CSE494 Project Report - Part C

J. Benton

April 29, 2004

1 Introduction

In this project we compare two clustering algorithms, K-Means and Buckshot, on documents using vector space similarity as a distance measure. We will briefly discuss the design of the GUI before presenting the relationship between the number of clusters (K) and cluster quality.

2 Design and GUI

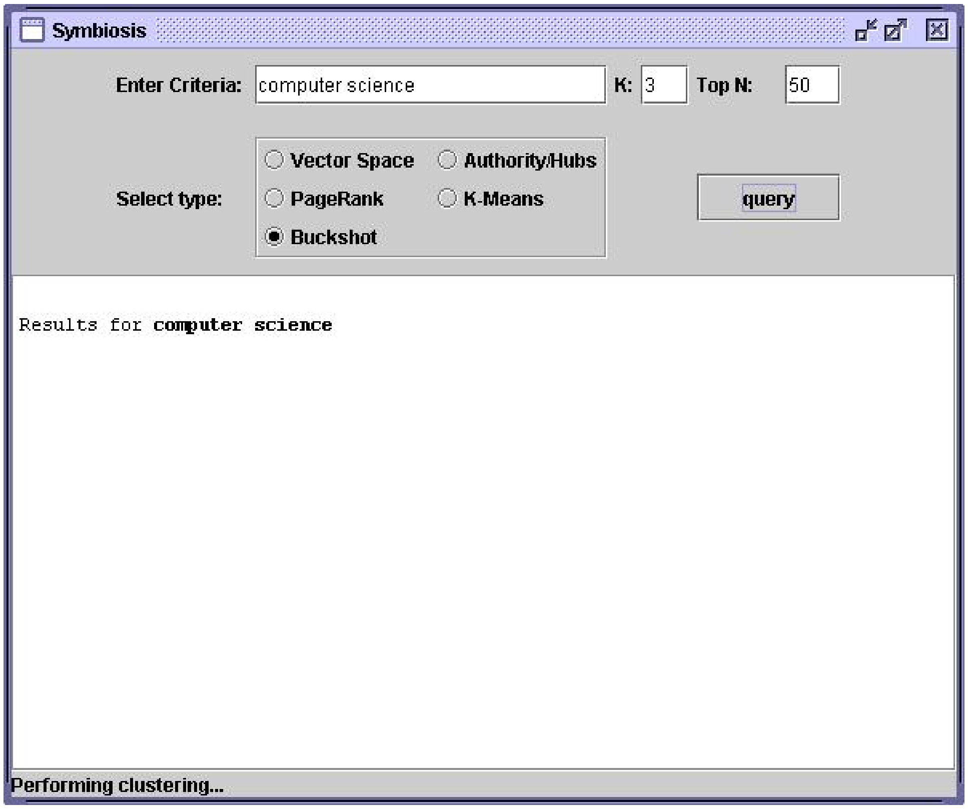
K-Means and Buckshot are each implemented as a separate Java class. Each is invoked with a single method call and requires search results from a vector space search as an input parameter. The executing classes indicate the current status of the clustering algorithm via a message that is pulled by the GUI and displayed on a status line.

Figure 1 illustrates the GUI design. When performing a query search (such as Vector Space or Authority/Hubs) a progress bar indicates the number of iterations performed in relationship to the number required. The progress bar is removed from view when it is impossible to determine the progression of an algorithm, such as in the case of the clustering algorithms. The bar is also made invisible when no algorithms are executing to allow more room to view results.

3 Analysis of K-Means

The K-Means algorithm requires a measure of similarity between documents to determine how to group them. For our project we used vector space similarity which is based upon the frequency of terms within a document.

The algorithm begins by chosing K documents at random, where K is the number of clusters we have chosen to obtain. Each document is assigned to a cluster. This process is called "seeding" and gives us starting documents for our clusters. Since this is random, different choices can lead to different final clusterings. Therefore, it is important that we take the *average* of several runs of the algorithm when measuring quality. So, for all of the experimental results we have taken the average of 3 runs for each metric.



Enter Criteria: computer science K: 3 Top N: 50

Select type: ○ Vector Space ○ Authority/Hubs ○ PageRank ○ K-Means ● Buckshot

```
Results for computer science
```

Performing clustering...

