Ontology Based Document Data Analysis

Ambreen Zafar, Muhammad Awais and **Muhammad Ahmad Aftab**

Dept. Software Engineering, GCUF, Faisalabad, Pakistan

Summary

A vast amount of data is generating at a rapid pace over the internet by means of blogs, online forums and emails etc. The huge volume and complex semantics of unstructured data initiates the need of effective management for efficient retrieval. It is intricate for users to find right keywords for search to retrieve relevant search results. There also exist polysemous words in the vocabulary of every natural language i.e. words contributing different meaning according to the context. Additional relations among words such as super-subordinate relation (hypernym/hyponym) and part-whole relation (meronym/holonym) can also be incorporated to capture the semantics of user's query. The concept of document clustering along with ontology provides users with the opportunity to overcome difficulties associated with traditional keyword based search. It intends to reduce search time and enhance the retrieval of relevant documents. This research proposes a semantic-based document clustering technique by applying K-means clustering algorithm over concept weight matrix, computed using modified TF-IDF approach. The weights are calculated specifically for the features and their relations extracted from WordNet ontology. Silhouette coefficient is used as a measure of cluster purity.

Key words:
Clustering, Ontology, WordNet, Concept Weight, TF-IDF

1. Introduction

In this age of IT, the concept of ontology is widely utilized in a variety of applications like Information Retrieval (IR), Natural Language Processing (NLP), Artificial Intelligence (AI) and Internet of Things (IoT) etc. The fields of medical, agriculture, finance, civil engineering, education and computer science are making extensive use of domain specific ontologies [1].

Ontology is defined as a technique that provides domain specific descriptions of shared conceptualization in a formal and explicit manner. The concepts provided by ontology are machine interpretable and constrained (as per their application) [2].

Ontology development differs from creating classes and relations in object-oriented paradigm. Class methods in object-oriented paradigm depend upon the operational characteristics, while ontology is based on the structural properties of a class [3].

1.1 Components of Ontology

The components of ontology include basic concepts of a domain, their definitions and relationships among them described by means of expressions/terminologies of targeted domain that constraint their interpretation. Ontology is comprised of:

- Basic concepts about a domain i.e. classes
- Individuals i.e. class instances
- Relations between concepts
- Attributes (name/type-value) to characterize concepts and relations; slots/roles
- Logical expressions that restrict or constrain attributes and relations (axioms); facets [4].

Ontology is a knowledge base containing objects of classes. In order to add more specifications to a class – the concept of sub-class is used.

Some common terms related to ontology along with their definitions are [5]:

- **Concept:** An idea that represents some class or entity related to a specific domain.
- **Holonym:** It denotes a concept as a whole; part of which is denoted by another concept.
- **Meronym:** A concept that denotes part or member of another concept.
- **Hypernym:** A concept having a broader meaning; parent or super-ordinate.
- **Hyponym:** A concept having more precise meaning than its corresponding super-ordinate.
- **Semantic:** Related to meaning with respect to context or specific language.
- **Synonym:** A word or expression that conveys exactly or almost similar meaning to some other word of similar language.

1.2 Levels of Ontology

Ontology can broadly be placed into three categories as per the content of business domain knowledge.

- Light weight
- Light heavy weight
- Heavy weight

Light weight ontology specifies the terms (concepts) and their hierarchy using particular associations like is-a

Manuscript received November 5, 2018
Manuscript revised November 20, 2018

hierarchy and part-whole relations etc. Light heavy weight ontology has added constraints which enforce restrictions on concepts' values and associations like constraints on length of a value and cardinality constraints etc. Heavy weight ontology contains axioms that constraint complex relations between concepts and interpretation of those concepts.

A business system modeling mostly incorporates heavy weight as well as light heavy weight ontologies, because they mostly have constraints and axioms, which are to be applied for business regulations [4].

1.3 Ontology Construction Methodologies

Ontology can be constructed in three possible ways:

- **Manual:** Ontology is created manually with the help of some tool like Protégé or OntoEdit in an Integrated Development Environment (IDE).
- **Semi automatic:** Human involvement exists during ontology development process for validation of concepts and relations suggested by the algorithm.
- **Automatic:** The system handles complete process of ontology construction without any human interaction.

Automatic and semi-automatic ontology generation tools like OntoLearn, OntoLT, Text2Onto, OntoGen, OntoGain and OntoPlus etc make use of various input sources like plain text, html, xml etc and generate ontology by applying linguistic or statistical approaches [6].

