In: *Proceedings of the ECIR 2012 Workshop on Task-Based and Aggregated Search (TBAS2012)*. Larsen, B., Lioma, C. & de Vries, A. P. (red.), 1 Apr. 2012: 19-23.

Relationship between Usefulness Assessments and Perceptions of Work Task Complexity and Search Topic Specificity: An Exploratory Study

Peter Ingwersen* and Peiling Wang**

*Royal School of Library and Information Science, Birketinget 6, DK 2300 Copenhagen S, Denmark

*University Carlos III Madrid, Calle Madrid, 126, 28903 Getafe (Madrid), Spain,

**University of Tennessee Knoxville, TN 37923, USA

*pi at iva.dk; **peilingw at utk.edu

ABSTRACT

This research investigates the relations between the usefulness assessments of retrieved documents and the perceptions of task complexity and search topic specificity. Twenty-three academic researchers submitted 65 real task-based information search topics. These task topics were searched in an integrated document collection consisting of full text research articles in PDFs, abstracts, and bibliographic records (the *i*Search Test Collection in Physics). The search results were provided to the researchers who, as task performers, made assessments of usefulness using a four-point sale (highly, fairly, marginally, or not useful). In addition, they also assessed the perceived task complexity (highly, fairly, and routine/low) and the perceived specificity of the search topic (highly, fairly, and generic/low). It is found that highly specific topics associate with all degrees of task complexity, whereas highly complex tasks tend to associate with search topics of high specificity. Although bibliographic records show better precisions than full text PDF documents, the latter contributed more useful documents. Suggestions are made for further studies in naturalist IR experiments.

Categories and Subject Descriptors

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

General Terms

Performance, Human Factors.

Keywords

Task-based IR; Integrated search; Task complexity; Task specificity

1. INTRODUCTION

Information retrieval (IR) is a sub-process of information-seeking (IS) processes, during which a selected retrieval system is used to find useful information for the task at hand [1]. It is useful to differentiate work tasks from IR 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 on-going dissertations, funded research projects, etc. The relevant information will support the task performers in moving forward with the work tasks [2]. In this study, we analysed real users in work task situations, their judgments of usefulness of retrieved documents, and their perceptions of their work tasks.

In professional settings, Bystrom & Jarvelin argue that a central property of work tasks is the perceived complexity that influences task performers' information-seeking behaviours [3]. The more complex a work task is, the greater the probability that the retrieval result is meagre and the greater the tendency that the task performer will seek information beyond formal IR systems. In IR, the perceived specificity of search topics plays an important role in the task performers' evaluations of the search results [2]. The more specific the topic is perceived to be, the more selective the task performer will be. Therefore, the nature of the work task and the search task can be measured by task complexity and topic specificity.

There has not been extensive research on how relevance assessment is influenced by perceived task complexity and topic specificity in integrated digital IR systems. Integrated digital IR systems, empowered by aggregated search engines, provide access to a variety of information objects from multiple sources. The purpose of this research is twofold: to observe real users' evaluations of integrated system search results and to relate their evaluations with the nature of their work tasks. Specifically, this research analysed the data in the *i*Search[1] collection (see section 2) to address the following research questions:

Q1. How do task performers perceive the nature of their work tasks in terms of task complexity and topic specificity?

Q2. How do task performers assess the usefulness of retrieved documents with regard to task complexity? What is the relationship between the usefulness and task complexity?

Q3. How do task performers assess the usefulness of retrieved documents with regard to topic specificity? What is the relationship between the usefulness and topic specificity?

Q4. How do task performers' assessments of usefulness relate to the nature of the work task? In other words,

34th ECIR Conference, April 1-5, 2012, Barcelona, Spain.

[1] *i*Search data are freely available to researchers at http://itlab.dbit.dk/~isearch

how do task complexity and topic specificity *jointly* relate to usefulness assessments?

Q5. 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?

2. DATA, MEASUREMENT, and ANALYSIS

The *i*Search Collection is an integrated IR system on Physics supported by the Denmark Electronic Library and consists of 18,000+ English bibliographic records, 160,000+ articles in full-text PDF, and 291,000+ abstracts and bibliographic records harvested from the open access portal arXiv.org and library catalogs. The system developed a search engine powered by a vector space model. Additionally, the collection includes the test searches for the 65 real-world tasks from 23 researchers (hereafter *task performers*) who were faculty and graduate students in Physics from three universities in Denmark. The data were collected in several stages: the task performer filled out an online form with five-questions to describe the task performers' information needs in their work contexts:

1. What is the search topic?
2. What is the task situation?
3. What are the key search terms?
4. What is the task performer's knowledge status of the topic?
5. What are the expected (ideal) search results?

