An Exploratory Study into Perceived Task Complexity, Topic Specificity and Usefulness for Integrated Search

Peter Ingwersen[1,2], Christina Lioma[3], Birger Larsen[1] and Peiling Wang[4]

[1]Royal School of Library and Information Science, Birketinget 6, DK 2300 Copenhagen S, Denmark
[2]Oslo University College, St. Olavs Plass, 0130, Oslo, Norway
[3]Department of Computer Science, Copenhagen University, Njalsgade 128, 2300 Copenhagen S, Denmark
[4]University of Tennessee Knoxville, TN 37923, USA

[1]pi at iva.dk; [3]c.lioma at diku.dk; [1]blar at iva.dk; [4]peilingw at utk.edu

ABSTRACT

We investigate the relations between user perceptions of work task complexity, topic specificity, and usefulness of retrieved results. 23 academic researchers submitted detailed descriptions of 65 real-life work tasks in the physics domain, and assessed documents retrieved from an integrated collection consisting of full text research articles in PDF, abstracts, and bibliographic records [6]. Bibliographic records were found to be more precise than full text PDFs, regardless of task complexity and topic specificity. PDFs were found to be more useful. Overall, for higher task complexity and topic specificity bibliographic records demonstrated much higher precision than did PDFs on a four-graded usefulness scale.

Categories and Subject Descriptors

H.3.3 [**Information search and retrieval]**

General Terms

Performance, Human Factors.

Keywords

Task-based IR; Task complexity; Search topic specificity

1. INTRODUCTION

In today's digital environment, information retrieval (IR) commonly takes place in search environments consisting of diverse document and information types. It is thus of interest to observe how searchers evaluate the variety of potentially relevant or useful information for carrying out the task at hand. In this study, we analyze real users in academic work task situations, their judgments of usefulness of retrieved documents (bibliographic book records and PDFs), and their perceptions of their work tasks.

IIIX'12 Nijmegen, The Netherlands.
Copyright 2012 ACM 978-1-4503-1282-0/2012/08 ... $15.00.

We differentiate between relevance and usefulness. Relevance is referred to as topical relevance [7], and usefulness is regarded as a quality of information that relates directly to the task at hand [8]. We also differentiate work tasks from information or search tasks. Information tasks are called upon when the task performer encounters difficulties in completing work tasks. Work tasks are performed over a period of time, during which information searches are conducted when information needs arise. Examples of work tasks include both job-related tasks, like writing a paper for a conference, and daily-life tasks, like cooking a meal. Any retrieved relevant information will support the task performers in clarifying the topic area and useful information will support them moving forward in their work task solution [2].

There has not been extensive research on how usefulness assessment is related to perceived task complexity and topic specificity in integrated IR systems [4]. Integrated IR systems provide access to a variety of information objects from multiple sources through aggregated search engines. The purpose of the present research is two-fold: (1) to analyze how task performers assess different information types retrieved simultaneously and represented as bibliographic records and PDFs; and (2) to understand the relationship between graded usefulness assessments, work task complexity and search topic specificity, as perceived by the same task performers.

Earlier studies (see Section 2) did not investigate in-depth the influence of information object types on assessments of degree of usefulness in relation to topic specificity and task complexity. Specifically, in this paper we analyzed the data in the *i*Search[1] collection (see section 3) to address the following research questions: (Q1) How do task performers assess the usefulness of different document types in the *i*Search collection? In other words, how do different types of documents contribute to search results? (Q2) What is the relationship between perceived task complexity, search topic specificity, graded usefulness assessments and document types? In other words, how are different types of documents assessed for usefulness under influence of perceived properties of work tasks?

2. RELATED WORK

Studies of task complexity and task performers' information-seeking behaviour in professional settings [1] found that the degree of perceived work task complexity influenced the seeking

[1] http://itlab.dbit.dk/~isearch

process and constitutes a central property of tasks [11]. Recently, Ingwersen and Wang [4] investigated the association between perceived specificity of search topics and work task complexity; as well as their respective or joint contributions to usefulness assessments. They found that both topic specificity and task complexity played important roles in the task performers' evaluations of the search results. The more specific the search topic was perceived, the lower the retrieval performance and mean number of useful retrieved documents. The same was observed for increasing level of work task complexity. However, the perceived specificity of the search topic had more influence than perceived complexity of the work task. Further, when the documents were assessed as highly useful, they were likely proportionally related to the highly complex tasks.

