iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 2 **Issue:** III **Month of publication:** March 2014

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

ISSN: 2321-9653

INTERNATIONAL JOURNAL FOR RESEARCH IN APPLIED SCIENCE AND ENGINEERING TECHNOLOGY (IJRASET)

Optimal Sentence Clustering Using An Innovative Hierarchical Fuzzy Clustering Algorithm

Christy Maria Joy[1], S. Leela[2]

[1]PG Scholar, Computer Science and Engineering, Karunya University, Tamil Nadu, Coimbatore

[2]Research Scholar, Computer Science and Engineering, Karunya University, Tamil Nadu, Coimbatore

Abstract*-The role of data clustering is inevitable in many text processing activities .Many proceedings are going on in this area since it has wider applications. Sentence clustering is a challenging task when compared with other data clustering, because a sentence is able to represent same ideas in different ways. For E.g. some people see a glass as half empty and some others see half full. Due to this variability fuzzy relational clustering algorithms are more able to produce optimal results in sentence clustering applications, since they allow the data elements to blend in more than one cluster with different cluster membership values. This paper presents a novel hierarchical fuzzy clustering algorithm which is capable of identifying the sub clusters within a cluster using a partitioning method.*

Index Terms: Fuzzy Clustering, Hierarchical Clustering, Membership Values, Sentence Clustering, Membership values

I. INTRODUCTION

Information Technology has been going through tremendous development during the last decade. This explosion in the IT industry paved the way towards the intensification of amount of data. But most of these data are not lying in a useful manner. The role of data mining is very significant in this scenario. Clustering is an important technique that can be used in the data mining process for extracting the information which underlying the data. The unsupervised clustering technique is able to categorise the data based on different notion, which enables us to understand various hidden themes under the data. There are many methods in data mining process to retrieve relevant information from datas.

Clustering can be applied in many fields such as medical diagnosis, pattern recognition, business applications, information retrieval (IR) document summarization etc. But clustering text at document level poses differences with that of sentence–level text clustering. Document clustering, break down the documents into different clusters based on some main themes. It won't consider the semantics of each sentences within that document .But a sentence with only small fragment of text can able to represent more than one theme within a document. So the roles of fuzzy clustering methods are vital in this area of sentence clustering, because fuzzy clustering techniques allow a prototype to fall in all clusters. Cluster membership value is a matter of degree in all fuzzy based methods. Various studies are showing that inclusion of sentence clustering in extractive multidocument summarization helps avoid problems of content overlap, leading to better coverage. [1] [2] [3] [4].Similarity function has played a great role in every clustering process and it can be find out by distance functions.

In addition to theme based summarization of documents, sentence clustering can be also employed in web mining, micro-level contradiction analysis, event classification in unstructured texts etc. Experimental Results are showing that inclusion of fuzzy techniques in sentence clustering domain provides better results than other hard clustering methods. In this paper we present a new fuzzy clustering method for sentence clustering using a hierarchical means. The experimental results are showing that hierarchical fuzzy clustering outperforms the other clustering techniques. This algorithm is a combination of various techniques such as Page Rank algorithm [5] , Expectation –Maximization [6] ,FRECCA algorithm [1] , Hierarchical Clustering Algorithm [x] etc.

ISSN: 2321-9653

The paper is structured as follows: Related work is described in section 2.Section 3 describes the algorithm. Section 4 presents the results and discussion. Section 5 draws the conclusion.

II. RELATED WORKS

Sentence clustering is a type of text classification since they are small sized texts, it posses many challenges than other text classification process. In the paper "Sentence Clustering via projection over term clusters" Lili Kotlerman et al says that language variability is the main challenge of sentence clustering. Most of the standard clustering algorithm can be used to cluster sentences. But fuzzy based sentence clustering algorithms shows improved performance evaluation results.

The evolution of fuzzy based clustering techniques can be traced from early 1960's.But the Hathaway et al.'s Relational Fuzzy C-Means algorithm [7] became more popular among other methods at that period. Later many alternative algorithms such as ARCA [8] based on Fuzzy C-Means came which overcomes the challenges of of RFCM. K Medoids [9] algorithm which is an advanced version of K Means algorithm is more robust in the presence of noise and outliners. In 1999 Geva presented a hierarchical unsupervised fuzzy clustering algorithm. But the general hierarchical clustering algorithm [11] is defined by S.C. Johnson in (1967). Horng et al. (2002) presented fuzzy information retrieval techniques using fuzzy hierarchical clustering and fuzzy inference techniques [18]. Shyi-Ming Chen et al. (2007) introduced a fuzzy hierarchical clustering method for clustering documents based on dynamic cluster centers [16]. Fuzzy swarm diversity hybrid model for text summarization [17] is a successful method introduced by Mohammed Salem Binwahlan et al in 2010.

