w_2 for document D_1. The sample set of w1 and w2 are shown corresponding to each iteration. Further, the sample set of new pair of weights obtained from crossover and mutation are also given corresponding to the iteration shown. The table also gives the similarity computed using our approach for documents D_1 for some of the iteration and shows that in iteration 20 the final similarity score is obtained giving the error difference of 0.02 with the similarity computed by human analysis. In Table 3.5 (b) the final weights computed for all documents in sample space are given with the computed similarity score and the human analysis based score. In this table the conceptual and descriptive scores are also given corresponding to each document in the sample space. Next, in Table 3.5(c) the computation of final weights for the sample space having seven documents is shown which will be used for computing the similarity of other documents with respect to the query.

Table 3.5(a): Semantic Similarity computed using GA for Document D_1

S. No.	**Document D1 with iteration number**	**Sample set of W1**	**Sample set W2**	**New sample pair of weights using Crossover**	**New sample pair of weights using Mutation**	**Similarity computation w.r.t. query with Conceptual value=0.32 & Descriptive value=0.41 (Human calculated Similarity)**
1.	D_1 iteration:1	0.64 0.65 0.66 0.67	0.75 0.76 0.77 0.78	0.67 0.8 0.69 0.56 0.62 0.67	0.56 0.67	0.45(0.74)
2.	D_1 iteration:4	0.60 0.7 0.68 0.69	0.67 0.79 0.76 0.80	0.63 0.74 0.64 0.75 0.65 0.76	0.36 0.47	0.51 (0.74)
3.	D_1 iteration:8	0.56 0.71 0.62 0.43	0.67 0.73 0.32 0.54	0.33 0.23 0.63 0.47 0.56 0.62	0.88 0.89	0.65 (0.74)
4.	D_1 iteration:20	0.63 0.62 0.31 0.23	0.45 0.74 0.54 0.30	0.62 0.43 0.63 0.56 0.99 0.99	0.77 0.65	0.72 (0.74) Now error is less than 0.02

Note: D1:en.wikipedia.org/wiki/education.

D2:www.teach_kids_attitude_1st.com/definition of education.html

D3:www.motivation_tools.com/youth/what_is_education.html.

D4:education.svtution.org/2011/06/what_is_education.htm.

D5:Dictionary.reference.com/brouse/education.

D6:psychology.about.com/od/educationalpsychology/educational_psychology.htm.

D7:press.chicago.edu/ucp/books/Chicago/w.html.

Table 3.5 (b): Similarity score computed for the sample set of documents

Document (Conceptual, Descriptive)	**W_1**	**W_2**	**Sim Computed**	**Sim Human**
D_1 (.32,.41)	0.99	0.99	0.72	0.74
D_2 (.67, .87)	0.44	0.66	0.87	0.86
D_3 (.57, .62)	0.40	0.52	0.55	0.57
D_4 (.54, .67)	0.33	0.48	0.50	0.49
D_5 (.48, .62)	0.47	0.50	0.56	0.57
D_6 (.54, .44)	0.41	0.41	0.40	0.42
D_7 (.23, .45)	0.39	0.47	0.30	0.31

Table 3.5(c): Final computed weights w_{1f} and w_{2f}

Weight	**Sum**	**Final average value**
W1f	Sum of all optimized w1 (0.99+0.44+0.40+0.33+0.47+0.41+0.39=3.43)	0.49
W2f	Sum of all optimized w2 (0.99+0.66+0.52+0.48+0.50+0.41+0.47=4.03)	0.58

The results obtained in Table 3.6 are giving the set of ranked documents according to the semantic score computed for each document with respect to query by using the pairs of w_{1f} and w_{2f}. Also, these set of documents were ranked according to human analysis rating. The variance of each set of ranked documents obtained from the proposed GA based approach is also compared with the set of ranked documents as per human rating. It has been found statistically that the documents ranked by using GA based approach give minimum score of variance as compared to lexical matching, and relational matching. The results scored from the discussed Genetic Algorithm based semantic similarity detection are presented in Table 3.6. The set of documents used in Table 3.6 are given in Appendix 1 Table 1.1 shown as the plain text retrieved from Google search engine after preprocessing.

It has been found through empirical analysis of the technique, that the *Relation based Similarity measure provides better results. It has also been found through statistical analysis that the results of similarity computation of documents with respect to query are further enhanced by applying Genetic Algorithm to perform the deep processing of the documents through number of iterations according to the given scheme.*

Table 3.6: Result-set of Ranked Documents

S. No.	**Conceptual weight corresponding to each ranked document**	**Descriptive weight corresponding to each ranked document**	**Ranked set according to $\text{Sim}_{\text{Computed}}$ Using w_{1f} and w_{2f}**	**Variance from $\text{Sim}_{\text{Human}}$**
1.	.5, .67, .43, .22, ,.12, .63, .01	.34, .22, .67, .41, .37, .64, .11	D_{17},D_{15},D_{12},D_{11},D_{13},D_{14}, ,D_{16}	6
2.	.70, .41, .46,.68, .98, .33, .21	.33, .56, .76, .81, .19, .34, .02	D_{23},D_{25},D_{24},D_{21},D_{22},D_{26}, D_{27}	4
3.	.63, .14, .45, .76, .88, .90, .55	.47, .80, .19, .04, .88, .56, .91	D_{33},D_{36},D_{34},D_{32},D_{31},D_{37}, D_{35}	16
4.	.22, .67, .88, .99, .02, .06, .88	.80, .99, .16, .73, .37, .45, .67,	D_{42},D_{45},D_{47},D_{41},D_{43},D_{46}, D_{44}	26
5.	.83, .56, .63, .12, .63,.01,.5	.21, .22, .66, .40, .37, .11, .64	D_{30},D_{19},D_{10},D_{20},D_{50},D_{39}, D_{18}	8

The improvised technique provides the much better and optimal solution which is shown by giving the ranking of these documents close to human analysis. In maximum number of cases, the detection of similarity score is better giving much more semantics as compared to the traditional similarity approach showing the superiority of the given techniques.

3.4 SUMMARY

The semantic comparison techniques additionally improve the searching of relevant information from web pages present on WWW. Many similarity computation algorithms have been given and used in the field of information retrieval make use of the concepts and relationships that may subsist between the concepts as discussed in Chapter 2. The approaches presented in this chapter, takes the benefits of ontology to calculate the similarity between the documents by extracting the relevant related concepts for a document to get better similarity score between documents which further provides improved and relevant result-set for a query specified to the search engine by the user. Although the techniques discussed in this chapter, provides meaningful information for the document by giving better similarity score but there is still some issues related to the design of ontology, extraction of related concepts using ontology, construction of the data structure for a document which provides the maximum meaningful information to a reader as conveyed by the author of the document.

Chapter 3 discusses the advanced techniques by giving more meaningful information of the document with the help of document semantic analysis. This will further help the search engine in retrieving relevant result-set. Our upcoming effort and the proposed methods which we will discuss in Chapter 4 considers major and extensive semantic web pages, to find the efficient and relevant semantic similarity between the web pages by using a knowledge base already formed and constructing a new knowledge base in form of a graph for each document.

CHAPTER IV

DOCUMENT SEMANTIC SIMILARITY USING CHAIN OF CONCEPT'S RELATIONSHIPS AND CURRENT TRENDS

4.1 INTRODUCTION

In general, the retrieval of information from web is tedious task as the web document is written in plain text using natural language which is difficult to understand and further process by machine efficiently. Thus, in this Chapter two techniques are given to understand the document using dictionary like WordNet and knowledge base like Ontology. Ontology is a structured system considered to classify and analyze the relationships between different concepts of knowledge which is widely accepted by the computational field.

Lamberti F. et. al. [4] proposed Relation based Page Rank algorithm has already used the ontology for ranking of documents by semantic web search engine. The ranking of page is done by computing its relevance by exploiting the relations available in the page and the query defined by a user. [26] Computed the semantic similarity of documents using ontology extraction algorithm which helps in finding similarity between documents which were dissimilar using keyword approaches. Use of multi-tree model using ontology by combining two trees of documents considered for semantic similarity computation [8].

The approaches which are using ontology for semantic similarity computation represents a document as Bag of Concepts (BOC) [27]. Even the document ontology is expanded using schemes available like set spreading and semantic network.

An approach which computes semantic similarity for paraphrase identification [42] also uses the formula for similarity computation between any two obtained sentences as:

$$\mathrm{Sim}(a, b)=aWb/|a||b|$$

In the above formula, W is a semantic similarity matrix which takes the information about the similarity of words.

It has been found that the researcher's main aim for finding semantic similarity score between texts close to human analysis is by considering ontology for identification of related concepts. It can be either consideration of a document into chunks which can be extended using WordNet for adding meaningful information for analysis of a text.

4.2 ONTOLOGY BASED SEMANTIC SIMILARITY FOR DOCUMENTS

For processing of a document it is initially parsed using Stanford Parser to extract the words or phrases from document. These words are then extended using WordNet to inculcate the related words of the document which are not physically and literally present in a document. These extended words are then represented in the form of ontology as nodes which are connected with each other using edges representing relationships that exist between them.

The extraction of keywords is done with the help of a parser from a document that further helps in finding noun, verb, preposition, adjective, adverb etc. Few researchers have stored only noun, verb and adjective out of extracted words in a database removing rest of the keywords. The database is then compared using an ontology constructed related to a domain to find the relations between the words obtained and stored in the database [1].

We have given an approach for computation of semantic similarity which relies on the structured knowledge related to a domain stored in form of ontology. The detailed architecture of ontology based approach is shown below in Figure 4.1. The key mechanism of the given architecture of the approach is Ontology Processor, Graph Construction Module, Ranker Module and Document Processor.

In this ontology dependent approach, first step is processing of document for extraction of words present in a document using syntactic analysis techniques. These extracted words helps in making the Vector Space Model for document having extracted words with the frequency based on number of existence of the each word in a document. The relations which exist between the words present in a document are extracted and stored in a relationships repository.

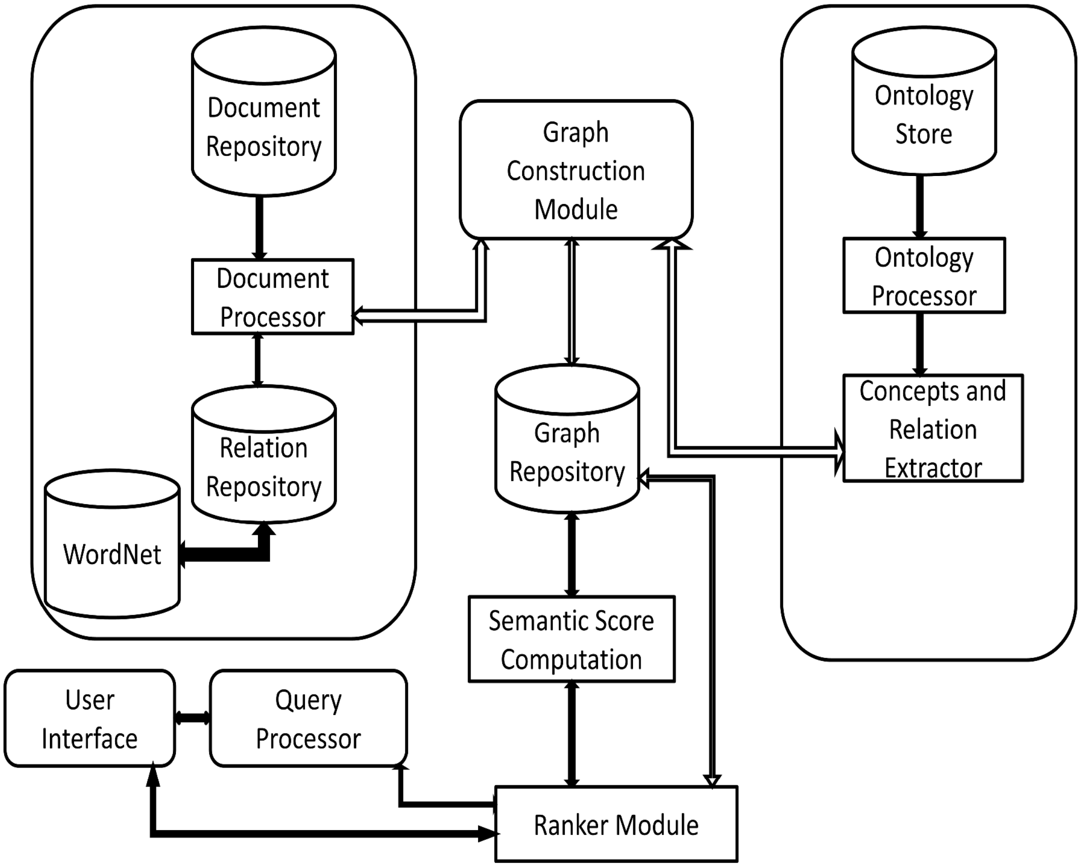


Figure 4.1: Architecture of Proposed Semantic Similarity Model

The relation repository consists of extracted relations along with the weights assigned to each relation by applying fuzzy set theory which shows its importance and relevance between words in document. According to fuzzy set theory, a relation is also a fuzzy set where each relation is given a weight from interval [0, 1] indicating relationship grades of these weights to each element which is considered as relation between words present in a document. The more the value is closer to upper range will indicate high degree of association and the value which is closer to lower range will denote low degree of association. The relations along with the weights assigned are stored in a database as shown in Table 4.1.

It is an assumption that relational repository constructed is considered as the fuzzy set which is defined by the association of each relation between words. The constructed relational repository and a base ontology are then used for processing of a document which helps in retrieving the concepts and relationships that exist between the concepts.

Table 4.1: Relation Table having Weight along with the Description

SNO	Relation	Weights	Description
1	type of	1	--------
2	is a	1	--------
3	Of	.8	--------
4	part of	1	--------
5	kind of	1	--------
6	Using	.5	--------
7	At	1	--------
8	Has	.9	--------
9	Through	.9	--------

The extracted concepts and relations are then spreaded using available spreading techniques like semantic networks or frame networks to build a knowledge representation network representing semantic relations between the concepts. This network can be undirected or directed and it consists of nodes and edges where nodes represent concepts and edges represent relationships. In our approach, we have considered the spreaded document knowledge representation as undirected so that all possible concepts and relations can be considered to capture the knowledge contained in any document to the deepest level. The structure used to represent the knowledge obtained by our approach can be any like link list, graph, matrix representation etc. But, for simplicity and embedding computational efficiency we have taken graph representation to represent our document in terms of extracted concepts and relations from relation repository. The construction of graph also involves the use of a domain dictionary, which is containing the words from a domain along with the synonyms of the words. This domain dictionary is said to be the lexical database for our approach as it helps in extracting concepts from a document to construct the ontology for that document.

Now, the document graph called ontology is used to find the semantic score between any two documents as now the computation is done for both nodes and edges incorporate maximum knowledge of documents. The resemblance between any two graphs of the documents is computed by using the probability intersection computation as it helps in detecting the common concepts and relationships between them that occur in both the documents.

$$\mathrm{P}(\mathrm{A} \cap \mathrm{B}) = \frac{1-(\mathrm{n}(\mathrm{G}(\mathrm{A} \cap \mathrm{B}))+\mathrm{r}(\mathrm{G}(\mathrm{A} \cap \mathrm{B}))}{\mathrm{n}(\mathrm{G}(\mathrm{A}))+\mathrm{n}(\mathrm{G}(\mathrm{B}))+\mathrm{r}(\mathrm{G}(\mathrm{A}))+\mathrm{r}(\mathrm{G}(\mathrm{B})} \qquad 4.1$$

Where, n (G (A∩B)) and r (G (A∩B)) symbolize the common number of nodes and relations from the graphs of the two documents for which computation of similarity is to be computed. The n(G(A)), n(G(B)) correspond to the total numeral of nodes in the graph of the documents A and B. Likewise r(G(A)) and r(G(B)) corresponds to the numeral of the relationship that exists in the constructed graphs of two documents. Figure 4.2(a) and Figure 4.2(b) shows the graph of document A and document B respectively where document A is containing the contents as={Android based phones are better than Window based phone} and document B is containing the content as={Samsung based mobiles are better than Nokia based mobiles}.

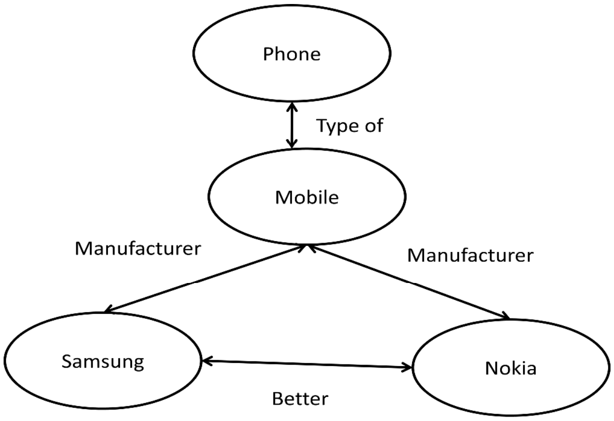


Figure 4.2(a): Graph of Document A

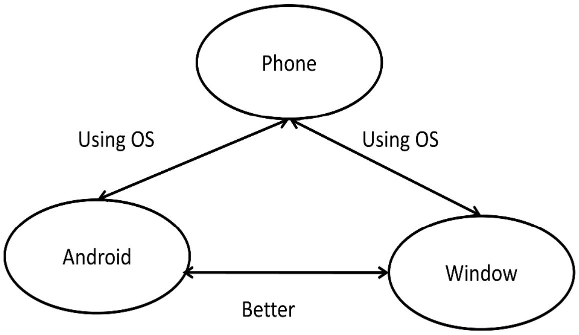


Figure 4.2(b): Graph of Document B

These constructed graphs are extended with the help of spreading process by using ontology as shown in Figure 4.4 and the extension process also considers the domain dictionary.

The different representation for spreaded graph cannot be captured by using lexical techniques as these approaches are incompetent to capture the conceptual view and therefore they cannot find implicit concepts. The extended graph of document A and document B, obtained by using the spreading technique of semantic networks and base ontology are shown in Figure 4.3(a) and Figure 4.3(b).

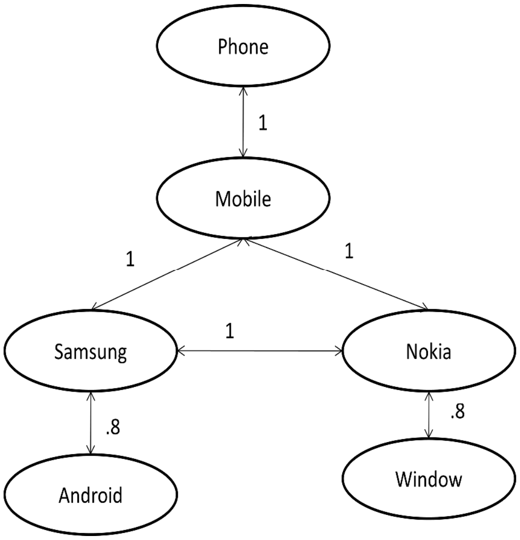


Figure 4.3(a): Graph of Document A after Spreading

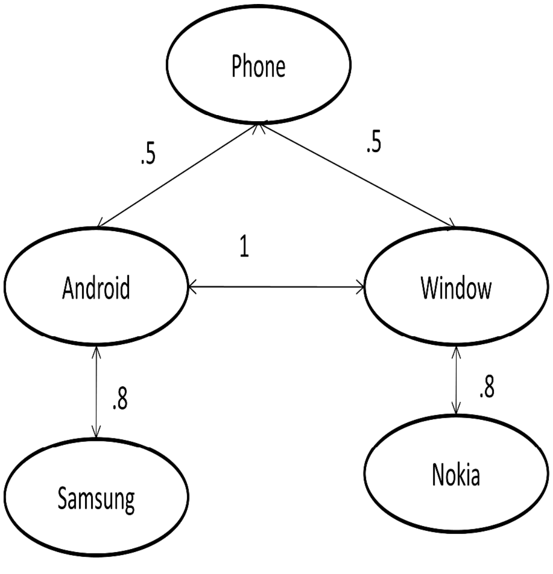


Figure 4.3(b): Graph of Document B after Spreading

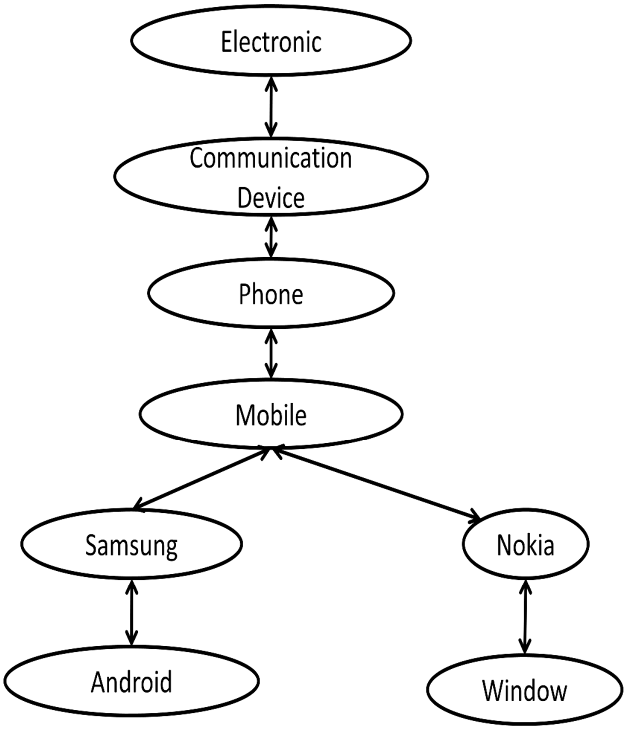


Figure 4.4: Given Ontology O

Final computation of semantic similarity is done by finding the value of number of nodes and relations as in our example n (G (A∩B)) =1 and r (G (A∩B)) =2. The count for number of nodes and number of relations for each document retrieved from graphs of Figure 4.3(a) and Figure 4.3(b) is as follows n(G(A))=6, n(G(B)) =5 and r(G(A))=6, r(G(B))=5. Using equation 4.1 the similarity is denoted by P (A∩B) = .86. This approach is also applied on set of 50 more documents related to mobile domain given in Appendix I Table 1.2 containing the content from which implicit concepts are extracted to capture the user view about the document written by an author. Thus, the semantic similarity of documents cannot be identified only by using the lexical based matching techniques available as there are documents present on web where the documents convey the same idea but may use different representation.

4.3 CONCEPTUAL SEMANTIC SIMILARITY DETECTION TECHNIQUE

The technique presented in section 4.1 is using ontology to analyze a document with the help of words and the relationships that exist between these words in a document. The conceptual semantic similarity focuses on only concepts rather than words and also the relationships between these concepts that exist which are stored in a base ontology. It is necessary to visualize and process the document using related concepts with a designed technique which is capable of capturing the intention of author while writing of the document.

Since, enormously large amount of information is available on the web, so there is also a requirement of technique which helps in organizing and utilizing the organized information by the different users of web. The technique which provides the means of organizing and utilizing the information should consider the fact that the information is presented mainly in natural language on web and the same is targeted to the reader/user for whom it is completely understandable as compared to machine. The information written in natural language is extracted by using the search engine as a tool, but the result-set produced is not up to the expectations of user as it contains many web pages which are not or of least interest of user.

As discussed, similarity can be of two types, one is detecting similar documents on the basis of attributes while other is detecting documents on the basis of relationships. In the second type of similarity computation which is considering relationships present between words in a document/text, we are focusing on understanding the meaning of a document or the information which is present in the document and the author of the documents wants to convey to the user. There are some issues which need to be considered while analyzing/addressing the relationships between words/concepts of document. First, issue is related to the number of relations that may exist between two words and identification of the actually present relationships in a document. Second, issue is dealing with the representation of relationships as they may be represented by different authors in one or the other manner giving same meaning. Last but not the final issue is related to the nature of the relationship and its variation according to time and requirement of new era. It is a well known fact that information related to domain is not constant so some of the relationships and the words are also dynamic in nature according to the requirement of outside world.

The architecture of semantic web is thus given in the form of layers by Tim Berner Lee, which is designed to consider concepts and relations between concepts from a document for its understanding and processing by the machine [2]. Using the NLP techniques and considering the issue of identifying the related concepts from a document, a conceptual semantic similarity technique has been introduced. In this technique, concepts and the relationships between the concepts are considered for a domain, in view of the idea that, each word can be replaced by a set of concepts by considering the inherent property of each word.

Although, there are many methods of finding the similarity between web documents by using available NLP techniques, Lexicography techniques, Ontology etc. While processing documents using these techniques the information is selected by analyzing documents syntactically and then the whole document is analyzed on the basis of the disambiguation of all the extracted words which have various meaning in different context. In comparison to these techniques, documents are analyzed and processed using semantic analysis along with the syntactic analysis of the document which helps in considering words synonyms, concepts representing a word, and to make it more efficient also the relationships that can exists between concepts. All these semantic analysis attributes can be represented by means of graph theory, relational algebra. In our approach of conceptual semantic similarity we have considered the representation as graph theory, means that we have represented the information stored in ontology, and document, in form of graph where each concept is represented by the nodes and relationships between these stored concepts are represented by edges of the graph. After processing of a document and representing of the same processed information of the document in form of graph the similarity between the constructed graphs can be easily computed by using graph comparison techniques [29]. There exist various ontology like Sweet, Gene, etc. and ontology building tools available like Protégé etc [21]. The basic parameters that are linked with ontology taxonomic hierarchy are length of shortest path, depth of most precise recurrent subsumer, density of the concepts from the root to the most exact recurrent subsume, density of concepts of the shortest path [59].

The most common approach for semantic similarity computation is Latent Semantic Analysis (LSA), Latent Relational Analysis (LRA) [38]. The LRA approach helps in computing the relational analysis between texts by extending VSM and applying SVD. Also, the extraction of concepts from the documents has been done by using heuristic rules for building the content which provides the relevant information of the document [31]. The ranking of documents can also be done by combining the approaches of keyword and semantic information [32]. [3] Considered the lexical patterns and numerous semantic relations that exist between the words for detection of similarity between semantic relations.

It has been seen that many different approaches have already been given for semantic similarity by considering related concepts. But, there is still a requirement of techniques which can be applied on the semantic web documents to provide more relevant results with less complexity. The conceptual semantic similarity is the technique which is designed to cover maximum related concepts of a domain for processing of the documents of that domain.

4.4 DETAILED CONCEPTUAL SEMANTIC SIMILARITY MODEL

To reflect on the numerous issues in the semantic similarity detection techniques this approach helps in understanding a document from the author intention or point of view of writing that particular document. According to our supposition any written document communicates the author vision or perspective about an activity/event which he/she wants to communicate to the reader. An activity is a series of interaction between various entities. An entity is a physical perceivable object or logical conceivable concepts. When an author writes a document, the entities are represented by concepts and interactions between these entities are represented by relationships. Therefore, in order to find intention of author completely we need to identify the various concepts and relationships between them from a document. These identified interrelated concepts of the activity or event which will be called as chain of concepts representing the intention of author regarding writing of a document is important to understand and hence it can also be easily used for relevant information retrieval task.

For understanding the concepts explained in above paragraph, it is necessary to understand that a document is a collection of words and relationship between these words which form the meaningful information. These collection of words are not the only way and also not sufficient to capture the purpose of the document written by an author. For, the deep analysis of document/text the concepts are considered which is set of ideas to represent a word. Each word can represent one or more concepts which demonstrate the probable idea behind that word. Therefore, first concepts are identified on the basis of given words. After this consideration of concepts, the relationships between these concepts are extracted from the document by using a base ontology. These extracted relationships between the concepts of the document constructs the multiple chains of related concept for the same document. The multiple chains extracted are connected with the relationships by using ontology that will further represent the document in the form of ontology which is called document ontology. This document ontology will convey the information and the idea behind that information written by the author which is definitely be understood by the document.

For the technical aspect of the given technique, two data bases are maintained which are, a base ontology and a dictionary. The dictionary is constructed and maintained with the help of all possible concepts available for a domain and also the words used in that domain documents for representing these respective concepts. On the other hand, the base ontology is constructed with the help of all concepts related to a domain analyzed while storing in the dictionary and the relationships between these concepts related to the same domain. Using these maintained databases, a document can now be processed. First, the document is processed using Stanford parser to extract the words and relationships between the words and the same is represented in the form of graph. Each extracted word is searched in the dictionary and its corresponding concepts are extracted to replace that word. This process will represent the document as bag of concepts. After obtaining the probable bag of concepts their relationships are extracted from the document and established by using the base ontology. In order to preserve consistency between both the databases i.e. dictionary and base ontology the literal string used to symbolize a concept is kept same in both the databases. The combination of concepts and relationship construct the document ontology for each document. Then, to find semantic information from any document, the document ontology constructed is analyzed for extraction of the longest chains of concepts as per the heuristic applied. According to the heuristic rule, the longest chains of concepts of a document represent prime/major intention of the author which he wants to convey to the reader. Finally, to find the semantic similarity between any two documents the longest chains extracted from each document are analyzed to extract the common longest chain between them. This common longest chain extracted from both the documents will give maximum interrelated concepts present in both the documents. The conceptual semantic similarity approach is given in Algorithm 4.1 and Algorithm 4.2. Algorithm 4.1 gives the process of the construction of document ontology. The computation of semantic similarity score between the two given documents is given in algorithm 4.2.

```
Algorithm 4.1:
Input: Set of Documents $D_s$, Base Ontology O, Dictionary $D_{CW}$.
Output: Document Ontology $D_O$.
    1. Select a document D from $D_s$ for which $D_O$ is to be constructed.
    2. For the document D
        i.   Extract words from D to construct BOW.    //BOW is vector representation of Bag of words.
        ii.  Construct BOC by replacing each extracted word by respective concepts present in $D_{CW}$.    //BOC is vector representation of Bag of concepts.
        iii. Construct set of chains connecting the concepts obtained using O.
        iv.  Obtain Document Ontology $D_O$ for D from the set of chains obtained.
```

```
Algorithm 4.2:
Input: Set of Documents $D_s$
Output: Semantic similarity score Sc between two given documents.
    1. Select two documents $D_1$ and $D_2$ from $D_s$.
    2. For each $D_1$, $D_2$
        i.   Construct document ontology $D_{1o}$, $D_{2o}$ for $D_1$, $D_2$ respectively using algorithm 4.3.
        ii.  Select longest chain $D_{C1}$, $D_{C2}$ from $D_{1O}$ and $D_{2O}$ respectively.
        iii. Select common longest chain $C_l$ from both $D_{C1}$, $D_{C2}$.
        iv.  Compute semantic similarity score Sc using
                 $Sc = Nr + Nc/(1 + Nrm + Nrc)$
```

Where N_r, N_c are number of relation and concepts present in matched C_l.

N_{rm}, N_{cm} are number of relations and concepts present in mismatch part.

4.4.1 Conceptual Semantic Similarity Implementation and Explanation with Example

In this sub-section, the proposed approach is given by using example. The constructed base ontology is represented as G(C, R).

Where, C is the set of concepts {c_1, c_2, c_3,……....c_n} existing for the particular domain .R is set of edges in the graph representing the relationships between two concepts from C.

The relationship R_{ij} represents the relationships that exist between the concepts c_i and c_j. The sample part of graph for base ontology present in knowledge base is shown in Figure 4.5. From empirical point of view the base ontology is constructed in Excel sheet which details are given in Appendix II Table 2.2.

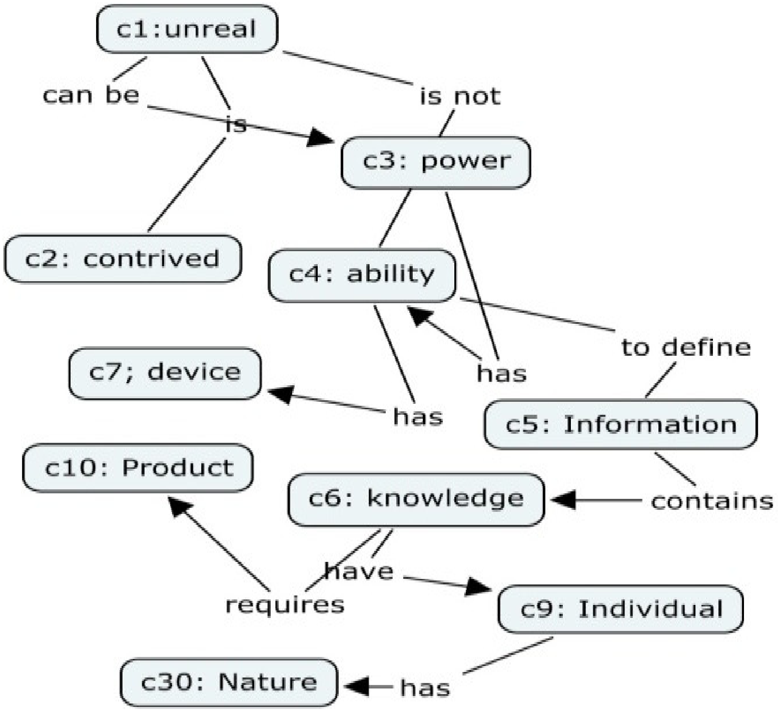


Figure 4.5: Sample Graph for Base Ontology

The above sample part of base ontology has been constructed by using the concepts that are actually present in the dictionary. The dictionary stored and maintained is having the words related to a domain along with all the representing concepts that can exist in a domain corresponding to each of these words. The sample component of the domain dictionary is shown in Table 4.2 and the detailed domain dictionary is given in Appendix II Table 2.1.

