ISSN 2348 – 8034
Impact Factor- 3.155

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

KNOWLEDGE DISCOVERY TECHNIQUES THROUGH TOPIC MODEL IN CITATION INFLUENCE

J Jayapriya[*1] and M John Basha M.E (Ph.D)[2]

[*1]PG Student, Department of Computer Science and Engineering, PTR College of Engineering and Technology, Tamilnadu, India

[2]Head of the Department,Department of Computer Science and Engineering, PTR College of Engineering and Technology, Tamilnadu, India

ABSTRACT

Information retrieval is one of the major topics among the researchers regarding data mining. This depends on accurate analysis of data and minimum time consumption as the main motto. In this paper we propose an index based cluster validity and cluster algorithm along with centroid ratio. In which the mechanism of centroid ratio is meant by the comparing the results of two clustering's such as clustering algorithm and cluster validity. To make this mechanism prominent, gained centroid ratio is implemented in prototype-based clustering known as Pair wise Random Swap clustering algorithm. It effectively overcomes the k- means local optimum problem by enabling swapping between simple perturbations to the solution and makes intersection on the nearest optimum solutions. The centroid ratio is calculated by means of relationship between the clustering and calculating mean square error (MSE) along with other external indices. The implementation proves the proposed system is simple and accurate for calculation when compared with other traditional approaches like Random Swap, Deterministic Random Swap, repeated k-means or k-means++. This prototype is implemented successfully on document clustering and result accuracy proves its efficiency than the other mining methods.

Keywords- *Data clustering, clustering evaluation, k-means, and random /deterministic swap.*

I. INTRODUCTION

Data mining is the widely discussing topic in information management system as information plays a vital role in most of the application. Initially data mining is the concept of extracting required information from a vast database. There are various methodologies were introduced and researched in order to improve the performance in data mining. The accurate analyzing with minimum computational cost should require that makes mining approach an efficient one. Because every search is based on the query by means of keywords the process begins. To make this in a better way a concept of clustering is evolved which makes connection between the keywords as a result it shows faster analysis in data extraction. The clustering can be classified in two form similar objects and dissimilar objects. Some of the problems are still facing today in clustering are, the most of the traditional methodology are in adequate in addressing the requirements. Next thing is dealing with large database and number of dimensions results in poor performance and time consumption. The current techniques failed to address the data duplication and distance measuring in a document. In traditional data mining technology there are various clustering algorithm are discussed such as exclusive, overlapping, hierarchical and probabilistic clustering each one is suffers from serious disadvantages.

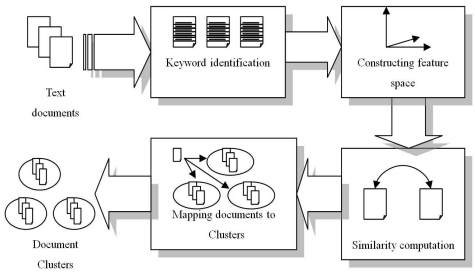


Fig 1: A sample document clustering process

Among this document clustering is most familiar one and it wide application grabs the researcher's attention towards it. The fig 1 shows how generally the data clustering works in which the input data is processed as

sample text documents, the keywords are indentified based on the queries which is paused on the searching approach based on the similarity between them and relationship among them results in clustering. The goal of clustering is attained during clustering evaluation it happens only by achieving the high intra-cluster similarity among the documents within a cluster which are similar and low inter-cluster similarity in the documents which are dissimilar. More than this a better method to evaluation is time taken for completing the overall performance.

II. RELATED WORKS

On discussing about clustering prototype based methodology are most familiar due to its best fits schema even in an unknown structure. Based on this a well known tradition technique in data mining is k-means [1] which represents by single prototype. It is widely applicable for data grouping in most of the real time applications. It shows some effective results on the aspect of computation and memory consumptions but those are highly sensitive to initializations. This initialization problem is analyzed in repeated k-means (RKM) [2] by running it in multiple times with a randomly chosen of parameters. But it is not effectively applicable for the real time applications.

On this continuation Swap-based clustering algorithm [3] is evolved it is based on finding the optimal centroids according to the k-means convergence property. The swap strategy in way of handling pair of centroid by means one is inserted and other to be removed in order to achieve a better solution. The swap is only made when a best prototype is found during the process and it obtains some good quality results independently. In the research, next some stochastic global optimization methods are developed such as simulated annealing [4] and genetic algorithms [5].

