Distributed K-Median Clustering with Application to Image Clustering

Aiyesha Ma and Ishwar K. Sethi

Oakland University
Rochester, MI, USA

http://iielab-secs.secs.oakland.edu/

Abstract

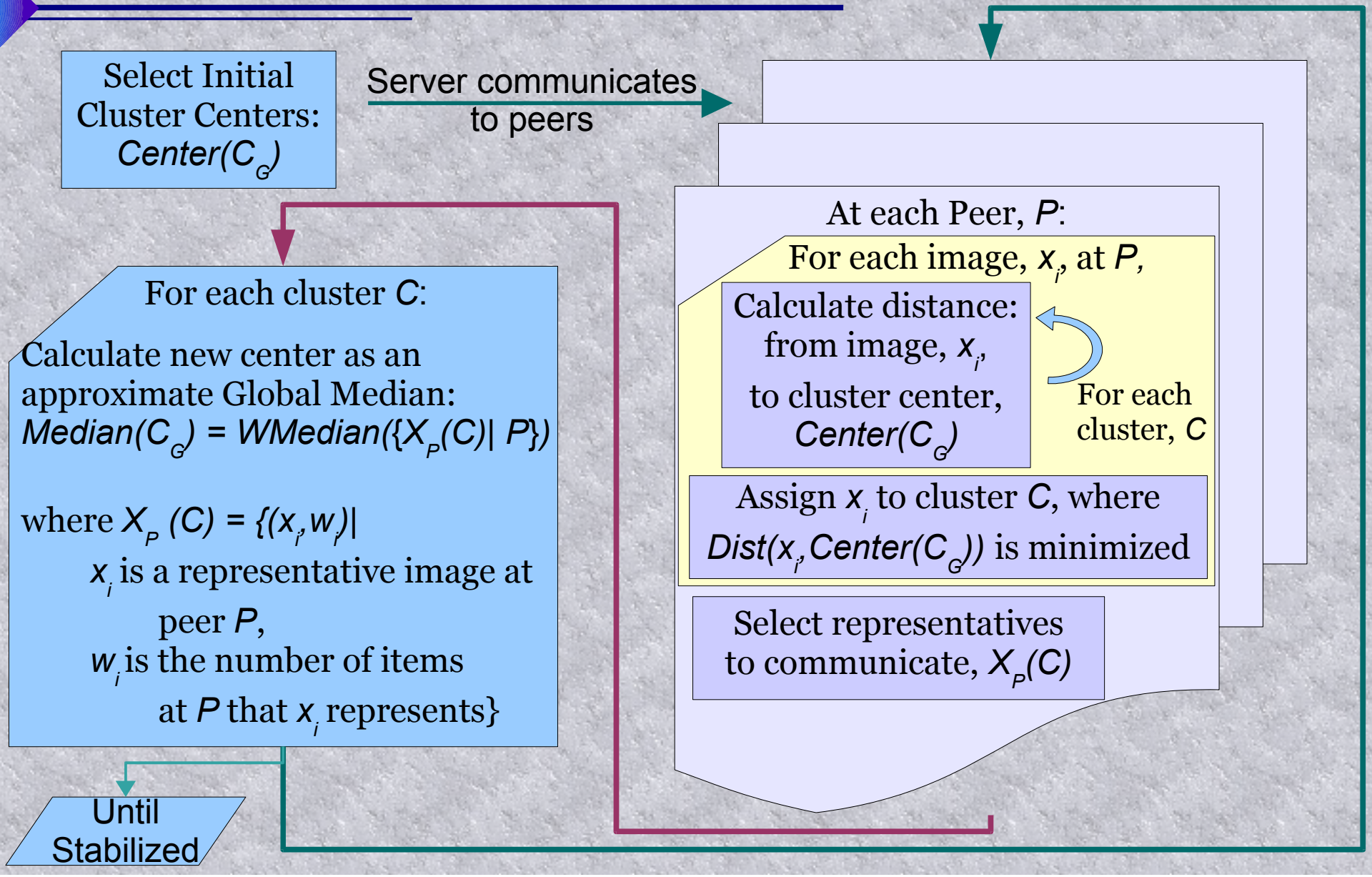
Developing algorithms suitable for distributed environments is important as data becomes more distributed. This paper proposes a distributed K-Median clustering algorithm for use in a distributed environment with centralized server, such as the Napster model in a peer-to-peer environment. Several approximate methods for computing the median in a distributed environment are proposed and analyzed in the context of the iterative K-Median algorithm.