Figure 1 [12] depicts the classic model of a real world assessment situation, in which the document information elements (DIEs) of retrieved documents were evaluated according to certain user criteria (e.g., quality, novelty) to derive at judgments of values for the task at hand (e.g., functional, conditional), which forms the bases for decision (e.g., accept, maybe). Further the task performers' states of knowledge and situations influence their assessments of usefulness of retrieved documents. In today's digital IR systems, both document representations and full texts are accessible in an integrated system, which may affect the amount of information viewed during the interactive assessment process and influence the assessment outcomes.

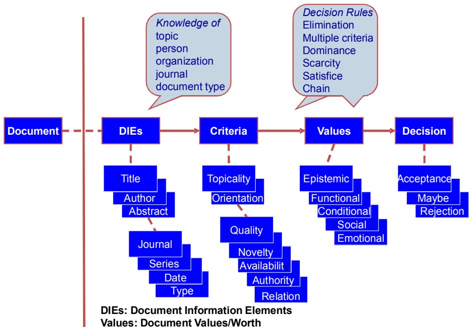

Figure 1: Factors involved in assessment of retrieved document [12].

3. DATA COLLECTION and ANALYSIS

For this study, we used the *i*Search test collection [6], which consists of 18,442 English bibliographic records harvested from library catalogs, 143,571 articles in full-text PDF, and 291,246 abstracts and metadata from the open access portal arXiv.org. In this study we focus on the bibliographic book records and the PDFs to identify the difference in usefulness assessments between document representation types. The *i*Search also includes descriptions of 65 real-world work tasks from 23 academic researchers (hereafter task performers) in Physics from 3 universities in Denmark. The data were collected in several stages: the task performer filled out an online form with 5 questions to describe the task performers' work task situation in context: (1) What are you looking for? ("What do you want the system to find, e.g. in arxiv.org?") (2) What is the motivation of the topic? "Why are you looking for this, what problem/task can be solved with the information + in what context did the problem arise?") (3) What central search terms would you use to express your situation and information need? ("Please provide 2-3 relevant search terms:") (4) What is your background knowledge of this topic? ("What do you know that might further help the system understand your situation?) (5) What would an ideal answer look like? ("What should a perfect answer contain to solve your problem or task?")

Based on participants' descriptions of their information needs, the *i*Search team searched and retrieved documents for all tasks. To make the assessment manageable, up to 200 documents were provided to the task performer. Wherever possible, the distribution of retrieved document types was proportional to the collection distribution [6]. The search results were made available to the task performers who had submitted the descriptions of the search tasks and topics. The task performers assessed the usefulness of each retrieved document within one week using a dedicated Website. They were trained to use the Website. Usefulness was measured as highly, fairly, marginally, or not useful, based on Sormunen's relevance measurement scale [8].

As part of the post assessment questionnaire, task performers judged the degree of work task complexity and the level of topic specificity [4]. Perceived task complexity was measured as high, fair, marginal, or routine task (least complex); marginal or fairly complex tasks were combined into one category in analysis. Perceived topic specificity was assessed as high, fair, or generic (least specific). Therefore, both task complexity and topic specificity were measured in three levels. The following examples illustrate how a participant perceived topic specificity. Strehl ratio was regarded as highly specific, gouge group fairly specific, and solar wind as generic topic.

Overall, participants submitted between 2 to 5 search tasks. A total of 65 search tasks were carried out resulting in 11,066 documents that were assessed by the 23 participants; the mean number of assessed documents per search task was 170 (ranged from 18 to 200). In this paper the analysis was carried out on the total number of the assessed *i*Search documents.

4. FINDINGS AND DISCUSSION

We briefly present distributions of document usefulness (Tables 1 & 2) from Ingwersen and Wang [4] as a context for interpreting the new results in Tables 3-4.

