Aggregated Probabilistic Fuzzy Relational Sentence Level Expectation Maximization Clustering Algorithm for Efficient Text Categorization

V.L.Kartheek
II M.Tech
Department of CSE
S.R.K.R Engineering College

V.Chandra Sekhar, Ph.D.
Associate Professor
Department of CSE
S.R.K.R Engineering College

ABSTRACT
Now a days, Text clustering becomes an important application to organize the data and to extract useful information from the available corpus. Many previous clustering techniques have difficulties in handling extreme outliers but fuzzy clustering algorithms tend to give them very small membership degree in surrounding clusters. In this paper we proposed an aggregated probabilistic Fuzzy relational sentence level expectation maximization clustering algorithm for efficient text categorization. It will give the accurate and maximum similarity by finding the relevance of sentences which belongs to a particular cluster. This technique leads to a fuzzy partition of the sentences and find out the accurate probability of the words belongs to a cluster. This algorithm is particularly used in finding maximum likelihood estimates of words in a given sentence. It gives the low search results with highest accuracy. The practical results show that the proposed method obtains better and accurate results for getting best sentence-wise text classification when compared with the existing methods.

Keywords
Fuzzy clustering, corpus, outliers.

1. INTRODUCTION
Cluster analysis is useful tool for finding required information from available large sets of text. Cluster analysis is a technique for classifying data [1], i.e., to division of a given set of objects or things into a set of classes or clusters based on similarity. The goal is to divide the sentence into words and then know the probability of belonging to a particular class or cluster [2]. It is a method of finding the relevance of sentences to some class or cluster. The hard clustering methods restrict each point of the data set to exactly one cluster [1]. These methods yield exhaustive partitions of the example set into non-empty and pair wise disjoint subsets. Fuzzy cluster analysis, [3] allows accurate relevance of words to clusters in the range of [0, 1]. This tells the flexibility to express that data points belong to more than one cluster at the same Time. Furthermore, these relevance of words offer accurate estimation. Clustering is the process of grouping or aggregating of data items. Sentence level clustering used in different applications such as classify and categorization of documents and organizing the documents, etc [4]. In text processing, sentence clustering plays an important role and is used in various text mining activities [5]. Size of the clusters may change from one cluster to another [6]. The previous clustering algorithms have some problems in clustering the input dataset [3] and also not identifying the outliers. Against the drawbacks of these clustering algorithms, Later various clustering algorithms can be developed for the clustering of sentences [7]. In those, Contents present in text documents contain hierarchical structure and there are many of the terms present in the documents which are related to more than one point [8]. But in the previous algorithms, the accuracy of belonging to a particular class or cluster is very low. Hence we proposed aggregated probabilistic Fuzzy relational sentence level expectation Maximization clustering algorithm.

The various previous algorithms can be facilitates some poor performance. The fuzzy algorithms find all the possibilities of relevance. From this method a large variety of clustering techniques was derived with more complex prototypes, which are mainly interesting in data analysis applications [9]. However, the generalization of these techniques to clustering uncertain data or objects is not yet explored. The sentence can be accurately predicts the level of matching to a particular cluster.

The Text classification plays major role today for all fields [10]. Recently, fuzzy set theory is more and more frequently used because of its simplicity and usage in different applications [11]. The theory has been successfully applied to use in many fields. The Fuzzy concept is very popular to get accurate and efficient results.

In this context, a generalization of the previously methods in order to be used in clustering of fuzzy data [9] would be a meritorious research. In this work, a new Aggregated probabilistic fuzzy relational clustering algorithm, based on the Sub systems concept. The fuzzy sub systems are used to find the relevant information regarding similar to a class. [12]. This clustering process divides the words in a given sentence of a Fuzzy System into a set of classes or clusters of fuzzy system based on similarity. From this new strategy, a flat fuzzy system $S(x)$ can be organized into a hierarchical structure of fuzzy systems. This fuzzy modeling is a trusted method to identify fuzzy models of target systems with many input variables or/and with different complexity interrelation. This method easily identify the correct class and gives accurate results. In, this proposal of a new technique, the Clustering of words in a given sentence, based on following methodology to decompose a original fuzzy system. Let us take $S(x)$ is the original function. Then it is divided into a set of n fuzzy sub-systems $S_1(x)$, $S_2(x)...S_n(x)$, which is organized in a fuzzy system. Each of these sub systems contains the information of the system $S(x)$. The proposed algorithm allows grouping a set of words into some subgroups (clusters) of similar relevance. It is a generalization of the Probabilistic Clustering Algorithm [8], here applied to words instead of points. With this algorithm, the system obtained from the data is transformed into a new system, organized into several subsystems, in system structures. This can be done in a several steps. Firstly, a brief introduction to fuzzy systems is presented. But much of these data are of potentially not useful. In order to make it useful one, we need to extract the

