Assessors' Search Result Satisfaction Associated with Relevance in a Scientific Domain

1st Author
1st author's affiliation
1st line of address
2nd line of address
Telephone number, incl. country code
1st author's email address

2nd Author
2nd author's affiliation
1st line of address
2nd line of address
Telephone number, incl. country code
2nd E-mail

3rd Author
3rd author's affiliation
1st line of address
2nd line of address
Telephone number, incl. country code
3rd E-mail

ABSTRACT

In this poster we investigate the influence between perceived ease of assessment of situational relevance by a four-point scale, perceived satisfaction with retrieval results and the actual relevance assessments made by test collection assessors based on their own genuine information tasks. Ease of assessment, satisfaction and number of relevant documents are cross tabulated with retrieval performance measured by Normalized Discounted Cumulated Gain. Results show that when assessors find small numbers of relevant documents they tend to regard the search results with dissatisfaction and, in addition, they obtain lower performance for all document types involved.

Categories and Subject Descriptors

H.2.4 [**Information retrieval experiment**]: The ACM Computing Classification Scheme: http://www.acm.org/class/1998/

General Terms

Performance, Human Factors.

Keywords

Relevance assessment, Information retrieval, Search satisfaction.

1. INTRODUCTION

A main challenge in IR evaluation is to assess retrieval performance, observe interactive IR processes and understand searcher behavior in context of the searcher situation. So far the sequence of TREC evaluations of IR systems has provided tracks and corresponding test collections mainly belonging to information domains and document types such as newswire documents, genomics, the web, etc. [1]. Very few collections include academic publications with reference lists and derived citation networks. The INEX collection from 2002-05 constitutes such a test collection for XML IR in the field of Computer Science. However, it is a small collection (approx. 16,000 documents) [2]. The large *iSearch* tests collection on Physics seeks to alleviate this problem. We describe iSearch below [3].

The TREC test collections are commonly providing a set of 'topics' that are constituted by a title, description and a narrative describing the kind of documents that are deemed relevant for n any given topic. Relevance assessments are made *posteriori* by pooling the top retrieval results per topic across many different retrieval models, removing the duplicates and presenting a selected list of full text documents to a human assessor. Typically, the assessments are made as 'topicality' judgments in binary form but they may also be done by means of a graded relevance scale, e.g., as proposed and tested in [4-6]. Performance is commonly measured by standard measures like Mean Average Precision (MAP) or measures belonging to the Cumulated Gain family [7]. Characteristically, relevance assessment consistency across several assessors has been investigated [8] in TREC. Notwithstanding, the assessment process and its behavioral aspects have scarcely been studied in connection with test collection design that applies genuine information task situations. In INEX the information requests were designed as simulated work task situations made from natural information situations, with some subsequent analysis of the natural tasks [9].

The present paper focuses on assessor behavioral observations and correlations to retrieval performance. Later contributions will seek to analyze relationships between information task features and relevance assessments. It is structured as follows. First the research design is described including a brief outline of the *ISearch* collection. This is followed by the result sections and a discussion of our findings.

2. RESEARCH DESIGN

The *ISearch* collection [3] consists of approx. 18,000 English monographic records from Danish digital libraries, 160,000 papers and articles in full-text PDF as well as 275,000 abstracts with a varied set of metadata and vocabularies captured from the open access portal arXiv.org. The collection currently contains a set of 65 genuine information tasks generated by 23 test persons from Physics university departments (Ph.D. and experienced M.Sc. students and Associate Professors). Each information task consists of an information need statement, a description of the underlying work task and a formulation of the current state of knowledge of the task captured from the persons through an online question form. In addition, the form also elicits statements on the ideal answer of a search (like the narrative in TREC) as

3rd *IIiX Conference*, Aug. 19-21, 2010, New Brunswick, NJ, USA.

perceived by the test person, as well as on search keys perceived appropriate by the person. In total, the extracted data from each information task serve as *contextual evidence* of the information situation of the person with a task at hand. The various kinds of extracted evidence may later be used in the *ISearch* test collection for experiments, e.g., as assigned simulated work task situations (cover stories) in search jobs [10] or in line with the research design by Kelly & Fu [11], who applied combinations of similar kinds of extracted contextual evidence on search tasks in interactive retrieval experiments. For each tests person/assessor a questionnaire on personal data was filled out.