4.1 Usefulness, task complexity & topic specificity

As in Table 1, only 26.0% of the 11,066 assessed documents were judged as useful in various degrees. Table 1 also suggests that the level of task complexity might have affected the performer's assessment of usefulness of the retrieved documents. Highly useful documents seemed to intersect with the highly complex search tasks (mean = 8.3, precision = 4.8%); overall, marginally useful documents showed a substantially better mean and higher precision than fairly useful documents; fairly useful documents showed a better mean and higher precision than highly useful documents. The results in Tables 3-4 indicate that both book records and PDFs contributed to this overlapping of highly complex tasks with highly useful documents.

There is, however, a slightly different association between document usefulness and search topic specificity as compared to task complexity. The most useful and the highest precision were consistently associated with the fairly (not highly) specific search topics. This might suggest that the perceived specificity of the search topic affected the performers' judgments of usefulness.

4.2 Effect of document type on precision

We compared the precision of the bibliographic records and PDF documents across different degrees of task complexity and topic specificity. The documents assessed as routine (least complex) and generic (least specific) were excluded due to the fact that no book records (or PDFs) were retrieved for these cases. The total number of search requests in Table 2 is thus smaller than in Table 1.

Overall, search precision for bibliographic records is much higher than for PDFs: 44% vs. 22.4% for tasks of high and fair complexity; 38.6% vs. 23.2% for tasks of high and fair specificity. Using the average precision of 26% as a benchmark (precision, last row, Table 1), the precision for bibliographic records strongly exceeds this comparison; the precision for full text PDFs is slightly below this comparison. On the other hand, the PDFs made much greater contributions to the useful documents than the bibliographic records; the ratio of contributions ranges between 3.3 and 4.6 to one. This should be observed against the fact that the ratio of PDFs versus book records in *i*Search is almost 9:1. Therefore, the task performers' judgment of more useful books was not in line with the odds for random selection.

In this study, the average precision (the percentage of useful retrieved documents) was generally low (26%), similar to the earlier study of real users' document selection behaviour [12]. Generally speaking, the more complex the tasks, or the more specific the topics were, the fewer the retrieved useful documents for both document types – Table 2.

In terms of precision of the documents assessed as highly useful, the more complex the tasks were, the higher the precision, Tables 3-4. This finding may seem counter intuitive; and might be interpreted that the task performers tended to be *less discriminative* when they perceived the tasks as highly complex. Perceived complexity may be influenced by knowledge of the task. It is worth further study to investigate if there is a causal relationship between the perception and knowledge state [4].

Additionally, the task performers evaluated fewer document information elements available in bibliographic records than in PDF metadata. There were both abstracts and full texts readily accessible for PDFs. Thus, participants were likely *less discriminative* when evaluating bibliographic records while the availability of the full text for PDF documents allowed the task performers to reach informed assessments of usefulness. Further studies should closely examine this point because it has implications for both system design and understanding of human behavior.

4.3 Document type, usefulness, nature of task

Tables 3 and 4 provide detailed summary of the assessments of usefulness associated with the two levels of task complexity and search topic specificity. For both document types we found that the highest mean number of highly useful documents (also the highest precision) is associated to the work tasks perceived as highly complex. This corroborates the finding (see Table 1) that there was an overlapping between highly useful documents and highly complex work tasks. For search topics of high specificity, a similar pattern was found, but only for the bibliographic records (Tables 3-4). In comparison, for book records, the precision for all levels of usefulness assessments was substantially higher than that for PDFs at all levels of complexity: P =4.9%; 13.6% and 25.4% for book records at the three usefulness levels, Table 3 vs. P = 2.0%; 5.1% and 17.35% for PDFs, Table 4. However, the number of assessed book records is rather small; thus it is fair to assume that some kind of *scarcity effect* may have occurred during the assessment process [10]. In terms of search topic specificity, the overlapping between highly useful documents and highly specific topics for books is also higher than that of PDFs (P = 20% and mean = 3 vs. P = 1.8% and mean = 1.8 for PDFs), which did not occur for other levels

The effect of document type (bibliographic references vs. full text) on relevance assessments has been reported in a previous study [9]. Our findings on topical specificity are inclusive, which calls for further studies.

