Journal of Documentation

Queries in authentic work tasks: the effects of task type and complexity

Miamaria Saastamoinen Kalervo Järvelin

JDOC
72,6

Received 30 September 2015
Revised 5 June 2016
Accepted 5 June 2016

Queries in authentic work tasks: the effects of task type and complexity

Miamaria Saastamoinen and Kalervo Järvelin
School of Information Sciences, University of Tampere, Tampere, Finland

Abstract

Purpose – The purpose of this paper is to investigate information retrieval (IR) in the context of authentic work tasks (WTs), as compared to traditional experimental IR study designs.

Design/methodology/approach – The participants were 22 professionals working in municipal administration, university research and education, and commercial companies. The data comprise 286 WTs and 420 search tasks (STs). The data were collected in natural situations. It includes transaction logs, video recordings, interviews, observation, and daily questionnaires.

Findings – The analysis included the effects of WT type and complexity on the number of STs, queries, search keys and types of queries. The findings suggest that simple STs are enough to support most WTs. Complex WTs (vs more simple ones) and intellectual WTs (vs communication, support and editing WTs) include more STs than other WT categories.

Research limitations/implications – Further research should address the problems related to controllability of field studies and enhance the use of realistic WT situations in test-based studies, as well.

Originality/value – The study is an attempt to bring traditional IR studies and realistic research settings closer to each other. Using authentic WTs when studying IR is still rare. The representativeness of the WT/ST types used in interactive IR experiments should be carefully addressed: in the work flow, people seldom consciously recognise separate "STs". This means that STs may mainly be an academic construct even to the point that studying IR without a decent context does violence to the further understanding of the phenomenon.

Keywords Information retrieval, Searching, Field studies, Queries, Work task complexity, Work tasks

Paper type Research paper

Introduction

Many jobs and tasks have information use and knowledge creation at their core. Knowledge work has a number of descriptive features; it is hard to delimit to comprise only some professions. (Pyöriä, 2005). The tasks of teachers, researchers, human resource experts and financial managers differ in their contents but they also share several features such as the varying degree of complexity, urgency and salience, and varying types of the final output or outcome (see e.g. Li and Belkin, 2008). Knowledge work is formed of work tasks (WTs) that are building blocks of one's job (Byström and Hansen, 2005). In the present study, WTs are understood as concrete sections of time that include actions to proceed towards a goal, the task outcome.

Information retrieval (IR) as a practical activity supports task performance in authentic environments. However, much of IR systems development takes place through laboratory experiments using test collections. While traditional test-collection based IR evaluation studies are relatively easy to execute and serve the development of IR methods well, they are not quite sufficient to learn about IR effectiveness in authentic environments. Without studies of authentic use situations, studies based on

The authors wish to thank Heljä Franssila (University of Tampere) for her cooperation in the data collection phase.

Journal of Documentation
Vol. 72 No. 6, 2016
pp. 1052-1077

0022-0418
DOI 10.1108/JD-09-2015-0119

simulated WTs or narrower search tasks (STs) conducted in a laboratory remain only as hypothetical approximations of the situations. Effective IR systems, or other information systems, should not be designed solely on such a basis (Vakkari, 2003). IR systems and methods are studied and developed continuously (see e.g. Ruotsalo *et al.*, 2013) but as a real-life phenomenon, bound to its context, IR has not been of much research interest (Järvelin *et al.*, 2015; Vakkari, 2003). Traditional server-side log analyses are perhaps adequate when developing a single search system; but they are not enough when developing the information environment or the ways of acting in it.

The purpose of the present study is to analyse IR conducted in real life in connection to WTs. WTs are categorised by their complexity and type and they act as the context in which IR is performed. The dependent variables representing IR are the number of STs, the number of queries, the number of search keys and the query types. The study is explorative and quantitative. The main data consist of client-side transaction logs and screen videos that are supported by electronic questionnaires, interviews and observation. This study is one of the few studies so far that aim at capturing authentic IR in context of authentic WTs with objective real time data collection methods. These kinds of investigations are important in order to understand what IR is used for, how it is performed and what factors affect it. Otherwise the refinement of search engines and designing of user tests are based on mere assumptions that may or may not add to the actual usefulness of information systems. The present paper focuses on the query aspects of the rich data set. The results concerning information needs, search media and information search processes are reported elsewhere (Saastamoinen and Järvelin, in press) for conciseness and readability of the paper.

Related research

In this section, earlier research on IR, STs and WTs are discussed. This includes empirical, theoretical and methodological points. WTs and STs are the two most important concepts of the present study, since it concerns the effects of WTs on IR which is organised into STs. IR studies on the one hand and studies on WTs on the other are many. This selective literature review is meant to (a) give an overview of tasks as theoretical constructs in IR and information seeking research; (b) discuss the concept of task complexity within these fields; (c) present earlier empirical findings and methods that connect task features to IR and aim at realism and validity in the study design. Especially the point (c) narrows the empirical focus of the section and excludes a vast body of information seeking studies, IR laboratory studies and survey studies as non-relevant to the approach of the present paper. Vakkari's (1999) review is recommended to the readers interested in older studies about the subject.