FRECCA algorithm [1] proposed by Andrew Skabar et al (2013) is the recent innovation in this area of fuzzy based sentence clustering. In his paper Andrew Skabar performed sentence clustering using a challenging quotation data set with different algorithm such as Spectral clustering algorithm [10], K-medoids algorithm, ARCA algorithm, and FRECCA algorithm etc. The performance evaluation results precisely shows that FRECCA algorithm outperforms other algorithm and is capable of identifying softer clusters. The FHC method is an integration of fuzzy relational clustering and hierarchical clustering into sentence clustering domain.

III. A COLLABORATIVE HIERARCHICAL FUZZY CLUSTERING ALGORITHM

This section presents a new collaborative fuzzy hierarchical algorithm for sentence clustering. We first describe the general hierarchical clustering algorithm, then a Fuzzy Relational Clustering Algorithm and a combined fuzzy hierarchical algorithm (FHC) for optimal sentence clustering. Since the new algorithm is a combination of different algorithms it able to acquire the advantageous properties of the various methods. The overview of the FHC framework is depicted in figure 1.

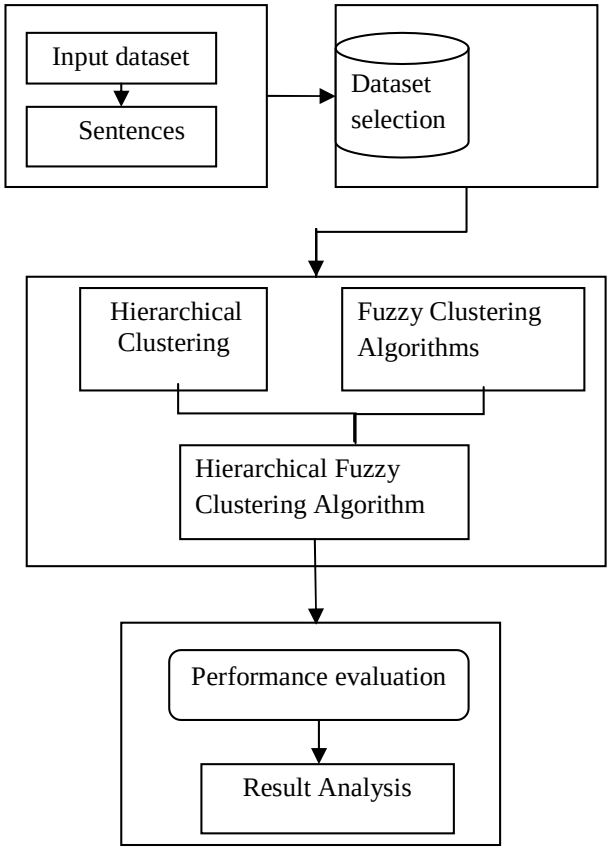


Figure 1. Overview of Hierarchical Fuzzy Clustering

A. Hierarchical Clustering Algorithm

Clustering technique will classify data elements based on some properties. Hierarchical clustering uses the distance matrix as clustering criteria and will build a hierarchy of clusters. It does not require the number of clusters ***k*** as an input. Actually its working is based on the union between the two nearest clusters. This clustering is also capable to generate sub clusters of a cluster. For Example, it can partition Sports cluster into two sub clusters such as Games and Athletes.

The general hierarchical algorithm [11], can be describe as follows: Start with each item to a cluster, so that each containing just one item. After that their similarity will measure, then find the most similar pair of clusters and merge them. Repeat the process of finding similarity values with new cluster until all data elements belongs to the same cluster.

B. Fuzzy Relational Clustering (FRC)

As we explained in section II, FRECCA algorithm [1] is a recent renowned algorithm for sentence clustering and is capable of identifying softer clusters. PageRank [5] is using here as a measure of centrality. And the algorithm proceeds based on the Expectation-Maximization [6] frame work. The Expectation Step finds out the PageRank value and also the cluster membership probabilities and the maximization step is only updating the mixing coefficient which is initialised during beginning phase of algorithm.

C .A combined Hierarchical Fuzzy Relational Clustering Algorithm

The Hierarchical Fuzzy Relational Clustering Algorithm is a hybrid method which is a combination of general hierarchical clustering concepts [11] with fuzzy relation models that is existing fuzzy clustering algorithms [1] [16].The fuzzy clustering algorithms which we used here are from references [1] and [16].