knowledge or information underlying the data. Data mining is a process of taking the valuable information from the huge amount of data. Clustering techniques can help in this data discovery and data analysis. APFRSEC algorithm is mainly useful in retrieving the information. Clustering text [7] at the sentence level and document level has many differences. Document clustering partitions the documents into several parts and cluster those parts based on the overall understanding of content. It doesn't give much importance to the meanings of each sentence in the document. So; there may be a content overlap or finding the hard data. Each data element in hard clustering method belongs to exactly one cluster.

1.1 Clusters of data

There are many algorithms already there for clustering of text or data [10]. Each algorithm will group the data objects based on some metrics or measure. The useful data can be classified for better way of getting things. Clustering is used in many different applications. The text mining [10] is the process of extraction or getting data efficiently. The retrieving of information is challenging today [13]. The similarity of words in a given sentence will decide the accuracy of belonging to a cluster. Sentence Clustering mainly used in variety of text mining applications. Clustering is one of the most [11] important concept for group of objects. When searching for a required data this technique is very useful.

2. BACK GROUND AND RELATED WORK

2.1 Efficient Sub Division of a System:

The subdivision of a given system always avoids the difficulty and problem . The sub division of a system concept simplifies the system easily and conveniently. The words can be taken from a particular related sentence and then solve it by getting aggregates. Based on this the estimation of nearest class can be identified. The text classification is very important for not only for the country but also for the entire world. The Clustering is one of the data mining techniques [13] that are used for classifying text. In this paper, we are going to present aggregation of probabilities of words in a given sentences. Then by finding the aggregated probabilities we will get Accurate and efficient results.

2.2 The Relevance Similarity of Sentences:

This proposal is a new approach for measuring the similarity between the collection of words and then the sentences. The relevance of words in a sentence can specify how much similar they are to the given classes of clusters. The finding of this important similarity gives the proof using this method [13]. The combining of sub systems gives the better results. The proposed approach outperforms the similarity between the words by find out the probabilities with the use of sub systems of a given system.

2.3 Experiments on Probabilistic Sentence Level Clustering

Identifying required data plays an important role in Text mining. The proposed method is based on the concept of the aggregated probability of sentences [15]. It used to find the relevant information or data sentences from a collection of documents. It uses the concept of sub division of a system.

2.4 Aggregated Expectation Maximization Clustering Algorithm

Nowadays, large amount of data is available in the form of texts. It is very difficult for the people to find out useful and significant data [16]. For getting useful data we have many different algorithms. The useful data can be taken from the large amount of data which is available and this data is in short and concise form. This proposal describes a system, which consists of two steps. In first step, we are finding out the relevance of words in a given sentence [17]. In second step, we are implementing Expectation Maximization Clustering Algorithm [18] to find out sentence similarity between the sentences based on the value of aggregated relevance of sentences, we can easily estimate the matching of text data to a class.

2.5 Fuzzy Logic

Fuzzy Logic concept was used in many applications. Basically, it permits the values to be defined for use. This logic is very popular now a day because of its effective identification of T or F [6]. This can be used to find out the True or false. This can be used in different operations. This concept can be shown with an example below. This diagram shows the importance of fuzzy concepts in various applications. Let take three age groups young, middle aged and old. It shows below with Fuzzy concept [11]. The fuzzy concepts are used in the different applications today. It can add the better identification of data or text to the users who can expect the efficient data[11]. In the following example the persons who are in between middle aged and old can also be identified efficiently with the usage of fuzzy concepts [6].