The underlying tasks are important starting points of IR research as stated by, for example, Kuhlthau (2005) and Vakkari (1999). Byström and Hansen (2005) name two conceptions of WTs as study objects. WTs can be abstractions as defined in task assignments, or concrete steps taken in order to proceed (Byström and Hansen, 2005). The latter view is applied in the present study. WTs contain subtasks (Vakkari, 1999). One subtask type is information STs. Based on earlier research, it seems that WT is becoming an increasingly acknowledged context of studies in IR. However, the term "WT" is used synonymously to STs in some studies.

IR and tasks are highly interconnected as shown by several studies. Järvelin (Ingwersen and Järvelin, 2005; Järvelin, 1986) presented a model for information access where task goals, task processes, the needed and used information, information searching, as well as information systems are bound in a complex interaction and mutual adaptation. The results of the empirical study by Pharo (2002) interestingly suggest that IR may even affect larger task goals, not only the other way round. Pharo (2002, 2004) also states that WT goals are directly connected to the criteria of relevance assessment. Similarly, Vakkari (2000, 2001) has shown that task performance stages are connected to the type of needed information, search terms, tactics and relevance criteria.

Since the seminal paper by Byström and Järvelin (1995), task complexity is a widely recognised task feature that can affect task performance and information seeking. Task complexity is related to, for example, task difficulty, task extent and the performer's level of knowledge. The concept is, however, used differently in different studies. The operationalisations vary, and sometimes task complexity is understood as ST complexity, sometimes as WT complexity (Liu and Li, 2012; Wildemuth *et al.*, 2014). Authentic WT complexity's effects on IR, such as the number of search terms and types of queries, have not been systematically analysed in past studies. The present paper analyses only WT complexity, but research discussing ST complexity is also a noteworthy reference point. Thus both are discussed below.

A dichotomy is often made between task complexity's objectivity and subjectivity (perceived complexity). In reality, these two are connected. Liu *et al.* (2015) controlled objective ST complexity (operationalised as the number of activities and information sources needed to succeed in the task) in an experiment and found that these complexity features are connected to the post-search difficulty estimated by the participants. Hansen (2011) found that some WT types in a patent office are often perceived to be difficult. Tasks that included handling patent applications in their early form were considered mostly difficult perhaps because of the low amount of case-related information available (Hansen, 2011).

Bell and Ruthven (2004) tested the relations between objective and perceived (subjective) ST complexity. They found that perceived task complexity is linked to how much and what types of information are required in the task assignment. Perceived task complexity was also affected by how clearly information requirements were stated in the assignment and how well they were understood by the task performers (Bell and Ruthven, 2004). However, in real-life situations, the issues concerning the task performer's ability to understand ST assignment may be of minor importance, depending on the profession. STs following information needs are often internally generated during the flow of a WT rather than ready-made for the task performer. Furthermore, even if the ST was externally generated, the task assignment does not have to be identical to the information needs it generates. The task performer may have the option to use several resources to complete the task and even the option to ask for further instructions from the initial ST generator, which may not be possible in a test situation.

Expertise or the amount of knowledge has connections to task complexity and task performance in interesting ways. Vakkari (2001) proposed a theory of task-based IR which is based on his longitudinal study on students' IR processes and Kuhlthau's (1993) information seeking model. The theory shows that task stage is connected to, for example, the number and type of search terms used. (Vakkari, 2001). Though Vakkari's (2001) theory explicitly discusses the stage of the task, that is the amount of knowledge the participants have, it is parallel to perceived complexity of perceived WTs: a WT, understood as a process within a longitudinal work duty, always has a stage dimension and prior knowledge which are reflected in the task performer's estimate of perceived complexity.

Haerem and Rau (2007) analysed the connections between objective task complexity and expertise. They defined objective task complexity based on whether a programme coding problem can be most efficiently solved by modification of the surface level (inputs and outputs) or by modifying the underlying process of the task (or both). They found task complexity an important factor in task performance. In their study, participants had different levels of expertise suitable for assigned programming tasks at hand. It was found that experts and novices may literally see tasks differently depending on task complexity. Novices even outperformed experts in objectively simple tasks because their superficial knowledge was suitable and enough for successfully performing the task (Haerem and Rau, 2007). Similarly, the findings of Vakkari *et al.* (2003) suggest that domain knowledge affects information searching if the participants have enough system knowledge as well. Hansen (2011) found in a patent information interaction study that low domain knowledge leads to the use of fewer search terms. Saastamoinen *et al.* (2012) findings suggest that with increasing subjective WT complexity, the number of queries per WT increases. Task complexity predicted some information searching phenomena better than others, and actually perceived complexity seemed to affect WT performance more clearly than a priori knowledge of the task features. However, a priori knowledge was also measured subjectively (Saastamoinen *et al.*, 2012).

Marchionini (2006) differentiates between exploratory STs that are connected to the aims of learning and discovering, and lookup searches that are, for example, factual searches or searching for known items or other well-structured or well-known objects. Exploratory STs can actually be considered complex STs (Wildemuth and Freund, 2012). Wildemuth and Freund (2009) call for a definition for exploratory searching. They state that explorativeness is formed of several factors, such as generality, need of several documents and some sort of overall obscurity (Wildemuth and Freund, 2012).

