Physicists' Information Tasks: Structure, Length and Retrieval Performance

Marianne Lykke
Royal School of Library and Information Science
Fredrik Bajers Vej 7k, 9220 Aalborg Ost, Denmark
mln@iva.dk

Peter Ingwersen, Toine Bogers, Haakon Lund and Birger Larsen
Royal School of Library and Information Science
Birketinget 6, 2300 Copenhagen S, Denmark
{pi,tb,hl,blar}@iva.dk

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. Results show variety in structure and length across task descriptions and task purposes. The results indicate effect of length and, in particular, of task purpose on retrieval performance of different document description levels that should be examined further.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: *Query formulation*

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 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]. 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. 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 in relation to document types in the integrated test collection?

2. RESEARCH DESIGN
The task descriptions were captured in online forms via computers located in participants' own university environment. The task description form had five questions, in line with the form used by [4]:

IIiX 2010, August 18–21, 2010, New Brunswick, New Jersey, USA.

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.

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. The research team performed test searches manually for each information task as exhaustively as possible, based on the suggested search terms and other terms in the original task descriptions, primarily from task description question e). Two months after task creation, access to a web-based relevance assessment system was opened for the participants. 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 type 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. Table 1 provides an overview of the study variables applied to answering the research questions.

We based categorization of main purpose on central elements of scientific work that require information retrieval, e.g. theoretical background, research methodology, previous results [6]. The facet analysis was conducted using Sormunen's (2002) definition that a search facet is a concept (or a family of concepts) identified from and defining an exclusive aspect of a search topic [7, 8]. In order to get an understanding of the vocabulary used to express the search facets, we classified the terms into five semantic categories, adopted from [9]. A synonym (ST) is a term that is interchangeable with equivalent task description term. A broader term (BT) refers to a term, which is broader in hierarchy. A narrower term (NT) refers to a term narrower in hierarchy. A related term (RT) is a term, which is associated and represents a related perspective. The information tasks have been divided into 1) verificative needs, 2) conscious, topical needs, and 3) muddled needs to determine level of complexity [10]. 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. With the aim to gain insight whether the different document types in the integrated test collection affect retrieval performance, we compared nDCG scores for task purposes and task length.

3. RESULTS

The 65 information tasks originated from 23 physicists from three different universities; 12 from Copenhagen University (UNI1), 32 from Technological University of Denmark (UNI2), and 21 from Aalborg University (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 looked for *Theoretical background*, e.g. "Descriptions of models and theory concerning passive mode-locking in linear cavities" (Task17), 26% looked for *Previous results*, 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% looked for *Research methodology*, e.g. "Tables, graphs and figures with comparisons of different energy harvesting techniques" (Task5).

The structure of task descriptions varied regarding number and types of search facets. Only 4 types of search facets appeared in all descriptions, and were elicited by all five questions. As shown in Table 2 these were *common topic* (e.g. nano spheres), *method* (e.g. dielectrophoresis), *information type* (e.g. articles), and

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
Task structure	Types of search facets used to describe the information task, e.g. common topic, method used, time, type of information. One count per single facet
Task length	Number of single search facets per task form question
Search vocabulary	Semantic categories used to express search facets, e.g. synonym terms, hierarchical broader terms or hierarchical narrower terms
Educational level	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 primarily from question e)

possible *applications* (e.g. intended for biomedical use). The facet type *research groups* appeared 16 times and by three questions. The facet types *specific reference, source, year, location,* and *disciplinary field* appeared less than 3 times and only by one or two questions. Question c) Background knowledge elicited the largest set of single search facets (10 different types), followed by questions b) Work task and d) Ideal answer each bringing out 7 facet types. Questions a) and e) elicited only the 4 different types. 98% of the e) descriptions included only 2 search facets: *common topic* and *method*.

Table 2: Types of search facets per task description questions (total number)

	Task description 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

Table 3: Number of search facets per task purpose (average)

	Task description question (n=325)					
Task purpose	a)	b)	c)	d)	e)	All
Theoretical 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
Research 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

All five questions brought out the search facet type *information type*, but as expected question d) Ideal answer elicited the most. Search facet *Information type* was often expressed from two perspectives: 1) document type, e.g. "books", "articles", "reports"), and 2) graphical representation, e.g. "diagram", "graphs", "codes", and "rates"). Sometimes the participants asked for personal information sources such as "people" and "research groups".

The information type facet was most frequently used when the task purpose concerned the task purpose *Research methodology,* 17.3% of facets in this category, whilst *Theoretical background* tasks contained 11.4%, and *Previous results* tasks 10.7%.

For 85% of the tasks there was variety regarding semantic categories across task descriptions a) to e). The participants used a combination of broader and narrower terms when expressing 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).

Question b) Work task situation provided the lengthiest average description at 10.2, whereas question e) Search terms had the lowest average at 4.6. The average length of questions a) Information need, c) Background knowledge, and d) Ideal answer were almost identical. The length also varied across task purpose, see Table 3. In general, the descriptions were lengthier for *Theoretical background* tasks (on average 35.2 facets) and shortest for *Previous results* tasks (31.8 facets). The e) Search term descriptions were almost 1.0 shorter for *Research methodology* tasks.

As shown in Figure 1, nDCG scores for task purpose and document types showed that book records performed better for *Previous results* and *Research methodology* compared to full-text articles and metadata records, whereas metadata records performed better for *Theoretical background* tasks. Full-text articles and metadata records showed notably lower performance for *Research methodology* tasks.

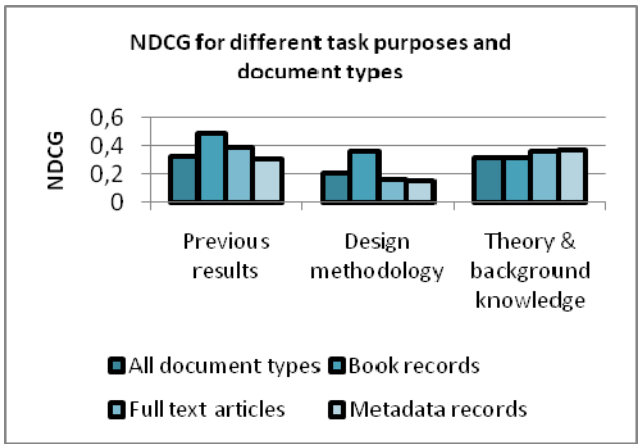


Figure 1: nDCG for task purposes and document types

Document types showed minimal differences on retrieval performance of task length, with book records performing slightly better, specifically for short tasks, see Figure 2. In general, short tasks of 2-4 facets performed better compared to longer tasks of 5-12 search facets.

4. DISCUSSION AND CONCLUSION

The participants described the information tasks insightfully and detailed. The structure of task descriptions was rich. The participants used a large variety of facet types to express the search topic. 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 descriptions according to the five questions in the task description form. Consequently, the descriptions deriving from the five task description questions a) to e) reflect and describe different perspectives on the information task. This result is in line with [4] that users are able to articulate different aspects of their information needs.

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

The participants articulated the tasks with use of combinations of narrower and broader terms, presumably varying the vocabulary in order to explain and clarify the information need. Synonyms were rarely used, mostly abbreviations along with their full form, presumably to explain and clarify the abbreviation.

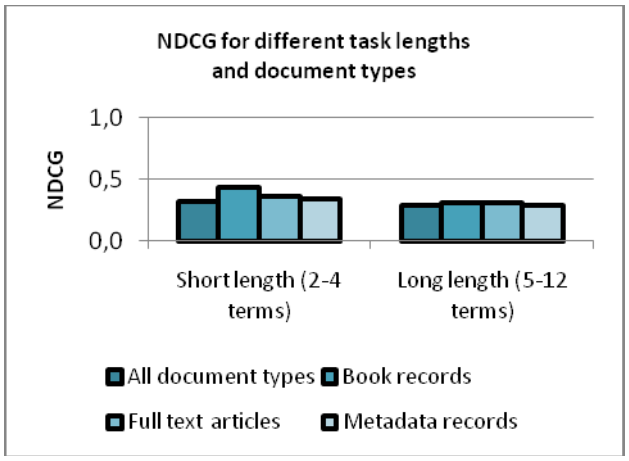


Figure 2: nDCG for task length and document types

Retrieval performance primarily based on question e) Search terms showed better performance for short descriptions. The results contrast findings by [7, 8, 11] that search success depends on searchers' ability to articulate and structure the search facets, and cover them exhaustively. The findings support later results that selection of few, key search facets are more important compared to lengthy, exhaustive coverage [12]. The results might be explained by the fact that participants in the present and latter study were subject experts with strong and work-related interest in the search topic as opposed to previous studies where study subjects were students with lesser connection to search topics. Due to a small sample the results represent merely indications, but support the importance of examining the effect of number of search facets present in the search queries in IR evaluation experiments, including the effect of subject expertise and interest.

Variations in task length also appeared across the three task purposes. On average 35.0 single search facets were used to articulate the *Theoretical background* tasks, and about 32.0 to express *Previous results* and *Research methodology* tasks. The largest differences appeared in question a) descriptions between *Theoretical background* on one side and *Previous results* and *Research methodology* on the other, in question c) between *Research methodology* and *Theoretical background* and *Previous results*, in question d) between all three task types, and in question e) between *Theoretical background* and *Previous results* and *Research methodology*.

The observed differences between task purposes were also reflected in the results comparing retrieval performance of task purpose in relation to document types. The findings indicate that book records perform better for *Previous results* and *Research methodology* tasks compared to *Theoretical 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 of the present data, including more experiments testing retrieval performance based on terms from the other task description questions are needed to explain the differences. Nevertheless, the findings indicate the importance of examining the consequence of variations in length and structure, e.g. whether number and types of search facets influence the search results. The results also indicate the merit of investigating whether tasks with specific purposes should be handled differently in the searching. Furthermore, the fact that the participants chose to express search facets using combinations of narrower and broader and abbreviations with its full form indicate that it might be worthy to examine whether uses of certain semantic categories influence the search results.

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 descriptions deriving from different questions and between tasks with different purposes, including differences in retrieval performance in relation to different document types in the integrated test collection. Future test searches should examine in detail how type and number of search facets influence the search result, and whether some facets and some task types are better searched by specific description levels and metadata.

6. 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: *Advances in Information Retrieval. Proceedings of 32nd European Conference on IR research, ECIR 2010, Milton Keynes, UK, March 28-3*. Springer, Berlin, Germany. 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.

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

[6] Rienecker, L & Stray Jørgensen, P (2005)

[7] 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.

[8] 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.

[9] Wang, P (1997). User's information needs at different stages of a research project : a cognitive view. In: P Vakkari, R Savolainen & B Dervin, *Information seeking in context*. London : Graham Taylor. 307–318.

[10] Ingwersen, P (1992). *Information Retrieval Interaction.* London: Taylor Graham. 246 p.

[11] 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.

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

Physicists' Information Tasks: Structure, Length and Retrieval Performance

Marianne Lykke, Peter Ingwersen, Toine Bogers, Birger Larsen & Haakon Lund
Royal School of Library and Information Science, Denmark

SUMMARY

Aims and challenge

Insight into the nature and characteristics of the tasks in order to guide design of information retrieval experiments

Research questions

- What is overall purpose of tasks?
- How is the structure of tasks concerning types of search facets?
- How is the length of tasks concerning number of search facets?
- How is the retrieval performance of tasks?

Setting

- Test collection for the evaluation of integrated search
- Domain-specific test collection in the physics domain
- System-driven as well as user-oriented evaluation
- Collection of approx. 18.000 monographic records, 160.000 papers and articles in full-text, and 270.000 abstracts
- Varied set of metadata and vocabularies
- Real information tasks, based on genuine work task situations
- Non-binary relevance assessment

CONTACT

Marianne Lykke
Royal school of LIS, Denmark
mlr@iva.dk

RESEARCH DESIGN

The tasks were captured in online forms via computers located in participants' own university environment.

In line with Kelly & Lu (2007) the participants described the tasks from 5 different perspectives:

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

The research team performed test searches manually for each task as exhaustively as possible. The queries were based on the suggested search terms and other terms in the task descriptions. Search performance was measured by nDCG, based on task creators' 4-point scale relevance judgments: highly, fairly, marginally, and non-relevant (Sormunen, 2002).

Pre-questionnaire for demographic data and post-assessment questionnaire on satisfaction with assessment procedure and research results.

Task characteristics were analysed coding central aspects of the five task descriptions:

Study variable	Definition
Task purpose	Mail goal of information task. Exclusive coding
Task structure	Types of search facets. One count per single facet
Task length	Number of single search facets
Vocabulary	Semantic category used to express facets synonym, broader or narrower terms
Info need type	Known item, known topic or muddled topic
Retrieval performance	Discounted Cummulative Gain (DCG) per task

RESULTS

68 natural information tasks were captured from 23 senior researchers, PhDs, and experienced MSc students from three different university departments of physics.

100% topical, conscious information needs.

Task purpose	%
Theoretical background	54
Previous results	26
Research methodology	20

Varied task structure

Search facets	Task description questions					
	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	n/a	15	11	5	n/a	31

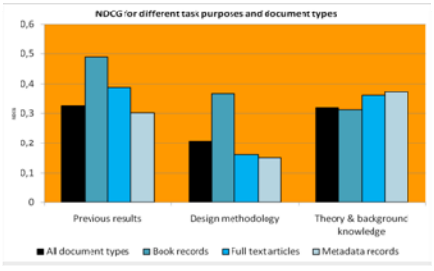
Vocabulary
Topics expressed with combination of hierarchical broader and narrower terms, e.g. chemical coating (an organic thin film polymer).

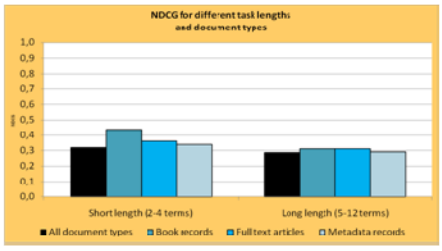
Few synonyms, most abbreviation with full term, e.g. "N=4 Supersymmetric Yang-Mills Theory (often called SYM)"

Varied task length per task purpose

	Task description questions (average)					
Task purpose	a)	b)	c)	d)	e)	All
Theoretical 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
Research 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

Retrieval performance for task purposes and task length





Take home message – Information tasks

- Known topic information needs
- Task purpose: theory, previous research, methodology
- Structure: many different search facets
- Varied length: question c) shortest with fewest facets
- Vocabulary – combinations of broader and narrower terms

Take home message – IR experiments

- Some facets perform better than others?
- Short length perform better that longer?
- Some semantic categories provide different search results?
- Some document types perform better for certain task purposes in integrated search?