1.4 K-means Clustering

Document Clustering is a technique to group together documents belonging to a specific category. Documents having similar concepts are placed in the same group, while others are placed in different group. This grouping of similar set of information reduces search time and assists users to retrieve their required information more effectively and precisely.

It is an un-supervised type of machine learning which enables a computer to train developer through learning of data patterns. These algorithms are intended to be used in scenarios where even the experts are un-aware of content to be searched in data. There exists no label or category upon which the algorithm can model interconnection between input and output data patterns.

This research applies K-means clustering algorithm over concepts extracted from WordNet ontology, specifically for the domain of "finance audit". The incorporation of WordNet ontology adds more semantics to the clusters and thus facilitates users to search with more versatility. This algorithm works by assigning data points to their closest centroid and iterating the process till final convergence. The number of clusters "K" is to be explicitly specified (see Fig 1).

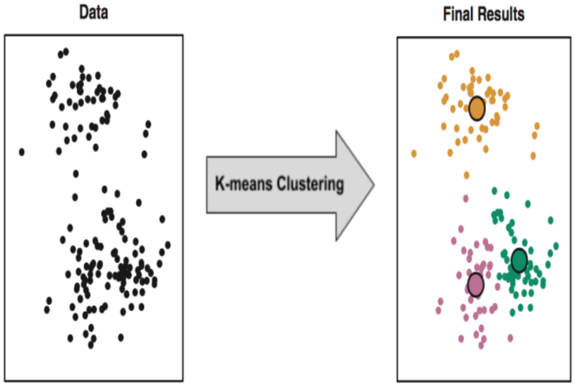


Fig. 1 K-means Clustering

1.5 WordNet

WordNet is a lexical catalog of English language which groups together various Parts Of Speech (POS) including nouns, adjectives, verbs and adverbs into sets of equivalent words called synsets. It additionally provides short definition called gloss against each synset and semantic relations (hypernym, hyponym,holonym and meronym and so forth) between synsets. It has been developed by Cognitive Science Laboratory of Princeton University.

WordNet evidently takes after a thesaurus, as words are gathered together in light of their implications. WordNet interlinks particular classes of words and deals with semantic relations among words, while the groupings of words in a thesaurus do not take after any explicit illustration other than concept similitude. WordNet's structure makes it a valuable resource for computational semantics and NLP.

A large number of relations in WordNet bond words from similar POS along with few cross-POS pointers. Cross-POS relations consolidate the "morphosemantic" affiliations that hold among semantically comparable words imparting a stem to a similar importance like observe representing verb, observation is nouns and observant representing adjective [7].

1.6 Concept Weight Matrix

The weights of the concepts extracted from WordNet ontology in a corpus are computed using TF-IDF approach. It is a numerical measurement that reflects how vital a word is to a report in an accumulation or corpus. It computes the weight of a term in a document of corpus such that:

- If a term occurs frequently in a limited number of documents, its weight will be the highest.

- If a term occurs in a limited number of times in a document or exists in a large number of documents, its weight will be lower.
- If a term exists in a large number of documents its weight will be the lowest.

Documents are first preprocessed using NLTK. It is a top framework to develop software for representation of human language information. It provides users with a friendly interface and a large number of packages that cover up to 50 corpora and lexical resources like WordNet (for handling of semantic, lexical and morphological associations of concepts) along with support of multiple NLP tasks like text preprocessing and classification etc [8]. NLP is a domain of AI that manages the connections between machines and natural languages. It makes computers capable of analyzing and processing natural language data i.e. making machines intelligent enough to capture the semantics behind natural languages.

NLP deals with two major characteristics of natural language which are [9]:

- **Phonology:** Organized arrangement of sounds in a language.
- **Morphology:** Technique of formation of words and relationships among words.

The key features of a natural language that must be taken into account are as follows [9]:

- **Lexical Ambiguity:** Polysemous words i.e. words with more than one meaning.
- **Syntactic Ambiguity:** Sentence generating several parse trees.
- **Semantic Ambiguity:** Sentence with more than one meanings
- **Anaphoric Ambiguity:** A Phrase or expression whose meaning depends upon some other phrase or expression.

TF-IDF calculated using traditional method is further modified as per WordNet relations.

2. Related Work

In [10], suffix array algorithm is incorporated to identify the common phrases from the document. Vector Space Model (VSM) is used along with Singular Value Decomposition (SVD) which is implemented over Latent Semantic Indexing (LSI) for document representation.TF-IDF technique is used to compute the weights of terms in the document. The terms matrix incorporates key concepts among collection of documents. Cluster labels are formed by means of WordNet ontology.

