

Contribution-Based Clustering Algorithm for Content-Based Image Retrieval

Harikrishna Narasimhan

Department of Computer Science and Engineering
College of Engineering, Guindy
Anna University Chennai
nhari88@gmail.com

Purushothaman Ramraj

Department of Computer Science and Engineering
College of Engineering, Guindy
Anna University Chennai
rpn2142@gmail.com

Abstract—Clustering is a form of unsupervised classification that aims at grouping data points based on similarity. In this paper, we propose a new partitional clustering algorithm based on the notion of ‘contribution of a data point’. We apply the algorithm to content-based image retrieval and compare its performance with that of the k-means clustering algorithm. Unlike the k-means algorithm, our algorithm optimizes on both intra-cluster and inter-cluster similarity measures. It has three passes and each pass has the same time complexity as an iteration in the k-means algorithm. Our experiments on a bench mark image data set reveal that our algorithm improves on the recall at the cost of precision.

Keywords—Content-based image retrieval (CBIR); clustering; contribution; k-means algorithm; game theory

I. INTRODUCTION

A cluster is a collection of data points that are similar to one another within the same cluster and dissimilar to data points in other clusters [1]. Clustering is a method of unsupervised classification, where data points are grouped into clusters based on their similarity. The goal of a clustering algorithm is to maximize the intra-cluster similarity and minimize the inter-cluster similarity.

Clustering algorithms can be broadly classified into five types: 1. Partitional clustering, 2. Hierarchical clustering, 3. Density-based clustering, 4. Grid-based clustering and 5. Model-based clustering [1]. Partitional and hierarchical clustering are the most widely used forms of clustering. In partition clustering, the set of n data points are partitioned into k non-empty clusters, where $k \leq n$. In the case of hierarchical clustering, the data points are organized into a hierarchical structure, resulting in a binary tree or dendrogram [2].

In this paper, we propose a new clustering algorithm, which would come under the category of partitional clustering algorithms. Two commonly used methods for partitioning data points include the k-means method [3, 4] and the k-medoids method [5]. In the k-means method, each cluster is represented by its centroid or the mean of all data points in the cluster. In the case of the k-medoids method, each cluster is represented by a data point close to the centroid of the cluster. Apart from these methods, there has been lots of work on fuzzy partitioning methods [6] and partition methods for large-scale datasets [7].

We use the notion of ‘contribution of a data point’ for partitional clustering. The resultant algorithm requires only three passes and we show that the time complexity of each pass is same as that of a single iteration of the k-means clustering algorithm. While the k-means algorithm optimizes only on the intra-cluster similarity, our algorithm also optimizes on the inter-cluster similarity.

Clustering has widespread applications in image processing. Color-based clustering techniques have proved useful in image segmentation [13]. The k-means algorithm is quite popular for this purpose. Clustering based on visual content of images is an area that has been extensively researched for several years [14]. This finds application in image retrieval.

Content-based image retrieval (CBIR) aims at finding images of interest from a large image database using the visual content of the images. Traditional approaches to image retrieval were text-based, where individual images had to be annotated with textual descriptions [8]. Since this is a tedious manual task, image retrieval based on visual content is very essential.

Organizing the retrieved search results into clusters is an intuitive form of content representation [14] and facilitates user’s browsing of images [15]. Image clustering can also be used to optimize the performance of a CBIR system [16]. While the performance of a number of clustering algorithms in image retrieval have been analyzed in previous works [17, 18, 19, 20], we apply our proposed algorithm to CBIR and compare its performance with that of the k-means clustering algorithm.

The rest of the paper is organized as follows. In Section II, we explain the notion of contribution of a data point. The proposed algorithm is outlined in Section III and its time complexity is analyzed in Section IV. In Section V, we describe how clustering can be applied to content-based image retrieval and present our experiments and results in section VI. Section VII contains some conclusions along with future directions of research.

II. CONTRIBUTION-BASED CLUSTERING

Partitional clustering aims at partitioning a group of data points into disjoint clusters optimizing a specific criterion [2]. When the number of data points is large, a brute force

enumeration of all possible combinations would be computationally expensive. Instead, heuristic methods are applied to find the optimal partitioning. The most popular criterion function used for partitional clustering is the sum of squared error function given by

$$E = \sum_{i=1}^k \sum_{x \in C_i} (x - m_i)^2$$

where k is the number of clusters, C_i is the i^{th} cluster, x is a data point and m_i is the centroid of the i^{th} cluster.

