ISSN: 0975-766X
CODEN: IJPTFI
Available Online through *Research Article*
www.ijptonline.com

HARMONIZING SEMANTICALLY SIMILAR SENTENCES BASED ON LEXICAL DATABASE –WORD NET AND FUZZY CLUSTERING FRAMEWORK

M. Uma Devi[1*] , G.Meera Gandhi[2]
[1]Assistant Professor , SRM University ,Chennai, Tamil Nadu, India.
[2]Professor, Faculty of Computing, Sathyabama University ,Chennai, Tamil Nadu, India.
Email: umadevi.as2006@gmail.com

Received on: 20.10.2016 *Accepted on: 25.11.2016*

Abstract

This paper is focusing on statistical measure of matching Similar Sentences using a Lexical Database (WordNet) and Fuzzy clustering framework is implemented that systematize data from one or more documents into different clusters. The conventional fuzzy clustering approach is not applicable to sentence clustering and so a WordNet and Fuzzy clustering algorithm is developed and used over the sentence of document datasets to match the related sentences. The term which donates the semantics of the sentence is examined on the sentence which can professionally find considerable matching concepts between documents. The semantic similarity of two sentences is calculated using information from a structured lexical database (WordNet) . This system is applied over Quotations dataset that find the Similarity measure of matching Semantically Similar Sentences. Our proposed system out performs the baseline method and contribute 25 % higher in similarity scoring of related sentences. The clustering performance in terms of Entropy and Purity is analyzed and this system yields more Purity and less Entropy.

Our investigational outcome reveals that this method is capable of relating the semantically similar sentences. The proposed method can be used in a variety of applications that involve text knowledge representation and discovery

Keywords: Lexical Database , Similar Sentences , Fuzzy Clustering, Entropy,Purity.

1. Introduction

Information overload is a tedious problem with the rapid growth of World Wide Web. Finding similar sentences is an essential issue for many applications, such as text summarization, snippet extraction, image extraction, question-answer model , social media retrieval, document retrieval and on. For a given document collection, one can determine how to effectively and efficiently identify the top- 'n' semantically similar sentences to a query. Multiple sentences often may

contain duplicate information containing the same event. Use the clustering method for the task of grouping the text spans in multiple documents that refer to the same event.

Sentence clustering plays an important role in many text processing activities and can also be used within most of the text mining tasks. For a given query term, to discover some novel information from a set of documents clustering at sentence level can be done. By clustering the sentences of these documents, we would instinctively anticipate at least one of the clusters to be directly associated to the concepts described by the query terms.

Certain sentences repeat some of the information present in other sentences and may, therefore, be considered Similar. If the information content of sentence x (denoted as i(x)) is contained within sentence y, then the content of y is said to subsum that of x: and it is represented as follows.

i(x) ⊂ (elements of) i(y)

A murder of a person may be

Johnny was found guilty of the murder.	Document -1.
The court found Johnny guilty of the murder	Document -2.

In the example above, Document (2) subsumes (1) therefore, (1) and (2) can be considered Similar. The above two sentences contain the same event.

This work depicts, how Semantically Similar sentences are matched and evaluated based on a Fuzzy sentence clustering scheme. Document Clustering is a well-established process in the Information Retrieval (IR) literature. The documents typically represent the data points in a high dimensional vector space in which each dimension corresponds to a unique keyword. This leads to a rectangular representation in which each rows represent the documents and each columns represent attributes of the documents. In sentence clustering, a sentence is likely to be related to more than one theme or topic present within a document or set of documents. In hard clustering data are grouped in an exclusive way so that a data can belong to a single cluster, whereas in fuzzy clustering each data can belong to more than one clusters with some degrees of membership. Each clustering has a prediction error on the predictions. The best clustering is the one that minimizes this prediction error. Most documents will contain interrelated topics irrespective of the specific task such as summarization, text mining and many sentences related to some degree to a number of documents. The work described

in this paper is being able to capture the fuzzy relationships which lead to an increase in the scope of problems where sentence clustering shall be applied.

