Physicists’ information tasks: structure, length and retrieval performance

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

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

3rd Author
3rd author's affiliation
1st line of address
Telephone number, incl. country code
3rd author's email address

4th Author
4th author's affiliation
1st line of address
Telephone number, incl. country code
4th author's email address

ABSTRACT

In this poster, we describe central aspects of 65 natural information tasks from 23 senior researchers, PhDs, and experienced MSc students from three different university departments of physics. We analyze 1) the main purpose of the information task, 2) which and how many search facets were used to describe the tasks, 3) what semantic categories were used to express the search facets, and 4) retrieval performance. Result show variety in structure and length across tasks and task purposes. The results indicate effect of length and, in particular, task purpose on retrieval performance of different document description levels that should be examined further.

Categories and Subject Descriptors

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

General Terms

Performance, Human Factors.

Keywords

Information tasks, retrieval performance, search facets

1. INTRODUCTION

As digital libraries offer access to increasingly large and diverse information sources there is a need to evaluate integrated search that cover various document types, levels of metadata, and vocabularies.

IR systems evaluation is addressed from two quite different perspectives: the system-driven and the user-oriented perspectives [1]. Systems-oriented evaluation takes place in laboratory environments with predesigned queries; expert-generated, static, binary relevance assessments and with experimental control. The user-oriented evaluation takes a semi-laboratory/semi-real-life approach, uses both simulated and genuine user information needs, non-binary relevance judgements, and seeks realism as well as experimental control. In both cases a test collection for experiments with integrated search requires the following as a minimum: a corpus with several different document types, several levels of descriptions, appropriate information tasks from users with real needs (for greater realism), and relevance assessments with adequate amount of relevant documents for each type and optionally graded relevance assessments.

We have developed a test collection that supports system-driven as well as user-oriented evaluation, based on genuine work task situations, real information tasks, and non-binary relevance judgements [2]. The test collection consists of approx. 18,000 book records, 160,000 full-text articles, and 275,000 metadata records with varied set of metadata and vocabularies from the physics domain. The scientific domain of physics comprises a realistic case with longstanding traditions for self-archiving of research publications in open access repositories and information sharing between scholarly and professional environments [3].

We elicited 65 natural information tasks from 23 senior researchers, PhD students, and experienced MSc students from three different university departments of physics. For each task a set of up to 200 documents per task was retrieved for relevance assessments with each document type represented proportional to the corpus distribution. Participants were asked to fill out a post-assessment questionnaire on satisfaction with the assessment procedure and search results for each task.

The present paper investigates central aspects of the captured information tasks. The purpose is twofold. We want to extend our understanding of physicists’ information tasks in general, and more specifically we want insight into the nature and characteristics of the tasks in order to guide design of IR experiments in the test collection. We particularly address the following research questions:

1) What was the overall purpose of the information tasks?
2) What types of search facets were used to articulate the information tasks (structure)?
3) How many search facets were used to articulate the information tasks (length)?
4) How was the retrieval performance?

2. RESEARCH DESIGN

To answer the questions, we analysed the captured task descriptions. The information description form had five questions, in line with the form used by [4]:

a) What are you looking for?
b) Why are you looking for this?
c) What is your background knowledge of this topic?
d) What should an ideal answer contain to solve your problem or task?
e) Which central search terms would you use to express your situation and information need?

Questions (b) – (c) correspond to questions asked in [4], with (b) being about the underlying work task situation or context, and (c) about the current knowledge state. Question (a) asks about the formulation of the current information need, and (d) correspond to the 'Narrative' section common to TREC topics whilst (e) asks for perceived adequate search terms.

The task descriptions were captured in online forms via computers located in participants' own university environment. Prior to describing their task details the participants were briefed about the project objectives and the structure and purpose of the form. After filling out the forms they answered an online questionnaire concerning their personal data, domain knowledge, and retrieval experience with IR systems. Two months after task creation, access to a web-based relevance assessment system was opened. This system allowed 1) access to the set of documents to be assessed (sorted randomly within each document type), presented in overview form and with the possibility of opening full text PDFs where available, and 2) assigning relevance scores according to the following 4-point scale: highly, fairly, marginally, and non-relevant [5]. The assessment period was set to one week. Documents could be re-assessed if the test person chose to. A post-assessment questionnaire on satisfaction with the assessment procedure and search results was filled out for each task.

