Image Retrieval using Canopy and Improved K mean Clustering

S.Sabena
Assistant Professor
Anna University of Technology, Tirunelveli

Dr.P.Yogesh
Associate Professor
Department of IST,
Anna University, Chennai

L.SaiRamesh
Teaching Fellow
Department of IST,
Anna University, Chennai

ABSTRACT

In a typical content based image retrieval (CBIR) system, target images are sorted by feature similarities with respect to the query. These methods fail to capture similarities among target images and user feedback. To overcome this problem existing methods combine relevance feedback and clustering. But clustering requires more number of expensive distance calculations. To remedy this problem we propose a new technique that combine canopy method, relevance feedback and improved k mean clustering. Canopy method reduces expensive distance calculation by measuring exact distances between points that occur in a common canopy. Improved k mean clustering automatically compute number of cluster and uses max min distance to reduce computational complexity. Relevance feedback captures exact user interest. The experiments show that our method is highly effective for image retrieval..

General Terms

Canopy method, improved k-mean clustering.

Keywords

Clustering, Image retrieval, learning methods, Relavance feedback

1. INTRODUCTION

Content based image retrieval (CBIR) systems analyze the visual content description to organize and find images in databases. The retrieval process usually relies on presenting a visual query to the systems, and extracting from a database the set of images that best fit the user request. This method is known as query by example, requires the definition of an image representation and need some similarity metrics to compare query and target images. But this method has a lot of problems. First, how good is the description provided by the adopted feature set, i.e., are the selected features able to provide a good clustering of the requested images, retrieving a sufficient number of desired images and avoiding false positives? Second, can the query completely capture the user interest? Third, how can cluster relevant images in a reliable manner?

According to this, several additional mechanisms have been introduced to achieve better performance. Among them, relevance feedback (RF) technique is an interactive strategy which is effective to improve the accuracy of information retrieval systems, in particular CBIR systems [4], [5], [6]. RF has a short-term memory. It is adapting the retrieval process for a specific user and a specific query. The user first submit a query, then sees some results and interacts in order to modify them by asking the system to change the weights of parameters or to modify the query itself for adapting the result to the real user's intents. Short term memory means that during that interaction time, the system can remember the results, but once it is finished, the system cleans its memory and the next user starts from scratch. RF approaches so far proposed show some critical issues yet unsolved. First, user interaction is time consuming and it is therefore desirable to reduce as much as possible the number of iterations to convergence.

Users evaluate the effectiveness of a CBIR system by the ranked results, but the relevant images are often not at the top of the rankings. Even if a system finds particular relevant images, many others may be substantially further down in the ranking. We call this the rank inversion problem. Existing research shows that this problem can be improved by analyzing the retrieved images using clustering method [7, 8]. Cluster analysis is an unsupervised learning method that constitutes a cornerstone of an intelligent data analysis process. It is the process of grouping objects into clusters such that the objects in the same cluster are similar where as objects in different clusters is different. K-Means [9] is a prototype-based, partitioned clustering technique that attempts to find a user-specified number of clusters (K). The problem with this method is results are depend initial number of cluster and initial condition. It also becomes computationally expensive when the data set to be clustered is large.

In this paper, a novel method is proposed which combine relevance feedback, canopy method and improved k mean clustering for efficient retrieval of images. Canopy method reduces the computation complexity by measuring exact distances between points that occur in a common canopy. Relevance feedback is used to identify the relevant and irrelevant clusters. Improved k mean cluster [1] automatically detect number of cluster for better solution.

2. RELATED WORK

2.1 Relevance Feedback

As already mentioned, the proposed technique is based on the well-known concept of relevance feedback. Relevance feedback is a powerful technique used in traditional text-based information retrieval systems. It is the process of automatically adjusting an existing query using the information fed back by the user about the relevance of previously retrieved objects such that the adjusted query is a better approximation to the user's information need [10], [11], [12]. In the relevance-feedback-based approach [13] the retrieval process is interactive between the computer and the human.