Ruotsalo *et al.* (2013) studied university students and staff performing assigned exploratory STs whose idea was to find relevant scientific papers on given subjects. Their results indicate that, compared to only typing queries, participants gained a lot of benefit from using a system, which visualised the search and enabled direct control over the importance of each key word (Ruotsalo *et al.*, 2013). Unfortunately, the study concerned only exploratory STs. It would have been interesting to see whether a highly interactive search system supports different types of tasks in distinct ways; that is, for example, when searching for facts or known items. Though lookup searches are often important in daily searching, IR systems should support more explorative needs as well (Marchionini, 2006). Lookup and exploratory STs are compared, for example, by White and Marchionini (2007). They found that explorative searching includes using more unique query terms per task (White and Marchionini, 2007).

Varying empirical approaches to authentic IR

Collecting anonymous search logs from search engines is an efficient way to get a large data set. Though quantitatively pre-eminent, these data do not include any explicit information about the information needs or tasks of the person behind queries. A good review of past log studies and a presentation of ways to exploit logs is found, for example, in Silvestri (2010) and Jansen (2009).

A typical way to study interaction between users and IR systems (called interactive information retrieval (IIR)) is to conduct a user study in controlled settings (see e.g. Kelly, 2009), often using simulated WTs that were proposed by Borlund and Ingwersen (1997; Borlund, 2003). Test settings enable systematic control over some features (i.e. independent variables) while studying their effects on participants' task performance. For example, Ruotsalo *et al.* (2013) conducted a rather typical user test.

The STs or simulated WTs tend to be used in studies without actually knowing their validity. Li and Hu (2013) tested on a small scale whether performing a simulated WT differs from performing an authentic task. The authors concluded that simulated WTs and authentic tasks did not differ essentially in task performance or participants' perceptions of the tasks, but that this holds only if simulated WTs are carefully designed regarding task complexity and relevance (motivation) to the participants (Li and Hu, 2013). This is also the original idea behind simulated WTs by Borlund (2003). Though Li and Hu's (2013) study was a fair attempt to empirically address question, it had some validity issues: the participants were undergraduate students, they were allowed to use only one information system, and the simulated WT seemed quite difficult in the short 15 minutes they had to accomplish it.

Vakkari's (2001) longitudinal study combined features of a controlled user study and naturalistic approach. The participants were students writing an authentic research proposal. During a four-month period, they had three sessions where they searched for relevant documents. The participants were interviewed, they were asked to think aloud while searching, and transaction logs were collected. The participants also assessed the relevance of the documents found. Between these search sessions, they were keeping diaries about their progress in the task (Vakkari, 2001). Similarly, Pharo (2002) combined controllability and authentic study assignments in his study: the participants contacted the researcher when they wanted to search the web for their theses. The participants were interviewed for their tasks, and their search sessions were recorded on video and observed. Combining controllability and authentic tasks is a potentially fertile and valuable approach. However, these studies only included a small number of students as participants. The participants had only one, quite large and demanding task to perform in somewhat artificial situations. Despite the inherent complexity of the task in question, IR was logged from the perspective of using a single information source; a bibliographic database or the web.

The field has a limited number of studies on authentic larger tasks' effects on IR (see e.g. Kellar *et al.*, 2006; Kelly, 2006; Kumpulainen, 2014). Or, if authentic work is studied, the data are often self-reports by the participants, that is interviews (Li, 2009), diaries (Du, 2012, 2014) or questionnaires that often involve *ex post facto* rationalisations and other limitations on the data. Traditionally, studies of information seeking or other information practices (e.g. Lingel, 2015; Robinson and Yerbury, 2015) do not discuss IR in detail, such as queries. Nicholas *et al.* (2009, 2010) combined questionnaires, interviews and observation to log analysis when studying information seeking and IR of researchers. A problem was, however, that the logs could not be linked to the participants of the survey; the people searching and telling about their searching were similar but likely not the same (Nicholas *et al.*, 2010).

Hansen (2011) conducted a thorough study of patent engineers' information seeking and IR behaviour. His data contained interviews, observational data, electronic diaries and interaction logs. Saastamoinen *et al.* (2012) studied municipal workers' information searching. The data were collected observing authentic WTs with varying perceived complexity (Saastamoinen *et al.*, 2012). Xie (2000) conducted a field study of library users. She used interviews, logs and observation to find out how the participants shift in their strategies and information goals while seeking for information. However, the study only analysed the shifts, not the factors affecting them (such as task complexity) (Xie, 2000).

In summary: STs in real life are part of larger tasks and thus more evolving and endogenous than assigned test tasks. It is possible that short, clearly targeted IR sessions in a single system are somewhat similar in real life and in simulated situations.

This would encourage evaluation studies using assigned STs. However, such tests can be far away from authentic WTs where several systems can be used side by side, deadlines can vary and people actually use information and documents after finding them – i.e. when IR is not an end in itself (Järvelin *et al.*, 2015). Naturalistic studies are resource-intensive because they mean letting participants do their own work, their own tasks, with their own information resources and collecting simultaneously data about their tasks, information resources and IR. This may be considered prohibitive. Naturalistic studies are nevertheless necessary: they provide invaluable knowledge about the phenomena studied. This knowledge can be used to design more effective and realistic laboratory and user studies.

