A Comparison of Interactive and Ad-Hoc Relevance Assessments

Birger Larsen[1], Saadia Malik[2], and Anastasios Tombros[3]

[1] Royal School of Library and Information Science, Denmark
blar@db.dk
[2] University Duisburg-Essen, Germany
malik@is.informatik.uni-duisburg.de
[3] Queen Mary University of London, UK
tassos@dcs.qmul.ac.uk

Abstract. In this paper we report an initial comparison of relevance assessments made as part of the INEX 2006 Interactive Track (itrack'06) to those made for the topic assessment phase of the INEX 2007 ad-hoc track. The results indicate that that there are important differences in what information was assessed under the two different conditions, but it also suggests a certain level of agreement in what constitutes relevant and non-relevant information. In addition, there are indications that the task type has an influence on the distribution of relevance assessments.

1 Introduction

In this paper, we report on a comparison of relevance assessments made as part of the INEX 2006 interactive track [7] (itrack'06) and those made as part of the topic assessment phase for the INEX 2007 ad-hoc track. Our analysis is based on eight topics that were assessed as part of both tracks.

The conditions under which the eight topics were assessed were significantly different, with searchers in itrack'06 assessing the usefulness of elements in addressing information seeking tasks, while topic assessors for the ad-hoc track focused on providing comprehensive assessments for each retrieved document. These different conditions provide the main motivation for carrying out this research. More specifically, we are primarily interested in investigating:

- The extent to which the different conditions affect the relevance of document elements, as perceived by itrack'06 searchers and ad-hoc topic assessors.
- The overlap of the assessed information, i.e. to what extent the information that searchers and assessors perceived as being useful in their respective tasks was similar.

In addition, the eight topics used in the study are classified into different task types [7,12], providing thus the opportunity to also study the effect of different topic types. Further, in itrack'06 two versions of an XML IR system were used (more details in section 2.1 and in [7]), allowing us to also study the effect of system type perception of document element relevance.

N. Fuhr et al. (Eds.): INEX 2007, LNCS 4862, pp. 348–358, 2008.

There has been significant work on the study of relevance assessments and agreement between assessors in the context of the Text Retrieval Conferences - TREC [1, 9, 13, 14, 15]. The main emphasis has been on binary relevance assessments, since this has been the basis for evaluation in TREC. In one of the few studies that have used multi-scale relevance assessments, Voorhees [14] used the TREC-9 web track data and a three-point relevance scale (not relevant, relevant, highly relevant) in order to examine the effect in evaluation stability of considering only highly relevant documents. Voorhees found that there is a negative effect on stability by the consideration of only highly relevant documents.

Most of the past work on relevance assessments in the context of TREC has also focused relevance judgements made by the TREC assessors, not by online searchers. Some exceptions involve interactive searching and judgement are the work by Cormack et al. [1], and Sanderson and Joho [9], interactive searching, judging and query reformulation are used for forming relevance assessments. In the study by Cormack et al., it was reported that an agreement level of 40% existed between relevance assessments made by interactive searching and by TREC assessors. Voorhees [13] has also examined inter-assessor agreement for a subset of the TREC-4 data (only between TREC assessors), and found similar levels of agreement. Inter-assessor agreement has generally been considered a problem area in IR evaluation in the context of TREC.

In the remaining of this paper, we first describe some methodological issues in section 2, we then present some initial results and analysis in section 3, and we conclude and outline our further plans for analysis in section 4.

2 Methodology

In this section we describe the methodology of our study. First in sections 2.1 and 2.2 we briefly summarise the frameworks under which relevance assessments were made for itrack'06 and the INEX 2007 ad-hoc track, respectively, and in section 2.3 we discuss the methodology by which the assessments in the two tracks were compared.

2.1 Interactive Track 2006

In the INEX 2006 interactive track (itrack'06) searchers from various participating institutions were asked to find information for addressing information seeking tasks by using two interactive retrieval systems: one based on a Passage retrieval backend[1] and one on an Element retrieval backend[2]. Both versions had similar search interfaces but differed in the returned retrieval entities: The passage retrieval backend returned non-overlapping passages derived by splitting the documents linearly. The element retrieval system returned elements of varying granularity based on the hierarchical document structure. The frontend was a modified version of the Daffodil system [3],

[1] The Passage retrieval backend was based on CSIRO's Panoptic™/Funnelback™ platform. See http://www.csiro.au/csiro/content/standard/pps6f,,.html for more information.