In [11], a document clustering technique based on ontology is implemented and its performance is compared with conventional clustering techniques. The system makes use of ontology specifically for e-learning domain. Ontology is constructed by utilizing a huge group of papers from well-known conferences. The clustering proposed here consists of three phases which are, preprocessing of documents, weight calculation of concepts and finally generation of clusters according to computed weights. A modified TF-IDF approach is used for weight calculation. A two-phase clustering approach is applied for clustering. An approach which resembles k-means is implemented in first phase. The input is a set of vectors containing weights of concepts. In second phase, the clusters generated in first phase are transformed to homogenous clusters.

In [12], bisecting k-means is implemented using MapReduce framework. The purpose behind it is to propose a system which resolves the clustering issue of data concentrated documents. Along with this, bisecting K-means clustering algorithm is integrated with WordNet to capture semantic relation between words in order to improve clustering process. For the purpose of experimentation, Elastic MapReduce is used to deploy bisecting k-means algorithm. The use of semantic relations of WordNet reduces the dimension of features and clustering of big data became visible because of reduction in number of dimension. Lexical categories of WordNet are also incorporated for nouns which enhance the internal performance measure.

The system in [13] incorporates Medical Subject Headings (MeSH) ontology for the implementation of concept retrieval. This ontology is published by the National Library of Medicine. It is a controlled vocabulary of biomedical literature which contains many different types of terms i.e. subject headings called as descriptors, short descriptions and links between these terms, terms similar to descriptor and synonym list. Information can be retrieved at any level of hierarchy. Proposed indexing scheme uses descriptors and entry terms. Main concepts and headings are descriptors while synonyms or terms related to main concepts and headings are entry terms. K-mediod clustering algorithm is finally applied to perform clustering that forms k clusters from a dataset containing n objects.

In [14], the concepts in documents are located in the MeSH ontology for their presence. If concept exists in the ontology then all of its relations are extracted. Their complete path is traced from root of ontology till leaf. All the terms in the path are allocated dynamic weights. Children terms of the concept contribute more towards semantics as compared to parent terms as they are more specific compared to parent/generic terms. Cosine similarity is used to measure similarity between documents. K-means clustering algorithm is finally applied to cluster similar documents.

The essential goal of [5] is to comprehend the essential ideas behind ontology with explicit emphasis on its application to clustering process. The algorithm first computes the distance between objects by means of techniques specified by ontology. Each object incorporates a distinct cluster. The clusters which are meant to be closest corresponding to distance matrix are merged. Distance matrix is computed again using merged objects. This process continues till the achievement of required number of clusters.

3. Methodology

The process of document clustering proposed in this research is represented in Fig 2.

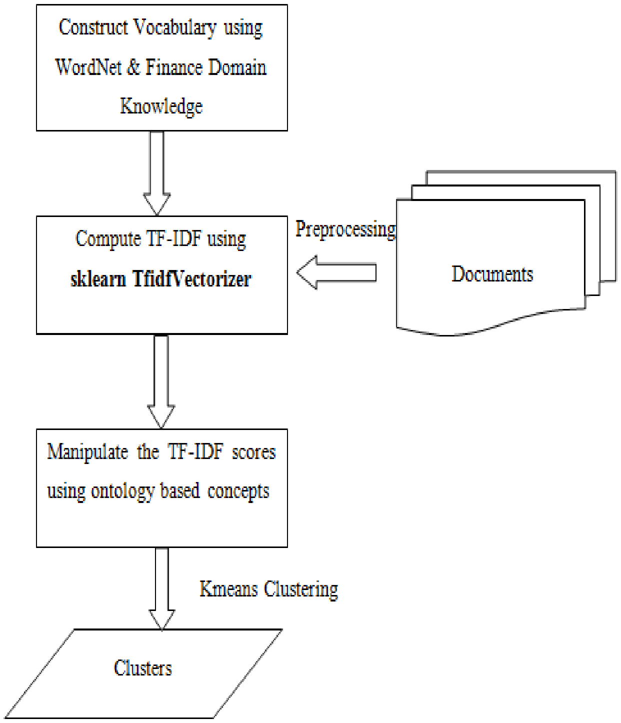
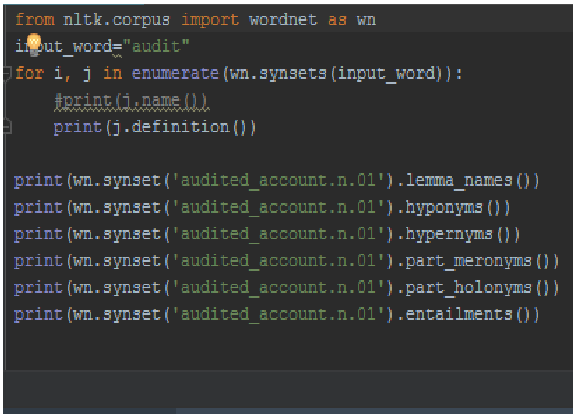