For each task a set of up to 200 documents per task were retrieved for situational relevance assessments made by the assessors based on their task descriptions. Each document type was represented in the set in proportion to their representation in the corpus. The retrieval was performed manually by the research team in the corpus using a vector space-based search engine and primarily by application of the search keys proposed by the test persons/assessors in the online form. No TREC-like pooling was done. The assessments were based on the Sormunen four-point relevance scale [6]: highly; fairly; marginally; and not relevant. The nature of situational relevance (usefulness to task situation) as well as the four-point scale were explained and illustrated to the assessors. They did the assessments on a dedicated web-based program and were allowed one week for the judgmental activity. A post-assessment questionnaire on satisfaction with the assessment procedure and search results was filled out for each task.

2.1 Research Questions

The present paper investigates measures of retrieval performance associated with central aspects of the relevance judgment behavior *and* the assessors' perceptions of the assessment situation and search results. This is partly based on the actual assessments done across the different document types in the collection, and partly captured from the post-relevance questionnaire. We operate with the following three research questions

1. Do human assessors find it easy to judge documents for situational relevance and do their relevance assessments influence the degree of easiness?

2. Does the number of (graded) relevance judgments made by human assessors influence the perceived degree of satisfaction with the search outcome?

3. Does retrieval performance vary significantly in relation to degree of satisfaction with search outcome and document type?

Research question one was based on two assumptions: The first assumption is that experienced assessors will find it easy to judge documents for situational relevance according to a four-graded scale. The second assumption is that the more comprehensive the judgments, the easier the test persons will find the assessment activity. Here, we measure comprehensiveness of relevance judgments by use of the relevance grading and number of positive relevant documents per information task.

Research question two is based on the hypothesis that with decreased volume of positive relevant documents and number of highly relevant documents per information task, dissatisfaction with the search outcome increases. Research question three assumes that performance varies in association with degree of satisfaction and document type. PDF full text documents are assumed to perform better than arXiv.org metadata records and book records owing to their diversified informativeness in the text volume. The outcome of the research questions can serve to better qualify the design of the test collection features in the future.

2.2 Analysis Methods

The relevance assessments per information task were captured and the distribution of the set of all positively relevant documents over all 65 information tasks was calculated. Highly, fairly and marginally relevant documents constitute 'all positively' relevant items. Two central questions from the post relevance questionnaire (PRQ) were selected concerning: (1) ease of assessing documents for situational relevance and the use of the four-point scale; (2) satisfaction with the search output and work task fulfillment. For each question descriptive statistics were generated and cross tabulations were made between relevant documents and (a) the degree of easiness of situational relevance assessment, and (b) degree of satisfaction with search output. In all cases retrieval performance was measured by NDCG [7] and statistical significance tests were performed in the form of two-tailed Student's t-tests with an α of 0.05.

3. RESULTS

Table 1 demonstrates the distribution of relevant documents regardless of document type over the 65 information tasks. Half of the tasks (33) contain 15-74 relevant documents. 12 tasks hold more than 74 relevant documents, whilst 20 information tasks contain less than 15 relevant documents.

Table 1. Distribution of relevant documents over tasks.

	No. of tasks
Range of relevant docs.	N = 65
> 100	9
75 - 100	3
50 - 74	8
25 - 49	13
15 - 24	12
10 - 14	8
< 10	12

We find the following distribution of graded relevance assessments across the tasks. All three positive relevance grades are found for 46 of the information tasks; 13 tasks account for the combination of fairly + marginally relevant, one task for the combination of highly + marginally whereas 5 tasks possess only one of the positive grades. A closer analysis reveals that 13 of the 20 tasks with less than 15 relevant documents, Table 1, have only two (fairly + marginally) or one positive relevance grade. This means that 7 information tasks contain all three positive relevance grades, albeit in scarce numbers.

3.1 Ease of Assessments and Satisfaction with Search Outcome

Table 2 displays the general results from the replies to the two selected questions from the PRQ. In general, the assessors had no difficulty performing the situational relevance assessments, but they are only 'somewhat satisfied' or quite 'dissatisfied' with the search outcome and the task fulfillment.

Table 2. Assessors' degree of ease, satisfaction and task fulfillment doing relevance assessments.