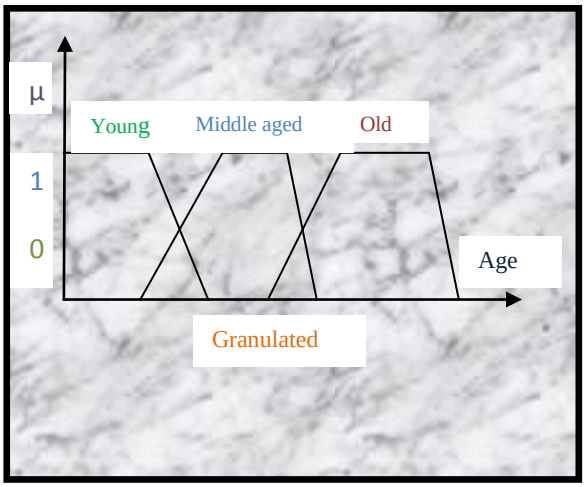

Figure1. Fuzzy vs Non Fuzzy classification

Fuzzy Classification Example1:
Let us take an Example to classify the text using Fuzzy logic and fuzzy set concept which is very advantageous.

Tall(x) = {
0, if height(x) < 5 ft.,

(Height(x)-5ft.)/2ft., if 5 ft. <= height (x) <= 7 ft.,

1, if height(x) > 7 ft.

}

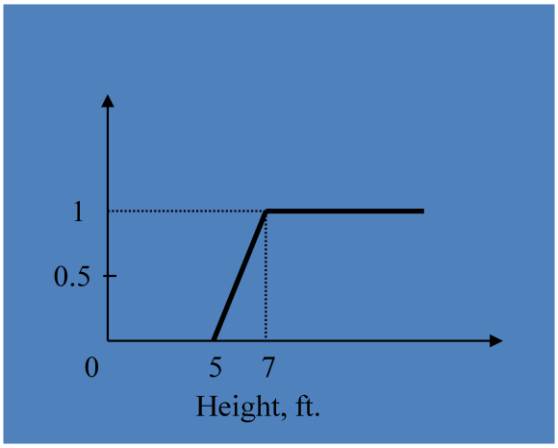

Figure 2: Fuzzy classification for Example1

In this Figure 2: text can be classified efficiently using fuzzy logic.

2.6 Relation by Relevance

In this aggregated probabilistic algorithm the relation between the sentence and a particular class can be find out by calculating the relevance and aggregation of words in a given sentence. The highest probability of sentence can have a highest priority to belong into a particular class of cluster.

3. PROPOSED WORK

In this wok, the analysis of one can take advantage of the efficiency and stability of clusters, when the data to be clustered are available in the form of similarity relationships between pairs of words. More precisely, we propose a new aggregated probabilistic Fuzzy relational sentence level expectation Maximization clustering algorithm which does not require any restriction on the relation matrix. This APFRSEC algorithm is applied for the clustering of the text data which is present. APFRSEC will give the output as clusters which are grouped from text data which is present in a given documents. In this APFRSEC algorithm, Page Rank algorithm is used as similarity measure.

3.1 Page Score Value

The description the application of the algorithm to data sets, and shows that the algorithm performs better than other fuzzy clustering algorithms. In the proposed algorithm, the use of Page score algorithm is to calculate the number of page hits and traffic. Page score algorithm is used to determine the importance of a particular node or thing to visit for usage. This algorithm specifies the most occurrences of things in use. This score is known as Page rank Score. Sentence is represented by system in this assumption. Then find out the most possibility of nearer representing similarity between sentences. It can represent the important relevance to a sentence. The aggregation concept here used to data reduction and better performance. Sentence in a document is represented by a node in the directed graph and the probabilities specify the similarity to a class. The page score algorithm retrieves the data based on the number of hits.

3.2 Emax Algorithm

It is a method that tries to find aggregated probabilities of a sentence that has the maximum likelihood of getting nearer results. Its main role is to calculate nearest estimation. It is an important method, which is mainly used to finding the maximum aggregated probabilities of the model. The E-step consists the calculation of aggregated probabilities. The probabilities calculated from E-step are compared in Max-step for maximum values.

3.3 APFRSEC

The Main idea of aggregated probabilistic Fuzzy Relational sentence level Expectation maximization clustering Algorithm clustering algorithm is used in this work to implement the separation of information among the various subsystems, which are organized into subsystems. Each of these subsystems of a given system can be calculated for relevance and then find out the matching. A System structure with n sub models fuzzy systems is depicted in Fig. 2. Each of this system structure has correspondent fuzzy system relevance $R_i(x)$. In the following, the system is explained with the perspective of sentence and the words are taken as sub systems to reduce the burden and identify efficiently.

This fuzzy concept system describes the importance of the method to be used. Therefore, the output of the above model is the aggregation of the each sub system component of each fuzzy system. Let the classes be c_1, c_2, c_3,…….c_m. The sentence can be divided into appropriate words like w_1, w_2, w_3,….w_j.

