

Ontology based Semantic Similarity Measure using Concept Weighting

S.Anitha Elavarasi,
Assistant Professor,
Department of Computer
Science and Engineering,
Sona College of Technology,

J.Akilandeswari, Ph.D
Professor & Head,
Department of Information
Technology,
Sona College of Technology,

K.Menaga,
PG Scholar,
Department of Computer
Science and Engineering,
Sona College of Technology,

ABSTRACT

Semantic similarity between the documents is essential when it is extracted from free text document. Representing the presence and absence of concept in binary format may not provide perfect accuracy. Concept weighting through term frequency will increase accuracy of clustered document. Concept weight is determined using term frequency and semantic distance. Semantic similarity of a concept is derived using ontology extracted from swoogle. Vector space model with parent-child (is-a) relationship ontology are exploited using protégé. Term frequencies for the extracted concepts are calculated using text processing. In this paper Cosine similarity using concept weight measure is applied to find similarity between different documents. According to the similarity score, documents are clustered. In this paper a sample walkthrough for the proposed system has been discussed by comparing two documents.

Keywords

Ontology, Text processing, Semantic distance, Term Frequency, Concept weight, Cosine Similarity, Clustering.

1. INTRODUCTION

Text mining plays an important role now a day, because substantial portion of information is stored in text or document format. Text mining is about knowledge discovery from large collections of unstructured text. Text mining [12] can help an organization derive potentially valuable business insights from text based content such as word documents, email and postings in social networks. Information extraction is a major component of text mining which extracts contents and structured information from unstructured text. Information extraction includes semantic web [13] (Semantic web is characterized by the use of graph and networked data). Semantic web is processed with implicit metadata and the data should be in hierarchical structure. Implicit metadata is attached with documents pointing to concepts in the ontology by graphical format.

Semantic similarity measure aims at providing meaningful document cluster. Semantic measure is obtained through Vector Space Model (VSM). It is an algebraic model used to represent the text documents as vectors. VSM processing is categorized into three stages [14]. The first stage is text processing where the content bearing terms are extracted from the document. Second stage is assigning weight measure for the extracted contents. And the third stage is to measure (semantic) similarity between each document. VSM uses binary and term frequency representation for various concepts. Term frequency (TF) is a statistical weight measure

which is used to evaluate the occurrence of words. It is measured by number of times a word occurs in the document.

Ontology represents knowledge as a set of concepts within a domain [15]. To measure similarity between documents, a number of approaches have been proposed. Although many approach do not consider semantic relations. To overcome this drawback, ontology based semantic similarity measure is defined which measures the semantic relationship between documents based on the likeliness of their meaning. Similarity computation can be categorized based on the methodology such as Lexicon and Syntax based method, Structure based method, Information based method, Feature based method and Hybrid methods [18]. Structure based method is used in this paper where the components are compared based on their distance (root to respective node) in ontology.

Clustering methods have been used in various scientific fields including statistics, machine learning and neural networks [16]. Clustering can be done either with hierarchical or iterative method. Clustering is unsupervised classification where it groups a set of data objects into clusters. Clustering will produce high quality clusters with high intra-class similarity and low inter-class similarity. The proposed system is to address the problem of binary representation of semantic similarity measure for various concepts present in the document. Concept weight with term frequency and semantic distance is used to increase the accuracy rate of the documents that are clustered.

Section 2 describes about related works. Section 3 describes architectural design of proposed system. Section 4 describes methodologies used in the proposed system. Section 5 describes about implementation detail. Finally section 6 concludes the paper.

2. RELATED WORK

Mihalcea and Corley[2] proposed a method of semantic similarity with corpus based and six knowledge based metrics which is applicable only for short texts. This uses bag of words approach which ignores many important relationships in sentence structure. Aygul and Nihan[3] proposed a query based keyphrase approach to find semantic similarity between words. SemKPsearch tool is designed for processing the query. The tool calculates word to word similarity using wordnet; an index is created for the given query and stored in semKPindex. Using index semKPsearch searches for relevant documents. Rimma Pivovarov and Elhadad [4] combined ontological and corpus based approaches by proposing a hybrid scheme, where similarity is determined using context

vectors through information from usage patterns. Similarity measure used is dependent on the Information content. Melton GB [5] explained the use of five similarity measures for medical domain to determine inter patient similarity. Ontology and information content principles are the potentially helpful tools for distance metric in similarity development, where those similarities can be applied only for medical related applications because texts are processed by MedLEE (Medical Language Extraction and Encoding system). Input data are converted to SNOMED CT codes using MedLEE with UML codes.

