ISSN 2347 - 3983

Volume 8. No. 9, September 2020

International Journal of Emerging Trends in Engineering Research

Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter175892020.pdf

https://doi.org/10.30534/ijeter/2020/175892020

# An Efficient Document Clustering in Enhancing Fuzzy Logic for Uncertain Data using Big Data in Cloud Storage

[1]G.Parimala , [2]Dr.R.Kayalvizhi, [3]S.Nithiya,

[1]Assistant Professor, Department of IT,SRM institute of Science and Technology,Chengalpattu,
parimalg@srmist.edu.in

[2]Assistant Professor, Department of CSE,SRM institute of Science and Technology,Chengalpattu,
kayalvir@srmist.edu.in

[3]Assistant Professor, Department of IT,SRM institute of Science and Technology,Chengalpattu,
nithiyas@srmist.edu.in

## ABSTRACT

The presence of data in documents has become to be bigger and unearthing such datasets are a provocative assignment. The objective of Enormous data is to store, recover and examination different content archives. The recovery of the indistinguishable data over huge databases is of major concern. Existing issue is fathomed by proficient document clustering in which suggests design coordinating methods that permits looking of numerous catch phrases at a particular time. In this paper, we consider different content data as input and handled utilizing pre-processing procedures like Key Express extraction, Stemming for tokenizing and Improved Fuzzy rationale. The ultimate result is to create clusters and list them for the modern dataset. In this way, this proposed algorithm comes about in progressing the execution of document clustering on comparison.

**Keywords:** big data, document clustering, fuzzy logic, algorithms

## 1 . INTRODUCTION

The rising components of social organize stages, cutting edge methods for changing common data into advanced organize are worked as bits and bytes. Working stage regions like wearable gadgets, sensors get to web in conjunction with created chucks of data to improve trade and its working for persistent handle[1-4]. The essential require and utilize of data over the web frame distinctive sources around the world cannot not be blocked. Typically totally transformative mechanical stage. The created data should be beated for legitimate capacity utilizing categorical classification, clustering and relapse procedures. Document clustering/categorization is one of the stick forms in mining content, sound, video data's. Utilizing fuzzy rules to measure the degrees of having a place of objects with regard to clusters, fuzzy clustering gives a common way to capture collaboration among document clusters.

In order to successfullyprotectsufficientdata of the initialdata, distinctive representation models have been created for documents. The progression of Big data in cloud computing is being risen with different sources utilizingprogressed processors now-a- days. Most utilized applications through web like twitter, facebook, whatsappdetailed that every day is been enacted by numerous million portableclients from distinctivenations having numerousdialects uploading billion of gallery's each day and keeping up billions of companionassociations and exchanging bundles of datafrequently.

There are a few occasions in which data of blended sort may actually emerge. A self-evident case is that of persistent data with either lost or censored values comparing to a discrete category. Whereas taking care of such data could be a standard include in numerous clustering algorithms, a more challenging circumstance emerges on the off chance that a few discrete categories such as “removed”, “unknown”, “incorrect”, or “censored”, all of which may carry data pertinent for finding an optimal clustering arrangement, are at the same time show within the data. Another normal case of blended data is unearthly data, emerging for occasion in numerous chemical and natural examinations. These sort of data comprise of non-negative genuine values with an intemperate extent of zeros, meaning the nonappearance of a measured include.

## 2. RELATED WROK

The system recommended the reason of opinion investigation has come to colossal thought in content mining. With unstructured arrange of data, copious commotion evacuation will has got to be done which is costly to evacuate[5-7]. Quick development in advanced strategies and computer program apparatuses, are utilized to preprocess the loud data utilizing opinion analysis.Phases like data procurement, preprocessing, include extraction and representation, naming of data. Algorithms like Characteristic Dialect Handling as well as machine learning calculations are actualized. Text Mining with Lucene and Hadoop (TMLH)connected these algorithms to the unstructured content records for opinion investigation in content mining.

The systemutilizesdataconnected on pictures to calculate the non-negative factorization values based on spatial and plenitudelimitations[8,9]. Blends of pixels are organized through restrictions of spatial determination and resultsgotten against protestacknowledgment and classification. The pixels are analyzed in straightblenddemonstrate, nonlinear blendshow, BiLinearblendshow in ghastlyunmixingmethod. It coordinates to analyze the most materials and finds the comparable divisionscreated from the hyper ghastlysymbolism of a location. Additionally, NMF is the leadingactualizeddemonstrate for the straightghostlyblend. It finds the edge focuses and decides the plenitudes at the same time. Through the compression or extraction of data, pixels are specificallydecayedneighborhoodleast and decrease the meeting. The creators have executed the NMF algorithm by considering it as unusedlimitation based smoothness and extraction of highlights with adequacy.

