**Title: The effects of indexing strategy-query term combination on retrieval effectiveness in a Swedish full text database**

**Abstract:**

This thesis deals with Swedish full text retrieval and the problem of morphological variation of query terms in the document database. The study is an information retrieval experiment with a test collection. While no Swedish test collection was available, such a collection was constructed. It consists of a document database containing 161,336 news articles, and 52 topics with four-graded (0, 1, 2, 3) relevance assessments.

The effects of indexing strategy-query term combination on retrieval effectiveness were studied. Three of five tested methods involved indexing strategies that used conflation, in the form of *normalization*. Further, two of these three combinations used indexing strategies that employed *compound splitting*. Normalization and compound splitting were performed by SWETWOL, a morphological analyzer for the Swedish language. A fourth combination attempted to group related terms by right hand truncation of query terms. A search expert performed the truncation. The four combinations were compared to each other and to a baseline combination, where no attempt was made to counteract the problem of morphological variation of query terms in the document database.

Two situations were examined in the evaluation: the *binary relevance situation* and the *multiple degree relevance situation*. With regard to the binary relevance situation, where the three (positive) relevance degrees (1, 2, 3) were merged into one, and where precision was used as evaluation measure, the four alternative combinations outperformed the baseline. The best performing combination was the combination that used truncation. This combination performed better than or equal to a median precision value for 41 of the 52 topics. One reason for the relatively good performance of the truncation combination was the capacity of its queries to retrieve different parts of speech.

In the multiple degree relevance situation, where the three (positive) relevance degrees were retained, retrieval effectiveness was taken to be the accumulated gain the user receives by examining the retrieval result up to given positions. The evaluation measure used was nDCG (normalized cumulated gain with discount). This measure credits retrieval methods that (1) rank highly relevant documents higher than less relevant ones, and (2) rank relevant (of any degree) documents high. With respect to (2), nDCG involves a discount component: a discount with regard to the relevance score of a relevant (of any degree) document is performed, and this discount is greater and greater, the higher position the document has in the ranked list of retrieved documents.

In the multiple degree relevance situation, the five combinations were evaluated under four different *user scenarios*, where each scenario simulated a certain user type. Again, the four alternative combinations outperformed the baseline, for each user scenario. The truncation combination had the best performance under each user scenario. This outcome agreed with the performance result in the binary relevance situation. However, there were also differences between the two relevance situations. For 25 percent of the topics and with regard to one of the four user scenarios, the set of best performing combinations in the binary relevance situation was disjunct from the set of best performing combinations in the multiple degree relevance situation. The user scenario in question was such that almost all importance was placed on highly relevant documents, and the discount was sharp.

The main conclusion of the thesis is that normalization and right hand truncation (performed by a search expert) enhanced retrieval effectiveness in comparison to the baseline, irrespective of which of the two relevance situations we consider. Further, the three indexing strategy-query term combinations based on normalization were almost as good as the combination that involves truncation. This holds for both relevance situations.

**Keywords:** base word form index, full text retrieval, indexing strategies, inflected word form index, morphological analysis, normalization, Swedish, SWETWOL, truncation, user scenarios