5. CONCLUSIONS

In this exploratory study, we analyzed the data in the *i*Search Collection to identify relationships between usefulness assessment, the nature of work tasks and document types. The nature of work tasks has two dimensions: the perceived work task complexity and the perceived search topic specificity. Despite of the small number of book records, they contributed to a larger than expected proportion of the useful documents for both highly specific topics and highly complex tasks. In general, higher task complexity and topic specificity seemed to lead to high number of highly useful Bibliographic records. However, regardless of document types in general tasks of fair complexity as well as topics of fair specificity tended to have more fairly and marginally useful documents retrieved.

We found in addition that task performers tend to be less discriminative in their assessments when 1) they perceive the tasks as highly complex and 2) they evaluate bibliographic records, probably owing to fewer available document information elements compared to PDFs.

The preliminary findings in this study extend the previous results [3, 4]. The main limitation is the data collected from the naturalist experiment with little or no control. Therefore, not all potential factors were present or observed. Further studies should collect longitudinal data on how perceived nature of tasks changes over time and how these changes affect assessment. Precision is a classic measurement of IR; its value in the naturalistic IR should be evaluated. Alternative performance measurements such as task completeness, task satisfaction, and mean average precision (MAP) or normalized cumulated gain of ranked output from retrieval runs [5] should be considered.

6. REFERENCES

[1] K. Byström and K. Järvelin. Task complexity affects information seeking and use. *IPM*, 31(2):191-213, 1995.

[2] P. Ingwersen and K. Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series).* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[3] P. Ingwersen, M. Lykke, T. Bogers, B. Larsen, and H. Lund. Assessors' search result satisfaction associated with relevance in a scientific domain. In *Proceedings of the third symposium on Information interaction in context, IIiX '10*, pages 283-288, New York, NY, USA, 2010. ACM.

[4] P. Ingwersen and P. Wang. Relationship between usefulness assessments and perceptions of work task complexity and search topic specificity: An exploratory study. In B. Larsen, C. Lioma, and A. P. de Vries, editors, *TBAS2012*, pages 19-23. http://ceur-ws.org, 2012.

[5] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422-446, 2002.

[6] M. Lykke, B. Larsen, H. Lund, and P. Ingwersen. Developing a test collection for the evaluation of integrated search. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. M. Rüger, and K. van Rijsbergen, editors, *ECIR*, volume 5993 of *Lecture Notes in Computer Science*, pages 627-630. Springer, 2010.

[7] T. Saracevic. Relevance reconsidered. In P. Ingwersen and N.O.Pors, editors, *Information science: Integration in perspectives. Proceedings of the Second Conference on Conceptions of Library and Information Science.* Copenhagen (Denmark), pages 201-218, 1996.

[8] E. Sormunen. Liberal relevance criteria of trec -: counting on negligible documents? In *SIGIR*, pages 324-330. ACM, 2002.

[9] P. Vakkari. Relevance and contributing information types of searched documents in task performance. In *SIGIR*, pages 2-9, 2000.

[10] P. Wang. *A cognitive model of document selection of real users of information retrieval systems*. PhD thesis, Information Science, University of Maryland, 1994.

[11] P. Wang. Information Behavior and Seeking, In D. Kelly and I. Ruthven, editors, *Information Retrieve Interaction: Interactive Information Seeking and Retrieval,* pages 15-42. London: Facet, 2011.

[12] P. Wang and D. Soergel. A cognitive model of document use during a research project. Study I. Document Selection. *JASIS*, 49(2):115-133, 1998.

Table 1: Document usefulness, task complexity, and task topic specificity (from [4]). prec. = precision. Bold = max. value per usefulness degree. The last row gives the total, mean or average.

Task Complexity		**Document Usefulness**								
		high			***fair***			***marginal***		
degree	tasks	# useful docs.	mean # useful docs.	precision	# useful docs.	mean # useful docs.	precision	# useful docs.	mean # useful docs.	precision
High	24	199	**8.3**	**4.8%**	226	9.4	5.4%	561	23.4	13.4%
Fair	36	129	3.6	2.1%	428	**11.9**	**6.9%**	1295	**36.0**	**20.9%**
Low	5	9	1.8	1.3%	12	2.4	1.7%	19	3.8	2.7%
	65	337	5.2	3.0%	666	10.2	6.0%	1875	28.8	16.9%

