Indexing

March 18, 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

In Information Retrieval, it is often necessary to interpret natural text where a a large amount of text has to be interpreted, so that it is available as a full text search and is represented efficiently in terms of both space (document storing) and time (retrieval processes) requirements.

It can also be regarded as : process of incorporating a new document into an information retrieval system.

2.1 Tokenization

Tokenization is the task of chopping it up into pieces, called tokens, perhaps at the same time throwing away certain characters, such as punctuation. Input: "John_DavenPort #person 52 years_old #age"

John	DavenPort	person	52	years	old	age

2.2 Stopwords

Stopwords refer to the words that have no meaning for "Retrieval Purposes". E.g.

· **Articles** : a, an, the, etc.

· **Prepositions** : in, on, of, etc.

· **Conjunctions** : and, or, but, if, etc

· **Pronouns** : I, you, them, it, etc

· **Others** : some verbs, nouns, adverbs, adjectives (make, thing, similar, etc.).

Stopwords can be up to 50% of the page content and not contribute to any relevant information w.r.t. retrieval process. Removal of these can improve the size of the index considerably. Sometimes we need to be careful in terms of words in phrases! e.g.: Library of Congress, Smoky the Bear!

Word	Occurrences	Percentage
the	8,543,794	6.8
of	3,893,790	3.1

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