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

Ingwersen, Peter; Lykke, Marianne; Bogers, Antonius Marinus; Larsen, Birger; Lund, Haakon

DOI:
10.1145/1840784.1840826

Publication date:
2010

Document Version
Preprint (usually an early version)

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

Peter Ingwersen, Marianne Lykke, Toine Bogers, Birger Larsen & Haakon Lund
Royal School of Library and Information Science
Birketinget 6, DK 2300 Copenhagen S, Denmark
+45 3258 6066
{pi, mln, tb, blar, hl}@iva.dk

ABSTRACT
In this poster we investigate the associations between perceived ease of assessment of situational relevance made by a four-point scale, perceived satisfaction with retrieval results *and* the actual relevance assessments and retrieval performance made by test collection assessors based on their own genuine information tasks. Ease of assessment and search satisfaction 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, except for monographic records.

Categories and Subject Descriptors
H.3.3 [**Information search and retrieval]**

General Terms
Performance, Human Factors.

Keywords
Relevance assessment, Information retrieval, Search satisfaction.

1. INTRODUCTION
Main challenges in IR evaluation are 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 domain 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. Both the CACM and the INEX XML IR test collections in the field of Computer Science constitute such compilation. However, they are small collections (INEX approx. 16,000 documents) [2]. The large *i*Search test collection on Physics seeks to alleviate this problem. We describe *i*Searc*h* below [3].

3rd IIiX Conference, Aug. 18-21, 2010, New Brunswick, NJ, USA.

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 any given topic. Relevance assessments are made *a posteriori* by pooling the top retrieval results per topic across a number of different retrieval engines, removing the duplicates, and presenting a selected list of full text documents to the same human assessor who originally created the topic. 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 in TREC [8]. 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 problems, with some subsequent analysis of the natural tasks [9; 10].

The poster focuses on assessor behavioral observations and correlations to retrieval performance. It is structured as follows. First the research design is described including a brief outline of the *i*Search collection. This is followed by the result sections and a discussion of our findings.

2. RESEARCH DESIGN
The *i*Search 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 assessors 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 as perceived by the assessor (like the narrative in TREC), 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 assessor with a task at hand. The various kinds of extracted evidence may later be used in the *i*Search test collection for experiments, e.g., in line with the research design by Kelly &

Fu [11]. 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 – not on its topical contents. 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 assessors in the online form. 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 prior to experiments. They did the assessments on a dedicated web-based program and were allowed one week for the judgmental activity. A post-assessment questionnaire (PAQ) on satisfaction with the assessment procedure and search results was filled out for each task.

2.1 Research Questions

The analyses are based on the assessments done across the three document types in the collection and selected data captured from the PAQ. We operate with three research questions

1. Do human assessors find it easy to judge documents for situational relevance?

2. Does the number of positively graded relevant documents influence the assessors' perception of satisfaction with their 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 the assumption that domain expert assessors will find it easy to judge documents for situational relevance, i.e. in relation to their work task situation, according to a four-graded scale.

The second research question assumes that the more *comprehensive* the judgments, the more satisfied the test persons will find the retrieval result. Here, we measure comprehensiveness of relevance judgments by the use of the relevance grades *and* average number of relevant documents per information task. The underlying hypothesis is that as the number of relevant documents per information task decreases—in particular the number of highly and fairly relevant documents per information task—the perceived satisfaction of the retrieval result also decreases.

In research question three we hypothesize that retrieval performance will be higher on tasks with a higher degree of search result satisfaction. As to document types full text PDF documents are assumed to perform better than arXiv.org metadata records or book records owing to their larger number of access points 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 PAQ were selected concerning: (1) ease of assessing documents for situational relevance; (2) satisfaction with the retrieval result. 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] applying the log(rank+1) version for discounting as in TREC evaluations. 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.

Range of relevant docs.	No. of tasks 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 of those 20 information tasks contain all three positive relevance grades, albeit in scarce document 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. As shown, the assessors had no difficulty performing the situational relevance assessments, but were only 'somewhat satisfied' or quite 'dissatisfied' with the search outcome. Research question 1 is thus answered affirmatively.

