Reading Report - Week 6 **Varad Meru**
CS 221 - Information Retrieval Student # 26648958
Prof. Cristina Lopes Due Date: 02/13/2015

A Vector Space Model for Automatic Indexing*: Summary

By Gerard Salton, Anita Wong, and Chung-Shu Yang. (Cornell University).

Summary

Authors start explaining the context of terms in a document and its representation in a t-dimensional vector space. Any document can be distributed into t-dimensional space where t is the maximum number of dimensions it can have. Once such an index vector for two documents is built, it is possible to find the similarity coefficient between them using the corresponding terms and weights. For example, a similarity measure might be the inner product of the two vectors, giving rise to the *Jaccard's or Tanimoto coefficient*, or alternatively an inverse function of the angle between the corresponding vector pairs; thus, calculating the *cosine coefficient*. In the absence of any special knowledge about the documents under consideration, considering the complete retrieval history of the given collection, optimum configuration is difficult to generate. In this case, authors suggest to minimize the function F given in (1).

$$F = \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} s(D_i, D_j), \quad c_j = (1/m) \sum_{\substack{i=1 \\ D_i \in K}}^{m} d_{ij}, \quad Q = \sum_{i=1}^{n} s(C^*, D_i) \tag{1}$$

where $s(D_i, D_j)$ is the similarity function between i and j. C denotes the centroid, with the structure $C = (c_1, c_2, ..., c_t)$, where t is the number of dimensions (or index terms).

The expression (1) is difficult to compute in practice since the number of vector comparisons is proportional to n^2 for a collection of n documents. Two things that are important for the success of a clustered document space are (a) the average similarity between pairs of documents within a cluster should be maximized, or vice-versa, the distance between the documents should be minimized and, (b) the average similarity between different cluster centroids be minimized.

The frequency of occurrence of each term k in document i, given by f_i^k can help in weighting. A better metric of weighting is multiplying the standard term frequency weight f_i^k by a factor inversely related to the document frequency d_k of the term k. (2) shows the mathematical views of IDF_k.

$$(IDF)_k = \lceil log_2 n \rceil - \lceil log_2 d_k \rceil + 1, \quad (IDF)_k = log(N/d_k) \tag{2}$$

A term weighting system proportional to $(f_i^k \cdot IDF_k)$ assigns largest weight to those terms which arise with high frequency in an individual document are rare across the collection. Further studies between the correlation between space density and indexing performance helps in understanding the transformations between terms and cluster of documents.

Commentary: This paper gave me a deep insight into the thinking behind information indexing and weighing while retrieving, especially in the point of view of clusters and terms. Got a first hand look on the functioning of the famed *tf-idf* and the cluster space model.

*Salton, Gerard, Anita Wong, and Chung-Shu Yang. *A vector space model for automatic indexing*, Communications of the ACM 18.11 (1975): 613-620.