Below is an example of the description of task and search topic, and the perceived task complexity and search topic specificity:

> 1. I'm looking for information on possible ways to achieve significant slip-lengths in micro- and nanochannels.
> 2. In two months I'll begin my PhD studies. Part of my studies is in collaboration with an experimental group at the University of California, Santa Barbara. In this group, they envision using pressure driven flows in nanochannels to generate electricity. The central effect in this project is streaming current, where charge in the electric double layer is converted by the pressure flow. The main dilemma is that the electric double layer is situated close to the nanochannel walls, whereas convection from the pressure driven flow peaks in the middle of the channel. Usually, there is a no-slip condition at the nanochannel walls and thus the convection is smaller in this region. Therefore, if we can increase the slip-length we will increase the convection close to the walls and in the electric double layer. Consequently, we can increase the efficiency of the pressure-to-electricity conversion, which at present is too low for practical use of the technology.
> 3. Graphene, slip-length, slip-velocity, nanochannel, streaming current, power generation
>
> Task complexity: *High*
>
> Topic specificity: *High*

The 23 participants, as task performers, provided descriptions of 65 search tasks, of which 8 people submitted 2 tasks; 12 people submitted 3 tasks; 2 people submitted 4 tasks; and 1 person submitted 5 tasks. The *i*Search team conducted searches and provided the results for all 65 tasks. To make the assessment manageable, the top 200 documents in search output were provided to the task performer. Thus, the sets of search results ranged from 18 to 200 documents.

Each set of search results was provided to the task performer who had submitted the description of the search task and topic. Each task performer then assessed the *usefulness* of each retrieved document in the provided set within one week. This assessment was completed using a web program which the task performers were trained to use. The degree of usefulness is measured as *highly, fairly, marginally,* or *not useful*. This scale is based on Sormunen's relevance measurement [5; 6]. Additionally, the task performers assessed the nature of their tasks: the degree of task complexity and the level of topic specificity. The degree of perceived task complexity was measured as *high*, *fair*, *marginal*, or *routine task* (least complex); *marginal* tasks were merged with *fairly* complex tasks in the analysis of this paper. Routine tasks are assumed to have the lowest degree of complexity. The level of perceived topic specificity was assessed as *high*, *fair*, or *generic* (least specific).

In summary, the collected data included the 65 search tasks from 23 real users who submitted between 2 to 5 search tasks. A total of 11,066 documents were retrieved; the mean number of retrieved documents per search task is 170 (ranges from 18 to 200). The retrieved documents were assessed by the task performer who submitted the search request. For each search request, the raw data included the degree of task complexity, the level of topic specificity, the total number of retrieved documents, and the assessment of usefulness of each document.

It is important to point out that because the participants of this research were real users, the collected data were naturalistic. The nature of their work tasks varied from highly complex tasks with highly specific topics (21 cases) to routine tasks with generic topics (2 cases); but there was no case of a highly complex task with a generic topic (Table 1). Due to the varied numbers of cases across different situations, the analysis reported in this paper is exploratory, rather than hypothesis-driven. The analysis aims to identify the relationships between the assessed usefulness of the retrieved documents and the perceived nature of the tasks.

3. RESULTS

The results will be reported in the order of the research questions.

3.1 Perceived task complexity and topic specificity

Table 1 summarizes the nature of the 65 task-based search requests: 24 *highly complex* tasks (37%), 36 *fairly complex* tasks (55%), and 5 *routine* tasks (8%); 43 *highly specific* search topics (66%), 19 *fairly specific* search topics (29%), and 3 *generic* topics (5%). Notably, the majority of these tasks fall into the four cells that combine high and fair task complexity with high and fair topic specificity (the shadowed 59 cases in Table 1). This set of tasks will be further analyzed below and in Section 3.4.