[2] The Element retrieval backend was based on Max Planck Institute for Informatics' TopX platform. See [11] for more information.

and the document collection used was the INEX Wikipedia corpus [2]. For a full description of the systems used in itrack'06 the reader can refer to [6].

Twelve search tasks of three different types [12] (*Decision making*, *Fact finding* and *Information gathering*), further split into two structural kinds (*Hierarchical* and *Parallel*), were used in the track [7]. The tasks were split into different categories allowing the searchers a choice between at least two tasks in each category, and at the same time ensuring that each searcher will perform at least one of each type and structure.

An important aspect of the study was to collect the searcher's assessments of the relevance of the information presented by the system. We chose to use a relevance scale based on work by Pehcevski et al. [8]. Searchers were asked to select an assessment score *for each viewed piece of information* that reflected the usefulness of the seen information in solving the task. Five different scores were available, expressing two aspects, or dimensions, in relation to solving the task: How much ***relevant information*** does the part of the document contain, and how much ***context is needed*** to understand the element? This was combined into five scores as follows:

- **Not relevant (NR).** The element does not contain any information that is useful in solving the task
- **Relevant, but too broad (TB).** The element contains relevant information, but also a substantial amount of other information
- **Relevant, but too narrow (TN).** The element contains relevant information, but needs more context to be understood
- **Partial Relevant answer (PR).** The element has enough context to be understandable, but contains only partially relevant information
- **Relevant answer (R).** The element contains highly relevant information, and is just right in size to be understandable.

In the interactive track, the intention is that each viewed element should be assessed with regard to its relevance to the topic by the searcher. This was, however, not enforced by the system as it may be regarded as intrusive by the searchers [6]. Note that in contrast to the assessments made for the ad-hoc track, there is no requirement for searchers to view each retrieved element as independent from other components viewed. Experiences from user studies clearly show that users learn from what they see during a search session. To impose a requirement for searchers to discard this knowledge would create an artificial situation and will restrain the searchers from interacting with the retrieved elements in a natural way.

Overall, 88 interactive track searchers made 2170 relevance assessments for the eight tasks analysed in this paper. Table 1 in Section 3 gives a detailed account of this data.

2.2 INEX 2007 Ad-Hoc Assessments

The purpose of the INEX 2007 ad-hoc track is to create a test collection consisting of a corpus of documents, a set of questions directed at the documents (called topics) and a set of relevance assessments specifying which documents (or the elements that are part thereof) that are relevant to each topic [4]. The elements to be assessed were

identified by pooling the output of multiple retrieval systems following the method first proposed in [10]; the pool of retrieved elements for each topic was then assessed by the topic author.

In INEX 2007 the assessment process focussed on the notion of specificity, that is, the extent to which the element focuses on the information need expressed in the topic [4]. A highlighting approach was taken, where the assessor first skims the document and then highlights any parts that contain only relevant information. From this, the specificity of any element with highlighted content can be calculated automatically. This may be done by computing the ratio of relevant content (rsize) to all content (size), measured in the number of characters.

All twelve topics that were used in itrack'06 were also submitted as topics for the ad-hoc track. Up to the point of writing this paper, full assessments for eight of these topics were available – we use these as the basis of our result presentation and analysis in section 3.

2.3 Mapping Ad-Hoc and Interactive Track Assessments

Whereas the interactive track assessments are given in terms of one of the five categories in section 2.1, the ad-hoc assessments are of a continuous nature. Thus a mapping between them is needed for comparisons. As mentioned above, there was a difference in the scope of the two types of assessments: where the ad-hoc track aimed at getting comprehensive assessments for each retrieved document, the interactive track searchers were free to assess as much or as little information as they saw fit. In addition, no attempt was made to control learning effects across a search session in the interactive track, while ad-hoc assessors were explicitly asked to assess each element on its own merit.

In the interactive track, non-relevant elements could be specified explicitly (by selecting the NR assessment), as well as implicitly (by searchers viewing an element but not giving any assessment). As such, there is a good correspondence with the ad-hoc track, where only relevant information was highlighted and the rest ignored.

The notion of relevant information (R) in the interactive track would correspond in the ad-hoc assessments to elements that are either fully highlighted or have a large ratio of highlighted content, for example elements with more than 75% relevant content might be considered as being relevant. Following the same line of argument, the interactive track notion of Too Broad (TB) would correspond to elements that in the ad-hoc assessments have a relatively small amount of highlighted content, for example, elements with less than 25% relevant content might be considered as being Too Broad.