Judgments; N = 65 (%)	Doing assessments	Understand 4-scale Rel.	Search result satisfaction	Task fulfillment
Extremly easy/satisfied	31 (47.7)	35 (53.8)	5 (7.7)	3 (4.6)
Somewhat easy/satisfied	33 (50.8)	25 (38.5)	26 (40.0)	25 (38.5)
Not easy/satisfied	1 (1.5)	5 (7.7)	34 (52.3)	37 (56.9)

3.2 Combining Performance, Assessment Easiness and Result Satisfaction

Tables 3 and 4 show the association between degrees of ease of assessment and the result satisfaction, respectively, *and* the actual relevance assessment made prior to their answers to the PRQ.

Table 3. Relevant document distribution over ease of assessment.

Relevance assessments	*Highly Rel.*		*Fairly Rel.*		*Marginally*		*All relevant*		Not relevant		Total	All relevant
Search result satisfaction	No.	%	No.	%	No.	%	No.	%	No.	%	No.	Avr. No.
Extremly easy (N=31)	132	2.5	318	6.2	874	17.0	1324	25.7	3820	74.3	5144	42.7
Somewhat essy (N=33)	205	3.5	339	5.9	952	16.4	1496	26.1	4226	73.9	5722	45.3
Not easy(N=1)	0	0	9	4.5	49	24.5	58	29	142	71.0	200	58.0
Total (N=65)	337	3.0	666	6.0	1875	16.9	2878		8188	74	11066	
Mean (N=65)	5.2		10.2		28.8		44.3		126		170.25	44.3

The descriptive statistics, Tables 3-4, also include the number and percentage of 'all relevant' as well as 'non-relevant' documents that were assessed across the three degrees of easiness/satisfaction. The average numbers and percentage of the graded relevance categories are also shown for *all* 65 information tasks. The (average) numbers of 'All relevant' grades for the two positive degrees of easiness of assessment are quite substantial and rather similar (42.7 – 45.3), Table 3, although not significantly *higher* for the level of 'somewhat' easy. Also the percentage figures are similar, regardless of perception of ease.

For Table 4, it is evident that when assessors perceive being presented with an insubstantial number of relevant documents, relatively speaking, they find the retrieval result unsatisfactory. However, a detailed analysis shows that 20 of the 34 non-satisfactory information tasks actually contain *all three grades* of positive relevance assessments out of which the 12 tasks observed above, Table 1, albeit contain rather few relevant items. Note that the average number and percentage of 'All relevant' documents are significantly lower in the category of 'not satisfied'. A similar picture is demonstrated concerning task fulfillment (not shown).

Table 4. Relevant document distribution over result satisfaction.

Relevance assessments	*Highly Rel.*		*Fairly Rel.*		*Marginally*		*All relevant*		Not relevant		Total	All relevant
Search result satisfaction	No.	%	No.	%	No.	%	No.	%	No.	%	No.	Avr. No.
Extremly satisfied (N=5)	106	17.2	63	10.2	139	22.5	308	49,9	309	50.1	617	61.6
Somewhat satisfied (N=26)	149	3.3	383	8.5	811	18.0	1343	29,8	3159	70.2	4502	51.7
Not satisfied (N=34)	82	1.4	220	3.7	925	15.6	1227	20,6	4720	79.4	5947	36.1
Total (N=65)	337	3.0	666	6.0	1875	16.9	2878	26	8188	74	11066	
Mean (N=65)	5.2		10.2		28.8		44.3		126		170.25	44.3

3.3 Retrieval Performance and Result Satisfaction

Table 5 provides NDCG scores [7] for the three values or degrees of search satisfaction crossed with the document types constituting the iSearch test collection. The expected performance differences between the 'somewhat' and the 'not satisfied' values in the PDF and Metadata+Abs. document types are statistically significant. The latter satisfaction value displays the lowest performance scores. No difference in performance can be detected between the PDF and the metadata records for the 'not satisfactory' value. The higher performance scores for PDF over metadata and book records in the 'somewhat satisfactory' category over the different DCVs are only statistically significant in relation to the book records.

It is interesting to observe the quite high NDCG scores for book records (.42; .25; .30; .32) in tasks that are being perceived as providing 'not satisfied' search results after the relevance assessment activity. They are not statistically significant.

