Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 57 (2015) 1149 – 1159

Procedia Computer Science

3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)

An Enhanced Fuzzy Clustering and Expectation Maximization Framework based Matching Semantically Similar Sentences

M.Uma Devi[a] , Dr. G. Meera Gandhi [b,]*

[a]*Research Scholar , Sathyabama University ,Chennai, Tamil Nadu, India, umadevi.as,2006@gmail.com*
[b]*Professor, Faculty of Computing, Sathyabama University ,Chennai, Tamil Nadu, India, meeragandhi.cse@sathyabamauniversity.ac.in*

Abstract

Statistical measure of finding Similar Sentences using a novel Fuzzy clustering algorithm framework is developed which organizes text from one or more documents into different clusters . The traditional fuzzy clustering approaches are not applicable to sentence clustering because most sentence similarity measures do not represent sentences in a common metric space. An enhanced Fuzzy clustering algorithm is applied in the sentence of datasets to group the related sentences. Page Rank algorithm highlights the more relevant inter clusters which interprets the Page-Rank score of an object . Expectation- Maximization (EM) framework has been developed in order to predict the overlapping clusters of semantically related sentences. Quotations dataset and News article dataset empirically implies the Similarity measure of matching Semantically Similar Sentences in which our system out performs the baseline method and projection methods. Our proposed method performs 34 % higher in similarity scoring of related sentences. It also analyzes the clustering performance in terms of Entropy and Purity which yields more Purity and less Entropy. Our Experimental results demonstrates that our method is capable of identifying the overlapping clusters of semantically related sentences, and can be used in a variety of text mining tasks.

Peer-review under responsibility of organizing committee of the 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)

Keywords: *Similar Sentences , Expectation , Fuzzy Clustering, Maximization, Page Rank.*

* Corresponding author. Tel.: 91-9444536571; fax: +0-000-000-0000 .
E-mail address: umadevi.as2006@gmail.com

Peer-review under responsibility of organizing committee of the 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015)
doi:10.1016/j.procs.2015.07.406

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 so 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. Consider the following examples.

Example 1

Bomb Blasting in Bombay may be described in two different documents as follows.

At least 105 people were killed	Document-1
120 people were found dead	Document-2

In the example above, Document (1) and (2) gives the same meaning , therefore, (1) and (2) can be considered Similar.

Example 2

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 -3.
The court found Johnny guilty of the murder of James last August and sentenced him to life	Document -4.

In the example above, Document (4) subsumes (3) therefore, (3) and (4) can be considered Similar. The above two sentences contain the same event.

This paper depicts , how Semantically Similar sentences are matched and evaluated based on a Fuzzy sentence clustering scheme. 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 it. The best clustering is the one that minimizes this prediction error. Most documents will contain interrelated topics irrespective of the specific task like summarization, text mining and many sentences will be related to some degree to a number of these. 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. Related Work

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 web mining , where the specific objective is to discover some novel information from a set of documents in response to some query. While clustering the sentences we would expect at least one of the clusters to be closely related to the concepts described by the query terms; however, other clusters may 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[15]. Claude Pasquier applied the standard clustering algorithms to group sentences into clusters[23].

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[14]. 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.G have analyzed the different approaches towards Measuring Semantic Similarity between Words for Semantic Similarity Search[14]. The Similarity Measures using Page Count used the popular Co-Occurrence measures Jaccard, Overlap (Simpson), Dice, and Point wise Mutual Information (PMI). 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.

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]. First, semantic similarity is derived from a lexical knowledge base and a corpus. The lexical knowledge base models common human knowledge about words in a natural language; this knowledge is usually stable across a wide range of language application areas. 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.G proposed a new method to find similar words by using Bag of Word (BOW) and Extended Entity Description (EDs) concept[12]. 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 is being extended using Ontology. This Ontology based Semantic Similarity can be used to increase the Precision and Recall rate and thus to improve the performance of the search result.

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)[5]. 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[10]. In word sense disambiguation, such an association with the context is frequently a sufficient basis for selecting or rejecting candidate senses; in our malapropism corrector, a word should be considered non anomalous in the context of another if there is any kind of semantic relationship at all apparent between them. 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[11] . 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[13]. 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 3. Section 4 describes the methods of matching similar sentences. Implementation of the scheme for the quotations datasets and News article datasets, followed by evaluation results in Section 5. Section 6 lists conclusion and directions for future research.

