CLUSTERING APPROACH FOR CLASSIFICATION OF RESEARCH ARTICLES BASED ON KEYWORD SEARCH

Dr. S. K. Jayanthi[1], C. Kavi Priya[2]

[1] Head and Associate Professor, Dept. of Computer Science, Vellalar College for Women, Erode, Tamilnadu, India
[2] Research Scholar, Department of Computer Science, Vellalar College for Women, Erode, Tamilnadu, India

ABSTRACT: Growth of research articles publication in various streams of research is exponential. Searching for a particular article from the research repository is considered to be a tremendous one and also time consuming. Research articles classification based on their respective domain plays an important role for researchers to retrieve articles in a fast manner. Hence a popular search mechanism, namely keyword search has been applied to retrieve appropriate articles, documents, texts, graphs and even relational databases. When new domains of documents are added to the repository it has to identify keywords and add to the corresponding domains for proper classification. A numerical statistic called TF-IDF has been proposed to determine the relevance of word to a document corpus. Clustering algorithms namely Hierarchical, K-Means and Fuzzy C-Means have been used to cluster articles based on the relevance factor TF-IDF. The strength of Fuzzy C-Means clustering has been validated using Silhouette Cluster Validation technique. Finally, performance has been evaluated using Precision, Recall and F-measure and demonstrated that Fuzzy C-Means clustering depicts better accuracy compared to K-Means and Hierarchical clustering.

Index Terms: Classification, Fuzzy C-Means, Hierarchical, K-Means, TF-IDF, Silhouette Cluster Validation

I. INTRODUCTION

Text mining techniques are applied on static repository of research documents. Now a days the number of documents in repository increases with time. New domain of documents related to various areas may be added, which in turn requires identification of keywords for classification of each document. When large numbers of research articles are received, it is necessary to organize them according to their similarity in research domain. Due to large volume of published articles, classification is extremely difficult and also a time consuming one. Hence various clustering approaches have been used to reduce time consumption.

This paper is to classify a stream of research articles inwards to the research repository. Then keyword list is constructed for each domain of articles. The main objective is to extract similar group of articles based on the given set of keywords using Term Frequency - Inverse Document Frequency (TF-IDF). Finally Hierarchical, K-Means and Fuzzy C-Means clustering approaches separate the documents from dataset using the TF-IDF score values.

The rest of this paper is organized as follows: Research inference is made based on the related work shown in Section II. Methodology is described to classify articles using clustering approaches in Section III. Results and performance evaluation is done in Section IV. Conclusion and future scope is given in Section V.

II. RELATED WORKS

Worarat Krathu et.al (2016) discussed about the optimal data mining workflow for the classification task. The main contribution includes (i) the application of data mining for discovering success factors and their relationships, and (ii) the optimal workflow as a standardized flow for further similar classification tasks. The major challenge of this work is that there exists no mature corpus in this context, and hence the approach is implemented without a supporting corpus. The result shows that the Support Vector Machine(SVM) performs better than other classifiers.

Ramanpreet Kaur et.al (2016) proposed a new method supporting clustering and classification, using k-means with feed forward neural networks to reduce time consumption. A model based on supervised as well as unsupervised technique has been developed to achieve the similarity between documents.

Snehal Shivaji Patil et.al (2015) had discussed about the tasks of searching similar patterns of text that is to be more effective, efficient and interactive. The current method for research paper selection is based on similarities of keywords and frequencies based on ontology. Research papers in each domain are clustered using text mining technique. Grouped research papers are reviewed systematically by

appropriate reviewer or domain experts and finally ranked based on the review results.

Twinkle Svadas1 et.al (2015) explained that the exponential growth of data has led to an information explosion era, where the data cannot be easily maintained. The system has been proposed to categorize the text documents and form the clusters. The document collection is obtained and pre-processing techniques are applied to remove stop words and to do tokenization. A document clustering algorithm is used for categorizing the news articles on the basis of topic. The results are improved by using the document clustering algorithm rather than simple clustering algorithms.

Ning Zhong et.al (2012) discussed about the processes of pattern deploying and pattern evolving, to improve the effectiveness of using and updating discovered patterns for finding relevant and interesting information using the approaches of D-Pattern Mining algorithm and inner patter evolution. Substantial experiments on RCV1 data collection and TREC topics demonstrate that the proposed solution achieves better performance.

Based on the review given above, it is observed that Hierarchical, K-Means and Fuzzy C-Means clustering have been used to classify the research articles. To improve the efficiency, TF-IDF has been calculated to optimize the search results.