But these techniques are not achieved effective response as it seriously had the drawback of time consumptions. The k- means is improved as global k-means algorithm (GKM) [6] that adds one cluster center at a period by means of global search procedure. According to k++ [7] means it chooses the initial values and also increases the speed as well as accuracy rather than the k-means. Some works [8] shows that data characteristics affect the clustering as those data are highly influenced in dimension, sparse, noise, outliers, types of attributes, data sets, and scales of attributes. As per conventional k-means [9] it uses Euclidean distance in calculating distance between data points and its dimensions. But these traditional indexing and algorithmic techniques are failed to achieve the accuracy.

Another popular method for clustering is spherical k-means [10] and it is widely accepted for document clustering. As the clustering validity is based on two factors such as external [11] and internal validity [12] by means of Rand index and the Jaccard coefficient however it shows high accuracy but time complexity also. Some works are taken in improving the clustering results based on standard k-means [13] it shows simple to implement but not so effective. In order improve the accuracy some works [14] [15] [16] handles the linear dependency on the number of data vectors as well as the clusters. It also applies agglomerative clustering by merging two existing clusters but these swapping methods requires the location to be relocate but it will not work properly on smaller size or variance and not suitable for real time practices.

In the papers [17] & [18] they have discussed about the finding the distance metric for clustering by means of MSE (Mean square error) values. Some research works like BIRCH sets [19], R15 [20], Data Aggregation (A7) [21] and fine-tuning [22] where discussed about clusters in regular grid structures by generating 2-D Gaussian distributions and data aggregation based on the distinct groups of points. But here the drawback is determining the number of clusters is outside the range still in to be in development stage.

III. EXISTING SYSTEM

The existing system is based on Bernoulli process topic (BPT) model which implies citation on document clustering. There is a huge repository is available by means of digital database with the help of internet and other mass media. As most of the documentations are connected by means of citation but the corpus plays two different roles in a document which leads to serious disadvantage in scientific articles in differentiating the two roles. This limitation is effectively handled by BPT it captured the citation network and distributing the parameters by means of variation approximation approach. Generally BPT is a corpus probabilistic model with the availability of citation information's from those documents. By using variation inference approach it distributes the documents over topic from the mixture of the distributions associated with the related documents. Earlier to this it was discussed by LDA model [23] but it is effective in capturing the essential properties of citation. Below fig2 shows the topic distribution of document and citation level.

ISSN 2348 – 8034
Impact Factor- 3.155

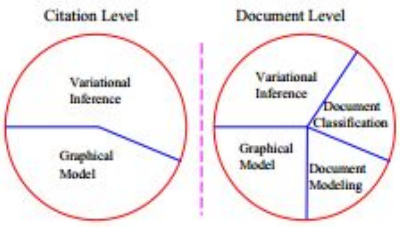


Fig 2: Topic distribution between document level and Citation level

Bernoulli process is a discrete-time stochastic process which takes only 0 and 1 by a sequence of independent identically distributed Bernoulli trials. It is a finite or infinite sequence of independent random variables X_1, X_2, X_3,

From which each i, the value of x_i is either 0 or 1;

For all values of i, the probability that $x_i = 1$ is the same number p.

However, the BPT is effective for document clustering on scientific articles it has some major disadvantages like memory space and time complexities are still to be improved. It does not have universal function for all clustering problems. It is not widely appreciated due to its greater computational complexities. These are not effectively applicable for high dimensional data's.