2. Materials and Methods Used

2.1 Sentence Clustering

Many text processing activities uses Sentence clustering over extractive multi-document summarization to avoid the problems of context overlapping [1,2,3,4]. It can also be used within most of the text mining tasks. For a given query term, to discover some novel information from a set of documents, clustering at sentence level can be done. By clustering the sentences of these documents, we would instinctively anticipate at least one of the clusters to be directly associated to the concepts described by the query terms. However, other clusters might contain information pertaining to the query in some way that may be unknown to us. Chao Shen, Tao Li, and Chris H. Q. Ding represented the sentences as vectors in term space and applying the K-means clustering algorithm[5]. Claude Pasquier, R. Mihalcea, P. Corsini, T. Geweniger applied the standard clustering algorithms to group sentences into clusters[6.15,16,17].

2.2 Vector Space Method of Similarity

The vector space model is able to adequately capture much of the semantic content of document-level text, because documents that are semantically related are likely to contain many words in common, based on cosine similarity[7]. The semantic similarity can be measured in terms of word co-occurrence at the document level not in sentences, since two sentences may be semantically related despite having few, if any, words in common. A number of sentence similarity measures have recently been proposed to solve this problem. Uma Devi. M and Meera Gandhi have analyzed the different approaches towards Measuring Semantic Similarity between Words for Semantic Similarity Search[8,20]. The Similarity Measures using Page Count used the popular Co-Occurrence measures Jaccard, Overlap (Simpson), Dice, and Point wise Mutual Information (PMI) [19]. The Snippet based Similarity Measures are using a lexical syntactic patterns extracted from the text Snippets which are used to compute the Semantic Similarity between words.

2.3 Similarity Based on Word Order Information

Yuhua Li and David McLean proposed the method for measuring the semantic similarity between sentences or very short texts, based on semantic and word order information[9]. The lexical knowledge base models are common human knowledge about words in a natural language; this knowledge is usually stable across a wide range of language

application areas[14,18]. Their semantic similarity not only captures common human knowledge, but it is also able to adapt to an application area using a corpus specific to that application. Uma Devi . M and Meera Gandhi proposed a new method to find similar words by using Bag of Word (BOW) and Extended Entity Description (EDs) concept[10]. This work is being able to find the similarity between words using Cosine Similarity and Ontology. First the Bag of Word is created for all the terms which are being extended using Ontology. This Ontology based Semantic Similarity can be used to increase the Precision and Recall rate and thus improves the performance of the search result.

2.4 Document Summarization

Dingding Wang and Tao Li proposed a new multi-document summarization framework based on sentence-level semantic analysis (SLSS) and symmetric non-negative matrix factorization (SNMF) [11]. SLSS is able to capture the semantic relationships between sentences and SNMF can divide the sentences into groups for extraction. Alexander Budanitsky and Graeme Hirst proposed a resource-based measures of lexical semantic distance, or, equivalently, semantic relatedness, for use in natural language processing applications[12]. Lexical semantic relatedness is sometimes constructed in context and cannot always be determined purely from an a priori lexical resource such as WordNet.

Andrew Rosenberg and Julia Hirschberg proposed a new external cluster evaluation measure, V-measure, and compared it with existing clustering evaluation measures[13]. V-measure is based upon two criteria for clustering usefulness, homogeneity and completeness, which captures a clustering solution's success including all and only data points from a given class in a given cluster. We have also demonstrated V-measure's usefulness in comparing clustering success across different domains by evaluating document and pitch accent clustering solutions. Uma Devi . M and Meera Gandhi.G proposed a Query Expansion Algorithm for Semantic Information Retrieval in Sports Domain(SIRSD) to do Semantic Search to improve search over large document repositories[12,21]. This algorithm reformulates user queries by using Word Net and Domain Ontology to improve the returned results. SIRSD reduces the issue of Semantic Interoperability during the user query search. The results show its effectiveness in generating a suitable number of query search with an accuracy of 87.1% compared to other competitors of generic search engines. The schematic diagram of our suggested clustering scheme to match the similar sentence is presented in Section 2.5 Section 3 describes Implementation of the scheme for the quotations datasets followed by evaluation results. Section 4 lists conclusion and directions for future research.

