Inex 2003 Working group report: Relevance

Jaana Kekäläinen
Dept. of Information Studies
33014 University of Tampere
Finland

jaana.kekalainen@uta.fi

In Inex 2003, relevance judgements were based on exhaustivity and specificity as dimensions of topical relevance. Both these dimensions were assessed on 4-point scale (i.e., not, marginally, fairly, highly specific / exhaustive; see [1]). This definition of relevance was chosen to suit the need to retrieve and rank elements of different granularity typical for structured document retrieval. The workshop, which attracted about 20 people, was not so much concerned with the concept or definition of relevance; rather the consequences of the chosen relevance definition on the assessment process were discussed. The practical experiences participants had on working with relevance assessments played a vital role in discussions. Four main themes came up during the sessions:

- How useful the dimensions of relevance are?
- What is the least meaningful unit to be assessed for relevance?
- Are the relevance assessments reliable?
- What is the validity of the assessment of VCAS and SCAS topics?

First, the issue of judging relevance along dimensions of exhaustivity and specificity was raised. The argument against these dimensions, and dimensions in general, was that it would be easier for the assessor to give only one relevance figure for each element to be assessed. This especially in case the used metric returns only one performance figure. Another opinion – which gained more support – was that the named dimensions help the assessor to become aware of the factors affecting the assessment, and thus help him to be more consistent. Should more dimensions of relevance be considered in assessment? Perhaps, but the question is how many balls the assessor can play with? This ballgame is still in the area of topicality.

Second, many of the participants were frustrated when assessing relevance of some minor elements that cannot really carry relevant meaning alone (e.g. article number or references). This was due to the assessment system forcing to judge all ascendant / descendant elements of any relevant element. This procedure is in accordance with relevance assessment rules which try to ascertain that *all* relevant elements are identified. However, the general opinion was strongly for making a list of elements that should neither be retrieved alone nor judged for relevance. Another argument in this discussion was that some elements could not be judged alone because 'a whole can be more than its parts'. Here the solution seemed to be that an element should be assessed on the basis of its relevance as an alone standing unit. This debate also touched upon the rules for assessment consistency in the online assessment tool.

Third, the consistency and reliability of relevance assessments were considered. Some participants thought that elements, which should be relevant, were judged non-relevant, i.e. they were not missed in assessment process but they were consciously assessed non-relevant. In the discussion it was obvious that people with different background had different understanding about the relevance that should be used. Those active in information retrieval were for topicality, but those working with DBMS were for system relevance (for manifestations of relevance, see [2]). This issue could not be agreed upon, yet the workshop made a suggestion for getting multiple relevant assessments for some topics in order to check the consistency of assessments. Later on it turned out that there already are multiple assessments for some topics, only the analysis of consistency is lacking.

Fourth, what is the role of 'vagueness' and 'strictness' in relevance assessment of content and structure (CAS) queries? This question seemed to divide opinions and practices: others had tried to assess the relevance according to whether the structural conditions were met or not; others had ignored the structural conditions because they were difficult to check. The relevance assessment guide gives support to both interpretations (see [1]). The whole matter is even more complicated because it is not quite clear how to implement 'vagueness' in retrieval and evaluation. The organizers investigate this matter.

The workshop made some suggestions for the INEX projects to come:

- It could be useful to re-use the old topics later on – with new / elaborated systems – to see whether any progress is made.
- The number of topics should be raised for better reliability of data. This, however, should be achieved without increasing the assessment load for individual groups. Two obvious possibilities were suggested: the number of participants could be higher, and the evaluation task could

be made easier (for example, by the list of elements not to retrieve / assess).

1. REFERENCES

[1] Kazai, G., Lalmas, M. & Piwowarski B. (2004). INEX'03 Relevance assessment guide. In *INEX 2003 Workshop PreProceedings*, 154-159. Available at: http://inex.is.informatik.uni-duisburg.de: 2003/.

[2] Saracevic, T. (1996). Relevance reconsidered '96. In P. Ingwersen & N. O. Pors (Eds.), *Proceedings of the Second International Conference on Conceptions of Library and Information Science: Integration in Perspective*. Copenhagen: The Royal School of Librarianship, 201–218.