The proposed algorithm allows the clustering of multivariate data while ensuring that each cluster representative remains an item in the collection. This facilitates exploratory analysis where retaining a representative in the collection is important, such as imaging applications.

Introduction and Background

- K-Means clustering is a well known and popular clustering technique.
  - Creates a new mean vector, which may not be meaningful in many applications
- Using the centroid of a cluster rather than the mean is one variation to the basic K-Means algorithm.
  - This is also known as the L1 Multivariate Median
- Dhillon and Modha first proposed a distributed K-Means clustering algorithm.
  - Computing the distributed median is more complicated

Distributed K-Median Clustering Algorithm



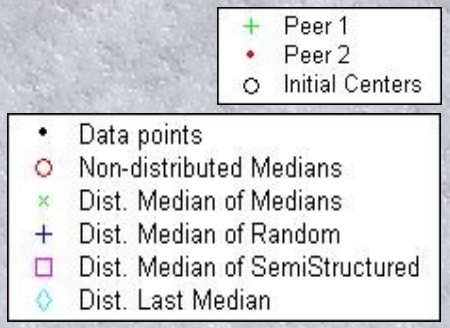
Approximate Global Median Approaches

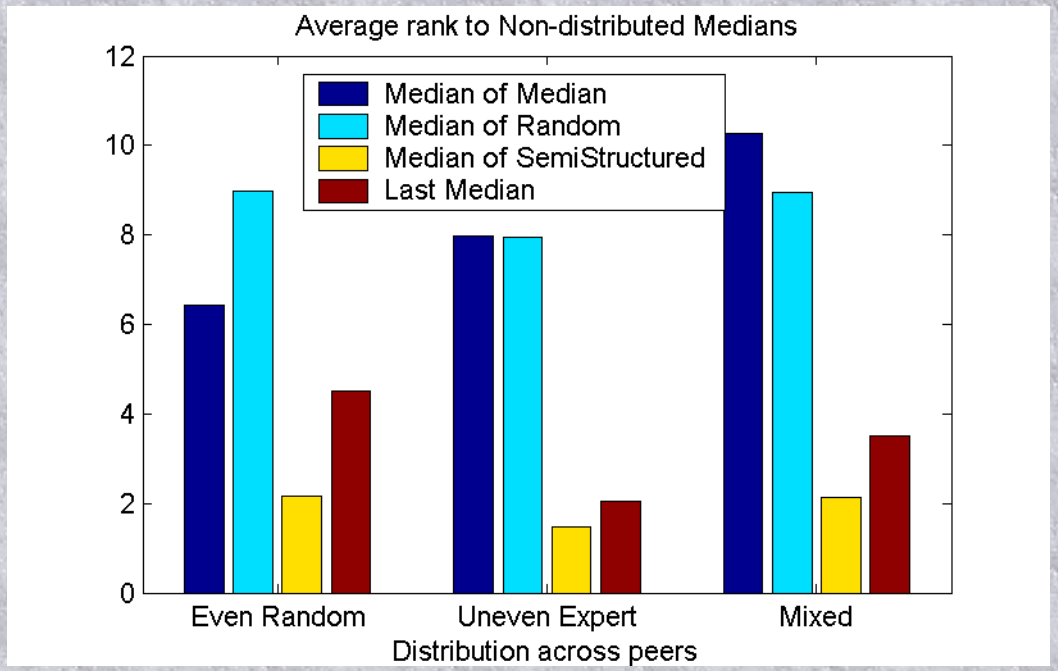
Selecting representatives to communicate