The basic RF mechanism consists in iteratively asking the user to discriminate between relevant and irrelevant images on a given set of results. The collected feedback is then used to drive different adaptation mechanisms which aim at better separating the relevant image cluster or at reformulating the query based on the additional user input. In the first case, we may apply feature re-weighting [14] or adaptation [15] algorithms, which modify the solution space metrics, giving more importance to some features with respect to others. A binary RF is used to train neural network systems as in PicSOM [16]. In [17], a fuzzy RF is described, where the user provides the system with a fuzzy judgment about the relevance of the images. By this kind of feedback user interest is captured and efficient result is achieved.

2.2 Canopy

Canopy Clustering is a very simple, fast and surprisingly accurate method for grouping objects into clusters. All objects are represented as a point in a multidimensional feature space. The algorithm uses a fast approximate distance metric and two distance thresholds T1 > T2 for processing. The basic algorithm is to begin with a set of points and remove one at random. Create a Canopy containing this point and iterate through the remainder of the point set. At each point, if its distance from the first point is < T1, then add the point to the cluster. If, in addition, the distance is < T2, then remove the point from the set. This way points that are very close to the original will avoid all further processing. The algorithm loops until the initial set is empty, accumulating a set of Canopies, each containing one or more points. A given point may occur in more than one Canopy.

Canopy Clustering is often used as an initial step in more rigorous clustering techniques, such as K-Means Clustering [3]. By starting with an initial clustering the number of more expensive distance measurements can be significantly reduced by ignoring points outside of the initial canopies. Canopy technique is more useful when the data set is large. The key idea is to perform clustering in two stages, first a rough and quick stage that divides the data into overlapping subsets we call canopies, then a more rigorous final stage in which expensive distance measurements are only made among points that occur in a common canopy.

2.3 K mean Clustering

The k-means algorithms, like other partition clustering algorithms, group n data points into k clusters by minimizing a cost function that has been pre-designed. H. Friguiand and O. Nasraoui [18], Y. Chan and W. Ching [19] introduce the degree of membership for each object belonging to every cluster and the weight for each dimension of a cluster on contributing to clustering. However, their algorithm is not Computable if one of weight is happens to be zero. Generally, the K-Means algorithm has the following important properties: (i) it is efficient in processing large data sets, (ii) it often terminates at a local optimum, (iii) the clusters have spherical shapes, (iv) it is sensitive to noise [20]. The following steps involved in k mean clustering.

Input: the number of cluster k, containing n data object.
Output: k cluster.

(1) From the n data objects, randomly select k data objects as the initial cluster centers.
(2) Calculate the distance of each data object to each cluster center, and assign it to the nearest cluster.
(3) The distribution of all data objects is completed. Re-calculate the center of k cluster.
(4) Comparison of the last, respectively, the corresponding cluster center, if the cluster centers change, to (2), otherwise, to (5).
(5) Output the results of clustering.

The important three steps involved in k mean clustering: First, select the initial cluster center. Second, assigning the data objects into desired cluster. Third, recalculate and adjust the cluster centers.

3. IMPROVED K MEANS CLUSTERING (IKC) ALGORITHM

In this section, we give a brief description of the IKC algorithm [1].

Traditional K-Means algorithm first randomly select k data objects as the initial cluster center, initial each data object to represent the average of a cluster or center. Each of the remaining data objects, according to its distance from the cluster center, from which it is assigned to the nearest cluster, and then re-calculated the average of each cluster. This process is repeated until the convergence criterion function. K mean algorithm has the following disadvantage:
(1) Sensitive to initial clustering center, and different initial centers often correspond to different clustering results; (2) sensitive to the order of data input; from different initial cluster centers will be different from the results of the cluster and not the same as accuracy; (3) Vulnerable to the impact of noise and isolated points.

To remedy this problem, Improved K mean cluster take the following steps:

1) Calculate the distances of the data points between each other in the data set U;

2) Select two nearest data points to form a subset A1, and remove them from the data set U;

3) Calculate the distances between the subset A1 and the remaining data points in data set U, and put the data that is nearest to subset A1 into A1.Repeat this process until the data points in Al reach a certain number. The distances between the subset Al and a data point can be calculated by the following formula:

$$d (x , A1) = \min(d (x , y), y \in A 1)$$

4) Repeat steps (2) and (3) until K subsets A1, A 2... A k is formed;

5) Calculate the average values of the K subsets: ml, m2 ... mk. We choose ml, m2 ... mk as initial centers.

Disadvantage of this method is, the computation become expensive when data set is large.