A widely used squared-error based algorithm is the k-means clustering algorithm [2]. In this paper, we propose a clustering algorithm similar to the k-means algorithm. We define the contribution of a data point belonging to a cluster as the impact that it has on the quality of the cluster. This metric is then used to obtain an optimal set of ' k ' cluster from the given set of data points.

The notion of contribution has its origin in game theory [9]. A recent work by Garg [10] focuses on the merger of game theory and data clustering. Garg mapped cluster formation to coalition formation in cooperative games and used the solution concept of Shapely value to find the optimal number of clusters for a given set of data points. While his work uses the concept of contribution to find the optimal cluster number, we use it in a different manner for optimal partitioning of the data points into a fixed number of clusters.

Given a cluster C_i with n points and centroid m_i , the average intra-cluster dispersion is given by

$$dispersion(C_i) = \frac{1}{n} \sum_{x \in C_i} (x - m_i)^2$$

The contribution of a point $x \in C_i$ is measured as

$$contribution(x, C_i) = dispersion(C_i - \{x\}) - dispersion(C_i)$$

Clearly, if the contribution of a data point is negative, it has an adverse impact on its cluster. On the other hand, a positive contribution indicates that the removal of the point from the cluster would degrade its quality.

In our work, we propose a clustering algorithm that treats points with negative contribution different from those with positive contribution. In the case of a negative contribution point, the point is shifted to a cluster, where its contribution is the highest, possibly positive. On the other hand, for a positive contribution point, a multi-objective optimization criterion is taken, where we try to optimize both the intra-cluster and inter-cluster dispersion measures.

III. ALGORITHM

We now outline the proposed contribution-based clustering algorithm. It optimizes on two measures, namely the intra-cluster dispersion given by

$$\alpha = \frac{1}{n} \sum_{x \in C_i} (x - m_i)^2$$

and the inter-cluster dispersion given by

$$\beta = \frac{1}{k} \sum_{i=1}^k (m_i - \bar{m})^2$$

where k is the number of clusters and \bar{m} is the mean of all centroids. The algorithm tries to minimize α and maximize β .

The three steps (passes) in the algorithm are described below.

Step 1: Initialization

Randomly select k centroids (m_1, m_2, \dots, m_k)

For each point x

Find $1 \leq l \leq k$ such that distance(x, m_l) is minimum

Add x to cluster C_l and update centroid m_l .

End For

Step 2: Negative Contribution Points

For each cluster C_l

For each point $x \in C_l$

If contribution(x, C_l) < 0

Move x to a cluster C_p such that contribution(x, C_p) is maximum

Update centroid m_p

End If

End For

End For

Step 3: Positive Contribution Points

For each cluster C_l

For each point $x \in C_l$

If contribution(x, C_l) ≥ 0

Move x to a cluster C_p such that $\frac{\alpha - \alpha_{new}}{\alpha} +$

$\frac{\beta_{new} - \beta}{\beta_{new}}$ is maximum

Update centroid m_p

End if

End for

End for

Note that α_{new} and β_{new} are values of α and β after the point x is moved to cluster C_p .

IV. COMPLEXITY ANALYSIS

Consider a cluster C_l with n points. Let x_i denote the i^{th} point in the cluster and let d be the dimension of each point. Then, its centroid is given by

$$m = \left(\frac{1}{n} \sum_{i=1}^n x_{i1}, \frac{1}{n} \sum_{i=1}^n x_{i2}, \dots, \frac{1}{n} \sum_{i=1}^n x_{id} \right)$$

Let us consider a point x_{n+1} outside the cluster. The contribution of the point if it were added to the cluster would be

$$contribution(x_{n+1}, C_l) = dispersion(C_l) - dispersion(C_l \cup \{x_{n+1}\})$$