Table 1. Nature of Tasks

Specificity \ Complexity	High	Fair	Routine	*Total*
High	**21** (7)	**21** (6)	**1** (1)	***43*** *(14)*
Fair	**3**	**14** (3)	**2** (2)	***19*** *(5)*
Generic	0	**1**	**2** (1)	***3*** *(1)*
Total	***24*** *(7)*	***36*** *(9)*	***5*** *(4)*	***65*** *(20)*

Number in ( ) is the number of difficult retrieval tasks

The numbers in parenthesis note the number of difficult search tasks, 20 (32%) that retrieved fewer than 15 useful documents [4].

There is obviously an overlapping of highly complex tasks with highly specific search topics as well as fairly complex tasks with

highly specific search topics (both cells have 21 cases, Table 1). There is a reasonable overlapping between fairly complex tasks and fairly specific topics (14 cases). In other words, the highly and fairly complex tasks tend to be highly specific topics, while fairly complex tasks tend to be fairly specific topics. This trend can be visualized also in Figure 1, in which the size of the data points corresponds to the number of cases that intersect with degrees of task complexity and levels of topic specificity.

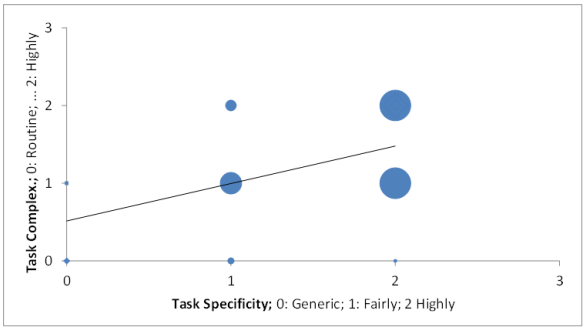

Fig. 1. Scatter Plot of Cases of Complexity and Specificity. Size of data point indicates number of cases.

With regard to the 20 difficult search tasks, the results indicate that routine tasks were the most difficult to retrieve in the *i*Search collection (4 out of 5, or 80%). The rest of the 16 difficult search tasks were more or less evenly distributed: 7 were tasks of high complexity and high specificity (33%); 6 were tasks of fair complexity and high specificity (28%); and 3 were tasks of fair complexity and fair specificity (23%).

3.2 Assessment of document usefulness and task complexity

Table 2a summarizes the assessment results of usefulness across different task complexities. The mean number of retrieved documents for different complexity degrees of tasks is 170 (ranges between 140 and 174). Only 26.0% of the total retrieved documents were assessed as useful of various degrees. This means that the search precision is substantially low, especially for routine tasks: only 5.7% useful.

Table 2a. Assessment of Document Usefulness and Task Complexity

Task Complexity		*All Retrieved*		*All Useful*			*All Not Useful*		
Degree	Cases	Total	Mean	Total	Mean	Precision	Total	Mean	False drop
High	24	4178	174	986	41.1	23.6%	3192	133	76.4%
Fair	36	6188	172	1852	51.4	29.9%	4336	120.4	70.1%
Routine	5	700	140	40	8.0	5.7%	660	132.2	94.3%
Σ / Mean / P	**65**	***11066***	**170**	**2878**	**44.3**	**26.0%**	**8188**	**126.0**	**74.0%**

The mean number of useful documents for highly complex tasks is 41 (precision = 23.6%); for fairly complex tasks is 51 (precision = 29.9%); for routine tasks is 8 (precision = 5.7%). This suggests that the level of task complexity might have affected the performer's assessment of usefulness of the retrieved documents. The *routine* tasks resulted in the smallest mean number of useful documents and the lowest precision. Chi square testing (χ^2 = 341.718; alpha = 0.03) indicates that a correlation does exist between retrieval performance, measured by search precision, and work task complexity, measured by degree of complexity.

Table 2b presents the degree of usefulness across the various task complexities. There were fewer documents assessed highly useful (mean = 5; precision 3.0%) than either fairly (mean = 10; precision = 6.0%) or marginally useful (mean = 29; precision = 16.9%). In each category of usefulness assessment results, the means and precisions vary across different degrees of task complexity. Therefore, the highest mean and precision can be identified as follows: the highly useful documents intersect with the highly complex search tasks (mean = 8, precision = 4.8%); fairly useful documents intersect with the fairly complex search tasks (mean = 12, precision = 6.9%); marginally useful documents intersect with the fairly complex search tasks (mean = 36, precision = 20.9%). Overall, marginally useful documents show a substantially better mean and higher precision than fairly useful documents, and fairly useful documents show a better mean and higher precision than highly useful documents.