III. METHODOLOGY

Initially, keywords for research articles based on different domain such as Data Mining, Networking, Cloud Computing and Image Processing have been collected from multiple web sources. Then the keyword list created by expert is used for classification process based on binary representation. Finally, TF-IDF is further taken for article classification process.

Under the classification phase, three types of clustering namely Hierarchical, K-Means and Fuzzy C-Means Clustering have been carried out to classify relevant articles. Once all the clustering process are completed, validation metrics such as precision, recall and F-measure has been computed for each clusters. Hence, Silhouette Cluster Validation Technique is used to validate the clusters. At last, the cluster value were plotted as graph for visual representation of results. The workflow of research articles classification is illustrated in **Fig.1**.

A. DATASET COLLECTION

The dataset for proposed work is downloaded from the websites http://scihub22266oqcxt.onion.link/ and http://liinwww.ira.uka.de/bibliography/Misc/Citeer/. 100 papers from the list of reviewed papers at random have been considered as the dataset. These papers are likely to be in four domains namely Data Mining, Networks, Cloud Computing and Image Processing with 25 papers in each domain. Keywords have been generated for each domain.

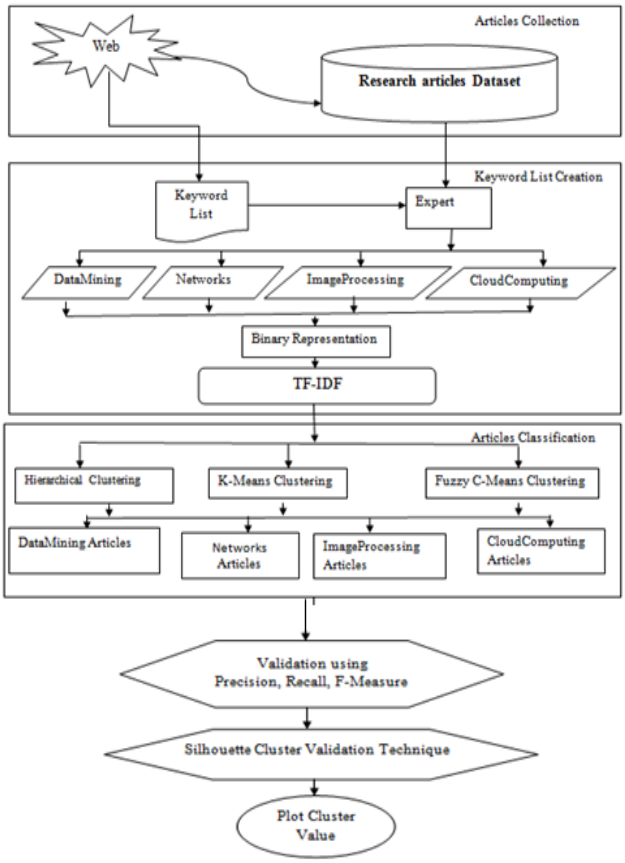

Fig.1 Research Framework

B. BINARY REPRESENTATION

Classification is based on binary representation in which the papers that contain keywords are said to be in the positive class(1) and the rest in negative class(0). 25 research articles per domain has been considered as the dataset. Keywords for each domain namely Data Mining, Networking, Cloud Computing and Image Processing have been considered are 16, 13, 12 and 14 respectively.

Binary representation of sample articles based on the sample Data Mining domain keywords is show in Table 1.

Table 1 Binary Representation of Articles based on Keywords

	1.txt	2.txt	3.txt	4.txt	5.txt
Clustering	1	1	1	1	1
Classification	1	1	0	0	1
Data	1	0	1	1	1
Analysis	1	0	1	1	0
Pattern	0	0	0	1	1

To refine categorization further, by setting a threshold value for average term frequency for each document.

C. TF–IDF

The Term Frequency tf(t,d) is the number of times that term t occurs in document d and is given as follows

$$\mathbf{tf}(t,d) = f_{t,d}$$

IDF is calculated by the number of document in the dataset divided by the number of document where a specific term appears and is given as follows

IDF(key) = log(Total number of documents / Number of documents with term in it)

TF-IDF is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining. TF-IDF value increases proportionally to the number of times a word appears in the document. TF-IDF is one of the most popular term-weighting schemes.

tfidf(t,d,D) = tf(t,d) * idf(t,D)

D. CLUSTERING APPROACHES

Clustering is one of the most essential approach for dealing massive amount of research articles from repository. Clustering algorithms have been used to cluster similar articles into one cluster. In this paper three clustering algorithms namely Hierarchical, K-Means and Fuzzy C-Means have been used to classify articles.