2.5 The Proposed System of Matching Similar Sentences

The proposed method finds the similar sentences from semantic and syntactic information contained in the sentences. The new model of matching semantically similar sentences based on Lexical database and Fuzzy clustering Framework is depicted in Figure. 1. The main components of the architecture of the system are Text processing , Wordnet based Analysis, Lexical database based document similarity and fuzzy clustering method. In this approach first the document processing is done by extracting the keywords using syntactic analysis and making a VSM for the document representing the terms along with the frequency.

A raw text document is the input to the proposed model. Each document has well-defined sentence boundaries. The text document considered here for this system consists of a sequence of words with some meaningful/useful information. The words along the sentence structure, express a specific meaning.

This proposed method vigorously forms a joint word set only using all the distinct words in the pair of sentences. For each sentence, a raw semantic vector is derived with the help of a lexical database. Using information from the lexical database, a word order vector is formed for each sentence.

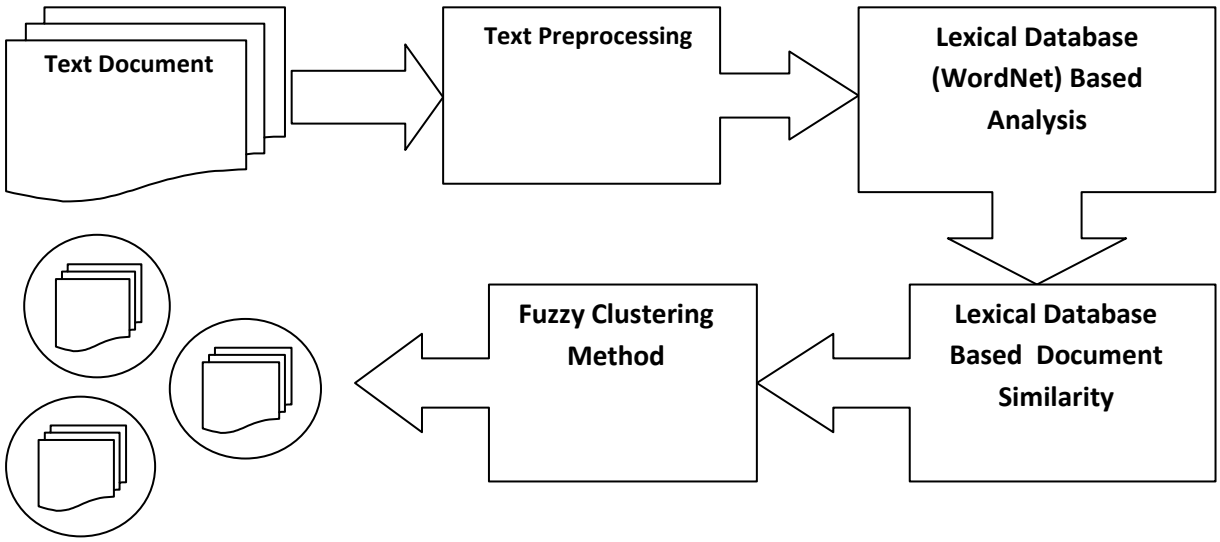

Figure.1. Lexical Database Based Matching Similar Sentences.

Clustered Sentences

Each word in a sentence donates differently to the meaning of the whole sentence. The importance of a word is weighted by using information content derived from a corpus. A semantic vector is obtained for each of the two sentences by combining the raw semantic vector with information content from the corpus. Semantic similarity is calculated based on

the two semantic vectors. An order similarity is deliberated using the two order vectors. At last , the sentence similarity is obtained by combining semantic similarity and order similarity.

