A Comparison of Document Clustering Techniques

Michael Steinbach George Karypis Vipin Kumar
Department of Computer Science / Army HPC Research Center, University of Minnesota
4-192 EE/CSci Building, 200 Union Street SE
Minneapolis, Minnesota 55455
steinbac@cs.umn.edu karypis@cs.umn.edu kumar@cs.umn.edu

ABSTRACT

This paper presents the results of an experimental study of some common document clustering techniques: agglomerative hierarchical clustering and K-means. (We used both a "standard" K-means algorithm and a "bisecting" K-means algorithm.) Our results indicate that the bisecting K-means technique is better than the standard K-means approach and (somewhat surprisingly) as good or better than the hierarchical approaches that we tested.

Keywords

K-means, hierarchical clustering, document clustering.

1. INTRODUCTION

Hierarchical clustering is often portrayed as the better quality clustering approach, but is limited because of its quadratic time complexity. In contrast, K-means and its variants have a time complexity that is linear in the number of documents, but are thought to produce inferior clusters. Sometimes K-means and agglomerative hierarchical approaches are combined so as to "get the best of both worlds." For example, in the document domain, Scatter/Gather [1], a document browsing system based on clustering, uses a hybrid approach involving both K-means and agglomerative hierarchical clustering. K-means is used because of its run-time efficiency and agglomerative hierarchical clustering is used because of its quality.

However, during the course of our experiments we discovered that a simple and efficient variant of K-means, "bisecting" K-means, can produce clusters of documents that are better than those produced by "regular" K-means and as good or better than those produced by agglomerative hierarchical clustering techniques. We have also been able to find what we think is a reasonable explanation for this behavior.

We refer the reader to [2] for a review of cluster analysis and to [4] for a review of information retrieval. For a more complete version of this paper, please see [6].

The data sets that we used are ones that are described more fully in [6]. They are summarized in the following table.

Table 1: Summary description of document sets.

Data Set	Source	Documents	Classes	Words
re0	Reuters	1504	13	11465
re1	Reuters	1657	25	3758
wap	WebAce	1560	20	8460
tr31	TREC	927	7	10128
tr45	TREC	690	10	8261
fbis	TREC	2463	17	2000
la1	TREC	3204	6	31472
la2	TREC	3075	6	31472

2. Evaluation of Cluster Quality

We use two metrics for evaluating cluster quality: entropy, which provides a measure of "goodness" for un-nested clusters or for the clusters at one level of a hierarchical clustering, and the F-measure, which measures the effectiveness of a hierarchical clustering. (The F measure was recently extended to document hierarchies in [5].) Our results will show that the bisecting K-means algorithm has the best performance for these two measures of cluster quality.

3. Bisecting K-means

The bisecting K-means algorithm starts with a single cluster of all the documents and works in the following manner:

1. Pick a cluster to split.
2. Find 2 sub-clusters using the basic K-means algorithm.
3. Repeat step 2, the bisecting step, for a fixed number of times and take the split that produces the clustering with the highest overall similarity. (For each cluster, its similarity is the average pairwise document similarity, and we seek to minimize that sum over all clusters.)
4. Repeat steps 1, 2 and 3 until the desired number of clusters is reached.

We found little difference between the possible methods for selecting a cluster to split and chose to split the largest remaining cluster.

4. Agglomerative Hierarchical Techniques

We used three different agglomerative hierarchical techniques for clustering documents.

Intra-Cluster Similarity Technique: This hierarchical technique looks at the similarity of all the documents in a cluster to their cluster centroid and is defined by $\text{Sim}(X) = \sum_{d \in X} cosine(\boldsymbol{d}, \boldsymbol{c})$, where $\boldsymbol{d}$ is a document in cluster, X, and $\boldsymbol{c}$ is the centroid of cluster X, i.e., the mean of the document vectors. The

choice of which pair of clusters to merge is made by determining which pair of clusters will lead to smallest decrease in similarity.

Centroid Similarity Technique: This hierarchical technique defines the similarity of two clusters to be the cosine similarity between the centroids of the two clusters.

