Does Degree of Work Task Completion Influence Retrieval Performance?

Peter Ingwersen, Toine Bogers
Royal School of Library and Information Science
Birketinget 6, DK 2300
Copenhagen S, Denmark
+453258 6066
{pi, tb}@iva.dk

Marianne Lykke
Aalborg University
Kroghstraede 1 DK 9220
Aalborg OE, Denmark
+452125 1854
mlykke@hum.aau.dk

ABSTRACT
In this contribution we investigate the potential influence between assessors' perceived completion of their work task at hand and their actual assessment of usefulness of the retrieved information. The results indicate that the number of useful documents found by assessors does *not* influence their perception of task completion. Also, with the exception of full text records and across all document types, both measured at rank 10, no statistically significant correlation is observed with respect to retrieval performance influenced by degrees of perceived work task completion or individual types of documents.

Keywords
Relevance assessment, usefulness measures, work task completion, assessor behavior, IR interaction.

INTRODUCTION

The present poster investigates assessment behavior with respect to degree of perception of task completion and how this factor influences retrieval performance as measured by degree of usefulness of retrieved documents in a collection of different document types, the *i*Search collection.

RESEARCH DESIGN

The *i*Search collection (Lykke et al., 2010) integrates approximately 18,000 English monographic records from Danish digital libraries without abstracts, 160,000 papers and articles in full-text PDF format as well as 275,000 abstracts with a varied set of metadata and vocabularies captured from the open access portal arXiv.org. The full text documents are much longer (4,422 words on average) than the metadata records (272 words on average). The collection currently contains a set of 65 richly described information tasks including genuine work task statements created by 23 test experts from Physics university departments. The same 23 experts assessed the usefulness, not topicality, of retrieved documents (up to 200 randomly distributed over the different document types per task) to their actual work task situation in relation to the 65 information tasks.

For degree of usefulness we applied the four-graded relevance scale as proposed by Sormunen (2002): highly; fairly; marginally; and not useful, as well as Normalized Discounted Cumulated Gain (nDCG) measurements (Järvelin & Kekäläinen, 2002). A post assessment questionnaire (PAQ) on satisfaction with the assessment procedure and search outcomes was filled out for each task.

Research Questions

The analyses are partly based on the actual assessments done across the different document types in *i*Search, and partly captured from the PAQ. We operate with the following two research questions

- Does the number of useful documents influence the assessors' perceived degree of task completion?
- Does retrieval performance vary significantly as to degree of task completion and document type?

Research question one assumes that the higher the number of useful documents, the more complete the test persons will perceive the work task.

For research question two we hypothesize that retrieval performance, as measured by usefulness, will be higher the more complete the task is perceived. With respect to document types and performance, we expect full-text PDF documents to perform better than arXiv.org metadata and book records owing to their higher informativeness.

Analysis Methods

After all the assessments were captured per information task we calculated the distribution of the set of all positively useful documents over all 65 information tasks and document types. Highly, fairly and marginally useful documents constitute 'all positively' useful items. One central question from the PAQ was selected concerning the degree of work task completion, measured by a three-point scale: extremely complete; somewhat complete; and not complete. We generated descriptive statistics and cross-tabulated between retrieval performance and the perception

.

ASIST 2010, October 22–27, 2010, Pittsburgh, PA, USA.

of task completion. Statistical significance tests were performed as a two-tailed Student's t-tests ($\alpha = 0.05$).

RESULTS

Influence of Assessment of Usefulness on Task Completion

Table 1, located at the end of the paper, shows the association between the actual assessments of useful documents made *prior* to the assessors' answers to the PAQ *and* the perceived degree of task completion captured by PAQ. We observe that in the category of work tasks perceived as 'Not complete', as expected the assessors *do* obtain a slightly smaller portion of highly useful documents (2%), compared to the category of tasks perceived 'somewhat complete' (3.2%) and the mean (3.0%). However, there is no significant difference in the average number of useful documents seen by the assessors across the categories of perceived task completion, although the figures indicate a slight negative trend for the 'Not complete' category. This is also the case for the 'All Useful' percentages and documents, Table 1. Thus, research question 1 is not answered positively: a lower number of useful documents retrieved *does not* (statistically for this sample) entail a similar sense of (less) task completeness.