Figure 1: The GUI used for performing searches. Note the status bar indicates the current status of the search. Also, text boxes exist to change search parameters.

The quality metric on a set of clusters is measured depending upon what trait is most desired. One important trait in text documents is the similarity of one document to another. We measure the overall similarity by using the centroids of the clusters to compute the average similarity between all documents within the cluster (i.e. the square of the centroid's magnitude). This is shown for two queries in Figure 2. We see a steady reduction in quality as the number of clusters increases. Another way to measure the quality is by observation of each cluster. We can assume that a particular cluster refers to a subject by reviewing a sample of documents in that cluster. We can then measure the precision on the remaining documents. In our case, the clusters contained 100 documents in total during two executions of the algorithm for 3 and 7 clusters. We reviewed three documents per cluster to determine the topic. Figures 3 and 4 show that, at least in the case of our example execution, document precision tends to increase as the number of clusters increases over a particular set of documents.

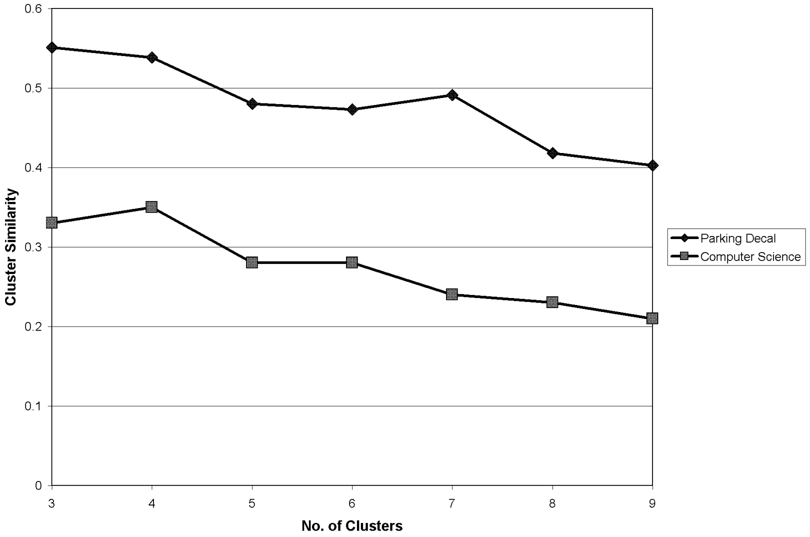


Figure 2: The average similarity between documents follows a steady decline as the number of clusters increases independent of the query being performed. Here we see the "computer science" and "parking decal" queries.

The difference between the calculated quality metric and the observed quality of the clusters is notable. The similarity between documents within clusters may be decreasing as the number of clusters increase, but the ability to split the topics into different categories provides for better precision across the clusters. It appears that clusters fall into natural categories when using K-Means. This may be because most of our web crawl covers only Arizona State University, where there is a particular format (e.g. menu options) for each department. Our results could differ if we included many different web sources.

4 Analysis of Buckshot

The Buckshot algorithm provides the K-Means algorithm with a set of K clusters consisting, in total, of $\sqrt{N}$ documents (where N is the number of documents to be clustered). These documents act as the seed for the K-Means algorithm and are found using Hierarchical Agglomerative Clustering (HAC). The idea is