- Local Median
  - The local median of C
  - Size of C at P
- Semi-structured
  - Local median of C and additional samples, such that no element of C is represented by an example further than *RepDist* away
  - Number of samples within *RepDist*
- Random Sampling
  - Randomly select n examples from C
  - Size of C at P / n
- Last Median
  - Local median of C, and sample in P_C nearest $Median(C_G)$
  - Number of samples in P_C closest to the representative

Analysis of Approximate Global Medians

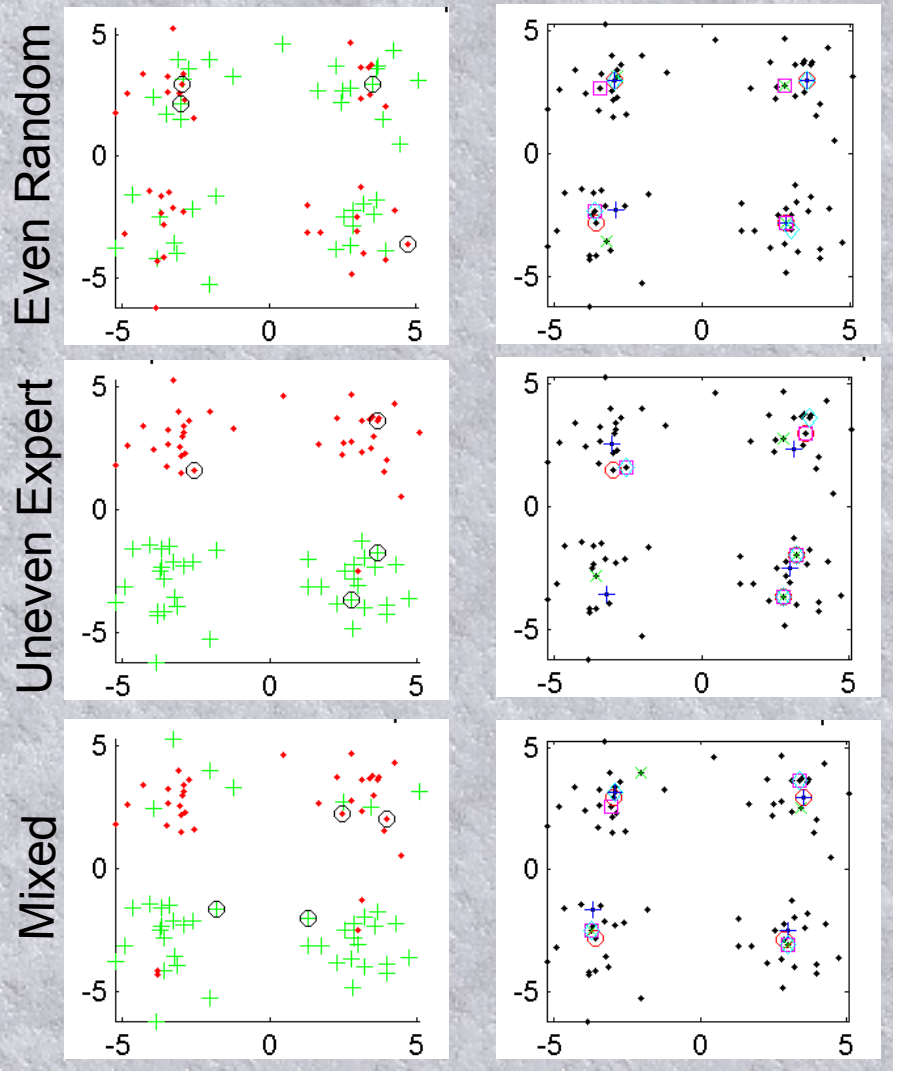
Performance tested over 100 test runs. A single example run is shown at right.