Before applying fuzzy relational algorithm, the dataset or sentences will partition into N items and then finally return the N number of clusters .Then find the similarity between the sentences in the text, and finally we will apply the fuzzy relational model. Cosine similarity is used here. Since we are merging different sentence clusters during the initialization stage of the FHC method the clusters centers will change each time .The input to the algorithm are set of sentences and its similarity values. And the desired outputs of the algorithm are clusters with membership values.

Algorithm

Step 1: Let each sentence be a cluster and let the membership degree of each clustering belonging to its another sentence level text cluster be equal to 1

Step 2: Find the proximity matrix of distances (similarity value)

Step 3: Assign sequence numbers to clustering and then perform fuzzy relational clustering algorithm for sentence clustering.

Step 4: After performing the fuzzy relational algorithm check whether the membership value is smaller than the particular threshold value or not using the value obtained from FRC. If yes, then compute the new membership value and update the mixing coefficient. Otherwise go to next step.

Step 5: Combine the first sentence cluster with second cluster into a new sentence cluster. Suppose first cluster contains n sentences and second cluster contains m sentences

Step 6: Then again find the degree of similarity between each pair of sentence clusters

Step 7: If the degree of similarity between any two sentence clusters is smaller than the threshold value, then Stop.

Stop 8: Otherwise, go to Step 5 and repeat the process.

IV .EVALUATION METRICS

The general clustering evaluation criteria are used here. Partition Entropy Coefficient (PE) [12] is an unsupervised cluster evaluation criterion. But Purity and Entropy [13] are most commonly used external evaluation criteria for clustering. If clustering has purity values close to 1, then it is good clustering, if it is close to zero ,then we can say that it is bad clustering. Also for a good clustering the purity values should be high and entropy value should be low. When the number of clusters is large, high purity can be easily achieved. Because of this we can't use purity as the basic criterion for clustering evaluation. So here we are also using a more reliable V-measure [14] for evaluation. This V-measure is considering both homogeneity and completeness. Rand Index [15] and F-Measure are taking into account the true positives (TP), true negatives (TN), false negatives (FN) and false positives (FP) values of objects and is commonly used in Information literature.

ISSN: 2321-9653

V. RESULTS AND DISCUSSION

Table 1 shows the results of applying FRC and FHC algorithm. The criteria of a good clustering are high intra class similarity and low inter class similarity. The cluster quality can determined by similarity function and the object representation used in it.

Table 1: Cluster evaluation on Sentences Text

Metric Used For Cluster Evaluation	FRC	FHC
PE	0.9947	0.7946
Purity	0.6250	0.7247
Entropy	0.5542	0.8975
V-Means	0.2038	0.63019
Rand	0.7530	0.7530
F-Means	0.5435	0.8576

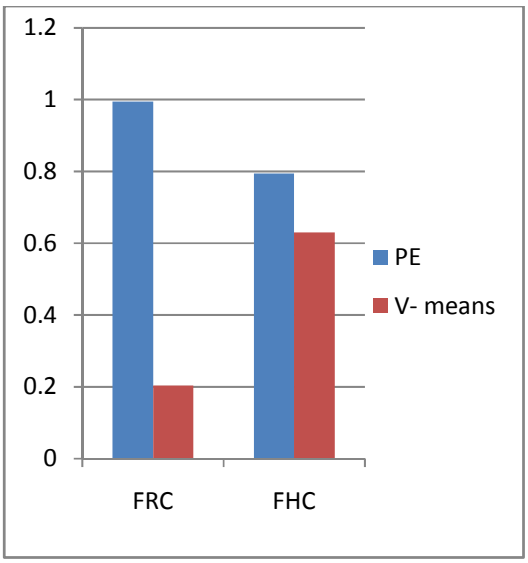


Figure 2: Comparison of FRC and FHC with PE and V-means

Figure 2 represents the comparative graph of PE values and V measure of two algorithms. From the graph it is clear that FHC has lower PE values than FRC. If the values of PE are minimum, it implies a good partition in the meaning of a more crisp partition.FHC has also a high V-measure value.

