1. Summary

With the background introduction about document space configurations, the paper discusses (1) correlation between indexing performance and space density, (2) correlation between space density and indexing performance. Last but not least, it presents a discrimination value model with transformation of terms (based on document frequencies).

A document could be identified by a vector of weighted index terms and a set of document vectors build a space. People are interested in optimum document space configuration for retrieval performance. Due to difficulty on anticipations of user relevance assessment, a suboptimal way is to achieve a maximum possible separation between documents. For practical measurement, a *clustered* document space with O(1) time complexity to measure space density is considered.

Then the existence of correlation between document space density and retrieval performance (and vice versa) is explored:

(1) for exploring the correlation between indexing performance and space density, 1) frequency of occurrence of a term in a document and 2) *inversed* document frequency of a term are used (together) for weighting of terms. Experiment on 2 different cluster organizations of document collections supports that improved (worsened) recall-precision performance is associated with decreased (increased) document space density.

(2) for investigating the correlation between space density and indexing performance, with the introduction of 1) skewness of occurrence frequency of a term for a cluster and 2) *inverse* of the number of clusters in which a term occurs, factors with spreading-out/compression effect are applied in index term weights to *artificially* modify document space density. And then it turns out that the outcome retrieval performance is inversely related with document space density.

According to above observation, term discrimination models is beneficial for retrieval performance by maintaining a document space of low similarity/density. “Discrimination value” which measures how well a term could decrease the similarity among documents offers a good point to determine index terms document vectors. With detailed investigation, it is found that best discriminators have neither too low nor too high document frequencies. Thus, to maintaining medium document frequencies, terms with very high document frequencies are transformed into phrases, while those with very low document frequencies are transformed into thesaurus. And these transformations lead to performance enhancement. Though lacking in conclusive proof, the proposed model performs well for document collections in several fields.

2. Comments

The paper shows strict and detailed work. It explores not only the correlations between indexing performance and space density, but also correlations between space density and indexing performance. Additionally, for each case, experiments are done for both forward and reverse proposition. And it also provides detailed explanations on experiment results. For example, in Section 2, it is mentioned that given better retrieval performance, the average similarity between the documents and the corresponding cluster centroids (factor *x*) and the average similarity between pairs of cluster centroids (factor *y*) are both decreased. However, there is *greater* “spreading out” of clusters than that of documents inside each cluster. Thus, the composite ratio *y/x* is decreased, implying smaller *overall* space density.

3. Questions

(1) Why different parameters/factors are used for Section 2 (based on documents) and Section 3 (based on clusters)? My explanation is that: weighting factor based on clusters might more significantly change the configuration (e.g., density) of a document space, which leads to more observable change in retrieval performance.

(2) How can we understand Figure 7? Specifically, why “Left-to-Right” (very low to medium document frequencies of terms) corresponds to “Recall Improving”? (Why “Right-to-Left” (very high to medium document frequencies of terms) corresponds to “Precision Improving”?) My (very possibly incorrect) understanding is that, using an unusual term as discriminator, then many relevant documents will be mistakenly classified into irrelevant classes because of lack of this term, and the recall becomes low. In contrast, if a term with very high document frequency is picked as a discriminator, a lot of irrelevant documents including this term will be retrieved due to the popularity of this term, which results in low precision. By shifting to medium document frequency, extreme cases should be greatly avoided.