Research design

This is a descriptive and explorative study presenting empirical findings concerning task-based IR. Since there was no ready-made methodological toolkit to be used in the study, the study also contributes to the methodology within the field. The purpose of the paper is to improve the research community's understanding of authentic IR, especially queries, in task context. This is important because (a) the understanding can increase realism in user tests; (b) querying is related to the effectiveness of IR; and (c) better knowledge facilitates improving effectiveness of information systems, their user interfaces and ways of use. The present paper discusses the query features and the number of STs in WTs.

Research questions

The research questions are as follows:

RQ1. How is WT complexity connected to IR features?

RQ2. How are WT types connected to IR features?

RQ3. How is WT complexity connected to IR features within each WT type?

The IR features discussed in the present paper are: (a) the number of STs in WTs; (b) the number of queries in STs; (c) the number of search keys of queries in STs; and (d) the types of queries in STs. WT complexity means here perceived complexity, an aggregated variable formed of subjective estimates given by the participants. WT types form an ad hoc categorisation based on the present data set. The categories are communication, support, editing and intellectual tasks. The IR features, WT complexity and WT types are defined below.

Data and data collection

Next, the data and the data collection process are presented. The exact figures of data are presented in Table I. A total of 22 subjects working in six different organisations participated in the study. Note that these six organisations of three organisation types

Table I. The profile of the data set

Organisation type	Organisations	Participants	Days	Work tasks	Search tasks
Municipal administration	1	10	10	47	55
University	3	5	30	101	173
Company	2	7	37	138	192
Total	6	22	77	286	420

were a convenience sample. The participants were 17 females and five males between their mid-twenties and mid-fifties three of whom were in managerial positions. The participants' WTs and information sources are briefly discussed below in order to give the reader an understanding of the organisations in question. However, please note that the differences between these organisations are not analysed further but their data are merged.

In municipal administration, the participants worked in the sectors of communication and human resources. They used several internal databases to find, e.g., people and decisions. In addition, they used the organisation's intranet in order to stay up-to-date. Several of these participants had WTs that involved updating the organisation's intranet or external website.

In the universities studied, the participants had research and teaching related WTs as well as small administrative tasks, such as filling in working hours in an electronic system. The systems they used included library catalogs, electronic libraries, dictionaries and powerful scientific calculation programmes. Teaching included planning courses and lectures as well as assessing study assignments on electronic platforms.

The participants in the companies had varying roles and tasks. They worked for example in design and supply of consumer goods and services, marketing, financial administration and data administration. These tasks typically included making decisions as well as typical administrative tasks. The key information was found in several internal databases but extra-organisational information was also searched in the web.

The majority of participants, 12, were asked to participate for five working days, although more was allowed and also welcomed. The municipal administrative organisation, with ten participants, had only one working day in the data per person. This was due to practical issues discussed and agreed with the organisation.

The first author of the paper performed all data collection. Several data collection methods were used (see Figure 1). This enabled triangulation of the data. An initial interview was held in order to get general information about the participants' working role, their tasks, the information systems they used and their overall information searching behaviour. After the data collection of each participant, the researcher made preliminary analysis on the data. After that, an exit interview was held with those participants, who had data for several days. The purpose of the exit interview was to clarify possible unclear parts in the data to the researcher. All interviews were held individually with each participant.

During all actual data collection days, the data were collected through transaction logging, screen capture video and questionnaires. In the morning of each day, the participants were requested to list in an electronic questionnaire the tasks they thought they are going to perform that day. They also estimated the complexity of the tasks in the questionnaire. At the end of the day they described their tasks similarly and stated how complex the tasks felt afterwards (see Appendix for the questionnaires).

A screen video recorder and a logging programme were installed on the primary computer of each participant. The logging software started recording automatically but the participant had to start the screen capture and switch it off. Logging could also be

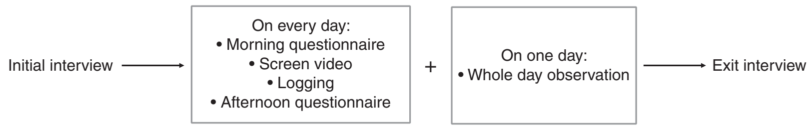


Figure 1.
The data collection methods

stopped and re-launched. The participants were advised to inform the researcher if there was something they did not want to be recorded. If they afterwards wanted something to be removed from the video or the logs, the passage in question was either cut or blurred so that all texts became unreadable. In addition, one data collection day included observation conducted by the researcher. All participants were offered a small-scale feedback describing their personal information searching behaviour or the overall results concerning the organisation in question; and the ones that had participated for several days also got movie tickets as an appreciation of their efforts.

Analysis

The WTs that the participants named in the daily questionnaires were accepted for further analysis if these tasks were recognisable in the log or in the videos or both. This means that tasks actually not performed, while planned to be performed in the morning, were naturally excluded as well as tasks where the computer was not used at all (some meetings, for instance). After data categorisation, the variables were analysed quantitatively. In the Findings section, the mean, median, Spearman's correlation marked with ϱ (rho), its statistical significance marked with p, and distributions are discussed. The generally accepted limit $p < 0.05$ is used for statistical significance. Statistical significance guides but does not strictly limit the interpretation of correlations.