The system proposed NMF in enormous data stage for document clustering. They have considered computer program instruments like Apache Hadoop and Apache Lucene to handle different content documents[10-12]. They have distinguished the issue with gigantic data in unstructured shape. Data that has been dumped into the database or framework must be classified accurately and adjusted legitimately. It'll gather the unlabeled data in clusters. A demonstrate that has been as of now in presence like SVD and LSI has diminished the data in a efficient way. Expansion to them, an upgraded rules of NMF brought raise in document clustering with an intrigued. They have concentrated on the NMF rules that underpins the k-means clustering approach. Within the particular setting, the content documents are preprocessed and executed based on Key include extraction and content documentation in Characteristic Dialect Preparing. Afterward, comes about are produced utilizing disseminated parallel execution through Hadoop.

The systemexecuted k-means clustering approach for modeling. It may be ahypotheticaloutline of the genuine world mechanisms. Expecting the similar characteristic strategieswithinthedata[13-20]. The foremostwinningstrategies for finding covered updesigncalculations are classified into two approaches, unsupervised (clustering) and supervised (classification). Where classes are not characterized for unsupervised and classes are characterized for directed. They have centered on clustering calculations by considering the numerical comes aboututilizing probability density function where objects are distinguished and set into comparativebunches called clusters. This algorithm is utilized to anticipateclimatedata reports. Expectationcoursenames are either yes or no. This has demonstrated that k-means clustering calculation is effective. In addition, many cloud based security systems have been developed by various authors by incorporating the auditing systems [22-24].

## 3. PROPOSED METHODOLOGY

i. Document Pre-processing
ii. Enhanced Fuzzy Logic for Uncertain Data (EFL-UD)

### 3.1 Document Pre-processing

Data preprocessing could be adata mining procedures that includeschangingcrudedata into anjustifiableorganize. Real world data is frequentlyinadequate, conflicting or missing in certain behaviors or patterns and is likely to contain numerousblunders[13-19]. Data preprocessing could be ademonstrated method of settling such issues. Data preprocessing prepares raw data for encouragepreparing. Datawithinthegenuine world is messyfragmentedthat's missingtrait values, missing certain attributes of intrigued, or containing as it weretotaldata. Data are loudthat'sit contains blunders or exceptions. They are conflictingthat'sit contains disparities in codes or names. No quality data is found so there will be no quality mining comes about. Quality choices must be based on quality data. Datawarehouse needs steady integration of quality data.The flow of data process is given in the figure1.

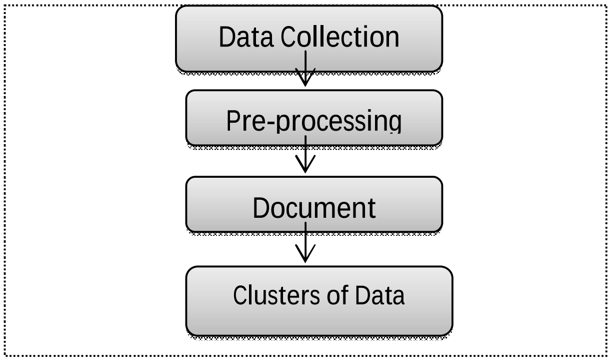

**Figure 1:** Flow diagram of data processing

**Stages of Document Pre-processing**

- Tokenization (lexeme, tokens)
- Lemmatization (word elimination)
- Data Optimization ( Computing frequency)

### 3.1.1 Tokenization

Tokenization is basicallypartaexpress, sentence, passage, or a completecontentdocument into littler units, such as person words or terms. Each of these littler units are called tokens[21-27]. Tokenization is the method of turning delicate data into nonsensitive data called "tokens" that can be utilized in a database or internal system without bringing it into scope. Tokenization is the method of tokenizing or part a string, content into a list of tokens. One can think of token as parts like a word could be a token in a sentence, and a sentence could be a token in a paragraph. The fundamental lexical unit of a dialectcomprising of one word or a few words, the components of which don't independentlypass on the meaning of the total.