Task Topic Specificity		**Document Usefulness**								
		high			***fair***			***marginal***		
degree	tasks	# useful docs.	mean # useful docs.	precision	# useful docs.	mean # useful docs.	precision	# useful docs.	mean # useful docs.	precision
High	43	223	5.2	3.0%	362	8.4	4.8%	1124	26.1	15.0%
Fair	19	102	**5.4**	**3.1%**	291	**15.3**	**8.7%**	731	**38.5**	**22.0%**
Low	3	12	4.0	4.6%	13	4.3	4.9%	20	6.7	7.6%
	65	337	5.2	3.0%	666	10.2	6.0%	1875	28.8	16.9%

Table 2: Comparison of Performances of Bibliographic Records and PDF Documents. Max. values in bold.

Task Complexity		**Bibliographic Records**				**PDFs**				
degree	tasks	# assessed docs.	# useful docs.	mean # useful docs.	precision	tasks	# assessed docs.	# useful docs.	mean # useful docs.	precision
High	20	314	135	6.8	43.0%	23	2470	489	21.3	19.8
Fair	31	589	262	**8.5**	**44.5%**	34	3161	873	**26.5**	**27.6**
	51	903	397	7.8	44.0%	57	5631	1362	23.9	22.4

Task topic specificity										
degree	tasks	# assessed docs.	# useful docs.	mean # useful docs.	precision	tasks	# assessed docs.	# useful docs.	mean # useful docs.	precision
High	35	525	176	5.0	33.5%	40	4037	809	20.2	20.0%
Fair	17	374	171	**10.1**	**45.7%**	18	1806	567	**31.5**	**31.4%**
	52	899	347	6.7	38.6%	58	5933	1376	23.7	23.2%

Table 3: Book records. Degree of usefulness vs. task complexity and topic specificity. Max. values in bold.

Task Complexity		**Document Usefulness (Book Records)**								
		high			***fair***			***marginal***		
degree	tasks	# useful docs.	mean # useful docs.	precision	# useful docs.	mean # useful docs.	precision	# useful docs.	mean # useful docs.	precision
High	20	23	**1.2**	**7.3%**	34	1.7	10.8%	78	3.9	24.8%
Fair	31	22	0.7	3.7%	89	**2.9**	**15.1%**	151	**4.9**	**25.6%**
	51	45	0.9	4.9%	123	2.4	13.6%	229	4.5	25.4%

Task Topic Specificity		**Document Usefulness (Book Records)**								
		high			***fair***			***marginal***		
degree	tasks	# useful docs.	mean # useful docs.	precision	# useful docs.	mean # useful docs.	precision	# useful docs.	mean # useful docs.	precision
High	35	105	**3.0**	**20.0%**	21	0.6	4.0%	50	1.4	9.5%
Fair	17	20	1.2	5.3%	67	**3.9**	**18.1%**	84	**4.9**	**22.5%**
	52	125	2.4	13.9%	88	1.7	9.8%	134	2.6	14.9%

Table 4: PDF full texts. Degree of usefulness vs. task complexity and topic specificity. Max. values in bold.

Task Complexity		**Document Usefulness (PDFs)**								
		high			***fair***			***marginal***		
degree	tasks	# useful docs.	mean # useful docs.	precision	# useful docs.	mean # useful docs.	precision	# useful docs.	mean # useful docs.	precision
High	23	70	**3.0**	**2.8%**	108	4.7	4.4%	311	13.5	12.6%
Fair	34	33	1.0	1.0%	177	**5.2**	**5.6%**	663	**19.5**	**21.0%**
	57	103	1.8	2.0%	285	5.0	5.1%	974	17.1	17.3$

Task Topic Specificity		**Document Usefulness (PDFs)**								
		high			***fair***			***marginal***		
degree	tasks	# useful docs.	mean # useful docs.	precision	# useful docs.	mean # useful docs.	precision	# useful docs.	mean # useful docs.	precision
High	40	72	1.8	1.8%	151	3.8	3.7%	586	14.7	14.5%
Fair	18	41	**2.3**	**2.3%**	136	**7.6**	**7.5%**	390	**21.7**	**21.6%**
	58	113	1.9	1.9%	287	4.9	4.8%	976	16.8	16.5%