Compression Report

Jeremy Greenfield <jsgreenf>
Alex Hills<ahills>

1. Nonpositional inverted indices are compressed via the following protocol: each line contains the first corresponding docID, then a space, then the value of the first docID subtracted from the value of the second docID, then a space, then the value of the third docID subtracted from the value of the second docID, and so forth for each query. Similarly, the positional inverted indices are compressed via the following protocol: docID1 position1 (position2-position1) (position3-position2)… (docID2-docID1) position1 (position2-position1) (position3-position2)… and so forth for each query. For example, The original line '1:1,3,6 2:1 6:1,5,20' translates to: '1 1 2 3 1 1 4 1 4 15'. Such schemes save space because the numbers needed to represent the docIDs are significantly small enough that less bytes are needed overall. Additionally, characters were encoded into bits via a Huffman encoding scheme. In this scheme, the document is analyzed and an encoding is created such that common characters are represented with the fewest number of bits. The bits are then packed together in a stream and parsed during querying and decompression.

2. Non-positional index: 15,871,567 bytes.
   Positional index: 61,105,142 bytes.
   Non-positional index ratio is about 22.76% and positional index ratio is about 31.17%. While the compression ratio for non-positional indices is higher, our encoding is particularly effective for compressing positional indices relative to other compression encodings/algorithms.
3. Non-positional index:
   a. 7zip: 13,568,351
   b. Bzip2: 17,122,370

   Positional index:
   c. 7zip: 69,788,951
   d. Bzip2: 74,423,056

   Our non-positional index was better than bzip2 but worse than 7z in terms of degree of compression. Our positional index was better than both bzip2 and 7zip. They can't be used for compressing postings lists because they cannot function as an inverted index (i.e. they cannot be queried without decompressing the entire file).

4. None at the moment.