Indexing

February 17, 2015

1 Documents and query representation

* term incidence matrix
    * About retrieval Models

1.1 bag of words representation

* TF, DF, DLength, AVG(DLength), V, N, IDF
    * subsection what and how to get from index

2 Preprocessing

2.1 Tokenization

2.2 Stopwords

2.3 Stemming

2.4 Term Positions

3 Index Construction

3.1 Inverted lists and catalog/offset files

3.2 Memory Structure, and limitations

3.3 option1: Multiple Passes

3.4 option2: Partial inverted lists

3.5 option3: preallocate the right amount of space

3.6 Updating an inverted index

4 Other things to store in the index

5 Proximity Search

virgil
$\sum_{i=1}^{5} a_i$

6 Compression

*probabilities as matching evidence

6.1 Basics of Compression, Entropy

6.2 Restricted Variable-Length Codes

6.3 Huffman codes

6.4 Lempel Ziv

7 Encoding integers

8 Distributed Indexing

9 Map-Reduce

10 Big Table

11 Query Processing