Retrieval Evaluation in Practice

Ricardo Baeza-Yates

Yahoo! Research and Universitat Pompeu Fabra
Barcelona, Spain
rbaeza@acm.org

Abstract. Nowadays, most research on retrieval evaluation is about comparing different systems to determine which is the best one, using a standard document collection and a set of queries with relevance judgements, such as TREC. Retrieval quality baselines are usually also standard, such as BM25. However, in an industrial setting, reality is much harder. First, real Web collections are much larger – billions of documents – and the number of all relevant answers for most queries could be several millions. Second, the baseline is the competition, so you cannot use a weak baseline. Third, good average quality is not enough if, for example, a significant fraction of the answers have quality well below average. On the other hand, search engines have hundreds of million of users and hence click-through data can and should be used for evaluation.

In this invited talk we explore important problems that arise in practice. Some of them are: Which queries are already well answered and which are the difficult queries? Which queries and how many answers per query should be judged by editors? How we can use clicks for retrieval evaluation? What retrieval measure we should use? What is the impact of culture, geography or language in these questions?

All these questions are not trivial and depend in each other, so we only give partial solutions. Hence, the main message to take away is that more research in retrieval evaluation is certainly needed.