Reading Report - Week 2
CS 221 - Information Retrieval
Prof. Cristina Lopes

Varad Meru
Student # 26648958
Due Date: 01/17/2015

Simple, proven approaches to text retrieval*: A Summary

By S.E. Robertson, and K. Sparck Jones

Summary

The document lists some of the first techniques used in text retrieval. The discussion starts with the a description of the terms and the matching techniques. This section highlights the usage of stop words list and the use of stems rather than full words while building the index. Techniques such as Porter's stemming algorithm may be used for reducing the index size and help users find the better search results. Matching is done with consideration of all the query words and then decreasing subsets of the query words.

The discussion then moves to the topic of weights of the terms, which helps in selecting more relevant documents. Three techniques are used when deciding the weights of the terms: 1) Collection Frequency, 2) Term Frequency and 3) Document Length. The collection frequency helps highlighting terms that occur in fewer documents in the corpus. The collection frequency weight of a term is $\mathbf{CFW}(i) = log(N) - log(n)$, where n is the number of documents the term $t(i)$ occurs in, and N is the number of documents in the corpus. The term frequency for the term $t(i)$ in document $d(j)$ is $\mathbf{TF}(i,j)$ = number of occurrences of term $t(i)$ in document $d(j)$. The document length is $\mathbf{DL}(j)$ = the total number of term occurrences in document $d(j)$. The combined weight for one term $t(i)$ and one document $d(j)$ is given by (1). **K1** and b are tuning constants, tuning the effects of term frequency and document length respectively.

With the theory of term weights, the paper introduces the concept of Iterative Searching, with two major techniques of 1) Relevance weighs and 2) Query Expansion discussion in detail. Relevance weights is the relation between relevant and non-relevant documents for a search term modulated by its collection frequency. The relevance weight is $\mathbf{RW}(i) = log[((r+0.5)(N-n-R+r+0.5)/((n-r+0.5)(R-r+0.5))]$. All terms taken from relevant documents are ranked according to their Offer Weight $\mathbf{OW}(i) = r * \mathbf{RW}(i)$. Then the top terms are included in the search. The **RW** can replace the **CFW** to give Combined Iterative Weight, given by **CIW** in (2).

$$\mathbf{CW}(i,j) = \frac{\mathbf{CFW}(i) * \mathbf{TF}(i,j) * (\mathbf{K1}+i)}{\mathbf{K1} * ((1-b) + (b * (\mathbf{NDL}(j)))) + \mathbf{TF}(i,j)} \quad (1)$$

$$\mathbf{CIW}(i,j) = \frac{\mathbf{RW}(i) * \mathbf{TF}(i,j) * (\mathbf{K1}+1)}{\mathbf{K1} * ((1-b) + (b * (\mathbf{NDL}(j)))) + \mathbf{TF}(i,j)} \quad (2)$$

It is preferable to start with at least 5 terms to give some scope for the engine to eliminate much data. Also, for longer queries, i.e. ones in which stem may occur with different $\mathbf{QF}(i)$, then for each query term - document match compute the Query Adjusted Combined Weight or Query Adjusted Iterative Weight: $\mathbf{QACW}(i) = \mathbf{QF}(i) * \mathbf{CW}(i,j)$ or $\mathbf{QACIW}(i) = \mathbf{QF}(i) * \mathbf{CIW}(i,j)$. In some cases, it is better to index multi-word or compound terms. If a suitable lexicon is available, it can assist in building the inverted file. On the other hand, discovering it is complex task. But overall, it is always important to use multi-word terms, but such elaborations are tricky to manage and not recommended fro beginners.

Commentary

I learned a lot of techniques which can be applied in the context of the project for indexing and ranking documents based on search terms, queries and potential search results.

*Robertson, Stephen E., and Karen Sparck Jones. *Simple, proven approaches to text retrieval.* UK: Computer Laboratory, University of Cambridge, 1997.