Construct Vocabulary using WordNet & Finance Domain Knowledge
Compute TF-IDF using **sklearn TfidfVectorizer**
Preprocessing
Documents
Manipulate the TF-IDF scores using ontology based concepts
Kmeans Clustering
Clusters

Fig. 2 Document Clustering Using WordNet

1. Concept vocabulary is generated with the help of WordNet Ontology. All the senses of "audit", their hypernyms, hyponyms, meronyms and holonyms along with finance related terms in their respective glosses are enumerated to generate a list of concepts as presented in Fig 3.



```
from nltk.corpus import wordnet as wn
input_word="audit"
for i, j in enumerate(wn.synsets(input_word)):
    #print(j.name())
    print(j.definition())

print(wn.synset('audited_account.n.01').lemma_names())
print(wn.synset('audited_account.n.01').hyponyms())
print(wn.synset('audited_account.n.01').hypernyms())
print(wn.synset('audited_account.n.01').part_meronyms())
print(wn.synset('audited_account.n.01').part_holonyms())
print(wn.synset('audited_account.n.01').entailments())
```

Fig. 3 Concept Vocabulary Generation Using WordNet

2. Corpus over which clustering algorithm is applied is built using annual reports of banks (MCB and HBL) of Pakistan.Preprocessing of documents is performed in following steps
   - A document is first splitted into sentences i.e. "sentence tokenization". NLTK "PunktSentenceTokenizer" is employed for this purpose. It uses an unsupervised machine learning algorithm and is trained on a number of European languages including English.
   - Words are then extracted from sentences referred to as 'word tokenization" using NLTK "WordPunctTokenizer" and stop words are eliminated.
   - From list of words obtained, POS tagging is performed. NLTK POS tagger is used here for attaching respective POS to each token of the documents. This tagger makes use of Penn Treebank tagset available in English language.
   - After identification of respective POS for each word; nouns and verbs are extracted by means of regular expressions.
   - Lemmatization is performed at the end. We preferred lemmatizer over stemmer since stemming calculations work by removing the end or the start of the word, considering a list of basic prefixes and postfixes that exist in the inflected word for example the word "organization" is stemmed into "organ". Lemmatization, on the other hand, considers the morphological investigation of the words. To do such, the algorithm must refer some thesaurus or dictionary to bring back the word to its actual root form. The lemmatizer used here makes use of WordNet to return back a word to its original lemma. The steps for document preprocessing are presented in Fig 4.

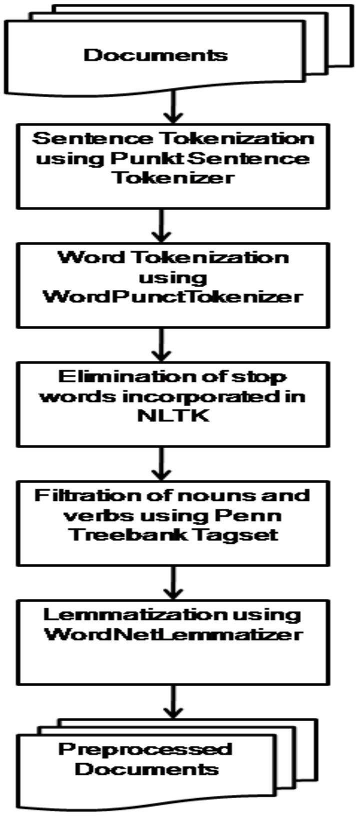

Documents
Sentence Tokenization using Punkt Sentence Tokenizer
Word Tokenization using WordPunctTokenizer
Elimination of stop words incorporated in NLTK
Filtration of nouns and verbs using Penn Treebank Tagset
Lemmatization using WordNetLemmatizer
Preprocessed Documents

Fig. 4 Document PreProcessing

Weight of a concept/term "t" in a document is calculated as:

$$W(t) = TF(t) * IDF(t) \qquad (1)$$

Where

$$TF(t) = \frac{number\ of\ times\ "t"\ appear\ in\ a\ document}{total\ terms\ in\ the\ document}$$

$$IDF(t) = \log\ \left(\frac{total\ number\ of\ documents}{documents\ having\ term\ "t"}\right)$$

In order to check the impact of WordNet based vocabulary, different approaches were applied and their impact on resulting clusters was analyzed.