WT types and complexity. Each WT was assigned a complexity estimate and task type. These are defined below. WT complexity refers to perceived complexity, that is, the subjective view of the task performer. Task complexity is simply the average of three percentage estimates (0-100 per cent) given by the task performer. One is the task complexity estimate given before the task, the second is the estimate given after the task, and the third is the complement percentage of participant's knowledge of the task process asked in the morning questionnaire. The last one means that the more the participant knows, the simpler the task; that is if she knows, e.g., 80 per cent of the task process, the complexity estimate based on this figure would simply be 20 per cent. If any of the three estimates was missing, the average was calculated of the estimates that were available. In total, 22 per cent of all potential estimates were missing, for instance due to participant's carelessness or an unforeseen task being performed.

The operationalisation of task complexity is adapted from earlier research (Saastamoinen *et al.*, 2012; Byström, 1999; Kumpulainen, 2014). Three task complexity estimates (pre-task, post-task and knowledge of task process) are chosen instead of one in order to increase the validity of the task complexity estimate. There is no reason to predict that one of these estimates would alone be more valid than others; they represent different aspects of WT complexity.

Continuous task complexity was used as the base for categorising complexity. WTs were categorised in two different ways into complexity classes. In one, the complexity classes are in 10 per cent intervals (i.e. tasks from 0 to 9.99 per cent complexity and so on). This is the one used when calculating correlations. In the other categorisation, the WT complexity is divided into four classes so that each includes about one quarter of the WTs; I (tasks from 0 per cent complexity to 21.7 per cent), II (21.8-38.3 per cent), III (38.4-50 per cent) and IV (50.1-100 per cent). Four classes are used for calculating the mean and the median. The categorisations are here called equidistant and equal-number categorisations, respectively.

The ways of categorising task complexity are based on extensive initial analysis of the data. Categorising the complexity was a necessity in order to calculate descriptive figures.

Using categories of approximately same sizes (equal-number categorisation) is an established statistical solution to get reliable analytical results. The previous experiences (Saastamoinen *et al.*, 2012) dividing tasks to only three categories showed the problematic interpretation of the middle category that was neither complex nor simple. Four categories enable summing up the data into complex and simple half when necessary. Using five categories was also tested but it was not found more useful than a four-category solution. The connections between WT complexity and IR features in this real-life data are seldom clearly linear; using too few categories leads to no variation between categories, and using too many leads to difficulties in finding the overall trend between several figures.

The equidistant categorisation maintains enough detail of the original complexity estimates but tidies up a little the variations found, and makes the distribution of WT complexity near normal. Equidistant complexity represents the absolute perceived complexity of the tasks, whereas the equal-number categorisation proportions the complexity of each task to other tasks (simple, semi-simple, semi-complex and complex) exploiting information about data distribution. It was necessary to use these ways of categorising task complexity because they allow complementary analyses. The categorisations were chosen following basic principles of quantitative analysis. Also a qualitative counterpart to task complexity was used, task types, which are explained next.

The researchers formed a WT type classification to all WTs in the data. The task type classification is based on written descriptions of tasks given by the participants. It is another task dimension that complements task complexity so that the possible effects of task complexity on IR could also be compared to another meaningful independent variable. The task type categorisation is data-driven while abstract and therefore usable across the tasks of the organisations studied. Four types of tasks were found in the data set: communication, support, editing and intellectual tasks. Note that since these task types were data-driven, the labels and descriptions were formed *post hoc*.

Communication tasks require distributing and/or receiving information. These tasks typically include handling e-mail or discussing with colleagues in a meeting. Real time teaching is also considered as communication task.

Support tasks are tasks that support the main tasks of the organisation. Typically, the process of these tasks is well-defined: they do not include a creative process *per se*. Support tasks are such as accounting and filling in structured forms.

When performing intellectual tasks, the participants create something new, either in the sense of a concrete output or creating knowledge. Also solving significant problems is counted into intellectual WTs.

Editing tasks are similar to intellectual tasks, but they often heavily build on some previous work by the task performer herself or others thus being "semi-intellectual". They can also be described as parts of doing some bigger tasks, such as starting or finishing something. Editing tasks have clear hints of incompleteness, partial work or the editing phase in the task description so they were quite easily recognisable, such as going over some new instructions and commenting on them, or to start writing a memorandum.

In support tasks, information is often handled as an object in the sense that it does not and is not intended to turn into "knowledge" but rather moved, applied in a straightforward manner, or mechanically revised. These tasks are simple in interacting with information content. In editing and intellectual tasks, people generate new information and knowledge. The output may be formal, such as an essay, or the information gained is used in order to create new knowledge to support decision making or to solve a problem the task performer has. The difference between editing and intellectual tasks is that intellectual tasks require more intellectual input as something is done from scratch; editing tasks are "partial", for example, the task performers build on an existing base.