Table 5. NDCG for search satisfaction values. Statistical significance (p=.001- .03) in bold/italics in rel. to *italics*.

Record type(s)	Value	# tasks	NDCG	NDCG@10	NDCG@20	NDCG@30
	Extr. Satisfied	5	0,43	0,37	0,33	0,34
All doc. Types	*Somewhat*	24	0,34	*0,30*	*0,27*	*0,26*
	Not Satisfied	33	0,25	***0,12***	***0,11***	***0,11***
	Extr. Satisfied	5	0,53	0,38	0,42	0,44
Book record	*Somewhat*	21	0,29	0,17	0,19	0,20
	Not Satisfied	21	0,42	0,25	0,30	0,32
	Extr. Satisfied	3	0,36	0,23	0,20	0,20
PDF full text	*Somewhat*	24	0,40	*0,35*	*0,33*	*0,32*
	Not Satisfied	29	0,28	***0,11***	***0,13***	***0,14***
	Extr. Satisfied	4	0,46	0,35	0,33	0,32
Metadata+Abs.	*Somewhat*	24	0,35	*0,25*	*0,24*	*0,24*
	Not Satisfied	31	0,26	***0,10***	***0,11***	***0,12***

Diagram 1 demonstrates the development of the retrieval performance as measured by NDCG over DCVs from 5 over 100 to 1500 for the 3 document types and the two satisfaction values.

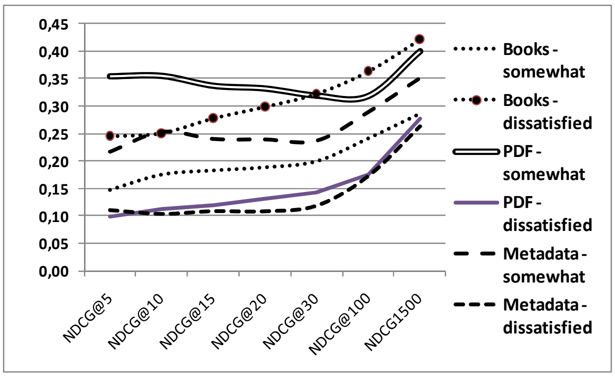

Diagram 1. NDCG scores for iSearch document types associated with result satisfaction.

Clearly, the 'not satisfied' assessors judging PDFs and arXiv.org metadata records both display the lowest NDCG scores over all DCVs. ' Not satisfied' assessors judging books constantly score book records .10 NDCG scores above the 'somewhat satisfied' assessors' scores. From NDCG30 the 'not satisfied' category for book records is the best performing document type. Assessors being 'somewhat satisfied' after judging PDFs and metadata records obtain the best performance scores at the start of result rankings, with the PDFs as the best performing document type.

4. DISCUSSION and FUTURE WORK

The distribution of relevant documents over the information tasks, Table 1, suggests that approximately 12-20 of the current tasks in the *iSearch* test collection are difficult to retrieve, owing to quite few (1-14) relevant documents found. Information task difficulty plays a role for the searching agent's behavior during relevance assessment and feedback [12] as well as for the total performance result. The test collection may hence be qualified into at least four *levels of information task difficulty*: (1) very difficult tasks with 1-9 relevant documents; (2) difficult tasks with 10-14 relevant documents; (3) fairly easy information tasks containing 15-74 relevant documents; and (4) easy information tasks containing more than 75 relevant documents.

The assessors in general find it easy to judge documents for situational relevance as well as understanding the four-point relevance scale [7], Table 2. Table 3 demonstrates that the assessors' relevance judgments, as measured in terms of number of relevant documents found and distribution over relevance grading did *not influence* the degree of easiness. The first part of research question one is thus answered positively. The second part, and the underlying assumption of an existence of correspondence between number of relevant documents found and degree of easiness cannot be answered, since the number of relevant documents in the 'extremely' and 'somewhat easy' are almost the same.