2.5.1 Semantic Similarity between words

The hierarchical structure of the knowledge base (i.e.) Lexical Database is important in determining the semantic distance between words. Given two words, w_1 and w_2, we need to find the semantic similarity S (w_1, w_2), this can be done by the analysis of the Lexical knowledge base (WordNet) as follows:

One direct method for similarity calculation is to find the shortest path connecting the two words and the depth of word in the hierarchy should also be taken into account. So similarity between words is a function of path length and depth as follows:

$$\text{sim}(\text{word1}, \text{word2}) = \text{f}(\text{len}, \text{ht}) \qquad (1)$$

where 'len' is the shortest path length between $word_1$ and $word_2$, 'ht' is the depth of the words $word_1$ and $word_2$ in the hierarchical Lexical Database. The above equation (1) can be rewritten using two independent functions as follows:

$$\text{sim}(\text{word1}, \text{word2}) = \text{f1}(\text{len}).\text{f2}(\text{ht}) \qquad (2)$$

Where f1 and f2 are transfer functions of path length and depth, respectively.

2.5.2 Semantic Similarity between Sentences

Sentences are made up of words, so it is easy to represent a sentence using the words in the sentence. Unlike Traditional methods that use a precompiled word list containing hundreds of thousands of words, this method dynamically frames the semantic vectors solely based on the compared sentences. Given two sentences, S_1 and S_2, a joint word set is framed as:

$$\begin{aligned} s &= s1 \cup s2 \\ &= \{ w1, w2, \ldots\ldots wm) \end{aligned} \qquad (3)$$

The joint word set S contains all the distinct words from S_1 and S_2. Since inflectional morphology may cause a word to appear in a sentence with different forms that convey a specific meaning for a specific context, we use word form as it appears in the sentence. For example, boy and boys, woman and women are considered as four distinct words and all included in the joint word set.

Since the joint word set is purely derived from the compared sentences, it is compact with no redundant information.

The joint word set, S, can be viewed as the semantic information for the compared sentences. Each sentence is readily represented by the use of the joint word set as follows:

Thus, the joint word set for two sentences:

a. **S_1: Marriage is joining of the two soul woman and man that has many paints.**
b. **S_2: The woman cries before the marriage and the man afterward.**

S = {Marriage is joining of the two soul woman and man that has many paints cries before afterward }

The lexical semantic vector, denoted by **~s** is the vector derived from the joint word set. The dimension of the semantic vector equals the number of words in the joint word set as each entry of the semantic vector corresponds to a word in the joint word set. The value of an entry of the semantic vector is:

$$Si = \sim S.\, I(Wi).\, I(\sim Wi) \qquad (4)$$

where $\mathbf{w_i}$ is a word in the joint word set**,** $\mathbf{\sim w_i}$ is its associated word in the sentence $\mathbf{I(\ w_i\ )}$ and $\mathbf{I(\ \sim w_i\ )}$ is the associated information.

The semantic similarity between two sentences is defined as the cosine coefficient between the two vectors:

$$Ss = \frac{S1\ .S2}{|S1|\ .|S2|} \qquad (5)$$

The following four standard document clustering techniques are chosen for testing the effect of the Lexical Database based similarity on clustering: 1) FRECCA, 2) ARCA, 3) K-Means, 4) K-Medoids.

3. Results and Evaluation

Fuzzy Clustering Framework algorithms can be evaluated in many ways, but the choice of evaluation algorithms is being implemented. For example, The researchers from AI might be using mutual information, while the some researchers from the field of IR would choose F-measure. Two instinctive concept of performance (accuracy) are precision and recall. In the area of Information Retrieval, recall is the proportion of relevant documents that are retrieved out of all relevant documents available, while precision is the proportion of retrieved and relevant documents out of all retrieved documents. Because it is insignificant to get the accurate recall by retrieving all documents for any query input, the F-measure, which combines both recall and precision, is introduced. Let R be recall and P be precision, then the generalized F-measure is defined as