Garcia-Molina et.al [6] proposed a vector space model to recover the drawback of finding similarity with bag-of-words and bag of concepts approach by constructing hierarchical structure without using ontology. David Sanchez and Montserrat Batet [7] proposed semantic similarity measure in terms of concept and information content. This approach is to mimic human judgments about semantic similarity. Final accuracy depends on the completeness and coherency of taxonomical knowledge. Ming Che Lee[8] evaluates the semantic similarity between irregular sentences based on similarity evaluation algorithm using part of speech approach. Exponential balance coefficient for evaluation algorithm is flexible but not accurate for long sentences. Varun chandola et.al [9] made comparative evaluation of different similarity measure used for categorical data. The measures are described based on three different classifications such as diagonal entries, off diagonal entries and both diagonal & off diagonal entries. In each classification fourteen different similarity measures for categorical attributes were discussed.

David Sánchez[10] introduced a new methodology to find ontological concept attributes through web where the methodology is automated and unsupervised. This proposed method is domain independent and doesn't depend on corpus structure. Batet M and Snachez D[11] proposed a new measure based on the taxonomical structure and Information Content (IC) of a biomedical ontology. Depth and path length of concept in the ontology are considered to measure least common subsumer (LCS). Then IC similarity measure with the LCS is founded and applied to cosine measure which finds the similarity based on IC and context vector.

Thusitha Mabotuwana [1] proposed a concept to determine similarity between the concepts and thus semantic similarity is used to measure similarity between documents by assigning weight to concepts. Vector Space model is the standard measure for measuring similarity between documents. Here vector is in binary format which contains 1's and 0's for representing presence and absence of concept in the document respectively. Concept weight is measured by including concept from ontology in the document. While measuring weight of a concept from the document, ancestors (the concepts in the graph) from seed to concept in document are assigned with weight. Then cosine similarity is assigned to measure similarity between the documents. However the similarity measure is not much accurate without measuring the term frequency.

3. ARCHITECTURAL DESIGN

The architectural design of the proposed system is shown in Fig.1. It addresses the problem of binary representation of semantic similarity measure for various concepts present in the document. Semantic similarity of a concept is derived using ontology extracted from swoogle which act as a knowledge source. The concepts are extracted from

documents through text processing. Vector space model with parent-child (is-a) relationship ontology are exploited to

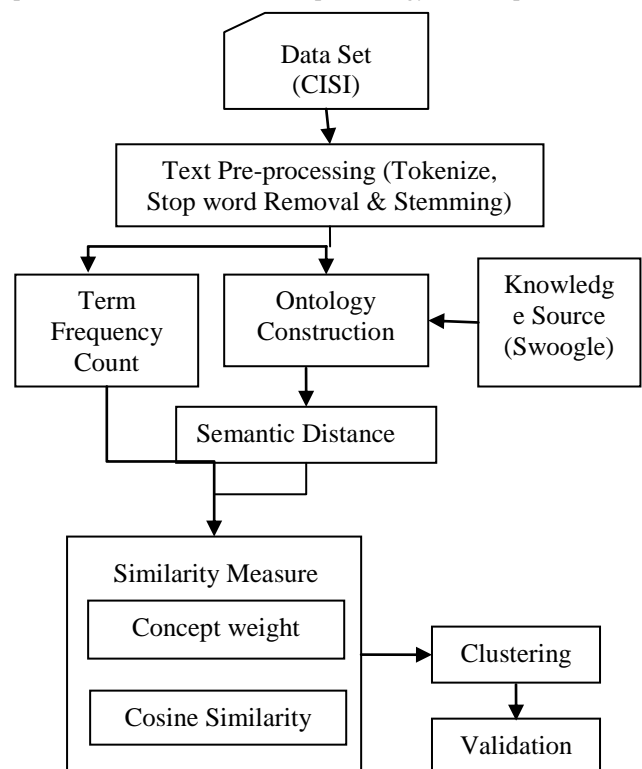


Figure 1: Architectural design of proposed system

measure the semantic distance. Term frequencies for the extracted concepts are calculated. Concept weight is determined using term frequency and semantic distance. Cosine similarity using concept weight measure is applied to find similarity between different documents. According to the similarity score, documents are clustered. Concept weight based clustering increase the accuracy rate of the documents. The module for the proposed system includes