Note that

$$\begin{aligned} dispersion(C_l) &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^d \left(\mathbf{x}_{ij}^2 + \frac{1}{n^2} \left(\sum_{k=1}^n \mathbf{x}_{kj} \right)^2 - \right. \\ &\quad \left. \frac{2}{n} \mathbf{x}_{ij} \sum_{k=1}^n \mathbf{x}_{kj} \right) = \\ &= \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^d \mathbf{x}_{ij}^2 + \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^d \left(\sum_{k=1}^n \mathbf{x}_{kj} \right)^2 - \\ &\quad \frac{2}{n^2} \sum_{i=1}^n \sum_{j=1}^d \mathbf{x}_{ij} \sum_{k=1}^n \mathbf{x}_{kj} \end{aligned}$$

and

$$\begin{aligned} dispersion(C_l \cup \{\mathbf{x}_{n+1}\}) &= \frac{1}{n+1} \sum_{i=1}^{n+1} \sum_{j=1}^d \mathbf{x}_{ij}^2 + \\ &\quad \frac{1}{(n+1)^2} \sum_{i=1}^{n+1} \sum_{j=1}^d \left(\sum_{k=1}^{n+1} \mathbf{x}_{kj} \right)^2 - \frac{2}{(n+1)^2} \sum_{i=1}^{n+1} \sum_{j=1}^d \mathbf{x}_{ij} \sum_{k=1}^{n+1} \mathbf{x}_{kj}. \end{aligned}$$

Substituting, we get

$$\begin{aligned} contribution(\mathbf{x}_{n+1}, C_l) &= \sum_{i=1}^n \sum_{j=1}^d \mathbf{x}_{ij}^2 \left(\frac{1}{n} - \frac{1}{n+1} \right) - \\ &\quad \frac{1}{n+1} \sum_{j=1}^d \mathbf{x}_{n+1,j}^2 + \sum_{i=1}^n \sum_{j=1}^d \left(\sum_{k=1}^n \mathbf{x}_{kj} \right)^2 \left(\frac{1}{n^3} - \frac{1}{(n+1)^3} \right) - \\ &\quad \frac{1}{(n+1)^3} \sum_{j=1}^d \left(\sum_{k=1}^{n+1} \mathbf{x}_{kj} \right)^2 - 2 \sum_{i=1}^n \sum_{j=1}^d \mathbf{x}_{ij} \left(\frac{1}{n^2} \sum_{k=1}^n \mathbf{x}_{kj} - \right. \\ &\quad \left. \frac{1}{(n+1)^2} \sum_{k=1}^{n+1} \mathbf{x}_{kj} \right) = SS \left(\frac{1}{n} - \frac{1}{n+1} \right) - \frac{1}{n+1} \sum_{j=1}^d \mathbf{x}_{n+1,j}^2 + \\ &\quad n \left(\frac{1}{n^3} - \frac{1}{(n+1)^3} \right) \sum_{j=1}^d (S_j)^2 - \frac{1}{(n+1)^3} \sum_{j=1}^d (S_j + \mathbf{x}_{n+1,j})^2 - \\ &\quad 2 \sum_{j=1}^d S_j \left(S_j \left(\frac{1}{n} - \frac{1}{(n+1)^2} \right) - \mathbf{x}_{n+1,j} \right), \end{aligned}$$

where $SS = \sum_{i=1}^n \sum_{j=1}^d \mathbf{x}_{ij}^2$ and for $1 \leq j \leq d$, $S_j = \sum_{i=1}^n \mathbf{x}_{ij}$.

It can be clearly seen that the contribution of a point in a cluster can be calculated in $O(d)$ time complexity provided the values of SS and S_j , $1 \leq j \leq d$, are recorded and updated after the transfer of each point. Similarly, it can be shown that the inter-cluster and intra-cluster dispersions can be calculated in $O(d)$ time complexity. Since the algorithm spends most of its time in calculating contributions, the time complexity of each step is $O(Nkd)$, where N is the total number of data points. This is same as the time complexity of a single iteration of the k-means algorithm. Note that the complexity becomes near-linear when N is quite large compared to k and d [2]. The space complexity of the algorithm is $O((N+k)d)$.

V. CONTENT-BASED IMAGE RETRIEVAL

Content-based image retrieval is a technique which uses visual content to search images from large-scale databases according to users' interest [8]. A common method of querying a content-based image retrieval system is to provide an example image. The system then retrieves all images in the database that are similar in content with the query image.