4. PROPOSED ALGORITHM

Proposed algorithm combine canopy [2] technique, Relevance feedback and Improved K mean clustering [1]. In solution space every data object is represented by different features. In our proposed method color histogram, Edge histogram, Color moment and Color and Edge Directivity Descriptor (CEDD) features are used to represent each images. Color histogram is used to find the distribution of different colors. Edge histogram finds the edge strength and direction. Canopy Clustering is a fast way to cluster the data objects by using threshold value T1 and two distant metrics. Initially query image is considered the first seed point. The distance from seed point for each data object in data sets is calculated using Euclidean distances. If the distance lies within threshold T1 then data object is assigned to that cluster. Here cheapest distance measure is used which consider only important features of images. If distance exceeds the threshold T1 then that data object will act as a initial seed for the next canopy.

After creating canopies improved k mean clustering which does not depend the initial number of cluster is applied. This is done by expensive distant measures which consider all the features of data object. Expensive distance measure is never applied to the data objects which don't lie in the same canopy. In proposed method Canopy Clustering is used as an initial step in improved K-Means Clustering [3]. By starting with an initial clustering the number of more expensive distance measurements can be significantly reduced by ignoring points outside of the initial canopies. Proposed technique has the following steps.

1. Create canopies using two threshold values T1 .Here cheapest distance metric is used.
2. Select two nearest data points from any canopy to form a subset A1.
3. Calculate the distances between the subset A1 and the remaining data points that occur in same canopy and put the data that is nearest to subset A1 into A1. Repeat this process until the data points in Al reach a certain number.
4. Repeat steps (2) and (3) until K subsets AI, A 2... A k is formed.
5. Calculate the average values of the K subsets: ml, m2... mk. We choose ml, m2, ... mk as initial centers

After forming clusters, the representative images in each cluster are shown to the user for feedback. The representative images will be more similar to the query image. In the feedback process user is requested to tick whether the cluster is relevant or irrelevant. Initially all the features have the same weight. If a cluster is considered as relevant then more weight is assigned. If cluster is irrelevant then weight of feature is reduced. Based on the user feedback the first N more relevant images are retrieved and shown to user.

We summarize the whole algorithm process as follows:

Input: Query Image.
Output: First 'N' relevant images.

1. Query image is segmented and given for user feedback
2. Based on feedback feature weight are updated.
3. Through applying IKC with canopy clustering, we partition the top ranked images from the initial retrieved results into K clusters
4. The user is required to label the representative images of the K clusters as relevant or irrelevant.
5. By the second feedback the distance between the query image and the initial retrieved images are calculated.
6. Based on new distance, the first 'N' images are retrieved and shown to user.

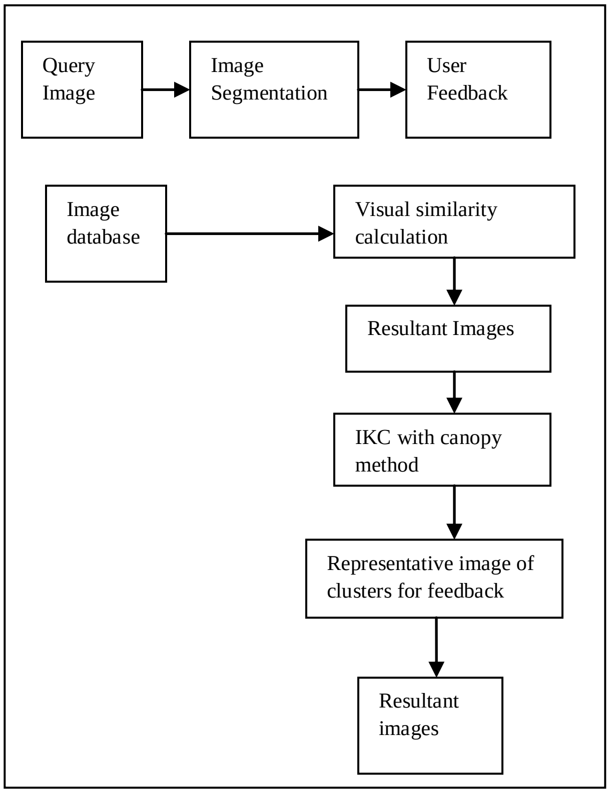

Fig 1: Flowchart of system architecture

5. EXPERIMENTS AND RESULTS

In order to evaluate the algorithms, we use 500 images from Corel experimental data set. The database includes a variety of images such as apples, forest, and flowers etc. In the initial retrieval, we use a color histogram, edge histogram and color moment to represent the image content. We design experiments similar to that taken in [2] to evaluate the algorithms.