### 3.1.2 Lemmatization:

It is the form of gathering the archedshapes of a word so they can be examined as a single thing. Lemmatization is the algorithmic approach of deciding the lemma of a word based on its aiming meaning. Unlike stemming, lemmatization depends on accuratelydistinguishing the expectingportion of discourse and meaning of a word in a sentence, as well as inside the biggersettingencompassing that sentence, such as neighboring sentences or indeedacompletereport. Lemmatization is closely related to stemming. The distinction is that a stemmer works on a single word without data of the setting, and so cannot separate between words which have distinctiveimplications depending on portion of discourse. Be that as it may, stemmers are regularlysimpler to actualize and run speedier.

### 3.1.3 Data Optimization

Data optimization implies collecting all the data at your transfer and overseeing it in a way that maximizes the speed and comprehensiveness with which basicdata can be extricated, analyzed and utilized. The data optimization process to get to, organize, and cleanse data, anything the source, to maximize the speed and comprehensiveness with which germanedata can be extricated, analyzed, and put to utilize. Data Optimization may be aprepare that plans the consistentconstruction from the data seepattern. A design of system architecture is given in figure 2 as flow diagram.

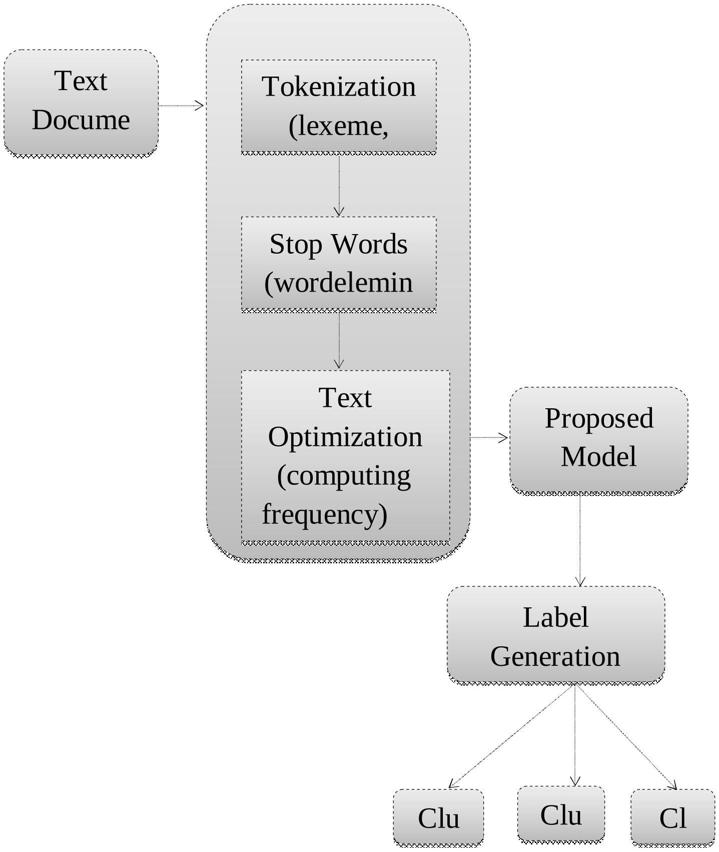

**Figure 2**: A flow of Proposed System Design

### 4.EFC-UC Algorithm:

```
Input : dataset of N token
        k, clusters
        T, iterations
Output: Document matrix $U_{n×k}$,
        word matrix $V_{k×m}$.
        $N= [n_1, . . . ,n_m]$
Random sampling documents $N_t$ from N
        $N_t$;
        $N ← N / N_t$ ;
        Set equal weights to documents w = [1, . . ,M]T ;
        if t > 1 then
        $N_t ← [V_{t−1},N_t]$;
        $w ← [w_{t−1}$ center, w];
        end
        $[V_t,U_t] ←$ EFL-UC (w, $N_t$, $V_{t−1}$, k);
        $w_t$
        center = $U_t$w;
        t ← t + 1;
        end
        U ← Non-iterative Extension $(N,V_{t−1})$;
```

### 3.2 Enhanced Fuzzy Logic – Uncertain Data (EFL- UD):

We summarized the issues that have been utilized to upgrade the fuzzy clustering for taking care of huge data in the literature. Here, the dataset of $N$ tokens, $N= [n_1, . . . ,n_m]$ with $n_i \in k$, $k$ *i*s thefuzzygathering. The most contrast among the strategies is the path to create the*k*centersin whole dataset, i.e., $\Delta_k= [k_1, . . . ,k_n]$. Each centres is an'*s*'directed points that represents the mid of a cluster[28-31]. Once the $k$ centroids are identified, non iterativegowthis embraced to induce fuzzy participations for tokens. In our discussions, we center the route of producing the ultimate $k$ centroids to distinguish the clustering.