Table 2b. Degree of Document Usefulness and Degree of Task Complexity

Task Complexity		*Highly Useful*			*Fairly Useful*			*Marginally Useful*		
Degree	Cases	Count	Mean	Precision	Count	Mean	Precision	Count	Mean	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%
Routine	5	9	1.8	1.3%	12	2.4	1.7%	19	3.8	2.7%
Σ / Mean / P	**65**	**337**	**5.2**	**3.0%**	**666**	**10.2**	**6.0%**	**1875**	**28.8**	**16.9%**

3.3 Assessment of document usefulness and topic specificity

Table 3a summarizes the assessment results of usefulness across different levels of topic specificity. The mean number of retrieved documents for different topic specificity levels of tasks is 170 (ranges between 88 and 174). In comparison to Table 2a, Table 3a has the same aggregated results (the last row).

The mean number of useful documents for highly specific topics is 40 (22.9% precision); for fairly specific topics is 59 (33.8% precision); for generic topics is 15 (17.1% precision). This suggests that the specificity of the topic may affect the performer's judgments of usefulness. The generic topics resulted in the smallest mean number of useful documents and the least precision. The fairly specific topics resulted in the greatest mean number of useful documents and the highest precision. Chi square testing (χ^2 = 153.868; alpha = 0.01) indicates the existence of a correlation between retrieval performance, measured by search precision, and search topic specificity, measured by level of specificity.

Table 3a. Assessment of Document Usefulness and Task Topic Specificity

Task Topic Specificity		*All Retrieved*		*All Useful*			*All Not Useful*		
Degree	Cases	Total	Mean	Total	Mean	Precision	Total	Mean	False drop
High	43	7476	174	1709	39.7	22.9%	5767	134.1	77.1%
Fair	19	3327	175	1124	59.2	33.8%	2203	115.9	66.2%
Generic	3	263	88	45	15.0	17.1%	218	72.7	82.9%
Σ / Mean / P	**65**	***11066***	**170**	**2878**	**44.3**	**26.0%**	**8818**	**126.0**	**74.0%**

Table 3b presents the degree of usefulness across the various levels of specificity. Although fewer retrieved documents were highly useful (see also Table 2b), the means and precisions of highly useful documents across different levels of specificity are less diverse (mean ranges between 4.0 and 5.4; precision ranges between 3.0% and 4.6%). In each category of usefulness

assessment results, the highest mean and precision can be identified as follows: the highly useful documents intersect with the fairly specific search topics (mean = 5, precision = 3.1%); fairly useful documents intersect with the fairly specific search topics (mean = 15, precision = 8.7%); marginally useful documents intersect mostly with the fairly specific search topics (mean = 39, precision = 22.0%). This suggests that the greatest mean and the highest search precision for all the degrees of usefulness were associated with the fairly specific search topics.

Table 3b. Degree of Document Usefulness and Level of Task Topic Specificity

Task Topic Specificity		*Highly Useful*			*Fairly Useful*			*Marginally Useful*		
Degree	**Cases**	**Count**	**Mean**	**Precision**	**Count**	**Mean**	**Precision**	**Count**	**Mean**	**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%
Generic	3	12	4.0	4.6%	13	4.3	4.9%	20	6.7	7.6%
Σ / Mean / P	**65**	**337**	**5.2**	**3.0%**	**666**	**10.2**	**6.0%**	**1875**	**16.9**	**16.9%**

3.4 Usefulness assessment of retrieved documents in relation to the nature of tasks

This analysis focuses on the relationship between the usefulness assessment of search results and the perception of the tasks. As mentioned in 3.1, this analysis focuses on a subset of 59 tasks excluding all routine tasks and all generic topics (See Table 1), which counted for a total of 6 cases. These tasks are classified by both task complexity and topic specificity (Table 1):

Category 1. *Highly complex* tasks with any topic specificity
Category 2. *Fairly complex* tasks with *highly specific* topics
Category 3. *Fairly complex* tasks with *fairly specific* topics

The summary data for the three categories are in Table 4a. Each category retrieved similar numbers of documents (mean = 175; range between 174 and 178). With regard to useful documents, Category 3 has the highest mean (67) and highest precision (37.8%) while the other two categories have similar means (41 and 42 respectively) and similar precisions (23.6% and 24.3% respectively).

