

# Advanced Approaches to the Statistical Analysis of TREC Information Retrieval Experiments

**David A. Hull**

**Paul B. Kantor, Kwong Bor Ng**

Rank Xerox Research Centre  
6 Chemin de Maupertuis  
38240 Meylan, France  
hull@xerox.fr

APLab, SCILS, Rutgers University  
4 Huntington St.  
New Brunswick, NJ 08903 USA  
kantorp@cs.rutgers.edu

August 13, 1997

## **Executive Summary**

This document is a progress report on a two-pronged study of the potential for the TREC procedure as an instrument for measuring the performance of text retrieval systems, in much the same way that one calibrates measuring devices or other scientific instruments. In Part I of the report we concentrate on the issue of detecting meaningful differences in a specific measure of performance, the average precision of retrieved ranked lists of documents. In Part II, which will be transmitted separately, we concentrate on the issue of detecting and measuring similarity (or difference) between retrieval systems, without regard to any choice of measures of performance.

Resolving power is expressed in terms of the smallest difference between systems that can be reliably detected by the procedure investigated. Because the TREC situation involves what are presumed to be a random selection of retrieval tasks (topics) from some much larger universe of such tasks, the assessment of resolving power necessitates the use of statistical reasoning. Because, in fact, TREC examines a number  $S$  of systems simultaneously, the statistical issues involved lead to an embarrassment of riches. There are many ways of addressing the issue, and, as might be expected, they lead to different conclusions. Statistical theory does not, perhaps surprisingly, lead to a clear preference for one method over another.

Broadly the methods can be placed in three classes. Most well-developed are the parametric tests, which assume that the effects of both topic and system are strictly additive, when precision is used as the score, and which assume that all excursions from the predictions of this additive model are drawn from a single, normal distribution which describ