$$\text{Precision} = (\text{relevant} \cap \text{retrieved})/\text{retrieved}.$$

$$\text{Recall} = (\text{relevant} \cap \text{retrieved})/\text{relevant}.$$

Figure 2 show the precision of the 2 methods for the first 50 images retrieved. For five iterations the average precision of both IKC and IKC with canopy is obtained. The average precision of IKC with canopy is significantly high when compared to IKC method.

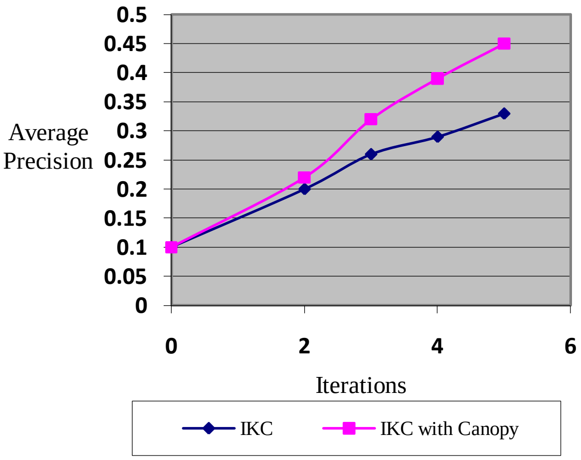

Fig 2: The average precision of IKC and IKC with canopy method.

We compare the different ranking results using Average Precision (AP) as the performance metric, which is the average of precisions computed after each relevant image, is retrieved. When compared to improved k mean cluster method, canopy with IKC take advantage due to the reduced computation complexity. This is shown in figure 3. Cluster validation is a technique to find the optimized number of clusters. Our method is independent of initial number of cluster. So it is necessary to find the optimized cluster. It is done by measuring intra cluster distance and inters cluster distance. To achieve efficiency intra cluster distance is minimized and inter cluster distance is maximized.

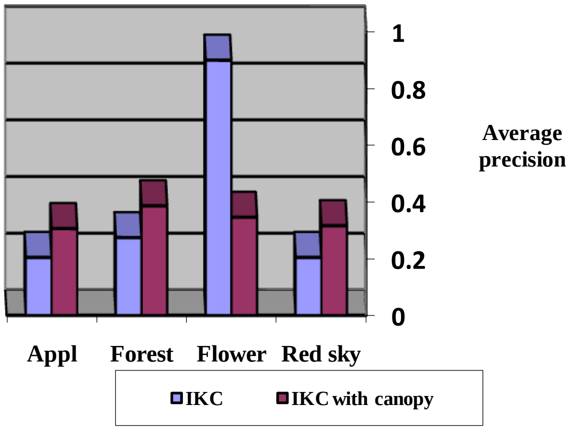

Fig 3: Performance analysis based on average precision.
The silhouette width is the average of each observation's silhouette value. The silhouette value measures the degree of confidence in the clustering assignment of a particular observation, with well-clustered observations having values near1.

For observation i, it is defined as

$$s(i)= (b_i-a_i) / \max(b_i, a_i).$$

where a_i is the average distance between i and all other observations in the same cluster, and b_i is the average distance between i and the observations in the nearest neighboring cluster.

The goodness of each type of cluster is shown in the figure 4. When number images increase the silhouette width of each cluster type also increases.

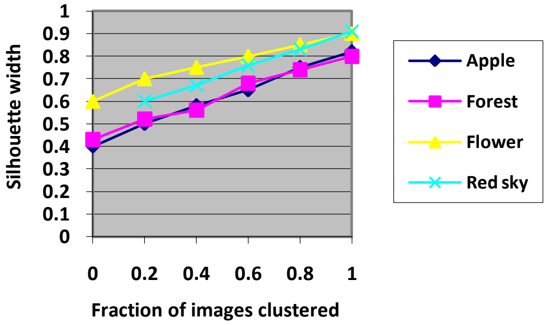

Fig 4: Performance analysis based on Silhouette width.