- Computation of TF-IDF scores without defined vocabulary
- Computation of TF-IDF scores with defined vocabulary
- Summation of TF-IDF scores of relevant concepts.
- Selection of the highest TF-IDF score among relevant concepts.

3. K-means Algorithm performs clustering by grouping together samples having same variance. We need to specify the number of clusters to be formed. The output is "k" clusters represented by centers of clusters. It is an iterative process which repeats its steps till final convergence. Its aim is to maximize inter cluster similarity and minimize intra cluster similarity. Clustering is performed in the following steps [15]:

1. Selection of initial centroids.
2. Samples are allocated the closest centroid.
3. New centroids are computed by taking mean of samples allocated to previous centroid.
4. Centroids are computed until their difference from previous ones' is minimized i.e. maximum convergence.

Performance of the proposed approach is measured by means of silhouette coefficient. It is used to measure the purity of a cluster to indicate how closely an object is associated with its respective cluster. The formula of silhouette coefficient is [16]:

$$SC = \frac{y-x}{max\ (x,y)} \qquad (2)$$

Where,
x: average distance between sample and remaining points of same class
y: average distance of sample and points of the closest class

4. Results

The resulting concept weight matrices are represented in Fig 5, Fig 6 and Fig 7.

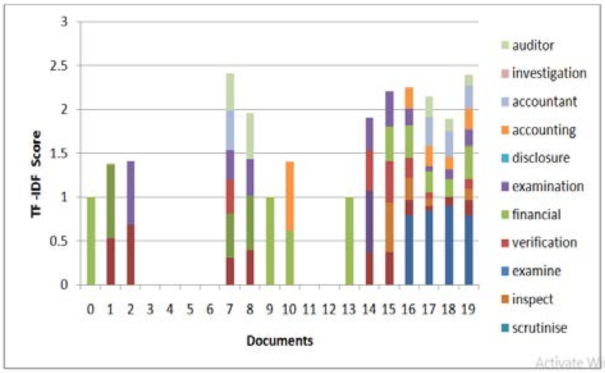


Fig. 5 Concepts' Weights for Defined Vocabulary

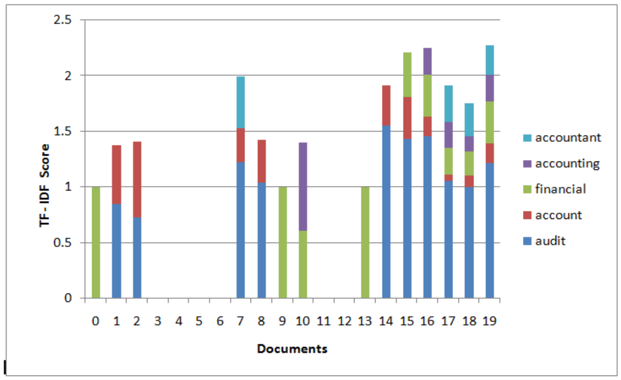


Fig. 6 Concepts' Weights after merging relevant concepts

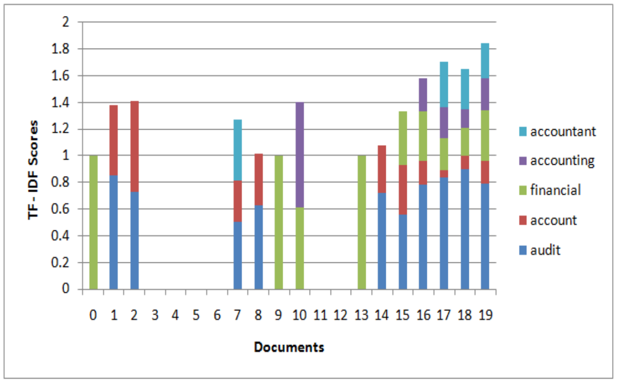


Fig. 7 Concepts'Weights by selecting the highest score among relevant concepts

Since clustering is an un-supervised type of machine learning so, we have computed silhouette coefficient by varying the number of clusters to determine the purity of clusters obtained as presented in Fig 8.