In this paper, we focus on the application of clustering to content-based image retrieval. A large collection of images is partitioned into a number of image clusters. Given a query image, the system retrieves all images from the cluster that is closest in content to the query image. The overall system is shown in Fig. 1. We apply the proposed contribution-based clustering algorithm to image retrieval and compare its performance with that of the k-means algorithm.

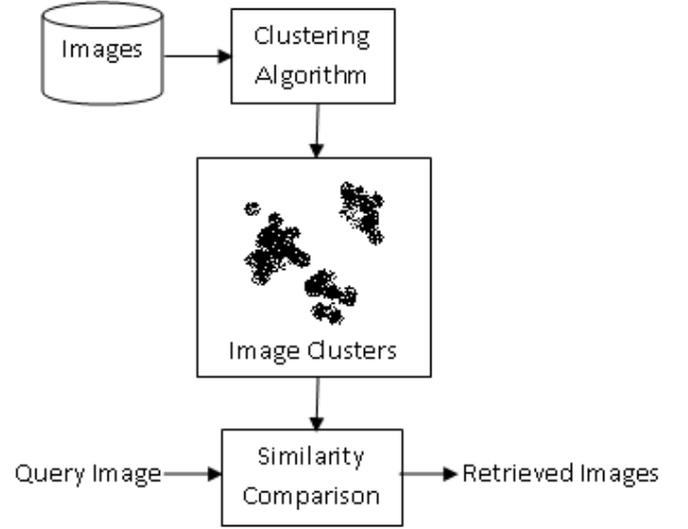


Fig. 1. Content-based Image Retrieval (CBIR) System.

Each image in the database is represented by a visual content descriptor consisting of a set of visual features [8]. A similarity/dissimilarity measure is then used to retrieve images whose features are closest to that of the query image. A common distance/dissimilarity metric is the Euclidean distance, which is used in our work. To represent the visual content of an image, we use a RGB color histogram. The color coordinates of the RGB color space are uniformly quantized into a number of bins. In our work, we use 8 bins each for the Red, Green and Blue coordinates, resulting in 512 bins/features.

VI. EXPERIMENTS AND RESULTS

Our test data consisted of 777 images belonging to 18 categories obtained from the University of Washington's Object and Concept Recognition for CBIR research project image dataset [21]. Each category contained varying number of images. All the images contained a textual description mentioning the salient foreground objects.

The images were clustered using our algorithm with the initial centroids chosen at random. The cluster whose centroid was closest in distance to the given test image was determined and the images belonging to the cluster were retrieved. The results were then compared with images retrieved using the k-means clustering algorithm with the same set of initial centroids.

Some of the retrieved images for sample test images are given in Table I.

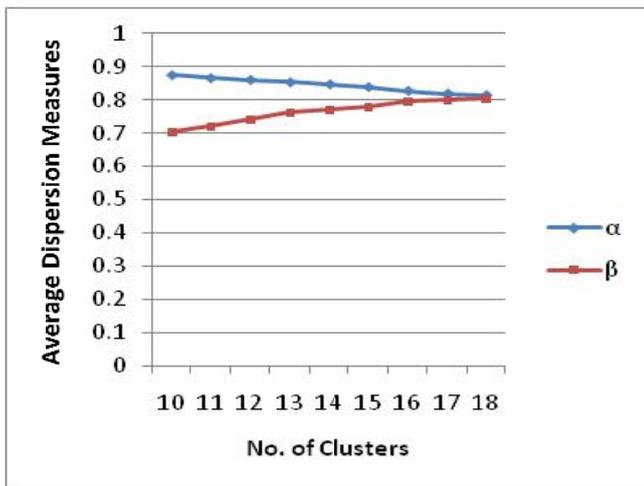
The following performance measures were used to evaluate the performance of the algorithm.

$$\text{Precision} = \frac{\text{Total number of retrieved relevant images}}{\text{Total number of retrieved images}}$$