$$F\alpha = \frac{(1+\alpha)\ RP}{\alpha P + R} \qquad (6)$$

Where α is an integer. Precision and recall are typically given equal weight for $\alpha = 1$, but variations exist which weight them differently, e.g. precision twice as much as recall for $\alpha = 0.5$, or vice versa for $\alpha = 2$. While F-measure addresses the total quality of the clustering in terms of retrieval performance, it does not address the composition of the clusters themselves. The experimental result conducted on the Quotations dataset is illustrated in Table 2. and Figure. 2. From this figure we can see that the result of our algorithms indicates that our proposal can obtain the higher precision than the Baseline technique. The clustering performance is shown in Table-3. and the comparisons over different clustering algorithms are shown in Figure.3.

Table 1: Extract from Famous Quotations Data Set

Knowledge

1. Our knowledge can only be finite, While our ignorance must necessarily be finite.
2. Everybody gets so much useful information all day long that they lose their commonsense.
3. Little minds are interested in the extraordinary; great minds in the commonplace.
....

Marriage

11. A husband is what is the left of lover, after the nerve has been extracted.
12. Marriage has many pains, but celibacy has no pleasures.
13. The woman cries before the wedding; the man afterward.
....

Nature

21. I have called this principle, by which each sight variation, if useful, is preserved, by the term natural selection.
22. Nature is reckless of the individual; when she has points to carry, she carries them.
23. I wanted to say something about the universe; there's God, angles, plants and horsiest.
....

Peace

31. There is no such thing as inner peace, there is only nervousness and death.
32. Once you hear the details of victory, it is hard to distinguish it from a defeat.
33. They sicken of the calm who know the storm.

The data sets of this algorithm are shown in Table 1**.** Two additional measures are cluster purity and entropy. Purity measures the percentage of the dominant class members in a given cluster (larger is better), while entropy looks at the distribution of documents from each reference class within clusters (smaller is better). These are written in equation (7) and (8).

Table 2: Similarity measure of Fuzzy Clustering system and Baseline Method.

Similar sentences words of Quotation datasets	Similarity	
	Fuzzy Clustering	Baseline
Knowledge, Useful Information	0.5	0.3
Marriage, Wedding	0.7	0.4
Nature, Plants	0.6	0.2
Peace, Calm	0.4	0.1
Diet, food	0.8	0.3
common sense, common place	0.5	0.4
husband, woman	0.4	0.31
wedding, husband	0.6	0.5
nervousness, victory	0.7	0.6

$$\text{Purity} = \sum_{j} \frac{n_j}{n} \operatorname{argmax}_{i} P(i,j) \quad (7)$$

$$\text{Entropy} = -\frac{1}{\log k} \sum_{j} \frac{n_j}{n} \sum_{i} P(i,j) \log p(i,j) \quad (8)$$

Table-3: Clustering Performance on the Quotation datasets.

Clustering Algorithm	Purity	Entropy
Fuzzy Clustering	0.752	0.355
FRECCA	0.713	0.335
ARCA	0.608	0.403
K-Means	0.689	0.335
K-Medoids	0.666	0.337