IV. PROPOSED SYSTEM

To provide a prominent solution for knowledge management system in data mining, in this paper we propose a Pair wise Random Swap Clustering Algorithm in order to overcome the drawbacks of existing systems and other problems faced by the traditional methodology. The proposed system has a cluster-level validity criterion called centroid ratio that has minimum time consumption in finding the unstable or incorrectly located centroids. The centroid ratio is mainly evolved to overcome the local optimum problem faced in k-means algorithm. It computes two set of centroids by which the pair wise ratio is calculated using swap algorithm. In this mechanism the centroid indices is evolved which are based on three factors known as point level partitions, data set, and centroids. By using these factors Mean square error (MSE) is calculated for evaluating the clustering. Here the swapping is done by randomly selected two locations using fine tuned approach to enable a nearest pairing. The nearest pairing of two centroid set is stated by graph-theoretic analysis which is the minimum matching of a given bipartite graph. In this the nodes correspond to the centroids are connected according to the edges from different clustering, and edge cost stands for the centroid distance.

```
SmartSwap(X,M) → C,P
C ← InitializeCentroids(X);
P ←PartitionDataset(X, C);
Maxorder ← log2M;
order ← 1;
WHILE order < Maxorder
   ci, cj ←FindNearestPair(C);
   S ← SortClustersByDistortion(P, C);
   cswap ←RandomSelect(ci, cj);
   clocation ←sorder;
   Cnew ← Swap(cswap, clocation);
   Pnew ← LocalRepartition(P, Cnew);
   KmeansIteration(Pnew, Cnew);
IF f(Cnew) < f(C), THEN
   order ← 1;
   C ←Cnew ;
ELSE
   order ← order + 1;
   KmeansIteration(P, C);
```

Fig 3: A pseudo code for pair wise swapping algorithm

ISSN 2348 – 8034
Impact Factor- 3.155

V. IMPLEMENTATION DESIGN

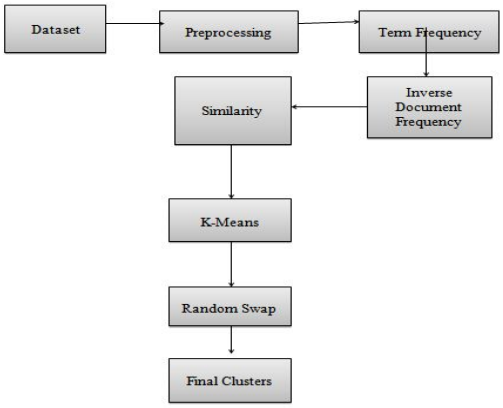


Fig 4: Implementation of proposed system

As discussed the dataset is processed as per shown in the fig 4, in which initially the dataset undergoes preprocessing. In this stage it undergoes stop word and stemming;

Stop word removal: Stop words are the little words which hold small information or conjunction content in a document. By removing these stop words the index size is made smaller without affecting the accuracy of the searching query.

Stemming removal: It is the process of removing the various grammatical forms / word forms of a word like its noun, adjective, verb, adverb etc. By doing this is the word comes to its root form. Generally it is done by removing any attached suffixes and prefixes from index terms. It makes possible of variants of words have similar semantic interpretations in a document.

The preprocessed data is moved for frequency which is calculated by counting the number of terms in document corpus. The word similarity is calculated by using TF*IDF and IDF which was drawn as a result of Inverse Document Frequency. After that the clustering is done by applying the k-means algorithm using vector quantization and the cluster is formed by the observation belongs to k-means which has nearest mean along with the similarity distance threshold value. The clustering is done among the nodes by following below mentioned pseudo code;

- Make initial guesses for the means $\mathbf{m}_1$, $\mathbf{m}_2$, ..., $\mathbf{m}_k$
- Until there are no changes in any mean
  - Use the estimated means to classify the samples into clusters
  - For i from 1 to k
    - Replace $\mathbf{m}_i$ with the mean of all of the samples for cluster i
  - end_for
- end_until

The pair wise swapping is done by calculating MSE mean square error value by enabling cluster centroid randomly. Then the centroid ratio is analyzed between the two clusters by means of index value. This swapping process is continued still achieving the best fit centroids in the process.

Select a proper centroid for removal:

There are M clusters in total: $p_{removal}=1/M$.

Select a proper new location:

ISSN 2348 – 8034
Impact Factor- 3.155

There are N choices: $p_{add}=1/N$

Only M is significantly different: $p_{add}=1/M$

In total:

M^2 significantly different swaps

Probability of each different swap is $p_{swap}=1/M^2$

The efficiency of pair wise random swap is calculated by applying;

Total time to find correct clustering:

Time per iteration * Number of iterations

VI. RESULT & DISCUSSION

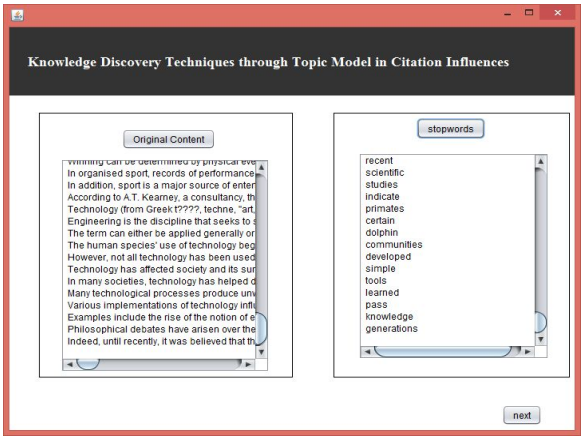


Fig 5: Stop word removal

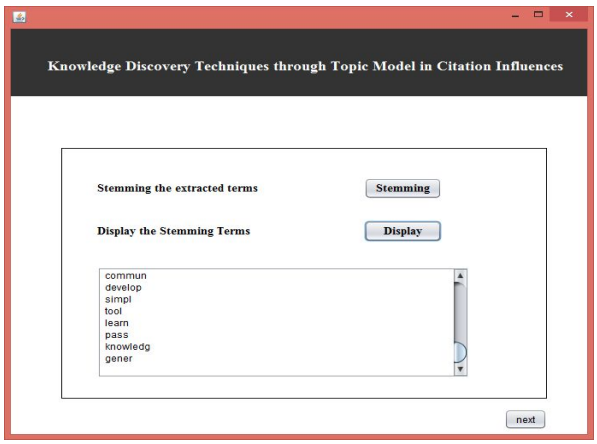


Fig 6: Stemming removal

ISSN 2348 – 8034
Impact Factor- 3.155

The data set according to the real time application is selected which is passed into framework in which the preprocessing is done, the fig 5 & fig 6 shows the result of preprocessing as stop words and stemming are removed from the document which selected as input data.

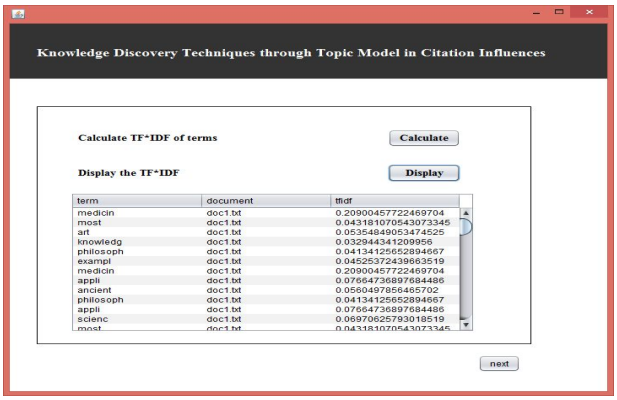


Fig 7: Calculate TF*IDF

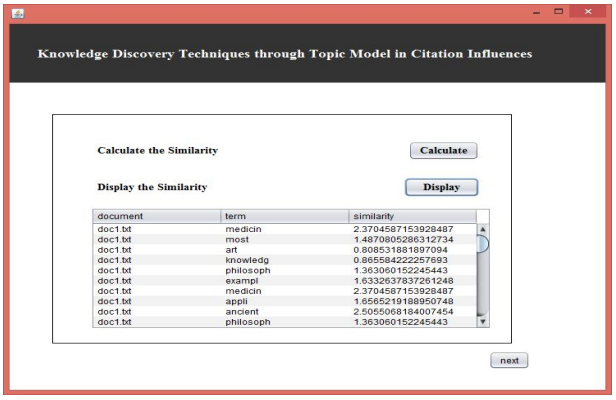


Fig 8: Calculating the similarity between TF*IDF and IDF

The fig 7 & fig 8 shows the result of term frequency and inverse document frequency, it is the statistical weight which is calculated to analyze the importance of a term in a text document collection by defining the number of documents in which a term appears.

It is calculated by applying;

$$\mathrm{tf}(t,d) = 0.5 + \frac{0.5 \times \mathrm{f}(t,d)}{\max\{\mathrm{f}(w,d) : w \in d\}}$$

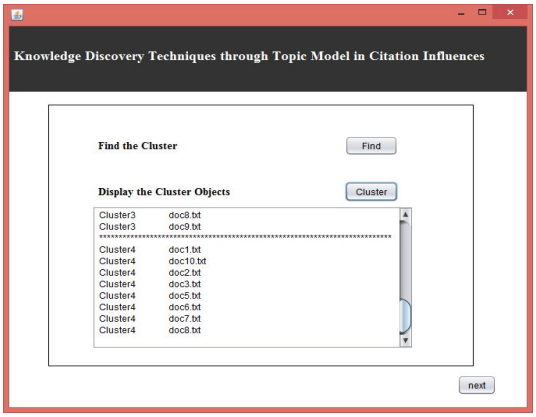
The frequency tf(t,d) = 1 if t occurs in d otherwise 0

For Calculating IDF;

IDF=log2NDF

Where N is the number of documents in the collection, and DF is the document frequency of the term, i.e., the number of documents in which the term appears.