3. System Description

The schematic diagram of the system is depicted in **Fig.1.** The Raw semantic and order vector is calculated first. Semantic vector is calculated from the Raw Semantic. The order vector is used to find the order Similarity. These order Similarity and the Semantic Similarity can be used to find Similar Sentences. We first describe the use of PageRank as a general graph centrality measure, and review the Gaussian mixture model approach. We then describe how PageRank can be used within an Expectation-Maximization framework to construct a complete relational fuzzy clustering algorithm.

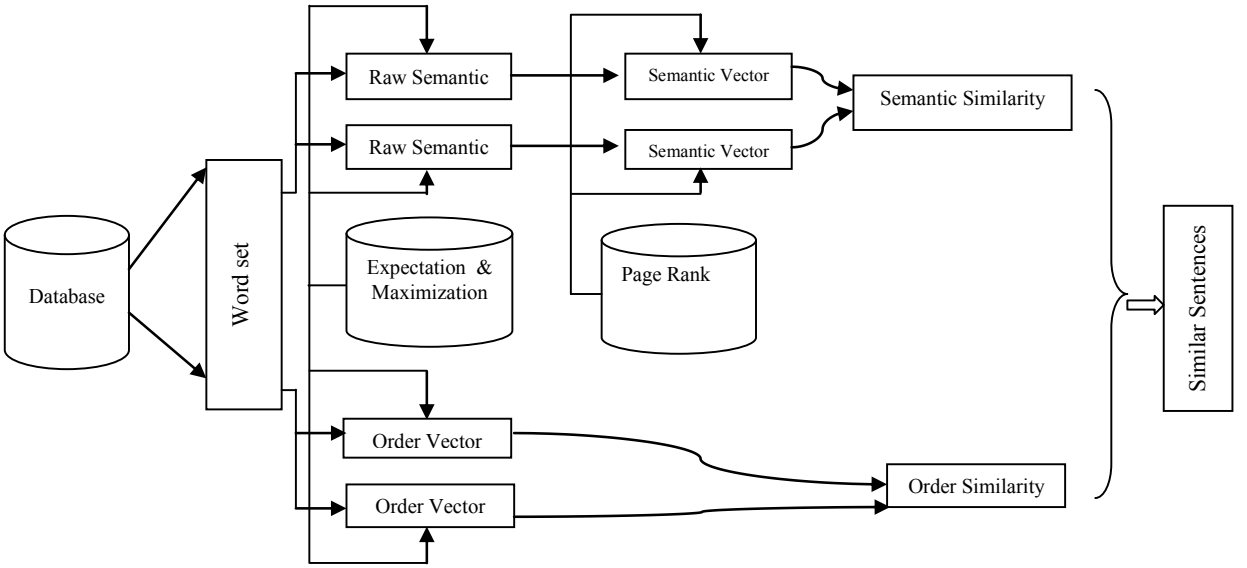

Fig .1. Schematic Diagram of the System

3.1 Assigning Page Rank Score

Sentence is represented by node on a graph and edges are weighted with value representing similarity between sentences. Page Rank is value assigns between 0 and 1. The importance of a node within a graph can be determined by using the Page Rank. This can be determined by taking into account global information recursively computed from the entire graph, with connections to high-scoring nodes Page Rank assigns to every node in a directed graph a Numerical score between 0 and 1, known as its Page Rank score (PR), and defined in (1) as

$$PR(V_i) = (1\text{-}d) + d \times \sum_{j \in In\,(Vi)} \frac{1}{|\,Out\,((Vj)|}\ PR(V_j) \qquad (1)$$

where In(Vi) is the set of vertices that point to Vi, Out(V_j) is the set of vertices pointed to by V_j, and d is a damping factor, typically set to around 0.8 to 0.9 . Using the analogy of a random surfer, nodes visited more often will be those with many links coming in from other frequently visited nodes, and the role of d is to reserve some probability for jumping to any node in the graph, thereby preventing situation of getting stuck in a disconnected part of the graph.