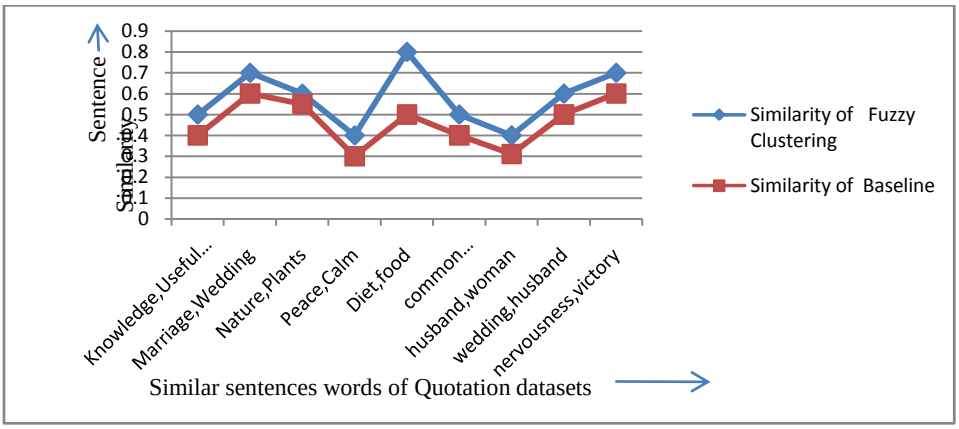

Figure . 2. Sentence Similarity of the Quotations Datasets.

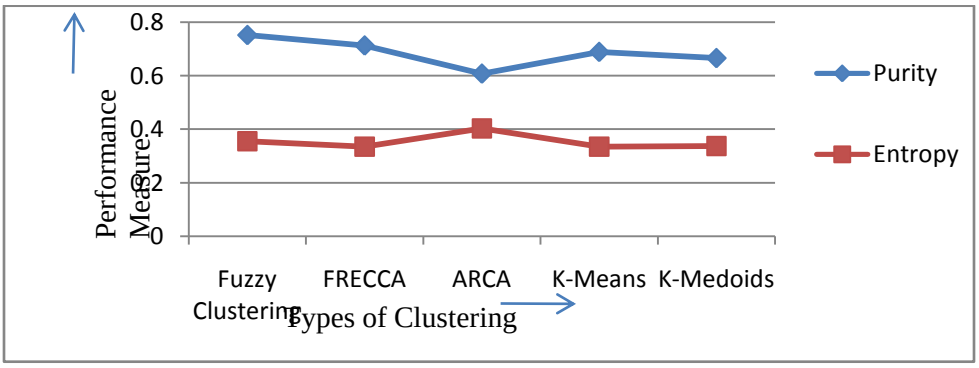

Figure . 3. Comparison of different clustering Algorithm with Fuzzy Clustering.

4. Conclusion and Future Enhancement

In this paper, we presented a new method of finding Semantically Similar Sentences using a Lexical Database based Fuzzy Clustering Algorithm. The comprehensive experimental evaluation demonstrates the efficiency of the proposed techniques with baseline technique of finding similar sentences. We conducted extensive experiments on Quotations datasets and trained two parameters (Entropy and Purity) for the efficiency evaluation. This method uses information from the centroids of the clusters to select sentences that are most likely to be relevant to the cluster topic. To understand the trade-off, we evaluated different combination of features between the baseline and our proposed method. The concepts present in natural language documents usually display some type of hierarchical structure, whereas the algorithm we have presented in this paper identifies only flat clusters. Our main objective in future is to extend these ideas to the development of a hierarchical Fuzzy relational clustering algorithm by incorporating Ontology concept to get more accurate Similar Sentences.

References

1. V. Hatzivassiloglou, J.L. Klavans, M.L. Holcombe, R. Barzilay, M. Kan, and K.R. McKeown. SIMFINDER: A Flexible Clustering Tool for Summarization. Proc. NAACL Workshop Automatic Summarization, pp. 41-49, 2001.
2. H. Zha. Generic Summarization and Keyphrase Extraction Using Mutual Reinforcement Principle and Sentence Clustering. Proc. 25th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 113-120, 2002.
3. D.R. Radev, H. Jing, M. Stys, and D. Tam. Centroid-Based Summarization of Multiple Documents. Information Processing and Management: An Int'l J., vol. 40, pp. 919-938, 2004.
4. R.M. Aliguyev. A New Sentence Similarity Measure and Sentence Based Extractive Technique for Automatic Text Summarization. Expert Systems with Applications, vol. 36, pp. 7764-7772, 2009.
5. Chao Shen, Tao Li, and Chris H. Q. Ding. 2011. Integrating clustering and multi-document summarization by bi-mixture probabilistic latent semantic analysis (plsa) with sentence bases . In AAAI
6. Claude Pasquier. 2010. Task 5: Single document keyphrase extraction using sentence clustering and latent dirichlet allocation. In Proceedings of the 5th International Workshop on Semantic Evaluation, SemEval '10, pages 154–157, Stroudsburg, PA, USA. Association for Computational Linguistics.