ISSN 2348 – 8034
Impact Factor- 3.155



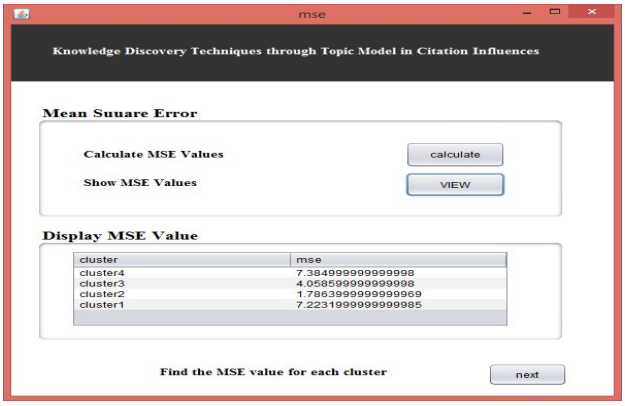


Fig 10: Result of MSE

On fig 9 it the result obtained after implementing the k means algorithm and fig 10 shows the calculation result of MSE (Mean square error) which calculate variance in the document.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2.$$

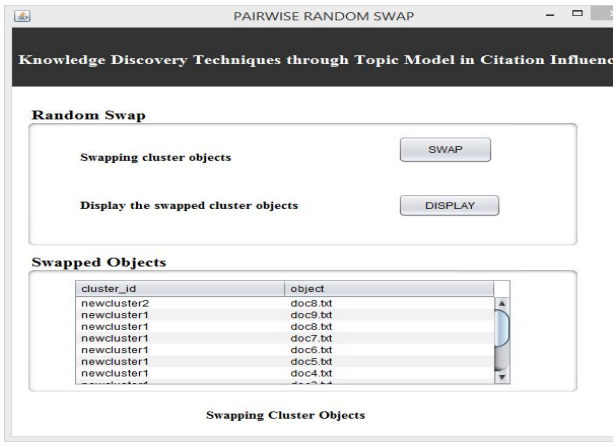
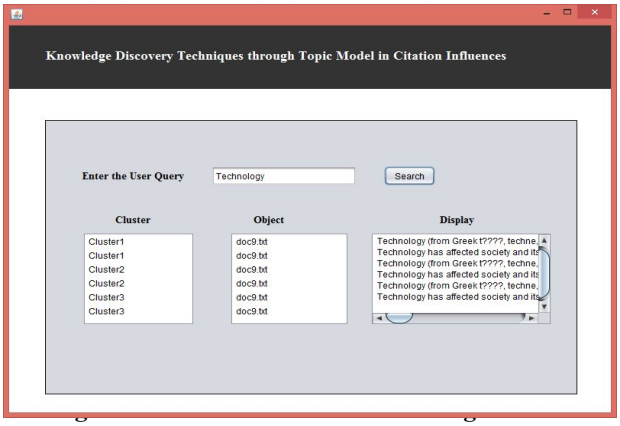


Fig 11: Swapping the cluster objects

ISSN 2348 – 8034
Impact Factor- 3.155



In fig 11 & fig 12 the result shows the implementation of pair wise random swap algorithm thus pairing the nearest cluster which was analyzed by the above the result. The final result shows the document clustering by which the showing the appropriate result according to the user query which is given. Thus proves the successful implementation of our proposed system according to the real time applications.

VII. CONCLUSION

We propose novel pair wise random swap algorithm to the clusters according to centroid ratio which was best fits. The index value plays an important role in clustering which determined the near pair to cluster. The obtained result proves the accuracy and detection of unstable centroids. The proposed algorithm results are compared to the traditional methods such as Random Swap, Deterministic Random Swap, Repeated k-means, and k-means++ which proves a better result. The proposed work effectively results in important factors like accuracy, memory space and time consumptions according to the real time applications. The best result is applicable for high dimensional data and simple for implementation. The future work is carried out in dealing with high variance in future.