Task Completion and Retrieval Performance

Observing Table 2 for the aggregated level named 'All document Types', the trend is clear up to rank 30, and statistically significant at nDCG10: when work tasks are perceived 'Not Complete' the usefulness score of the retrieved documents is indeed lower than for tasks felt 'Somewhat Complete'.

Record type(s)	Value	# tasks	NDCG	NDCG@10	NDCG@20	NDCG@30
	Extr. Complete	3	0.39	0.28	0.32	0.34
All doc. Types	*Somewhat*	23	0.33	***0.29***	0.25	0.24
	Not Complete	36	0.27	*0.16*	0.14	0.14
	Extr. Complete	3	0.46	0.30	0.33	0.36
Book record	*Somewhat*	19	0.34	0.21	0.24	0.25
	Not Complete	25	0.39	0.24	0.27	0.29
	Extr.Complete	1	0.06	0.00	0.00	0.00
PDF full text	*Somewhat*	23	0.38	***0.32***	*0.30*	*0.29*
	Not Complete	32	0.31	*0.16*	*0.17*	*0.18*
	Extr.Complete	2	0.32	0.32	0.32	0.32
Metadata+Abs.	*Somewhat*	23	0.31	0.21	0.19	0.19
	Not Complete	34	0.31	0.16	0.16	0.17

Table 2. NDCG scores for task completion. Statistical significance (p=.001- .03) in gray+*italics* in rel. to *italics*.

However, when we observe the usefulness scores at *document type level* the analysis displays a much fuzzier picture. For Book records the usefulness scores are actually *higher* when work tasks are perceived 'Not Complete', compared to the 'Somewhat' category, across all the DCVs. Only for PDFs the difference in usefulness score between the 'Somewhat' and 'Not Complete' categories is marked (in *italics*), but only statistically significant at nDCG10 (.32 vs. .16). For research question two we may state that the perceived degree of work task completion *does* influence negatively the retrieval performance measured by degree of usefulness of the retrieved documents – but *only* significantly within rank 10 for PDF documents or when individual document types are integrated.

CONCLUDING DISCUSSION

These observations concerning the research questions are quite interesting from an interactive IR evaluation point of view. The results show that the number of useful documents observed by assessors' prior to their perception of degree of task completeness does not influence statistically the feeling of task completion.

Further, and most important: when searchers feel that their work task is incomplete this perception *does influence* their sense of document usefulness at the *top rankings only*, in particular for PDFs; but for other document types and at lower rankings the analyses do not confirm the hypothesis. In fact for book records a less degree of task completeness shows better performance in terms of useful documents retrieved.

REFERENCES

Järvelin, K and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions of Information Systems (ACM TOIS)*, 20(4), 422-446.

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, (p. 627-630). Berlin, Germany: Springer. DOI-10.1007/978-3-642-12275-0_63.

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*, (p. 320-330), New York , NY: ACM Press.

.

Assessments of Usefulness	*Highly Useful*		*Fairly Useful*		*Marginally Useful*		*All Usefull*		Not Useful		Total	All Useful
Work task completion	No.	%	No.	%	No.	%	No.	%	No.	%	No.	Avr. No.
Extremly complete (N=3)	69	26.2	18	6.8	36	13.7	123	46.8	140	53.2	263	41
Somewhat complete (N=25)	134	3.2	282	6.8	804	19.4	1220	29.4	2930	70.6	4150	48.8
Not complete (N=37)	134	2.0	366	5.5	1035	15.6	1535	23.1	5118	76.9	6653	41.5
Total (N=65)	337	3.0	666	6.0	1875	16.9	2878	26.0	8188	74.0	11066	
Mean (N=65)	5.2		10.2		28.8		44.3		126.0		170.2	44.3

Table 1. Distribution of useful documents over task completion.