1. Text pre-processing
2. Term frequency computation
3. Ontology Construction
4. Semantic similarity computation
5. Concept weight based Cosine Similarity
6. Clustering
7. Validation

4. METHODOLOGY

4.1. Text Pre-processing

Text processing converts text to produce a set of indexing terms [17]. Text processing is implemented by rapid miner tool [19]. Steps followed in text processing are:

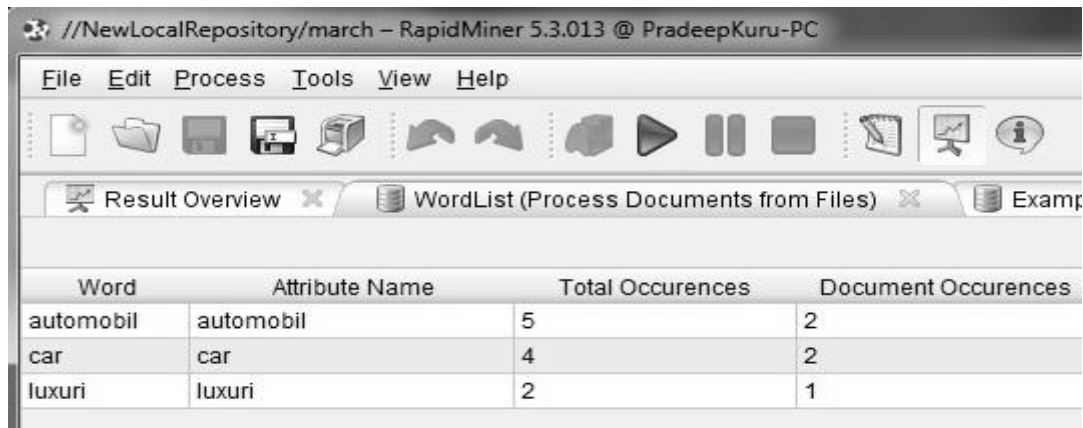
1. Tokenize the document by listing every word in the document as tokens (Token construction).
2. Stop word removal.
3. Stemming using port stemmer

SAMPLE WALKTHROUGH:

Consider two documents X and Y. Contents in the documents are,

X → Car is an automobile. Car Automobile can have luxury cars. Luxury is used.

Y → Automobile is simple term which also mention car. Automobile car works well. Automobile is more famous.



Word	Attribute Name	Total Occurrences	Document Occurrences
automobil	automobil	5	2
car	car	4	2
luxuri	luxuri	2	1

Figure 2: Text preprocessing

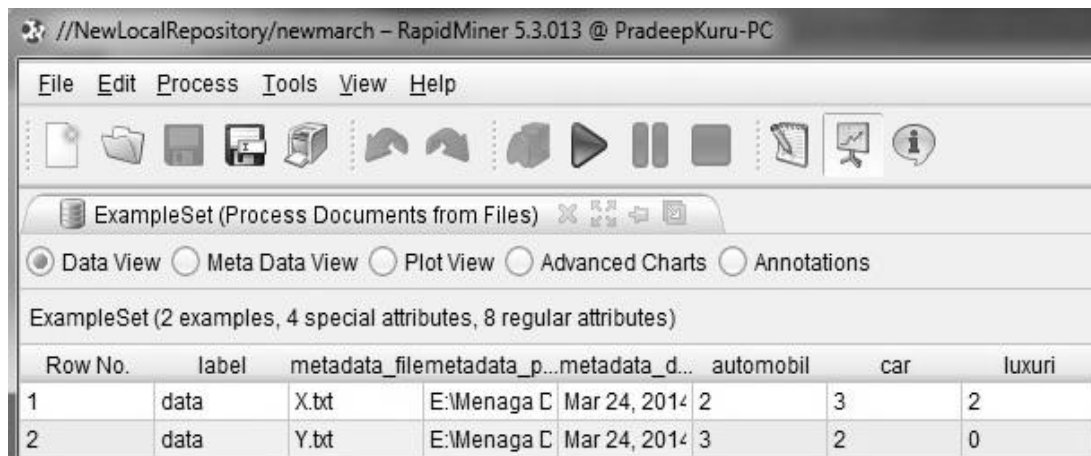
At the end of text processing following terms are extracted,
 X → {automobile, car, luxury} → (doc 1)
 Y → {automobile, car} → (doc 2)

4.2. Term Frequency Computation

Term frequency represents the number of occurrence of concept in documents. Concepts extracted from text processing are represented as vectors, where they are measured with TF. TF counts for the documents enable to get higher accuracy rate. The term frequency for the document X and Y are shown in table 1.