Support, editing and intellectual tasks can be seen as forming one interpretation of task complexity. Communication tasks cannot be inherently included in the continuum. However, task types and task complexity are, by definition, independent of each other in the present study. Task types are based on labels given by the participants and categorised by the researcher, whereas task complexity is an average of numeral task complexity estimates given by the participants.

Two researchers from the same research centre reclassified a random subset of tasks. They had the same brief descriptions of the task types as presented above. The first one knew the study but had not handled the data. He classified 82 per cent of tasks to the same categories as the original classifier. The other reclassifier did not know anything about the study beforehand. The similarity of his and the original classifier's decisions was 74 per cent. The intra-classifier reliability was 88 per cent. Cohen's Kappa for inter-classifier reliability was 0.69, for intra-classifier 0.83.

Information STs. An ST is a temporally continuous (i.e. two STs cannot overlap) subtask that includes a query or queries and has a somewhat uniform motivation ("an information need") to search for information. Thus, STs are not self-standing search assignments as in IIR experiments but materialised processes judged afterwards. This is one reason for the decision that STs cannot overlap; in the definition applied here, there do not exist self-standing, abstract information needs that could be returned to. Instead, information needs are formed seamlessly in the work flow as a motivation to conduct the immediate concrete search actions. Although STs cannot overlap each other, a random interruption, such as an incoming call, does not necessarily break the ST. The latter happens only if this call, e.g., results in a recognisable switch of WTs in the logs. In practice, finding the technical boundaries between STs was not an issue. Most often STs were quick and imminent. It was utterly rare that an identical ST (i.e. information need as defined here) would continue after performing another search or WT.

While the above ST definition may seem loose, information STs as understood in laboratory-like settings seldom have a clear equivalent in real life. Even if they had, a pre-constructed ST in a laboratory still remains an experimental artefact, whereas what people do in their work places is an IR phenomenon that is strictly dependent on the task at hand. In the field, people did not seem to analyse or plan their STs; there were no explicit traces of planning in the data. A query was issued in an information system if it was considered the optimal solution for the current information need. As the main purpose of the present paper is to describe real-life IR, the IR features to be studied had to be selected based on the data so that they are reliably analysable, potentially responsive to WT features, and of general interest in the IR community. Knowing the distributions of these IR features helps design and calibrate IIR experiments, for instance based on simulated WTs.

In the present paper, four IR features are analysed:

(1) The number of STs in a WT:

- Motivation: to understand if some WT categories involve more searching than others and why.

(2) The number of queries in an ST:

- Motivation: to understand if some information needs take longer, or require multiple approaches, to be fulfilled and why.

  The queries need not be unique; the searcher may as well type an identical query several times in different information systems or even in the same system and they are all counted. In reality, repeated queries in a single ST were rare, so the practical importance of this decision is small. However, for the searcher herself, a repeated query is as important as a unique one. Thus, there is no need to consider only unique queries here. A typical repeated query was, for example, a search for a translation of a word while reading a text. The word may recur after a while and it is not uncommon that the participant already forgot the meaning and has to search again.

(3) The median number of search keys:

- Motivation: query length may reflect the vagueness vs clarity of information needs.

  The number of search keys in each query of an ST was calculated and a median was taken across the queries in each ST. The number of search keys was calculated manually; space acted as an absolute word boundary and also other punctuation marks if they separated two words as judged intellectually. Also clear filtering conditions were calculated as search keys, such as selecting a retail chain from a drop-down menu (cf. point 4). The median was used instead of mean because it is not so sensitive to a few extremely long queries in the data.

(4) Query type:

- Motivation: query types may reflect the clarity of the information need which may further reflect the features of WTs.

Query types are intended to describe the exactness and versatility of queries in STs. Each query belongs to one of the following, mutually exclusive types: *v* (predefined attribute value), *f* (figure), *c* (common noun) or *p* (proper name). First, if the query is performed without any free-form typing of search keys, it is of type *v* (e.g. using only drop-down menus). The remaining query types are free-form to some extent. Second, if the search key is only a figure, the query type is *f*. "Figure" includes here both numerical sequences and pictorial queries; however, the only "pictorial queries" in the data were searching with molecular structures. Third, queries including common nouns form type *c*. If a query includes at least one proper name, it is judged "specific", that is, it belongs to the fourth type, *p*. A query is of only one type but an ST having several queries can include several query types, respectively. Query types are analysed as binary variables. Each ST therefore either has or does not have queries of each type (*v*, *f*, *c*, *p*). The number of different query types can range from 1 to 4. An ST gets value *c*, if all queries in it are type *c*, *cp* if some queries are type *c*, some *p*, and so on. Thus the most varied ST could be *vfcp*; it should have at least four queries, all of different types.

Below are some examples of query types in authentic STs:

(1) *v*: Selecting attribute values in various search fields in an internal database without typing any free-form query terms, such as when searching for sales figures.

Typical attributes in the present data set are a geographical district, a specific store, and time period.

(2) *f*: Checking product information in order to reply to an e-mail. Three queries are typed in an internal database: first, an identification number of purchase, then a product code, and last, another identification number of purchase. These are all type *f* because they are typed numbers only.

(3) *c*: Searching for a fact in a general web search engine using query "malonate".