Table 2. Assessors' degree of ease doing relevance assessments and retrieval result satisfaction.

Judgments: N = 65 (%)	Doing assessments	Search result satisfaction
Extremely easy/satisfied	31 (47.7)	5 (7.7)
Somewhat easy/satisfied	33 (50.8)	26 (40.0)
Not easy/satisfied	1 (1.5)	34 /52.3)

3.2 Combining Relevance Judgments and Retrieval Result Satisfaction

Table 3 (displayed at the end of the paper) deals with research question 2. It demonstrates the association between degrees of search result satisfaction *and* the actual relevance assessments made prior to their answers to the PRQ. The descriptive statistics also include the number and percentage of 'all relevant' as well as 'non-relevant' documents that were assessed across the three degrees of satisfaction. The average numbers and percentage of the graded relevance categories are also shown for *all* 65 information tasks.

For Table 3 it is evident that when assessors perceive being presented with an insubstantial number of relevant documents, relatively speaking, they find the retrieval result unsatisfactory. A detailed analysis of distribution of relevance grades over the information tasks shows that 21 of the 'somewhat' satisfactory tasks and 20 of the 34 non-satisfactory tasks actually contain *all three grades* of positive relevance assessments. However, of the 34 tasks in the 'non-satisfied' category 12 belong to the tasks observed above, Table 1, containing rather few (below 10) relevant items. Note also that the number and percentage of 'highly' and 'fairly' relevant documents, as well as the average number and percentage of 'All relevant' documents, are significantly lower in the category of 'not satisfied'.

One hypothesis behind research question 2 is thus confirmed: a relatively low *number* of relevant documents observed entails dissatisfaction with retrieval result. However, in terms of assessment comprehensiveness the applied number of relevance grades does *not* seem to influence the degree of satisfaction.

3.3 Retrieval Performance and Result Satisfaction

Table 4 answers the third research question. It provides the NDCG scores for the three categories of retrieval satisfaction crossed with the document types constituting the *i*Search test collection. The expected performance differences between the 'somewhat' and the 'not' satisfied categories for 'all document types' and in the PDF and Metadata+Abs. document types are indeed statistically significant. The latter category displays the lowest overall and @10+@20 NDCG scores.

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

No difference in performance can be detected between the PDF and the metadata records for the 'not' satisfactory category. The higher performance scores for PDF over metadata and book records in the 'somewhat' satisfied category across the different document cutoff values (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' satisfactory* search results. However, they are not statistically significant.

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

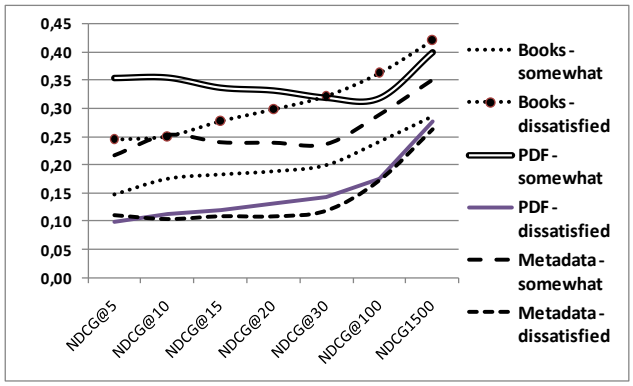


Figure 1. NDCG scores for *i*Search document types associated with retrieval result satisfaction.

Clearly, the dissatisfied assessors judging PDFs and arXiv.org metadata records display the lowest NDCG scores over all DCVs. The dissatisfied assessors judging *books* constantly score book records .10 NDCG scores above the 'somewhat satisfied' assessors' scores. From NDCG30 the dissatisfied category for book records is the best performing document type. Assessors being 'somewhat' satisfied when judging PDFs and metadata records obtain the best performance scores at the *start* of *result rankings*, see also Table 4, 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 *i*Search test collection may provide low retrieval performance, also at low DCVs, owing to quite few (1-14) relevant documents found. Information task retrieval difficulty plays a role for the assessors' behavior during relevance assessment and feedback [12] as well as for the total performance result. We are presently seeking more information task situations from researchers in physics with the aim of obtaining more tasks with a substantial (> 15) number of relevant documents.