The probability of a word belonging to a class can be calculated from the following:

$$\text{Relevance } R_i(x) = P(c_m / w_i) = \frac{\sum_{q=1}^{n} Yqi * Fj}{\sum_{q=1}^{n} yqi} \quad \text{,------(1)}$$

yqi specifies the number of occurrences of word wi in document yq.

Where Fj =1 ,if document yq belongs to class c_m

0 ,otherwise .

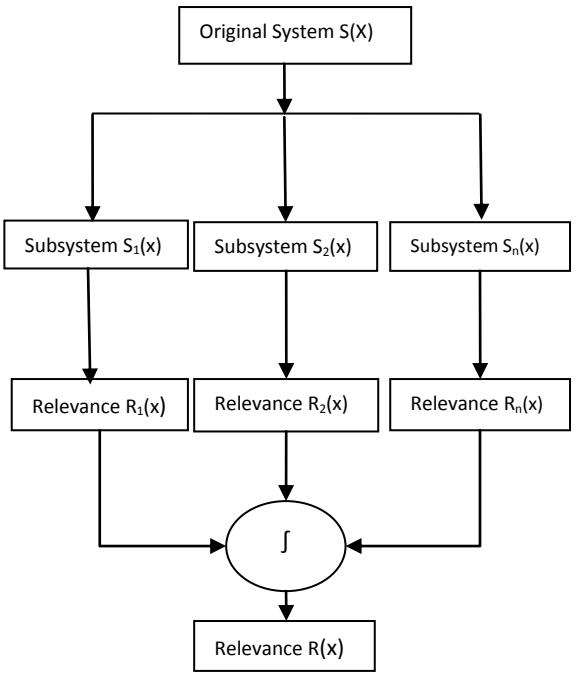

Fig 3. Aggregated Relevance of Sub Systems

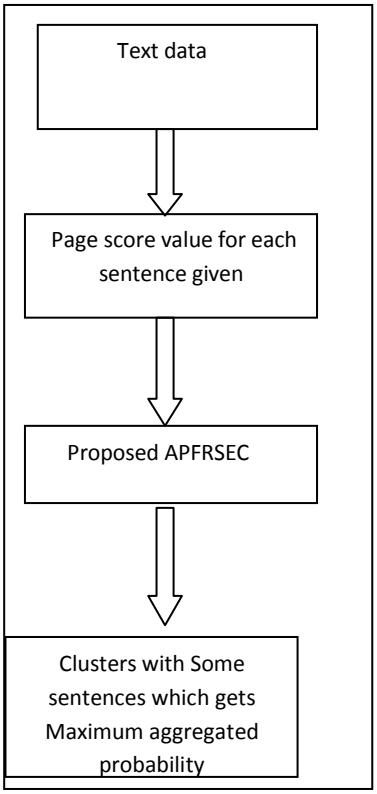

Fig4: APFRSEC Clustering Process

The sentence level classification is definitely giving some accurate results by finding out the aggregate relevances of different words in a given sentence.Where $R_i(x)$ represents the relevance function of the i th fuzzy subsystem covering the point **x** of the Universe of Discourse. The relevance $R_i(x)$ reveals the accurate contribution to the respective original fuzzy system. This should be considered in the aggregation of all Sub systems for getting effective results. Aggregation means combining two or more attributes into a single attribute. The main purpose of using this aggregation method is to data reduction and also aggregated data tends to have less variability. For example, cities aggregated into regions, states, countries etc. Aggregation is also called as summation. The relevance of an aggregated system is calculated from the following equation.

$$^{R}_{i}(x) = \cup_{i=1}^{n} Ri(x) \quad \text{,----------(2)}$$

By using this formula, it is to find out the relevance of each word in a given sentence and then it is to be decided that sentence belongs to which class. Because of finding out the probability, we get accurate results.

3.4 The Division of a System into Sub Systems

The Clustering is a a way to separate a set of data or text X into some subsets that represent some sub structures of **X**. The system structure can be divided into sub systems. Then solve for the probabilities for each subsystems and aggregating the probabilities of all subsystems in a given system. Then we get the maximum probabilities to get into a particular cluster. If we get the same probabilities for two clusters. Then it belongs to both the clusters. The Division of a system into sub systems is very important method to solve any complicated problem. By dividing it into sub systems, the process can be done easily with accuracy. In this proposed Method we are taking care about the aggregation of probabilities which gives accurate Results for classifying efficiently. Finally the sentences can be belongs to a particular cluster based on the accurate aggregate relevance of sentences. The procedure for this is shown below with an algorithm. The reason for usage of this concept is to accurate specification of results and also simplicity in searching. With this the search results are low when compared with word wise searching. If the searching is based on sentence then the search results are low and give accurate outcome.