PageRank can be used more generally to determine the importance of an object in a network. For example, TextRank and LexRank both use PageRank for ranking sentences for the purpose of extractive text summarization. In both TextRank and LexRank, each sentence in a document or documents is represented by a node on a graph. However, unlike a web graph, in which edges are unweighted, edges on a document graph are weighted with a value representing the similarity between sentences. The PageRank algorithm can easily be modified to deal with weighted undirected edges, resulting in:V_i which is defined in (2) as

$$PR(V_i) = (1\text{-}d) + d \times \sum_{j=1}^{N} \left(Wji \frac{PR(Vi)}{\sum_{k=1}^{N} Wjk} \right) \quad (2)$$

Where w_{ji} is the similarity between V_j and V_i, and we assume that these weights are stored in a matrix W=$\{ w_{ji}\}$, which we refer to as the "affinity matrix." Before describing how PageRank can be used to determine centrality within a mixture of components, leading to the proposed relational fuzzy clustering algorithm, we briefly review Gaussian mixture models and the EM algorithm.

3.2 Expectation and Maximization Algorithm

It is an unsupervised method, which does not need any training phase; it tries to find the parameters of the probability distribution that has the maximum likelihood of its parameters. Its main role is to parameter estimation. It is an iterative method, which is mainly used to finding the maximum likelihood parameters of the model. The E-step involves the computation of cluster membership probabilities. The probabilities calculated from E-step are estimated with the parameters in M-step.

Assuming that the parameters of each component are represented by a parameter vector θ_m, the problem is to determine the values of the components of this vector, and this can be achieved using the Expectation-Maximization algorithm. Following random initialization of the parameter vectors θ_m, m=1….C, an Expectation step (E-step), followed by a Maximization step (M-step), are iterated until convergence. The E-step computes the cluster membership probabilities. For example, assuming spherical Gaussian mixture components, these probabilities are calculated using (3) as follows.

$$P(m/Xi) = \frac{\pi_m p(X_i / \beta_m, \sigma_m)}{\sum_{k=1\ldots C} \pi_k p((X_i | \beta_m, \sigma_m)} \text{ , m=1,.....,C} \quad (3)$$

Where β_m, and σ_m are the current estimates of the mean and standard deviation, respectively, of component m. The denominator acts as a normalization factor, ensuring the value in (4) as follows.

$$0 \le P(\mathrm{m}|X_i) \le 1 \; and \; \sum_{m=1}^{C} P(\mathrm{m}|X_i) = 1 \quad (4)$$

In the M-step, these probabilities are then used to re-estimate the parameters. The spherical Gaussian case (5),(6) and (7) are used to evaluate the likelihood clusters.

$$\beta_m = \frac{\sum_{i=1}^{N} P(\mathrm{m}|X_i) X_i}{\sum_{i=1}^{N} P(\mathrm{m}|X_i)}, \qquad \mathrm{m} = 1, \ldots, \mathrm{C}, \quad (5)$$

$$\sigma^2_m = \frac{\sum_{i=1}^{N} P(\mathrm{m}|X_i) ||X_i - \mu_m||^2}{\sum_{i=1}^{N} P(\mathrm{m}|X_i)}, \quad \mathrm{M} = 1,2, \ldots, \mathrm{C}, \quad (6)$$

$$\pi_m = \frac{1}{N} \sum_{i=1}^{N} P(\mathrm{m}|X_i), \quad \mathrm{m} = 1, \ldots, \mathrm{C}. \quad (7)$$

3.3 Fuzzy Relational Clustering

Cluster membership values are initialized randomly, and normalized such that cluster membership for an object sums to unity over all clusters. Mixing coefficients are initialized such that priors for all clusters are equal. The E-

step calculates the PageRank value for each object in each cluster. PageRank values for each cluster are calculated, with the affinity matrix weights w_{ij} obtained by scaling the similarities by their cluster membership values as in (8).

$$w™_{ij} = s_{ij} \times p™_i \times p™_j, \tag{8}$$

This maximization step involves only the single step of updating the mixing coefficients based on membership values calculated in the Expectation Step.

4. Methods to find Similar Sentences

4.1 *Global Recursive Computation*

The importance of a node within a graph can be determined by taking into account global information recursively computed from the entire graph, with connections to high-scoring nodes contributing more to the score of a node than connections to low-scoring nodes. PageRank assigns to every node in a directed graph a numerical score between 0 and 1. Using the analogy of a random surfer, nodes visited more often will be those with many links coming in from other frequently visited nodes.