Table 1. Term Frequency Count

	Concept	TF
Doc X	Automobile	2
	Car	3
	Luxury	2
Doc Y	Automobile	3
	Car	2
	Luxury	0



Row No.	label	metadata_file	metadata_p...	metadata_d...	automobil	car	luxuri
1	data	X.txt	E:\Menaga C	Mar 24, 2014	2	3	2
2	data	Y.txt	E:\Menaga C	Mar 24, 2014	3	2	0

Figure 3: TF measure through Rapid Miner

4.3. Ontology Construction

Ontology represents knowledge as the set of concept. Swoogle is a semantic web search engine. The knowledge source (OWL file) is extracted from Swoogle for constructing ontology in the form of Ontograph using protégé [20] to measure semantic distance.

Semantic weight is measured by finding shortest distance for each concept in the constructed ontograph. Thus a fragmentation of automobile ontograph is shown below.

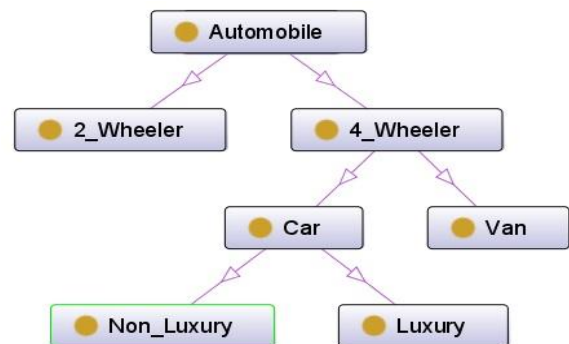


Figure 4: Construction of OntoGraf from OWL file

4.4. Semantic Similarity Computation

fileNme	word	Term Counts
med000865.txt	phosphate	2
med000865.txt	plasma	2
med000865.txt	disease	1
med000865.txt	found	1
med000865.txt	hypertrophy	1
med000865.txt	osteodystrophy	1
med000865.txt	plasmacalcium	1
med000865.txt	prove	1
med000865.txt	renal	1
med000865.txt	significantly	1
med000865.txt	survey	1
med000905.txt	strain	35
med000905.txt	viru	29

Founded Records : 2214

Figure 8: Calculated TF for the dataset

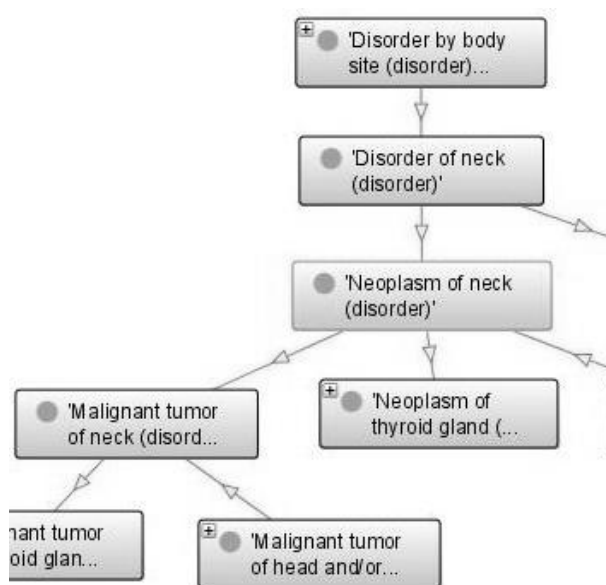


Figure 9: Ontograph for SNOMED-CT ontology using Protégé

6. CONCLUSION

In this paper semantic similarity with vector space model is used to measure similarity between the documents. Binary representation does not consider the frequency of occurrence for a concept. In the proposed system concept weighting through term frequency is used to get higher similarity which includes every significant terms and their number of occurrence in the document. For measuring shortest distance only concept present in the document are assigned with weight instead of assigning weight to every term in the particular ontograph. By concept weight based similarity measure accuracy rate and time consumption is increased.

Text processing and CWCS measure sample walkthrough is discussed in previous sections. The proposed system has been implemented for the benchmark dataset (CISI-Chartered Institute For Securities And Investment) and results are yet to be validated.

