Patrick Trinkle
CMSC676
Spring 2009
Project 5 Write-Up

Background

Document clustering solves a set problems utilizing the information associated with the clustered documents. Clusters of documents represent documents that are similar in content. Similar document can often satisfy the same information need (or query). A corpus can be navigated topically by cluster; also, news stories can be clustered by topic and then routed to interested parties without human intervention.

When documents are represented via the vector space model, centroids are used as a simple way to represent a cluster of document. A centroid is typically calculated as the mean of the documents in the cluster; or the "center." Documents with a high similarity score to each other are clustered together. The similarity score algorithm we used is the cosine similarity, which is defined as the dot product of the document vectors divided by the products of the vector magnitudes.

Because the cosine similarity between documents is symmetric the similarity matrix is a triangle matrix and because the similarity of a document with itself is 1; it has no diagonal as this provides no useful information.

Implementation

To implement the document clustering system with a smaller memory footprint hash tables were used to both represent document vectors and store similarity scores; versus an actual matrix. Because the system always needs to merge the two documents (or clusters) with the highest similarity scores it would be optimal to store the scores in a max heap (binary tree).

Initially the dictionary and postings files are read into the system. From this data, document vectors for each document are built. These document vectors are then built into trivial centroids (centroids of vector count 1). All memory is freed from no longer needed data structures to reduce the memory footprint of this program once a resource is no longer needed.

As documents and centroids are merged they need to be removed from the matrix and the new centroid added to the matrix. This is fairly straightforward with hash tables, because we just remove the entries with the centroid names involved. With the keys from the similarity hash table it is fairly simple to determine which values to remove from the similarity matrix. It is also trivial to compute the new values without needing to update the entire table. The similarity scores are keyed as "doc1 + doc2" and the documents are named by their components in a comma separated list: "doc1,doc2,doc3." Because the information of each document is represented in a centroid the source documents can be removed and the memory freed.

A centroid-centroid merger involves two averages being averaged; each centroids weight in the merger is calculated from how many vectors each represents, thus not losing any information in the merger.

The documents were merged without a threshold value; therefore all document eventually were merged into one centroid, in the order of the most similar pairing. This last centroid represents the

average of the entire corpus. Interestingly enough if a threshold was used in merging clear clusters will appear in our corpus.

Conclusion

The program performs a somewhat minimalist approach to the number of calculations required however it can be optimized to reduce the number of comparisons required to find the maximum similarity. The total number of similarity calculations required is actually the minimum required for the process: 252,004. To complete processing on the 503 documents in the corpus took 2 minutes 41.65 seconds. The time required was not measured over a series of incrementally larger corpora because of the known time sink in the algorithm--it's slow.

A special centroid was created from the entire corpus. This centroid represented the center or average document in the corpus. The document in the corpus with the largest cosine similarity to this centroid was document 462 with a similiarity of 0.42388. There were duplicate documents in the corpus; documents 102 and 130. Because they are exact duplicates their similarity score was 1 and they were merged first. Because they are the same document; the centroid created represents either of them. A set of document s with a similarity score of 0 have no terms in common from our dictionary. Documents 005 and 068 have a score of 0.

As mentioned previously there are clear clusters in the corpus. The system output the details of each merger in order, and a quick glance reveals that certain sets of documents lend themselves to clustering. As listed below in a sample output, documents 340 through 370 all appear to be related.

Merging: 102 and 130 of similarity: 1
Merging: 434 and 435 of similarity: 0.99915577928128
Merging: 436 and 437 of similarity: 0.998318821368518
Merging: 433 and 434,435 of similarity: 0.994257886960587
Merging: 436,437 and 433,434,435 of similarity: 0.989991376964755
Merging: 443 and 441 of similarity: 0.983885330038762
Merging: 442 and 443,441 of similarity: 0.982616879931887
Merging: 440 and 439 of similarity: 0.975257968745463
Merging: 438 and 440,439 of similarity: 0.980299410513474
Merging: 442,443,441 and 438,440,439 of similarity: 0.959897168792568
Merging: 444 and 442,443,441,438,440,439 of similarity: 0.962478095746579
Merging: 065 and 081 of similarity: 0.896731816243516
Merging: 492 and 498 of similarity: 0.881412310076653
Merging: 346 and 368 of similarity: 0.850779217705977
Merging: 345 and 346,368 of similarity: 0.862582535386588
Merging: 370 and 345,346,368 of similarity: 0.870828478141997
Merging: 372 and 370,345,346,368 of similarity: 0.88190269679475
Merging: 360 and 372,370,345,346,368 of similarity: 0.890458817887435
Merging: 341 and 360,372,370,345,346,368 of similarity: 0.880309372857407
Merging: 351 and 341,360,372,370,345,346,368 of similarity: 0.880222822216505
Merging: 355 and 351,341,360,372,370,345,346,368 of similarity: 0.881123300662761
Merging: 362 and 355,351,341,360,372,370,345,346,368 of similarity: 0.880037187135309
Merging: 347 and 362,355,351,341,360,372,370,345,346,368 of similarity: 0.881884076629119
Merging: 358 and 347,362,355,351,341,360,372,370,345,346,368 of similarity: 0.869432736967697
Merging: 375 and 358,347,362,355,351,341,360,372,370,345,346,368 of similarity: 0.870425566128286

Merging: 367 and 375,358,347,362,355,351,341,360,372,370,345,346,368 of similarity: 0.871456735804187
Merging: 337 and 367,375,358,347,362,355,351,341,360,372,370,345,346,368 of similarity: 0.869890635072087
Merging: 343 and 337,367,375,358,347,362,355,351,341,360,372,370,345,346,368 of similarity: 0.868666442850909
Merging: 369 and 343,337,367,375,358,347,362,355,351,341,360,372,370,345,346,368 of similarity: 0.864741719191751
Merging: 349 and 369,343,337,367,375,358,347,362,355,351,341,360,372,370,345,346,368 of similarity: 0.852655398810971