4.2 *Document Graph Construction*

PageRank can be used more generally to determine the importance of an object in a network. In both Text Rank and Lex Rank, each sentence in a document or documents is represented by a node on a graph. However, unlike a web graph, in which edges are unweighted, edges on a document graph are weighted with a value representing the similarity between sentences.

PageRank score of an object within a cluster as a measure of its centrality to that cluster. These PageRank values are then treated as likelihoods. Since there is no parameterized likelihood function as such, the only parameters that need to be determined are the cluster membership values and mixing coefficients. PageRank value for each object in each cluster. The intuition behind this scaling is that an object's entitlement to contribute to the centrality score of some other object depends not only on its similarity to that other object, but also on its degree of membership to the cluster.

4.3 *Renormalization of Duplicate Cluster*

The number of initial clusters must be specified as input to the algorithm. If this number is too high, then duplicate clusters will be found. While it might appear at first sight that duplicate clusters can simply be removed after the algorithm has converged, and membership subsequently renormalized to sum to one, this is not possible because of the coupling between membership values and PageRank values. To perform a check for duplicate clusters at the completion of each Maximization step. If duplicate clusters are found, membership values are renormalized, and the algorithm is allowed to proceed until a stage at which convergence has been achieved and no duplicate clusters exist.

4.4 *Measuring Sentence Similarity*

This approach is similar to that used to calculate document similarity in the IR literature; however, rather than using a common vector space representation for all sentences, the two sentences being compared are represented in a reduced vector space of dimension n, where n is the number of distinct non stopwords appearing in the two sentences. Semantic vectors, V1 and V2, representing sentences S1 and S2 in this reduced vector space are first constructed. The semantic similarity between two sentences is defined as the cosine coefficient between the two vectors as in (9).

$$S_s = \frac{S1 \,.\, S2}{||s1|| \,.\, ||s2||} \tag{9}$$

Algorithm:

Input: sentences i and j where i=1,….,N j=1,…..,N
Pair wise similarity values 'S' is the similarity between sentences i and j
Number of clusters C.
Output: Cluster membership values

1. Create initially an undirected graph with sentence-set terms as nodes and use lexical resources to extract semantically-related terms for each node.
2. Initialize and normalize membership values
3. for i = 1 to N
   {
4. for m=1 to C
5. Assign random numbers between 0 & 1.
6. for m=1 to C
7. $p_i^m = p_i^m / \sum_{j=1}^{C} P_j^m$
   }
8. for m= 1 to C
9. $\pi_m = 1/C$
10. Repeat until convergence
11. for m=1 to C
    {
12. for i=1 to N
13. for j=1 to N
14. calculate page rank scores for cluster m
15. Repeat until convergence
16. Assign page rank scores to likelihoods
17. $l_j^m = PR_i^m$
18. }
19. for i=1 to N
20. for m = 1 to C
21. Calculate new cluster membership values
22. Represent term clusters and sentences as vectors in term space and calculate the similarity of each sentence with each of the term clusters.
23. Assign each sentence to the best-scoring term cluster.

5. Results and Evaluation

Clustering algorithms have been evaluated in many ways. The choice of evaluation methods frequently depends on the domain in which the research is being conducted. For example, an AI researcher might favour mutual information, while someone from the field of IR would choose F-measure. These metrics, and others, will be discussed here.

Two intuitive notions of performance (accuracy) are precision and recall. In the field of IR, recall is defined as 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 trivial to get perfect recall by retrieving all documents for any query, 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 in (10) as

$$F_{\alpha} = \frac{(1+\alpha)\,RP}{\alpha P + R} \quad (10)$$

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. Two additional measures are cluster purity and entropy. These are written in equation (11) and (12). 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).

$$\text{Purity} = \sum_{j} \frac{nj}{n} \underset{i}{argmax}.P(i,j) \quad (11)$$

$$\text{Entropy} = -\frac{1}{\log k} \sum_{j} \frac{nj}{n} \sum_{i} P(i,j) \log p(i,j) \quad (12)$$

The data sets of this algorithm is shown in **Table 1 and Table 2.**

Table 1: Extract from Famous Quotations Data Set

Knowledge
1. The true Sign of intelligence is not knowledge but imagination.
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. Marriages are like fingerprints; each one is different and each one is beautiful
12. A husband is what is the left of lover, after the nerve has been extracted.
13. Love is the greatest gift when given. It is the highest honor when received.
….
Nature
21. I like this place and could willingly waste my time in it.
22. Nature is reckless of the individual; when she has points to carry, she carries them.
23. Live in each season as it passes; breathe the air, drink the drink, taste the fruit, and resign yourself to the influence of the earth.
….
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.
….
Food
41. After a good dinner one can forgive anybody, even one's own relations.
42. Food is an important part of a balanced diet.
43. Dinner, a time when one should eat wisely but not too well, and talk well but not too wisely.