In relation to *research question one* the assessors in general find it easy to judge documents for situational relevance, Table 2.

With respect to *research question two* it is evident that with an decreasing number of relevant documents found by the assessors, or perceived as small, (but not necessarily the number of graded relevance grades used) the degree of satisfaction with the retrieval result also decreases. Table 3 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 (36.1 vs. 51.7).

With respect to the *third research question* the general trend is that information results perceived as '*dissatisfying'* are also those that obtain the *least performance* scores (.25 vs. .34 for 'somewhat satisfied' in the NDCG column, Table 4). At short result rankings (NDCG10-30) the performance difference is even larger (.11 vs. .30) and statistically significant. However, a more detailed analysis, Table 4 and Figure 1, reveals that full text PDF and arXiv.org metadata with abstracts for the 'somewhat' satisfactory 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 full text PDF documents perform better than other document types is hence confirmed. The result have implications for the future design of retrieval rankings integrating different document types.

The second statistically significant difference of retrieval performance is found between the 'somewhat' and 'not' satisfying categories for *all* document types (in italics and bold, Table 4). With one exception a robust association exists between *low performance* scores and *dissatisfaction* with retrieval result. The exception is the *book type*, which displays the highest NDCG scores for the 'not' compared to the 'somewhat' satisfying category of results. One explanation might well be that a number of science monographs recently catalogued in Danish digital libraries contain quite substantial table-of-contents data as a new standard and thus are easier retrieved.

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 *i*Search 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.

Table 3. 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	49,92	309	50.1	617	61.6
Somewhat satisfied (N=26)	149	3.3	383	8.5	811	18.0	1343	29,83	3159	70.2	4502	51.7
Not satisfied (N=34)	82	1.4	220	3.7	925	15.6	1227	20,63	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

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

Peter Ingwersen, Marianne Lykke, Toine Bogers, Birger Larsen & Haakon Lund

Royal School of Library and Information Science, Denmark

SUMMARY

We investigate the associations between perceived **ease of assessment of situational relevance**, perceived **satisfaction with retrieval results** *and* the actual relevance assessments and **retrieval performance** made by test collection assessors based on their own genuine information tasks.

The retrieval performance is measured on a four-point scale by *Normalized Discounted Cumulated Gain [3].*

Results show:

(1) When assessors find small numbers of relevant documents they tend to regard the search results with dissatisfaction.

(2) In addition, they obtain lower performance when dissatisfied for all document types involved, except for monographic records.

(3) At short result rankings (NDCG10-30) the performance difference between 'somewhat' and 'not' satisfied search results is even larger (.11 vs. .30) and statistically significant.

Test searches were based on 65 **genuine and realistic search tasks**, from 23 lectures, PhDs and experienced MSc students. Performance measures based **situational relevance assessments, creating the *i*Search test collection** [1].

***i*Search:**

- **the test collection on Physics**:
160,000 full text PDFs,
275,000 arXiv.org Abstracts,
18,000 library records
3,750,000 references

CONTACT

Peter Ingwersen
Royal school of LIS, Denmark
pi@iva.dk

1. INTRODUCTION

Main challenges in IR evaluation are to assess retrieval performance, observe interactive IR processes and understand searcher behavior in context of the searcher situation.
Very few test collections include academic publications with reference lists and derived citation networks.

The large ***i*Search tests collection on Physics** seeks to alleviate this problem

In *i*Search each information task consists of ***five statements*** (*contextual evidence*) captured from the assessors through an online questionnaire on the:

- information need contents;
- underlying work task;
- current state of knowledge of the task;
- ideal answer of a search as perceived by the assessor (like the narrative in TREC), and
- search keys perceived appropriate by the person.

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 – not on its topical contents. Each document type was represented in the set in proportion to their representation in the corpus. The assessments were based on the Sormunen four-point relevance scale [2]: highly; fairly; marginally; and not relevant.