In order to obtain insight about task characteristics we coded central aspects of the descriptions: 1) the main purpose of the information task, 2) which and how many single search facets were used to describe the tasks, 3) what semantic category of terms were used to express the search facets, and 3) retrieval performance of task length and purpose.

We performed test searches for each information task, based on search terms from task form question e). Search performance for each task was measured by use of the metric normalized discounted cumulative gain (nDCG). We based the calculation of on the task creator's relevance judgments. The following variables applied to answering the research questions.

Table 1: Variables of the study

Variable	Definition and measurement
Task purpose	Main goal of information task; e.g. finding background information, information about techniques and methods. Exclusive coding, identification of one main purpose per task
Structure of information task description	Search facets used to describe the information task, e.g. common topic, method used, time, type of information. One count per single facet
Length of information task description	Number of single search facets per task form question
Semantic difference	Semantic differences across task form questions, e.g. introduction of different search facets in different questions
Vocabulary variation	Vocabulary variations used to express the facets, e.g. synonym variations, expressions at broader or narrower hierarchical level
Educational level of task	Educational level for the information task, e.g. information task in relation to master thesis, PhD. dissertation, or senior research
Institutional affiliation	Participants' affiliation, expressed by university
Information need type	Type of information need: known item, known topic, or muddled topic information need
Retrieval performance	Discounted cumulative gain (DCG) per information task. Search based on search terms from task form question e)

3. RESULTS

The 65 information tasks originated from 23 physicists from three different universities; 12 from <anonymised> University (UNI1), 32 from the <anonymised> (UNI2), and 21 from <anonymised> (UNI3). 4 tasks derived from 2 senior researchers, 25 from 8 PhD students, and 36 from 13 experienced MSc students.

The tasks were all topical, conscious information needs (100%). The tasks represent three categories of task purpose. 54% of the participants look for *theoretical background information*, e.g. "Descriptions of models and theory concerning passive mode-locking in linear cavities" (Task17), 26% look for *previous results and findings*, e.g. "In particular I look for results obtained by using a Fourier Split Step Method for solving the non-linear Schrödinger equation" (Task3), and 20% for *design of research methodology*, e.g. "Tables, graphs and figures with comparisons of different energy harvesting techniques" (Task5).

The participants described the tasks with structural variation and detail. Only 4 facet categories appeared in all descriptions, and were elicited by all task form questions. As shown in table 2 these were *common topic* (e.g. nano spheres), *method* (e.g. dielectrophoresis), *information type* (e.g. articles), and possible *applications* (e.g. intended for biomedical use). The facet *research groups* appeared 16 times and by 3 description questions. The facets *specific reference, source, year, location,* and *disciplinary field* appeared less than 3 times and only by one or two questions.

Task form question c) State of knowledge elicited the largest set of single facet categories (10 categories), followed by b) Work task and d) Ideal answer containing 7 categories. Task form question a) and e) elicited only the 4 main categories. 98% of the e) descriptions included only 2 facets: the common topic and the methodology facet.

All task form questions brought out the facet information type, but as expected description type d) Ideal answer elicited most. Information type was often expressed at two levels, 1) document type: "books", "articles", "reports", and 2) graphic representation: "diagram", "graphs", "codes", and "rates". Sometimes the participants asked for personal sources as "people" and "research groups".

Table 2: Search facets per task form questions (total)

	Task form questions n=325					
Search facet	a)	b)	c)	d)	e)	All
Common topic	316	545	310	234	242	1647
Method	47	73	66	37	48	271
Info type	38	26	29	145	5	243
Application	1	7	1	1	1	11
Other	-	15	11	5	-	31

The information type facet was most frequently used when the task purpose concerned *design methodology*. 17.3% of facets in this category, whilst *theory and background* tasks contained 11.4%, and *previous results* tasks 10.7%. The participants from UNI2 used the information type facets mostly, 17.6% of facets used. UNI3 participants used the facet 10.8%. The facet was least used by UNI1, 4.3%.

For 85% of the tasks there was vocabulary variety between task form descriptions. The participants used a combination of broader and narrower terms to explain topics of interest, e.g. they looked for a "chemical coating (an organic thin film polymer)" (task 25). Synonyms were scarcely used, mostly abbreviations along with their full form, e.g. "N=4 Supersymmetric Yang-Mills Theory (often called SYM" (Task 36). Synonym variations were most frequently used by UNI1 participants.

Table 3: Search facets per university (average)

	Task form question N=325				
Facets	a)	b)	c)	d)	e)
UNI1	7.91	13.66	5.91	7.66	5.25
UNI2	5.53	8.15	5.75	5.87	4.28
UNI3	6.19	11.47	7.71	6.76	4.57
All	6.18	10.24	6.41	6.49	4.55

Task form question b) Work task provided the lengthiest average description at 10.2, whereas question e) Search terms had the lowest average at 4.6. The average length of a) Information need, c) State of knowledge, and d) Ideal answer were almost identical. The length also varied across task purpose, see table 4. In general, the descriptions were higher for *theory and background* tasks (on average 35.2 facets) and lowest for *previous results tasks* (31.8 facets). The e) Search term descriptions were almost 1.0 lower for *design methodology* tasks.

Table 4: Search facets per task purpose (average)

	Task form question n=325					
Task purpose	a)	b)	c)	d)	e)	All
Theory and background	6.7	10.5	6.1	7.1	4.8	35.2
Previous results	5.5	10.1	6.1	5.4	4.7	31.8
Design methodology	5.8	9.6	6.9	6.3	3.8	32.4
All	6.2	10.2	6.4	6.5	4.6	33.8

nDCG scores for task purpose and document types showed that book records performed better for *previous results* and *design methodology* compared to full-text articles and metadata records, whereas metadata records performed better for *theory and background* tasks. Full-text articles and metadata records showed notably lower performance for *design methodology* tasks.

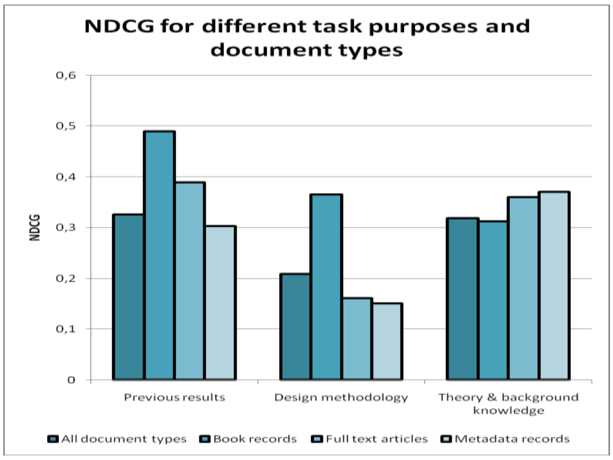

Figure 1: nDCG for task purposes and document types

Document types showed minimal differences on retrieval performance of task length, with book records performing slightly better, specifically for short tasks, se figure 2.

3. DISCUSSION AND CONCLUSION

The task structure was rich with use of a varied set of search facets. The participants described the information tasks insightfully and detailed. They articulated the tasks with use of very specific terminology, and were at the same time capable of varying the vocabulary by broader, narrower and synonym terms in order to explain and clarify the information need. They referred to a large extent to specific research methods, specific information types (different textual and graphical forms), and possible applications for the research. To some degree they described related research groups and locations, related literature and disciplines, and specific sources. They were consistent in varying the task form descriptions according to the questions.

Consequently, the five task form descriptions a) to e) reflect and describe different perspectives on the information task.

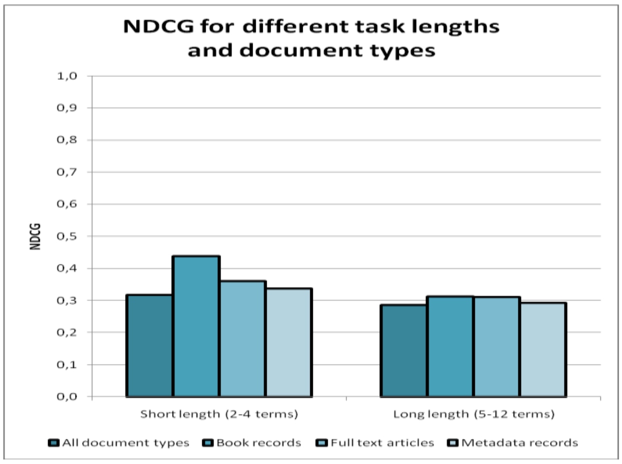

Figure 1: nDCG for task length and document types

The descriptions varied in length, with b) Work task providing highest task length and e) Search terms lowest. The findings differ from [4] where background questions elicited most terms and work task least. However, the task length cannot be compared directly across the two studies, as we coded single search facets (concept level), whereas Kelly & Fu (2006) counted terms (term level).

Retrieval performance based on e) search terms showed better performance for short descriptions. The results contrast findings by [5, 6, 7] that search success depends on searchers' ability to articulate the facets of topics in query terms, structure the facets in the query, and cover them exhaustively. The findings support later results that selection of few, key search facets are more important than exhaustive coverage [8]. Due to a small sample the results represent merely indications, but support the importance of examining effect of task structure and length in IR evaluation.

Variations in length and structure were also present across the three task purposes. On average 35.0 single facets were used to articulate the *background and theory* tasks, and about 32.0 to express *previous results* and *design methodology* tasks. The largest differences appear in a) descriptions between *theory and background* on one side and *previous results/design methodology* on the other, in c) between *design methodology* and *theory and background/previous results,* in d) between all three task types, and in e) between *theory and background/previous results* and *design methodology*.

The retrieval performance supports the observed differences between task purposes. The findings indicate that book records perform better for *previous results* and *design methodology* tasks compared to *theory and background.* Unfortunately, the present data does not explain the observations, and may only be explained due to the smaller collection of book records. Deeper qualitative studies and more retrieval experiments testing the other task descriptions are needed to explain the differences.

Nevertheless, the findings indicate the importance of examining the consequence of variations of length, structure and vocabulary, e.g. whether use of facets influence the search results, whether use of evidence from the different task form questions influence the search results, whether some search facets are better searched by specific description levels and metadata, and whether tasks with specific purposes should be handled differently in the searching.

5. CONCLUSION

The purpose was to extend our understanding of physicists' information tasks in general and more specifically to gain insight into the nature and characteristics of the tasks in order to guide design of IR experiments in the test collection. The analysis showed large variation in structure, length and vocabulary between tasks and between tasks with different purposes.

Future test searches should examine, in general and more specifically in relation to the present test collection how task purpose, choice and number of search facets influence the search result, and whether some facets are better searched by specific description levels and metadata.

5. REFERENCES

[1] Järvelin, K. (2007). An analysis of two approaches in information retrieval: from frameworks to study designs. *JASIST* 58(7), 971-986.

[2] Lykke, M., Larsen, B., Lund, H. & Ingwersen, P. (2010). Developing a test collection for the evaluation of integrated search. In: *Proceedings of 32nd European Conference on IR research, ECIR 2010, Milton Keynes, UK, March 2010*. Springer, Berlin and Heidelberg. 627-630.

[3] Gómez, N.D. (2004). Physicists' information behaviour: a qualitative study of users. In: *70th IFLA Council and General Conference IFLA, Buenos Aires, 22-27 August, 2004.*

[4] Kelly, D., & Fu, X. (2007): Eliciting better information need descriptions from users of information search systems. *Information Processing & Management*, 43(1), 30-46.

[4] Sormunen, E. (2002a): Liberal relevance criteria of TREC – Counting on negligible documents? In: *Proceedings of SIGIR 2002*. ACM Press, New York, 320-330.

[5] Sormunen, E. (2002b) A retrospective evaluation method for exact-match and best-match queries applying an interactive query performance analyser. In: *Advances in Information Retrieval: Proceedings of the 24th European Colloquium on IR Research,* Springer, Berlin and Heidelberg. 334-352.

[6] Vakkari, P., Jones, S. & MacFarlane, A. (2004). Query exhaustivity, relevance feedback and search success in automatic and interactive query expansion. *Journal of Documentation,* 60(2). 109-127.

[7] Kekäläinen, J. & Järvelin, K. (1998). The impact of query structure and query extension on retrieval performance. In: *Proceedings of the SIGIR'98,* ACM, New York (NY). 130-137.

[8] Lykke, M., Price, S. L., Delcambre, L. M. L. & Vedsted, P. (2010). How doctors search: a study of family practitioners' query behaviour and the impact on search results. (In press).