Information Retrieval

María Carrasco Rodríguez (16874129)
Fabian Lindenberg (74076658)
Lea Voget (45869178)

Assignment 4

University of California Irvine

February 13, 2011

3
Index the data

We indexed 518,643 unique terms and compressed the index in the following way: First of all, we used delta encoding to encode the positions of a term within one document. That is, instead of saving the absolute positions of the terms, we stored the difference to the previous position. Therefore, we were able to store smaller numbers which needed less decimal digits.

Moreover, we used short document IDs instead of the long human readable document identifier strings. Then, we delta encoded the doc IDs as well. So, if a term occurred in three documents, we represented the ID of the second document as the difference to the first document and the third document's ID was encoded as the difference to the second document's ID.

This approach led to the necessity of a data structure mapping the document IDs to the actual string identifiers. This mapping has a size of 16,976 KB. Using the described compression techniques, we reduced the size of the index from 1,324,559 KB to 437,414 KB (including the mapping). Thus, although we needed to maintain a the document ID mapping, we were able to decrease the size of the index data by 65.7%.