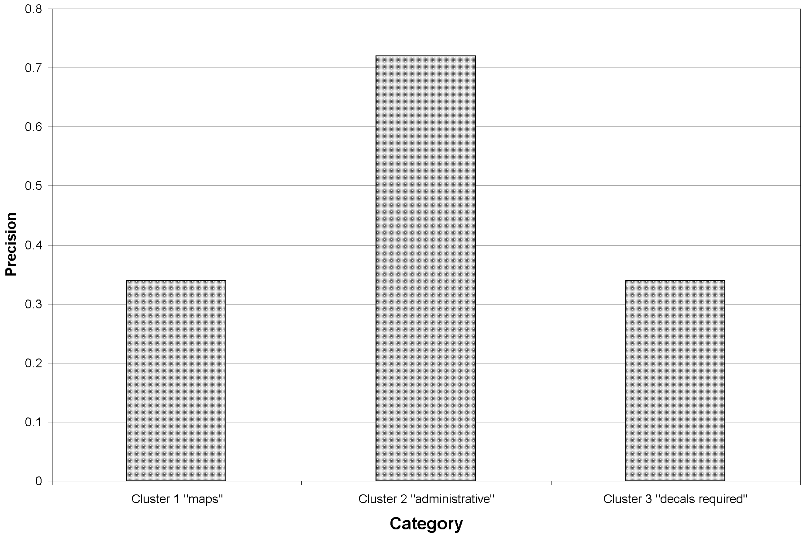


Figure 3: The top 100 “parking decal” documents were placed into 3 clusters. The above illustrates the precision on a subject matter determined after reviewing two sample documents from each cluster.

to avoid selecting bad seed documents such as outliers.

We use the same quality measures as in K-Means with varying cluster size. It is important to note that the effects of selecting a bad seed in the K-Means algorithm would be diminished by averaging similarity over runs. We note that there were no observed outliers during any of the runs. Considering this, Figure 5 shows that for one query, “parking decal”, the quality remains steady as the number of clusters increase. Compare this to K-Means (see Figure 2), where this query gives decreasing quality. Another query, “computer science”, has decreasing quality just as it does in K-Means.

The Buckshot algorithm also has the effect of providing greater precision with the “parking decal” query as the number of clusters increases. Figures 6 and 7 show a more pronounced increase in precision with the Buckshot algorithm when compared with that of K-Means (see Figures 3 and 4). Again, as with the K-Means analysis, this may be due to the structure of the crawled web pages. It is common practice to include the same terms per department. Many of the cluster topics relate directly to an Arizona State University department.

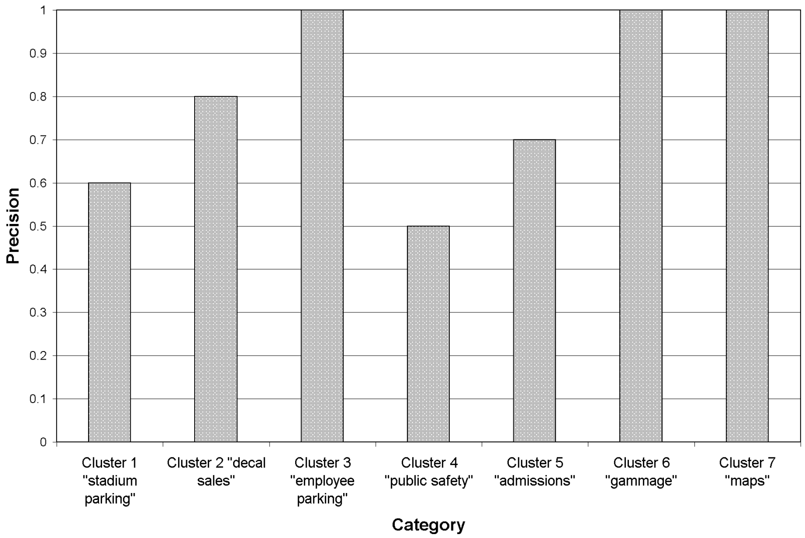


Figure 4: The top 100 “parking decal” documents were placed into 7 clusters. The above illustrates the precision on a subject matter determined after reviewing two sample documents from each cluster.