The experimental results conducted on the Quotations dataset is illustrated in **Table .3.** and **Fig. 2.** The Similarity measure on News Article datasets are depicted in **Fig.3. and Fig.4.** 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 .4.** and the comparisons over different clustering algorithms are shown in **Fig.5.**

Table 2: Samples from News Article Data Set

1. Eighteen decapitated bodies have been found in a mass grave in northern Algeria, press reports said Thursday, adding that two shepherds were murdered earlier this week.

2. Security forces found the mass grave on Wednesday at Chbika, near Djelfa, 275 kilometers (170 miles) south of the capital.

3. It contained the bodies of people killed last year during a wedding ceremony, according to Le Quotidien Liberte.

4. The victims included women, children and old men.

5. Most of them had been decapitated and their heads thrown on a road, reported the Es Sahafa.

6. Another mass grave containing the bodies of around 10 people was discovered recently near Algiers, in the Eucalyptus district.

Table 3. Similarity measure of enhanced Fuzzy system and Baseline Method

Similar sentences words of Quotation datasets	Similarity	
	Enhanced Fuzzy	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

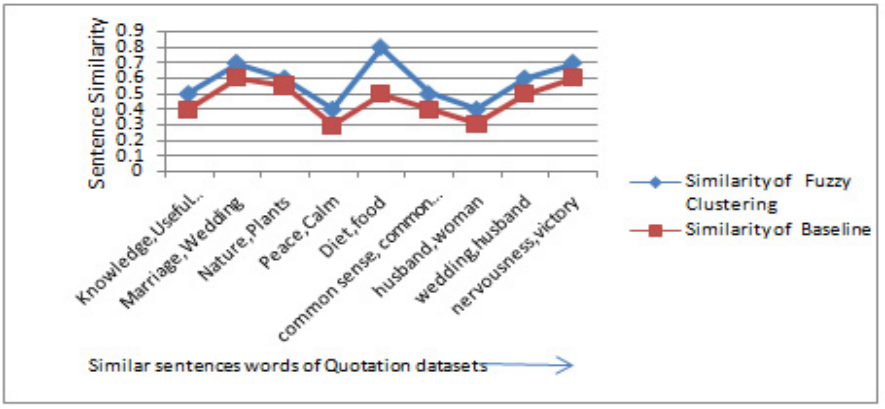

Fig . 2. Sentence Similarity of the Quotations Data Set

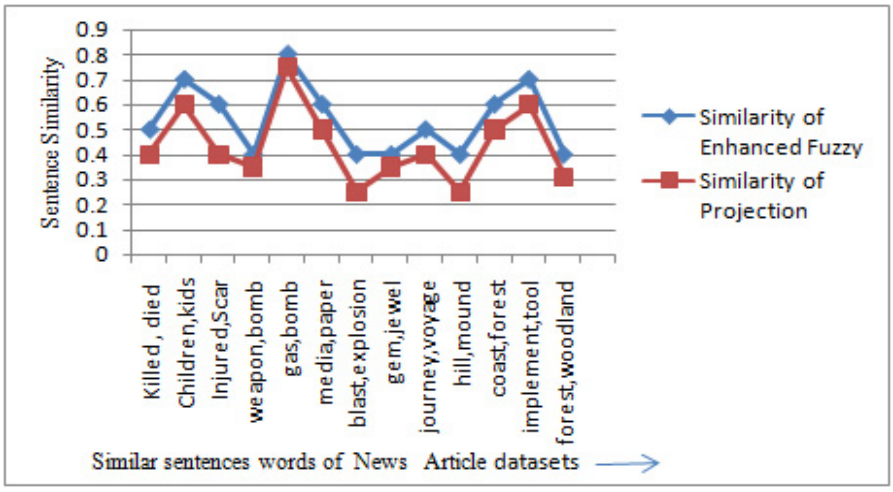

Fig . 3. Sentence Similarity of the News Article Data Set

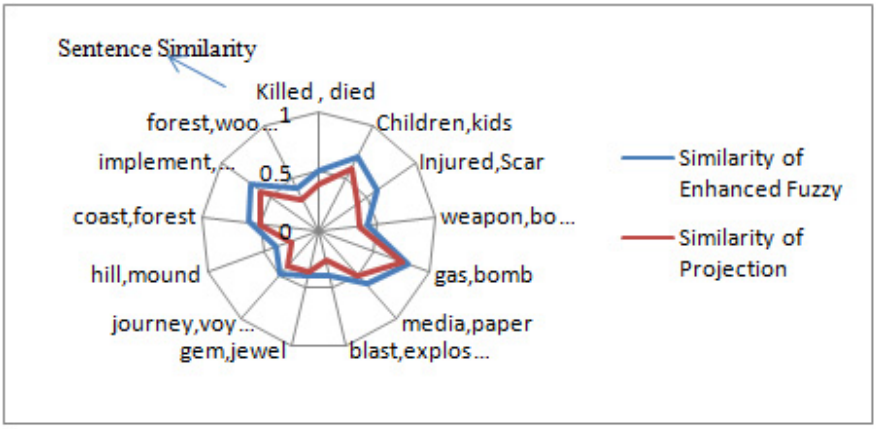

Fig . 4. Clustered sentences and Sentence Similarity of the News Article Data Set

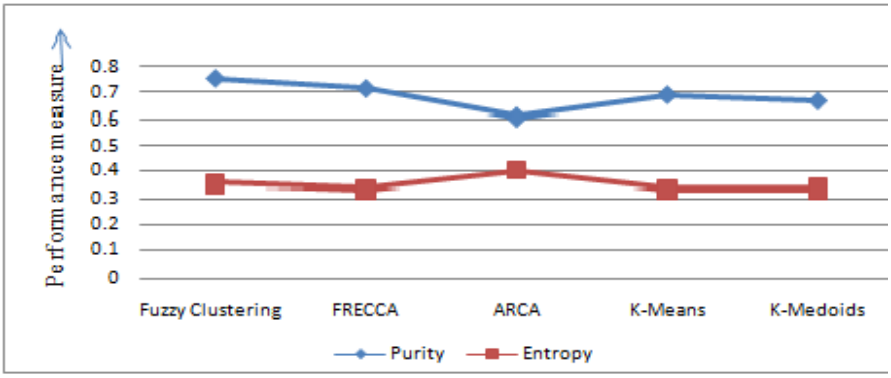

Fig . 5. Comparison of different clustering Algorithm with Enhanced Fuzzy

Table 4. Clustering Performance on the above datasets

Clustering Algorithm	Purity	Entropy
Enhanced Fuzzy	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