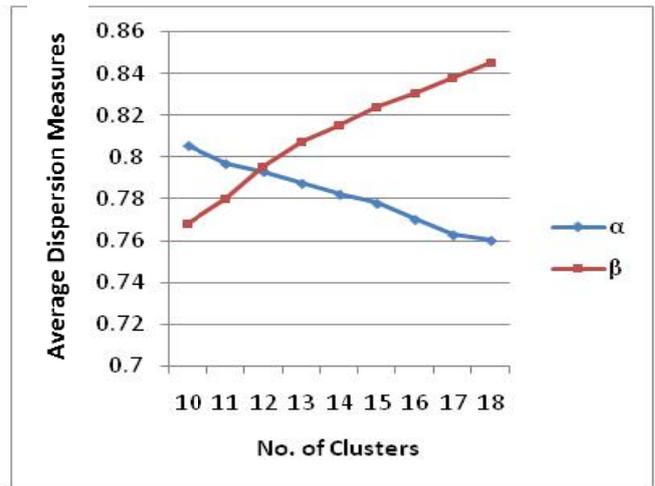
$$\text{Recall} = \frac{\text{Total number of retrieved relevant images}}{\text{Total number of relevant images}}$$

TABLE I. OUTPUTS FOR SAMPLE TEST IMAGES

| Test Image | Some Retrieved Images |
|---|--|
|  |      |
|  |      |
|  |      |
|  |      |
|  |      |