7. REFERENCES

- [1] Thusitha Mabotuwana, Michael C.Lee, Eric V.Cohen-Solal. An ontology-based similarity measure for biomedical data-Application to radiology reports. Journal of Biomedical Informatics; 2013.<http://dx.doi.org/10.1016/j.jbi.2013.06.013>
- [2] Mihalcea R, Corley C, Strapparava C. Corpus-based and knowledge-based measures of text semantic similarity. Proceedings of the 21st national conference on artificial intelligence, vol. 1. Boston, Massachusetts: AAAIPress; 2006. p. 775–80.
- [3] Aygul I, Cicekli N, Cicekli I. Searching documents with semantically related keyphrases. In: Sixth international conference on advances in semantic processing; 2012. p. 59–64.
- [4] Pivovarov R, Elhadad N. A hybrid knowledge-based and data-driven approach to identifying semantically similar concepts. Journal of Biomedical Informatics. 2012;45(3):471–81.
- [5] Melton GB et al. Inter-patient distance metrics using SNOMED CT defining relationships. Journal of Biomedical Informatics. 2006;39(6):697–705.
- [6] Ganesan P, Garcia-Molina H, Widom J. Exploiting hierarchical domain structure to compute similarity. ACM Trans InfSyst 2003;21(1):64–93.
- [7] David Sanchez , Montserrat Batet. Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. Journal of Biomedical Informatics 44 (2011) 749–759.[doi:10.1016/j.jbi.2011.03.013](https://doi.org/10.1016/j.jbi.2011.03.013)
- [8] Ming Che Lee. A novel sentence similarity measure for semantic-based expert systems. Expert systems with application 38 (2011) 6392–6399. [doi:10.1016/j.eswa.2010.10.043](https://doi.org/10.1016/j.eswa.2010.10.043).
- [9] ShyamBoriah, VarunChandola, Vipin Kumar. Similarity Measures for Categorical Data: A Comparative Evaluation. Department of Computer Science and Engineering. University of Minnesota.<sboriah,chandola,kumar@cs.umn.edu>
- [10] David Sánchez. A methodology to learn ontological attributes from the Web. Proceedings at Intelligent Technologies for Advanced Knowledge Acquisition (ITAKA). 26.43007 Tarragona, Spain. Journal of Data and knowledge Engineering. 69 (2010): 573–597. [doi:10.1016/j.datak.2010.01.006](https://doi.org/10.1016/j.datak.2010.01.006).
- [11] Montserrat Batet , David Sánchez, Aida Valls. An ontology-based measure to compute semantic similarity in biomedicine. Proceedings at Intelligent Technologies for Advanced Knowledge Acquisition (ITAKA) Research Group, Department d'Enginyeria Informatics Matemàtiques, Universitat Rovira Virgili, Tarragona, Catalonia, Spain. Journal of Biomedical Informatics 44 (2011): 118–125.
- [12] <http://searchbusinessanalytics.techtarget.com/definition/text-mining>.

- [13] Text mining and the Semantic Web - http://gate.ac.uk/sale/talks/text_mining_manchester05.ppt&sa=U&ei=B_xpUtuH0S4rgeH_oDABw&ved=0CBwQFjAA&usg=AFQjCNG_wjZkFbjkqENVQtLcrdgQ5CobYA.
- [14] Vector Space Model: <http://cogsys.imm.dtu.dk/thor/projects/multimedia/textmining/node5.html>.
- [15] What is an Ontology? - Stanford Knowledge Systems Laboratory: <http://www-ksl.stanford.edu/kst/what-is-an-ontology.html>.
- [16] http://www.stat.columbia.edu/~madigan/W2025/notes/clustering.pdf&sa=U&ei=Yf5pUumrKpHJrAeVuYHYDw&ved=0CBoQFjAA&usg=AFQjCNEqfrRx_Q4oAwusW3K3-DUHIFqV2Q:What is Cluster Analysis?
- [17] http://www.csee.umbc.edu/~ian/irF02/lectures/03Text-Processing.pdf&sa=U&ei=j_5pUtiyLYazrgePqIHYCQ&ved=0CCMQFjAB&usg=AFQjCNGCfUzmAQArYRIPcJdyJ0nnW6qpHQ:Text Processing
- [18] Paris D. HosseinZadeh, Marek Z. Reformat. Assessment of semantic similarity of concepts defined in ontology. *Journal of Information Sciences* (2013) Elsevier Inc. doi:10.1016/j.ins.2013.06.056.
- [19] Rapid miner 5.3 download: http://rapid-i.com/content/view/17/211/lang,en/installation_guide.
- [20] http://protege.stanford.edu/download/protege/4.3/installanywhere/Web_Installers/developer.doc.
<http://protegewiki.stanford.edu/wiki/Protege4DevDocs>.