It is, however, more difficult to identify a direct parallel to the notion of Too Narrow (TN) in the ad-hoc assessment data. It might be argued though that it is unlikely that small elements would have been relevant to the itrack'06 topics. Pragmatically, such small elements can be filtered out by excluding elements smaller than a given absolute size, e.g., 125 characters[3]. A similar reasoning based on absolute size could

[3] Based on that a typical sentence length in English text is around 125 characters (http://hearle.nahoo.net/Academic/Maths/Sentence.html).

be applied as a supplemental criterion to the notion of Relevant (R): elements that contain, e.g., 500 characters of highlighted content could be deemed Relevant, regardless of the ratio of highlighted content.

The notion of Partial Relevant Answer (PR) is also difficult to translate to the ad-hoc assessments, because only relevant information was highlighted in the assessment process.

3 Results and Analysis

In the interactive track 88 searchers were recruited by 8 research groups, and overall they completed 334 search sessions[4]. Table 1 presents some basic statistics for the assessments provided as part of itrack'06. For the eight topics analysed in the present paper, 2170 elements were assessed. As different searchers would often assess the same elements for the same topic, the number of unique assessed elements was 1039 (an average of 2.1 assessments per element). For 177 of these uniquely assessed elements, two or more different assessments (e.g. R, TB and TB) were given by searchers. These present a particular challenge in our study, because we need to arrive at a single assessment for each element in order to compare it to the ad-hoc assessments.

Table 1. Basic statistics on the relevance assessments provided by the INEX 2006 interactive track searchers (including elements that were viewed, but not assessed)

Total number of assessments (including elements assessed more than once)	2170
Unique elements assessed	1039
Unique elements with two or more different assessments	177

In Table 2, we provide details about how these different assessments are distributed among the 1039 uniquely assessed elements. Both rows and columns list the relevance categories and the table shows how many elements have been assessed under both categories by any number of different searchers. There are for instance 57 elements that have been assessed both as Relevant and as Too Broad.

The distribution of values in Table 2 is fairly uniform, with the maximum value being the 10% of the elements marked as NA and R. This largest value corresponds to searchers viewing, but not assessing (NA), elements that other searchers had assessed as relevant. Overall, elements that were not assessed by some searchers but were assessed by other searchers (i.e. the NA row) correspond to the largest percentage in Table 2. Elements assessed as non-relevant (NR) are noteworthy as they correspond to cases where searchers have explicitly indicated that the elements are particularly ill-fitted to the topic. Elements assessed as non-relevant overlap with relevant of any category in 3-5% of the cases. In the heuristics applied to derive a single assessment for the 177 elements, special weight is given to those that were explicitly assessed as non-relevant.

[4] Due to system problems, logs of some search sessions had to be excluded.

Table 2. Details of how different assessments are distributed among document elements in raw counts (left) and percentages over the 1039 unique assessed elements (right)

	R	NA	NR	PR	TB	TN
R	-	103	52	68	57	36
NA	103	-	77	75	59	34
NR	52	77	-	47	32	19
PR	68	75	47	-	35	20
TB	57	59	32	35	-	18
TN	36	34	19	20	18	-

	R	NA	NR	PR	TB	TN
R	-	9.9%	5.0%	6.5%	5.5%	3.5%
NA	9.9%	-	7.4%	7.2%	5.7%	3.3%
NR	5.0%	7.4%	-	4.5%	3.1%	1.8%
PR	6.5%	7.2%	4.5%	-	3.4%	1.9%
TB	5.5%	5.7%	3.1%	3.4%	-	1.7%
TN	3.5%	3.3%	1.8%	1.9%	1.7%	-

We applied the following heuristics to arrive at a single category of relevance for each of the 177 elements that were assessed differently by different searchers:

1. For elements that were viewed, but not-assessed, the explicit assessments are given priority.
2. If there was a majority vote, the majority category was chosen regardless of the difference.
3. If there was a tie with an element assessed as non-relevant, NR was chosen.
4. In remaining ties, any elements assessed as Relevant were categorised as relevant.
5. Any outstanding ties (i.e., between PR, TB and TN in any combination) were left as ties (indicated as -tie- below).

Table 3 shows the resulting distribution of the interactive track assessments in total and over the eight topics. Less than 25% were Partially Relevant, Narrow or Broad including only 10 ties. The rest are roughly divided into three equally sized groups of relevant, non-relevant and non-assessed elements, each of around 25%.