Hierarchical Clustering

Hierarchical clustering works by grouping data objects into a tree of cluster. The hierarchy can be constructed in top-down (called divisive) or bottom-up (called agglomerative) fashion. Hierarchical clustering algorithms are one of the Distanced-based clustering algorithms, i.e. using a similarity function to measure the closeness between text documents. In the top-down approach it begin with one cluster which includes all the documents. Then recursively split cluster into sub-clusters. In the agglomerative approach, each document is initially considered as an individual cluster. Then successively the most similar clusters are merged together until all documents are grouped in one cluster.

K-Means Clustering

K-Means clustering is one of the partitioning algorithms which is widely used in the data mining. This clustering partitions n documents in the context of text data into k clusters. It is a easy way to classify a given article dataset through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. The next step is to take each point belonging to a given article data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early groupage is done. At this point, re-calculate k new centroids resulting from the previous step. After calculation of k new centroids, a new binding has to be done between the same article data set points and the nearest new centroid. A loop has been generated. As a result of this loop one can notice that the k centroids change their location step by step until no more changes are done.

Fuzzy C-Means Clustering

Fuzzy C-Means (FCM) is one of the most popular fuzzy clustering algorithms. Fuzzy C-Means (FCM) is a method of clustering that assign membership levels and use them to assign data elements to one or more clusters. It uses reciprocal distance to compute fuzzy weights. Every element of the universe can belongs to any fuzzy set with a degree of membership that varies from 0 to 1. FCM introduces the fuzziness for the belongingness of each object and can retain more information of the dataset.

This algorithm works by assigning membership to each data point of article dataset corresponding to each cluster center on the basis of distance between the cluster center and the data point. After each iteration membership and cluster centers are updated.

IV. RESULTS AND DISCUSSION

Classification of articles of various domain has been done using Hierarchical clustering, K-Means clustering and Fuzzy C-Means clustering.

A. EXPERIMENTAL ANALYSIS

The Process of proposed work contains the following steps:

Step 1: First, the research articles and keywords are collected from multiple websites http://liinwww.ira.uka.de/bibliography/Misc/Citeer/ & http://scihub22266oqcxt.onion.link/.

Step 2: 100 articles selected randomly from the list of reviewed articles have been considered as the dataset. These papers are likely to be in four domains namely Data Mining, Networks, Cloud Computing and Image Processing with 25 articles in each domain.

Step 3: Keywords have been generated for each domain by the expert.

Step 4: The articles' that contain keywords are said to be in the positive class(1) and those that does not contain keywords are said to be in negative class(0) under binary representation.

Step 5: Next to the binary representation process, TF-IDF value is computed to find the number of occurrences of the keywords in each document. The TF-IDF produces a final result which is further taken

for article classification process.

Step 6: The classification is done using the clustering approaches namely Hierarchical, K-Means and Fuzzy C-Means clustering by using the training and testing dataset.

Step 7: Finally Precision, Recall and F-measures values are calculated to compute the accuracy.

Step 8: Validate clustering approaches using Silhouette cluster validation technique.

Table 2 Classification of Articles using Clustering Approaches

Domain	Number of Articles (100)	No. of Articles Classified based on Keywords		
		Hierarchical	K-Means	Fuzzy C-Means
Data Mining	25	10	10	11
Networks	25	3	15	16
Image Processing	25	8	21	23
Cloud computing	25	4	13	13

From Table 2 it has been observed that Fuzzy C-Means clustering performs better classification compared to Hierarchical clustering and K-Means clustering.

B. PERFORMANCE EVALUATION

To measure the accuracy of clustering, Precision, Recall and F-measures values are calculated.

Precision is a measure of exactness or quality. It is a fraction of documents that are relevant among the entire retrieved documents and it is given by

$$\text{Precision} = |Ra| / |A|$$

where,
Ra: Set of relevant documents retrieved
A: Set of documents retrieved

Recall is a measure of completeness or quantity. A fraction of the documents that is

retrieved and relevant among all relevant documents is defined as recall. Basically, it gives coverage of result.

$$\text{Recall} = |Ra| / |R|$$

where,
Ra: Set of relevant documents retrieved
R: Set of all relevant documents

F-measure is the harmonic mean between precision and recall, and is defined as

$$F = \frac{2(P \,.\, R)}{P+R}$$

Table 3 Result Analysis

	PRECISION			RECALL			F-MEASURE		
Cluster / Domain	HCLUST	KMEANS	FCM	HCLUST	KMEANS	FCM	HCLUST	KMEANS	FCM
Data Mining	1	1	1	0.4	0.4	0.44	0.57	0.57	0.61
Networks	1	1	1	0.12	0.6	0.64	0.21	0.75	0.78
Image Processing	1	1	1	0.32	0.84	0.92	0.57	0.91	0.95
Cloud Computing	1	1	1	0.16	0.52	0.52	0.27	0.68	0.68

From Table 3 it shows that Fuzzy C-Means has higher Precision, Recall and F-measure values when compared to the Hierarchical clustering and K-Means clustering. Through the performance evaluation, it is shown that the proposed Fuzzy C-Means is capable to achieve best accuracy.