7. C.D. Manning, P. Raghavan, and H. Schu¨ tze. Introduction to Information Retrieval. Cambridge Univ. Press, 2008.
8. M.Uma Devi and G.Meera Gandhi. An Approach towards Measuring Semantic Similarity between Words using Ontology for Semantic Similarity Search. In Proc NC4t'13 held at Sathyabama University Chennai, on 29 August 2013.
9. J.J. Jiang and D.W. Conrath,. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. Proc. 10th Int'l Conf. Research in Computational Linguistics, pp. 19-33, 1997.
10. M.Uma Devi and G.Meera Gandhi. A Survey on Different Methods of Semantic Similarity and Semantic Similarity Search using Ontology. In Proc CCIIS'13 held at VIT University , Vellore, from 21-11-13 to 23-11-13 .
11. D. Wang, T. Li, S. Zhu, and C. Ding. Multi-Document Summarization via Sentence-Level Semantic Analysis and Symmetric Matrix Factorization. Proc. 31st Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 307-314, 2008.
12. M.Uma devi and G.Meera Gandhi. WordNet and Ontology Based Query Expansion for Semantic Information Retrieval in Sports domain. J.Comput.Sci., 11(2):361-371,2015,ISSN:1549-3636,DOI:10.3844/jcssp.2015.361.371 ,http://www.thescipub.com/jcs.toc.
13. A. Rosenberg and J. Hirschberg. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. Proc Conf. Empirical Methods in Natural Language Processing (EMNLP '07), pp. 410-420, 2007.
14. Y. Li, D. McLean, Z.A. Bandar, J.D. O'Shea, and K. Crockett. Sentence Similarity Based on Semantic Nets and Corpus Statistics. IEEE Trans. Knowledge and Data Eng., vol. 8, no. 8, pp. 1138-1150, Aug. 2006.
15. R. Mihalcea, C. Corley, and C. Strapparava. Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity. Proc. 21st Nat'l Conf. Artificial Intelligence, pp. 775-780, 2006.
16. P. Corsini, F. Lazzerini, and F. Marcelloni. A New Fuzzy Relational Clustering Algorithm Based on the Fuzzy C-Means Algorithm. Soft Computing, vol. 9, pp. 439-447, 2005.
17. T. Geweniger, D. Zu¨ hlke, B. Hammer, and T. Villmann. Median Fuzzy C-Means for Clustering Dissimilarity Data. Neurocomputing, vol. 73, nos. 7-9, pp. 1109-1116, 2010.
18. A. Budanitsky and G. Hirst,. Evaluating WordNet-Based Measures of Lexical Semantic Relatedness. Computational

Linguistics , vol. 32, no. 1, pp. 13-47, 2006.

19. M.Uma devi and G.Meera Gandhi.An Enhanced Ontology Based Measure of Similarity between Words and Semantic Similarity Search. @ Springer International publishing Switzerland 2015,Emerging ICT for bridging the future ,Volume-1 , Advances and Intelligent Systems and Computing 337 , DOI : 10.1007 / 978-3-319-13728-5_50.
20. M.Uma devi and G.Meera Gandhi. Enhanced Fuzzy Clustering and Expectation Maximization Framework based MatchingSemantically Similar Sentences. http://www.sciencedirect.com/science/article/pii/S1877050915019353, Procedia Computer Science 57 (2015) 1149 – 1159, doi: 10.1016/j.procs.2015.07.406.
21. Priya, T.,Justin Samuel, S (2016), " Priority based multi sencar technique in Wireless Sensor", Indian Journal of Science and Technology, Vol 9(21),pp. 1-6.