Table 3. Distribution of interactive track assessments over topics after application of heuristics on elements with two or more different assessments

Topic	T1	T3	T4	T5	T7	T8	T9	T12	Total
R	15	52	11	26	21	37	67	50	279
NA	21	31	27	23	16	55	60	35	268
NR	13	31	16	60	20	71	42	10	263
PR	4	16	9	14	11	25	15	11	105
TB	5	16	4	7	6	5	18	3	64
-tie-	1	5		1	1	1	1		10
TN	9	7	3	5	11	5	4	6	50
Total	68	158	70	136	86	199	207	115	1039

In order to compare the interactive assessments to those of the ad-hoc track, we applied the mapping heuristics discussed in Section 2.3 to the ad-hoc assessments. We regard any element with 75% or more highlighted content as relevant (R), and any with less than 25% as Too Broad. We thus arrive at a set of inferred assessments

where the ad-hoc assessments are mapped to the interactive track Relevant and Too Broad relevance categories as shown in Table 4. 801 elements that were assessed in the interactive track but not assessed in the ad-hoc track are also shown (the NA column). In addition, the 39 assessments that fall outside the range defined by the inferred R and TB categories are shown distributed over 5 intermediate bins according to the rsize/size ratio. Excluding 23 elements that were viewed but not assessed in the interactive track (NA, second row) leaves only 215 elements that were assessed in both tracks.

The data from Table 4 suggest that there is little agreement in what kind of information interactive and ad-hoc assessors deem as useful for the same information-seeking tasks, since there is relatively small overlap in the 215 common elements assessed. A further observation from the data is that, with regards to the commonly assessed elements, there is a certain degree of agreement on relevant and not relevant information, as demonstrated by the level of agreement in the R and NR[5] rows. For instance, of the 129 elements assessed as relevant in the interactive track, 75 were relevant in the ad-hoc assessments and 12 more had between 50% - 75% relevant content as measured by the rsize to size ratio. In addition, looking at marginal cases such as TB and TN in the interactive assessments, we notice that relatively few of these are Relevant in the ad-hoc data.

Table 4. Distribution of inferred relevance categories (Relevant and Too Broad) of ad-hoc assessments as well as non-assessed ad-hoc elements over interactive track assessments

		Ad-hoc data: Inferred relevance categories & non-assessed elements									
	rsize / size	0.3-0.4	0.4-0.5	0.5-0.6	0.6-0.7	0.7-0.8	R	TB	Total	NA	Grand total
Interactive track data	R	7	2	3	5	4	**75**	**33**	129	150	279
	NA						**13**	**10**	23	245	268
	NR	2	3			1	**6**	**13**	25	238	263
	PR	1	1	2	1	2	**11**	**5**	23	82	105
	TB			2		1	**9**	**12**	24	40	64
	-tie-						**1**	**1**	2	8	10
	TN	1		1			**4**	**6**	12	38	50
	Total	11	6	8	6	8	119	80	238	801	1039

The rather small overlap between the two sets of assessments indicates that each set contains significant numbers of elements not assessed in the other set. To investigate the nature of the unique contribution by the interactive track, we have checked

[5] Especially so given that non-assessed (NA) elements in the ad-hoc track are an explicit indication of non relevance.

how many of the 801 elements not assessed in the ad-hoc track were actually present in the ad-hoc assessment pools. Table 5 shows that 510 of the interactive track elements were not even included in the ad-hoc track pools, that is, they were not found by any of the systems of the ad-hoc track participants. In slightly more than half of these cases, the interactive track searchers found these elements either non-relevant or not worth assessing. However, in 117 cases (23%) they did find the elements fully relevant and in another 114 cases (22%) relevant to some degree (i.e., PR,TB, TN or a tie between these). Thus, at least from the perspective of interactive track searchers, there were much more relevant information to be found for these 8 tasks than identified in the ad-hoc track.