3.5 Proposed APFRSEC Algorithm

Our proposed aggregated probabilistic Fuzzy Relational sentence level Expectation maximization clustering Algorithm is developed using Sub system concept with the application of Probability. It is used in this work to implement the separation of information among the various subsystems, which are organized into a original System structures. Each of these subsystems may contain information related with particular aspects of the system.

Initialization:

# of original word patterns : m

# of sentences formed : s

# of classes : p

Initial #of clusters : k=0

Input:

$X_i = <x_{i1}, x_{i2}, x_{i3},\ldots\ldots x_{ip}>, 1<=i<=s>$

Output:

Clusters G1, G2; . . .; Gk

Procedure for APFRSEC Algorithm:

For Each word pattern in a Sentence S_i,1<=i<=m

Probability Relevance R_i is finding out by (1);

Then aggregating the probability relevance of each word in a given sentence by (2).

The Sentence which gets maximum aggregated Probability of Relevance can belongs to a Class C_i .

Return with the Created K clusters;

End Procedure.

4. EXPERIMENTAL RESULT

Table 1: Document Set D1

DOC	AU (w1)	M.Tech (w2)	First (w3)	Sem (w4)	Result (w5)	Class
d1	1	1	0	1	0	C1
d2	1	1	1	1	1	C1
d3	0	0	0	0	1	C2
d4	0	1	1	0	0	C2
d5	1	0	0	1	0	C2

For example,

Suppose there are two classes C_1 and C_2. The five documents d1, d2, d3, d4 and d5 belonging to c1, c1, c2, c2, c2 respectively. See the Table 1. For values of words appearing no. of times and belongs to a class. Let the occurrences of w1 in these documents be 1, 2, 3, 4 and 5 respectively. Then, the probability of word w1 belongs to a class c1 is calculated as

$$R_1(x) = P(c1/w1) = \frac{1*1+2*1+3*0+4*0+5*0}{1+2+3+4+5} = 0.2\,,$$

$$R_2(x) = P(c1/w2) = \frac{1*1+2*1+3*0+4*0+5*0}{1+2+3+4+5} = 0.2\,,$$

Similarly we have to find the $R_3(x)$ or P(c1/w3), $R_4(x)$ or p(c1/w4) and $R_5(x)$ or P(c1/w5). Then the finding of aggregating these all words probabilities belonging to a particular class C1 . Then getting a result of aggregated relevance of sentence S_1 belongs to a class C1 as follows.

Then the values found are,

$$R1(x) = 0.2$$

$$R_2(x) = 0.2$$

$$R_3(x) = 0.4$$

$$R_4(x) = 0.1$$

$$R_5(x) = 0.4$$

Agg $_{R(x)}$ of S1 belongs to a class C_1 =

$$R_1(x) + R_2(x) + R_3(x) + R_4(x) + R_5(x)$$

$$= 0.2 + 0.2 + 0.4 + 0.1 + 0.4$$

$$= 1.3$$

After that take sentence S_1 and dividing it into possible words like w1, w2, w3, w4 by sub systems concept. Then calculate the following probabilities of relevance of words w_i in a given sentence Like $R_1(x)$ or P (c1/w1), $R_2(x)$ or P (c1/w2), $R_3(x)$ or P (c1/w3), $R_4(x)$ or P (c1/w4). Then the finding of aggregating these all words probabilities belonging to a particular class C_2. Then getting a result of aggregated relevance of sentence S_1 belongs to a class C_2 as follows.

Agg $_{R(x)}$ of S1 belongs to a class C_2 =

$$R_1(x) + R_2(x) + R_3(x) + R_4(x)$$

$$= 0.3 + 0.2 + 0.3 + 0.4$$

$$= 1.2$$

If Agg $_{R(x)}$ of S_1 belongs to a class C_1 > Agg $_{R(x)}$ of S_1

Belongs to a class C_2. Then the sentence S1 is Belongs to Class C1.