Now, there is an assumption that each word in a domain is linked to the interrelated concepts. The documents exceptionally include a word that is not related to any other word or concept. Firstly, the approach is applied on small text of D_1 and D_2 which is as follows:

D_1: Artificial intelligence is intelligence of machine and robot and branch of computer science that aims to create it.

D_2: Artificial Intelligence is branch of computer science concerned with making computers behave like humans.

Table 4.2: Dictionary having Words and Related Concepts

Words	**Related Concepts**
w_1: artificial	c_1: unreal, c_2: contrived
w_2: intelligence	c_3: power, c_4: ability, c_5: information, c_6:knowledge
w_3: machine	c_7: device, c_{11}: mechanism
w_{11}: human	c_9: individual
w_{10}: behave	c_{30}: nature, c_4: ability, c_{31}: living way

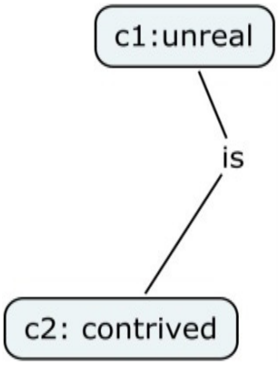
To construct the document ontology for D1 the words actually present in D1 are extracted. The set of extracted words from D1 is represented as Bag of Words (BOW). This BOW is further represented by using vector having the set of elements as given below:

BOW= ({w_1: Artificial, w_2: Intelligence, w_3: Machine, w_4: Robot, w_5: Branch, w_6: Computer, w_7: Science, w_8: Aim, w_9: Create}).

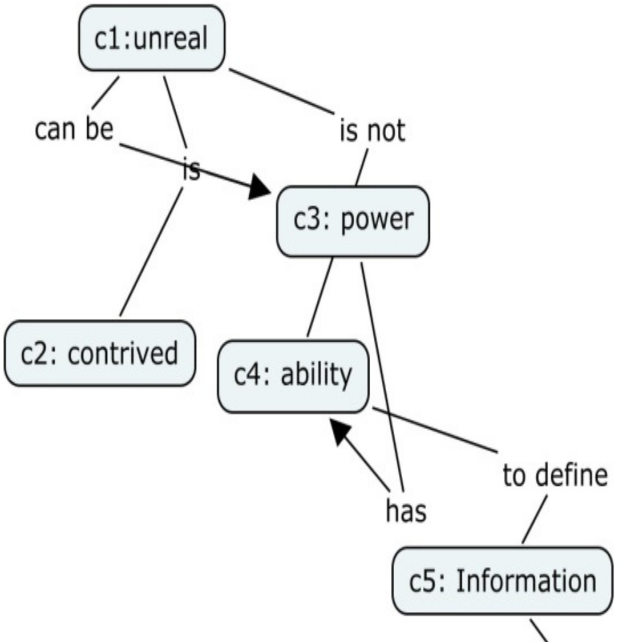
Next, these words are replaced by relative concepts by using the dictionary database to symbolize the view of author and user exclusively. So, the Bag of Concepts (BOC) is symbolized as the vector having set of the elements as follows:

BOC= ({c_1, c_2}, {c_3, c_4, c_5, c_6}, {c_7, c_8, c_9, c_{10}, c_{11}}, {c_7, c_{11}}, {c_{12}, c_{13}, c_{15}, c_{14}, c_{16}}, {c_7, c_{18}, c_{19}, c_{20}}, {c_{13}}, {c_{14}, c_{15}, c_{12}}, {c_{21}, c_{22}, c_{23}, c_{25}, c_{26}}, {c_{27}, c_{28}, c_{29}}).

In the next pace the concepts attained for the document is interrelated by using the base ontology O for which sample content is shown in Figure 4.5. In the process of document ontology construction the concepts c1, c2 retrieved for the word w1 and their relationship according to the original document and the base ontology is represented as follows:



Next the word w2 representing the concepts c3, c4, c5, c6 is interconnected as and the set of chains obtained after the interconnection of all the concepts obtained till this step is represented as follows:



In the same way, other concepts are also obtained along with the relationships extending the set of chains of related concepts to form the complete document ontology for D1 which is shown in Figure 4.6. There is also an assumption that hardly ever the processing of document will start forming the chain which is not considering the intention of the author. Such chain will definitely be no longer continued if it is not associated to the conceptualization of the individual.

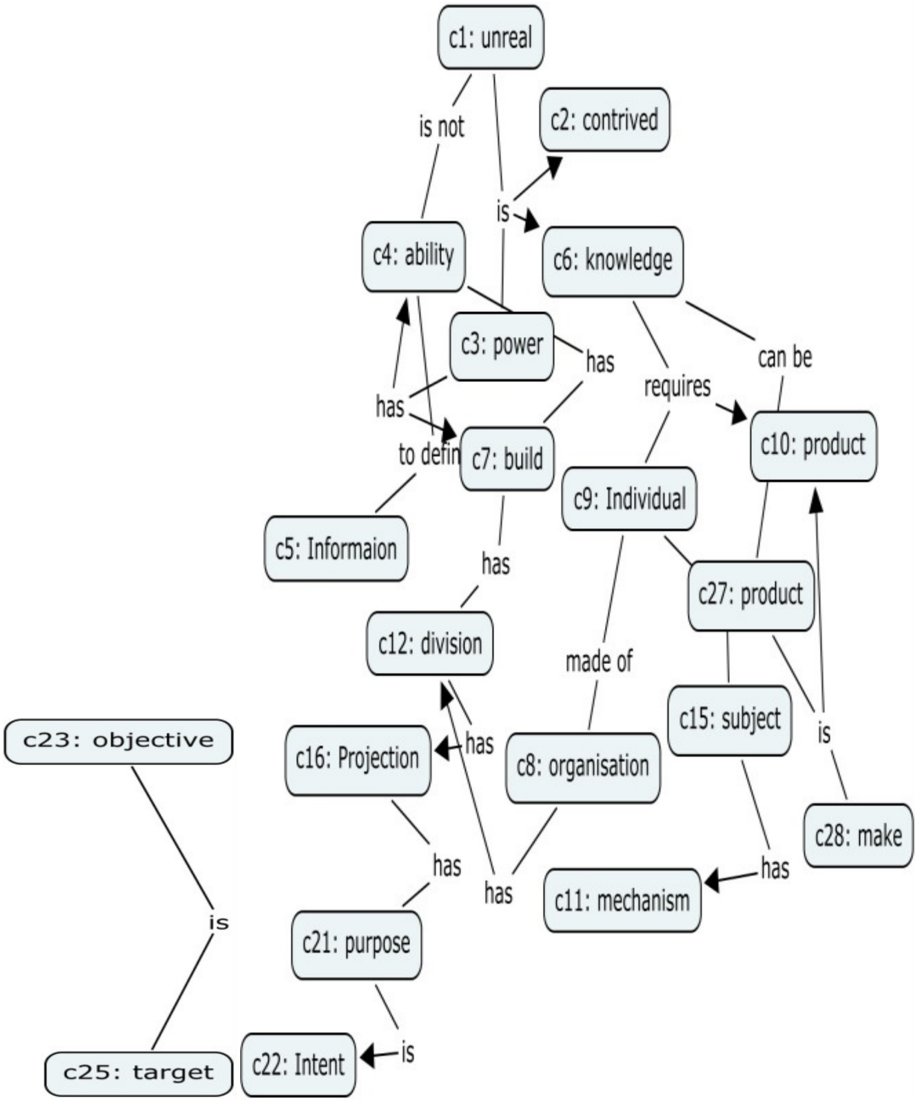


Figure 4.6: Document Ontology for D1

Following the same procedure we can construct the ontology of the document D2 shown in Figure 4.7. In the above document ontology construction process we have finally got two ontology for D1 as shown in Figure 4.6, as while connecting the concepts and forming the chains of the document c_{23}, c_{25} are the two concepts obatined while replacement of words by respective concepts are not connected to any other concepts retrieved for the document. This is due to the fact, that a document ontology may have chains representing the idea of the author. To analyze the document with efficiency the heuristic rule according to which the longest chain reprents the prime/major intention of the author of the document is applied. So, the extraction of the longest chains among all the set of chains retained in both the document ontology is done.

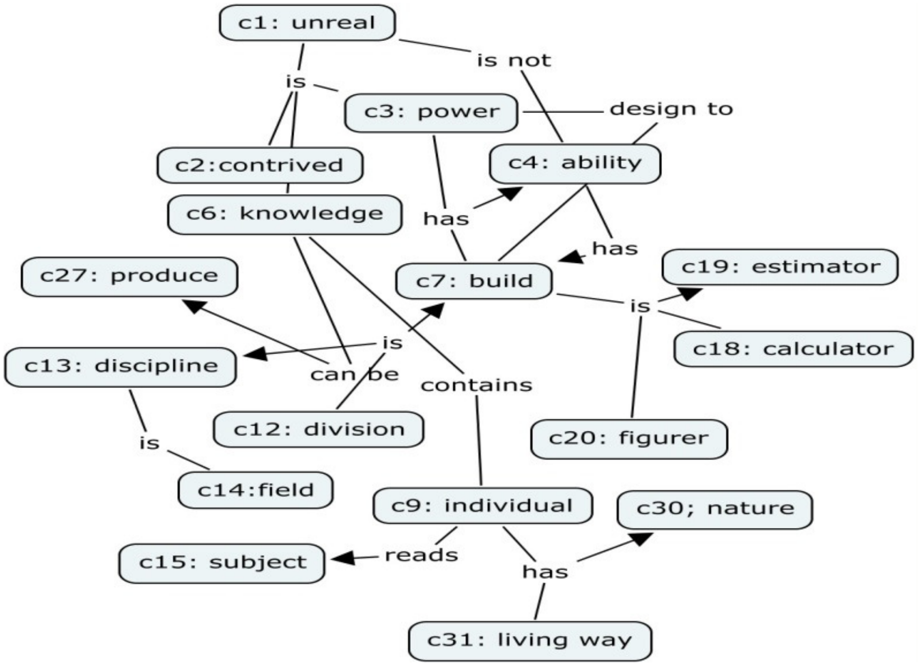


Figure 4.7: Document Ontology for D2

Finally, the longest chains found in both the documents are compared to find the common longest chain present in both the documents for computation of the semantic score between documents. From the above two document ontology constructed for D1 and D2 the common longest chain is extracted and the common sub-graph is obtained from both the documents is shown in Figure 4.8.

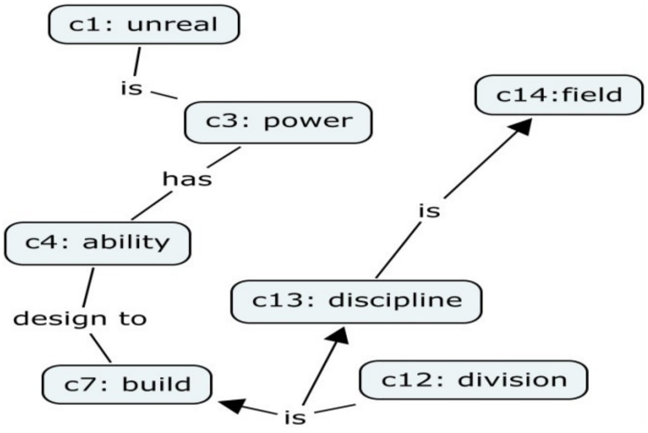


Figure 4.8: Common Longest Chain obtained from Document Ontology D1 and D2

The common longest chain obtained as shown in Figure 4.8 represents the maximum interrelated concepts available in both the documents considered to calculate the semantic similarity between them. This common longest chain will help in conveying the prime intention of the author for writing a document to the reader of the documents. Now, if the similarity between three given documents D1, D2 and D3 is to be calculated then the longest chains present in all the set of documents (D1, D2), (D1, D3), (D2, D3) are extracted and again the common longest chain is obtained to represents the semantic similarity score between any given sets of documents. This gives the idea that the longer the length of the common longest chain the more is the semantic score between the documents. To attain the efficiency of the given approach, the set of 50 documents given in Appendix I Table 1.1 related to the artificial intelligence domain is also analyzed and processed. From empirical point of view, the document content is parsed using the Stanford Parser. The tree of each document is constructed by using the library Stanford-parser.jar and lexicalized parser class as is also discussed in Chapter 3. The document graph is constructed for each tree obtained and the same is displayed by using the jgraphx.jar library. Next, the words extracted from each document are replaced by the set of respective concepts already stored in the dictionary which is stored in Excel sheet. The data stored in the dictionary is extracted by using the Apache POI library. After the replacement process, the relationships are extracted by using the base ontology which is also stored in Excel sheet and the same Apache POI library is used to extract the data from base ontology. The document ontology is constructed for each document and it is also displayed in the form of a graph having concepts as nodes and relationships between concepts as edges. The analysis and processing of artificial intelligence related documents by the given approach provides more meaningful information of relatedness between any two documents which is closer to human analysis. The same approach is also applied on the set of documents related to domain mobile given in Appendix I Table 1.2. The details of results and outcomes are given with the next proposed technique which indeed an extension of the current given techniques.

4.5 SEMANTIC SIMILARITY COMPUTATION BY EXTENDING DOCUMENT ONTOLOGY

Concretely, the semantic similarity computation has been done by constructing document ontology in above explained technique by using knowledge base ontology.

The conceptual semantic similarity approach of semantic computation between the documents have already improved the searching and matching of the relevant information from the set of documents. But, there exist documents which need further processing to extract the relevant semantics by adding the implicit information that are not actually present in the document content. Thus, there is a need to design a technique which considers maximum relevant information from the documents by extracting the implied semantic information for the content of the document and is not covered by the traditional approaches or the approaches given above and in Chapter 3.

In view of the above requirement, a technique of extending the constructed document ontology is given which helps in considering implicit information to provide the user the maximum relevant information as per the recent trends of that same information. The basic idea is to cover all the content and provide the information as per the demand of the user of the document and outside world technology. Using this approach, the document ontology constructed is extended by using the recent trends available for the domain to which the set of documents which are processed belongs. Then, the extended ontologies of each document are compared to give user the benefit of the proposed approach of extension. This technique gives three major contributions to the field of semantic analysis which are as follows:

1. It provides a way of constructing a document ontology which helps in representing the main idea of that document content.
2. The technique provides a way of extending the constructed document ontology by using the hidden/implicit related concepts stored in a separate trend database.
3. Finally, it gives the way of finding the semantic similarity depending on the depth of extension of the document ontology.

The ontology construction and its efficient use is prime requirement for processing the semantic web document as the semantic web is a layered architecture considering the content/information of a document in form of related concepts. Ling S. et. al. [59] has given the semantic similarity computation technique based on the fuzzy logic. Fuzzy logic is an approach to computing based on "degrees of truth" rather than the usual "true or false" (1 or 0) Boolean logic on which the modern computer is based.

The fuzzy based approach helps in computing the similarity by considering the information theory field. The information theory considers the structure of ontology in form of hierarchical and non-hierarchical organization of information. This approach is basically for content-based information retrieval system and intelligent environment of question-answering. The content based information retrieval is computing the similarity by defined as extended semantic fuzzy set for each concept along with the paths of semantic information.

Shixiong et. al. [60] has given the importance of four major factors that has high impact on the semantic analysis of text. The four main factors are related distance, related coincidence, related depth and the related density. Similarly, automatic method of finding concepts with the consideration of hypernym relationships using a given ontology like WordNet has also been done by Aditi et. al. [61]. The algorithm in [61] also helped in clustering of documents on the basis of concepts.

Madalina et. al. [64] focuses on the conceptual graph for computation of dissimilarity score. The dissimilarity score is basically based on the number of cliques of the analogous graph and a configuration encoding the two graphs projection in sequence which will support in establishing knowledge relationships. This will further help in developing and designing reasoning oriented techniques giving more accurate results for semantic similarity. Another technique proposed in the same area again by considering a domain specific dictionary for mapping of ontology and integration of system depending on the query service of ontology [63]. In this method, the rules of mapping for generating the query were applied to uncover the concepts not identified by the traditional dictionary WordNet. Likewise, the Wikipedia as ontology in combination with the spreading activation technique has been widely utilized for related information retrieval [62]. The technique extracts each keyword categories of a query based on the title of document from the database where all the documents which need to be processed are stored. Lastly, all the categories extracted in this manner are represented as the category tree nodes of Wikipedia for application of the spreading activation technique.

In summary of all these techniques, it can be said that all these methods are either based on Fuzzy based Reasoning, NLP, Conceptual Graph, Spreading methods etc for relatedness detection of texts. In similar manner, a prototype model [65] given by using ontology for extending the original words by using ontology.

But, after so many efforts there is a requirement of enhancing the techniques so that the processing technique is capable of analysing and understanding the text from the author point of view which considers all relevant and important concept of the written text, so that they can be applied for efficient information retrieval.

4.6 FORMAL MODEL FOR EXTENDED DOCUMENT ONTOLOGY

Formally, the techniques given above extract the available words from a document and then consider the set of probable concept representing each word from that document. Then the concept relationships are extracted from the document/text by using a base ontology already constructed for a particular domain or topic. In general, as it has already seen that words alone cannot be sufficient to provide the idea behind that word, so there is a need to consider the concepts to understand the conceptual view of a particular text in some context or the other context. There is also an assumption, that each concept which is considered for processing of document related to one or the other concept, as there rarely may exist any concept which does not connect to the another concept. Even if such concept exists then it is considered as the one which does not provide the relevant information about the text, so it may be ignored during the natural analysis of document/text. During the deep analysis or understanding of the documents it is also noted, sometimes the author conveys his/her idea by using the set of words and concepts which are not considered by using the traditional dictionary like WordNet. Basically, the idea behind this is that these concepts or words are related to the recent trends of our modern language. For example, in the text of android mobile phones are better than windows, there is implicit information that occurs in reality but not appearing in words of text is that Samsung mobile phones are better than Nokia. As in this context android is an operating system by Samsung and Windows is operating system by Nokia. Thus, according to our hypothesis and the deep understanding of semantic analysis, it has been discovered that all existing concepts are related to the current trends of the web information needs to be considered for deep semantic analysis of the document.

In our approach of construction of extended document ontology, firstly the concept relationships chains are extracted from a document by making the use of the base ontology already stored. These extracted chains of related concepts are further extended by using the recent trends related to the domain. These recent trends which occur for a domain according to requirement of knowledge of the dynamic world are stored in a database which is maintained separately from the domain specific dictionary. From the procedural point of view of the scheme, the base ontology is constructed and implemented by using two data structures similar to the above approach of conceptual semantic similarity. One is a dictionary, having the related words from a domain accumulating along with the respective set of concepts available from traditional dictionary like WordNet for each domain. Second is the Ontology Graph, which is the graphical representation of the concepts identified while storing in the dictionary along with the set of probable relationships that exists between identified concepts. One more database is also constructed and maintained in which all the recent trends related knowledge and understanding of the corresponding related concepts regarding a particular domain.

The extended document ontology construction scheme also works in the same manner of extraction of the words from a document by using NLP technique which is the prime requirement of the techniques given in this research work as discussed in Chapter 3. Then, the replacement of these extracted words with the concepts is done by using the data structure named as dictionary. This is done, so that these concepts can be connected with the relationships which can be obtained by already maintained data structure named as ontology graph. After, connecting the relationships between the concepts, this will give the graphical representation of a document which is termed as document ontology. This document ontology will have concepts as nodes and the relationship between these concepts as edges. Now, this document ontology is further extended by using the already maintained database having recent trend related concepts. The purpose of extension of the constructed document ontology is to enhance the conceptual view of the document to the next level of understanding i.e. deep/hidden analysis. This deep/hidden analysis allows us to consider implicit knowledge that is already embedded in the content of the document but is not actually presented by the set of related words. Lastly, the constructed and maintained extended document ontologies are compared with each other by considering the sets of the longest chains available in the extended document ontology. The Algorithm 4.3 and Algorithm 4.4 shows the process of construction of the document ontology extended with the current trends of the domain and computation of similarity between the given set of documents by considering the longest chains of related concepts.

Algorithm 4.3

```
Input: Set of Documents Ds, Base Ontology O, Dictionary DCW, Database DB.

Output: Document Ontology Do

    1. Select a document D from Ds for which DO is to be constructed.
    2. For the document D
        i.  Extract words from D, and replace each word by respective concepts
            present in DCW.
        ii. Construct document ontology DO for D by connecting the concepts
            obtained using O.
    3. Extend the DO by using DB.
```

Algorithm 4.4

```
Input: Set of Document Ontology SDo.

Output: Semantic similarity measure SSm between given two Extended Document
Ontology Doi and Doj.

    1. For each pair of documents ontology Doi and Doj
    2. Find the longest chains of related concepts available in Doi and Doj.
    3. Extract the common longest chains available in the longest chains extracted
       above from both the extended document ontology.
    4. Compute the semantic similarity between documents using common longest
       chains $C_l$ obtained in step 2 using formula given below:
                 $Sci = Nr + Nc/(1 + Nrm + Nrc)$

Where, Sci is semantic similarity between pair of chains and i= {1, 2….n}
representing each pair,

Nr, NC is number of relation and concepts present in matched Cl.

Nrm, Ncm are number of relations and concepts present in mismatch part.

    5. Compute semantic similarity score SSc between extended Document
       Ontology's using
SSm (Doi, Doj) =max {Sci, where i=1, 2…….n}.
```

4.7 IMPLEMENTATION AND DESCRIPTION USING EXAMPLES

In this section, the scheme of extended document ontology is demonstrated with the help of examples. In our example the two documents named D_1 and D_2 are considered which is having content related to the mobile phones of Samsung and Nokia. The part of the content from D_1 and D_2 is as follows:

D_1: Samsung and nokia are organizations and manufacturer of mobile phones. In addition to mobile phones and related devices, the company also manufacturers things such as televisions, cameras, and electronic components. Samsung mobiles phones are better than nokia based mobile phones.

D_2: Mobile phones are manufactured by different organizations have operating system like android or windows. Android based mobile phones are better than windows based mobile phones.

To start with implementing the scheme first these two documents i.e. D_1 and D_2 are processed for extraction of words from them by making the use of Stanford Parser. With the use of Stanford Parser, the words or phrases which are relevant for each sentence of the documents like NN, VP, ADJP, and NP etc. are retrieved from the tree constructed by using the Stanford-parse.jar and lexicalized parser class to parse the document. These retrieved words are replaced by using the domain dictionary, part of which is shown in Table 4.3. This domain dictionary is having words and related concepts is stored in Excel sheet and the complete dictionary is given in Appendix II Table 2.4

Table 4.3: Domain Dictionary

Words	Related Concepts
w_1: electronic components	c_1: electronic element, c_2: electronic ingredient, c_3: electronic constituent
w_2: devices	c_8: instrument, c_9: machine
w_3: manufacturer	c_{22}: maker, c_{23}: producer
w_9: model	c_9: simulation, c_{10}: framework
w_{10}: source	c_{30}: origin, c_{31}: informant, c_{32}: root

As discussed above, the set of words for D1 and D2 obtained by using the Stanford parser is shown with the help of the unique identification number assigned while construction of dictionary as follows:

D1 = {w_{11}, w_{12}, w_{13}, w_{10}, w_9, w_3, w_2 …….}

D2= {w_{10}, w_3, w_{13}, w_8, w_6………}

Next, is the step of replacement of these words again represented with the help of corresponding concept present in the constructed dictionary. So, the updated document set is the Bag of Concepts as shown below:

D_1={C_1,C_2,C_4,C_6,C_{11},C_{12},C_{17},C_{18},C_{22},C_{23},…..}

D_2= { C_3,C_4,C_8,C_{22},C_{23},C_5,C_{16},C_6,C_{31},C_{32},……..}

Next, these obtained concepts are then used to construct the document ontology graph by establishing the relationships between these concepts by making the use of the data structure i.e. ontology graph. The sample part of the constructed and maintained base ontology is shown in the Figure 4.9 which is made with the help of CMap tool. The constructed document ontology for document D1 and D2 in first step is shown in Figure 4.10 and Figure 4.11 respectively.

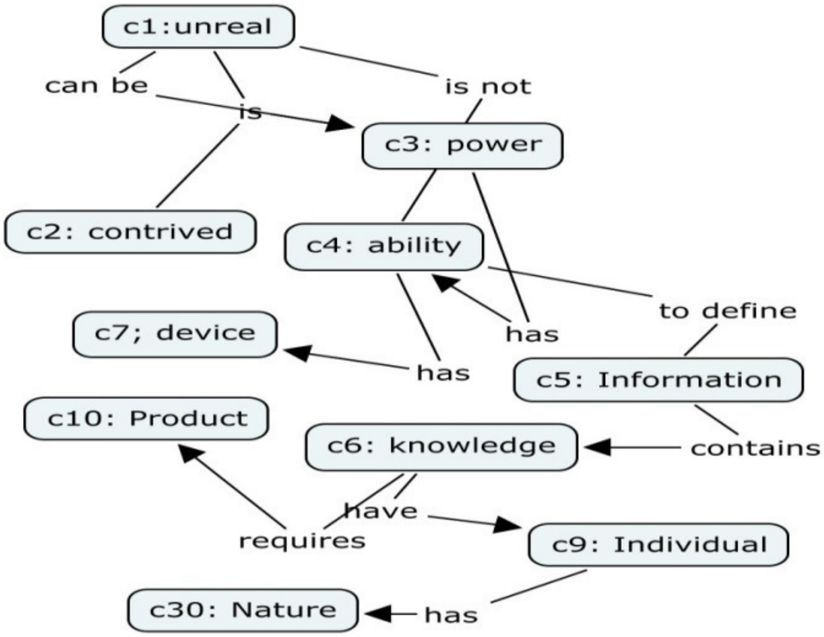


Figure 4.9: Base Ontology

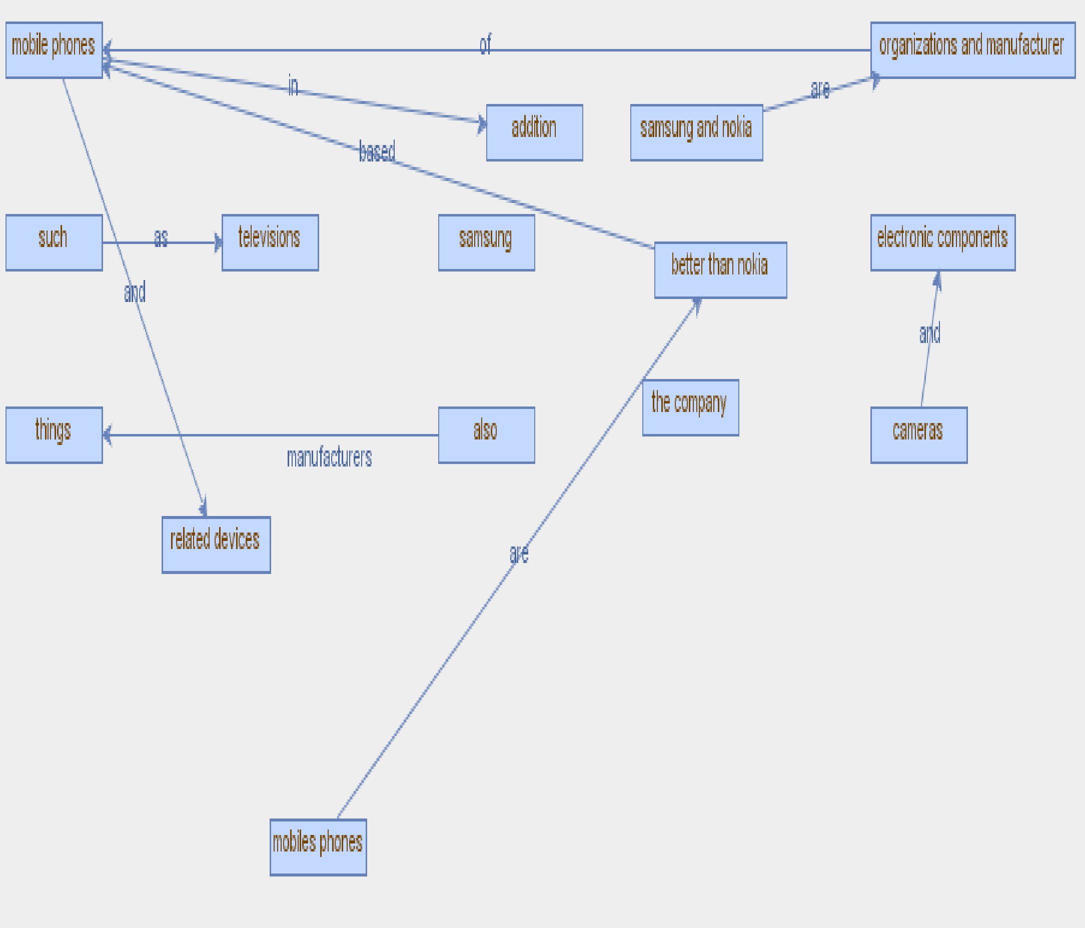


Figure 4.10: Document Ontology for D1

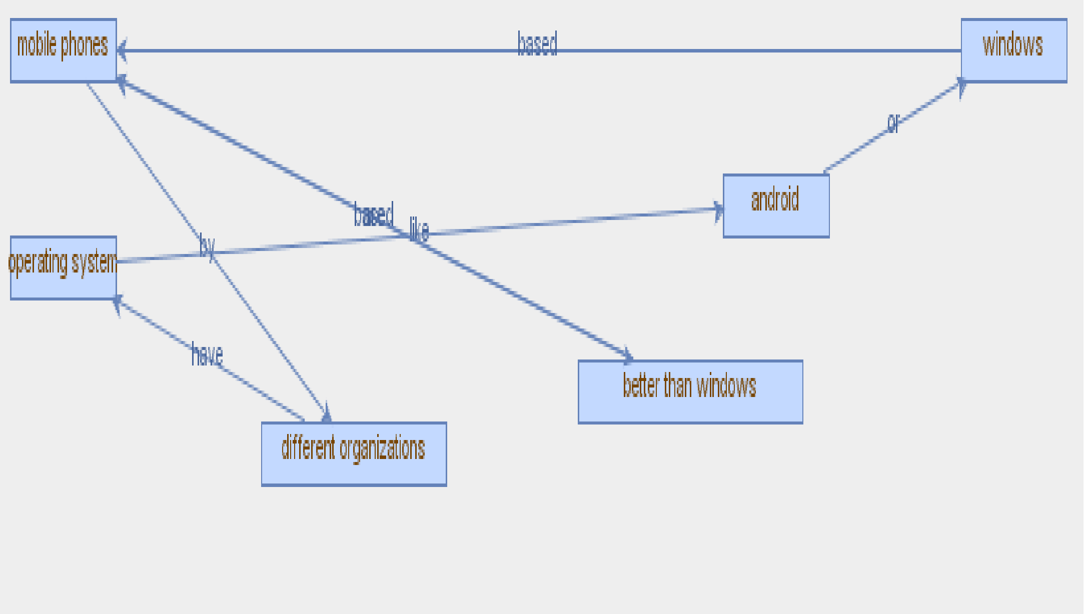


Figure 4.11: Document Ontology for D2

Now, the constructed document ontology for D1 and D2 are extended by using the maintained database of current trends concepts and their relationships which is shown in Table 4.4. The recent trend database has been kept separately from base ontology to retain the data independence property. So, the database can be updated easily without changing the base ontology as per the requirement of changing world like technology enhancement, some product become obsolete; some product may require development of one or the other product etc. For extending the already constructed document ontology, the concept pairs between which trend relationships exists are identified and thus these trend relationships are added to the document ontology. This will help in embedding the semantic similarity by giving the conceptual view of the document with respect to the implicit information about the domain to which the documents belong. The extended document ontology for document D1 and document D2 is shown in Figure 4.12 and 4.13 respectively.