Table 5. Distribution of non-assessed elements from the ad-hoc track over interactive track assessments, including and excluding elements in the ad-hoc pools

	NA	NA, not in ad-hoc pool
R	150	117
NA	245	150
NR	238	129
PR	82	49
TB	40	25
-tie-	8	7
TN	38	33
Total	801	510

Finally, we investigate if there were any differences in the perceived relevance depending on the task type, and depending on the type of backend used. As the number of mutually judged elements in the ad-hoc and interactive track is quite small the full set of interactive track assessments are used for this analysis. Figure 1 shows the distribution of inferred relevance categories over the three tasks types used in the study. Comparing across task types there are indications that the searchers found a larger proportion of Relevant and a smaller proportion of Non-relevant elements for the Information gathering tasks. For the Fact finding tasks, the trend is the opposite. The Decision making tasks lie in the middle of these two extremes, with a relative low proportion of Non-relevant and a slightly lager proportion of Too broad and Too narrow than either of the other two task types. The element and passage backend systems thus performed better for the more general Information gathering tasks, and somewhat poorer for the more specific Fact finding tasks. This may seem counter intuitive, bearing in mind that the goal of XML element retrieval is to support more focused retrieval. It may, however, be partially explained by the fact that keyword only queries with no structural hints were used in the study.

Figure 2 below shows the distribution of the inferred relevance categories on the two backend systems. The distribution is quite similar, with only a slight tendency for more Relevant and less non-assessed elements in the passage system.

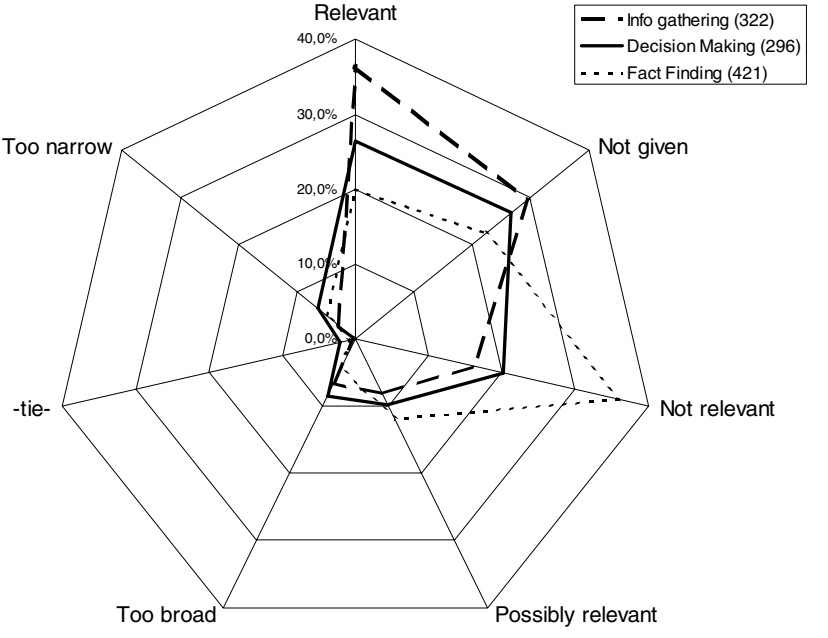

Fig. 1. Distribution of inferred relevance categories over the three task types in the study: Information gathering, Decision making and Fact finding

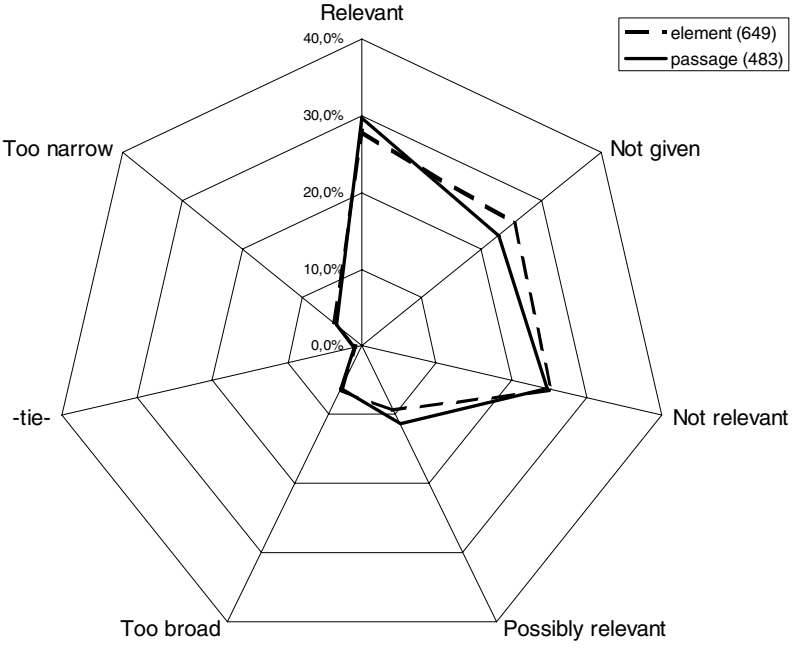

Fig. 2. Distribution of inferred relevance categories over backend systems: element and passage

