QaDAIRE: Quantitative Data Analysis for Information Retrieval Evaluation

Supervisors: Lorraine Goeuriot, Philippe Mulhem

Objectif :

Information retrieval has a strong tradition of experimental evaluation, and for decades the most widely-used evaluation approach has been the Cranfield methodology, a system-oriented method where effectiveness is measured using test collections [1]. Test collections are standardized and re-usable resources consisting of a representative set of search topics, a document collection to be searched, and a set of relevance assessments indicating whether documents are responsive to a search topic. The evaluation and relative ranking of systems, given a test collection, is then guided by the use of various effectiveness metrics.

The performances of a system is directly linked to the relevance of returned documents. Specifically, the relevance assessments are manual judgements of the relevance of a given document to a certain topic. While relevance has been studied for a long time, both in the information retrieval and information science communities, it has been shown to be a complex concept, in particular with regards to operationalizing relevance for practical IR system evaluation.

EIRAP project aims at investigating the measurement of relevance, both from the assessor and the system point-of-view. A variety of methods for the measurement of relevance have been proposed in previous work, and their applicability for IR evaluation has been demonstrated. However, little attention has been paid to the practicalities of making the relevance judgments in these different modes. We argue that this is a vital dimension to consider, as the gathering of relevance judgments is the principle (and often very large) cost that is incurred when creating a test collection.

This project will investigate three relevance judging techniques:

1. *Ordinal relevance.* Absolute graded judgements, where a relevance score is given to each topic-document pair on a pre-defined scale. We plan to use a 5-level scale, in line with recent evaluation campaigns such as the TREC Web Tracks.
2. *Magnitude estimation.* Judges assign scores to topic-document pairs based on their subjective perception of the level of relevance. Judgements are made for multiple documents in response to a topic, such that the ratio between assigned numbers reflects the ratio of the rater's perception of differing relevance content.
3. *Preference judgements.* Judges give preference orderings to indicate which document in a pair is more relevant with regard to the search topic being considered.

The purpose of the internship is a quantitative comparison between the three assessment techniques, where we will analyse these measurements in terms of:

- time taken to perform the assessment;
- inter annotator agreement over the 3 methods;
- consistency of the three methods with the official TREC judgments;
- impact of using the different methods on the relative ranking of IR system effectiveness, using runs that participated in the TREC Web Track;
- comparison of two assessors groups: supervised in-lab judges, and crowd-sourced workers.

We expect that the results of this project will offer a substantial contribution towards a better understanding of the relevance judging process, and in particular of the relative costs, both cognitive and economic, of the various relevance judging techniques, and the type of judges (supervised or crowd-sourced).

[1] Cleverdon, C.W. (1967). The Cranfield tests on index language devices. Aslib Proceedings, 19(6), pp. 173-192

Competencies required: strong programming skills, statistics, information retrieval