C. Silhouette Cluster Validation Technique

This describes about the strength of the clustering. The cluster value is calculated by Silhouette Plot value. Silhouette refers to a method of interpretation and validation of consistency within clusters of data.

Silhouette cluster validation technique is applied to cross validate the Fuzzy C-Means clustering. The range of silhouette cluster validation is -1 to 1

- 0.71 – 1.00 excellent split
- 0.51 – 0.70 reasonable structure has been found
- 0.26 – 0.50 weak structure, could be artificial
- ≤ 0.25 horrible split

Table 4 Silhouette Value of Fuzzy C-Means Clustering

S. No.	Domain	Silhouette Value
1	Data Mining	0.72
2	Networks	0.68
3	Image Processing	0.69
4	Cloud Computing	0.72

Table 4 shows cluster validation of Fuzzy C-Means Clustering for the above mentioned domains which provides excellent split for Data Mining and Cloud Computing and reasonable split for Networks and Image Processing.

V. CONCLUSION AND FUTUREWORK

The current work utilized, Hierarchical, K-Means and Fuzzy C-Means clustering approaches to classified articles of various domain using keyword list based on TF-IDF value. As a result it is proved that Fuzzy C-Means clustering is better than Hierarchical clustering and K-Means clustering. Classification of domain articles based on clustering approaches help to increase searching efficiency and reduces searching time.

Text features that can be extracted automatically using natural language processing or information extraction tools can be used. In future, the research articles dataset can be of any type such as pdf, html as well as doc format. This work can also be extended for clustering the document based on phrases as well as documents of different languages. Since, this paper concentrates on four major domains, the work can be prolonged by including more documents in various domain such as health, entertainment.

REFERENCES

[1] Sindhu Antony, Rupali Wagh, *Study on Text Clustering For Topic Identification*, International Journal of Advanced Research in Computer Science, ISSN No. 0976-5697, Volume 8, No. 1, Jan-Feb 2017.

[2] Ramanpreet Kaur and Amandeep Kaur, *Text Document Clustering and Classification using K-Means Algorithm and Neural Networks*, Indian Journal of Science and Technology, Vol 9(40), DOI: 10.17485/ijst/2016/v9i40/97722, October 2016.

[3] Worarat Krathu, Praisan Padungweang, and Chakarida Nukoolkit, *Data Mining Approach for Automatic Discovering Success Factors Relationship Statements in Full Text Articles*, proceedings of the 8th International Conference on Advanced Computational Intelligence Chiang Mai, Thailand, 14-16th February-2016.

[4] Twinkle Svadas, Jasmin Jha, *Document Cluster Mining on Text Documents*, International Journal of Computer Science and Mobile Computing, ISSN 2320–088X, Vol.4 Issue.6, pg. 778-782, June- 2015.

[5] Snehal Shivaji Patil, S.A.Uddin, *Research Paper Selection Based on an Ontology and Text Mining Technique using Clustering*, IOSR Journal of Computer Engineering (IOSR-JCE),e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 17, Issue 1, Ver. I (Jan – Feb. 2015), PP 65-71.

[6] Pawar T. A., Karande N. D., *Effective Pattern Discovery for Text Mining Using Pattern Based Approach*, International Journal of Advance Research in Computer Science and Management Studies, ISSN: 2321-7782 (Online), Volume 2, Issue 9, September 2014.

[7] Ye Yuan, Guoren Wang, Lei Chen, and Haixun Wang, *Efficient Keyword Search on Uncertain Graph Data*, IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 12, December 2013.

[8] Czarnecki J., Nobeli I., Smith A. M., and Shepherd A. J., *A Text Mining System for Extracting Metabolic Reactions from Full-Text Articles, BMC bioinformatics*, Vol. 13, No. 1, p. 172, July- 2012.

[9] Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, *Effective Pattern Discovery for Text Mining*, IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 1, January-2012.

[10] Cohen A. M. and Hersh W. R., *A Survey of Current Work in Biomedical Text Mining, Briefings in Bioinformatics*, Vol. 6, No. 1, pp. 57–71, March-2005.

[11] Huang M., Zhu X., Hao Y., Payan D. G., Qu K., and Li M., *Discovering Patterns to Extract Protein–Protein Interactions from Full Texts, Bioinformatics*, Vol. 20, No. 18, pp. 3604–3612, July- 2004.