4 Concluding Remarks and Future Work

We reported an initial comparison of relevance assessments made as part of the INEX 2006 Interactive Track (itrack'06) to those made for the topic assessment phase of the INEX 2007 ad-hoc track. The data that we presented suggest that there are significant differences in what information was assessed under the two different conditions, but it also suggests a certain level of agreement in what constitutes relevant and non-relevant information for those elements that were assessed in both tracks. In addition, a noteworthy amount of additional relevant elements were identified by interactive track searchers. There are also indications that the task type has an influence on the distribution of relevance assessments, and that there were not great differences between the assessments given in the element and passage backend retrieval systems. For future work, we plan to investigate the effect of different relevance schemes (e.g. by removing the 'partially relevant' level), and we also plan to further investigate the effect that specific differences in the assessment conditions might have had in the relevance assessments. Kazai's initial analysis indicate [5], based on video recordings, that the process of giving assessments in itself may have affected the data, leading to increased interaction.

References

1. Cormack, G.V., Palmer, C.R., Clarke, C.L.A.: Efficient construction of large test collections. In: Proceedings of the 21st ACM SIGIR Conference, pp. 282–289 (1998)
2. Denoyer, L., Gallinari, P.: The Wikipedia XML corpus. SIGIR Forum. 40(1), 64–69 (2006)
3. Fuhr, N., Klas, C.P., Schaefer, A., Mutschke, P.: Daffodil: An integrated desktop for supporting high-level search activities in federated digital libraries. In: Agosti, M., Thanos, C. (eds.) ECDL 2002. LNCS, vol. 2458, pp. 597–612. Springer, Heidelberg (2002)
4. Fuhr, N., Lalmas, M., Trotman, A., Kamps, J. (eds.): Focused access to XML documents. In: INEX 2006. LNCS, vol. 4518. Springer, Heidelberg (2007) (to appear)
5. Kazai, G.: Search and navigation in structured document retrieval: Comparison of user behaviour in search on document passages and XML elements. In: Proceedings of the 12th Australasian Document Computing Symposium (ADCS 2007) (2007)
6. Larsen, B., Tombros, A., Malik, S.: Obtrusiveness and relevance assessment in interactive XML IR experiments. In: Trotman, A., Lalmas, M., Fuhr, N. (eds.) INEX 2005, pp. 39–42 (2005)
7. Malik, S., Tombros, A., Larsen, B.: The interactive track at INEX 2006. In: Fuhr, N., Lalmas, M., Trotman, A. (eds.) Proceedings of the 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, pp. 387–399 (2006)
8. Pehcevski, J., Thom, J.A., Vercoustre, A.M.: Users and assessors in the context of INEX: Are relevance dimensions relevant? In: Trotman, A., Lalmas, M., Fuhr, N. (eds.) INEX 2005, pp. 47–62 (2005)
9. Sanderson, M., Joho, H.: Forming test collections with no system pooling. In: Proceedings of the 27th ACM SIGIR Conference, pp. 33–40 (2004)
10. Spärck Jones, K., van Rijsbergen, C.J.:: Report on the need for and provision of an 'ideal' information retrieval test collection. British Library Research and Development Report 5266, University Computer Laboratory, Cambridge (1975)

11. Theobald, M., Schenkel, R., Weikum, G.: An efficient and versatile query engine for TopX search. In: Proceedings of the 31st International Conference on Very Large Data Bases (VLDB), pp. 625–636 (2005)
12. Toms, E.G., O'Brien, H., MacKenzie, T., Jordan, C., Freund, L., Toze, S., Dawe, E., MacNutt, A.: Task effects on interactive search – The Query Factor. In: INEX 2006. Springer, Heidelberg (to appear, 2008)
13. Voorhees, E.M.: Variations in relevance judgments and the measurement of retrieval effectiveness. In: Proceedings of the 21st ACM SIGIR Conference, pp. 315–323 (1998)
14. Voorhees, E.M.: Evaluation by highly relevant documents. In: Proceedings of the 24th ACM SIGIR Conference, pp. 74–82 (2001)
15. Zobel, J.: How reliable are the results of large-scale retrieval experiments? In: Proceedings of the 21st ACM SIGIR Conference, pp. 307–314 (1998)