Table 4.4: Trend Related Concepts

Concept	**Relationship**	**Concept**
samsung	technology	android
nokia	technology	windows
android	better	windows
android	technology	more
windows	based	less
android	based	mobile phones
windows	based	mobile phones
samsung	product	mobile phones
nokia	product	mobile phones
samsung mobile phones	has os	android
nokia mobile phones	has os	windows
android	usage	more free applications
nokia	usage	more paid applications
samsung	usage	more free applications

Concept	**Relationship**	**Concept**
samsung mobile os	source code	open source
nokia mobile os	source code	closed
samsung mobile os	latest version	lollipop
nokia mobile os	latest version	update
samsung mobile os	address	android.com
nokia mobile os	address	windowsphone.com
samsung mobile contacts	backup	gmail.com
nokia mobile contacts	backup	hotmail.com
android	owned by	google
windows	owned by	microsoft
samsung	origin	south korea
nokia	origin	finland
android mobile	address	samsung.com
windows mobile	address	nokia.com
android	user friendly	phones
linux	source code	open source
linux	is	open source

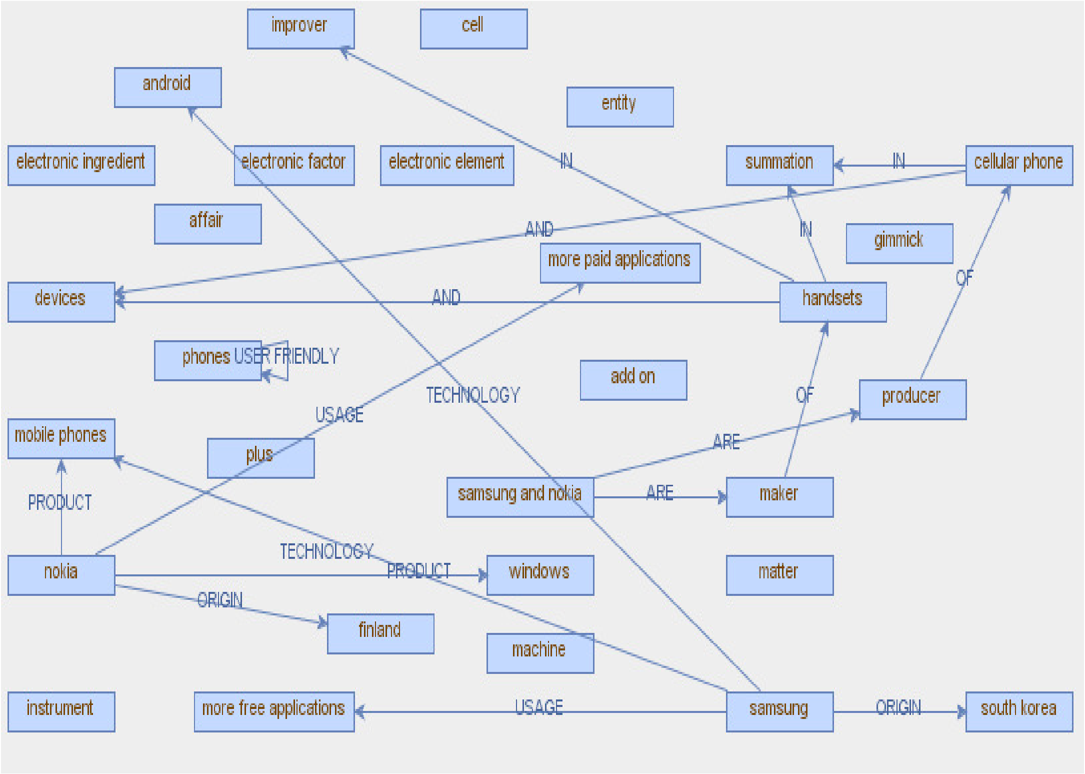


Figure 4.12: Extended Document Ontology of D1

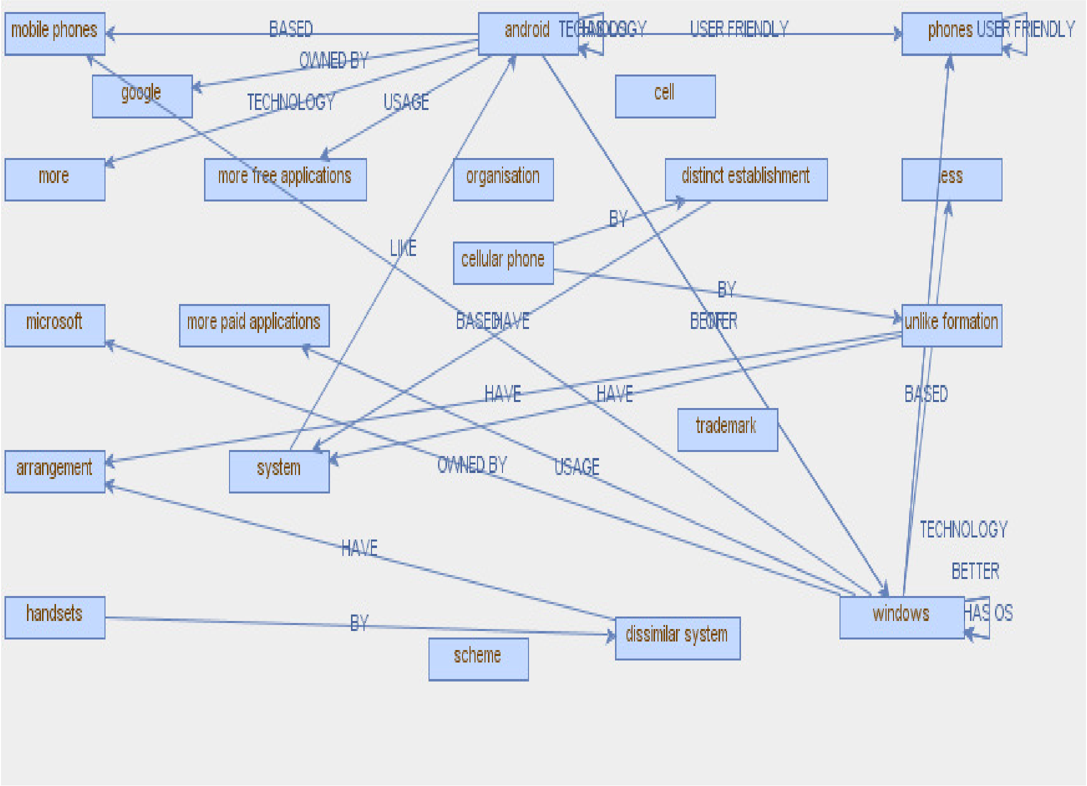


Figure 4.13: Extended Document Ontology of D2

Lastly, as per the scheme given above the constructed extended document ontology's are compared by extracting the longest chains from these extended document ontologies'. Then, by using the Algorithm 4.4 the common longest chains are obtained from the extracted longest chains to give the prime/major intention of the documents and thus computing the semantic similarity between them.

4.8 ANALYSIS OF ONTOLOGY BASED APPROACHES

In this section, the performance analysis of the ontology based approaches which are considering concepts, chains of concept relationship, and extended chains of related concepts is described. The above explained ontology based approaches definitely depends on many factors like extraction of relevant words from document, replacement of words by set of relevant concepts, relationships between concepts obtained, consideration of relevant recent trend related concepts. In the first approach, using ontology the concepts and the relationships are considered from a document to understand the information given in the document. Then, the process of spreading was applied to these concepts depending upon domain to domain along with the establishment of the relationships between the obtained concepts. For analysis of the given technique the set of documents were considered which were first analyzed by implementing the lexical matching approach for similarity, then the same set of documents were processed by considering related concepts. It was empirically found that the similarity between the given set of documents was improved to some extent. As, in ontology based approach the semantic information is considered with the help of concepts and relationships between these concepts from each document in the set. The results for this approach of related concepts are given in Table 4.5.

From the table it can be inferred that some sentences taken from document contains the concepts of product and their brands. Although, all these sentences give similarity on the basis of the product and brands but they include some semantic information like the types of products available in market. Even when these documents were analyzed by the human being it was noted that by understanding the related words of a document the author also want to convey more information which leads to the thinking of the comparisons between these concepts. But, at this level of consideration of concepts and relationships the comparison analysis is not captured. This is considered while developing and designing of the extended ontology approach whereas, the other ontology based approach help us to view a document from the other side i.e. semantic/meaningful information. The results shown in Table 4.5 give the betterment of the ontology approach by giving more similarity between the set of documents as compared with the lexical matching.

Table 4.5: Comparisons between Lexical Approach and Ontology Approach

SNO	Document A	Document B	Keyword Similarity	Novel semantic Similarity
1.	https://en.wikipedia.org/wiki/Diesel_engine	https://en.wikipedia.org/wiki/Petrol_engine	.04	0.1
2.	http://www.ibef.org/download/Samsung.pdf	http://www.ibscdc.org/Free%20Cases/BOS0010A.pdf	0.2	0.5
3.	http://macktribble44.tripod.com/id2.html	http://tsl.news/opinions/3579/	.01	0.3
4.	https://en.wikipedia.org/wiki/Android_(operating_system)	https://en.wikipedia.org/wiki/Windows_Phone	0.39	.5
5.	http://www.samsung.com/uk/discover/blu-ray-101	http://www.winxdvd.com/resource/dvd.htm	.012	0.5

The ontology based approach considered concepts and relationships between these concepts but they were considered as an independent entity set of document. But, still there is a requirement of considering these related concepts in continuous connecting chain form. It means that a document is not the set of related concepts but it is the set of related chains of concepts. The formation of such chains is done by making the use of the proposed conceptual similarity approach.

The conceptual semantic similarity helps us to view a document as an ontology graph from which the chains of related concepts can be easily extracted. Due to the complexity of implementation of the conceptual semantic similarity technique the longest chain of concepts relationships are considered which according to our assumption provides the prime intention of the author of the document.

The empirical evaluation of the conceptual semantic similarity approach is given in Table 4.6 with the VSM approach, LRA by Turney [48] also considering relationships.

Table 4.6: Analysis between VSM, LRA and Conceptual Semantic Similarity

Set of Documents	**Semantic Score using VSM**	**Semantic Score using LRA**	**Proposed Conceptual Semantic score**
D_1, D_7	.44	.54	.61
D_7, D_8	.34	.6	.6
D_3, D_5	.1	.16	.2
D_7, D_9	.12	.19	.23
D_5, D_8	.38	.44	.55

Note: The sample parts of documents are as follows:

D_1: Artificial intelligence is the area of computer science focusing on creating machine that can engage on behavior that human consider intelligent.

D_2: Artificial intelligence track focuses on fundamental mechanism that enable the construction of intelligent system that can operate autonomously, learn from experience, plan their actions and solve complex problems.

D_3: Knowledge representation and knowledge engineering are central to artificial intelligence research. Many of the problems machines are expected to solve will require extensive knowledge about the world.

D_4: Intelligent agent must be able to set goal and achieve them. They need a way to visualize future and be able to make choices that maximizes the utility of available courses.

D_5: Machine learning is central to artificial intelligence research. It is study of computer algorithm that improves automatically through experience.

D_6: Natural Language processing gives machine the ability to read and understand the languages that human speak.

D_7: Intelligence is ability to think to imagine, to create, memorize, understand, recognize pattern, make choice, adapt to changes and learn from experience.

D_8: Artificial intelligence textbook define the field as study and design of intelligent agent where an intelligent agent is system that perceives its environment and takes action that maximizes its chance of success.

D_9: Artificial intelligence includes game playing, expert system, natural language, neural network, and robotics. Currently no computer exhibit full artificial intelligence.

D_{10}: Applications of artificial intelligence robots that plan their own actions, web crawlers that efficiently locate information, intelligent assistant that help humans defect financial fraud and game playing system that perform better than human player.

D_{11}: Artificial branches include logical artificial intelligence, search, pattern recognition, representation, inference, common sense knowledge and reasoning, learning, planning, ontology, heuristic and genetic programming.

Despite of the complexity of the processing of the conceptual semantic similarity approach the results obtained are quiet close to the analysis performed by understanding of human being. One of the applications i.e. ranking of the set of documents between which semantic similarity is calculated by using the conceptual semantic similarity approach is also evaluated. The results of ranking of the given set of documents are shown in Table 4.7.

Table 4.7: Ranking of Documents by using LRA and Conceptual Similarity

S No	Actual Rank	Semantic Rank	LRA Rank	Variance by LRA Rank	Variance by Semantic Rank
1	D_1, D_3, D_5	D_3, D_5, D_1	D_1, D_5, D_3	10	3
2	D_1, D_2, D_7, D_8	D_1, D_7, D_2, D_8	D_2, D_1, D_7, D_8	14	9
3	D_3 D_4, D_5	D_3, D_5, D_4	D_4, D_3, D_5	8	5
4	D_1, D_6, D_7, D_8	D_6, D_1, D_7, D_8	D_1, D_6, D_7, D_8	12	11
5	D_1, D_5, D_9	D_5, D_1, D_9	D_5, D_9, D_1	38	24

The results presented in Table 4.6 and Table 4.7 gives the idea and confidence about considering the related concepts, then the formation of chains of related concept to construct the document ontology for semantic analysis of the document. The Figure 4.14 gives the analysis of lexical matching using cosine similarity with the proposed semantic conceptual matching showing the enhancement gained in the similarity score by the proposed approach. The graph shows the semantic similarity computed by the proposed semantic conceptual matching technique between the texts of documents given above.

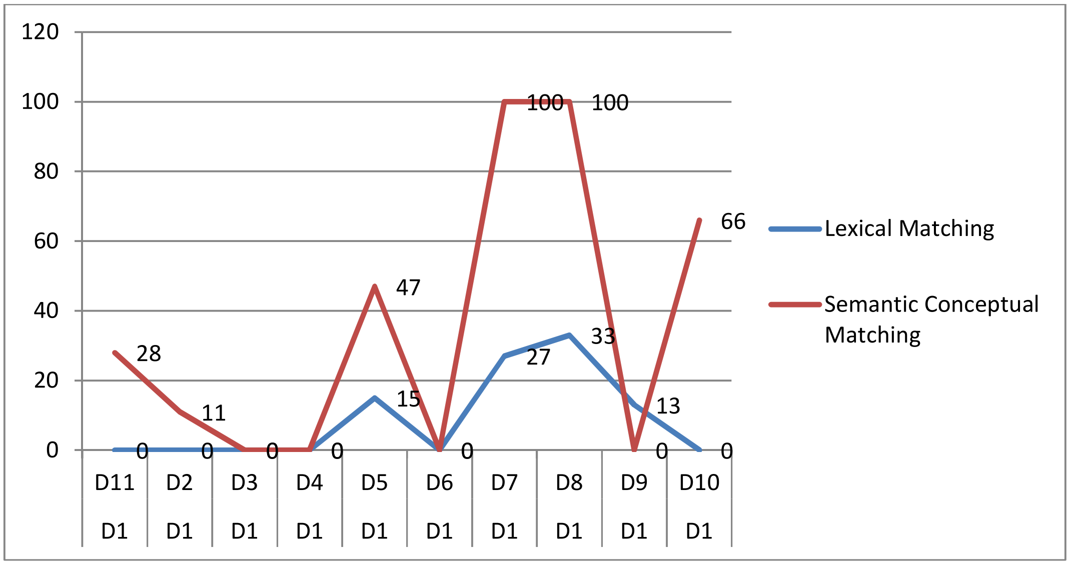


Figure 4.14: Analysis of Lexical Matching and Proposed Conceptual Matching

As discussed above, the word semantic is not only related to the meaningful information in terms of Natural Language Processing, but in the field of relevant information retrieval it should be taken one more step ahead. The idea is basically considering the implicit hidden information along with the semantic information which is also one of the major requirements of the user of WWW while searching the relevant information from web for a particular query given by them to a search engine. This major and implicit requirement of the user of web is taken into consideration by designing a technique which takes the technique of conceptual semantic similarity to this level. The chains of related concepts which are obtained by using conceptual semantic similarity are used to construct the document ontology as defined. This document ontology is then extended with the current trends concepts and their relationships that may exist regarding a particular domain. The consideration of the trend related concepts helps in giving maximum relevant and semantic information that persist in a document and also which the author of the document wants to convey to the user. Figure 4.15 gives the analysis of lexical matching using cosine similarity with the proposed extended document ontology conceptual matching between the same set of documents as mentioned in Table 4.6 showing the enhancement gained in the similarity score by the proposed approach.

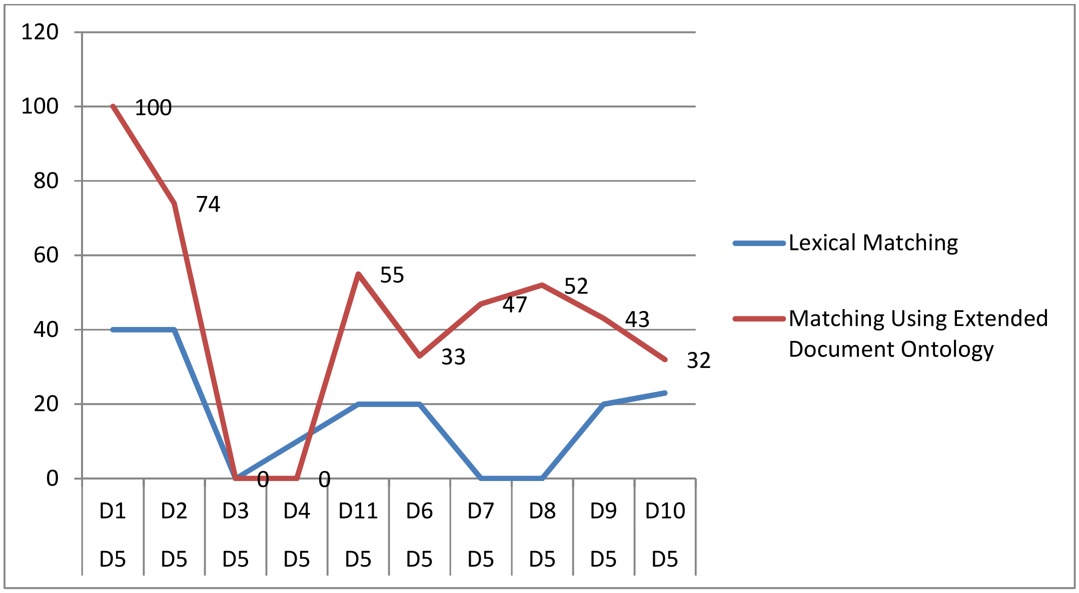


Figure 4.15: Analysis of Lexical Matching and Proposed Extended Document Ontology Matching

The technique of extended document ontology with the help of recent trends is implemented to analyze the results from the hidden information that is also embedded in the document.

During analysis of the documents by using the extended document ontology all the above approaches were also analyzed to verify the importance of spreading of the document ontology. *It has been empirically proved that the concept analysis, concept-relationships analysis, provides better results in terms of the semantic processing of the document, but the extended concept-relationship technique gives better results for the set of documents given in Appendix I Table 1.2 as compared to other techniques.* The results of all the designed techniques are given in Table 4.8 which shows that each level of improvement in the design and development of the above explained techniques gives better semantic score between two documents.

Table 4.8: Results of Proposed Semantic Similarity Computations Techniques

Set of Documents	**Lexical Analysis Score**	**Proposed Conceptual Analysis Score**	**Proposed Related Conceptual Ontology Based Analysis Score**	**Proposed Extended Related Conceptual Ontology Based Analysis Score**
D_1, D_2	.39	.40	.40	.72
D_3, D_4	.57	.59	.61	.66
D_2, D_3	.09	.09	.11	.20
D_4, D_1	.65	.69	.71	.71
D_3, D_5	.58	.60	.62	.66
D_4, D_5	.57	.58	.58	.63
D_5, D_6	.61	.64	.66	.66
D_4, D_6	.58	.59	.60	.66
D_2, D_5	.20	.21	.22	.37

Note: D_1: https://www.android.com/

D_2:www.microsoft.com

D_3:www.samsung.com

D_4: http://www.windowsphone.com/

D_5: https://en.wikipedia.org/wiki/Open-source_software

D_6: http://opensource.org/

For further deep assessment of the results and performance of above explained techniques, the set of 50 documents related to domain artificial intelligence given in Appendix I Table 1.1, and mobile devices given in Appendix I Table 1.2, retrieved from Google search engine were also processed. We have discovered that all the approaches are giving better results than the traditional approaches showing their superiority.

4.9 SUMMARIZATION OF ONTOLOGY APPROACHES

The layered structural design of semantic web makes available various mechanisms for giving better search approaches and extracting considerable web pages. The computation of semantic similarity between the documents further improves the searching of considerable and important web pages. Many similarity algorithms already exist that make the complete use of semantic annotations obtained with the assistance of ontology having concepts and relationships between them. The algorithms computing semantic similarity between the contents of web documents has many applications for improvement in the retrieval of IR like Indexing, Crawling and Ranking of the web pages. The approaches given in this chapter, helps in constructing a semantic view of the content of a web document by extracting the concepts of that document along with the relationships by making the use of knowledge structure i.e. ontology. This semantic view is further extended to the next level to embed the hidden information that is present in the document according to the domain requirements.

In next chapter, we will be giving the techniques that are developed and designed for the specific application of ranking of the set of web pages by a search engine. This will help the user of the web to retrieve the relevant set of web documents by making the use of a common tool i.e. search engine like Google, Ontolook, and Swoogle etc.

CHAPTER V

EFFICIENT RANKING OF WEB DOCUMENT BY COMPUTING SEMANTIC SIMILARITY

5.1 SEMANTIC WEB PAGE RANKING: AN INTRODUCTION

The web is congested with large amount of information which is complex and difficult to process by machine. The complexity of information is increasing day by day due to changing in size of web and technology. Thus, not only the searching of relevant information from web is difficult task but also it sometimes gives irrelevant information. The irrelevant information retrieval is caused by the structure of user query (which is set of keywords) and the way it is used to search the content in the index of search engine. However, it is necessary to consider the relations which are present in the mind of the user while specifying his/her query to the search engine. The traditional search engine considers only keywords from the query without considering or processing the relationships that exist between the query words. In other words, there is a major need of considering the semantic information contained in user query to provide him the most relevant result-set [86]. Therefore, the web is shifting towards semantic web in which we annotate the web document or the query with the semantic information which is the concepts and relationships between these concepts represented in the form of ontology.

A query having the keywords like “Volvo”, “Delhi” and “Chandigarh”, when specified by the user to a traditional search engine with the aim in mind of going from Delhi to Chandigarh by Volvo bus. The traditional search engine will provide the ranked web pages which includes pages like a PDF document which gives the schedule of buses from Chandigarh to Delhi and vice-versa as the first web page. The other ranked pages are having the content which gives the information of hotels that are available in Chandigarh, bus services available in Delhi and Chandigarh, roadmap from Delhi to Chandigarh etc. These web pages having the respective information in their content are not according to the necessity of the user who wants to go from Delhi to Chandigarh by Volvo. As the user requirement is not only for the schedule of buses between the mentioned source and destination but also to consider the availability and booking of seats in Volvo. After getting the result-set it is found that out of 10 retrieved web pages only 3 web pages were having the content which is of user interest according to the user specified query. The reason for retrieval of irrelevant web pages is the processing of information based on lexical approach as also stated above, thus showing the major need for incorporating the relation based approach in ranking process of the web pages.

To retrieve relevant web pages, we have to implement semantic search technique for the web pages. To do this, in addition to the content of web pages, it is necessary to include semantic information about the web pages which can be embedded in the page itself. To be more precise, the main aim of semantic web is the extension of the current web which is collection of unstructured documents into a web which is collection of structured documents". The semantic web pages can be annotated with the help of technology like resource description framework (RDF) so that they can be interpreted by using an additional resource like ontology [58, 87]. By using the ontology, the concepts in query are taken into consideration to rank the relevant web page. It is our hypothesis that there must exist at least one relation between a given pair of keywords in user query.

5.2 EARLIER WORK

The concept of semantic search means that we are incorporating semantics in searching techniques to improve the ranked result-set for a particular query. Many semantic search techniques already discussed in Chapter 2 exist which consider concepts from documents. Some of the techniques focus on ranking of the web documents like in [70], [71]. Almost all of the techniques rely on the prerequisite of processing of documents i.e. preprocessing, crawling, indexing of the web pages that deals with the semantic information [72] [73] [74]. Various ways of semantic search have also been given [75]. These semantic search approaches are used to explore the concepts and relationships in the field of IR but these approaches yet need to be improved to exploit the full potential of semantics in document.

The ranking of a set of the web documents is done by considering the relevance of a user query with documents. Boanerges et. al. [76] has also classified the ranking criteria based on statistical and semantic metrics. The statistical metrics depends on the statistical aspects of ontology like number and connectivity aspects of entities and relationships, whereas the semantic metrics are based on semantic aspects of ontology. The main aim of the research on ranking is basically to provide the relevant result-set to end user by analyzing the content from the semantic web documents and also to enable the existing techniques to uncover all the potential semantic associations between the known concepts [77].

A query is considered as a document in our proposed techniques. To find the similarity computation between the query document and web document, it is necessary to consider all the semantic associations among entities depending on their relevance with respect to the domain. This becomes necessary as ignorance of such processing may result in high number of irrelevant documents in user response. To avoid this, there is a requirement of a customizable criterion that only focuses on the relevant semantic associations which further assist in providing the user with the relevant ranked result-set of documents.

For considering the semantic annotations available in a web document, many semantic search engines have already been developed like Swoogle, Ontolook etc. These search engines provide user with the set of documents which is having maximum relevant documents according to their query.

The consideration of semantic relationships can be explicitly done by adding the semantic to the content of the document by using the schemes like Resource Description Framework (RDF). The RDF is basically a framework which helps in capturing the resource i.e. concept and classes of resources which indicates the relationships between the concepts. It also helps in many semantic technologies which are gaining high popularity and thus are used in wide variety of web applications [78] [79] [81].

Many ranking models have been developed based on the Boolean model [82], Statistical model [81], Hyperlink based model [35], Conceptual model and many more [83] [84] which have been widely used by many web searching techniques. Some ranking models based on Fuzzy set theory, Neural Network, Relevance feedback models also have been widely used for increasing the efficiency of the ranking methods for the search engine. Like [85] gives the integration of the constructed conceptual graph for developing the domain specific ontology which are compared to one or the other domain specific ontology for similarity detection. Danushka et. al. [86] represents the semantic relationships between words on the basis of retrieved lexical patterns clusters. The model given by [86] depends on the semantic associations between words. It uses Mahalanobis measure for computing the semantic relationship between documents as a feature of distance. Jiwei Zhong et. al. [87] has given the technique for Conceptual Graph (CG) matching widely applicable in semantic search. The CG matching handler module is designed in such a way that it takes query graph as the input and a candidate graph is also fetched from the major resource i.e. CG repository. The ranking of the candidates obtained above is returned to the user interface as an output.

Mehrnoush Shamsfard et. al. [56] has given a method of ranking of documents named Orank based on ontology. This new method of ranking of documents is processed by determining semantic similarity between a web document and a query specified by a user with the help of NLP techniques. The NLP techniques helps in stemming of words and extracting phrases from the content of document. The conceptual method based on ontology is then used to include the semantic information i.e. phrases etc. by annotating the web document. This method also expands the query by using the spread activation algorithm. This algorithm helps in expanding the query from various aspects so that more and more semantic information can be incorporated in it. Finally, the new expanded query and the document which is annotated with semantic information are used for finding similarity between them. This semantic similarity computation indicates the degree of relevancy by using available statistical techniques.

In next sub-section, the techniques used for semantic web searching are discussed.

5.2.1 Methods of Semantic Web Searching

The existing traditional web searching and ranking models are not appropriate for the semantic web for two main reasons. First, is that these models are not capable of differentiating between the annotated semantic web documents from ordinary web pages. Second, is that these models do not parse and process the internal formation of semantic web document and external semantic links present between them. Therefore, the concept of semantic web emerged [2, 75] by mainly using ontology based semantic annotations.

Many ranking approaches for semantic search engines retrieving information from semantic web have been described and already being used. Shaaojie et. al. [23] has given a ranking model named SimRank for detecting the semantic score associated with each web page so that they can be ranked according to the detected semantic web page score. The semantic web page score is obtained by considering the information present in the content of semantic web page by making partitioning of already constructed web database so that numerous social web networks can be constructed. The SimRank improved the common traditional ranking algorithm i.e. Page Rank by considering the semantic information contained in the content of query and also the relevance of a web page with respect to the query. Yet, the limitation of SimRank ranking algorithm is the time taken in computation and assignment of semantic score of web pages.

Hung et. al. [120] has given the measures to improve the similarity computed by the SimRank. In SimRank, according to the authors the similarity between two nodes is not computed accurately when they are reachable each other from the path of odd length. The new similarity measures Acoss and Ascoss++ address this problem and compute the efficient similarity score between any nodes by considering all the weighted edges in the given network.

Golub G. et. al. [89] discovered the design of model for finding relationships between a given set of concepts which are present in a query specified by a user. The idea is to use the concept of content similarity. The content similarity of the set of web page helps to construct the ontology of documents by using the basic techniques of pre-processing, normalization, Latent Semantic Analysis (LSA) by singular vector decomposition, graph construction and graphical user interface construction.

An approach to compute the similarity measure between a web page and a query by considering the semantic distance between the semantic descriptions of both the documents has been given by Rudi L. et. al. [88]. The basic requirement for semantic similarity computation by this approach is that the user needs to specify all the relations between the words of a query. Therefore, this basic requirement is not reasonable particularly for naive users of web. The applicability of this semantic similarity approach in actual context of information becomes very inadequate as the number and type of relationships between the words of query is sometimes not known to user itself, or they may have incomplete/ wrong information with them.

Another ranking model named SemRank has been given by Anyanwu K. et. al. [90]. The SemRank is based on the relevance score obtained between a web page and a query, thus it gives a novel technique for ranking of modular searches. The main focus of SemRank is to detect how much semantic information is associated with a web page which is required by the users according to the query specified by them. The approach also focuses on the complex relationships identified from the content of the web pages and their ranking. The semantic web search engine named Swoogle has already been given by Ding L. et. al. [91] which is actually a crawler based indexing and retrieval system for retrieval relevant semantic web pages from the semantic web. The process involved in Swoogle ranking of semantic web documents includes finding of appropriate ontology's, detection of instance data and characterization of semantic web and computation of semantic rank score. Hyunjung P. et. al. [92] discussed a link based ranking algorithm for semantic web resources which is independent of link direction between web pages of semantic web. As in the semantic web, the web page direction of Resource Description Framework (RDF) is known by a specific schema not by the process of voting process as it is done in current web i.e. WWW. The link based algorithm for ranking focuses on classes and the property weights are assigned to each resource available in the web page or query depending upon the importance of the resource in each identified class.

Li Y. et. al. [24] developed a system called OntoLook which reflects on all relevant relationships that exists between concepts for computation of semantic web ranking by a semantic web search engine. The OntoLook semantic web search engine processes not only the keywords but also the relationships between the identified entities which is already incorporated in the defined architecture of semantic web. The ranking given by OntoLook will have set of web pages which includes the identified concepts and relationships between them. The interface of the semantic search engine will give the user a means to identify and select the concept available for each keyword which the user wants to specify in their query. This is done because while processing of the query and the web pages the concepts and relationships need to be considered. Although the interface of the OntoLook helps user to specify the concepts and relationships which will definitely provide more information associated with a query but there is still a limitation existing in the ranking strategy. This limitation is covered to some extent in Lamberti L. et. al. [4] technique of relation based ranking of semantic web documents. This ranking strategy exploits the significance response and post method result-set to build up and design a ranking methodology which deeply considers relationships available between keywords from web pages. In this ranking methodology which is based on concepts and relationships, the graph for domain based ontology, search engine query, semantic annotation of web page and page sub graph are created. Then, a probability for web page which is to be selected according to the user query is calculated. This probability is used to rank the web pages. The major limitation linked with this approach is that while computation of probability there is a chance of zero score for a web page. Although the authors of this ranking strategy claim the computation of zero probability for a web page is common and it does not have an effect on significance of the web page to the query. The problem occurs when two or more than two pages are assigned zero score, because the zero score cannot be used to order between or among web pages.

The conclusion is that any ranking scheme like Page Rank [95] used by the Google [94] [57] [93] can arrange the result-set in proper formation which can meet the needs of the user. However, the above stated techniques appear promising but the effectiveness of the techniques can be measured by finding computational complexity and accuracy of result from a large size i.e. billions of indexed pages. These techniques when further improved with the help of semantic information processing technique they can provide the user a result-set which will have no or limited set of irrelevant pages.

In this Chapter, we have proposed two techniques for ranking of the web documents as per the query specified to a search engine by a user. These techniques are not based upon the lexical analysis of the document rather these techniques consider semantic associations between concepts from the web document and the query document.

5.3 RANKING MODEL USING WEIGHTED SEMANTIC ASSOCIATION

The semantic association ranking technique is designed to consider the view of user while giving a query to a search engine for retrieval of documents which are relevant with respect to the query. The model of semantic ranking basically focuses on the query. The query is processed by keeping the user broad view/intention into consideration. The selection of relevant documents can be done on the basis of this intention. This ranking model provides the result-set of the web pages on the basis of computation of semantic similarity between query and a web page. The higher the semantic similarity score of a web page with respect to query the higher is the rank of the web page showing its relevance and importance to user. The overall structural design of the semantic ranking model is given in Figure 5.1.

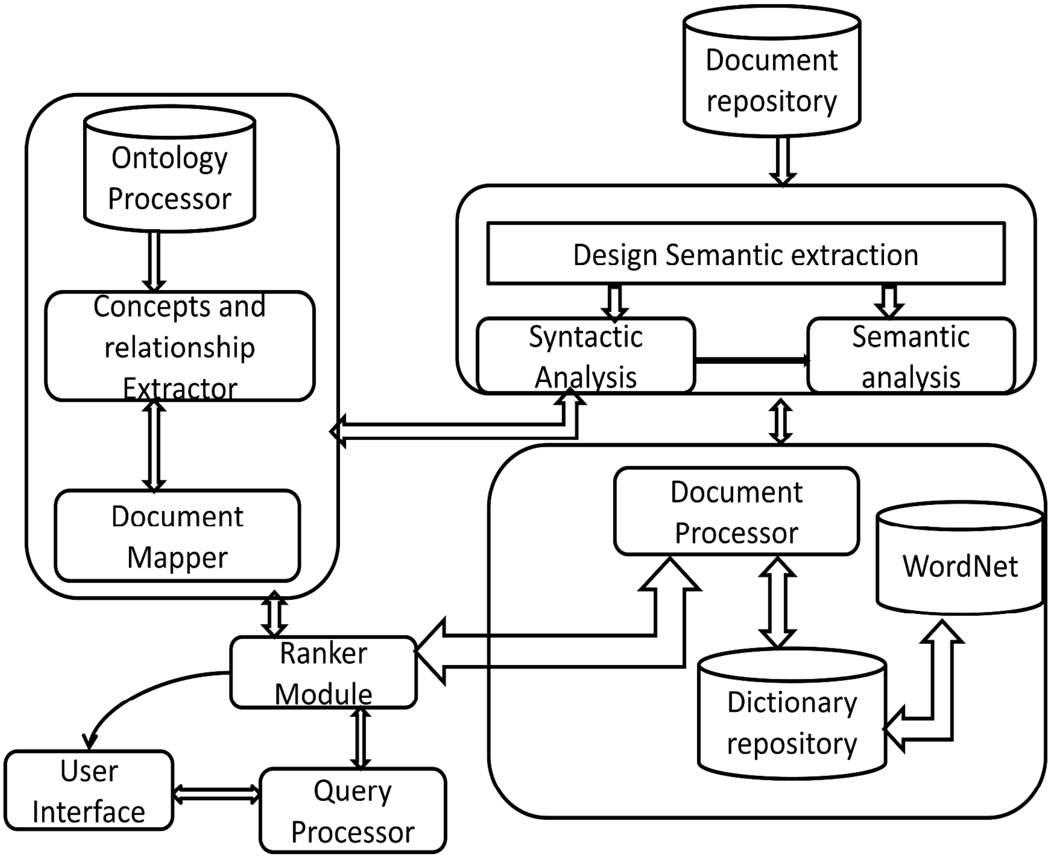


Figure 5.1: Architecture of Proposed Ranking Model

The major components of the ranking model are a Document Processor, Ontology Processor and Ranker Module. The technique first processes the document by extracting the words from the document by using the Document Processor which performs the syntactic analysis and thus constructs a Vector Space Model. A dictionary is also constructed and maintained in which the words that belong to a domain along with the synonyms and meaning are stored. The domain words stored in the dictionary are assigned weights inspired from fuzzy logic theory. These weights are decided according to importance of the word in the given class of words visualized as fuzzy set. Now, after processing of document by document processor, the processed form of document is given to the ontology processor which helps in extracting the concepts and relationships from a document by using concept analyzer and relation extractor. The domain related query given by the user to a search engine is also processed in the same manner.

Therefore, in our approach we are constructing two databases one is weighted word dictionary and other is weighted relations ontology. The weighted word dictionary database is having words along with the synonyms with assigned weights which indicate the level of relevance with the domain. It means more is the value of weight assigned more is the importance of the word and its synonyms in a given domain. Similarly, the weighted relations ontology is having the weighted relationships between the words/synonyms stored in dictionary database. The more is the weight assigned to a relation the stronger is the connection/association between words. These weights are decided manually by processing a set of documents in the given domain. The weighted word dictionary and ontology weighted relations database are shown in Table 5.1 and Table 5.2 respectively.

Table 5.1: Dictionary Based Weights

S No	Domain (Education)	Weight Assigned	Synonyms
1	process	.8	-,-,-
2	study	1	-,-,-
3	learning	1	-,-,-
4	experience	.8	-,-,-
5	social	.9	-,-,-
6	official	.8	-,-,-
7	dynamic	.8	-,-,-
8	starting point	.5	-,-,-
9	used every where	1	-,-,-
10	university	.6	-,-,-

To process a document, the words available in the documents are represented as vector space model. The complete structure of semantic ranking works in two stages. In the first stage of processing, mapping or association of each word that is available in the vector space is done by using the weighted word dictionary database. The mapping process will give semantics associated with each word i.e. set of synonyms, meaningful definition, weights associated with each word indicating the importance of the word with respect to the domain. This mapping process is done sentence by sentence.

Table 5.2: Ontology Based Weights

S No	Concept-Relationship Between Objects Represented in FOL	Weights Assigned
1.	of (education, man)	.8
2.	related to(education, study)	.7
3.	has(person, education)	.6
4.	at(education, college)	1
5.	at(education, school)	1
6.	at(education, university)	1
7.	is a(education, process)	1
8.	has(life, learning)	.9
9.	through(learning, experience)	.9

Then all the relevance value obtained for each sentence is aggregated by using statistical approach e.g. mean, median, variance computation to get the relevance score associated with each paragraph of the document. Further, the paragraph associated relevance score is integrated to compute the document relevance score with respect to the domain again by using available statistical models. Next, the query given by the user is processed by using the same approach of document processing and the relevance score with respect to the document is obtained. This relevance score of the query specifies the importance of the document with respect to the domain as the score obtained is with the help of already stored fuzzy weighted terms available in the semantic dictionary repository.

In second stage of processing, the other database named weighted relations ontology is used for analyzing the document with respect to the query in terms of concepts and relationships available in them.

In this stage the mapping process by ontology processor is done for the document and the query is to identify the relationships between the synonyms obtained in the first stage. These synonyms are considered as concepts available in the weighted ontology relations. This mapping process will compute all the weighted relationships available in the ontology corresponding to the concepts of the document and the query. The mapping process in second stage is also done in sequence i.e. first for sentence level, then for paragraph, which is further combine to get the value for the complete document again by using available statistical techniques.

Finally, the computed relevance score associated with the document with respect to the domain and query are compared and the maximum score obtained is considered as the final semantic score of the document with respect to the query. In this way all the document available in the document repository are processed and the final semantic score is obtained for each document with respect to a specific query given by the user to the search engine. According to the semantic score obtained, for each document ranking can be done i.e. the higher the semantic score of a document, the higher is the rank/priority of the document with respect to a particular query. The algorithm for the same ranking model explained is given below as Algorithm 5.1.

Algorithm 5.1

Step1: Create a Text-List (by links).

Step 2: Take query as a text: a String.

Step 3: For each Text in Text-List do:

(a) Construct Text-Vector-Space.
(b) Construct Domain-Dictionary of words.
(c) Using Statistical-Model () and Domain-Dictionary, Calculate relevance-value of Text with respect to Query.
(d) Construct Domain-Ontology of the Text.
(e) Calculate Domain-Similarity of Text value by using Domain-Ontology.

(f) Determine the utmost of Domain-Similarity score and relevance-value and call it Relevance-Score.

Step 4: Go to step 3 until no text is left in Text-List or no more texts are to be considered.

Step 5: Arrange the text (links) according to decreasing order of relevance-score and assign them ranks.

Step 6: Display the texts according to their ranks.

The details of basic terminologies used in above algorithm are given below:

Text-Vector-Space: Consists of text words their weight age.

Domain-Dictionary: Consists of text-words (nouns, pronouns, synonyms and their weightage).

Domain-Ontology: A graph containing concepts as nodes and relations as edges.

Domain-Similarity is calculated for the Text with respect to Domain-Ontology and Domain-Dictionary.

Statistical-Model is used to calculate the relevance score of text with respect to domain-Dictionary.

Further, the detailed working of given semantic ranking model is explained with the help of example. The set of four documents D1, D2, D3, and D4 which part of content is shown in Table 5.3 with respect to the query document as “What is education”.

Table 5.3: Relevance Semantic Score of Documents According to User Query

SNO	Document No.	Relevance value to domain specific dictionary (Dv)	Relevance value to ontology (Ov)	Final Relevance value=Max(Dv, Ov)
1.	D1	.85	.94	.94
2.	D2	.74	.88	.88
3.	D3	.45	.75	.75
4.	D4	.65	.6	.65

Note:

D1: Education is a lifelong process. A person learns through his experience. It goes on forever from his birth to death without any break or barrier.

D2: Education of man does not begin at school but begins at birth. It ends not when he graduates from university but ends at his death. Hence, Education is a lifelong process.

D3: Education is not only academics but social also. It is important in one's person life.

D4: In a person life everyone needs to be educated and social. Everyone learns through experiences gained in one's life.

The query "what is education" is related to the domain education which weighted word dictionary and weighted relations ontology are already identified and stored as shown above in Table 5.1 and Table 5.2. The entire four documents are processed in two stages. In first stage, the processing of document with respect to the domain will provide the semantic score associated with each document with respect to domain. In second stage, the relevance score of each document in respect to query document is computed. Finally the maximum value computed in both the stage is assigned to each document showing their semantic relevance with respect to the query and domain as shown in Table 5.3.

Now, the above relevance value obtained helps in giving the ranked list of the document as the higher semantic score associated with the document indicates that the document have the higher rank showing its relevance/importance according to the query. The ranking of all the above four document is shown in Table 5.4. Also, all the above four documents are ranked by Google search engine by sending the same query i.e. "What is Education". The ranking score for D1, D2, D3, and D4 given by Google Page Rank search are .62, 1.24, 1.11, and 1.13. The same set of documents is analyzed by human beings to get the human rating for these set of documents by considering the user view in the query given to the search engine. To show the superiority of the given semantic ranking model the variance of the obtained ranks, Google rank and human analysis ranking is calculated and shown in Table 5.4. Finally, the variance computed shows that the variance by semantic ranking model according to the human ranking is minimum in each case as compared to the variance obtained from Google rank when compared with human ranked list if documents which eventually shows the importance of the given semantic ranking model.

Table 5.4: Ranked Set of Documents Relevance to User Query

S No	Actual Rank	Google Rank	Variance by Google Rank	Our Rank	Variance by Our Rank
1.	D1, D2, D4, D3	D2, D4, D3, D1	10	D1, D2, D3, D4	2
2.	D21, D23, D25, D26, D22, D24	D21, D22, D23, D24, D25, D26	34	D21, D25, D26, D22, D23, D24	10
3.	D32, D33, D31, D34	D32, D33, D34, D31	18	D31, D32, D33, D34	6

The sets of documents represented by (D21, D22, D23, D24, D25, D26) and set (D32, D33, D34, D31) respectively shown in Table 5.4 are considered for processing in the same manner to check the efficiency of the semantic ranking model. The part of contents of each of the document present in the above two sets is given as follows:

D21: Education in its broadest sense is the means through which the aim and habit of a group of people lives on from generation to generation.
D22: Education means the process of becoming an educated person.
D23: Education means to know the knowledge.
D24: Education teaches lesson of humanity. It is very necessary for humans.
D25: Education is the act or process of imparting or acquiring particular knowledge, as for a profession.
D26: Education psychology involves the study of how people learn.
D32: Education is a learning process throughout the life.
D31: Education is a continuous process that comes through experience.
D33: Education is an active and dynamic process.
D34: Person goes on reconstructing experiences throughout the whole life.

The same processing is done for the set of 50 documents belonging to the domains like Artificial Intelligence, Mobile Devices etc. The ranked list of each domain is obtained for 50 queries given by user. *In maximum number of cases it has been found that the ranking is close to human ranking by given semantic ranking model.* The technique of ranking gives the semantic ranked set but still there is an improvement required from the query consideration point of view. A query given by the user is important to rank the document to give user a relevant result-set. In the technique discussed above the document is processed deeply according to the concepts and relationships available in weighted in ontology relations. But, there is a requirement for improving the query processing by understanding the implicit or hidden information which the user of query wants to provide.

In next section, a semantic web ranking technique has been proposed to provide user the set of semantically ranked web pages according to his intention. This technique relies on the semantic content available in base ontology, web page and query given by user, thus giving the ranked result-set which will be close to result-set obtained after human analysis.

5.4 BI-RELEVANCE BASED SEMANTIC WEB RANKING MODEL

The various complex issues discussed in section 5.2 needs to be considered to develop a new semantic search ranking strategy. According to our presupposition, a user wants to retrieve the web pages that are relevant not only to his/her query given to the search engine but also to the particular domain. Therefore, the basic idea of the proposed technique is to consider the maximum related concepts that are present in a web page, user query and base ontology. To realize this idea the relation probability depending on the relationships that are available between any two concept pair in the web page and base ontology is calculated. This relation probability in our technique gives the relevance of the web page with respect to the domain. In the same manner, the relation probability between the web page and the user query (which is considered as a document provisionally) is computed. The relation probability computed between web page and user query gives the relevance of web page with respect to user query. The relation probability as per our hypothesis is a measure of degree with which the relations between two sets of relations (one related to web page and base ontology and second related to query and page) are related. Finally, the joint relational probability is found which will be used to assign the score for each web page. This score will be used to rank the web pages later.

From computation point of view, there is a need to construct a base ontology which will be used initially to find relationships between concepts in user query or web page and later to calculate relational probability as indicated above. The design of base ontology is inspired from the one proposed in [4] with some necessary modification made to incorporate the different domains such as transportation, artificial intelligence etc. The ontology for each web page is also constructed by first pre-processing the web page and then normalization is done for constructing the structure in graphical form. The construction of ontology corresponding to web page is done in the same manner as described in [87] [89]. The proposed semantic ranking technique is not designed to provide altogether different techniques rather it is an important extension in existing one [95]. This enhancement will as per our hypothesis lead to improvement in the existing page ranking technique.

The structural design of semantic search engine is presented in Figure 5.2. The crawler, as all of us know, is used by a search engine to fetch the web pages which are then indexed by an indexer and the ranking techniques are applied on the indexed documents corresponding to a user query. The crawled web pages from the semantic web are stored in a web page database. The stored semantic web pages are annotated with the semantic content of the document by using scheme like RDF, OWL etc. The RDF or OWL parser interprets a web page and transforms it into a representation as required by search logic.

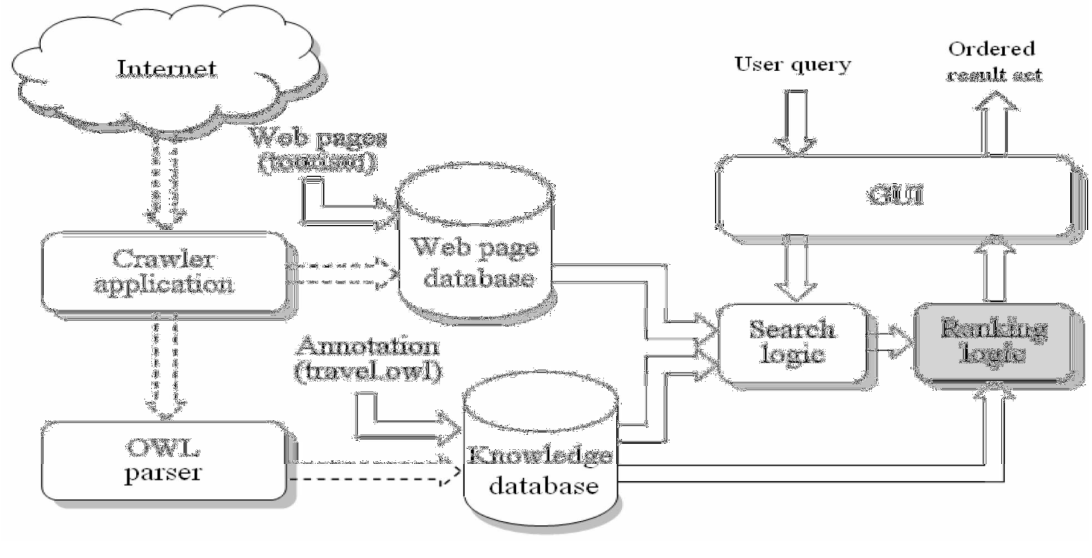


Figure 5.2: Architecture of Semantic Search Engine [92]

The knowledge database is used to store the transformed RDF/OWL documents. The base ontology is also the part of knowledge database and it is represented in the same form as that of semantic web pages. The search logic component of the architecture of the system is used to fetch/retrieve the significant result-set from the web page database. Then, the retrieved web pages are ordered according to the semantic score assigned to each web page as per the proposed technique. It is assumed and later verified that the higher the semantic score of a page, the most relevant is the web page according to given user query. Therefore, in summary, we can say that the proposed model is having two basic steps. First, relationships between concepts in user query and web page are found and similarly relationships between concepts in web page and ontology are found. Second, the relationships are used to find the relational probability between user query and web page which, as stated above is measure of degree with which the relations in the query or web pages are related. Similarly, the relationships between web page and ontology are used to find the relational probability between web page and ontology which, as stated above is measure of degree with which the relations in the web pages or ontology are related These two steps are discussed in details in coming two sub sections.

5.4.1 Identification of Relationships Among Concepts

The one of the major consideration is to find the relationships among concepts. In relation based search engine [24], while forming a query there is requirement to provide the keywords along with a particular concept associated with that keyword by selecting the same from the pull down menu available in the search engine. The pull down menu will provide all the concepts which can be constructed using the ontology web language (OWL).

The base ontology created for the semantic ranking model is in the form of graph. This graphical structure of ontology gives the concepts represented as nodes and the edges represented as the relationships between these concepts. The relationships edges are labeled with the number of relationships and the kind of associations that subsist between the concepts. In the similar manner, the query is processed for creation of query graph, to obtain the relationships among concepts by using the base ontology.

Next, a page graph is constructed by using the base ontology for each semantic web document with the help of OWL parser. The page graph includes the concepts and relationships between the concepts available for the semantic web document. The page graph constructed for each semantic web page is also called as page ontology. Finally, the semantic rank score is assigned to each semantic web page with respect to the query by computing the relational probability as discussed above.

In the base ontology the nodes represent concepts and the edges represents the relationships between the nodes which sample part is shown in Figure 5.3. This base ontology is constructed in the same manner like travel.owl for the domain traveler [87, 96].

Formally, the base ontology graph is represented as G (C, R), where:

C= set of vertices in constructed graph G. {C_1, C_2, C_3, C_4........, C_n} are the total number of n concepts which are present in the constructed domain base ontology, and

R=set of edges in constructed graph G. {R_{ij} | represents the relationships that is present between two concepts C_i and C_j, such that $i<j$}.

The base ontology graph G (C, R) is a weighted graph in which each edge is allocated a weight which defines the total number of relationships that exist between two nodes (here concepts) of the graph.

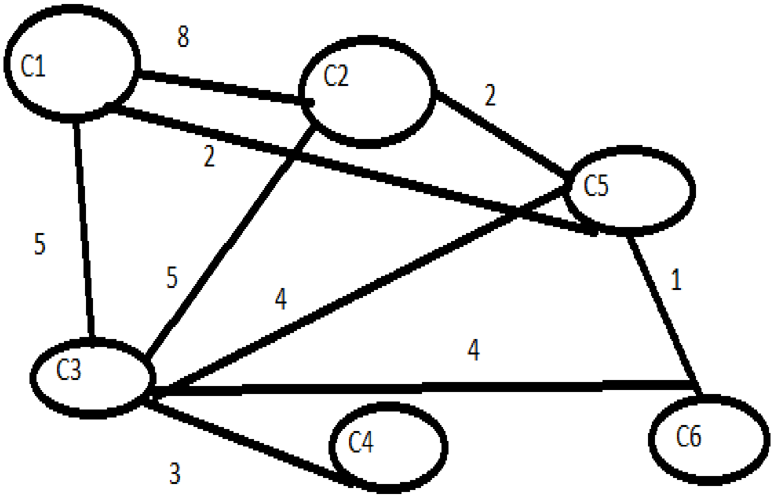


Figure 5.3: A Sample Ontology Graph with Six Concepts

For example, Figure 5.3 depicts an ontology consisting six concepts and number of relations between them. The six nodes named C_1, C_2, C_3, C_4, C_5 and C_6 are described as follows:

1. (C_1: Source),
2. (C_2: Destination),
3. (C_3: Accommodation),
4. (C_4: Accommodation Classes),
5. (C_5: Running Timings),
6. (C_6: Booking).

The description related to the relationships between any two above mentioned concept pairs of the underlying graph is shown in Table 5.5. In the table concept pairs, type of relationships between these concepts pairs, and total number of relationships between these concepts pairs is given depending upon the type of relationships. The detailed base ontology description is given in Appendix II Table 2.5. Now, when user provides query to a search engine, it is specified by using keywords and their relational concepts from the pull down menu of the search engine. After the query description by the user the query graph is constructed by using the OWL parser. Formally, the query graph is also defined as $Q= \{C_Q, R_Q\}$ where,

C_Q= (set of vertices in the query graph) which is collection of keywords/concepts again represented by $\{C_1, C_2, C_3,\ldots\ldots C_n\}$ from the query and,

R_Q= (set of edges) which is collection of relationships between the query keywords/concepts given by the user at the time of description. It is again represented as {R_{ij} | represents the relationships between the query concepts that is present in two query concepts C_i and C_j, such that $i<j$}.

The query graph constructed is also a weighted graph as is the case with the base ontology graph in which the edges are labeled with the number of relations between the set of concepts pairs that are present in the query. Next, each semantic web page related to query stored in the knowledge base are considered. The constructed page graph for each semantic web document is represented by: $P = \{C_P, R_P\}$, where

C_P is collection of concepts mentioned in web page and

R_P is set of relationships that exist between concept C_i and C_j.

For computation of the semantic rank score which is the aim of our semantic ranking technique, we have used the following symbols to calculate the probability of concepts relationships in a web page corresponding to the query and the ontology:

α : count of relationships between concept pairs present in the query graph,
δ: count of relationships between concept pairs in the page graph and
η: count of relationships between the concepts pairs in the ontology graph.

Table 5.5: Relationships between Concepts Pairs

Concept Pairs	**Relations between Concept Pairs**	**No. of Relations**
c_1, c_2	has part, has public transport, has volvo to, has train to, has flight to, has roadways to, from to, to from	8
c_1, c_3	has accommodation, is a way to, facility, organizes visit to, public transport	5
c_2, c_3	has accommodation, is a way to, facility, organizes visit to, public transport	5
c_3, c_5	day wise, hour wise, month wise, year wise	4
c_1, c_5	from to, to from	2
c_3, c_4	has types, has ratings, has classes	3
c_3, c_6	through credit, through cash, online booking, e-ticketing	4
c_2, c_5	from to, to from	2
c_5, c_6	booking for hours	1

In the computation process of semantic rank score, firstly the relation probability of relation R_{ij}, between the concepts C_i and C_j in web page and base ontology is calculated. It may be noted that higher the value of relation probability for a concept pair between a web page and the base ontology more will be the relatedness of the web page with respect to base ontology in the context of a given concept pair. Further, it is assumed that the number of relationships between concept pair C_i and C_j in base ontology will always be more than the number of relationships between same concepts pair present in the web page. The calculated relation probability between the web page and the base ontology is represented by τ_{ij} as defined below by the formula:

Relation-Probability $\tau_{ij} = \delta_{ij}/\eta_{ij}$

Likewise, the relation probability of relation R_{ij} , between the concepts C_i and C_j is calculated for the query graph with respect to the page graph. Again, it is assumed that the number of relationships between concepts pair present in the web page is more as compared to the number of relationships between that concepts pair present in the query. The calculated relation probability between the web page and the query is represented by Ω_{ij}, given below by the formula:

Relation-Probability $\Omega_{ij} = \alpha_{ij}/\delta_{ij}$.

This relation probability calculation is performed for each concept pair available in web page and base ontology to calculate cumulative relation probability, as discussed in next subsection. Same calculations will be performed in the context of user query and web page to calculate cumulative relation probability, also discussed in next subsection. Finally, these two cumulative relation probabilities are used to find out the joint relational probability which will be the indicator of the relatedness of a user query to the web pages in a specific domain. In other words, this joint probability computation as per our hypothesis will definitely give more semantically associated results corresponding to the user query, as all the relationships between the concepts that are available in the ontology; page and the query are considered.

5.4.1.1 Probability Computation for Ranking

In this sub section, the step by step computation of probability for obtaining the relevance semantic score of the web pages is given. The cumulative relation probability, designated as P (P_k, O) of kth page and ontology is calculated by multiplying the relation probabilities τ_{ij} corresponding to each concept pair in the page.

P (P_k, O) = $\Pi\tau_{ij,}$ where i and j range for all concept pair Ci and Cj in given kth document.

Similarly, the cumulative relation probability, designated as P (Q, P_k,) of kth page and query is calculated by multiplying the relation probabilities Ω_{ij} corresponding to each concept pair in the query.

P (P_k ,Q)= Π Ω_{ij} , where i and j range for all concept pair Ci and Cj in given kth document.

The joint relational probability which is a score calculated by adding the cumulative probability P (P_k, O) and P (P_k, Q) as given below:

P (Q, P_k) = P (P_k, O) + P (P_k, Q), where k ranges from 1 to N (i.e. total number of page relevant to user query).

For illustration, let us suppose that user enters the set of keywords and their related concepts as: [{keyword: Volvo, concept: accommodation}, {keyword: Delhi, concept: source}, {keyword: Chandigarh, concept: destination}]. It is also taken that as assumption that the intension of user while describing the query in above form of keywords and concepts is to go from Delhi to Chandigarh by Volvo bus at some chosen time. Now, the hypothesis is made that the user would rarely specify a sequence of keywords which do not relate with each other. So, there must exist at least one relation between the keywords/concepts specified by the user. If then also some keyword/concept do not relate to any of the keyword/concept then it is of no interest of the user, as it will automatically disconnect with all the rest of identified related concepts.

Now, presume that the semantic web contains only three semantic web documents related to domain of travel. These three semantic web documents are represented by web pages P_1, P_2, and P_3 respectively. Their corresponding graph is shown in Figure 5.4(a) and the corresponding constructed query graph is shown in Figure 5.4 (b).

From Figure 5.4 (a) the concepts related to the three web pages P1, P2, and P3 can be obtained and the score on the edges provides the total number of relations that exist between the two concepts of respective page.

Figure 5.4 (b) gives the concepts from the query graph revised with respect to each web page to consider all the related concepts from the query and the web pages.

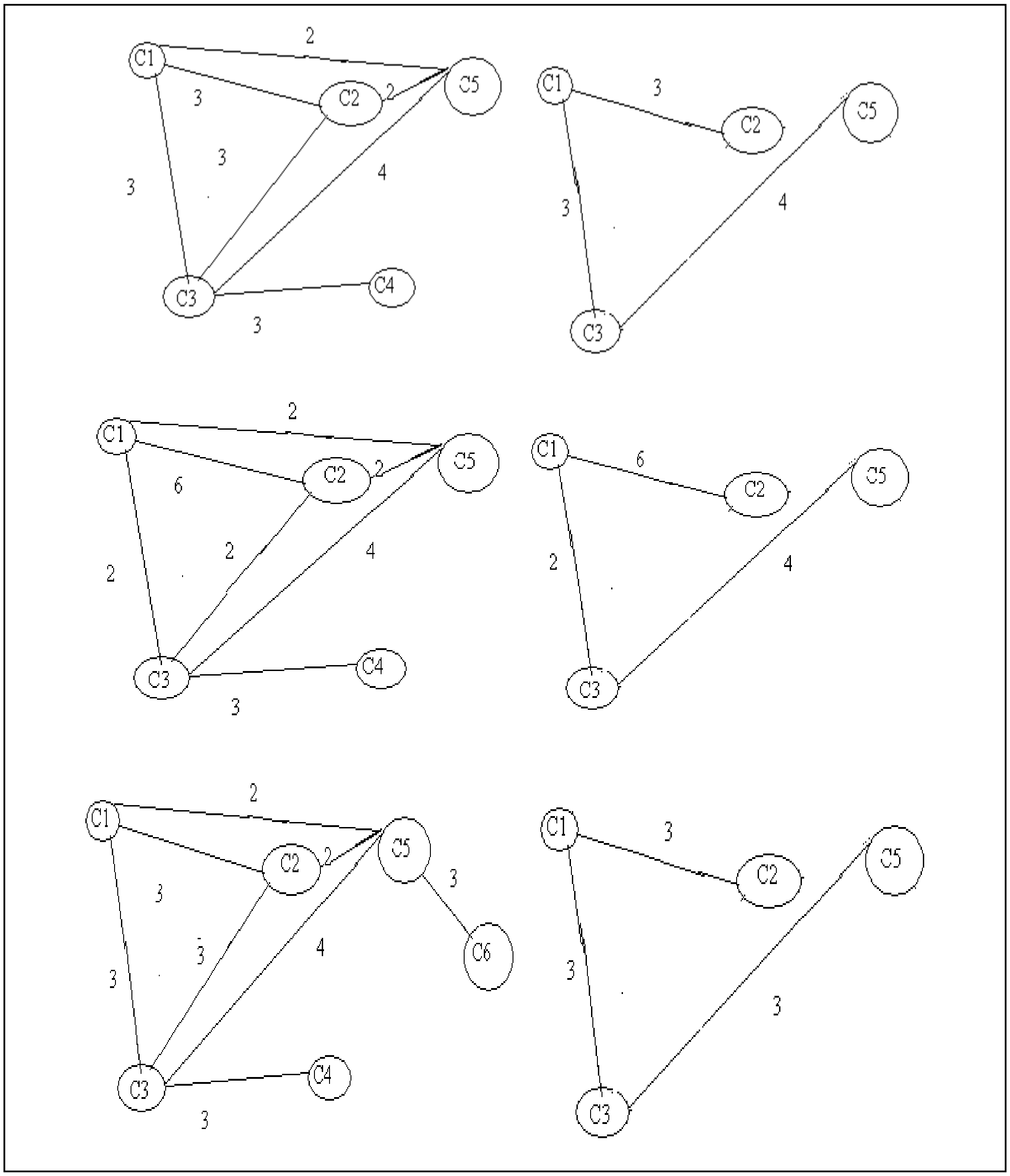


Figure 5.4: (a) Page Graph with respect to Ontology shown on left side, and (b) corresponding Query Graph with respect to Page shown on right side.

Now, after construction of the page graph, query graph and underlying ontology the computation of probability is done as follows:
For first web page P_1, the relation probabilities P (R_{12}, P_1), P (R_{13}, P_1) with respect to the underlying ontology where R_{12} is the relationships between C_1 and C_2 for page P_1 can be computed as:

$\tau_{12}= \delta_{12} / \eta_{12} =3/8$, and

$\tau_{13}= \delta_{13}/\eta_{13}=3/5$, and

Likewise, other relation probabilities

P (R_{23}, P_1), P (R_{15}, P_1), P (R_{25}, P_1), P (R_{35}, P_1) and P (R_{34}, P_1) are calculated. The complete relative probability of page P_1 with respect to ontology O is calculated as:

$P(P_1,O)=P((R_{12},P_1)\Pi(R_{13},P_1)\Pi(R_{23},P_1)\Pi(R_{15},P_1)\Pi(R_{25},P_1)\Pi(R_{35},P_1)\Pi(R_{34},P_1))$.

Where P (P_1, O) in above formula represents the probability computation of web page P_1 with respect to ontology O of the domain to which the page is related.

The reliability on ontology is only due to the knowledge that the concepts which are physically present in ontology and then searched in the web page are considered along with all the relationships count and type of relationships that exist between these common concepts of ontology and web page. Now the calculation of relevance semantic score is done which is to be associated to web page $P_{1,}$ further indicating the ranking score with significance to user query. For this computation the relative probability P (R_{12}, Q) of relationship R_{12} available in the web page with respect to the query Q is calculated as:

$\Omega_{12=} \alpha_{12}/\delta_{12} =1/3$.

Likewise, for all other relations that are present in web page and query are taken into consideration and the relative probability of each relation between pairs of concepts available in the web page and the query is computed. The total probability of web page P_1 with respect to the user query is computed by the formula as given below:

$P(P_1,Q)= P(R_{12},Q)\Pi P(R_{13},Q)\Pi P(R_{35},Q)$.

Finally, the joint relational probability between the user query and the web page is computed by the given below formula as:

$P\ (Q,\ P_1) = P\ (P_1,\ O)\ \upsilon\ P\ (P_1,\ Q).$

In view of the fact that the events are not correlated, therefore

$P\ (Q,\ P_1) = P\ (P_1,\ O) + P\ (P_1,\ Q).$

The above expression can be decomposed as:

$P\ (Q,\ P_1) = \Pi\ \tau_{ij} + \Pi\ \alpha_{ij}$ {for i, j=1, 2......n}

And, thus can be revised as:

$P\ (Q,\ P_1) = [\tau_{12}.\tau_{13}.\tau_{23}.\tau_{15}.\tau_{25}.\tau_{35}.\tau_{34}]+[\alpha_{12}.\alpha_{13}.\alpha_{35}].$

In conclusion, the relation probability and joint relational probability is computed for the three web pages P_1, P_2 and P_3. The $P(P_1,O)$=.375 and $P(P_1,Q)$=.028 and the joint probability is computed as: $P(Q,P_1)$=.375+.028=.403.

In the same manner, the joint relational probability of user query corresponding to web page P_2 is computed as $P\ (Q,\ P_2)$ = .162 and joint probability of user query in accordance with web page P_3 is computed as $P\ (Q,\ P_3)$ = .442.

Now, according to the semantic ranking model approach discussed above the ranking or order of available web pages P_1, P_2 and P_3 according to the user query is defined as $P_3>P_1>P_2$ providing more relevant web pages to the user. Additionally, other examples related to domain of hotels in which the user query is considered as set of keywords/concepts as hotel, Delhi and airport. Here, it is assumed that the user is giving query in this form with the aim or need of booking of hotel which is situated close to the airport in Delhi. After, computation of joint probability of each extracted and interpreted web page with respect to specific user query the relevant ranked semantic score web pages are presented to the user for the set of documents related to domain artificial intelligence given in Appendix I Table 1.1. It has been observed that the method of ranking explained above provides the ranking of the web pages in order which have more relevant web pages on the top of the list. The performance of the given method has also been evaluated on the set of documents given in Appendix I Table 1.1 and Table 1.2) to strengthen the work.

5.5 IMPLEMENTATAION AND ANALYSIS OF SEMANTIC RANK MODEL

The detailed processing of semantic rank model depends on the keywords/semantic concepts, their association with each other specified in the web page or the user query. This would further change or modified for domain to domain as per the information/knowledge associated with the domain itself. The comparison of the performance of the given ranking model based on semantic content is done with the ranking algorithm which is based on relations between concepts which is given by Lamberti et. al. [4]. In addition to this, the process-wise comparison is also done with results obtained by Page Rank Citation given by Berin, Motwani and Winograd [95]. Table 5.6 represents the first five URL's given by the Google search engine for the query as the set of keywords Volvo, Delhi, and Chandigarh. Further, these five URL's were ranked with the relation based ranking algorithms for semantic web, and the rank number corresponding to each URL is shown.

Table 5.6: URL's for the Query Volvo, Delhi, and Chandigarh

First five URL by Google	**Relation based Page Ranking**	**P(p,O)**	**P(p,q)**	**P(q,p)**	**Our Ranking**
http://www.sunrisevilla.in/chandigarh/delhi-chandigarh.asp	3	.003	.055	.058	**5**
http://www.makemytrip.com/bus-tickets/delhi-chandigarh-volvo-ac-seater.html	5	.24	.031	.271	**2**
http://www.scl.gov.in/pdf/bus-sch-pdf	4	.02	.083	.103	**4**
http://www.online-bus-tickets.in/delhitochandigarh-volvo.html	1	.03	.56	.135	**3**
http://hartrans-gov.in/online/index.asp	2	.32	.011	.33	**1**

Next, to find the rank order of these five URL's with the semantic rank model the probability of content present in each web page is computed with respect to the underlying ontology and the query. The semantic ranking number is also shown corresponding to each URL in set of five URL's.

In Table 5.6, the results obtained are analyzed by computing the joint probability for the first five URL results given by Google search engine. Also the computations done by relation based page ranking for ranking of same set of URL are analyzed for the query containing keywords Volvo, Delhi, Chandigarh and their corresponding concepts as accommodation, source and destination. Correspondingly, the results of first five URL given by Google search engine are also computed for another user query that is defined by collection of keywords as Hotel, Delhi and Airport with corresponding concepts defined as accommodation, destination and nearby hotel.

In Table 5.7, the results obtained are analyzed by computing the joint probability for the first five URL results given by Google search engine. Similarly, the computation done by relation based page ranking for ranking of same set of URL are analyzed for the query containing keywords Hotel, Delhi and Airport and their corresponding concepts as accommodation, destination and nearby hotel.

For the deep analysis of the performance of given ranking model the result set obtained by semantic rank model is being compared with the result-set obtained from the Relation based Ranking algorithm for given set of documents semantic web documents. The results are shown in Table 5.8, in which there are four types of queries having different set of keywords are given. Each query result set is examined and the corresponding ranked order is given for each query obtained by relation based ranking and our ranking technique. The actual ranking of each result set corresponding to the query is also considered based on *sample of 50 human rating of each web document*. From these web pages the actual relevance of the web pages are determined according to the intended query. Then the variance between the semantic ranking model and ranking algorithm based on relations present in semantic document is computed for each relevant result-set obtained corresponding to each unique user query as shown in Table 5.8.

Table 5.7: URL's for the Query Hotel, Delhi and Airport.

First five URL by Google	Relation based page ranking	P(p,O)	P(p,q)	P(q,p)	Our Ranking
http://www.cleartrip.com/hotels/india/newdelhi/locality/airport-zone/	4	.267	.33	.601	**2**
http://www.airporthoteldelhi.com	5	0	.33	.33	**4**
http://newdelhi.airporthotelguide.com	2	.533	.25	.78	**1**
http://www.newdelhiairport.in/eaton-smart.aspx	1	.237	.25	.487	**3**
http://www.newdelhiairport.in/travellers.aspx	3	.112	0	.112	**5**

Table 5.8: Comparison of Ranking of first five URL to Corresponding User Query

SNO	Query given in Google Search Engine	Relation based Ranking of first five URL' given by Google	Our Ranking for the URL's given by Google	Actual Ranking for the URL's given by Google	Variance by Relation based Ranking	Variance by Our Ranking
1	Volvo, Delhi, Chandigarh	3,5,4,1,2	5,2,4,3,1	5,2,1,3,4	30	18
2	Hotel,	4,5,2,1,3	2,4,1,3,5	2,3,1,4,5	22	2

SNO	Query given in Google Search Engine	Relation based Ranking of first five URL' given by Google	Our Ranking for the URL's given by Google	Actual Ranking for the URL's given by Google	Variance by Relation based Ranking	Variance by Our Ranking
	Delhi, Airport					
3	Hotel, Rome, Historical center	1,3,2,4,5,	3,5,2,1,4	3,4,2,1,5	14	2
4	College, Delhi, MBA	1,3,4,2,5	5,1,3,2,4	3,1,5,2,4	10	8

From the results shown in Table 5.8 it has been found that in each case of the ranked order of web documents the variance computed by semantic rank model method is much smaller further showing its superiority. Thus, it can be said after the analysis of the results obtained that the ranking order of semantic web documents given by the semantic rank seen that the results shown by *our approach gives better ranking to the web pages according to the user query relevancy.*

However, the computational complexity of the semantic rank model is due to the calculation of the joint probability of the web page with respect to the user query and the underlying ontology. Also, as per the requirement of the technique the user need to enter the query as collection of keywords and their concepts need to be selected by the user from a set of concepts available which is a time consuming method. But, still the result set extracted from the described ranking method are relevant in terms of semantics and they meet the need of the user to the maximum level which overcome the limitation of computational complexity and time.

5.6 SUMMARIZATION OF RANKING TECHNIQUES

The layered semantic web architecture provides various means of strategies for improving search techniques and retrieving the relevant web pages as per the needs of the web user. The efficient web page ranking method additionally improves the searching of relevant web pages. Many ranking algorithms have been given using different approaches of computing similarity and that also make use of the semantic annotations which technically deals with ontology-based concepts and relations. The ranking model presented and discussed in this chapter deals with the concepts and relationships between the concepts available in the web page and query along with the domain deep information stored in a knowledgebase. The ranking model using semantic association deals with the concepts and relationships in a web documents in accordance with the concepts and relationships that are given by the user in the form of query. The proposed ranking further considers the semantic information that is available in the web pages with respect to the ontology knowledge structure which is stored and maintained corresponding to the same domain. The maximum score of both the comparisons indicates the semantic measure between the web pages or between the web page and the query for ranking applications. The score gives the basis to find the degree of association between two given text.

Similarly, the probability based semantic web page ranking approach discussed in this chapter capture the information stored in the form of ontology, query and web page to extract the web pages which are relevant to the user in respect to the intended query given by them to the search engine. In the probability computation based ranking scheme the web page significance is measured by computing the joint probability of web page with respect to the ontology and web page with respect to the query. The consideration of the probability computation of all concepts and relationships that are present in the ontology, web page, and the query would definitely lead to the true semantic analysis computation between two texts i.e. whether it is between two web pages or between the web page and the query.

CHAPTER VI

CONCLUSION AND FUTURE WORK

6.1 CONCLUSION

In this thesis, we have discussed a lot of research work related to the semantic analysis of the natural language information/content. From the application point of view, we have given a few techniques to compute semantic similarity between two given text/documents. We studied two types of similarity in detail i.e. the attributional similarity and the relational similarity. Various challenges faced in the field of semantic analysis of natural language for relevant and efficient information retrieval were also deeply analyzed. In Chapter 2, we discussed the methods available for the similarity detection by using lexical approaches. We have found that the research work presented in this field lacks in finding true similarity measure between two given documents, as these approaches are purely based on keywords present in documents. The consideration of relationships between the words somehow increases the similarity matching between the two texts, but the results of matching of two documents are not up to the expectations of users. The matching of relationships between the words present in the texts is again dependent on the techniques of lexical matching. Whereas, there may be the possibility that a set of words used by an author in one document, can be replaced by the set of synonyms for the same set of words due to which the similarity measure cannot be detected according to the expectation level of the human being.

To overcome this limitation, the work related to consideration of interrelated concepts was also discussed. The consideration of interrelated concepts is mainly done by using most common knowledge structure i.e. ontology which is the base of semantic web. In the techniques of semantic similarity computation by using the concepts and relationships, it is again found that the results produced by them are not according to the human analysis rating. This human rating is being produced by the sample set of individuals selected to express their views about the content of the set of documents and thus giving the order of the set of documents according to the particular query given to them. So, there is need to design new techniques or enhance the existing techniques to compute the semantic similarity between documents which can further be applied in many fields.

In Chapter 3, we have given two techniques for measuring the semantic similarity which considers the combination of the lexical matching with the concept matching. In one proposed technique, the matching is not only done in terms of keywords by constructing the VSM, but also in terms of relationships between the word pairs of VSM by constructing the RSM similar to VSM. Further, the concept matching is done by considering the concepts corresponding to each word in VSM and also their weighted relationships. These weighted relationships provide the importance of the whole entity i.e. concepts pairs and the weighted relationship importance with respect to a particular user query that too further in correspondence with the domain. In addition to this a novel technique for ranking of the web documents is given by using Genetic Algorithm. The use of genetic algorithm for measuring the semantic score of the web document helps in retrieving the relevant result set ranked according to the fitness function based on the relevant result-set obtained by human analysis rating. In the technique of finding the similarity between the two given texts by using genetic algorithm, the given text is analyzed and processed at two different level i.e. conceptual levels and the descriptive level. Each level score is modified by the two weight constants w1 and w2 which value is defined by making the use of genetic algorithm. The techniques of finding the similarity between web documents described in Chapter 3 helps in detecting the semantic similarity by considering the words and the relationship between these words that are available in both the web documents.

In Chapter 4, three more enhanced techniques are given according to the requirement of semantic analysis computation. The knowledge structure i.e. ontology again plays a vital role in extraction of the concepts along with the relationships between them from the web document. In this chapter, the techniques focus on the extraction of words from the web document and then the replacement of these extracted words with the set of probable concepts which are stored in the dictionary. The dictionary is called domain dictionary as the words and their respective set of concepts belong to a domain of computer science field. The extraction and replacement process helps in getting the semantic information associated with the web documents to some extent. Then, to make more semantic information available, the relationships that exist between the pair of concepts of web documents are considered by using the ontology.

These extracted relationships help in constructing the chains of connected concepts, which further helps in the development process of the ontology for a document which is called the document ontology. Further, the constructed document ontology is extended by using the recent trends that are available for a particular domain to add the implicit information so that more semantic information is extracted from the content of a document. After, construction of document ontology for each of the two web documents between which we want to compute the semantic similarity, the document ontologies are compared. Considering the computational cost and the complexity of the document ontology's comparison, the longest chains of the connected concepts obtained from each documents are compared. The common longest chain extracted from these document ontologies reflect the major or prime intension of the author based on which the similarity can be calculated between the web pages. The results obtained by the approaches discussed in this chapter also shows that the given approaches are helpful in getting the semantic information associated with each web documents and thus the similarity computation obtained from the techniques discussed give true semantic score.

In Chapter 5, we have given slightly different processing techniques of semantic similarity detection and for ranking of the web documents. The two methods given in this chapter are based on the computation of semantic similarity by considering the attributional and relational similarity measures. Specifically, the computation of attributional and then the relational similarity helps in improving the semantic similarity measurements. Additionally, a technique which is based on the probability calculation between the web page and the query, then between the web page and the ontology provides another way of considering the most relevant concepts and the relationships between them. In next, section the future perspective directions for further research in the field of IR are given.

6.2 FUTURE SCOPE

The idea of similarity that is perceived by human beings is not yet completely known from the processing aspects of machine. Many researchers in the field of cognitive science, neural network, fuzzy logic, machine learning, psycholinguistics etc. have tried to learn several aspects of human thinking and the ways of analyzing the things of real world. In all the fields various issues are considered related to human thinking.

Any individual thinking may not match with other individual but the source of factual knowledge is same for all the individuals. To develop the knowledge from this factual information comes through learning and experience.

The way through which the content is analyzed by human being is different from the available processing techniques as it is really a difficult task to inculcate the process of human thinking into a machine processing technique. This is because of the reason that it is again difficult to understand computationally that which part/parts of brain works while analyzing and understanding the things of real world.

Chomsky [111] defines that similarity detection is an inherent ability in an individual as according to him the language of an individual is already encoded in his/her brain by birth. The encoded language of an individual would basically depend on the environment of an individual and the experience. The LSA given by Landuer and Duamis [110] does not consider any external source of information like dictionaries constructed according to the knowledge and trends in a field. The LSA computes the similarity based on the content structure of the given documents only. However, from the NLP perspective the conceptual view of language/knowledge is very attractive. Some applications of NLP techniques are based on supervised datasets and some on the unsupervised datasets. Like, a person learning the things with the help of supervisor is approximated as the process of supervised learning and learning of things without the help of supervisor is unsupervised learning. Moreover, it is really a difficult task to process the huge amount of information present in WWW with the intention of semantic understanding of information as it would require large computations power. Therefore, it is believed that certain amount of source of information needs to be stored in some suitable formalism like ontology used presently. However, the processing of information depends on the analysis of the content and the use of knowledge stored while analyzing the content. This process of analyzing the content is similar to detection of similarity measure which indicates the intelligence of machine. For example, the exam like SAT for selection of candidates for U.S universities, include the word-analogy questions which is considered as base for several relational similarity computation algorithms. To solve the question of word-analogy the individual not only needs to understand the question but also he needs to analyze the relationships that exists in each word-pair to get the true answer of the given question. This analysis when computed from machine, the machine requires high artificial intelligence processing skills to be inculcated into it in some form of algorithms. The same is the case in the papers of IQ tests like detections of things/objects that have relation with other entities, pattern recognition, etc. conducted by different organizations for different purposes.

Computing various such questions which requires intelligence needs to measure the relational similarity along with the attribution similarity. One test which is widely known for testing the intelligence of a machine which further assist in measuring the relational similarity is the Turing test given by Alan Turing [112]. In Turing test approach a human being cannot differentiate between the result which is produced by either a computer program or an individual, in this case the computer program is considered to be an intelligent program. From above discussions it can be concluded that embedding the human intelligence into a machine is difficult task which involve various field work like NLP, Machine Learning, Artificial Intelligence, Neural Network etc.

The main disadvantage of the keyword based search engine is its lack of ability to evaluate the relationships between the words present in query and further present in the documents. Semantic Search Engines are developed as extreme requirement of solution to this inability of traditional search engines. Semantic Search Engines application is still a challenging task due to several reasons like annotation of web documents, manage changes in web documents, level of semantic annotations on which the relevancy of retrieval of information depends, processing of RDF using different data structures etc. Even though the work in the field of retrieving relevant information from WWW by applying semantic analysis from the stages of data representation to similarity measure computation is vital. In each stage of IR the contributions can be enhanced and refined by modifying the techniques for efficient and effective semantic similarity computation between two given texts. Some of the extensions that can be done in the work related to semantic analysis of information of a web page which is near to the approach of analysis of human thinking are as follows:

- **Extension to the knowledge structure:** Most of the semantic similarity techniques makes the use of structured knowledge base i.e. ontology. This structured ontology is used for effective retrieval of the concepts and the relationships between them. The knowledge base created is related to a domain which needs to be modified with the changes involved in the information available according to the changes in the real world. There should also be the means of constructing this knowledge structure automatically which is capable of inculcating the changes in the information as and when required. This modification in the knowledge base is a vital task as it involves the intelligent system learning techniques which are not easy to embed.

- **Extension to the proposed semantic similarity techniques:** In one of our proposed technique the concepts and relationships retrieved are used to construct the chain of interrelated concepts. According to our assumption, the longest chain of interrelated concepts from a document represents the prime intention of the author. Also, due to computational complexity we have considered the longest chain of interrelated concepts for semantic score computation between any two web documents. Although, this work can be further extended by considering all possible interrelated concepts chains so that the secondary intention of the author can also be considered. The work done in another proposed technique of computation of semantic similarity using genetic algorithm is evaluated theoretically which can further be implemented to measure the performance of the same. Further, the performance of the proposed techniques has been compared with the basic approach of similarity computation. The other techniques of similarity detection can be considered to evaluate the proposed techniques performance but for that we need to consider the large corpus of web documents and also to enrich the base ontology as per the domain.
- **Empirical evaluation using benchmark dataset**s: The proposed techniques performance can also be measured by using the well-known benchmark datasets like M&C data set [26] which is a subset of Rubenstein-Goodenough's [27], WordSim203 [28] which is a subset of Wordsim353 [29]. The concepts pairs present in a dataset can also be classified on the basis of level of similarity like extremely similar, extremely different, moderately similar, moderately different, not analyzed etc. These classes of concept pairs in a benchmark datasets can further be translated to an equivalent numerical

similarity score which is used to compute the semantic similarity between interrelated concepts pairs.

- **Semantic Similarity application**s: There are many application areas where the concept of similarity detection on the basis of semantics is required like detection of duplicate pages, ranking of documents, crawling of document by search engine, indexing of documents by a search engine, plagiarism detection, etc. In our research work, the application of semantic similarity detection has been considered in the field of ranking of web documents by a search engine. However, the given techniques of semantic similarity computation between two web documents can be applied to any phase of IR or even it can be used to organize the information resource center i.e. semantic web. These algorithms can be modified according to the requirement of IR so that the results-sets retrieved for a user are incredible fulfilling all or maximum requirements.

The above future directions for research are limited as there are several ways to extend this work as per the demand or requirement of IR. All of the above directions aim in improving the result set extracted from WWW by a search engine as per the specified query. The approach is same in all of the techniques and i.e. semantic analysis. This could be done by exploiting existing techniques or developing new ones.

REFERENCES

[1] Amit Pisharody , Howard E Michel, “Search Engine Technique Using Keyword Relations”, *Proceedings of International Conference on Artificial Intelligence (ICAI)*, Page(s) 300-306, 2005.

[2] Tim Berners-Lee, James Hendler and Ora Lassila, “The Semantic Web,” *Scientific Am.*, 2001.

[3] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka,” Measuring the similarity between implicit semantic relations from the web”, *International World Wide Web Conference Committee (IW3C2), ACM*, Page(s) 651-659, 2009.

[4] Fabrizio Lamberti, Andrea Sanna and Claudio Demartini, “A relation-based Page Rank algorithm for semantic web search engines”, *IEEE Transaction on Knowledge and Data Engineering*, Volume 21, No. 1, Jan 2009.

[5] Giannis Varelas , Epimendinis Voutsakis, Paraskevi Raftpoulou, Euripides G. M., and Evangleos E. Milios., ” Semantic similarity methods in WorldNet and their applications to information retrieval on the web”, WIDM *Proceedings of the 7th annual ACM International Workshop on Web Information and Data Management, ACM Transactions,* Bermen, Germany, Page 10, 2005.

[6] Oleshchuk V. and Pederson A., “Ontology Based Semantic Similarity Comparison of Documents”, *Proceedings of IEEE 14th Workshop on Database and Expert Systems Applications*, Page(s) 735-738, ISSN 1529-4188, 2003.

[7] Quan Thanh Tho, Siu Cheung Hui, and Tru Hoang Cao, “Automatic Fuzzy Ontology Generation for Semantic Web”, *IEEE Transactions on Knowledge and Data Engineering*, Volume 18, No. 6, June 2006.

[8] Behnam Hajian and Tony White, “A Method of Measuring Semantic Similarity using a Multi-tree Model” *Proceedings IJCAI 9th Workshop on Intelligent Techniques for Web Personalization & Recommender Systems (ITWP'11),* Barcelona, Spain, 16 July 2011.

[9] Jan Kasperzak., “Systems for Discovering Similar Documents”, *Ph.D. Thesis proposal, Faculty of informatics Masaryk university*, September 2009.

[10] Peter D. Turney, and Patrick Pantel.,” From Frequency to Meaning: Vector Space Models of Semantics”, *Journal of Artificial Intelligence Research*, Page(s) 41–188, 2010.

[11] Jan Kasprzak, Michal Brandejs, Miroslav K ripac, Pavel Smerk, ”Distributed System for Discovering Similar Documents”, *ICEIS Proceedings of 10th International Conference on Enterprise Information System*, Volume 3 DISI, Page(s) 437-440, Setubal, Portugal, 2008.

[12] Sheetal A Takale and Sushma S. Nandgaonkar, “Measuring Semantic Similarity between Words Using Web Documents”, *International Journal of Advanced Computer Science and Applications, IJACSA*, Volume 1, Issue 4 October, 2010.

[13] James W. Cooper, Anni R. Coden, and Eric W. Brown, “Detecting Similar Documents Using Salient Terms”, *ACM transactions CIKM’02*, Mclean, USA. November 4-9, 2002.

[14] Thomas R. Gruber,” Toward Principles for the Design of Ontologies Used for Knowledge Sharing”, International Journal Human-Computer Studies 43, Page(s) 907-928, August 23, 1993.

[15] Dragomir R. Radev, Jahna Otterbacher, Hong Qi, and Daniel Tam “Mead reducs: Michigan at duc 2003”, *In Proceedings of DUC*, Edmonton, AB, Canada, 2003.

[16] Dnyanesh Rajpathak and Rahul Chougule, “A Generic Ontology Development Framework for Data Integration and Decision Support in a Distributed Environment”, *International Journal of Computer Integrated Manufacturing*, Volume 24, Page(s) 154-170, Issue 2, 2011.

[17] Priti Sriniwas Sajja and Rajender Akerker, “Intelligent Technology for Web Applications”, *CRC Press*, International Standard Book Number- 978-1-4398-7162-1.

[18] Mihalcea, R., Moldovan, D.I., “An Automatic Method for Generating Sense Tagged Corpora”, *In Proceedings of the 16th National Conference on Artificial Intelligence*, AAAI Press, 1999.

[19] Martin Volk, “Using the Web as Corpus for Linguistic Research” In Catcher of the Meaning Pajusalu, R., Hennoste, T. (Eds.). Department of General Linguistics 3, University of Tartu, Germany, 2002.

[20] Christiane FellBaum, “WordNet A Lexical Database”, Massachuests Institute of Technology”, International Standard Book Number- 0-262-06197-X, 1998.

[21] David Sanchez and Antonio Moreno, “Creating Ontologies from Web Documents”, Department of Computer Science and Mathematics, University Rovira I Virgili, Tarragona, 2004.

[22] Van Dam, K.H. and Lukszo, Z., "Modelling Energy and Transport Infrastructures as a Multi-Agent System using a Generic Ontology", *IEEE International Conference on Systems, Man and Cybernetics SMC*, Volume 1, Page(s) 890-895, 2006.

[23] Shaojie Qiao Tianrui Li ; Hong Li ; Yan Zhu ,Jing Peng and Jiangtao Qiu "SimRank: A Page Rank Approach based on Similarity Measure", *In Proceedings of 10th International Conference On Semantic Web,* IEEE, 2010.

[24] Yufei Li, Yuan Wang, and Xiaotao Huang, "A Relation-Based Search Engine in Semantic Web", *IEEE Transaction Knowledge and Data Engineering,* Volume 19, No. 2, Page(s) 273-282, February 2007.

[25] Mei Kobayashi and Koichi Takeda, "Information Retrieval on the Web", *ACM Computing Surveys,* Volume 32, No. 2, Page(s) 144-173, June 2000.

[26] Oleshchuk V., and Asle P., "Ontology Based Semantic Similarity Comparison of Documents", *In Proceedings of IEEE 14th Workshop on Database and Expert Systems Applications,* 2003.

[27] Rajesh Thiagarajan, Geetha Manjunath, and Markus Stumptner, "Computing semantic similarity using ontologies", *International Semantic Web Conference (ISWC),* Karlsruhe, Germany, 2008.

[28] Yuhua Li, Zuhair Bandar, David McLean and James O'Shea, "A method for measuring sentence similarity and its application to conversational agent", *In Proceedings Of 17th International Conference FLAIRS*, Florida, USA, AAAI Press, 2012.

[29] Yin Guisheng and Sheng Qiuyan," Research on ontology-based measuring semantic similarity", *International Conference on Internet Computing in Science and Engineering,* Page(s) 250-253, 2008.

[30] Jun Fang, Lei Guo, XiaoDong Wang and Ning Yang, "Ontology-based automatic classification and ranking for web documents", *Proceedings Of 4th IEEE International Conference On Fuzzy Systems and Knowledge Discovery,* Volume 3, 2007.

[31] Shahrul Azman Noah1 , Lailatulqadri Zakaria1 and Arifah Che Alhadi, "Extracting and modelling the semantic information content of web documents to support semantic document retrieval", *Proceedings Of 6th Asia-Pacific Conference On Conceptual Modelling*, Volume 96, Page(s) 79-86, 2009.

[32] Aleman Meza B., Budak Arpinar, Mustafa Nural, and Amit Sheth, "Ranking Documents Semantically Using Ontological Relationships", *Proceedings of IEEE 4th International Conference on Semantic Computing*, Page(s) 299-304, 2010.

[33] Jun Fang, Lei Guo, and Yue Niu, "Documents Classification by using Ontology Reasoning and Similarity Measure", *Proceedings Of 7th IEEE International Conference On Fuzzy Systems and Knowledge Discovery (FSKD),* Volume 4, Page(s) 1535-1539, 2010.

[34] Silva F., Girardi R. And Lucas D.," An Information Retrieval Model for the Semantic Web", *Proceedings of 6th International Conference on Information Technology: New Generation,* Page(s) 143-148, 2009.

[35] Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka, "A Web Search Engine based Approach to Measure Semantic Similarity between Words", *IEEE Transactions on Knowledge and Data Engineering,* Volume 23, No. 7, July 2011.

[36] Vincent Schickel-Zuber and Boi Faltings, "OSS: A Semantic Similarity Function based on Hierarchical Ontologies", *Proceedings of 20th International Joint Conference On Artificial Intelligence,* Page(s) 551-556, 2007.

[37] Juhum Kwon, O-Hoon Choi, Chang-Joo Moon, Soo-Hyun Park, and Doo-Kwon Baik, "Deriving Similarity for Semantic Web using Similarity Graph", *Journal of Intelligent Information System,* Volume 26, Issue 2, Page(s) 149-166, 2006.

[38] Peter D. Turney, "Measuring Semantic Similarity by Latent Relational Analysis", *Proceedings of 19th ACM International Conference on Artificial Intelligence (IJCAI),* Page(s) 1136-1141, 2005.

[39] Dino Isa, Lam Hong, V. P. Kallimani, and R. Rajkumar," Text Document Pre-Processing Using the Bayes Formula for Classification Based on the Vector Space Model", *Journal of Computer and Information Science*, Volume 1, No. 4, 2008.

[40] Vikram Singh and Balwinder Saini, "An Effective Pre-Processing Algorithm For Information Retrieval Systems", *International Journal of Database Management Systems (IJDMS),* Volume 6, No. 6, Page(s) 13-24, December 2014.

[41] Ronald R. Yager, “A Hierarchical Document Retrieval Language”, *Information Retrieval*, Volume 3, Page(s) 357-377, Kluwer Academic Publishers, 2000.

[42] Samuel Fernando and Mark Stevenson,” A Semantic Similarity Approach to Paraphrase Detection*”, Proceedings of the 11th Annual Research Colloquium of the UK Special Interest Group for Computational Linguistics,* 2009.

[43] Nicola Guarino, Daniel Oberle, and Steffen Staab, “What is an Ontology?”, Handbook on Ontologies, *International Handbooks on Information Systems*, DOI 10.1007/978-3-540-92673-3, Springer-Verlag Berlin Heidelberg 2009.

[44] Rekha Baghel and Renu Dhir, “A Frequent Concepts Based Document Clustering Algorithm”, *International Journal of Computer Applications*, ISSN 0975 – 8887, Volume 4, No. 5, July 2010.

[45] Kishor Wagh and Satish Kolhe, “Information Retrieval Based on Semantic Similarity Using Information Content”, *International Journal of Computer Science Issues (IJCSI),* ISSN (Online): 1694-0814, Volume 8, Issue 4, No 2, July 2011.

[46] Ngoc-Diep Ho and Fairon Cédrick, “Lexical Similarity based on Quantity of Information Exchanged - Synonym Extraction “, *International Conference RIVF*, February 2-5, Hanoi, Vietnam, 2004.

[47] JungAe Kwak and Hwan-Seung Yong, “Ontology Matching Based On Hypernym, Hyponym, Holonym, and Meronym Sets in WordNet”, *International Journal of Web & Semantic Technology (IJWesT),* Volume 1, No. 2, April 2010.

[48] Feiyu Lin and Kurt Sandkuhl, ”A Survey of Exploiting WordNet in Ontology Matching”, *International Federation for Information Processing IFIP*, Volume 276, Artificial Intelligence and Practice II, Max Bramer, (Boston: Springer), Page(s) 341–350, 2008.

[49] Nicholas J. Belkin and Bruce Croft, “Information Filtering and Information Retrieval: Two sides of the same coin?” *Communication of ACM-Special issue on information filtering*, Volume 35, Issue 12, Dec 1992.

[50] Ming Che Lee, Jia Wei Chang and Tung Cheng Hsieh, “A Grammar Based Semantic Similarity Algorithm for Natural Language Sentences”, *Research Article in Scientific World Journal,* Volume 20, Page(s) 17, 2014.

[51] Hongzhe Liu and Pengfe Wang, "Assessing Text Semantic Similarity using Ontology", *Journal of Software*, Volume 9, No. 2, February 2014.

[52] Mingxin Gan, Xue Dou and Rui Jiang, "From Ontology to Semantic Similarity: Calculation of Ontology Based Semantic Similarity", *Scientific World Journal*, Page(s) 11, 2013 .

[53] Yuxin Mao," A Semantic-based Genetic Algorithm for sub-ontology evolution", *Information Technology Journal,* 9(4), ISSN1812-5638, Page(s) 609-620, 2010.

[54] Razib M. Othman , Safaai Deris , Rosli M. Illias , Hany T. Alashwal and Rohayanti Hassan, "Incorporating Semantic Similarity Measure in Genetic Algorithm: An Approach for Searching the Gene Ontology terms", *International Journal of computational intelligence*, Volume 3, No. 3, 2006.

[55] Wang Wei , Payam M. Barnaghi and Andrzej Bargiela, "Search with Meanings: An Overview of Semantic Search Systems", *School of Computer Science,* University of Nottingham Malaysia.

[56] Mehrnoush Shamsfard , Azadeh Nematzadeh and Sarah Motiee "ORank: An Ontology based System for Ranking Documents", *International Journal of Computer Science*, Volume 1, No 3, ISSN 1306-4428, 2006.

[57] Sergey Brinand and Lawrence Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine", *Proceeding of 7th International Conference on World Wide Web (WWW)*, Page(s) 107-117, 1998.

[58] Li Ding, Pranam Kolari, Zhongli Ding, and Sasikanth Avancha, "Using Ontologies in the Semantic Web: A Survey", *Ontologies Integrated Series of Information Systems*, Volume 14, Page(s) 79-113, Springer, 2007.

[59] Ling Song, Jun Ma, Hui Liu, Li Lian, and Dongmei Zhang, "Fuzzy Semantic Similarity between Ontological Concepts", *Advances and Innovations in System Computing Sciences and Software Engineering,* Springer, Page(s) 275-280, 2007.

[60] Shixiong Xia, Zuhui Hu and Qiang Niu, "An Approach of Semantic Similarity Between Ontology Concepts Based on Multi Expression Programming ", *IEEE Sixth WISA Conference Web Information System and Applications*, Page(s) 184-188 , 2009.

[61] Aditi Sharan and Manju L. Joshi," An Algorithm for Finding Document Concepts using Semantic Similarities from WordNet Ontology" *International*

Journal of Computational Vision and Robotics, Volume 1, No 2, Page(s) 147-157, 2010.

[62] May Sabai Han, "Semantic Information Retrieval Based on Wikipedia Taxonomy", *International Journal of Computer Applications Technology and Research*, Volume 2, Issue 1, ISSN 2319-8356, Page(s) 77-80, 2013.

[63] Zin Thu Thu Myint and Kay Khaing Win, "Conceptual Similarity Measurement Algorithm for Domain Specific Ontology", *International Journal of Information Technology of Information Technology, Modelling and Computing (IJITMC),* Volume 2, No. 2, May 2014.

[64] Madalina Croitoru, Bo Hu, Srinandan Dashmapatra, Paul Lewis, David Dupplaw, and Liang Xiao," A Conceptual Graph Based Approach to Ontology Similarity Measure", *Springer ICCS*, Page(s) 154-164, 2007.

[65] Xiquan Yang , Ye Zhang, Na Sun, and Deran Kong, "Research on Method of Concept Similarity Based on Ontology", *Proceedings of International Symposium on Web Information Systems and Applications (WISA),* Page(s) 132-135, 2009.

[66] Guenther Goerz and Martin Scholz, "Adaptation of NLP Techniques to Cultural Heritage Research and Documentation", *Journal of Computing and Information Technology - CIT 18*, Volume 4, Page(s) 317–324, 2010.

[67] R. Subhashini and J. Akilandeswari, "A Survey on Ontology Construction Methodologies", *International Journal of Enterprise Computing and Business Systems*, Volume 1, Issue 1, January 2011.

[68] Tere Aaberge and Rajendra Akerkar, "Ontology and Ontology Construction: Background and Practices", *International Journal of Computer Science and Applications*, Volume 9, No. 2, Page(s) 32 – 41, 2012.

[69] Thomas R. Gruber, "Towards Principles for the Design of Ontologies Used for Knowledge Sharing", *International Journal of Human-Computer Studies*, 43(5/6), Page(s) 907-928, 1995.

[70] Jens Graupmann, Ralf Schenkel Gerhard Weikum, "The Sphere Search Engine for Unified Ranked Retrieval of Heterogeneous XML and Web Documents", *In Proceeding of 31st International Conference on Very Large Data Bases, Trondheim, Norway,* Page(s) 529-540, 2005.

[71] Hsinchun Chen, Lynch, K.J., Basu, K., and Ng, T.D., ”Generating, Integrating and Activating Thesauri for Concept-Based Document Retrieval”, *IEEE Intelligent Systems*, Volume 8, Issue 2, Page(s) 25-35, 1993.

[72] R.F. Mihalcea and S.I. Mihalcea, “Word Semantics for Information Retrieval: Moving One Step Closer to the Semantic Web”, *In Proc. 13th International Conference on Tools with Artificial Intelligence*, Dallas, Texas, Page(s) 280-287, 2001.

[73] Cartic Ramakrishnan, Krys J. Kochut, and Amit P. Sheth, “A Framework for Schema-Driven Relationship Discovery from Unstructured Text” *In Proc. Fifth International Semantic Web Conference, Athens, Georgia,* Page(s) 583-596, 2006.

[74] Arijit Sengupta, Mehmet Dalkilic, and James Costello, “Semantic Thumbnails: A Novel Method for Summarizing Document Collections” *In Proceedings of 22nd Annual International Conference on Design of Communication ACM The engineering of Quality Documentation,* Memphis, Tennessee, Page(s) 45-51, 2004.

[75] R.V. Guha, Rob McCool, and Eric Miller, “Semantic Search”, In Proceedings of 12th *International World Wide Web Conference ACM*, Budapest, Hungary, Page(s) 700-709, 2003.

[76] Boanerges Aleman-Meza, Christian Halaschek-Wiener, I. Budak Arpinar, Cartic Ramakrishnan, and Amit Sheth, “Ranking Complex Relationships on the Semantic Web”, *IEEE Internet Computing*, Volume 9, Issue 3, Page(s) 37-44, June 2005.

[77] Kemafor Anyanwu, and Amit Sheth, “ρ-Queries: Enabling Querying for Semantic Associations on the Semantic Web ACM”, *In Proceedings of 12th International World Wide Web Conference*, Budapest, Hungary, Page(s) 690-699, 2003.

[78] Amit P. Sheth, “From & Integration to Analytics”, *In Proceedings of Semantic Interoperability and Integration,* IBFI, Schloss Dagstuhl, Germany, 2004.

[79] E. Miller, “The Semantic Web is Here”, *In Keynote at the Semantic Technology Conference*, San Francisco, California, USA, 2005.

[80] Y.L. Lee, “Apps Make Semantic Web a Reality”, *SD Times*, 2005.

[81] R. Lempel and S. Moran. “The Stochastic Approach for Link-structure Analysis (SALSA) and the TKC e®ect”, *In the 9th International WWW Conference*, Volume 33, Page(s) 387-40, May 2000.

[82] E. Greengrass, “Information Retrieval: A Survey”. *DOD Technical Report TR-R52-008-001*, November 2000.

[83] Cosma G. and Joy Mike, “An Approach to Source-Code Plagiarism Detection and Investigation using Latent Semantic Analysis”, *IEEE Transaction on Computers*, Volume 61, Issue 3, Page(s) 379-394, March 2012.

[84] Iosif E., and Potamianous A., “Unsupervised Semantic Similarity Computation Between Terms using Web Documents”, *IEEE Transaction on Knowledge and Data Engineering*, Volume 22, Issue 11, Page(s) 1637-1647, November 2010.

[85] Li Zhanjun, and Karthik Ramani, “Ontology-based Design Information Extraction and Retrieval”, *Artificial Intelligence for Engineering Design, Analysis and Manufacturing,* Volume 21, Page(s) 137-154, 2007.

[86] Lei Y., Uren V., and E. Motta, “SemSearch: A Search Engine for the Semantic Web”, *Proceeding Of 15th International Conference on Managing Knowledge in a World of Networks (EKAW)*, Page(s) 238-245, 2006.

[87] Asuncion Gomer-Perez and Oscar Corcho, “Ontology Languages for the Semantic Web,” *IEEE Intelligent Systems*, Volume 17, Issue 1, Page(s) 54-60, Jan-Feb 2002.

[88] Rudi L. Cilibrasi and Paul M. B Vitanyi, “The Google Similarity Distance”, *IEEE Transactions on Knowledge and Data Engineering,* Volume 19, Issue 3, March 2007.

[89] Gene H. Golub, Franklin T. Luk, and A. Lanczos, “Method of Computing the Singular Values and Corresponding Singular Vectors of a Matrix”, ACM *Transactions on Mathematical Software,* Volume 7, Issue 2, Page(s) 149-169, 1981.

[90] Kemafor Anyanwu, Angela Maduko, and Amit Sheth, “SemRank: Ranking Complex Relation Search Results on the Semantic Web,” *Proceeeding Of 14th International Conference on World Wide Web (WWW) ACM*, Page(s) 117-127, 2005.

[91] Li Ding, Tim Finin, Anupam Joshi, Rong Pan, R. Scott Cost, Yun Peng, Pavan Reddivari Vishal Doshi and Joel Sachs, “Swoogle: A Search and Metadata

Engine for the Semantic Web", *Proceeding Of 13th ACM International Conference on Information and Knowledge Management (CIKM),* Page(s) 652-659, 2004.

[92] Hyunjung Park, Sangkyu Rho and Jinsoo Park, "A Link-based Ranking Algorithm for Semantic Web Resources: A class oriented approach independent of link direction", Volume 22, Issue 1, Page(s) 1-25, 2011.

[93] Nenad Stojanovic, Rudi Studer, and Ljiljana Stojanovic, "An Approach for the Ranking of Query Results in the Semantic Web", *Proceeding Of 2nd International Conference on Semantic Web (ISWC),* Page(s) 500-516, 2003.

[94] Tim Berners-Lee and M. Fischetti , "Weaving the Web", *Harper Audio*,1999.

[95] Page L., S. Brin, R. Motwani, and T. Winograd, "The Page Rank Citation Ranking: Bringing Order to the Web", *Stanford Digital Library Technologies Project*, 1998.

[96] Web Ontology Language, http://www.w3.org/2004/OWL/, 2004.

[97] Sanjay Ghemawat, Howard Gobioff, and Shun Tek Leung, "The Google File System", In *Proceedings of ACM Symposium on Operating Systems Principles*, Volume 37, Issue 5, Page(s) 29-43, 2003.

[98] Jeffrey Dean and Sanjay Ghemawat, "Mapreduce: Simplified Data Processing on Large Clusters", *In Proceeding of 6th Symposium on Operating System Design and Implementation (OSDI)*, 2004.

[99] Adam Kilgarriff, "Googleology is bad Science", *Computational Linguistics*, Volume 33, Issue 1, Page(s) 147–151, 2007.

[100] Pushpa C N, Thriveni J, Venu Gopal K R, and L M Patnaik, "Web Search Engine Based Semantic Similarity Measure Between Words Using Pattern Retrieval Algorithm", *IEEE Computer Science and Information Technology*, Page(s) 1-11, 2013.

[101] Nuno Vasconcelos, "On the Efficient Evaluation of Probabilistic Similarity Functions for Image Retrieval", *IEEE Transactions on Information Theory,* Volume 50, Issue 7, Page(s) 1482-1496, 2004.

[102] Peipei Li, Haixun Wang, Kenny Q. Zhu, Zhongyuan Wang, Xuegang Hu, and Xindong Wu, "A Large Probabilistic Semantic Network based Approach to Compute term Similarity", *IEEE Transaction of Knowledge and Data Engineering*, Volume 27, No. 10, 2015.

[103] Eduardo Blanco and Del Moldovan," A Semantic Logic based Approach to Determine Textual Similarity", *IEEE/ACM Transactions on Audio, Speech and Language Processing*, Volume 23, No. 4, Page(s) 683-693, April 2015.

[104] Nuno Vasconcelos, "On the Efficient Evaluation of Probabilistic Similarity Functions for Image Retrieval", *IEEE Transactions on Information Theory*, Volume 50, Issue 7, Page(s) 1482-1496, 2004.

[105] Costas P. Pappis and Nikos I. Karacapilidis, "A Comparative Assessment of Measures of Similarity of Fuzzy Values", *Fuzzy Sets and Systems*, Volume 56, Issue 2, Page(s) 171-174, 1993.

[106] Yuxin Peng, Chong Wah Ngo., Cuihua Fang, Xiaoou Chen, and Jianguo Xia, "Audio Similarity Measure by Graph Modelling and Matching" *In Proceedings of the 14th Annual ACM international Conference on Multimedia*, Page(s) 603-606, 2006.

[107] Xiaojun Wan and Yuxin Peng "A Measure Based on Optimal Matching in Graph Theory for Document Similarity", *Information Retrieval Technology LNCS 3411, Springer* Berlin / Heidelberg ISSN 0302-9743 (Print) 1611-3349 (Online), Pg(s) 227-238, DOI: 10.1007/b106653, 2005.

[108] Horst Bunke, "Graph matching: Theoretical Foundations, Algorithms, and Applications", *In Proceedings of Vision Interface*, Montreal, Page(s) 82–88, 2000.

[109] Bila Zaka, "Theory and Applications of Similarity Detection Techniques", *Dissertation, Institute for Information System and Computer Media*, Graz University of Technology, Austrai, February 2009.

[110] Thomas K. Landauer and Susan T. Dumais., "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of knowledge", *Psychological Review*, 104(2):211–240, April 1997.

[111] Soumen Chakrabarti, "Mining the Web: Discovering Knowledge from Hypertext Data", Morgan Kaufmann Publishers, ISBN 1-55860-754-4, 2003.

[112] A. M. Turing, "Computing Machinery and Intelligence", *Mind* 49, Page(s) 433-460, 1950.

[113] Khaled Shaban, "A Semantic Graph model For Text Representation And Matching in Document Mining", *Ph.D Thesis in Electrical and Computer Engineering Department*, Waterloo, Ontario, Canada, 2006.

[114] Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse and Geoffrey Zweig, "Syntactic Clustering of the Web," *Proceedings of the Sixth WWW Conference*. Santa Clara, CA, 1997.

[115] Michael O. Rabin, "Fingerprinting by Random Polynomials", *Centre for Research in Computing Technology,* Harvard University, Report TR-15-81, 1981.

[116] D. Minnie and S. Srinivasan, "Multi Domain Meta Search Engine with an Intelligent Interface for Efficient Information Retrieval on the Web", *In Proceedings of First International Conference on Computer Science Engineering and Information Technology*, *CCSEIT, Springer,* Tirunelveli, TamilNadu, Volume 204(CCIS), Page(s) 121-129, September 2011.

[117] Georgina Cosma and Mike Joy, " An Approach to Source-code plagiarism detection and investigation using latent semantic analysis", *IEEE Transactions on Computers*, Volume 61, No. 3, Page(s) 379-394, March 2012.

[118] Ziv Bar-Yossef and Maxim Girevich, "Efficient Search Engine Measurement", *ACM Transactions on Web*, Volume 5, Issue 4, Article 18, 48 Page(s), October 2011.

[119] Weifeng Su, Hejun Wu, Yafei Li, Jing Zhao, Fredrick, Hongmin and Tianqiang Huang, "Understanding Query Interfaces by Statistical Parsing", *ACM Transactions on Web*, Volume 7, Issue 2, Article 8, 22 Page(s), May 2013.

[120] Hung-Husuan Chen and C. Lee Giles, "Ascoss++: An Asymmetric Similarity Measure for Weighted Network to Address the Problem of SimRank", ACM *Transactions on Knowledge Discovery and Data,* Volume 10, Issue 2, Article 15, 26 Page(s), October 2015.

[121] Ronald Fagin, Benny Kimelfeld, Fredrick Reiss, and Stijn Vansumeren, "Document Spanners: A Formal Approach to Information Extraction", *Journal of ACM,* Volume 62, Issue 2, Article 12, 51 Page(s), April 2015.

APPENDIX I

Table 1.1 gives the sample content of set of 50 documents related to domain Artificial Intelligence. These set of documents were used for processing during implementation of our proposed semantic similarity based techniques to analyze the results given by our proposed techniques in comparison to other existing techniques as discussed in Chapter 3, Chapter 4 and Chapter 5.

Table 1.1: Set of 50 Documents Related to Artificial Intelligence Domain

Document Number	Document Content
D1	Artificial intelligence is the intelligence of machine and robot and the branch of computer science that aims to create it.
D2	Artificial intelligence textbook define the field as study and design of intelligent agent where an intelligent agent is system that perceives its environment and takes action that maximizes its chance of success.
D3	Knowledge representation and knowledge engineering are central to artificial intelligence research. Many of the problems machines are expected to solve will require extensive knowledge about the world.
D4	Intelligent agent must be able to set goal and achieve them. They need a way to visualize future and be able to make choices that maximizes the utility of available courses.
D5	Machine learning is central to artificial intelligence research. It is study of computer algorithm that improves automatically through experience.
D6	Natural Language processing gives machine the ability to read and understand the languages that human speak.
D7	Artificial intelligence is the area of computer science focusing on creating machine that can engage on behavior that human consider intelligent.
D8	Artificial intelligence is branch of computer science concerned with making computers behave like humans.
D9	Artificial intelligence includes game playing, expert system, natural language, neural network, and robotics. Currently no computer exhibit

Document Number	Document Content
	full artificial intelligence.
D10	Applications of artificial intelligence robots that plan their own actions, web crawlers that efficiently locate information, intelligent assistant that help humans defect financial fraud and game playing system that perform better than human player.
D11	Artificial intelligence track focuses on fundamental mechanism that enable the construction of intelligent system that can operate autonomously, learn from experience, plan their actions and solve complex problems.
D12	Artificial intelligence covers key challenges in computing such as how to represent human knowledge and mechanize thought process, how to use computational model to understand, explain and predict complex behavior of individual or group and how to make computer as easy to interact with as people.
D13	Intelligence is ability to think to imagine, to create, memorize, understand, recognize pattern, make choice, adapt to changes and learn from experience.
D14	Intelligence is the capacity to learn and solve problems. In particular it is ability to solve novel problems, to act rationally, to act like humans.
D15	Artificial intelligence involves ability to interact with real world which is to perceive, understand and act.
D16	Artificial intelligence includes reasoning and planning which is ability to deal with unexpected problem and uncertainties, solving new problem, planning and making decisions.
D17	Artificial intelligence also includes learning and adaptation. The internal models used are always being updated.
D18	Artificial intelligence has made substantial progress in recognition and learning, some planning and reasoning problem and many open research problems.
D19	Artificial branches include logical artificial intelligence, search, pattern recognition, representation, inference, common sense knowledge and

Document Number	Document Content
	reasoning, learning, planning, ontology, heuristic and genetic programming.
D21	Weak artificial intelligence refers to technology that is able to manipulate predetermined rules and apply the rules to reach a well defined goal.
D22	Strong artificial intelligence refers to technology that has the ability to think cognitively or is able to function in a way similar to human brain.
D23	Medical artificial intelligence is primarily concerned with construction of program that diagnosis and make therapy recommendation.
D24	A new study says that human are much better at controlling traffic in urban areas than current computer system, leading to development of new ones based on artificial intelligence.
D25	Artificial intelligence researchers have developed several specialized programming for artificial intelligence such as LISP, PROLOG, STRIPS, etc.
D26	Artificial intelligence applications are also often written in standard language like C++, MATLAB and LUSH.
D27	There are primarily two computer language used in artificial intelligence work LISP and PROLOG.
D28	Artificial intelligence is the ability of digital computer or computer controlled robot to perform task commonly associated with intelligent being.
D29	The ethics of artificial intelligence is the part of ethics of technology specific to robots and other artificial intelligent beings.
D30	Artificial intelligence combines science and engineering in order to build machine capable of intelligent behavior.
D31	Artificial intelligence as engineering is the system that often thought of as science fiction but in fact is all around us.
D32	Artificial intelligence as science helps us to answer the questions like what is intelligence and how it works.
D33	Social intelligence is the ability to get along with other, knowledge of

Document Number	Document Content
	social matters, and insight into words or underlying personality facts for others.
D34	Artificial intelligence is computational part of the ability to achieve goals in the world.
D36	Artificial intelligence is the use of computers to model the behavioral aspect of human reasoning and learning.
D37	Artificial intelligence is the art of making computers do smart things by using soft-computing instead of using traditional hard computing.
D38	Artificial intelligence lets computer learn things and ask questions with the help of fuzzy inference system.
D39	Human intelligence is the ability of humans to combine several cognitive processes to adopt the environment. Artificial intelligence is the field dedicated to developing machine that will be able to minimize and perform as humans.
D40	Human intelligence is defined as the quality of mind that is made up of capabilities to learn from past experience, adaptation to new situations, handling of abstract ideas and ability to change individual environment using gained knowledge.
D41	Artificial intelligence is the field of computer science dedicated to developing machine that will be able to perform same task as human world.
D42	Machine learning deals with designing and developing algorithm to evolve behavior based on empirical data. One key goal is to able to generalized from limited set of data.
D43	Artificial intelligence encompasses other areas apart from machine learning, including knowledge representation, natural language understanding , planning, robotics etc.
D44	The field of artificial intelligence strives to understand and build intelligent entities. The strong artificial intelligence is machine can think and act like human. The weak artificial intelligence is some thinking features can be added to machine.

Document Number	Document Content
D45	Artificial intelligence is branch of computer science dealing with symbolic, non-algorithmic methods of problem solving. Artificial intelligence works with pattern matching methods which attempts to describe objects, events or processes in terms of their qualitative features and logical and computational relationships.
D46	Intelligence is to make sense out of ambiguous message, to respond to situations very flexibly, to recognize relative importance of different elements of situations.
D47	Applications of artificial intelligence are : Expert System which is program designed to act as expert in particular domain. Natural Language processing which enable people and computer to communicate in natural language Speech recognition which is to allow computer to understand human speech. Automatic programming which is to create special programs that act intelligent tools to assist programmers and expedite each phase of programming processes.
D48	Artificial intelligence has increased understanding of the nature of intelligence and provided an impressive array of applications in wide range of areas. It has sharpened understanding of human reasoning and of the nature of intelligence in general.
D49	Artificial intelligence can have two purposes. One is to use the power of computers to augment human thinking. The other is to use a computer artificial intelligence to understand how human think.
D50	Artificial intelligence is the subfield of computer science concerned with understanding the nature of intelligence and constructing computer system capable of intelligent actions. It embodies the dual motives of furthering basic scientific understanding and making computers more sophisticated in the service of humanity.

Table 1.2 gives the sample content of set of 50 documents related to domain Mobile. These set of documents were used for processing during implementation of our proposed semantic similarity based techniques to analyze the results given by our proposed techniques in comparison to other existing techniques as discussed in Chapter 3, Chapter 4 and Chapter 5.

Table 1.2: Set of 50 Documents Related to Mobile Domain

Document Number	Document Content
D1	Android based mobile phones have more applications than windows based mobile phones.
D2	Windows based mobile phones have less application than android based mobile phones.
D3	Android source model is open source and in most devices with proprietary components.
D4	Samsung and Nokia are organizations and manufacturer of mobile phones. In addition to mobile phones and related devices, the company also manufacturers things such as televisions, cameras, and electronic components. Samsung mobiles phones are better than Nokia based mobile phones.
D5	Mobile phones are manufactured by different organizations have operating system like android or windows. Android based mobile phones are better than windows based mobile phones.
D6	Windows is written in C, C++. Windows source model is closed source.
D7	Latest android release is 5.1.1 lollipop and android official website is www.android.com.
D8	Windows latest release is 8.1 update and windows phones official website is www.windowsphone.com.
D9	Android is mobile operating system based on Linux kernel and developed by Google.
D10	Android is designed primarily for touch screen mobile devices.
D11	Windows is a family of mobile operating system developed by Microsoft.

Document Number	Document Content
D12	Windows phones official website is www.windowsphone.com.
D13	In android based mobile phones user can sync their contacts on gmail.com
D14	In windows based mobile phones user can sync their contacts on hotmail.com
D15	Samsung was founded in 1938. Samsung is a south Korean MNC having head quarter in Samsung towns Seoul.
D16	Nokia was founded in 1871. Nokia is Finnish MNC having head quarter in greater Helsinki.
D17	Samsung official website is www.samsung.com.
D18	Nokia official website is www.nokia.com.
D19	List of Samsung products are electronic component, home appliances, consumer electronics, medical equipments.
D20	List of Nokia product is limited to mobile phones and other services.
D21	It comprises numerous subsidiaries and affiliated businesses; most of them united under the Samsung brand, and is the largest South Korean business conglomerate.
D22	Nokia focuses on large-scale telecommunications infrastructures, technology development and licensing
D23	Nokia is a public limited-liability company listed on the Helsinki and New York stock exchanges
D24	Samsung comprises around 80 companies. It is highly diversified, with activities in areas including construction, consumer electronics, financial services, shipbuilding, and medical services
D25	Nokia Networks (previously known as Nokia Siemens Networks (NSN) and Nokia Solutions and Networks (NSN)) is a multinational data networking and telecommunications equipment company headquartered in Espoo, Finland.
D26	Samsung is recognized as one of the leading and most enduring names in the world of mobile technology
D27	Nokia Technologies develops and licenses innovations and

Document Number	Document Content
	the Nokia brand
D28	Nokia Technologies consists of an advanced development team. The development is done in wide areas from imaging, sensing, wireless connectivity, power management and advanced materials.
D29	Samsung Machine Tools of America is a national distributor of machines in the United States
D30	Samsung Medical Center incorporates Samsung Seoul Hospital, Kangbook Samsung Hospital, Samsung Changwon Hospital, Samsung Cancer Center and Samsung Life Sciences Research Center.
D31	Nokia Technologies also provides public participation in its development through a program Invent with Nokia
D32	Samsung Engineering is a multinational construction company headquartered in Seoul
D33	Samsung Electronics is a multinational electronics and information technology company headquartered in Suwon
D34	It was an important factors for Samsung in taking over the Market with the release of dual SIM phone
D35	Initially, Nokia was quite rigid till they finally launched their first Dual Sim Mobile Phone
D36	Samsung integrated with basic features like Color Display, VGA Camera, FM etc with its wide range of Mobile
D37	Initially Nokia concentrated on reliability. Lately, Nokia did also implement these features, but till that time Samsung had captured the section of society who were more interested in having basic features .
D38	Battery is undoubtedly the greatest strength of Nokia
D39	But over the years Samsung did quite a nice job with their R&D and improved their battery quality as well.
D40	Samsung introduced the smart phone world with galaxy series like Galaxy Y, Galaxy Ace, Galaxy Fit and Galaxy S Series. Samsung uses the much user friendly Android Operating System by Google.
D41	Nokia stuck to their simian OS and later with Windows OS. Such wide

Document Number	Document Content
	range of products with user friendly nature helped Samsung to capture the market in very short span of time**.**
D42	Nokia is known for the best build quality when it comes to cell phones
D43	Samsung on the other hand is known for using cheap plastic components and making fairly fragile smart phones by comparison.
D44	In Android, you can install any apps from outside of Google play store.
D45	In terms of security, windows phone is more secure than android. The reason for this is that windows phone doesn't allow installation of apps from unknown sources.
D46	Samsung did provide a lot of basic features in low prices and also introduced Smart Phones series with wide range of products for other section of mobile users.
D47	Microsoft Windows Phone is closed-sourced, meaning that it is owned and managed by Microsoft and developers do not have direct access to the operating system programming code
D48	Android is an open source platform, meaning that the operating system is available for modification by manufacturers to suit their respective needs and phones.
D49	The five major Windows Phone 8 smart phones out now are all high-end, high-quality devices built by Nokia, HTC and Samsung to showcase WP8
D50	Android has much the greater market share, and this is reflected in the amount of handsets from which you can choose.

APPENDIX II

Table 2.1 gives the domain dictionary constructed related to domain of Artificial Intelligence. This dictionary is having the words along with the probable concepts corresponding to each word. The domain dictionary is used for replacement of words by the set of probable concepts in our proposed semantic technique as discussed in Chapter 4.

Table 2.1: Domain Dictionary related to Artificial Intelligence Domain

Word	Set of Probable Concepts
artificial intelligence	unreal computing, contrived information, unreal ability
intelligence	information, knowledge, power, ability,
machine	device, product, mechanism, individual, organization
artificial	unreal, stilted, contrived
robot	device, mechanism, machine
branch	division, discipline, field, subject, projection
computer science	computing, field, discipline, division
science	branch, field, discipline, subject, division
aim	purpose, intent, objective, target, aspire
create	produce, make, build
textbook	text, text edition, school text, schoolbook, casebook
define	specify, delineate, delimit
field	discipline, domain, sphere, plain, subject
study	survey, work, report, field, discipline, sketch, analyze, examine, canvas
design	plan, blueprint, pattern, figure, intent, aim
agent	factor, broker
system	scheme, organization, arrangement

Word	Set of Probable Concepts
word	set of probable concepts
perceive	comprehend
environment	surroundings
action	activity, natural process, process, execute, carry,
success	win , prosperity, achievement
research	inquiry, search, explore, enquiry
problems	job, trouble, difficulty, question
solve	workout, clear, resolve, calculate, compute, figure, determine
world	universe, existence, creation, reality, domain
goal	end, finish, score, context
visualize	picture, image, see, watch, ideate,
future	later, next, succeed
choice	pick, selection, option, prime, prize, quality, select, alternative,
maximize	increase, exploit, tap
utility	public, goal, useful, substitute
courses	line, path, track, trend, row, class, flow
learning	discover, see, instruct, teach, determine, check, watch, hear
algorithm	rule, instructions, formula
experience	undergo, see, know, live, receive
natural language processing	human language technology
language	linguistic, terminology, words, speech
ability	power, quality, cognition, knowledge
read	interpret, talk, utter, indicate, learn, study
understand	infer, read, interpret, translate, realize
human	man, humanity, earthborn, homo, human being

Word	**Set of Probable Concepts**
word	set of probable concepts
speak	speech, utter, verbalize, address
focus	concentrate, center, focalize, sharpen
engage	pursue, absorb, occupy, engross, lease, rent, hire, mesh, wage, lock
behavior	conduct, doing, demeanor
concern	care, refer, pertain, relate, interest, occupy
make	do, create, induce, stimulate, produce, form, build, attain
neural network	computer architecture
mechanism	device, natural object
fundamental	central, profound, underlying
construct	build, make, manufacture, fabricate
operate	run, function, work
chance	opportunity, probability, prospect
intelligent agent	power factor, knowledge factor, information factor
knowledge engineering	cognition technology
knowledge representation	cognition state
many	more
computer	machine, device, computing device, electronic device
area	region, expanse, surface area, domain, field
behave	act, comfort, do
challenges	dispute, take actions
imagine	conceive, think, suppose, ideate, guess, envisage
memorize	learn, study
recognize	know, acknowledge, realize, greet, make out

Word	Set of Probable Concepts
pattern	form, shape, design, model, figure, blueprint, formula
adapt	adjust, conform, accommodate
changes	modification, alteration, variety, vary, switch, shift, exchange, transfer
capacity	capability, content, capacitance
rationally	right
act	human action, routine, bit, move, behave, do, play, represent, work, pretend
real	existent, actual, literal, tangible, material, substantial, genuine
reasoning	logical thinking, abstract thinking, reason out, conclude, intelligent, thinking
planning	preparation, provision, contrive, design
unexpected	unannounced, unpredicted, un hoped, un thought, upset, unscheduled, unplanned
uncertainty	unsure, unsealed, unsettled, changeable
make	create, doing, draw, produce, construct
decision	determination, conclusion, mind, result, outcome, termination, option, choice, selection
adaptation	adjustment, alteration, modification
internal	inner, home, interior
modes	manner, style, way, mood, fashion, modality
updating	change, modify, inform
substantial	significant, real, material, satisfy
progress	advancement, progression, build, work
inference	reasoning, logical thinking, abstract thinking
common	mutual, rough out, coarse
sense	signified, sensation, feel, common sense

Word	Set of Probable Concepts
ontology	metaphysics
heuristic	rule, formula
genetic	inherited, transmitted, genic, hereditary
programming	scheduling, planning, create by mental act
broken	separate, fall apart, violate, fail, erupt, interrupt, split up
group	radical, meet, gather, assemble, forgather
strong	stiff, substantial, firm, secure, un attackable, unassailable
weak	light, unaccented, decrepit, debile, feeble, infirm, frail
refer	mention, advert, pertain, relate, concern, consult, denote
technology	engineering, discipline, subject, field, branch of knowledge
manipulate	control, falsify, represent, rig
predetermined	bias, shape, mold, regulate, influence
rules	pattern, formula, principle, convention, dominate, normal, ruler
apply	use, utilize, hold, practice, implement, enforce
reach	range, scope, orbit, compass, stretch, make, attain, gain, achieve, accomplish
well	good, easily, considerably, intimately, comfortably
function	purpose, role, use, part, office, affair, routine, procedure, operate, work
similar	like, alike, exchangeable, interchangeable, standardize
brain	mind, learning ability, brainpower, head, mental capacity, psyche, master mind
medical	checkup, health check
diagnosis	identification, designation
therapy	medical care, medical aid
recommendation	good word, testimonial, advise, praise, characteristics
control	command, hold, contain, check, curb

Word	Set of Probable Concepts
traffic	collection, aggregate, accumulation, commence, merchandise, dealing
urban	metropolis, citified, city
area	country, sphere, orbit, domain, orbit, arena, field, region, expanse, surface area
current	stream, flow, course, line, electrical phenomenon
development	evolution, growth, exploitation, maturation
new	raw, fresh, novel, recent, modern
planner	contriver, deviser, notebook
lisp	programming language, articulate, enounce, enunciate
prolog	logic programming
strips	slip, clean, programming language
written	compose, pen, scripted, publish, incite
standard	criteria, measure, touchstone, stock
mathematics	math, scientific discipline
primary	chief, main, elemental, principal
digital	digit, figure, integer, whole number
task	project, job, undertaking, tax
associate	companion, fellow, familiar, relate, link, colligate, connect, consort, assort
ethics	moral, ethical motive, value system, ethical code, moral philosophy
fiction	fabrication, fable
around	about, close to, some, roughly, approximate,
fact	info, information, realness, realism, concept, construct, reality
answer	reply, response, solution, result, solvent, resolution
questions	inquiry, enquiry, query, interrogation, interview, motion
work	study, employment, act, function, operate, go on, exercise, process,

Word	Set of Probable Concepts
	bring, play,
social	mixer, culture, ethnic, interpersonal, friendly, elite
matters	substance, affair, thing, topic, subject
insight	penetration, perceptiveness, brainstorm, brain ware
mood	temper, humor, mode, modality
personality	attribute, celebrity, famous person
part	region, office, role, share, break, divide, partial
match	catch, peer, equal, fit, correspond, check, agree, mate, equate, oppose
dimension	property, attribute, proportion, mark, shape, form
fast	secured, firm, flying, degrade, dissolute, loyal
art	artwork, graphics, artistry, artistry creation
smart	ache, bright, fresh, impertinent
soft	voice, diffused, easy, gentle, flaccid, mild, easy going
traditional	conventional, orthodox, long standing, time honored
one	single, unity, solitary, individual, lone
fuzzy	foggy, burred, bleary
dedicated	give, commit, devoted
quality	caliber, timber, tone, choice, prime, select
mind	head, brain, judgment, thinker, idea, intellect
capability	capacity, potential, capable
past	preceding, by, retrieving,
abstract	outline, synopsis, non objective, sneak, lift
idea	thought, estimate, approximate, mind, theme
empirical	empiricism, quackery
data	information, datum, data point

Word	Set of Probable Concepts
generalize	infer, extrapolate, popularize
encompasses	embrace, comprehend, cover
features	characteristics, lineament
symbolic	emblematic
attempt	effort, try, endeavor, undertake, attack, seek
objects	aim, target, physical object, objective
events	outcome, result, consequence, effect, issue, upshot
process	procedure, operation, outgrowth, appendage, treat, work on, serve, march, action, litigate
qualitative	soft
logical	legitimate, coherent, consistent, order, lucid
relationship	relation, kinship
ambiguous	indeterminate, evasive, double, fork, oracular, unstructured
message	content, subject matter, substance
respond	react, answer, reply
situation	site, position, office, spot, post, place, state of affairs
flexible	elastic, pliable, whippy, flexible, compromising
vision	sight, visual sense, imagination, visual modality
communicate	pass, pass on, put across, convey, transmit, intercommunicate
increase	addition, gain, increment, growth, set up
impressive	telling
array	range, layout, set out align, regalia
sharpen	focus, focalize, point, acute, crisp, abrupt, astute, task
power	ability, might, king, force, office, top execution, leader
augment	grow, increase
embodies	incarnate, substantiate, personify

Word	Set of Probable Concepts
dual	double, tow fold, duple
motive	need, motor, motif
sophisticated	twist, pervert, convolute, advanced
service	over hard, inspection, pair, serve, avail, help
game playing	mettlesome act, mettlesome drama
fundamental mechanism	central device, key device, primal device, primal procedure, central phenomenon
information	data, entropy
assistant	helper, supporter, adjunct
application	diligence, coating, covering
memorize	learn
construction	structure, building, expression
model	pattern, simulation, framework
complex	composite, coordination, compound
compute	reckon, calculate
easy	gentle, lenient, tardily
people	populate, natives, citizens, community, group, inhabitants
way	manner, mode, style, fashion
particular	specific, peculiar, special
program	plan, course of study, syllabus, curriculum
order	ordination, edict, prescript, decree
dimensions	attribute, property, proportion
algorithmic problems	algorithm, rules
use	usage, utilization, role, purpose, employ, apply
thing	affair, matter, object, article, item

Word	Set of Probable Concepts
help	aid, assistance, service, avail, facilitate
handle	manipulate, treat, cover, deal, address
limited set	restricted set, confined set, specific set
think	thought, thought process, intellection, mutation, cerebration
relationship	relation, kinship
term	terminus, condition, full term, terminal figure
relative importance	relation grandness, relation importance
different elements	dissimilar components, unlike factors, dissimilar ingredients
situation	state of affair, position, site, place
process	treat, action, work, work on
programmer	coder, software engineer
subfield	subfield
domain	sphere, area, orbit, field, arena
phase	form, stage, period
expert	good, practiced, proficient, skillful
programming process	scheduling process
understand	apprehension, reason, intellect, interpret, translate
general	universal, worldwide, ecumenical, cosmopolitan, common
wide range	wide reach, wide orbit, broad orbit, broad scope
hard	difficult, severe, concentrated, strong, tough, unvoiced, laborious, intemperate
knowledge	cognition
representation	state, creation, activity
engineering	technology, direct, discipline, organize, mastermind, design, plan

Word	Set of Probable Concepts
central	exchange, telephone, key, cardinal, fundamental

The base ontology created for domain artificial intelligence is given in Table 2.2 which is used to extract the relationships between the concepts pairs obtained for each document by using the domain dictionary. These concepts pairs along with the relationships are used to construct the document ontology of each document which is further used in similarity computation.

Table 2.2: Base Ontology for Domain Artificial Intelligence

Concept/Word	Relationship	Concept/Word
intelligence	can be	maximize
artificial	can be	create
human	is	expert
textbook	define	science
human	has	network
action	depends on	environment
problem	can be	solve
solve	requires	knowledge
science	has	aim
success	is	natural
world	has	environment
solve	requires	processing
field	has	textbook
world	has	ability
agent	is	expert'
world	through	language
machine	can be	create
computer	is a	machine
study	has	courses
human	has	behavior
study	has	choice

Concept/Word	**Relationship**	**Concept/Word**
human	type of	machine
robot	type of	machine
knowledge	is	central
field	includes	problem
success	has	future
agent	is	system
human	has	knowledge
field	has	problem
world	has	network
science	includes	engineering
field	has	choice
language	is	natural
human	speak	language
success	requires	knowledge
study	requires	textbook
research	is	central
utility	increase	maximize
environment	is	visualize
machine	requires	process
science	is	branch
environment	is	natural
knowledge	is	visualize
human	type of	robot
computer	requires	science
artificial	is	unreal
robot	is	machine
human	is	experience
machine	is	device
intelligence	can be	machine
problem	define	algorithm
action	is	perceive

Concept/Word	**Relationship**	**Concept/Word**
game	is	play
system	can be	designed
language	is	visualize
world	had been	create
agent	can be	designed
intelligence	requires	knowledge
human	requires	intelligence
human	has	focus
knowledge	through	learning
problem	has	goal
success	requires	intelligence
perceive	from	environment
intelligence	is	visualize
intelligence	by	read
robot	can be	create
human	has	mind
problem	has	choice
textbook	has	knowledge
knowledge	can be	maximize
research	is	visualize
learning	through	experience
field	has	courses
environment	has	represent
world	is	visualize
unreal	is	contrived
power	is	unreal
power	can be	unreal
ability	is not	unreal
ability	has	power
power	used in	device
device	has	division

Concept/Word	**Relationship**	**Concept/Word**
division	made of	organization
organization	has	division
division	is	discipline
discipline	is	field
projection	has	discipline
ability	define	information
information	contains	knowledge
knowledge	can be	produce
produce	is	make
make	is	build
product	is	produce
product	requires	knowledge
individual	have	knowledge
organization	has	individual
organization	made of	individual
individual	reads	subject
subject	has	mechanism
individual	has	living way
infidel	has	nature
organization	has	discipline
projection	has	purpose
purpose	is	intent
device	is	calculator
device	is	figurer
device	is	estimator
calculator	has	objective
estimator	has	objective
figurer	has	objective
objective	is	target
target	to	aspire
make	is	build

Concept/Word	**Relationship**	**Concept/Word**
produce	is	make
intelligence	of	machine
branch	of	computer
system	from	experience
estimator	of	field
unreal computing	has	ability
intelligence	of	device
unreal computing	of	machine
contrived information	of	product
unreal ability	needs	knowledge
power	of	device
ability	of	device
information	related to	device
device	needs	knowledge
unreal computing	needs	ability
unreal computing	is	power
contrived information	is	ability
unreal ability	needs	information
unreal ability	needs	knowledge
power	of	product
power	of	organization
ability	of	mechanism
information	relates	product
mechanism	requires	knowledge
machine	and	robot
device	has	mechanism
unreal computing	is	division
unreal computing	is	discipline
contrived information	relates	subject
unreal computing	is	field
field	of	computing

Concept/Word	Relationship	Concept/Word
field	is	discipline
field	is	division
division	is	branch
division	is	discipline
division	has	subjects
division	means	branch
subject	has	objective
subject	has	purpose
field	has	target
intent	to	make
intent	to	produce
intent	`to	build
target	to	build
purpose	to	produce
work	of	power
ability	of	work
analyze	of	information
analyze	of	knowledge
examine	of	ability
intent	of	information
intent	of	knowledge
pattern	of	information
factor	has	arrangement
factor	need	arrangement
factor	has	organization
factor	need	organization
comprehend	its	surrounding
activity	take	scheme
mechanism	and	projection
device	and	discipline
device	and	division

Concept/Word	**Relationship**	**Concept/Word**
computing	of	subject
discipline	of	computing
scheme	perceives	surroundings
surroundings	takes	activity
activity	maximizes	opportunity
activity	maximizes	probability
opportunity	of	win
opportunity	of	achievement
plan	of	power factor
pattern	of	information factor
organization	is	information factor
arrangement	is	knowledge factor
discipline	as	pattern
subject	as	field
information factor	is	organization
more	of	job
more	of	task
cognition	about	universe
cognition	about	world
more	of	state
cognition technology	and	state
cognition technology	and	activity
cognition technology	and	creation
survey	of	machine
work	of	device
analyze	of	device
examine	of	machine
analyze	of	rule
examine	of	instruction
cognition	about	universe
cognition	about	world

Concept/Word	**Relationship**	**Concept/Word**
world	is	more
universe	is	more
more	of	product
more	of	job
more	of	device
knowledge	requires	job
cognition	is	creation
creation	and	cognition technology
cognition technology	are	fundamental
state	and	cognition technology
organization	perceives	surroundings
field	of	information factor
contrived information	is	field
unreal ability	is	field
contrived information	is	domain
unreal ability	is	domain
unreal computing	is	field
unreal computing	is	domain
field	of	division
computing	of	field
field	of	discipline
domain	of	division
domain	do	computing
domain	of	discipline
division	on	device
division	on	organization
discipline	on	computing
computing	on	product
device	on	conduct
device	on	doing
product	on	doing

Concept/Word	Relationship	Concept/Word
conduct	that	human being
doing	that	human being
human being	consider	power
human being	consider	ability
human being	consider	information
human being	consider	knowledge
division	is	discipline
division	of	computing
device	is	computing
computing	and	division
unreal computing	is	contrived information
device	has	ability
device	has	field
computing	is	discipline
discipline	of	computing
human being	has	ability
ability	needs	human being
ability	consider	human being
unreal computing	is	unreal ability
unreal computing	needs	contrived information
power factor	has	end
knowledge factor	needs	score
knowledge factor	has	end
useful	maximizes	quality
select	maximizes	goal
public	maximizes	selection
public	of	class
goal	of	line
goal	of	path
useful	of	trend
human language	gives	power

Concept/Word	**Relationship**	**Concept/Word**
technology		
human language technology	gives	cognition
human language technology	gives	knowledge
cognition	and	process
cognition	and	setup
dispute	in	discipline
take exception	in	field
dispute	in	field
discipline	and	individual
field	and	individual
computing	and	conduct
computing	and	doing
unreal computing	includes	mettlesome act
central device	enable	build
primal device	enable	fabricate
unreal computing	includes	mettlesome drama
diligence	of	knowledge
covering	of	knowledge
human language technology	gives	device
human language technology	gives	product
organization	gives	human language technology
individual	has	human language technology
unreal ability	includes	mettlesome act
mettlesome act	includes	contrived information
human being	that	helper
helper	that	power
human being	that	ability

Concept/Word	**Relationship**	**Concept/Word**
human being	that	knowledge
diligence	of	device
coating	of	mechanism
covering	of	machine
device	is	discipline
field	kind of	ability
device	as	discipline
discipline	as	computing
diligence	of	unreal computing
covering	of	contrived information
diligence	of	unreal ability
covering	of	device
diligence	of	mechanism
unreal computing	plan	natural process
contrived information	plan	execute
unreal ability	plan	process
unreal computing	plan	activity
data	that	power
entropy	that	knowledge
data	that	information
data	that	man
entropy	that	human being
data	that	earth born
activity	and	job
natural process	and	trouble
execute	and	difficulty
scheme	of	structure
structure	of	organization
power	of	organization
ability	of	arrangement
power	is	quality

Concept/Word	Relationship	Concept/Word
ability	is	cognition
knowledge	is	cognition
power	is	capability
power	is	capacitance
ability	is	capability
knowledge	is	content
computer	and	assemble
compound	and	computation
composite	and	computation
coordination	and	computation
simulation	and`	doing
conduct	and	pattern
process	and	cognition
operation	and	human being
appendage	and	man
dispute	in	compute
take exception	in	reckon
unreal computing	covers	dispute
unreal ability	covers	take exception
populate	as	gentle
populate	as	lenient
populate	as	tardily
power	involves	unreal ability
cognition	involves	contrived information
knowledge	involves	unreal computing
power	with	universe
cognition	with	existence
knowledge	with	creation
cognition	with	reality
power	with	domain
unreal ability	includes	logical thinking

Concept/Word	Relationship	Concept/Word
contrived information	includes	abstract thinking
unreal computing	includes	intelligent
unreal computing	includes	preparation
unreal computing	includes	provision
logical thinking	is	power
abstract thinking	is	power
intelligent	is	power
preparation	is	power
provision	is	power
logical thinking	and	job
abstract thinking	and	trouble
intelligent	and	difficulty
preparation	and	inquiry
provision	and	enquiry
logical thinking	and	search
abstract thinking	and	explore
unreal ability	has	advancement
contrived information	has	progression
unreal computing	has	build
abstract thinking	has	work
advancement	in	instruct
progression	in	see
build	in	discover
work	in	teach
common sense	and	discover
common sense	and	see
common sense	and	instruct
common sense	and	teach
legitimate	include	unreal
coherent	include	stilted
consistent	include	contrived

Concept/Word	**Relationship**	**Concept/Word**
unreal	include	ability
contrived	include	information
unreal	include	power
ordered	include	unreal
stiff	and	light
substantial	and	unaccented
firm	and	decrepit
secure	and	defile
unattackable	and	feeble
unreal ability	into	radical
contrived information	into	gather
unreal computing	into	assemble
unreal computing	into	meet
light	refers	engineering
unaccented	refers	discipline
decrepit	refers	subject
defile	refers	field
feeble	refers	branch of knowledge
unreal ability	refers	subject
contrived information	refers	field
unreal computing	refers	branch of knowledge
engineering	is	capable
discipline	is	capable
subject	is	capable
field	is	capable
branch of knowledge	is	capable
stiff	refers	discipline
substantial	refers	subject
firm	refers	field
secure	refers	branch of knowledge
unattackable	refers	engineering

Concept/Word	**Relationship**	**Concept/Word**
discipline	has	power
subject	has	power
field	has	power
branch of knowledge	has	power
engineering	has	power
power	is	capable
capable	in	manner
mode	in	capable
style	in	capable
fashion	in	capable
machine	than	metropolis
computing device	than	domain
arrangement	than	orbit
scheme	than	arena
organization	than	sphere
device	than	citified
domain	in	command
orbit	in	hold
arena	in	contain
sphere	in	check
citified	in	command
command	at	human being
hold	at	man
check	at	earth born
command	at	homo
survey	that	homo
work	that	human being
report	that	man
unreal ability	involves	power
contrived information	involves	knowledge
exchange	is	mechanism

Concept/Word	Relationship	Concept/Word
exchange	is	determine
key	is	instruct
fundamental	is	mechanism
telephone	is	product
cognition	about	universe
existence	about	cognition
cognition	about	domain
reality	about	cognition
cognition	are	key
cognition	are	fundamental
cognition	are	cardinal
key	are	discipline
fundamental	are	technology
cardinal	are	organize
cognition	are	job
cognition	are	trouble
cognition	are	difficulty
goal	maximizes	selection
goal	maximizes	prize
goal	maximizes	option
useful	maximizes	option
useful	maximizes	selection
activity	plan	power
information	plan	process
information	plan	natural process
knowledge	plan	activity
central	enable	build
profound	enable	make
underlying	enable	fabricate
build	of	scheme
make	of	scheme

Concept/Word	Relationship	Concept/Word
fabricate	of	arrangement
power	of	organization
power	of	scheme
cognition technology	are	exchange
telephone	are	cognition technology
key	are	cognition technology
cardinal	are	cognition technology
fundamental	are	cognition technology
activity	and	cognition technology
organization	as	field
device	as	information factor
division	include	win
product	kind of	achievement
field	as	ability
information factor	kind of	mechanism
computing	as	achievement
win	kind of	division
machine	like	man
device	like	human
man	like	computing
human being	like	computing
job	are	cognition
creation	and	cognition technology
state	and	cognition technology
activity	and	cognition technology
cognition	and	cognition technology
job	as	achievement
more	kind of	opportunity
more	as	discipline
job	as	computing
job	like	device

Concept/Word	**Relationship**	**Concept/Word**
more	is	product
select	as	probability
goal	kind of	opportunity
goal	as	more
job	as	line
device	is	cardinal
device	is	telephone
device	is	exchange
device	is	key
field	with	computing device
division	with	device
discipline	with	machine
computing	with	computing device
unreal	on	central phenomenon
information	on	primal device
knowledge	on	key device
pattern	and	conduct
computation	and	compound
simulation	and	doing
framework	and	composite
coordination	of	assemble
conduct	of	gather
assemble	and	machine
assemble	and	electronic device
capacitance	is	specific
capability	is	special
content	is	peculiar
structure	of	syllabus
manufacture	of	plan
fabricate	of	course of study
plan	make	characteristics

Concept/Word	**Relationship**	**Concept/Word**
course of study	make	advise
syllabus	make	testimonial
curriculum	make	testimonial
metropolis	than	flow
citified	than	course
city	than	arrangement
city	in	moderate
citified	in	aggregate
city	in	hold
metropolis	in	hold
field	that	human being
modern	that	homo
sketch	that	man
recent	that	homo
examine	that	humanity
humanity	at	aggregate
human being	at	merchandise
man	at	moderate
man	at	curb
homo	at	manipulate
raw	on	unreal ability
recent	on	contrived information
novel	on	unreal computing
modern	on	unreal computing
unreal	have	scheduling
information	have	scheduling
power	have	create by mental act
ability	have	planning
knowledge	have	scheduling
scheduling	for	unreal computing
planning	for	unreal computing

Concept/Word	**Relationship**	**Concept/Word**
words	in	information
terminology	in	employment
device	in	play
linguistic	in	study
knowledge	and	logic programming
linguistic	and	information
machine	and	articulate
device	and	play
computing device	and	information
unreal ability	is	power
contrived information	is	knowledge
unreal computing	is	cognition
power	of	digit
power	of	electronic device
cognition	of	computing device
information	of	computing device
power	of	machine
unreal ability	is	part
contrived information	is	partial
unreal computing	is	region
break	of	ethical code
divide	of	value system
role	of	moral
role	of	value system
value system	of	unreal computing
moral	of	unreal ability
moral	of	subject
ethical code	of	branch of knowledge
value system	of	engineering
value system	of	discipline
unreal ability	combines	mastermind

Concept/Word	**Relationship**	**Concept/Word**
contrived information	combines	technology
unreal computing	combines	field
direct	in	edict
division	in	prescript
field	in	decree
information	and	logic programming
capacitance	in	special
division	is	line
mechanism	is	path
unreal ability	as	direct
unreal computing	as	technology
contrived information	as	organized
direct	is	scheme
technology	is	scheme
unreal ability	is	division
unreal computing	is	field
contrived information	is	discipline
unreal ability	as	division
unreal computing	as	field
contrived information	as	discipline
interrogation	is	information
enquiry	is	knowledge
query	is	power
information	is	knowledge
ethnic	is	power
knowledge	is	cognition
power	is	cognition
ability	is	knowledge
ability	is	power
unreal ability	is	region
contrived information	is	partial

Concept/Word	**Relationship**	**Concept/Word**
unreal computing	is	computation
region	of	knowledge
computation	of	cognition
partial	of	cognition
attribute	of	human being
proportion	of	information
proportion	of	knowledge
knowledge	for	rules
information	for	rules
information	for	algorithmic problem
knowledge	for	algorithmic problem
power	for	rules
power	for	algorithmic problem
unreal ability	is	usage
contrived information	is	utilization
unreal computing	is	apply
usage	of	electronic device
purpose	of	machine
role	of	machine
utilization	of	electronic device
electronic device	do	impertinent
machine	do	impertinent
impertinent	by	soft computation
soft computation	of	concentrated
soft computation	of	intemperate
soft computation	of	calculate
unreal ability	is	artistry creation
unreal computing	is	artistry
artistry creation	of	electronic device
graphics	of	electronic device
artwork	of	machine

Concept/Word	Relationship	Concept/Word
matter	and	enquiry
affairs	and	query
matter	and	motion
affairs	and	interrogation
enquiry	with	aid
query	with	assistance
interrogation	with	assist
query	with	helper
service	of	arrangement
helper	of	organization
aid	of	scheme
avail	of	arrangement
unreal ability	lets	system
unreal computing	lets	machine
humanity	is	knowledge
humanity	is	power
information	is	power
knowledge	is	power
power	of	human being
power	of	man
cognition	of	man
cognition	of	human being
head	of	capacity
brain	of	capacity
intellect	of	potentiality
capacity	from	preceding
potentiality	from	undergo
humanity	as	prime
information	as	prime
humanity	as	caliber
knowledge	as	caliber

Concept/Word	**Relationship**	**Concept/Word**
man	as	select
select	of	intellect
finish	is	capable
content	is	capable
end	is	capable
score	is	capable
tone	of	thinker
caliber	of	intellect
manipulation	of	outline
treatment	of	sneak
cover	of	non objective
deal	of	outline
address	of	lift
non objective	and	knowledge
lift	and	power
outline	and	cognition
capable	from	restricted set
capable	from	confine set
capable	from	specific set
restricted set	of	information
confine set	of	data point
specific set	of	datum
secure	is	product
knowledge	is	mechanism
firm	is	organization
information	related to	individual
stilted	can be	mechanism
product	and	human being
subject	of	unreal computing
subject	of	contrived information
subject	of	unreal ability

Concept/Word	**Relationship**	**Concept/Word**
human being	owns	product
human being	know	mechanism
domain	of	unreal computing
domain	of	contrived information
domain	of	unreal ability
mechanism	and	man
device	and	man
organization	and	human being
organization	know	humanity
discipline	of	unreal computing
discipline	of	contrived information
discipline	of	unreal ability
unreal computing	and	knowledge
model	attempts	target
design	attempts	objective
design	attempts	aim
form	attempts	outcome
contrived information	and	information
unreal ability	and	ability
knowledge	is	intellection
decrepit	is	mutation
stilted	is	cerebration
information	is	thought
information	is	thought process
unreal	with	model
terminal figure	of	characteristics
condition	of	lineament
terminus	of	lineament
terminology	of	characteristics
information	with	design
knowledge	with	design

Concept/Word	**Relationship**	**Concept/Word**
contrived	with	form
formula	attempts	physical object
blueprint	attempts	effect
target	in	terminal figure
outcome	in	condition
aim	in	terminus
objective	in	terminology
characteristics	and	relationship
characteristics	and	kinship
field	with	emblematic
division	with	emblematic
computing	with	emblematic
emblematic	of	job
emblematic	of	trouble
relation grandness	of	dissimilar components
relation grandness	of	unlike factor
dissimilar components	of	place
dissimilar components	of	site
unlike factor	of	state of affairs
ability	is	common sense
information	is	signified
knowledge	is	signified
signified	of	in determine
feel	of	evasive
unreal ability	are	arrangement
contrived information	are	scheme
arrangement	is	plan
scheme	is	syllabus
course of study	is	sphere
plan	in	field
syllabus	in	area

Concept/Word	Relationship	Concept/Word
course of study	as	skillful
plan	as	good
syllabus	as	proficient
action	enable	populate
work	enable	device
action	enable	computing device
machine	of	stage
machine	of	form
device	of	stage
computing device	of	stage
coder	and	stage
software engineer	and	stage
coder	and	form
stage	of	scheduling processes
device	in	words
machine	in	linguistic
nature	of	ability
nature	of	power
nature	of	knowledge
computing device	in	terminology
unreal ability	has	reason
contrived information	has	apprehension
contrived information	has	intellect
unreal computing	has	apprehension
unreal computing	has	intellect
reason	of	nature
translate	of	nature
apprehension	of	nature
intellect	of	nature
knowledge	in	world wide
information	in	cosmopolitan

Concept/Word	Relationship	Concept/Word
ability	in	universal
power	in	world wide
power	and	range
knowledge	and	align
information	and	range
knowledge	and	range
information	about	range
range	of	coating
set out	of	covering
covering	in	broad scope
covering	in	broad ambit
diligence	in	wide reach
single	is	leader
unity	is	ability
unity	is	king
single	is	king
leader	of	machine
leader	of	electronic device
ability	of	computing device
office	of	electronic device
office	of	computing device
king	of	machine
ability	of	machine
unreal computing	is	subfield
contrived information	is	subfield
subfield	of	computing
subfield	of	discipline
subfield	of	division
field	with	nature
division	with	nature
computing	with	nature

Concept/Word	**Relationship**	**Concept/Word**
nature	of	ability
nature	related to	science
nature	of	power
nature	of	knowledge
information	and	arrangement
knowledge	and	arrangement
power	and	electronic device
power	and	computing device
information	and	device
convolute	in	inspection
advanced	in	serve
pervert	in	help
advanced	in	help
help	of	humanity
serve	of	humanity
inspection	to	humanity
path	includes	mechanism
telephone	comes from	knowledge
ability	is	key
projection	is	analyze
knowledge	is	field
job	is	achievement
win	includes	more
opportunity	is	goal
win	comes from	path
information factor	kind of	key
organization	is	cardinal
win	in	field
achievement	in	division
information factor	is	unreal computing
organization	is	field

Concept/Word	**Relationship**	**Concept/Word**
achievement	in	line
information factor	is	device
organization	is	key
contrived information	is	key
knowledge	is	cardinal
discipline	includes	goal
computing	is	path
contrived information	is	information factor
ability	in	organization
projection	is	pattern
device	kind of	information factor
discipline	same as	line
computing	means as	path
discipline	includes	more
computing	in	device
information factor	is	key
organization	is	cardinal
information factor	in	division
goal	from	more
line	is	job
more	includes	exchange
job	is	rule
field	comes from	more
job	is	division
line	is	exchange
path	is	rule
field	is	line
division	like as	path
contrived information	includes	key
discipline	is	cardinal
unreal computing	includes	key

Concept/Word	Relationship	Concept/Word
field	into	cardinal
projection	same as	more
device	derived from	job
ability	inherited	cardinal
field	in	analyze
device	derived from	division
more	is	path
job	is	line
more	in	analyze
job	derived from	device
domain	means as	analyze
division	includes	rule
computing	is	analyze
ability	comes from	mechanism
device	is	pattern
knowledge	like as	information factor
diligence	is	more
mechanism	includes	job
diligence	is	goal
mechanism	is	path
device	inherited	division
mechanism	is	computing
device	comes from	field
mechanism	includes	division
diligence	is	rule
device	part of	goal
mechanism	kind of	path
ability	means as	fabricate
mechanism	is	arrangement
discipline	derived from	fabricate
pattern	is	arrangement

Concept/Word	**Relationship**	**Concept/Word**
more	comes from	fabricate
job	like as	arrangement
goal	comes from	fabricate
path	is	arrangement
analyze	is	fabricate
rule	includes	arrangement
information	in	organization
primal device	is	device
field	inherited	fabricate
discipline	derived from	arrangement
device	is	fabricate
mechanism	is	arrangement
pattern	is	fabricate
information factor	in	arrangement
line	like as	fabricate
path	is	arrangement
device	comes from	arrangement
computing device	derived from	fabricate
diligence	into	fabricate
mechanism	is	arrangement
assemble	synonym	mechanism
electronic device	comes from	projection
achievement	in	assemble
win	comes from	coordination
more	comes from	coordination
job	in	assemble
coordination	is	goal
assemble	like as	path
coordination	includes	analyze
projection	is	device
field	is	coordination

Concept/Word	**Relationship**	**Concept/Word**
computing	is	assemble
division	is	coordination
coordination	includes	mechanism
assemble	in	device
fabricate	into	coordination
arrangement	is	assemble
rule	comes from	coordination
device	includes	assemble
device	in	coordination
diligence	is	coordination
mechanism	in	assemble
coordination	is	fabricate
assemble	is	arrangement
ability	comes from	win
cognition	is	arrangement
cognition	is	field
ability	includes	unreal computing
knowledge	includes	exchange
power	is	key
unreal computing	in	knowledge
field	is	cognition
cardinal	is	knowledge
key	derived from	cognition
information factor	means as	information
organization	includes	knowledge
power	includes	unreal ability
capability	is	ability
capability	in	domain
capability	in	field
capacitance	into	knowledge
unreal computing	is	power

Concept/Word	Relationship	Concept/Word
region	in	capacitance
field	into	capacitance
power	is	key
capacitance	includes	exchange
information factor	is	power
organization	includes	capacitance
domain	is	division
universe	is	device
universe	includes	computing
domain	includes	computing device
intelligent	is	special
power	includes	capability
logical thinking	in	cognition
intelligent	part of	mettlesome drama
logical thinking	kind of	unreal ability
logical thinking	kind of	unreal computing
power	derived from	field
key	synonym	intelligent
exchange	includes	power
logical thinking	includes	arrangement
power	in	organization
logical thinking	in	mettlesome drama
cognition	from	domain
cognition	from	region
power	from	region
instruct	in	special
capacitance	is	advancement
explore	into	machine
trouble	in	electronic device
abstract thinking	includes	mechanism
explore	is	projection

Concept/Word	Relationship	Concept/Word
mechanism	inherited	trouble
explore	is	projection
explore	into	machine
trouble	in	assemble
field	from	power
field	is	unreal ability
field	in	information factor
field	is	contrived information
capable	inherited	cognition
capable	is	capacitance
capable	in	field
capable	in	arrangement
capable	includes	knowledge
common sense	in	assemble
common sense	in	mechanism
see	from	machine
see	is	projection
capable	is	capability
capable	in	domain
capable	in	division
instruct	is	discipline
instruct	from	electronic device
firm	is	defile
field	is	unreal ability
field	from	power
power	in	logical thinking
power	is	advancement
power	is	capable
capable	is	capability
capable	is	ability
capable	in	words

Concept/Word	**Relationship**	**Concept/Word**
capable	in	advancement
contrived information	inherited	capable
unreal ability	inherited	capable
manner	includes	power
manner	includes	information
manner	includes	instruct
manner	in	discipline
manner	in	domain
structure	part of	more
structure	part of	coordination
structure	is	fabricate
analyze	into	structure
goal	like as	structure
field	derived from	structure
electronic device	derived from	structure
syllabus	into	assemble
syllabus	in	arrangement
device	means as	syllabus
line	kind of	syllabus
information factor	into	syllabus
organization	includes	syllabus
power	comes from	words
words	comes from	advancement
information	is	instruct
information	includes	logical thinking
information	in	device
information	is	capability
see	in	logic programming
logic programming	is	search
discipline	is	logic programming
unreal computing	includes	power

Concept/Word	**Relationship**	**Concept/Word**
field	is	unreal computing
logical thinking	includes	unreal computing
ability	includes	unreal computing
power	means as	capable
power	means as	ability
machine	like as	structure
machine	includes	coordination
machine	into	fabricate
electronic device	includes	syllabus
electronic device	in	assemble
electronic device	is	arrangement
power	comes from	field
power	like as	more
power	kind of	mechanism
power	includes	analyze
power	inherited	goal
machine	is	information factor
machine	in	domain
machine	in	knowledge
machine	like as	device
machine	in	line
unreal computing	is	contrived information
unreal computing	is	unreal ability
unreal computing	derived from	electronic device
power	includes	discipline
power	includes	domain
power	in	organization
unreal computing	includes	ability
cognition	includes	power
cognition	like as	fabricate
cognition	kind of	diligence