REFERENCE

1. *A. K. Jain, "Data clustering: 50 years beyond K-means," Pattern Recognit. Lett., vol. 31, no. 8, pp. 651–666, 2010.*
2. *A. Jain, M. Murty, and P. Flynn, "Data clustering: A review," ACM CSUR, vol. 31, no. 3, pp. 264–323, 1999. P. Fränti and J. Kivijärvi, "Randomized local search algorithm for the clustering problem," Pattern Anal. Applicat., vol. 3, no. 4, pp. 358–369, 2000.*
3. *G. Babu and M. Murty, "Simulated annealing for selecting optimal initial seeds in the k-means algorithm," Indian J. Pure Appl. Math., vol. 25, no. 1–2, pp. 85–94, 1994.*
4. *K. Krishna and M. Murty, "Genetic k-means algorithm," IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 29, no. 3, pp. 433–439, Jun. 1999.*
5. *A. Likas, N. Vlassis, and J. Verbeek, "The global k-means clustering algorithm," Pattern Recognit., vol. 36, no. 2, pp. 451–461, 2003.*
6. *D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," in Proc. 18th Annu. ACM-SIAM SODA, Philadelphia, PA, USA, 2007, pp. 1027–1035.*
7. *H. Xiong, J. Wu, and J. Chen, "K-means clustering versus validation measures: A data-distribution perspective," IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 39, no. 2, pp. 318–331, Apr. 2009.*
8. *C. Aggarwal, A. Hinneburg, and D. Keim, "On the surprising behavior of distance metrics in high dimensional space," in Proc. 8th ICDT, vol. 1973. London, U.K., 2001, pp. 420–434.*
9. *I. Dhillon, Y. Guan, and J. Kogan, "Iterative clustering of high dimensional text data augmented by local search," in Proc. IEEE ICDM, Washington, DC, USA, 2002, pp. 131–138.*
10. *J. Wu, H. Xiong, and J. Chen, "Adapting the right measures for k-means clustering," in Proc. 15th ACM SIGKDD Int. Conf. KDD, Paris, France, 2009, pp. 877–886.*
11. *Y. Liu, Z. Li, H. Xiong, X. Gao, and J.Wu, "Understanding of internal clustering validation measures," in Proc. 10th ICDM, Sydney, NSW, Australia, 2010, pp. 911–916.*

12. *S. Khan and A. Ahmad, "Cluster center initialization algorithm for k-means clustering," Pattern Recognit. Lett., vol. 25, no. 11, pp. 1293–1302, 2004.*
13. *P. Fränti, O. Virmajoki, and V. Hautamäki, "Probabilistic clustering by random swap algorithm," in Proc. 19th ICPR, Tampa, FL, USA, 2008, pp. 1–4.*
14. *P. Fränti and O. Virmajoki, "Iterative shrinking method for clustering problems," Pattern Recognit., vol. 39, no. 5, pp. 761–775, 2006.*
15. *H. Frigui and R. Krishnapuram, "Clustering by competitive agglomeration," Pattern Recognit., vol. 30, no. 7, pp. 1109–1119, 1997.*
16. *M. Meila, "Comparing clusterings—An information based distance," J. Multivar. Anal., vol. 98, no. 5, pp. 873–895, 2007.*
17. *A. Rakhlin and A. Caponnetto, "Stability of k-means clustering," in Advances in Neural Information Processing System, vol. 19. Cambridge, MA, USA: MIT Press, 2007.*
18. *T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: A new data clustering algorithm and its applications," Data Min. Knowl. Discov., vol. 1, no. 2, pp. 141–182, 1997.*
19. *C. Veenman, M. Reinders, and E. Backer, "A maximum variance cluster algorithm," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 9, pp. 1273–1280, Sept. 2002.*
20. *A. Gionis, H. Mannila, and P. Tsaparas, "Clustering aggregation," ACM Trans. Knowl. Discov. Data, vol. 1 no. 1, pp. 1–30, 2007.*
21. *T. Kaukoranta, P. Fränti, and O. Nevalainen, "A fast exact GLA based on code vector activity detection," IEEE Trans. Image Process., vol. 9, no. 8, pp. 1337–1342, Aug. 2000*
22. *E. Erosheva, S. Fienberg, and J. Lafferty, "Mixed membership models of scientific publications," in Proceedings of the National Academy of Sciences ess, 2004, p. 2004.*