

Project: Reordering of document ids using the Blelloch's approach to estimate the improvement in compression of index using gamma encoding.

Students:

1. Abhishek Seth (04329001)
2. Kaushal Mittal(04329024)
3. Sandesh Tawari(04329014)

We have implemented and tested the ideas presented in Blelloch's paper on document reordering.

Blelloch's Algorithm is about top down clustering of documents based on cosine similarity measure and reassigning the document ids such that the documents in the same cluster get near by doc ids.

We have used TFIDF scores for calculating the Center of mass and calculate (document,document) similarity using cosine measures. We have calculated the center of mass of a cluster of documents as an average of the TFIDF scores of documents in the cluster. But the downside of using TFIDF is that frequent terms get lower score and hence document pairs sharing these terms get low similarity scores. So we also experimented with use of Jaccard similarity measure instead of cosine measure.

Results

=====

Data set : `Reuters20578`

Similarity measure: Cosine

Corpus size (#documents)	Run1 (1100)	Run2 (1100)	Run3 (1100)	Run4 (1100)	Run1 (20578)	Run2 (20578)	Run3 (20578)
Inv-index size w/o any encoding	267KB	267KB	267KB	267KB	4883KB	4883KB	4883KB
Inv-index size w/ gamma encoding and using lucene assigned or random doc-ids	86KB	86KB	86KB	86KB	1990KB	1990KB	1991KB
Inv-index size w/ gamma encoding and using reordered doc-ids	79KB	80KB	80KB	81KB	1779KB	1802KB	1804KB

Similarity measure: Jaccard

Corpus size (#documents)	(1100)	(20578)
Inv-index size w/o any encoding	267KB	4883KB
Inv-index size w/ gamma encoding and using lucene assigned or random doc-ids	86KB	1991KB
Inv-index size w/ gamma encoding and using reordered doc-ids	82KB	1862KB

Data Set : `Newsgroup-18828`

Corpus size - 18822 documents

Without compressiion - 8941 KB

Compression with
Random doc ids - 3731 KB

Compression with
Reordered doc ids - 3305 KB

The improvement in the compression for the corpus with 18822 documents using the cosine similarity measure is 11.42 %.

Summary

Approximate improvement in compression(through gamma encoding and cosine measure) after document reordering for 1100 documents was 6.4% and for 20578 documents was 9.8%. It took on an average 7 minutes for 1100 documents and 3 hrs 58 mins for 20578 documents before optimisations. After code optimisation, use of ramfs for storing the temporary files generated for graphs, and using jaccard similarity, the running time reduced to 59 secs for 1100 documents and 53 minutes for 20578 documents.