Distribution Across Peers

Clustering Result

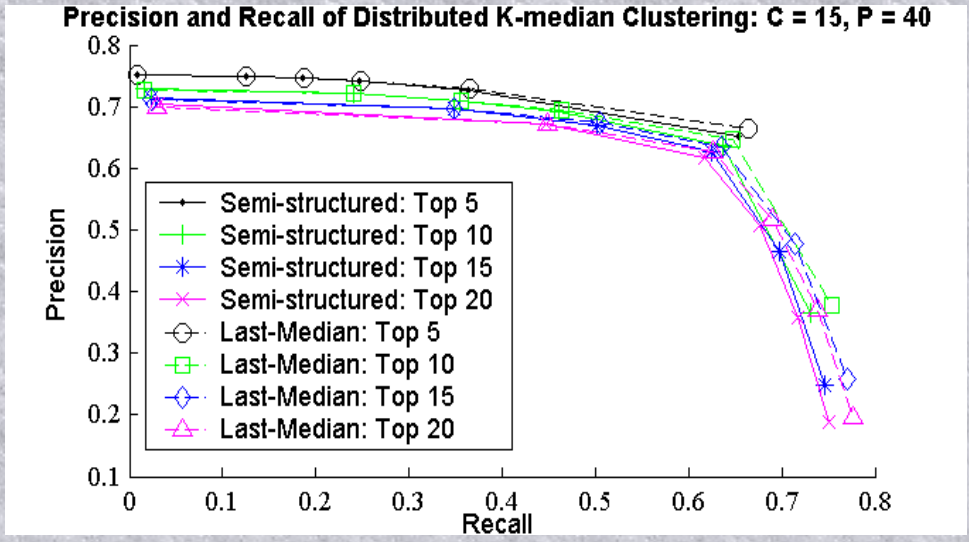


Application to Image Clustering

- 7100 Color photo images
  - KeyPhoto Collection
  - Benchathlon
  - University of Washington
- Feature vector: Global Histogram
  - 256 bins in the HSV (Hue, Saturation, Value) color space: H = 16 bins, S = 4 bins, V = 4 bins
- Images were assigned to 40 peers, using uneven expert and uneven random.

Analysis of Distributed K-Median Clustering

- Clustered images with
  - K = 9, 15, 40
  - Used Last Median, and Semi-structured approaches
  - Stabilized after 4 to 7 iterations
- Analyzed by using non-distributed clustering results as ground truth



Visual Results of Clustering





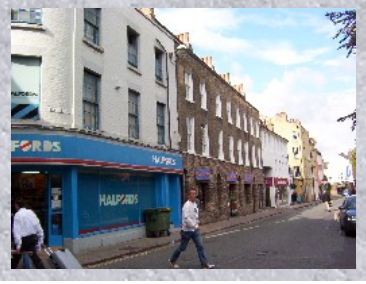










Visual Results of Clustering













Conclusion

- Proposed several schemes for passing peer information to a central server to calculate approximate global medians of the clusters.
- Analyzed the various proposed methods, and determined that two of the four methods performed better than the others. Distribution of the data affects which method may be more applicable.
- Demonstrated the use of the clustering algorithm on a test set of 7100 images. Results were analyzed relative to the results of non-distributed K-median clustering.

Selected References

- Dhillon, I.S., Modha, D.S.: A data clustering algorithm on distributed memory multiprocessors. Large-Scale Parallel Data Mining, Lecture Notes in Artificial Intelligence 1759 (2000) 245–260
- Jin, R., Goswami, A., Agrawal, G.: Fast and exact out-of-core and distributed k-means clustering. Knowledge and Information System Journal (2005)
- University of Washington Image Collection. http://www.cs.washington.edu/research/imagedatabase/groundtruth/
- Benchathlon Image Collection. http://www.benchathlon.net
- Manjunath, B.S., Ohm, J.R., Vasudevan, V.V., Yamada, A.: Color and texture descriptors. IEEE Trans. On Circuits and Systems for Video Technology, 11 (2001)
- Sikora, T.: The mpeg-7 visual standard for content description overview. IEEE Trans. On Circuits and Systems for Video Technology, 11 (2001)