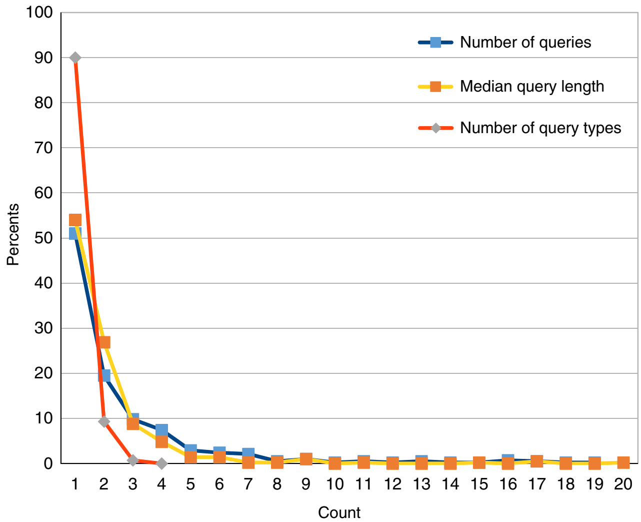
(4) *p*: Searching for a scientific article in a public web search engine. The query is formed of the name of the article, the writer's last name and a predefined attribute of estimated publication years. Because there are typed proper names involved, the query is of type *p*.

Results

Overall view of the data

Altogether 286 WTs were identified in the data. Their complexity varied from 0 to 90 per cent, the mean being 44 per cent. Task complexity was roughly normally distributed. Communication tasks formed 32.2 per cent (92 tasks) of the data, intellectual tasks 28.3 per cent (81) and editing tasks 24.1 per cent (69), respectively. Support tasks were the smallest category with 15.4 per cent (44) of the tasks. The 286 WTs contained in all 420 STs. However, 42 per cent of WTs did not include any STs. The maximum number of STs in a WT was 13, the next to maximum 11 STs.

Overall, STs are quite simple considering the features analysed in this study (see Figure 2). Over one half (51 per cent) of STs consists of only one query; and about 90 per cent of all STs include five queries or less. The maximum number of queries in a single ST is 42; the second largest number of queries is 22. Queries are short: the median query length of 54 per cent of STs is one search key. STs with median of



Notes: $n=420$; x-axis is cut and the maximum number of query types can be four

Figure 2. The distributions of number of queries, median query lengths, and number of different query types in STs

two-word queries are still quite common, about 27 per cent. The maximum median length of queries in an ST is 81, which is by far an outlier, since the second longest median query length is 17 keys. In order to smoothen the data, the outlier ST with 42 queries is from now on excluded from the analyses where the number of queries is discussed. Similarly, the ST with the median query length of 81 search keys will be excluded from the analyses concerning the median query length.

Query types *p* (proper names) and *c* (common nouns) are clearly most common. About 85 per cent of STs include queries of type *p* or *c* or both. Combinations of query types in STs are rare; this is again natural given the majority of short STs. Type *p* alone is the most common query type of ST, occurring in over half of STs. Below one fourth of STs are of type *c*.

IR in WTs of varying complexity

WT complexity's relationship to IR features is presented in Table II, which gives the means and medians of IR features across WT complexity levels I-IV. Moreover, it shows the correlation coefficients and their statistical significance. WT complexity is connected to the number of STs ($\varrho = 0.16$; $p = 0.006$): the more complex the WTs, the more STs it includes (see Table II (a)). The number of queries in an ST (see Table II (b))

Table II. The effects of WT complexity on STs and queries

	Work task complexity				
	I	II	III	IV	Total
(a) No. of search tasks in work tasks					
Mean	0.94	1.30	1.75	1.76	1.47
Median	0.00	1.00	1.00	1.00	1.00
n	63	67	84	72	N 286
ρ sig			$\varrho = 0.16$; $p = 0.006$		
(b) No. of queries in search tasks					
Mean	2.68	2.00	2.50	2.89	2.54
Median	2.00	1.00	1.00	2.00	1.00
n	59	87	147	*126*	N 419
ρ sig			$\varrho = 0.12$; $p = 0.017$		
(c) Median query length of search tasks					
Mean	2.22	1.99	1.95	1.83	1.96
Median	2.00	1.00	1.00	1.00	1.00
n	59	87	147	*126*	N 419
ρ sig			$\varrho = -0.12$; $p = 0.016$		
(d) No. of query types in search tasks					
Mean	1.07	1.06	1.09	1.18	1.11
Median	1.00	1.00	1.00	1.00	1.00
n	59	87	147	127	N 420
ρ sig			$\varrho = 0.15$; $p = 0.002$		
(e) Frequency of search tasks with each query type[a]					
p	35	66	91	66	258
c	23	24	37	54	138
f	0	2	25	19	46
v	5	0	7	11	23
n	59	87	147	127	N 420

Notes: Two figures missing an outlier are marked in italic. [a]ST can include several query types is weakly positively connected to WT complexity ($\varrho = 0.12$; $p = 0.017$). Mean and median show that the connection is curvilinear. However, studying the distributions more carefully shows that the longest STs (i.e. with more than three queries) form 29 per cent of STs in the most complex WTs (category IV). Though one-query STs form over a half of the data, their share is smallest in the most complex tasks (IV), 42 per cent. Moreover, 45 per cent of the longest STs (with more than three queries) take place in the most complex WTs (IV).