6. Conclusion and Future Enhancement

In this paper, We presented a new method of finding Semantically Similar Sentences using an enhanced Fuzzy Clustering Algorithm. The comprehensive experimental evaluation demonstrates the efficiency of the proposed techniques with baseline technique and Projection Method of finding Similar sentences. we conducted extensive experiments on Quotations datasets & News Article 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 and Wordnet 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," P*roc. 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-sBased 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. 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.
6. 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.
7. Y. Chen, E.K. Garcia, M.R. Gupta, A. Rahimi, and L. Cazzanti, "Similarity-Based Classification: Concepts and Algorithms," *J. Machine Learning Research*, vol. 10, pp. 747-776, 2009.
8. C.D. Manning, P. Raghavan, and H. Schu¨ tze, *Introduction to Information Retrieval*. Cambridge Univ. Press, 2008.
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. Budanitsky and G. Hirst, "Evaluating WordNet-Based Measures of Lexical Semantic Relatedness," *Computational Linguistics* , vol. 32, no. 1, pp. 13-47, 2006.
11. 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.
12. 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 .
13. M.Uma devi and G.Meera Gandhi," WordNet and Ontology Based Query Expansion for Semantic Information Retrieval in Sports ", *J.Comput.Sci*., 11(2):361-371,2015,ISSN:1549-3636,DOI:10.3844/jcssp.2015.361.371 ,http://www.thescipub.com/jcs.toc.
14. 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.
15. 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.*
16. 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.