Inter-assessor agreement at INEX 06

Nils Pharo[1], Andrew Trotman[2], Shlomo Geva[3], Benjamin Piwowarski[4]

[1]Faculty of Journalism, Library and Information Science, Oslo University College, Norway
nils.pharo@jbi.hio.no
[2]Department of Computer Science, University of Otago, Dunedin, New Zealand
andrew@cs.otago.ac.nz
[3]Faculty of Information Technology, Queensland University of Technology, Australia
s.geva@qut.edu.au
[4]Department of Computer Science, University of Chile
bpiwowar@dcc.uchile.cl

Extended abstract

A basic requirement of IR test collection is to have a pool of documents with relevance assessments. Such relevance assessments are typically performed by topic experts. The procedure at INEX is having the participating groups submitting topics which are later returned to topic creators for relevance assessment. Usually the participants perform relevance assessments of their own topics, but some topics are assessed by more than one assessor.

At the INEX 06 workshop an experiment was conducted in order to measure inter-assessor-agreement. Since the INEX 06 collection consists of articles from Wikipedia it is relatively easy to design topics (simulated work tasks) of general interest, thus making it easier for other than topic creators to perform relevance judgments. Participants at the workshop were asked to perform relevance assessments of a selection of topics, which were logged in the online assessment system. The assessors also answered questionnaires related to the experiment.

The main purpose of this experiment was to measure the level of agreement between many assessors judging the same topics. We present the findings and relate them to background factors.

The findings from our analysis can be used for several purposes: 1) they can help us design the procedures for relevance assessments in future INEX experiments; 2) they provide a valuable understanding of the general factors influencing relevance judgements used in test collections; and 3) they can be used to argue for the validity of using the test collection approach as a method for testing the retrieval efficiency of IR systems.