Table 4a. Assessment of Document Usefulness and Nature of Tasks

Joint task complexity and topic specificity		*All Retrieved*		*All Useful*			*All Not Useful*		
	Cases	**Total**	**Mean**	**Total**	**Mean**	**Precision**	**Total**	**Mean**	**False drop**
Category 1	24	4178	174	986	41.1	23.6%	3192	133.0	76.4%
Category 2	21	3656	174	889	42.3	24.3%	2767	131.8	75.7%
Category 3	14	2487	178	940	67.1	37.8%	1547	110.5	62.2%
Σ / Mean / P	**59**	**10321**	**175**	**2815**	**47.7**	**27.3%**	**7506**	**127.1**	**72.9%**

A close look at the degrees of usefulness across three categories of tasks (Table 4b) reveals: for the highly useful documents, Category 1 (the highly complex tasks with search topics of any specificity) has the best results (mean = 8; precision = 4.8%); for the fairly, Category 3 (the fairly complex tasks with fairly specific topics) has the best results (mean = 17; precision = 9.7%); similarly, for marginally useful documents, Category 3 also shows the best results (mean = 45; precision = 25%). Since both Category 2 and Category 3 have the same degree of task complexity (*fair*), the observed differences in degrees of usefulness assessments are likely due to topic specificity. In other words, for the *fairly* complex tasks, the higher the topic specificity, the lower the retrieval performances (See Table 4b, the two rows for Category 2 and Category 2). This observation corroborates the data in Table 3b.

Table 4b. Degree of Document Usefulness and Nature of Tasks

Joint task complexity and topic specificity		*Highly Useful*			*Fairly Useful*			*Marginally Useful*		
	Cases	**Count**	**Mean**	**Precision**	**Count**	**Mean**	**Precision**	**Count**	**Mean**	**Precision**
Category 1	24	199	8.3	4.8%	226	9.4	5.4%	561	23.4	13.4%
Category 2	21	60	2.9	1.6%	179	8.5	4.9%	650	31.0	17.8%
Category 3	14	65	4.6	2.6%	241	17.2	9.7%	634	45.3	25.5%
Σ / Mean / P	**59**	**324**	**5.5**	**3.1%**	**646**	**10.9**	**6.3%**	**1845**	**31.3**	**17.9%**

3.5 Effect of document type on retrieval precision

The *i*Search is an integrated IR system with a collection of different types of documents, including bibliographic records, full text PDF articles, etc. This analysis compares the performances of the two types of documents, bibliographic records and PDF documents, across different degrees of complexity and different levels of topic specificity.

Table 5a. Comparison of Performances of Bibliographic Records and PDF Documents

Nature of Tasks	Bibliographic Records					PDF Documents				
(a) Complexity	**Cases**	**Total Retrieved**	**Total Useful**	**Mean Useful**	**% of Useful**	**Cases**	**Total Retrieved**	**Total Useful**	**Mean Useful**	**% of Useful**
High	20	314	135	6.8	43.0	23	2470	489	21.3	19.8
Fair	26	491	201	7.7	40.9	28	2693	670	23.9	24.9
Total	46	805	336	7.3	41.7	51	5163	1159	22.7	22.4
(b) Topic Specificity	Cases	Total retrieved	Total Useful	Mean Useful	% of Useful	Cases	Total retrieved	Total Useful	Mean Useful	% of Useful
High	35	525	176	5.0	33.5	40	4037	809	20.0	20.0
Fair	17	374	171	10.1	45.7	18	1806	567	31.5	31.4
Total	52	899	347	6.7	38.6	58	5933	1376	23.7	23.2

Overall search precisions for bibliographic records are much higher than those for PDF documents: 41.7% vs. 22.4 % for tasks of high and fair complexity; 38.6% vs. 23.2% for tasks of high and fair specificity. Using the grand mean precision of 26% as a benchmark (Table 2a and Table 3a), the precisions for bibliographic records exceeded benchmark; the precisions for full text PDF documents were slightly below the benchmark.

On the other hand, the full text PDF documents made much greater contributions to the useful documents than the bibliographic records did; the ratio of contributions ranges between 3.3 and 4.6 (Table 5b).