Similarly, the aggregated relevance is find out for each and every sentence to get accurate outcome. The sentence gets the maximum relevance of belonging to a class can get the priority. The efficient text classification is very useful today for all. The sentence level text classification is performed very efficiently. For, example when searching for the required data, this algorithm gives the low search results with highest accuracy. So, the aggregated probabilistic algorithm can perform well and gives the accurate results. The performance of this algorithm is described with neat diagrams and also show the test sample and their net score level percentage for FSFC and APFRSEC algorithms. When discuss about any algorithm, the specification of its importance and efficient performance is very important. So, the following illustrations can be useful to better understand about the algorithm and its efficiency when performing the action.

Table 2: Various Clustering Algorithm performance

Techniques	Max class of objects assigned	Mixed quality	Combinatorial Index
APFRSEC	0.810	0.601	0.865
Filtered	0.672	0.461	0.503
Spec.Clus	0.671	0.482	0.799
Farthest first	0.421	0.458	0.578

In table1, the comparison is performed out for different numbers of clusters. We compare the performance of APFRSEC algorithm with previous algorithms to the dataset and evaluating using the external measures.

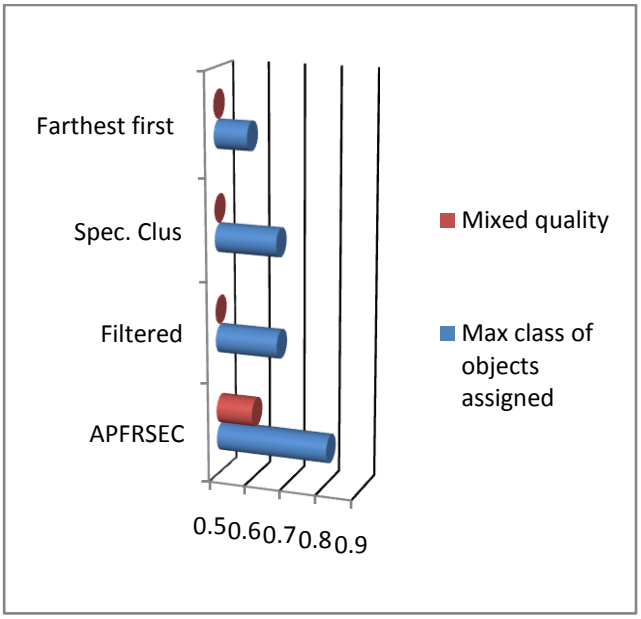

Fig5: Max class of objects assigned and mixed quality Comparison

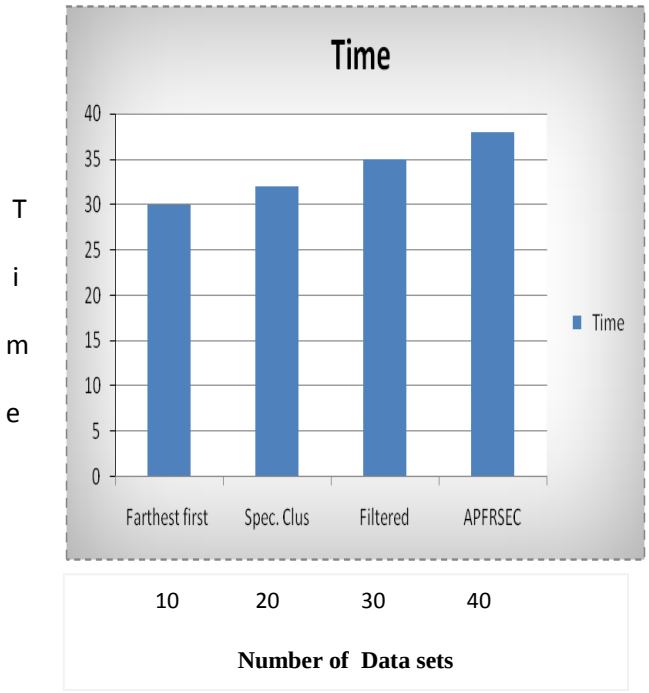

Fig 6: Comparison with Time

Table 3. Text sample vs. Average relevance % Level

Test	Average relevance percentage Level (%)	
	Fuzzy Self FC Algorithm	APFRSEC Algorithm
Text sample 1	87.69	92.30
Text sample 2	87.99	93.05
Text sample 3	88.73	92.68
Text sample 4	86.40	93.62
Text sample 5	88.49	95.01
Text sample 6	88.99	95.02
Text sample 7	91.00	96.00

In each algorithm, the affinity matrix was used and pair wise similarities also calculated for each of the method. But in APFRSEC algorithm, we find out the accurate aggregates of probabilities to get accurate and efficient Results. It is to be understood from the observations that the APFRSEC algorithm gives accurate Results.