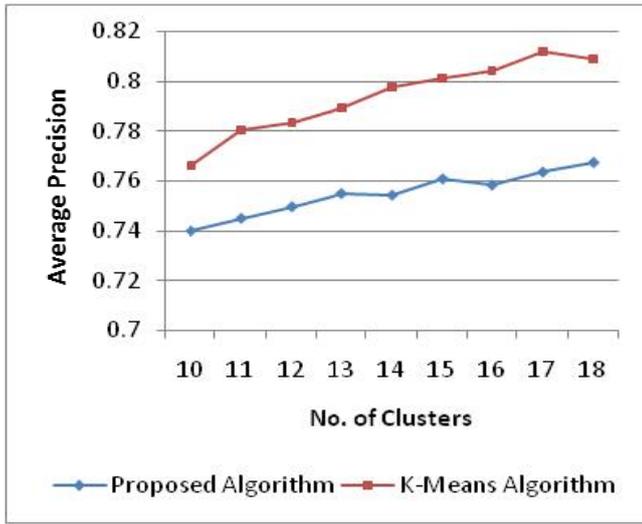


(a) Proposed Algorithm

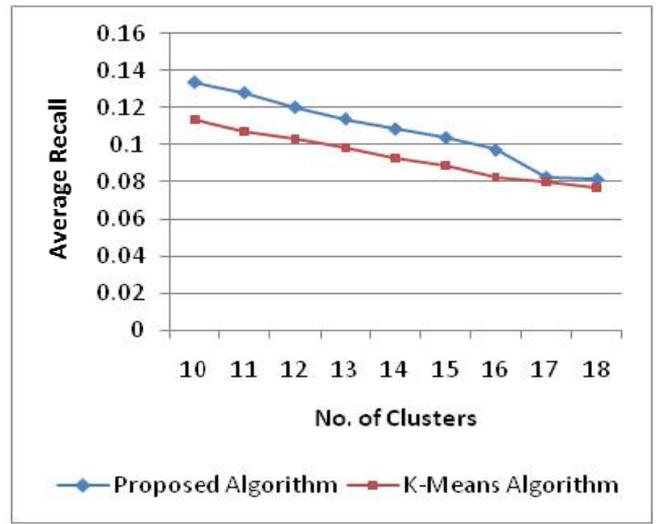


(b) K-Means Algorithm

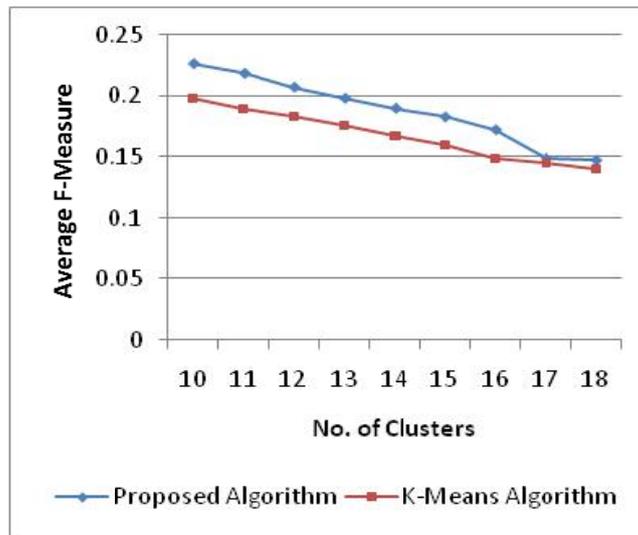
Fig. 2. Plots of average values of α and β against number of clusters.



(a)



(b)



(c)

Fig. 3. Plots of average Precision, Recall and F-Measure against number of clusters.

Note that an image is regarded as relevant if it shares at least one object common with the query image.

Since algorithms that improve on precision may degrade recall and vice versa, we use another performance evaluation measure called F-measure, which combines both precision and recall. F-measure is the harmonic mean of precision and recall and is given by

$$F - \text{Measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The experiment was repeated for 30 different randomly selected initial sets of centroids. Each set of clusters produced was then evaluated with 20 query images. The performance measures were averaged over $20 \times 30 = 600$ trials.

The average values of the dispersion measures, α and β were found for different number of clusters. The resultant

plots for the proposed contribution-based algorithm and the k-means algorithm are shown in Fig. 2(a) and Fig. 2(b) respectively. As the number of clusters increases, the average values of α and β seems to converge for the proposed clustering algorithm, while it shows a divergent trend in the case of the k-means algorithm. The difference in trend is due to the fact that our algorithm performs a multi-objective optimization on both the dispersion measures, while the k-means algorithm performs optimization only on the inter-cluster dispersion.

The average values of precision, recall and F-Measure were analyzed for different number of clusters. From Fig. 3(a) and 3(b), it is seen that the proposed algorithm gives lower average precision than the k-means algorithm, while it gives higher average recall. However, as shown in Fig. 3(c), the F-Measure is always higher for our algorithm when compared to the k-means algorithm.

The low recall in both cases is due to the fact that query output consists of images retrieved from a *single* cluster. Alternatively, the system could output a set of clusters ranked by relevance. This would improve the performance of the system and also, provide the user with a convenient interface for browsing through the retrieved images.

It has to be noted that the motivation behind our work is not to improve the performance of existing CBIR systems, but instead to show that the proposed algorithm performs better at image clustering when compared to the popular k-means partitioning clustering algorithm.

VII. CONCLUSIONS AND FUTURE WORK

We have thus proposed a new partitioning clustering algorithm based on the notion of ‘contribution of a data point’ Unlike the k-means algorithm, our algorithm optimized on both the intra-cluster and inter-cluster similarity measures and required fewer passes with each pass having the same time complexity as that of the k-means algorithm. We applied the clustering algorithm to content-based image retrieval and our experiments reveal that the algorithm improves on recall at the cost of precision.

As with many other clustering algorithms, a limitation with our algorithm is that it requires the number of clusters to be known in prior. Various methods exist to determine the number of clusters for a given dataset [11] including the one based on game theory [10]. A problem with the k-means and k-medoids clustering algorithm is that they do not perform well when the clusters are non-spherical in shape. Whether the proposed algorithm suffers from such a limitation is yet to be investigated.

Future lines of work would be to apply the concept of contribution to other clustering techniques such as hierarchical clustering. Our algorithm could also be extended to fuzzy partitioning of data points. The content-based image retrieval technique described in this paper uses only the RGB color histogram as the visual content descriptor of an image. The performance of the system with other visual features based on shape and texture and other distance metrics would have to be tested [8]. Also, learning through relevance feedback from the user could be incorporated in the proposed system [12].

ACKNOWLEDGMENT

The authors would like to thank Dr. S. Sridhar for a valuable discussion they had with him on data clustering.