The clustering strategy utiizes a '$n$' dubious tokens, '$t$' into $k$ clusters. Utilizing Uncertain Clustering ($UC_{clust}$)the disparity as similitude, a clustering moves to segment tokens into $k$ clusters and identify the leading *k*groups, one for each cluster, to play down the Uncertain cluster($UC_{clust}$) dissimilarity as below

$$P(j) = \sum_{j=1}^{k} \sum_{p \in L_j} D(P||L_i)$$

Where, '*j*' ranges form*1,2,….,k*. $L_i$ is the length of the cluster and *P(j)* is the probability of the clustering.

For a token, $t$ in cluster, $k$ the $UC_{clust}$ divergence $D(t||L_i)$ between $t$ and the representative $k_i$ measures the additional data required to build $t$ given $k_i$. Hence, $\sum_{t \in L_i} D\langle t|L_i \rangle$ captures the overall additional data required to construct using its representative, $L_i$. Adding overall $k$ clusters, the $UC_{clust}$ divergence in this way to count the quantity of the grouping. The bit the value of $UC_{clust}$, the better the clustering. Within the building stage, the *uncertain k-medoids*strategy uses an initial clustering by selecting *k*incharges one by one. The primary representative $L_i$is the one which has the lowest sum

$$L_i = \min \left\{ \sum_{P \in P',\ P \leftarrow P'} D(P'||\,P) \right\}$$

Where, $P$ has the probability that the comparable tokens are connect together to make a certain cluster. $P'$has the probability of uncertain data. The algorithm chooses the representative $L_i$ which diminishes the $UC_{clust}$divergence as much as conceivable. For each token, $t$ which has not been chosen, we test whether it tought to be chosen within the current circular. For any other nonselectedtoken $P'$, $P'$ will be allocted to the new representative $P$ if the divergence $D(P'||\,P)$ is lesser than the divergence between $P'$ and any orderly selected representatives.

In the proposed framework, a subset of tokens which are almost less sufficient for stacked into memory are to begin with chosen from the given dataset. Clustering is at that point performed upon this subset. Extension could be an ensuing handle where theclustergives the name non-sampled tokens so that all token within the unique dataset are clustered. In FCM is connected on a subset created with irregular samples to create the $k$ cluster centroids. The centres are utilized to analyse participations of all other tokens in anoniterative fashion.

Beneath the over common system of clustering, we presently examine more points of interest of Enhanced- FCM, i.e., the fuzzy c-means based approach. In this approach, Weighted Fuzzy C-Means (WFCM) is received to clustering each chunk by consolidating token weights, i.e., token are related with diverse weights to appear to project the set of significance by clustering. Expect that protest $n_i$ is related with a weight V, the objective of WFCM is to play down thefollowing function.

$$f(x) = \sum_{k=1}^{n} \sum_{i=1}^{m} Vu_i^c .\ dist(n_i, k_c) + C$$

Where, V denotes the weight of the token, C is the constant coefficient, $u_i^c$ is the fuzzy member, $dist(n_i, k_c)$is the distance between $n_i$and $k_c$.

As the primary term is consistent for a given $n_i$, as it were choose the cluster grant of the token . For the comfort of discourse, considering difficult task, i.e., each token is allocated to the cluster with the lowesttoken-to-centroid distance, i.e., $c = min\{\ dist\ (n_i\,, k_c)\}$ . By this, $c$ is a sequentially connected of all token that have a place to this cluster, i.e.,

$$k_c = \frac{1}{|z|} \sum_{j \in z} z_j \qquad (2)$$

The over disparity implies that inadequate and top vertex data, when the chunk measure is little, the displacement $dE(n_i, k_c)$ is dominated by $|k_c|^2$, which is independent on $n_i$. In "identical-cluster" issue, i.e., all tokens are assigned to one cluster, the centroid with small displacement.

## 4.RESULTS

We offer a broad exploratory think about of the proposed fuzzy logic for document clustering. The results appear with expanded adaptability, the system accomplishes noteworthy changes in the viability of document clustering with existing fuzzy and non-fuzzy sets. This illustrates the awesome potential of our approaches for expansive document clustering.

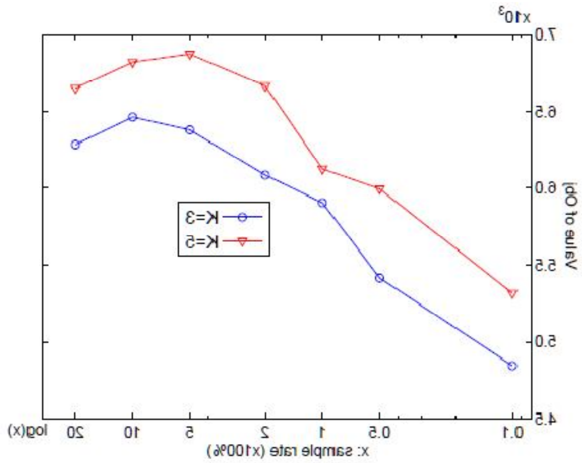

**Figure 3:** Sampling rate on object (token) value of EFL-UC

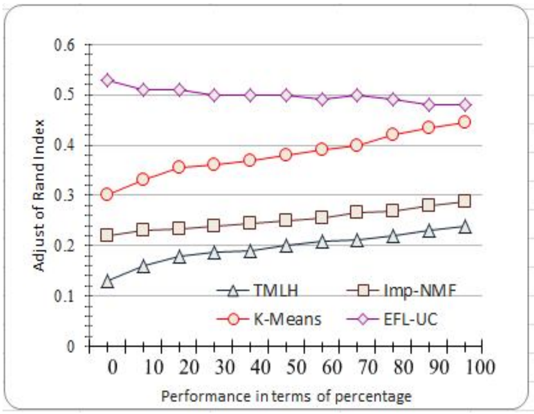

**Figure 4**: Comparison of the document clustering – Initialization of data set

Here, the ARI index is mapped with performance in terms of percentage which is high for EFL-UC when compare to other TMLH, Imp-NMF and K-means approach.

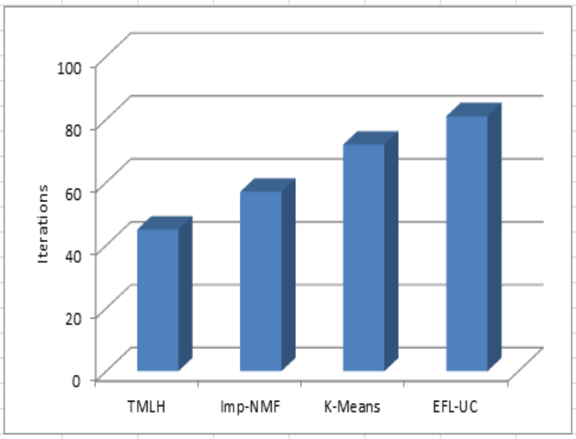

**Figure 5*:*** Similar data gathering in EFL-UC, TMLH, Imp- NMF & K-Means

Representation of similar data gathering (document clustering) in compared for TMLH, Imp- NMF, K-Means and EFL-UC. Figure 5 shows EFC-UC performs better than the existing systems.

## 5.CONCLUSION

In this paper, we propose an enhanced fuzzy logic for uncertain clustering (EFL-UC). This model is more likely discussed about the uncertainty of the document clustering by probabilistic approach. Once the uncertainty reached the certain probabilistic approach the propose system employs enhanced fuzzy logic to smoothen the uncertain document. Identifying the maximum similar tokens for clustering the data. The tests appear that our algorithm is exceptionally compelling on *document* × *term* parameter compared to conventional clustering and common apportioned co-clustering algorithms. This strategy shows up successful to characterize the document clusters. Our test results appear that the predominant viability of EFC-UC for handling huge data to represent scalability.

## REFERENCES

1. Yiheng Chen and Bing Qin “The Comparison of SOM and K-means of text clustering” School of Computer Science and Technology, Harbin Institute ofTechnology.
2. Porter M.F. “Snowball: A language for stemming algorithms”,2001.
3. K. Premalatha and A.M. Nataranjan,” A Literature Review on Document Clustering,” International Technology Journal,2010.
4. L.V. Bijuraj,”Clustering and its Applications,”Proceedings of National Conference on New Horizons in IT - NCNHIT2013.

5. DipeshShrestha,*"Text Mining with Lucene and Hadoop:Document Clustering with feature extraction"*, thesis - WakhokUniversity, 2009.

6. E. Laxmi Lydia and D. Ramya,*"Text Mining With Lucene And Hadoop: Document Clustering With Updated Rules Of NMF Non Negative Matrix Factorization"*, International Journal of Pure and Applied Mathematics, Volume 118, No.7 2018, pp 191-198.

7. SerhatSelcukBucak and Bilge Gunsel,*"Incremental Clustering via Nonnegative Matrix Factorization"*,2008 19th International Conference on Pattern Recognition. DoI: 10.1109/icpr.2008.4761104.

8. Manjula.K.S ,Sarvar Begum , D. VenkataSwethaRamana," Extracting Summary from Documents UsingK-Mean Clustering Algorithm," International Journal of Advanced Research in Computer and Communication Engineering,Vol. 2, August2013.

9. Dr.E.Laxmi Lydia, P.Govindaswamy, SK. Lakshmanaprabu, D. Ramya, "*Document Clustering based on Text Mining K-means algorithm using Euclidean Distance Similarity*", Journal of Advanced research in Dynamical & Control Systems, Vol.10, 02-Special Issue, 2018.

10. Abhay Kumar, Ramnish Sinha, Daya Shankar Verma and VandanaBhattacherjeeSatendra Singh, "Modeling using K-Means Clustering Algorithm",2012 1st International Conference on recent Advances in Information Technology.

11. Deepika Sharma," Stemming Algorithms: A Comparative Study and their Analysis," International Journal of Applied Information Systems (IJAIS)",Volume 4, September2012.

12. W.Sarada," A Review on Clustering Techniques and their Comparison," International Journal of Advanced Research in Computer Engineering &Technology (IJARCET) Volume 2, November2013.

13. MohitBhansali, Praveen Kumar, "*Searching and Analyzing qualitative data on personal computer*", IOSR Journal of Computer Engineering, e-ISSN: 2278-0661,p-ISSN:2278-8727 Volume 10, Issue2, April 2013, PP 41-45.

14. Jimmy Lin, DmitriyRyaboy, Kevin Wells, "*Full-text indexing for optimizing selection operations in large-scale data analytics*", ACM, San Jose, California, USA, 978-1-4503-0700-0/11/06, June, 2011.

15. I. S. Dhillon, S. Mallela, and D. S. Modha. Informationtheoretic co-clustering. In Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003), pages 89–98, 2003.

6. I.S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. KDD '01, pages 269–274, 2001.

17. I.S. Dhillon and D.S. Modha. Concept decompositions for large sparse text data using clustering. Mach. Learn., 42(1-2):143–175, 2001.

18. G. Govaert and M. Nadif. An EM algorithm for the block mixture model. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27(4):643–647, 2005.

19. G. Govaert and M. Nadif. Block clustering with bernoulli mixture models: Comparison of different approaches. Computational Statistics & Data Analysis, 52(6):3233–3245, 2008.

20. J. Tang, J. Tao, H. Urakawa, and J. Corander, "T-BAPS: A Bayesian statistical tool for comparison of microbial communities using terminal- restriction fragment length polymorphism (T-RFLP) data," Stat. Appl. Genet. Mol. Biol., vol. 6, no. 1, 2007.

21. J. Corander, M. Gyllenberg, and T. Koski, "Bayesian unsupervised classification framework based on stochastic partitions of data and a parallel search strategy," Adv. Data Anal. Classification, vol. 3, p.

22. K.Sreelatha and Dr. V. Krishna Reddy, "A Comprehensive review of Security Challenges for Data Deduplication and Integrity Auditing", International Journal of Emerging Trends in Engineering Research, Vol.7, No.11, pp. 725-732, 2019

23. Dr. JKR Sastry, B. TrinathBasu, "Extended OpenStack Architecture for Enforcing Comprehensive Security within Cloud Computing System",International Journal of Emerging Trends in Engineering Research, Vol.8, No.7, pp. 3271-3279, 2020.

24. Kantilal P Rane, "Design of Drone3dContour: A Novel Contouring System using Altitude Measurement and Cloud-Web Computing",International Journal of Emerging Trends in Engineering Research, Vol.8, No.6,pp.2395-2401, 2020.