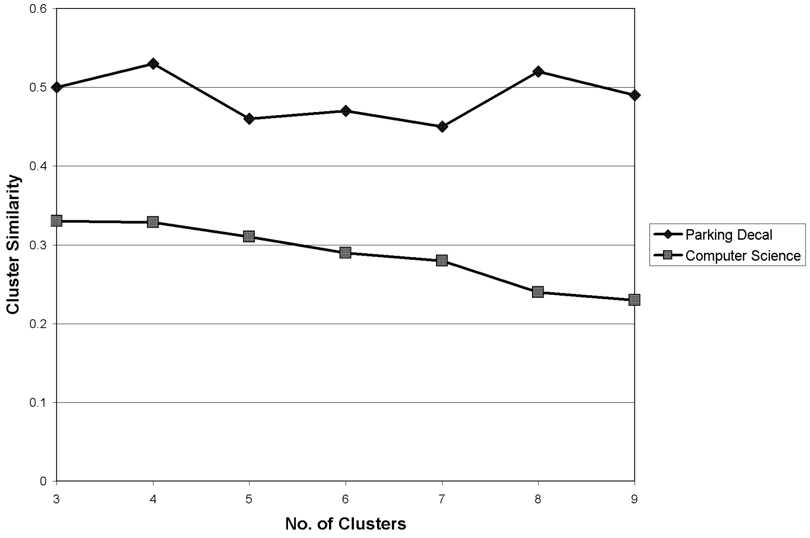


Figure 5: We see that "parking decals" quality remains steady while "computer science" falls in quality as the number of clusters increases.

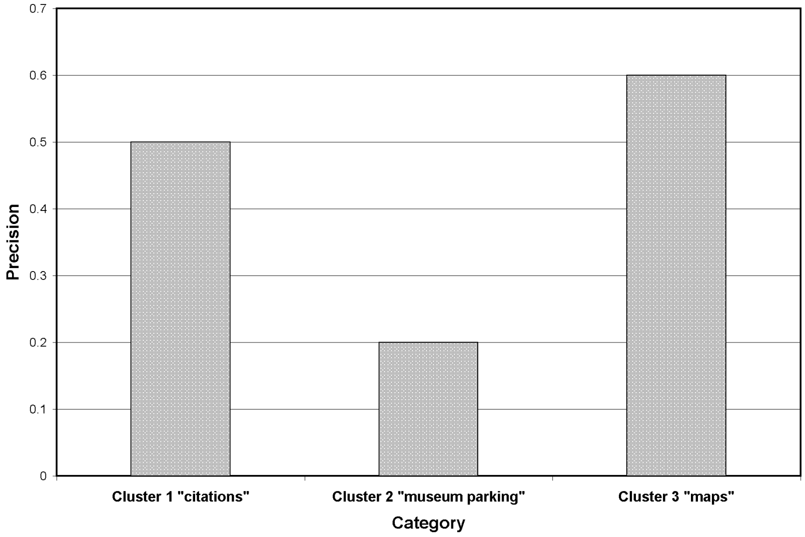


Figure 6: The top 100 “parking decal” documents were placed into 3 clusters. The above illustrates the precision on a subject matter determined after reviewing two sample documents from each cluster. (Using the Buckshot algorithm.)

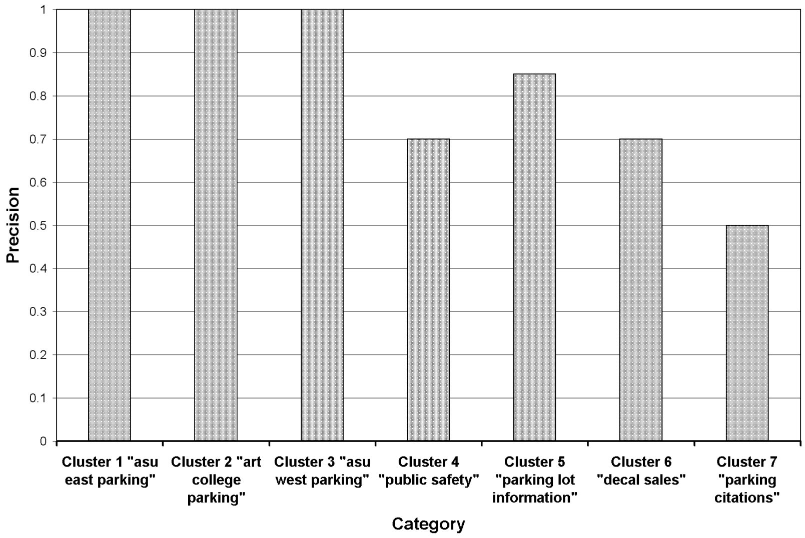


Figure 7: The top 100 “parking decal” documents were placed into 7 clusters. The above illustrates the precision on a subject matter determined after reviewing two sample documents from each cluster. (Using the buckshot algorithm.)