REFERENCES

- [1] J. Han and K. Micheline, “Data mining concepts and techniques,” Morgan Kaufman, 2006.
- [2] R. Xu and D. Wunsch, “Survey of clustering algorithms,” IEEE Transactions on Neural Networks, Vol.16, Issue 3, pp. 645– 678, May 2005.
- [3] E. Forgy, “Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications,” Biometrics, vol. 21, pp. 768–780, 1965.
- [4] J. MacQueen, “Some methods for classification and analysis of multivariate observations,” in Proc. 5th Berkeley Symp., vol. 1, 1967, pp.281–297.

- [5] L. Kaufman and P. Rousseeuw, “Finding groups in data: An introduction to cluster analysis,” Wiley, 1990.
- [6] F. Höppner, F. Klawonn, and R. Kruse, “Fuzzy cluster analysis: Methods for classification, data analysis, and image recognition,” New York, Wiley, 1999.
- [7] P. Bradley, U. Fayyad, and C. Reina, “Scaling clustering algorithms to large databases,” in Proc. 4th Int. Conf. Knowledge Discovery and Data Mining (KDD’98), 1998, pp. 9–15.
- [8] F.H. Long, H.J. Zhang, and D.D. Feng, “Fundamentals of content-based image retrieval,” in D.D. Feng, W.C. Siu, and H.J. Zhang (Eds), ‘Multimedia information retrieval and management—technological fundamentals and applications’, Springer-Verlag, New York, 2003, pp. 1–26
- [9] M.J. Osborne, “An introduction to game theory,” Oxford University Press, USA, 2007.
- [10] V.K. Garg, “Pragmatic data mining: Novel paradigms for tackling key challenges,” Project Report, Computer Science & Automation (CSA), Indian Institute of Science, 2009.
- [11] D. Pelleg and A. Moore, “X-means: Extending k-means with efficient estimation of the number of clusters,” in Proc. 17th Int. Conf. Machine Learning (ICML’00), 2000, pp. 727–734.
- [12] Y. Liu, D. Zhang, G. Lu, and W. Ma, “A survey of content-based image retrieval with high-level semantics,” Pattern Recogn., vol. 40(1), 2007, pp. 262-282.
- [13] R. Datta, D. Joshi, J. Li, and J.Z. Wang, “Image retrieval: Ideas, influences, and trends of the new age,” ACM Comput. Surv., 40(2), Apr. 2008, pp. 1-60.
- [14] H. Zhang, J.E. Fritts, and S.A. Goldman, “Image segmentation evaluation: A survey of unsupervised methods,” Comput. Vis. Image Underst., 110(2), May. 2008, pp. 260-280.
- [15] D. Cai, X. He, Z. Li, W. Ma, and J. Wen, "Hierarchical clustering of WWW image search results using visual, textual and link information," in Proc. 12th Annual ACM International Conference on Multimedia (MULTIMEDIA '04), Oct. 2004, New York, NY, pp. 952-959.
- [16] D. Kinoshenko, V. Mashtalir and E. Yegorova, "Clustering method for fast content-based image retrieval," Computer Vision and Graphics, 32, Mar. 2006, pp. 946-952.
- [17] Y. Chen, J.Z. Wang, and R. Krovetz, “Content-based image retrieval by clustering,” in Proc. 5th ACM SIGMM international Workshop on Multimedia information Retrieval, Berkeley, California, Nov. 2003, MIR '03, ACM, New York, NY, pp. 193-200.
- [18] P.J. Dutta, D.K. Bhattacharyya, J.K. Kalita and M. Dutta, "Clustering approach to content based image retrieval," in Proc. Conference on Geometric Modeling and Imaging: New Trends (GMAI), 2006, IEEE Computer Society, Washington, DC, pp. 183-188.
- [19] G. Liu and B. Lee, "A color-based clustering approach for web image search results," in Proc. 2009 International Conference on Hybrid information Technology (ICHIT '09), Daejeon, Korea, Aug. 2009, vol. 321, ACM, New York, NY, pp. 481-484.
- [20] K. Jarrah, S. Krishnan, L. Guan , "Automatic content-based image retrieval using hierarchical clustering algorithms," in Proc. International Joint Conference on Neural Networks (IJCNN '06), Oct. 2006, Vancouver, BC pp. 3532 - 3537.
- [21] “Object and concept recognition for content-based image retrieval,” [Online]. Available: <http://www.cs.washington.edu/research/imagedatabase/groundtruth/>.