Evaluation Measures Scales: From Theory to Experimentation

Nicola Ferro

University of Padua

Abstract. Evaluation measures are the basis for quantifying the performance of IR systems and the way in which their values can be processed to perform statistical analyses depends on the scales on which these measures are defined. For example, mean and variance should be computed only when relying on interval scales.
In this talk, we will present our formal theory of IR evaluation measures, based on the representational theory of measurements, to determine whether and when IR measures are interval scales. We found that common set- based retrieval measures – namely Precision, Recall, and F-measure – always are interval scales in the case of binary relevance while this does not happen in the multi-graded relevance case. In the case of rank-based retrieval measures – namely AP, gRBP, DCG, and ERR – only gRBP is an interval scale when we choose a specific value of the parameter p and define a specific total order among systems while all the other IR measures are not interval scales. We will also introduce some brand new set-based and rank-based IR evaluation measures which ensure to be interval scales.
In our previous work we defined a theory of IR evaluation measures, based on the representational theory of measurement, which allowed us to determine whether and when IR measures are interval scales. Finally, we will discuss the outcomes of an extensive evaluation, based on standard TREC collections, to study how our theoretical findings impact on the experimental ones. In particular, we conduct a correlation analysis to study the relationship among the above-mentioned state-of-the-art evaluation measures and their scales. We study how the scales of evaluation measures impact on non parametric and parametric statistical tests for multiple comparisons of IR system performance.

Biography. Nicola Ferro is associate professor in computer science at the University of Padua, Italy. His research interests include information retrieval, its experimental evaluation, multilingual information access and digital libraries. He is the coordinator of the CLEF evaluation initiative, which involves more than 200 research groups world-wide in large-scale IR evaluation activities. He was the coordinator of the EU Seventh Framework Programme Network of Excellence PROMISE on information retrieval evaluation. He is associate editor of ACM TOIS and was general chair of ECIR 2016.