6. CONCLUSION

In this paper we propose a new method which combines canopy cluster, improved k mean cluster and relevance feedback to re-rank the search result. Canopy cluster reduced computation time by more than an order of magnitude while also slightly increasing accuracy. Experimental results show that our re-ranking algorithm achieves a more rational ranking of retrieval results and it is superior to the method in [2]. In future the canopy and relevance feedback method can be used with Fuzzy c mean cluster for more accurate retrieval.

7. REFERENCE

[1] ZHANGXu-bo, PENGJin-ye.2010. "Re-ranking algorithm using clustering and relevance feedback for image retrieval".

[2]. Andrew McCallum, Kamal Nigam.2005. Lyle H. Ungar, "Efficient Clustering of High Dimensional Data Sets with Application to Reference Matching".

[3]. Nhu-Van Nguyen, Alain Boucher.2005."Clusters-based Relevance feedback for CBIR: a combination of query movement and query expansion".

[4]. M. Ortega, S. Mehrotra.2004. “Relevance feedbacktechniques in the MARS image retrieval system”, Multimedia Systems, vol 9, no 6, pp 535-547.

[5]. D. Kim, C. Chung, and K. Barnard.2005. “Relevance feedback using adaptive clustering for image similarity retrieval.” J. Syst. Softw. 9-23.

[6]. Y. Ishikawa, R. Subramanya, and C. Faloutsos. 1998.“MindReader: Querying databases through multiple examples. ” In Proc. Of the 24th Intl. Conference on Very Large Databases, pp 218–227.

[7]. G. Park, Y. Baek, and H.K. Lee.2005. "Re-ranking Algorithm Using Post retrieval clustermg for Content-based Image Retrieval", Information Processing and Management, 4 1(2), pp. 177- 194.

[8]. Y. Hu, N.H. Yu, Z.W. Li.2007 "Image Search Result Clustering and Reranking via Partial Grouping", Proc of ICME ,[S.L]:[s.n], pp . 603-606.

[9]. J. B. MacQueen.1967. "Some Methods for classification and Analysis of Multivariate Observations" , Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Vol 1, pp. 281–297.

[10]. C. Buckley and G. Salton,1965.“Optimization of relevance feedback weights,” in Proc. SIGIR’95.

[11]. G. Salton and M. J. McGill.1983. Introduction to Modern Information Retrieval. New York: McGraw-Hill.

[12]. W. M. Shaw, “Term-relevance computations and perfect retrieval performance,” Inform. processing Management.

[13]. C. J. Van Rijsbergen.1987 Information Retrieval, 2nd ed. London, U.K.: Butterworths.

[14]. M. L. Kerfi and D. Ziou.2004 “Image retrieval based on feature weighing and relevance feedback,” in Proc. IEEE Int. Conf. Image Processing (ICIP2004), vol. 1, pp. 689–692.

[15]. A. Grigorova, F. G. B. De Natale, C. Dagli, and T. S. Huang.2007. “Contentbased image retrieval by feature adaptation and relevance feedback,” IEEE Trans. Multimedia, vol. 9, no. 6, pp. 1183–1192.

[16]. M. Koskela, J. Laaksonen, and E. Oja.2004 “Use of image subsets in image retrieval with self-organizing maps,” in Proc. Int. Conf. Image and Video Retrieval (CIVR), pp. 508–516.

[17]. K.-H. Yap and K. Wu.2005. “Fuzzy relevance feedback in content-based image retrieval systems using radial basis function network,” in Proc. IEEE Int. Conf. Multimedia and Expo.

[18]. H. Friguiand, and O. Nasraoui,2004.“Unsupervised Learning of Prototypes and Attribute Weights,” Pattern Recognition, vol.37, no.3, pp.567-581.

[19]. Y. Chan, W. Ching, M. K. Ng, and J.Z. Huang.2004 “An Optimization Algorithm for Clustering Using Weighted Dissimilarity Measures,” Pattern Recognition, vol.37, no.5, pp. 943-952.

[20]. J. M. Pena, J. A. Lozano, and P. Larranaga.1999 “An empirical comparison of four initialization methods for the k-means algorithm,” Pattern Recognition Letters, vol. 20, pp. 1027–1040.