In this Aggregated probabilistic Fuzzy Relational sentence level Expectation maximization clustering Algorithm, It can forms the Sentence Level Clusters Using Aggregated Relevance of Sub systems And find out the Probabilities of each valuable Word In a given sentence .Our Method Find out The Relevance of each Word from above stated formula and find out the Aggregate of Probabilities of collection of words in a given Sentence. Then We Get the maximum likelihood estimates to belong to a particular Cluster Accurately.

5. CONCLUSION & FUTURE WORK

In this Work, aggregated probabilistic Fuzzy Relational sentence level Expectation maximization clustering Algorithm is proposed and applied to clustering fuzzy sets. . This algorithm gives the low search results with highest accuracy. It gives the More Accurate and efficient Results When compared to Existing System. Thus we found maximum likelihood estimates of parameters. When the number of data sets increases, then the APFRSEC algorithm takes more time to perform clustering. This proposal can be useful in future for research work.

6. ACKNOWLEDGEMENTS

Our sincere thanks to the experts who have contributed towards development of this paper.

7. REFERRENCES

[1] Jung-Yi Jiang, Ren-Jia Liou, and Shie-Jue Lee, Member, A Fuzzy Self-Constructing Feature Clustering Algorithm for Text Classification, March 2011, IEEE, VOL.23, NO. 3.

[2] N. Slonim and N. Tishby. "The power of Word clusters for text classification," 23rd European Colloquium on Information Retrieval Research (ECIR), 2001.

[3] Neha Mehta, Mamta Kathuria, Mahesh Singh, "Comparison of conventional and fuzzy ClusteringTechniques: A survey", April 2014 IJARCCE, Vol. 2, Issue 1.

[4] G.Thilagavathi, J.Anitha, K.Nethra,"Sentence Similarity based Document Clustering using Fuzzy algorithm", March 2014, IJAFRC, Vol1, Issue 3.

[5] K.Jeyalakshmi1, R.Deepa2, M.Manjula," An Efficient Clustering Sentence-Level Text Using A Novel Hierarchical Fuzzy Relational Clustering Algorithm", February2014, IJARCCE, Vol. 3, Issue2.

[6] M.S.Yang," A Survey of Fuzzy clustering", October 1993, Vol 18, No 11.

[7] F. Pereira, N. Tishby, and L. Lee."Distribution of A clustering of English words," 31st Annual Meeting of ACL, 1993, pages 183– 190.

[8] [8] Hathaway RJ, Bezdek JC Recent convergence Results for the fuzzy C-means clustering Algorithms, Oct 1988. J Class 5:237-247.

[9] S.M. Jagatheesan1, V. Thiagarasu2," Development of Fuzzy based categorical Text Clustering Algorithm for Information Retrieval", January 2014, vol 2, issue 1.

[10] K. Nalini Dr. L. Jaba Sheela," Survey on Text

[11] Classification", IJIRAE, July 2014, Vol 1, Issue6

[12] Roventa, E., Spircu, T."Averaging Procedures in De fuzzification Processes, Fuzzy Sets and Systems ", 2003,136, pp. 375-385.

[13] S.J.Lee and C. S. Ouyang. "A neuro-fuzzys System modeling with self-constructing rule Generation and hybrid svd-based Learning," IEEE Transactions on Fuzzy Systems, June 2003, 11(3):341– 353.

[14] G. Salton and M. J. McGill. Introduction to Modern Retrieva,. McGraw-HillBook Company, 1983.

[15] F. Sebastiani. "Machine learning in automat Edtext categorization," ACM Computing Surveys, 34(1):1–47,

[16] March 2002.

[17] L.X.Wang. A Course in Fuzzy Systems and Control.

[18] Prentic-Hall International, Inc., 1997.

[19] Y. Yang and J. O. Pedersen. "A comparative Study on feature selection in text categorization.

[20] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter, "Distributional Word Clusters Versus Words for Text Categorization," J.Machine Learning Research, 2003, vol. 3, pp.1183-1208

[21] L.D.Baker and A. McCallum, "Distributional Clustering of Words for Text classification, "Proc.ACM SIGIR, pp, 1998,Pages96-10.