UPGMA: This is the UPGMA scheme as described in [2]. It defines the cluster similarity as follows, where d_1 and d_2 are documents in cluster1 and cluster2, respectively.

$$similarity(cluster1,cluster2)= \frac{\sum cosine(\boldsymbol{d}_1,\boldsymbol{d}_2)}{size(cluster1)*size(cluster2)}$$

5. Results

In this paper we only compare the clustering algorithms when used to create hierarchical clusterings of documents, and only report results for the hierarchical algorithms and bisecting K-means. (The "standard" K-means and "flat" clustering results can be found in [6].) Figure 1 shows an example of our entropy results, which are more fully reported in [6]. Table 2 shows the comparison between F values for bisecting K-means and UPGMA, the best hierarchical technique. We state the two main results succinctly.

Bisecting K-means is better than regular K-means and UPGMA in most cases. Even in cases where other schemes are better, bisecting K-means is only slightly worse.

Regular K-means, although worse than bisecting K-means, is generally better than UPGMA. (Results not shown.)

6. Why agglomerative hierarchical clustering performs poorly

What distinguishes documents of different classes is the frequency with which the words are used. Furthermore, each document has only a subset of all words from the complete vocabulary. Also, because of the probabilistic nature of how words are distributed, any two documents may share many of the same words. Thus, we would expect that two documents could often be nearest neighbors without belonging to the same class.

Since, in many cases, the nearest neighbors of a document are of different classes, agglomerative hierarchal clustering will often put documents of the same class in the same cluster, even at the earliest stages of the clustering process. Because of the way that hierarchical clustering works, these "mistakes" cannot be fixed once they happen.

In cases where nearest neighbors are unreliable, a different approach is needed that relies on more global properties. (This issue was discussed in a non-document context in [3].) Since computing the cosine similarity of a document to a cluster centroid is the same as computing the average similarity of the document to all the cluster's documents [6], K-means is implicitly making use of such a "global property" approach.

We believe that this explains why K-means does better vis-à-vis agglomerative hierarchical clustering in the document domain, although this is not the case in some other domains.

7. Conclusions

This paper presented the results of an experimental study of some common document clustering techniques. In particular, we compared the two main approaches to document clustering, agglomerative hierarchical clustering and K-means. For K-means we used a standard K-means and a variant of K-means, bisecting K-means. Our results indicate that the bisecting K-means technique is better than the standard K-means approach and as good or better than the hierarchical approaches that we tested. In addition, the run time of bisecting K-means is very attractive when compared to that of agglomerative hierarchical clustering techniques - O(n) versus O(n^2).

Table 2: Comparison of the F-measure

Data Set	Bisecting K-means	UPGMA
re0	**0.5863**	**0.5859**
re1	**0.7067**	0.6855
wap	**0.6750**	0.6434
tr31	**0.8869**	0.8693
tr45	0.8080	**0.8528**
Fbis	**0.6814**	0.6717
la1	**0.7856**	0.6963
la2	**0.7882**	0.7168

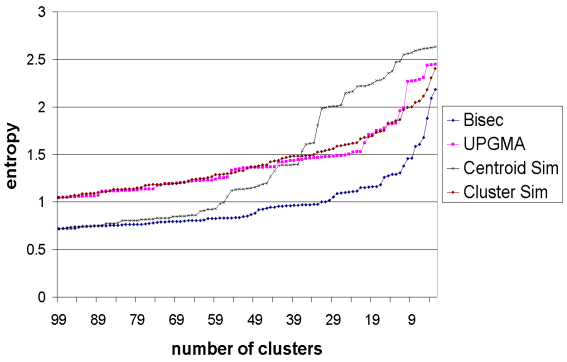

Figure 1: Comparison of entropy for re0 data set.

REFERENCES

[1] Cutting, D., Karger, D., Pedersen, J. and Tukey, J. W., "Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections," SIGIR '92, 318– 329 (1992).

[2] Dubes, R. C. and Jain, A. K., *Algorithms for Clustering Data*, Prentice Hall (1988).

[3] Guha, S., Rastogi, R. and Shim, K., "ROCK: A Robust Clustering Algorithm for Categorical Attributes," ICDE 15 (1999).

[4] Kowalski, G., *Information Retrieval Systems – Theory and Implementation*, Kluwer Academic Publishers (1997).

[5] Larsen, B. and Aone, C., "Fast and Effective Text Mining Using Linear-time Document Clustering," KDD-99, San Diego, California (1999).

[6] Steinbach, M., Karypis, G., Kumar, V., "A Comparison of Document Clustering Techniques," University of Minnesota, Technical Report #00-034 (2000). http://www.cs.umn.edu/tech_reports/