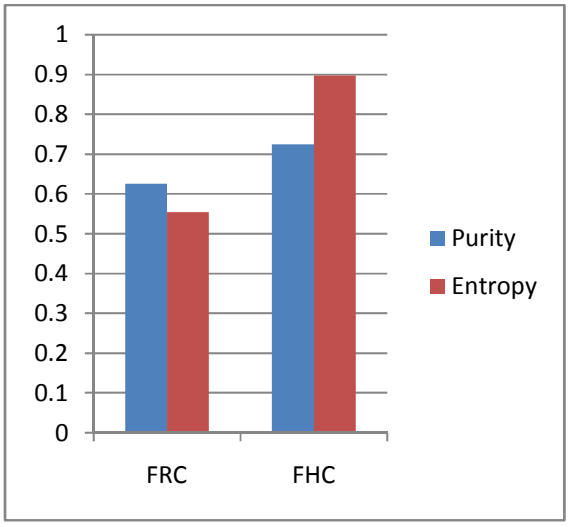


Figure 3: Comparison of FRC and FHC using Purity and Entropy

Figure 3 precisely shows the high purity value of the FHC. High purity indicates good clustering process. But it has also high entropy value and is not a favourable condition for good clustering.

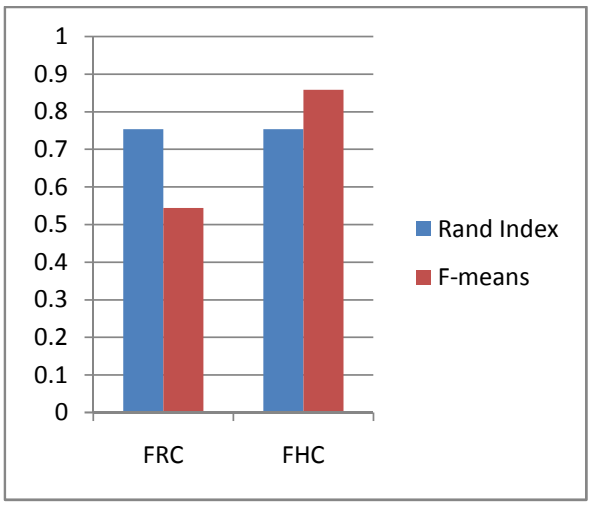


Figure 4: Comparison of FRC and FHC using Rand Index and F-means

From the fourth graph we can see that there is an improvement in the F –measure value by applying the FHC method.But it shows the same value for the Rand index in both existing and new FHC method.

VI .CONCLUSION

In this paper we introduced a hybrid Fuzzy Hierarchical Clustering algorithm for sentence- level text clustering. The effective feature selection, similarity measure selection and proper choice of algorithm will give good clustering of texts. Experimental results based on the FHC algorithm shows that it is capable of producing superior results than other existing clustering algorithms. When the number of input sentences increases, the corresponding time required for processing also increases a lot. But it has high purity value ,V-means ,F-means value etc. This algorithm can be applied in different applications.

VII. REFERENCES

[1] Andrew Skabar and Khaled Abdalgader, "Clustering Sentence-Level Text Using a Novel Fuzzy Relational Clustering Algorithm", IEEE Transactions On Knowledge and Data engineering, vol. 25, 2013

[2] 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.

[3] Ramiz M. Aliguliyev "A new sentence similarity measure and sentence based extractive technique for automatic text summarization" Expert Systems with Applications 36 , pp.7764–7772, 2009.

[4] 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.

[5] S. Brin and L. Page, " The Anatomy of a Large-Scale Hypertextual Web Search Engine," Computer Networks and ISDN Systems, vol. 30,pp. 107-117, 1998.

[6] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," J. the Royal Statistical Soc. Series B (Methodological), vol. 39, no. 1, pp. 1-38, 1977.

[7] R.J. Hathaway, J.W. Devenport, and J.C. Bezdek, "Relational Dual of the C-Means Clustering Algorithms," Pattern Recognition,vol. 22, no. 2, pp. 205-212, 1989.

[8] 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.

[9] L. Kaufman and P.J. Rousseeuw, "Clustering by Means of Medoids," Statistical Analysis Based on the L1 Norm, Y. Godge, eds., pp. 405-416, North Holland/Elsevier, 1987.

[10] U.V. Luxburg, "A Tutorial on Spectral Clustering", Statistics and Computing, vol. 17, no. 4, pp. 395-416, 2007.

[11] S. C. Johnson : "Hierarchical Clustering Schemes" Psychometrika 2 :241-254,1967

[12] J.C. Bezdek, "Cluster Validity with Fuzzy Sets," J. Cybernetics ,vol. 3, no. 3, pp. 58-72, 1974

[13] C.D. Manning, P. Raghavan, and H. Schu tze, Introduction to Information Retrieval. Cambridge Univ. Press, 2008.