With respect to research question two there is no doubt that when the number of relevant documents found by the assessor is low, or perceived as small, (but not necessarily the number of graded relevance categories used) the degree of satisfaction with the retrieval result is negative. Table 4 clearly indicates the connection between very few highly (1.4 %), fairly (3.7 %) and marginally relevant documents (15.6 %) *and* dissatisfaction, in comparison with the distribution of the three 'positive' relevance grades for tasks perceived 'somewhat satisfying' (3.3 %; 8.5 % and 18 %, respectively). In addition, the *average number* of documents found relevant according to the three grades is significantly lower among the search results perceived as dissatisfying.

With respect to the third research question the general trend is that information tasks perceived as '*dissatisfying'* are also those that obtain the *least performance* scores (.25 vs. .34 for 'somewhat satisfied' in the NDCG column, Table 5). At short result rankings (NDCG10-30) the performance difference is even larger (.11 vs. .30) and statistically significant. However, a more detailed analysis, Table 5 and Diagram 1, reveals that PDF full text and arXiv.org metadata with abstracts for the 'somewhat satisfied' category contain different but albeit not statistically significant performance scores, with the PDF type serving as the best performing type - also over several DCVs. The assumption that the PDF full text documents would perform better than other document types is hence realistic.

The only strong statistically significant difference of retrieval performance is found between the 'somewhat satisfying' and 'not satisfying' categories for the document types (in italics and bold, Table 5). With one exception there exists a robust and significant association between *low performance* scores and *dissatisfaction* with retrieval result.

The exception is the *book type*, which displays the highest NDCG scores for the 'not satisfied' category of search results compared to the 'somewhat satisfying' category. One explanation might well be that a number of recently catalogued science monographs in Danish digital libraries also contain quite substantial table-of-contents data as a new standard and thus are easier retrieved. The problems of perceived satisfaction concerning retrieval results and task fulfillment encountered by the assessors are realistic issues in everyday retrieval situations, as tried out here.

The intention is further to investigate factors captured both from the post work task questionnaire, the information task form and the post relevance questionnaire, in comparison with the actual relevance assessments and performance scores in order to better understand the assessment process and to qualify the information tasks, e.g. in relation to task difficulty or document types in the *ISearch* collection for future experimental use.

5. REFERENCES

[1] Voorhees, E.M and Harman, D.K. 2005. TREC: Experiment and Evaluation in Information Retrieval. MIT Press, Cambridge, MA.

[2] Kamps, J., Lalmas, M. and J. Pehcevski. 2007. Evaluating relevant in context: Document retrieval with a twist. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, New York, NY, 723–724.

[3] Lykke, M., Larsen, B., Lund, H. and Ingwersen, P. 2010. Developing a Test Collection for the Evaluation of Integrated Search. In: Advances in Information Retrieval. Proceedings of 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31. Springer, Berlin, Germany, 627-630. DOI - 10.1007/978-3-642-12275-0_63.

[4] Kekäläinen, J. 2005. Binary and graded relevance in IR evaluations - Comparison of the effects on ranking of IR systems. Inf. Proc.& Man., 41(5), 1019-1033.

[5] Kekäläinen, J. and Järvelin, K. 2002. Using graded relevance assessments in IR evaluation. J. Am. Soc. Inf. Sc. Tech., 53(13), 1120-1129.

[6] Sormunen, E. 2002. Liberal relevance criteria of TREC – Counting on negligible documents? In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, New York , NY, 320-330.

[7] Järvelin, K and Kekäläinen, J. 2002. Cumulated gain-based evaluation of IR techniques. ACM Trans. In. Syst. (ACM TOIS), 20(4), 422-446.

[8] Voorhees, E.M. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In: Proceedings of the 21th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, New York , NY, 315-323.

[9] Malik, S., Klas, H.-P., Fuhr, N., Larsen, B. and Tombros, A. 2006. Designing a user interface for interactive retrieval of structured documents — Lessons learned from the INEX interactive track. In: Research and Advanced Technology for Digital Libraries. Springer Verlag, Heidelberg, 291-302. DOI: 10.1007/11863878_25.

[10] Borlund, P. 2003. The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. Inf. Res., 8(3), paper no. 152.

[11] Kelly, D. and Fu, X. 2007. Eliciting better information need descriptions from users of information search systems. Inf. Proc.& Man., 43(1), 30-46.

[12] Arapakis, I., Jose, J.M. and Gray, P.D. 2008. Affective feedback: An investigation into the role of emotions in the information seeking process. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, New York, NY, 395-402.