2. RESEARCH DESIGN & QUESTIONS

The analyses are based on the assessments of satisfaction across the three document types **in *i*Search** *and* selected data captured from the Post Assessment Questionnaire. We operate with the three research questions:

1. **Do human assessors find it easy to judge documents for situational relevance?**
2. **Does the number of positively graded relevant documents influence the assessors' perception of satisfaction with their search outcome?**
3. **Does retrieval performance vary significantly in relation to degree of satisfaction with search outcome and document type?**

3. FINDINGS

Statistical significance tests were performed in the form of two-tailed Student's t-tests with an α of 0.05..

Table 1. Distribution of relevant documents over tasks regardless of dicument types.

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

A closer analysis reveals that 7 of the 20 tasks with less than 15 relevant documents, Table 1, have all three positive relevance grades, albeit in scarce numbers. 13 tasks have only one or two positive relevance grades.

3.1 Research question 1 is answered affirmatively.

Table 2. Assessors' degree of ease and satisfaction

Judgments: N = 65 (%)	Doing assessments	Search result satisfaction
Extremely easy/satisfied	31 (47.7)	5 (7.7)
Somewhat easy/satisfied	33 (50.8)	26 (40.0)
Not easy/satisfied	1 (1.5)	34 /52.3)

3.2 Retrieval Result Satisfaction & Number of Relevance Judgments

Table 3. IR results satisfaction over relevance assessment grades

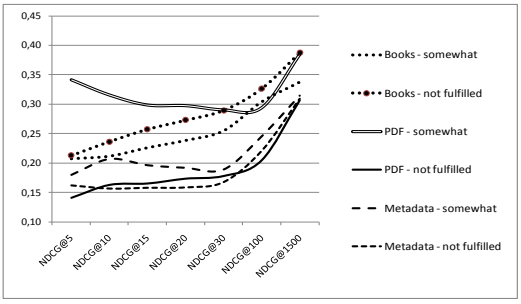
Relevance assessments	*Highly Rel.*		*Fairly Rel.*		*Marginally*		*All relevant*		Not relevant		Total	All relevant
Search result satisfaction	No.	%	No.	%	No.	%	No.	%	No.	%	No.	Avr. No.
Extremly satisfied (N=5)	106	17.2	63	10.2	139	22.5	308	49,92	309	50.1	617	61.6
Somewhat satisfied (N=26)	149	3.3	383	8.5	811	18.0	1343	29,83	3159	70.2	4502	51.7
Not satisfied (N=34)	82	**1.4**	220	**3.7**	925	**15.6**	1227	**20,63**	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

Findings show that when assessors perceive being presented with an *insubstantial number* of relevant documents, relatively speaking, they find the retrieval *result unsatisfactory*. One hypothesis **behind research question 2** is thus confirmed. However, in terms of assessment comprehensiveness the *applied number of relevance grades* does *not* seem to influence the degree of satisfaction.

3.3 IR Performance & Satisfaction

The expected performance differences between the 'somewhat' and the 'not' satisfied categories for 'all document types' and in the PDF and Metadata+Abs. document types are indeed **statistically significant for NDCG@10-30** – **research question 3**. The latter category displays the lowest overall and @10+@20 NDCG scores

Figure 1. NDCG scores for document types and retrieval satisfaction in *i*Search.



4. CONCLUSION

A central statistically significant difference of retrieval performance is found between the 'somewhat' and 'not' satisfying categories for *all* document types:

- A **robust positive association exists** between ***low performance*** scores and ***dissatisfaction*** with retrieval result.
- *Only the book type* displays the highest NDCG scores for the 'not' compared to the 'somewhat' satisfying category of results.

REFERENCES

[1] 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. Proc.of 32nd European Conference on IR Research. Springer, Berlin, Germany, 627-630.
[2] Sormunen, E. 2002. Liberal relevance criteria of TREC – Counting on negligible documents? In: Proc.of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, New York , NY, 320-330.
[3] Järvelin, K &Kekäläinen, J. 2002. Cumulated gain-based evaluation of IR techniques. ACM Trans. In. Syst. (ACM TOIS), 20(4), 422-446