Table 5b. Comparison of Contributions of Document Types

Nature of Tasks	Total Useful Documents	Useful biblio. records	Contribution	Useful PDF	Contribution	Ratio of Contributions
Complexity High	986	135	13.69%	489	49.59%	3.6
Complexity Fair	1852	201	10.85%	670	36.18%	3.3
Specificity High	1709	176	10.30%	809	47.34%	4.6
Specificity Fair	1124	171	15.21%	567	50.44%	3.3

4. DISCUSSION and CONCLUSIONS

In this exploratory study, we analyzed the data in the *i*Search Collection to identify relationships between usefulness assessments and the nature of work tasks. The nature of work tasks has two dimensions: the perceived task complexity and the perceived search topic specificity.

4.1 Perception of the nature of the tasks (Q1)

The majority of the submitted requests were for complex tasks (high or fair) with specific topics (high or fair). Highly complex tasks associate with highly specific topics; fairly complex tasks also associate with highly specific topics. This observation suggests that when researchers approach IR systems, most likely they need to perform complex tasks with highly specific topics. The fact that the difficult search tasks distributed over various complexities and specificities suggests that the *i*Search has stable search performances except for routine tasks. Almost all routine tasks were difficult retrieval tasks in the *i*Search.

4.2 Assessment of usefulness with regard to task complexity (Q2)

In this study, the percentages of useful retrieved documents are generally low, which is similar to the study of real users' document selection behavior [7]. Generally speaking, the more complex the tasks were, the fewer the retrieved useful documents. Routine tasks had the smallest means and the lowest precisions. However, the documents assessed as *highly useful* had a different pattern: the higher the task complexity, the higher the mean and the precision. One interpretation may be that the task performers tend to assess documents less discriminative when they perceive the tasks as highly complex. Perception of complexity may be influenced by knowledge of the task. Although a statistically significant correlation is found between the assessment of usefulness and the perceived complexity of tasks, it is not clear if there is a causal relationship between the two.

4.3 Assessment of usefulness with regard to topic specificity (Q3)

Evidently, the higher the topic specificity is, the fewer the retrieved useful documents with the exception of the generic tasks. Overall, the tasks with fairly specific topics performed the best: the greatest means and the highest precisions. One explanation may be that the *i*Search is more useful to retrieve fairly specific topics because of its design of indexes or access points. Although a statistically significant correlation is found between the assessment of usefulness and the perceived specificity of search topics, it is not clear if there is a causal relationship between the two.

4.4 Assessment of usefulness with regard to both task complexity and topic specificity (Q4)

The tasks of fair complexity with fair topic specificity show the best retrieval performance. Tasks of fairly specific topics have overall better performances. Although both dimensions of tasks influence assessments, the *perceived specificity* of search topics *has more influence* than the perceived task complexity.

4.5 Document types and contributions (Q5)

Although the bibliographic records achieved higher means and precisions than full text PDF documents, the latter contributed substantially more to the retrieved useful documents. This observation may suggest that the task performers are likely less discriminative towards bibliographic records due to the absence of the entire work. The availability of the full text PDF documents allows the task performers to reach more accurate assessments. The effect of document types on assessment results has been reported in a previous study [10].

The preliminary findings in this study extend the previous results of *i*Search Collection [8]. 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 [9] should be considered.

REFERENCES

[1] Wang, P. 2011. Information Behavior and Seeking. In Diane Kelly and Ian Ruthven (eds.), *Information Retrieve Interaction. Interactive Information Seeking and Retrieval.* London: Facet Publishing, 15 – 42.

[2] Ingwersen, P. and Järvelin, K. 2005. *The Turn: Intergration of Information Seeking and Information Retrieval in Context*. Springer.

[3] Byström, K. and Järvelin, K. 1995. Task complexity affects information seeking and use. *Information Processing and Management*, 31,191-213.

[4] 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.

[5] Kekäläinen, J. and Järvelin, K. 2002. Using graded relevance assessments in IR evaluation. *Journal of the American Society for Information Science and Technology*, 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] Wang, P. and Soergel, D. 1998. A cognitive model of document use during a research project. Study I: Document selection. *Journal of the American Society for Information Science 49*(2), 115-133.

[8] Ingwersen, P., Lykke, M., Bogers, T., Larsen, B. and Lund, H. 2010. Assessors' search result satisfaction associated with relevance in a scientific domain. In: *Proceeding of the third symposium on Information interaction in context (IIiX 2010).* ACM, New York, NY, USA, 283-287.

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

[10] Vakkari, P. 2000. Relevance and contributing information types of searched documents in task performance. In: Belkin, N.J. et al. (Eds.) *Proceedings of the 23rd ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM Press, New York, NY, 2-9.