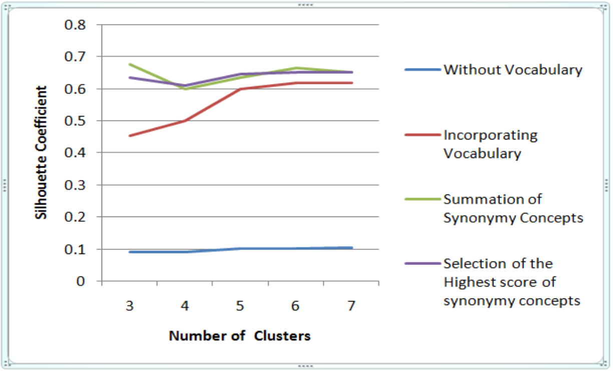


Fig. 8 Analysis of Silhouette Coefficient

5. Conclusion

On the basis of Silhouette coefficient, calculated over gradually increasing number of clusters by means of TF-IDF matrix it is concluded that clusters formed by summation of TF-IDF scores of synonymy concepts and selection of highest score among synonymy concepts have the highest and almost similar purity level. In Addition to this, clusters formed without any modification to TF-IDF scores,computed using predefined vocabularies of concepts are also extremely pure in contrast to clusters without any defined vocabulary. Hence the addition of relations among concepts as provided by WordNet ontology, to the document clustering process improves the purity of clusters. This research could be enhanced by incorporating other areas of finance domain in concept vocabulary like financial statements invoice and audit rules etc for further up gradation of clustering process. Along with this, the addition of more hierarchy levels of concepts in the vocabulary and calculation of their weights according to their position, for the generation of concept weight matrix can also be incorporated to analyze their impact on resulting clusters.

References

[1] Meenachi, N. M., & Baba, M. S. (2012). A survey on usage of ontology in different domains. Int. J. Appl. Inf. Sys, 9, 46-55

[2] Soe, T. L. (2014). Ontology-Based Indexing and Semantic Indexing in Information Retrieval Systems. International Journal of Research Studies in Computer Science and Engineering (IJRSCSE), 1, 1-9. Sridevi, U. K., & Nagaveni, N. (2009, October). Ontology based semantic measures in document similarity ranking. In Advances in Recent Technologies in Communication and Computing, 2009. ARTCom'09. International Conference on(pp. 482-486). IEEE.

[3] Noy, N. F., & McGuinness, D. L. (2001). Ontology development 101: A guide to creating your first ontology

[4] Kalibatiene, D., & Vasilecas, O. (2011, October). Survey on ontology languages. In International Conference on Business Informatics Research (pp. 124-141). Springer, Berlin, Heidelberg.

[5] Punitha, S. C., Mugunthadevi, K., & Punithavalli, M. (2011). Impact of ontology based approach on document clustering. International Journal of Computer Applications (0975–8887) Volume

[6] Esserhrouchni, O. E. I., Frikh, B., & Ouhbi, B. (2014, May). Building ontologies: a state of the art, and an application to finance domain. In Next Generation Networks and Services (NGNS), 2014 Fifth International Conference on (pp. 223-230). IEEE.

[7] https://wordnet.princeton.edu/

[8] https://www.nltk.org/

[9] https://www.xenonstack.com/blog/data-science/ai-nlp-big-deep-learning/

[10] Kolhe, S. R., & Sawarkar, S. D. (2017, January). A concept driven document clustering using WordNet. In Nascent

Technologies in Engineering (ICNTE), 2017 International Conference on (pp. 1-5). IEEE.

[11] Alaee, S., & Taghiyareh, F. (2016, April). A semantic ontology-based document organizer to cluster elearning documents. In Web Research (ICWR), 2016 Second International Conference on (pp. 1-7). IEEE

[12] Elsayed, A., Mokhtar, H. M., & Ismail, O. (2015). Ontology based document clustering using mapreduce. arXiv preprint arXiv:1505.02891.

[13] Gupta, S., & Chole, V. (2014). Document Clustering Using Concept Weight. International Journal of Computer Science and Mobile Computing, 3(5), 1207-1210.

[14] Logeswari, S., & Premalatha, K. (2013, January). Biomedical document clustering using ontology based concept weight. In Computer Communication and Informatics (ICCCI), 2013 International Conference on (pp. 1-4). IEEE.

[15] http://scikit-learn.org/stable/modules/clustering.html

[16] http://scikit-learn.org/stable/modules/generated/ sklearn.metrics.silhouette_score.html