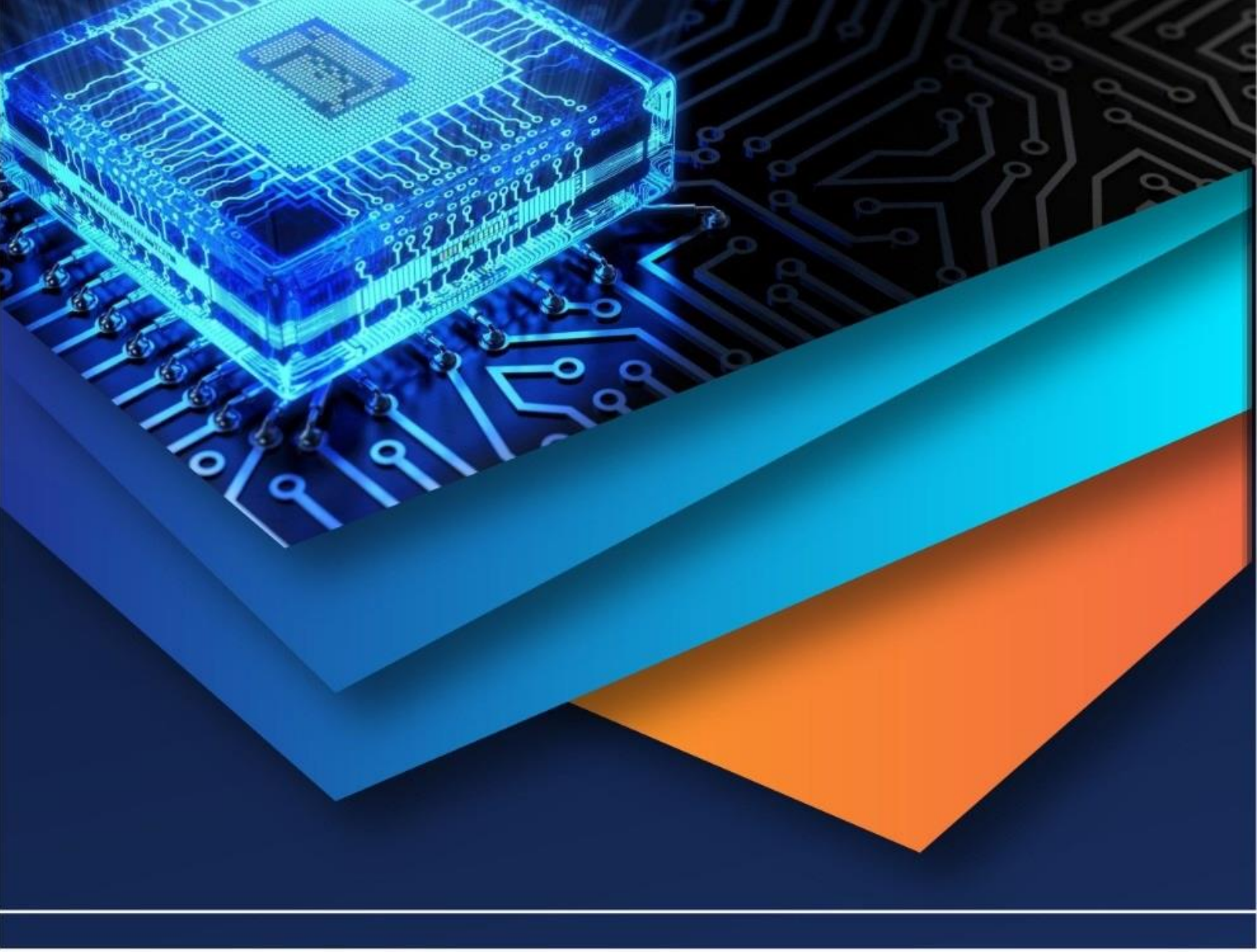
[14] A. Rosenberg and J. Hirschberg, "V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure," Proc Conf. Empirical Methods in Natural Language Processing (EMNLP '07),pp. 410-420, 2007.

[15] W.M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," Am. Statistical Assoc. J., vol. 66, no. 338, pp. 846-850,1971.

[16] Shyi-Ming Chen and Liang-Yu Chen," A fuzzy hierarchical clustering method for clustering documents based on dynamic cluster centers", Journal of the Chinese Institute of Engineers, Vol. 30, No. 1, pp. 169-172 , 2007

[17] Mohammed Salem Binwahlan , Naomie Salim , Ladda Suanmali, "Fuzzy swarm diversity hybrid model for text summarization", Information Processing and Management 46 pp.571–588,Elsevier 2010

[18] Horng, Y. J., Chen, S. M., and Lee. C. H., 2002, "Fuzzy Information Retrieval Using Fuzzy Hierarchical Clustering and Fuzzy Inference Techniques," Proceedings of the 13th International Conference on Information Management, Taipei, Taiwan, Republic of China, Vol. 1, pp. 215-222.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY