International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Performance Exploration on Various Document Clustering Techniques with K-Means Family

V.Kumaresan, R.Nagarajan

Abstract: ***Clustering performs a important position in numerous fields which include Computer science & packages, facts, pattern reputation, system studying technique and find out dating among the files. Clustering focuses on document clustering, and other related area. Increase within the extent of statistics saved in virtual form (text, photograph, audio) has improved the need for requirement of an automated tool, that allows people to find and manage the records in an efficient way. Usually clustering refer to document clustering technique investigates the documents and find its relation. This paper center of attention on the a range of clustering methods and evaluation its overall performance. This paper also categories the document clustering techniques as three major groups, namely Group K-means, Expectation Maximization and Semantic-based techniques (Hybrid method). Several experiments were conducted to analyse the performance accuracy and Speed.***

KEYWORDS — K means, K*, Hybrid, data set, bisection.

I. INTRODUCTION

Clustering may be a critical half within the application for a ramification of fields also as facts processing, applied math information analysis and compression Clustering has been advanced in numerous strategies. The elemental Clustering disadvantage is that of grouping alongside (clustering) facts matters that are much like each other. The major fashionable method to Clustering is to study it as a density estimation disadvantage. The document Clustering become inside the maximum essential investigated for up the exactitude structures accomplice diploma as a cheap manner of finding the closest neighbours of a report. The intention of this observe is to offer a well-known assessment of various Clustering strategies in information processing.

A method for grouping set of statistics gadgets into multiple agencies/clusters so objects inside the cluster have excessive similarity, however are extraordinarily assorted to matters within the alternative clusters is thought as "clustering".

Clustering has been utilized in lots of numerous areas and an oversized variety of diverse clustering algorithms profession to differing types of pc report and having various programs are latest.An extraordinary deal relies upon on a way to define the similarity among objects. Similarity among devices may be measured in phrases of their proximity (distance), or as a relation among the alternatives they exhibit.

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
V.Kumaresan*, Assistant Professor, Annamalai University,Tamilnadu, India.Email: kv.kumaresan@gmail.com.
Dr. R. Nagarajan, Assistant Professor, Annamalai University, Annamalai Nagar, India. Email: rathinanagarajan@gmail.com

Clustering is companion degree unattended mastering approach. Hence, there are not any predefined classes then the type is based on inherent applied arithmetic shape of the overall assortment of enter dataset.

Document Clustering offers with the unattended partitioning of a record collection into significant groups supported their count content, usually for the aim of topic categorization; i.e. A private cluster contains files on one subject matter while completely one-of-a-kind clusters can comprise files on exceptional topics. Not like record sort which will be a supervised gaining information of approach that wishes preceding records of file categories to instruct a classifier, file agglomeration is companion credentials unattended gaining information of technique that doesn't have faith in preceding categorization records.

The paper provides the revelation on the power of document clustering with k-means that. Even most of the clustering method the percentage of outliers is far beyond 60% and the outliers hide the true solution Experiments are created with MATLAB a development package. Section two details the background and motivation of the clustering techniques. Section three presents the categories of clustering techniques. Section four deals with the performance of varieties of clustering with k-means that and also the paper area unit ended in section five.

II. LITERATURE REVIEW

A replacement algorithmic rule capable of partitioning a group of documents or alternative samples supported an embedding during a high dimensional metric space [1].Document clustering has not been generally welcomed as scholarly degree information recovery device. Objections to its use comprise a pair of main categories: initial, that Clustering is solely too slow for big corpora (with amount typically quadratic at intervals the range of documents); and second, that clustering does now not notably improve retrieval. Usually they have a tendency to argue that these issues arise only Clustering is employed in a trial to boost typical search techniques.

In any case, seeing clustering as partner data get to apparatus in its case blocks these protests, and gives an examination new access worldview [2]. A loosely applicable algorithmic rule for computing most probability estimates from incomplete information is given at numerous levels of generality. Theory displaying the monotone behaviour of the possibility and convergence of the algorithmic rule comes [3].

In this analysis consists of each a entire experimental assessment concerning fifteen certainly one-of-a-kind datasets, furthermore as associate diploma assessment of the traits of the numerous criterion talents and their end end result at the clusters they produce [15].

III. CLUSTERING TECHNIQUES

A. *Hierarchical Clustering:*

Stratified Clustering builds a cluster hierarchy or a tree of clusters conjointly called a 'dendogram', associate degree inverted tree that describes the order during which points are incorporate or clusters are split. Every cluster node contains kid clusters and relative clusters partitions the purpose coated by their common oldsters. Such approach permits exploring information on completely different levels of roughness. during this each item is assigned to a cluster specified 'N' things will have 'N' clusters that finds and to combine of clusters and merge them into one cluster and work out distance between a replacement cluster and every of recent clusters. Repeat these steps till all things are clustered into K no. of clusters.

B. *Agglomerative:*

It starts with the factors as private clusters and, at every step, close by pair of the clusters is encompass. It begins with the elements as private clusters and, at every step, nearby pair of the clusters is consist of. This needs technique the perception of cluster distance. Collective stratified Clustering could be a bottom-up clustering technique wherever clusters have sub-clusters, that successively have sub-clusters, etc. It unearths the 2 clusters that are closest to each opportunity and type one.

C. *Divisive:*

It starts with one, panoptic cluster and, at every step; A cluster is split into a cohesion clusters of person factors. During this case, at every step that clusters is to be spited is set. Dissentious technique could be a top-down Clustering technique and is a smaller amount ordinarily used. It works during a similar thanks to collective Clustering however within the Opposite direction [3].

D. *Partitioning Clustering:*

Partitioning strategies are classified into 2 subcategories viz. center of mass and medoid algorithms. Focus of mass calculations speaks to each cluster by abuse the gravity focal point of the occurrences. The Mediod algorithmic principle speaks to each group by implies that of the events nearest to gravity focus Partitioning Clustering algorithms try and domestically improve an exact criterion. They education consultation the values of the similarity or distance, they order the outcomes, and select the simplest that optimizes the criterion. Hence, most people of them is probably idea-about as greedy-like algorithms.

E. *K-means:*

It's partitioning technique that discovers shared select cluster of circular structure and creates a specific assortment of disjoint, level (non-hierarchical) groups. Statistical technique may be wont to cluster and consequently the assign rank values to the cluster specific data.

Here particular information is reborn into numeric with the aid of the usage of project rank really worth. K-Means algorithmic rule makes objects into k-sections wherever each bundle addresses a gathering. The algorithmic rule starts off advanced out with initial set of method that and classifies instances supported their distances to their facilities. Next, it computes the cluster approach that another time mistreatment the instances which may be assigned to the clusters; then, supported the modern day set of approach that categorization of all cases is completed. This step continues repeating till cluster manner don't trade among successive steps. Finally, the way that of cluster is calculated once more and therefore the instances are assigned to their everlasting clusters.

F. *K-mediods:*

The target of K-medoid Clustering is to seek out a non-covering set of clusters specified every cluster encompasses a most representative purpose. during this algorithmic rule, instead of scheming the mean of the things in every cluster, a representative item or medoid is chosen for every cluster at each iteration. Medoids for each group are determined by discovering object 'i' inside the cluster that limits: $\sum d(i,j)$, where, 'Ci' is the cluster containing object 'i' and 'd(i, j)' is the separation between objects 'i' and 'j'. Medoids for each group are determined by discovering object 'i' inside the cluster.

The target of K-medoid Clustering is to are seeking for out a non-overlapping set of clusters distinct each cluster contains a maximum representative purpose. Throughout this algorithmic rule, in preference to scheming the mean of the matters in every cluster, a consultant item or medoid is chosen for each cluster at every new launch. Medoids for every cluster are calculated through finding object 'I' in the cluster that minimizes: $\sum d(i,j)$, in which, 'Ci' is the cluster containing object 'i' and 'd(i, j)' is the distance among items 'i' and 'j'. Medoids for every cluster are calculated via finding item 'i' within the cluster.

G. *Grid-based Clustering:*

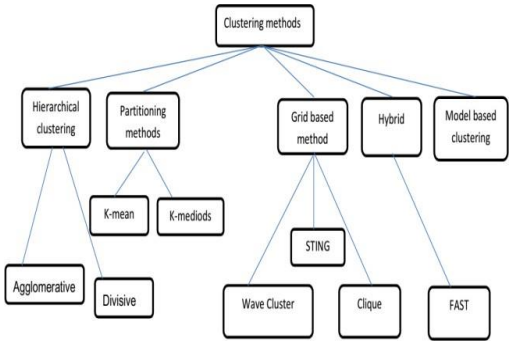
These algorithms in the main focuses on abstraction information, i.e., data that model the geometric structure of items in house, their connections, properties and activities The most objective of this algorithmic rule is to quantize the information set into variety of cells so work with objects happiness to those cells. They build many stratified levels of teams of objects rather than relocating points. During this sense, they're nearer to stratified algorithms however the merging of grids, and consequently clusters, doesn't rely upon a distance live however it's determined by a predefined parameter. A dense cell contains quite sure numbers of points are connected to make the clusters. Primarily grid-based Clustering algorithms are classified into: applied mathematics info Grid-based technique (STING), Wave Cluster, and Clustering In QUEst(CLIQUE) [6].

H. Wave Cluster:

A multi-resolution Clustering algorithmic rule, wave cluster is employed to seek out clusters for terribly massive abstraction databases. The goal of the algorithmic rule is to discover clusters from the given set of abstraction objects. The information is summarized by imposing a multi dimensional grid structure on to the data house [3]. Remodeling the first feature by applying riffle rework so finding the dense regions within the new house is that the main plan. The signal process technique that decomposes a proof into completely different frequency sub bands is understood as a riffle rework. The primary step of the riffle

Cluster algorithmic rule is to quantize the feature house. Within the second step, distinct riffle rework is applied on the measure feature house and therefore new units are generated. Wave Cluster works with an oversized variety of numerical attributes [8].

Figure1. Diverse Clustering procedures



I. STastical info Grid:

STING (STastical info Grid) could be a grid-based multi-resolution Clustering technique during which the embedded abstraction areas of input object are divided into rectangular cells. To applied mathematics info concerning the attributes in every grid cell, like the mean, maximum, and minimum values are hold on as applied mathematics parameters in these rectangular cells. Applied mathematics parameters of upper level cells will simply be computed from the parameters of lower level cells. The standard of STING Clustering depends on the roughness of rock bottom level of grid structure because it uses a multi goals way to cluster analysis. Moreover, for construction of a parent cell STING doesn't consider the reflection relationship between the youngsters and their neighbouring cells. Accordingly, the states of the resulting groups are isothetic for example all the bunch limits are either level or vertical, and 'np' corner to corner limit is recognized. Dense clusters will be known about mistreatment count and cell size info. STING divides the abstraction space into many levels of rectangular cells and immaterial cells are removed.

J. Clique:

A Clustering algorithmic rule that discovers high-thickness locales by partitioning the information house into cells (hyper-rectangles) and finding the thick cells is circle (CLustering In QUEst) clustering. Groups are found by taking the association of all nearby and high-thickness cells. Clusters are represented by communicating the group as a DNF (Disjunctive traditional form) expression so simplifying the expression for simplicity and easy use. The circle is predicated on the subsequent straightforward property of clusters: Since a cluster represents a dense region in some mathematical space of the feature space, there'll be dense areas similar to the cluster all told lower dimensional subspaces. The circle generates the doable set of k-dimensional cells which may probably be dense by viewing dense (k-1) dimensional cells, since every k-dimensional cell should be related to a group of k dense (k-1) dimensional cells.

K. Model-based Clustering:

These algorithms realize smart estimations of model parameters that best work the information. They will be either divided or stratified, counting on the structure or model they anticipate regarding the information set and therefore the manner they refine this model to spot partitioning. They're nearer to density-based algorithms; therein they grow explicit clusters so the perfect model is improved. However, it typically starts with a hard and fast variety of clusters and that they don't use an equivalent construct of density. Since, the gaps in input space from alternative units are large; the outlier will be simply detected in model-based Clustering.

IV. DIFFERENT DOCUMENT CLUSTERING ALGORITHMS

A. Bisection K Means

The bisection K-means algorithmic rule will manufacture either an un-nested (flat) Clustering or a stratified clustering. For un-nested clusters we'll usually refine• the clusters mistreatment the essential K-means algorithms; however we have a tendency to don't refine the nested clusters. We'll offer additional details later. To be precise, the bisecting K-means algorithmic rule could be a dissentious stratified Clustering algorithm, but, to avoid confusion, once we speak of stratified Clustering algorithms we shall mean collective hierarchical algorithms of the kind historically wont to cluster documents. Finally, note that bisection K-means encompasses a time complexness that is linear within the variety of documents. If the amount of clusters is massive and if refinement isn't used, then bisecting K means that is even additional economical than the regular K-means algorithmic rule (In this case, there's no ought to compare every purpose to each cluster centre of mass since to cut a cluster we tend to just consider the focuses inside the cluster and their separations to 2 centroids.)

The bisection K-way algorithmic rule will manufacture either an un-nested (flat) Clustering or a stratified clustering. For un-nested clusters we can normally refine the clusters mistreatment the important K-approach algorithms; however we will be predisposed to do not refine the nested clusters. We'll provide greater info later.

To be particular, the bisecting K-manner algorithmic rule will be a dissentious stratified Clustering set of regulations, however, to keep away from confusion, as soon as we communicate of stratified Clustering algorithms we shall mean collective hierarchical algorithms of the type historically wont to cluster documents. Finally, phrase that bisection K-way includes a time complexness this is linear inside the kind of files. If the amount of clusters is huge and if refinement isn't used, then bisecting K way that is even more low cost than the normal K-approach algorithmic rule (In this case, there's no should examine each purpose to each cluster centre of mass considering the truth that to cut a cluster we have a tendency to really don't forget the points in the cluster and their distances to 2 centroids.)

B. Spherical Gaussian EM algorithm (sGEM)

It's conceivable to refine the distributing results by reallocating new cluster investment. The essential plan of the reallocation technique (Rasmussen, 1992) is to start out from some initial partitioning of the information set, so proceed by moving objects from one cluster {to associate degree other| to a different} cluster to get an improved partitioning. Thus, any repetitive optimization-clustering algorithmic rule will be applied to try to such operation. The matter is developed as a finite mixture model, and applies a variant of the EM algorithmic rule for learning the model. The foremost crucial drawback is a way to estimate the model parameters. The information samples are assumed to be drawn from the variable traditional density and it's more conjointly assumed that the options are statistically freelance and a element generates its members from the spherical mathematician with an equivalent covariance matrix.

C. Linear Partitioning and Reallocation using EM Algorithm (LPR)

The matter of clustering a dataset into teams will be obtained by scheming the mean of the information set initial so compare every purpose with the mean value. If the purpose value is a smaller amount the mean, it's assigned to the primary cluster. Otherwise, it's assigned to the second cluster. The matter arises whereas considering high dimensional information set. supported the thought of the PDDP (Principal Direction dissentious Partitioning) algorithmic rule, this drawback will be dealt by projected all the information points onto the principal direction the principal eigenvector of the variance matrix of the data set, so the cacophonic method will be performed supported this principal direction.

D. Model Hybrid Scheme for Text Clustering (HSTC) Model

The Model Hybrid Scheme for Text Clustering Model exploits content-based procedure for efficient Clustering. The algorithmic standard performs Clustering on a dataset D containing 'n' archives diagrammatic as D={Dj; j = 1 .. nD} having a social illicit relationship of terms T = {ti; I = 1 ..nT} got when performing expressions pre-preparing. The pre-handling performs stop-word expulsion and stemming to evacuation intermittent and irrelevant terms.

A content-based distance live is employed as similarity measure the inclusion of activity characteristics includes document structure and magnificence info into similarity analysis, therefore on improve the general Clustering performance. Whereas scheming the document distance live, a document 'D' is diagrammatic mistreatment 2 vectors, V* and V**. V*(D) represents the content description of D and could be a set of terms wherever every term â€˜tâ€™ is related to its normalized frequency 'tf'.

E. Method Text Clustering with Feature Selection (TCFS) Method

The methodology employed by TCFS method is analogous thereto of HSTC method; however it differs in 3 ways. The primary is within the pre-processing stage, second is that the Clustering method and therefore the third is within the clustering algorithmic rule used. Each uses an equivalent similarity live, cosine distance. Inside the pre-preparing stage, beside stop word end and stemming, weight estimation plays out, that figures the term weight and phonetics weight are encased.Term weight is calculable mistreatment TF/IDF values that utilize info regarding term and variety of times (n) it seems within the document. Mistreatment the term weight worth a term cube is made.

A term cube could be a 3D model representing the document, term and n relationship. The linguistics weight is calculated by construct extraction, construct or linguistics weight calculation and construction of semantic cube. The construct extraction module is intended to spot concept in every document. This method is completed with the assistance of the metaphysics assortment.

The terms are matched with ideas, synonyms, meronyms and hypernyms within the metaphysics. The construct weight is calculable with the concept and its component count. The linguistics cube is made with ideas, linguistics weight and document. In cluster process that teams the documents, 2 techniques, namely, term Clustering and linguistics clustering technique are used. Term Clustering teams documents in keeping with the term weight, whereas linguistics Clustering teams documents in keeping with the semantic weight. For Clustering, a classical k-means algorithmic rule is employed.

V. RESULTS AND DISCUSSIONS

This section presents the results supported the comparative study on special characteristics of the varied document cluster technique. Conjointly associate degree experiment is administered during a real worth to discover the document cluster at the side of the prevailing acquainted strategies. Every one of the examinations were led utilizing an Intel(r) Core i5 machine with 8GB RAM. 3 performance metrics, namely, purity of a cluster, F-measure and C.P.U. execution time were used. The general purity obtained by the chosen seven algorithms for various varieties of clusters is shown in below table.

A. *DATA SET:*

Information Set: all told of the information sets, we've removed stop words, i.e., common words like “a”, “are”, “do”, and “for”. We’ve conjointly performed stemming mistreatment Porter's suffix-stripping algorithmic rule. we have a tendency to collected documents that have relevancy judgments so designated documents that have simply one relevance judgment. the category labels of la1, and la2 were generated in keeping with the section names of articles, like “Document”, “Clustering”, “Techniques”, “Methods”, “Means”, and “Returns. Information sets re0 and re1 are from Reuters-20141 text categorization check assortment Distribution one.0 [reut].

Table I: Outline Depiction of Report Sets

Information Set	Source	Number of Documents	Number of Classes	Minimum Class Size	Maximum Class Size	Average Class Size	Number of Words
Re0	Reuters-21578	1504	13	11	608	115.7	11465
Re1	Reuters-21578	1657	25	10	371	663	3758
WAP	WebAce	1560	20	5	341	78	8460
Tr35	TREC	927	7	2	352	132.4	10128
Tr45	TREC	690	10	14	160	69	8261
Fbis	TREC	2463	17	38	506	144.9	2000
La1	TREC	3204	6	273	943	534	31472
La2	TREC	3075	6	248	905	512.5	31472

Table II: Virtue of Cluster

No.of Clusters	Regular K means	Bisecting K Means	Bisecting K-means with refinement	K* Means	EM	Traditional E/M	sGEM	LPR	HSTC	TCFS	PAM	CLARA	CLARANS	IST	CST	UPGM Hierarchical A refin (UPGMA)	Hierarchical	AHC
15	1.3839	1.3305	1.1811	1.1425	1.9838	1.5601	1.5875	1.4488	1.6587	1.1122	1.3652	1.1147	1.2111	1.2231	1.2211	1.1111	1.4811	1.3344
30	1.6896	1.6315	1.7111	1.6654	1.0058	1.3198	1.4254	1.3954	1.4587	1.4422	1.3344	1.2243	1.1442	1.4001	1.4104	1.3958	1.7361	1.2584
60	1.5557	1.5494	1.5601	1.4589	1.3584	1.271	1.4422	1.4128	1.3945	1.3587	1.5874	1.2587	1.3564	1.3874	1.2874	1.1874	1.8028	1.3587
100	0.5228	0.4713	0.4722	0.4112	0.4107	0.4665	0.4258	0.4125	0.3987	0.4001	0.3856	0.41	0.4018	0.4258	0.3989	0.4101	0.5711	0.3987
150	1.3587	1.5874	1.4422	1.3344	1.1955	1.1254	1.2243	1.1442	1.4001	1.4104	1.3958	1.3874	1.2874	1.1874	1.1122	1.3652	1266 5	1.5874
200	1.3198	1.3708	1.4053	1.3958	1.8594	1.3985	1.3344	1.2243	1.1442	1.4001	1.4104	1.3958	1.3945	1.3587	1.5874	1.2587	1.3832	1.3708
250	0.071	0.657	0.9511	0.9121	0.4046	0.4225	0.2584	0.0022	0.2058	0.11	0.301	0.3344	0.2243	0.1442	0.4001	0.4104	0.3958	0.2243
500	0.9673	0.9799	0.9445	0.9222	1.5955	1.6101	1.3587	1.5874	1.4422	1.4589	1.3584	1.271	1.4422	1.4128	1.3945	1.2587	1.1392	1.3344

From Table II, it’s clear that the all the algorithm rule quality clusters followed by K means that compared with alternative algorithms. The performance of all the k means primarily based algorithms with relevancy purity of clusters may be a k-means variant, whereas while showing smart purity still is low whereas examination with alternative algorithms. k-means variant, while showing good purity still is low while comparing with other algorithms.

While thinking about the exactness of the calculation as far as accuracy, review and F measure, the pattern got is like that of group immaculateness. The greatest exactness is achieved by TCFS calculation, while K-means shows the least grouping precision.

While considering the time taken or speed of clustering, it was discovered that the conventional K-Means calculation was the quickest, trailed by HSTC and TCFS. The EM variant models were slow when compared with all the other algorithms very similar.

Table 3: Accuracy of the Algorithmic rule

Accuracy Type	Regular K means	Bisecting K Means	Bisecting K-means with refinement	K* Means	EM	Traditional EM	sGEM	LPR	HSTC	TCFS	PAM	CLARA	CLARANS	IST	CST	UP GM Hierarchical A refin (UP emen GM Hierarchical t A)	AHC
Precision	1.3839	1.3305	1.1811	1.5601	1.9838	1.3587	1.5874	1.2587	1.3564	1.3874	1.2874	1.1874	1.8028	1.5601	1.4589	1.3584	1.4811	1.6587
Recall	1.6896	1.6315	1.7111	1.3198	2.0058	1.3874	1.2874	1.1874	1.8028	1.3587	1.4589	1.3584	1.271	1.4422	1.4128	1.3344	1.7361	1.4587
F Measure	1.8557	1.5494	1.5601	1.4589	1.3584	1.271	1.4422	1.4128	1.3958	1.3945	1.3874	1.2874	1.1874	1.8028	1.3587	1.3958	1.8028	1.7911

VI. CONCLUSION

Increase among the amount of text info hold on in digital kind has inflated the need for automated tool that facilitate people understand degree manage these data in a cheap manner. Text clustering, that's that the tactic of grouping documents having similar properties supported linguistics and mathematics content, may be a very important component in many data organization and management tasks. among the gift analysis work seven approaches to document Clustering was thought-about and their ways and performance were analyzed. The chosen ways square measure K- implies that, K*-Means, EM algorithm, sGEM, LAR, HSTC and TCFS and different algorithms. the whole algorithms initial belong to the K-means family, the second algorithms belong to the EM family, whereas the last take into consideration linguistics and philosophy of the text. Experiments proven that everyone the techniques square measure economical in document clustering methodology, but the performance of other algorithm is slightly higher in terms clump quality and so the traditional K-Means algorithmic rule is improbably fast in producing cluster result . In future, work that fuses the K-means variants and EM variants algorithms square measure to be performed to combine the advantage of quality clustering throughout a fast manner.

REFERENCES

1. Balabantaray , Rakesh Chandra, Chandrali Sarma, and Monica Jha , " Document clump victimisation K-Means and K-Medoids." arXiv preprint arXiv:1502.07938,2015.
2. Boley, D., "Principal direction dissentious partitioning. data processing and data Discovery",1998, 2(4), 325–344.
3. BOTTOU, L. and BENGIO, Y, Convergence properties of the K-means algorithms. In Tesauro, G. and Touretzky, D. (Eds.) Advances in Neural information science Systems seven,1995,585-592, The Massachusetts Institute of Technology Press, Cambridge, MA.
4. Duda, R. O., Hart, P. E., & Stork, D. G., "Pattern classification". New York: Wiley,2001.
5. Fisher, D., "unvaried optimisation and simplification of hierarchical clustering's". Journal of computing analysis,1996, 4, 147–180.
6. Jain, A. K., & Dubes, R. C., "Algorithms for clump knowledge". Englewood Cliffs, NJ: Prentice-Hall.1998.
7. Karypis, G., C luto: "A clump toolkit. Technical report 02-017", Department of engineering, University of North Star State. out there at http://www.cs.umn.edu/~cluto,2002.
8. King, B., "Step-wise clump procedures. Journal of the yank applied math Association", 1967, 69, 86–101.
9. MacQueen, J. (1967). "Some strategies for classification and analysis of variable observations". In Proceedings of the fifth conference on mathematical statistics and chance , Berkeley, CA: University of Calif. Press,1967, pp. 281–297
10. K.Popat et al, "Review and Comparative Study of clump Techniques". (IJCSIT) International Journal of engineering and data Technologies, Vol. 5 (1) ,2014, 805-812.
11. U.S. Patki, Dr. P.G. Khot, "A Literature Review on Text Document clump Algorithms employed in Text Mining", Journal of Engineering Computers & Applied Sciences (JECAS) ,2017, 6(10), 1552-1564.
12. Twinkle Svadas, Jasmin Jha, "Document Cluster Mining on Text Documents", International Journal of engineering and Mobile Computing, ISSN 2320–088X, Vol.4 Issue.6, 2015,pg. 778-782.
13. TREC: Text REtrieval Conference. http://trec.nist.gov.
14. TREC: "Text REtrieval Conference connexion judgments". http://trec.nist.gov/data/qrels_eng/index.html.
15. Worarat Krathu, Praisan Padungweang, and Chakarida Nukoolkit , "Data processing Approach for Automatic Discovering Success Factors Relationship Statements fully Text Articles", proceedings of the eighth International Conference on Advanced procedure Intelligence Chiang Mai, Thailand, 2016, 14-16.
16. Yogapreethi.N,Maheswari.S, "A Review On Text Mining In data processing", International journal on soft computing (IJSC), 2016, 7(3), 145-160.
17. Zhang, S. and Pan, X., "A completely unique text classification supported Mahalanobis distance", third International Conference on pc analysis and Development (ICCRD), 2011,Pp. 156-158.

AUTHORS PROFILE

V.Kumaresan, Assistant Professor, Annamalai University.. He is pursuing his Ph.D., in Annamalai University in the field of Computer Application. He has a tutorial expertise of sixteen years. He has published Technical papers in International, National Journals and conferences His space of analysis interest includes data processing.

Dr.R.Nagarajan, is working as Assistant Professor in Annamalai University. He has revealed twelve analysis paper in International Journals. He has twenty one years of expertise in Application Programming. He's concerned in analysis activities for the past twelve years. His space of specialization includes data processing, Document clustering and Cloud Computing.