Concept/Word	Relationship	Concept/Word
cognition	part of	division
domain	in	cognition
rule	in	cognition
analyze	in	cognition
more	means as	cognition
cognition	includes	pattern
cognition	includes	power
computing device	includes	arrangement
computing device	is	device
division	includes	computing device
rule	includes	computing device
job	derived from	computing device
computing device	derived from	information factor
information	includes	common sense
instruct	in	logic programming
capacitance	in	words
information	in	special
diligence	kind of	structure
knowledge	comes from	syllabus
cognition	is	ability
cognition	is	diligence
cognition	includes	more
computing device	derived from	knowledge
computing device	derived from	job
structure	inherited	diligence
syllabus	includes	knowledge
more	includes	role
role	in	fabricate
role	in	coordination
role	in	field
role	from	goal

Concept/Word	**Relationship**	**Concept/Word**
value system	kind of	course of study
value system	part of	assemble
value system	in	arrangement
value system	from	diligence
value system	from	division
value system	from	analyze
line	from	value system
job	is	value system
discipline	comes from	knowledge
device	comes from	discipline
unreal computing	is	ability
unreal computing	is	logical thinking
unreal computing	is	power
field	is	unreal computing
achievement	is	unreal computing
region	from	power
region	from	capable
field	is	region
capability	comes from	knowledge
region	includes	capability
region	includes	knowledge
region	includes	electronic device
region	comes from	device
region	comes from	arrangement
field	includes	words
capable	includes	field
field	part of	build
field	part of	capacitance
decree	from	information
manner	in	decree
discover	in	special

Concept/Word	**Relationship**	**Concept/Word**
discover	in	decree
special	into	decree
discipline	as	knowledge
discipline	derived from	device
role	in	coordination
role	as	diligence
path	into	role
more	as	role
assemble	includes	moral
mechanism	includes	moral
line	inherited	moral
moral	inherited	job
moral	inherited	pattern
unreal ability	into	information factor
unreal ability	inherited	information
knowledge	derived from	region
fabricate	derived from	region
course of study	in	subject
arrangement	in	subject
value system	as	pattern
engineering	into	information factor
unreal computing	into	organization
query	inherited	unreal computing
query	inherited	unreal ability
ability	from	query
power	from	query
logical thinking	comes from	query
contrived information	comes from	query
device	comes from	query
information factor	comes from	query
power	from	region

Concept/Word	Relationship	Concept/Word
power	from	domain
power	as	capable
capability	from	power
cognition	from	power
discipline	in	power
key	as	power
arrangement	as	power
power	as	knowledge
ability	in	information factor
ability	into	device
ability	as	query
unreal ability	is	ability
unreal computing	is	ability
logical thinking	includes	ability
ability	is	contrived information
power	includes	knowledge
key	as	power
power	in	region
power	into	arrangement
power	in	field
discipline	includes	power
domain	includes	power
quality	inherited	capable
quality	derived from	capability
quality	derived from	cognition
unreal computing	is	ability
unreal computing	as	information factor
unreal computing	from	query
contrived information	is	unreal computing
unreal ability	is	unreal computing
region	as	knowledge

Concept/Word	**Relationship**	**Concept/Word**
region	as	pattern
region	in	analyze
region	in	diligence
arrangement	includes	region
line	in	region
power	from	region
discipline	in	region
more	as	region
moral	as	region
fabricate	in	region
field	as	region
coordination	includes	region
region	as	domain
information factor	as	knowledge
knowledge	derived from	device
knowledge	includes	mechanism
knowledge	in	arrangement
knowledge	as	path
knowledge	includes	unreal ability
job	derived from	knowledge
machine	derived from	knowledge
course of study	inherited	knowledge
knowledge	includes	power
capable	includes	knowledge
logical thinking	as	knowledge
assemble	in	knowledge
cognition	comes from	computation
computation	inherited	capable
computation	derived from	power
computation	derived from	capability
cognition	as	capable

Concept/Word	**Relationship**	**Concept/Word**
cognition	as	power
cognition	as	capability
proportion	includes	analyze
more	includes	proportion
computation	same as	proportion
pattern	kind of	proportion
proportion	includes	role
proportion	includes	power
proportion	in	fabricate
proportion	in	coordination
line	is	proportion
diligence	includes	proportion
information	derived from	device
information	inherited	job
information	derived from	cognition
information	derived from	information factor
information	in	moral
information	in	path
information	in	organization
information	into	machine
information	includes	course of study
information	in	assemble
information	into	arrangement
information	includes	mechanism
algorithmic problem	in	planning
rules	in	unreal computing
technology	as	unreal ability
contrived information	as	technology
information	as	technology
information factor	as	technology
device	comes from	technology

Concept/Word	**Relationship**	**Concept/Word**
technology	includes	ability
technology	includes	field
technology	from	unreal ability
technology	from	unreal computing
technology	from	power
technology	in	field
technology	includes	logical thinking
scheme	in	domain
scheme	as	key
scheme	in	discipline
scheme	as	arrangement
power	in	scheme
region	includes	scheme
capable	into	scheme
cognition	into	scheme
capability	into	scheme
proportion	includes	ability
proportion	includes	path
role	into	proportion
mechanism	includes	knowledge
line	in	knowledge
value system	as	knowledge
computation	as	path
cognition	in	line
device	derived from	technology
scheme	as	cardinal
technology	derived from	device
unreal ability	comes from	technology
special	as	technology
scheme	includes	path
cardinal	is	path

Concept/Word	Relationship	Concept/Word
field	in	path
capability	from	path
query	includes	field
power	from	capable
query	from	device
power	from	cardinal
contrived information	as	unreal ability
contrived information	as	unreal computing
contrived information	as	ability
contrived information	in	discipline
contrived information	in	power
contrived information	in	field
contrived information	as	logical thinking
contrived information	as	power
power	kind of	utilization
proportion	like as	utilization
artistry creation	like as	utilization
region	like as	utilization
knowledge	includes	utilization
query	from	utilization
arrangement	from	utilization
role	as	utilization
fabricate	as	utilization
capable	in	utilization
cognition	in	utilization
capability	includes	utilization
coordination	includes	utilization
diligence	from	utilization
line	as	utilization
field	as	utilization
domain	as	utilization

Concept/Word	**Relationship**	**Concept/Word**
device	derived from	utilization
more	derived from	utilization
electronic device	derived from	knowledge
electronic device	inherited	power
electronic device	inherited	value system
electronic device	inherited	course of study
electronic device	inherited	assemble
electronic device	as	arrangement
electronic device	as	device
electronic device	from	discipline
electronic device	is	key
electronic device	as	information factor
electronic device	includes	path
electronic device	includes	job
unreal computing	is	unreal ability
ability	is	unreal ability
query	from	unreal ability
discipline	as	unreal ability
power	into	unreal ability
field	is	unreal ability
logical thinking	is	unreal ability
artistry creation	from	region
artistry creation	from	usage
artistry creation	as	proportion
artistry creation	as	knowledge
artistry creation	includes	power
artistry creation	in	arrangement
artistry creation	is	role
artistry creation	from	fabricate
artistry creation	as	capable
artistry creation	as	cognition

Concept/Word	**Relationship**	**Concept/Word**
artistry creation	as	capability
artistry creation	in	coordination
artistry creation	in	fabricate
artistry creation	in	domain
artistry creation	in	diligence
artistry creation	from	field
artistry creation	from	device
artistry creation	in	line
artistry creation	in	more
electronic device	as	digit
affairs	in	information
affairs	in	common sense
affairs	as	abstract thinking
query	from	logic programming
query	from	teach
query	as	explore
query	from	power
helper	in	proportion
helper	in	region
helper	in	role
helper	includes	power
helper	derived from	fabricate
helper	inherited	universe
helper	in	coordination
helper	in	fabricate
helper	in	diligence
helper	as	field
helper	in	domain
helper	derived from	device
helper	into	line
helper	includes	more

Concept/Word	**Relationship**	**Concept/Word**
knowledge	as	organization
value system	in	organization
digit	in	organization
course of study	includes	organization
assemble	as	organization
information factor	as	organization
arrangement	as	organization
device	as	organization
computing	as	organization
discipline	as	organization
key	includes	organization
path	includes	organization
job	includes	organization
machine	from	organization
unreal ability	from	discipline
organization	as	usage
unreal ability	like as	power
organization	kind of	capacitance
artistry creation	from	organization
artistry creation	from	capacitance
artistry creation	includes	analyze
electronic device	as	rule
affairs	includes	information
interrogation	includes	logic programming
interrogation	as	power
assist	in	domain
ability	is	unreal ability
ability	includes	role
knowledge	includes	artistry creation
knowledge	includes	usage
knowledge	from	cognition

Concept/Word	**Relationship**	**Concept/Word**
knowledge	from	value system
knowledge	from	unreal computing
computing	as	power
field	in	power
logical thinking	includes	power
power	in	division
power	as	information factor
power	includes	knowledge
line	in	cognition
helper	comes from	cognition
field	is	cognition
cognition	is	proportion
cognition	is	information
cognition	is	discipline
cognition	includes	organization
cognition	as	fabricate
cognition	as	capable
capability	as	cognition
coordination	as	cognition
more	in	cognition
diligence	in	cognition
analyze	as	cognition
human being	as	path
human being	includes	organization
human being	includes	knowledge
computing	as	human being
organization	includes	human being
course of study	into	human being
assemble	as	human being
arrangement	includes	human being
device	derived from	human being

Concept/Word	**Relationship**	**Concept/Word**
job	from	human being
line	from	cover
cover	from	artistry creation
cover	in	region
cover	as	role
cover	as	more
cover	into	fabricate
cover	into	diligence
cover	same as	division
cover	in	field
cover	as	analyze
path	includes	non objective
electronic device	includes	non objective
knowledge	includes	non objective
value system	from	non objective
job	from	non objective
information	from	non objective
course of study	from	non objective
common sense	from	non objective
logical thinking	from	non objective
arrangement	from	non objective
device	from	non objective
computing	from	non objective
caliber	means as	cognition
helper	kind of	caliber
utilization	kind of	caliber
caliber	kind of	proportion
caliber	as	power
caliber	in	field
caliber	in	coordination
intellect	includes	human being

Concept/Word	**Relationship**	**Concept/Word**
intellect	includes	organization
intellect	from	electronic device
intellect	from	knowledge
intellect	from	digit
intellect	in	capable
intellect	in	assemble
intellect	into	product
logic programming	from	knowledge
teach	from	knowledge
search	as	knowledge
discipline	in	knowledge
discipline	as	subject
discipline	as	fabricate
unreal computing	is	unreal ability
unreal computing	includes	information
course of study	as	unreal computing
knowledge	as	unreal computing
knowledge	as	logic programming
knowledge	as	power
field	in	knowledge
mechanism	in	power
capable	in	mechanism
end	from	unreal ability
end	from	unreal computing
end	from	ability
end	from	discipline
end	includes	power
end	in	field
end	from	logical thinking
capable	in	artistry creation
capable	into	usage

Concept/Word	Relationship	Concept/Word
capable	includes	region
capable	includes	power
capable	kind of	organization
capable	is	capable
capable	is	capability
capable	as	knowledge
terminus	in	specific set
rule	in	specific set
discipline	in	specific set
more	as	specific set
cover	as	specific set
cognition	as	specific set
goal	includes	specific set
helpers	from	specific set
proportion	from	specific set
coordination	from	specific set
fabricate	from	specific set
diligence	from	specific set
field	from	specific set
datum	from	device
datum	from	lineament
datum	from	computing
datum	from	job
datum	includes	non objective
datum	includes	human being
datum	in	path
datum	in	organization
datum	in	information
datum	into	assemble
datum	from	arrangement
datum	from	device

Concept/Word	**Relationship**	**Concept/Word**
datum	from	discipline
information	as	power
design	as	universe
aim	as	field
words	from	aim
capable	from	aim
build	from	aim
capacitance	includes	aim
terminus	kind of	goal
terminus	kind of	rule
terminus	same as	field
terminus	derived from	fabricate
terminus	comes from	coordination
terminus	is	special
terminus	includes	more
terminus	includes	domain
terminus	includes	information
terminus	in	discipline
terminus	in	cover
terminus	as	manner
terminus	as	cognition
terminus	as	helpers
terminus	from	artistry creation
terminus	from	usage
terminus	from	proportion
terminus	includes	region
terminus	includes	decree
terminus	from	discover
terminus	from	role
terminus	from	region
terminus	from	power

Concept/Word	**Relationship**	**Concept/Word**
terminus	from	diligence
terminus	from	ability
path	from	lineament
device	from	lineament
division	derived from	lineament
discipline	derived from	lineament
arrangement	as	lineament
assemble	as	lineament
job	as	lineament
information factor	as	lineament
course of study	includes	lineament
contrived information	includes	lineament
computing	from	lineament
non objective	from	lineament
human being	includes	lineament
organization	from	lineament
electronic device	from	lineament
information	into	lineament
knowledge	from	lineament
device	derived from	lineament
machine	inherited	lineament
moral	from	lineament
value system	from	lineament
unity	from	information
unity	from	end
unity	from	unreal ability
unity	from	power
unity	in	field
diligence	from	kind
common sense	as	kind
discipline	as	kind

Concept/Word	Relationship	Concept/Word
capable	from	kind
artistry creation	from	kind
power	from	kind
fabricate	from	kind
capable	from	kind
machine	as	unreal ability
machine	as	contrived information
machine	derived from	computing
machine	derived from	digit
machine	derived from	course of study
unreal computing	as	information
unreal computing	as	terminus
unreal computing	as	ability
unreal computing	as	discipline
unreal computing	includes	power
unreal computing	includes	field
ability	from	intellect
diligence	from	ability
common sense	from	ability
lineament	from	ability
power	as	ability
organization	as	ability
capable	in	ability
capability	is	ability
knowledge	from	ability
goal	from	ability
more	includes	ability
mechanism	includes	nature
mechanism	as	unreal ability
mechanism	as	affairs
mechanism	as	common sense

Concept/Word	**Relationship**	**Concept/Word**
mechanism	in	line
mechanism	in	job
projection	derived from	query
projection	inherited	logic programming
projection	inherited	discover
diligence	in	power
unreal computing	as	structure
unreal computing	as	electronic device
logical thinking	as	unreal computing
power	in	syllabus
unreal ability	in	syllabus
arrangement	into	discipline
power	into	arrangement
score	as	arrangement
arrangement	in	ability
information	from	arrangement
logical thinking	from	arrangement
capability	from	arrangement
power	in	plan
capable	into	plan
device	as	plan
signified	derived from	plan
capacitance	comes from	plan
knowledge	comes from	plan
intellect	inherited	diligence
intellect	kind of	subject
intellect	kind of	cognition
intellect	like as	unreal ability
intellect	same as	usage
intellect	includes	proportion
intellect	includes	computation

Concept/Word	**Relationship**	**Concept/Word**
intellect	from	moral
intellect	means as	power
intellect	means as	build
nature	derived from	unreal ability
nature	derived from	signified
nature	inherited	contrived information
nature	in	discipline
nature	into	cover
nature	as	man
nature	as	artistry creation
nature	derived from	electronic device
nature	derived from	information
nature	includes	cognition
nature	includes	subject
nature	from	device
nature	from	structure
nature	from	coordination
nature	as	fabricate
nature	comes from	mechanism
nature	derived from	field
nature	derived from	domain
nature	inherited	analyze
nature	from	goal
nature	from	more
nature	from	field
ability	as	field
ability	as	words
ability	as	in determine
ability	as	score
ability	from	computing
ability	derived from	non objective

Concept/Word	Relationship	Concept/Word
ability	includes	syllabus
ability	from	capacitance
ability	from	assemble
ability	in	arrangement
ability	into	knowledge
ability	in	discipline
ability	derived from	device
ability	includes	path
ability	in	job
ability	from	information factor
information	from	universal
decree	from	universal
special	into	universal
capable	into	universal
range	in	interrogation
power	in	affairs
power	is	mettlesome act
unreal computing	is	contrived information
ability	is	mettlesome act
ability	is	mettlesome drama
the intelligence	kind of	science and engineering
combines	means as	is

The base ontology related to artificial intelligence domain having the concepts along with the relationship having weight associated with it is given in Table 2.3. This weighted ontology is used in our proposed relation based measuring of similarity to construct the relation space model as discussed in chapter 3.

Table 2.3: Ontology Based Weights for Set of Documents Related to Domain Artificial Intelligence

Concept/Word	Relationship	Concept/Word	Weight
artificial intelligence	is	intelligence	1
the intelligence	of	machine and robot	0.8
machine and robot	and	branch	0.8
the branch	of	computer science	0.6
computer science	aims	artificial intelligence	0.1
an intelligent agent	is	system	1
system	perceives	environment	0.9
its environment	takes	actions	0.3
action	maximizes	chance	0.4
its chance	of	success	0.3
the field	as	study	0.5
study and design	of	intelligent agent	0.8
representation	and	knowledge engineering	0.9
knowledge engineering	are	artificial intelligence research central	0.9
extensive knowledge	about	world	0.7
problem machines	expected	extensive knowledge	0.8
many	of	problem machines	0.2
artificial intelligence textbook	that	artificial intelligence	0.2
study	and	design	0.6
design	of	intelligent agent	0.6
intelligent agent	is	system	1
study	of	computer algorithm	0.7

Concept/Word	Relationship	Concept/Word	Weight
computer algorithm	improve	automatically	0.6
automatically	through	experience	0.5
machine learning	is	branch	0.9
input	from	environment	0.6
natural language processing	gives	machine	0.7
artificial intelligence	is	area	0.7
area	of	scientific discipline	0.5
problem	require	broad cognition	0.8
broad cognition	about	universe	0.8
focusing	on	creating	0.7
human	consider	intelligent	0.8
machine	on	behavior	0.6
behavior	that	human	0.5
scientific discipline	with	making	0.8
artificial intelligence	is	branch	0.9
branch	of	scientific discipline	0.7
computing machine	like	human	0.9
scientific computing	creates	intelligent machine	0.9
artificial intelligence	includes	game playing	0.6
artificial intelligence	is	subdivision	0.7
subdivision	of	scientific computing	0.7
neural network	and	robotics	0.6
computer	has	artificial intelligence	0.8
applications	of	artificial intelligence robots	0.7
artificial intelligence robots	plan	actions	0.5

Concept/Word	Relationship	Concept/Word	Weight
information, intelligent assistant	that	help human	0.8
financial fraud and game playing system	perform	better	0.5
better	than	human player	0.3
probability	of	win	0.5
actions	and	complex problems	0.3
autonomously	from	experience	0.7
artificial intelligence track	on	fundamental mechanism	0.6
the construction	of	intelligence system	0.45
fundamental mechanism	enable	construction	0.36
as easy	as	people	0.2
such	as	human knowledge	0.2
computational model	and	complex behavior	0.5
individual or group	and	computer	0.5
complex behavior	of	individual or group	0.6
artificial intelligence	covers	key challenges	0.9
human knowledge	and	thought process	0.6
key challenges	in	computing	0.6
intelligence	is	capacity to learn	1
capacity to learn	in	particular	0.3
ability	with	real world	0.5
artificial intelligence	involves	ability	0.7
artificial intelligence	includes	reasoning and planning	0.8
learning	and	internal models	0.6
internal models	are	always	0.2

Concept/Word	Relationship	Concept/Word	Weight
also	includes	learning	0.1
artificial intelligence	made	substantial progress	0.8
substantial progress	in	recognition and learning	0.8
research problems	in	planning and reasoning	0.7
artificial branches	include	logical artificial intelligence	0.7
common sense knowledge	and	reasoning, learning, planning, ontology, heuristic and genetic programming	0.8
artificial intelligence	into	two groups	0.9
weak artificial intelligence	refers	technology	0.5
technology	is	apply the rules	0.9
strong artificial intelligence	refers	technology	0.5
technology	has	think cognitively	1
think cognitively	related to	human brain	1
medical artificial intelligence	is	primarily	0.4
development	on	artificial intelligence	0.5
new study	that	human	0.4
human	are	better	0.6
better	than	computer system	0.7
artificial intelligence researchers	developed	several specialized programming	0.6
several specialized programming	for	artificial intelligence	0.6
language	as	lisp, prolog, strips	0.5
standard language	like	c	1

Concept/Word	**Relationship**	**Concept/Word**	**Weight**
often	in	standard language	0.2
artificial intelligence applications	are	computer science	0.1
computer language	in	lisp	0.3
lisp	are	primarily	0.3
lisp	and	prolog	0.5
artificial intelligence	is	ability	0.8
commonly	with	intelligent being	0.7
ability	of	digital computer	0.7
digital computer or computer	controlled	robot	0.8
robot	and	artificial intelligence	0.7
ethics	of	artificial intelligence	0.6
ethics	of	technology specific	0.7
artificial intelligence	is	part	0.7
part	of	ethics	0.6
artificial intelligence	combines	science and engineering	1
science and engineering	in	order	0.6
artificial intelligence	as	engineering	1
engineering	is	system	1
science function	in	fact	0.7
fact	is	all around	0.2
often	as	science function	1
artificial intelligence	as	science	1
science	helps	human	0.5
questions	is	intelligence	0.6

Concept/Word	Relationship	Concept/Word	Weight
intelligence	and	computer	0.1
social intelligence	is	knowledge of social matters	1
artificial intelligence	is	computational part	0.9
computational part	of	goals	1
all dimensions	of	human intelligence	0.9
human intelligence	for	algorithmic problem	0.8
artificial intelligence	is	use	0.7
use	of	computers	1
computers	do	smart things	1
smart things	by	using	0.6
instead	of	using	0.2
artificial intelligence	is	art	0.7
art	of	making	0.9
artificial intelligence	lets	computer	0.9
questions	with	help	0.5
things	and	questions	0.6
the help	of	fuzzy inference system	0.7
artificial intelligence	is	field	0.7
human intelligence	is	ability	0.8
ability	of	human	1
field	dedicated	development	1
mind	of	capabilities	0.8
capabilities	from	past experience	0.9
quality	of	mind	0.8
human intelligence	as	quality	0.7
handling	of	abstract ideas	0.6

Concept/Word	Relationship	Concept/Word	Weight
abstract ideas	and	change	0.6
field	of	computer science	0.9
artificial intelligence	is	field	0.7
goal	is	able	0.5
able	from	limited set	0.6
limited set	of	data	0.6
deals	with	designing	0.6
designing	and	developing	0.8
artificial intelligence	encompasses	areas	0.8
apart	from	machine learning	0.6
artificial intelligence	and	intelligent entities	0.6
field	of	artificial intelligence	0.7
weak artificial intelligence	is	some thinking	0.7
strong artificial intelligence	is	machine	0.7
machine	and	human	0.6
features	added	machine	0.6
artificial intelligence works	with	pattern matching models	0.8
pattern matching models	attempts	objects, events or processes	0.8
objects, events or processes	in	terms	0.5
terms	of	qualitative features	0.4
qualitative features	and	logical and computation features	0.5
dealing	with	symbolic, non-algorithmic methods	0.6
symbolic, non-algorithmic methods	of	problem	0.6
branch	of	computer science	0.7

Concept/Word	Relationship	Concept/Word	Weight
artificial intelligence	is	branch	0.7
different elements	of	situations	0.6
relative importance	of	different elements	0.6
intelligence	is	sense	0.8
sense	of	ambiguous message	0.7
expert	in	particular domain	0.8
applications	of	artificial intelligence	0.9
artificial intelligence	are	expert system	1
expert system	is	program	1
program	as	expert	1
automatic programming	is	special programs	1
special programs	as	intelligent tools	0.9
processing	enable	people and computer	0.7
people and computer	in	natural language speech recognition	0.9
programmers	and	phase	0.5
each phase	of	programming processes	0.6
human reasoning	of	nature	0.6
nature	of	intelligence	0.6
intelligence	provided	impressive array	0.7
impressive array	of	applications	1
applications	in	wide range	1
wide range	of	areas	1
intelligence	in	general	0.8
understanding	of	human reasoning	0.9
artificial intelligence	increased	understanding	1

Concept/Word	Relationship	Concept/Word	Weight
other	is	computer artificial intelligence	0.6
one	is	power	0.3
power	of	computers	1
basic scientific understanding	and	making	0.9
artificial intelligence	is	subfield	0.7
subfield	of	computer science	0.8
computer science	with	understanding	1
capable	of	intelligent actions	0.9
nature	of	intelligence	0.6
intelligence	and	constructing	0.7
dual motives	of	furthering	0.6

Table 2.4 gives the domain dictionary related to mobile domain having words and set of probable concepts. This dictionary is used for processing of set of documents of mobile domain so that the recent trends related to mobile domain can be extracted by using the recent trend database for constructing the extended document ontology as discussed in chapter 4.

Table 2.4 Domain Dictionary for Domain Mobile

Word	Set of Probable Concepts
mobile phones	phones, handsets, cell, cellular phone
system	organization, scheme, arrangement, system
official website	functionary internet site, prescribed site
phone	telephone, telephone set, headphone

Word	Set of Probable Concepts
windows	trademark
windows phone	trademark telephone, trademark telephone set, trademark handset
latest	recent
release	freeing, liberation, acquaintance
source	informant, root, beginning, origin, reference, generate
model	simulation, example, framework, model
source model	informant simulation, informant modeling, informant framework
different	unlike, distinct, dissimilar
organization	system, arrangement, establishment, formation
manufacturer	maker, producer
addition	improver, add on, summation, plus, accession
television	telecasting
things	matter, affair, entity
application	diligence, coating, covering, practical application, application
window	windows
device	instrument, gimmick, device, machine
component	element, factor, constituent
samsung	organization, samsung
open source	open resource
nosier	nosier, organization

Word	Set of Probable Concepts
scheme	system, arrangement, plan, method, idea, proposal
different organization	dissimilar system, distinct establishment, unlike formation
design	intent, aim, mean, devise, propose, contrive, plan
primarily	chiefly, mainly, principally, mostly
electronic components	electronic factor, electronic element, electronic ingredient, electronic constituent

The base ontology created having concept pairs and relationships among them related to domain travel is shown in Table 2.5 which is used in processing of web documents by our proposed probability based bi-relevance semantic rank model. This ontology is used to construct the ontology graph, page graph, and query graph.

Table 2.5 Base Ontology Related to Domain Travel

Concept Pairs	**Relation Between Concept Pairs**	**Number of Relations**
c1: destination, c2:source	from to, has part, has volvo to, has train to, has flight to, has roadways to, has public transport, to from	8
c1: destination, c3: accommodation	is a way to, has accommodation, facility, public transport, organizes visit to.	5
c2: source, c3: accommodation	is a way to, has accommodation, facility, public transport, organizes visit to.	5
c3: accommodation, c5: running	day wise, hour wise, month wise, year wise	4
c1: destination, c5: running	from to, to from	2
c3: accommodation,	has types, has ratings, has classes	3

Concept Pairs	**Relation Between Concept Pairs**	**Number of Relations**
c4:accommodation classes		
c3: accommodation, c6: booking	through credit, through cash, online booking, e-ticketing	4
c2: source, c5: running	from to, to from	2
c5: running, c6: booking	booking for hours	1
c7: tourists, c1: destination	visiting to, for education, for training, for friends, for relatives, for religion, for shopping, for business, for holiday, for profession, for health, for medical, for others	13
c2: source, c13: gurgaon	is, part of, kind of, type of	4
c2: source, c14: faridabad	is, part of, kind of, type of	4
c12: delhi, c13: gurgaon	distance, way to, hours, roadways, airways, timings	6
c1: destination, c12: delhi	is, part of, kind of, type of	4
c1: destination, c11: place	rural , urban, hilly, snowy, desert, beach, temperature, weather, hotels available, transport available, seasons	11
c12: delhi, c15chandigarh	distance, way to, hours, roadways, airways, timings	6
c12: delhi, c14: faridabad	distance, way to, hours, roadways, airways, timings	6
c2: source, c12: delhi	is, part of, kind of, type of	4
c7: tourists, c8: activity	sightseeing, sports, education, adventure, swimming, eating, enjoyment, playing	8
c3: accommodation,	has rating, one star rating, two star rating, three	10

Concept Pairs	Relation Between Concept Pairs	Number of Relations
c9: class	star rating, facility, extra benefits, three star rating, four star rating, five star rating, seven star rating	
c2: source, c15: chandigarh	is, part of, kind of, type of	4
c1: destination, c13: gurgaon	is, is, part of, kind of, type of	4
c13: gurgaon, c15: chandigarh	distance, way to, hours, roadways, airways, timings	6
c13: gurgaon, c14: faridabad	distance, way to, hours, roadways, airways, timings	6
c1: destination, c14: faridabad	is, part of, kind of, type of	4
c1: destination, c15: chandigarh	is, part of, kind of, type of	4
c2: source, c16: transport	by road, by air, bus, volvo, deluxe, train, indigo flight	9
c7: tourist, c3: accommodation	requires, booking, e booking, check in, check out, price, time, duration, facility, class, availability, type	12
c3: accommodation, c8: activity	related to, given by, incorporated, facility, price	5
c7: tourist, c16: transport	avails, facility, way to, choice, booking, running, e booking, tickets, seats, class, price	11
c3: accommodation, c6: booking	class, price, requires, needs, have to, part of	6
c6: booking, c16: transport	class, price, requires, needs, have to, part of	6
c6: booking, c9: class	includes, kind of, given by, consider, choice, availability, facility, requires	8

Concept Pairs	**Relation Between Concept Pairs**	**Number of Relations**
c4: accommodation classes, c6: booking	includes, part of	2
c14: faridabad, c15 chandigarh	is, part of, kind of, type of	4
c4: accommodation classes, c16: transport	has, includes, of, part of, given by	5
c9: class, c16: transport	includes, kind of, given by, consider, choice, availability, facility, requires	8
c16: transport, c17: schedule	timings, running, booking, from to, to from	5
c2: source, c16: transport	by road, by air, bus, volvo, deluxe, train, indigo flight, jet airways flight, rajdhani train	9
c3: accommodation, c10: budget	facility, extra benefits, breakfast, lunch, dinner, cab facility, driver facility, resources available	9
c2: source, c11: place	rural , urban, hilly, snowy, desert, temperature, beach weather, hotels available, transport available	10

BRIEF BIODATA OF RESEARCH SCHOLAR

Poonam Chahal was born in 1984. She received her Bachelor of Engineering in Information Technology in 2005 from Institute of Technology and Management affiliated to Maharishi Dayanand University Rohtak, and Master of Technology in Computer Science and Engineering in 2009 from Career Institute of Technology and Management affiliated to Maharishi Dayanand University, Rohtak. She has 10 years of teaching experience. Presently she is working as Assistant Professor in Department of Computer Science and Engineering at Faculty of Engineering and Technology, Manav Rachna International University, Faridabad.

LIST OF PUBLISHED PAPERS

SNO	Title of Paper	Name of Journal Where Published	No.	Volume and Issue	Year	Pages
1.	Web Documents Ranked using Genetic Algorithm	International Journal of Computer Applications Foundation of Computer Science	ISSN 0975-8887	Volume 70, Issue 22	2013	Pages 18-21
2.	An Ontology Based Approach for Finding Semantic Similarity between Web Documents	International Journal of Current Engineering and Technology Inpressco	ISSN 2277-4106	Volume 3, Issue 5	2013	Pages 1925-1931
3.	Comparative analysis of various approaches for Semantic Information Retrieval.	Manav Rachna International Journal of Engineering & Technology		Volume 5, Issue 2	2013	Pages 24-28
4.	An Efficient Web Page Ranking for Semantic Web	Journal of the Institution of Engineers: Series B, Springer	ISSN 2250-2106	Volume 95, Issue 1	2014	Pages 15-21
5.	Relation based measuring of Semantic Similarity of Web Documents, June 2015	International Journal of Computer Applications	ISSN 0975-8887	Volume 119, Issue 7	2015	Pages 26-19
6.	Ranking of Web Documents using Semantic Similarity.	International Conference on Information Systems and Computer Networks (ISCON), IEEE			2013	Pages 145-150

LIST OF ACCEPTED PAPERS

SNO	Title of Paper	Name of Journal	Present Status	Year
1.	Semantic Analysis Based Approach for Relevant Text Extraction Using Ontology	International Journal of Information Retrieval and Research, IGI Publications Indexed DBLP, ACM	In Press	2016
2.	Web Documents Semantic Similarity by extending Document Ontology Using Current Trends	International Journal of Web Sciences, Inderscience Indexed DBLP, ACM	In Press	2016

LIST OF COMMUNICATED PAPERS

SNO	Title of Paper	Name of Journal	Present Status	Year
1.	Semantic Similarity between Web Documents Using Ontology	Journal of Institution of Engineers: Series B (Springer)	Under Review	2016
2.	An Efficient Approach for Ranking of Semantic Web Documents Using Semantic Clustering	CSI Transaction on ICT Springer	Under Review	2017