Interestingly, the median query length is negatively albeit weakly connected to task complexity ($\varrho = -0.12$; $p = 0.016$). Thus, the more complex the task, the shorter the queries (Table II (c)). The drop in the mean seems quite small, but the difference is quite clear between the most simple and most complex WTs: data distribution reveals that though over one half of the queries have only one search key, the share of these short queries is largest (58 per cent) in the STs of the most complex WTs (IV), and the share of long-query STs (median query length of three or more search keys) is largest (24 per cent) in the most simple WTs (I). Perhaps it is easy to come up with search terms in simple WTs.

The number of different query types in an ST can vary between one and four but in our data the maximum was three (which forms only 0.7 per cent of the data). This aggregated variable correlates significantly with task complexity ($\varrho = 0.15$; $p = 0.002$). Table II (d) shows that in the most complex WTs (IV), there are a little more query types on average than in simpler tasks (I-III). Actually, STs with two or three query types are clearly more common in the most complex WTs (IV) than in simpler ones (I-III), based on data distributions.

Table II (e) shows the frequencies of STs containing queries of each type. The table is interpreted as follows: a query is of only one type, but an ST can include several queries and thus several different query types, respectively. The query type frequencies in Tables II (e), III (e) and IV (e) are "dichotomic". For example, 35 STs of the total 59 STs (59 per cent) in simple (I) WTs include *p* queries; the rest, 24 STs (41 per cent) do not. The same applies to the other query types. The text includes a few calculation examples.

V queries (predefined attribute value) are rare (found in only 23 STs). No linear trend is seen between task complexity categories despite the fact that the majority of *v*-query STs (78 per cent) ((7+11)/23)) happen in the more complex half of WTs (III-IV). *F* queries (figures) share a similar pattern: 96 per cent of them happen in the more complex half.

C- and *p* queries are more interesting because they are far more common. *C*-query STs pile up into the most complex WTs (IV): 43 per cent (54/127) of STs in complex WTs have *c* queries and 39 per cent of all *c*-queries appear in the most complex tasks. *P* queries differ from other query types. In all task complexity categories, STs including *p* queries are more common than STs without them. However, the share of STs with *p* queries is smallest (52 per cent) in the most complex tasks (IV).

IR in different types of WTs

The relationships of IR features and WT types are presented in Table III which is similar in structure with Table II. Note however that since WT types are qualitative features, calculating correlation was not possible. WT types' connections to the number of STs in WTs are shown in Table III (a). The average numbers of STs per WT in support, editing and intellectual tasks are 1.25, 1.43, and 1.65, respectively. Communication tasks have 1.43 STs per WT. Thus, intellectual tasks are most search-intensive WTs. Support tasks are interesting because they include the largest shares of both WTs with zero and with several STs. This means that typically, support

Table III.
The effects of WT type on STs and queries

	Work task type Communication	Support	Editing	Intellectual	Total
(a) No. of search tasks in work tasks					
Mean	1.43	1.25	1.43	1.65	1.47
Median	1.00	0.00	1.00	1.00	1.00
n	92	44	69	81	N 286
(b) No. of queries in search tasks					
Mean	2.05	2.59	2.12	3.30	2.54
Median	1.00	2.00	1.00	2.00	1.00
n	132	*54*	99	134	N 419
(c) Median query length of search tasks					
Mean	1.74	2.20	2.00	2.05	1.96
Median	1.00	2.00	1.00	2.00	1.00
n	132	55	99	*133*	N 419
(d) No. of query types in search tasks					
Mean	1.08	1.15	1.05	1.16	1.11
Median	1.00	1.00	1.00	1.00	1.00
n	132	55	99	134	N 420
(e) Frequency of search tasks with each query type[a]					
p	90	33	65	70	258
c	28	14	37	59	138
f	17	6	0	23	46
v	8	10	2	3	23
n	132	55	99	134	N 420

Notes: Two figures missing an outlier are marked in italic. [a]ST can include several query types

tasks are performed with a lot of searching or entirely without it; intermediate forms are rare in this data set.

In all WT types, having only one-query STs is more common than having two-three query STs; which then is more common than having more than three queries in an ST. However, in intellectual tasks, having one, two-three, or more queries is almost equally common, and the mean number of queries is higher than in other task types. It is notable that support tasks come close to intellectual tasks in the number of queries (Table III (b)). STs in communication tasks tend to have the shortest queries (Table III (c)). Again, support and intellectual tasks resemble each other in that they have longer queries in their STs. The sum of various query types (Table III (d)) reveals that in support tasks and intellectual tasks, queries are more variable than in other types of WTs. The differences are small, however.

The occurrence of query types is again analysed as binary variables (Table III (e)). The rare *v*-query searches happen almost exclusively in communication and support tasks (78 per cent in all, i.e. (8+10)/23)) whereas also the rare *f*-query searches take place in communication and intellectual tasks (87 per cent). *C*-query STs (as opposed to non-*c*-query STs) become more common in the continuum support, editing, intellectual tasks ranging from 25 per cent (14/55) to 44 per cent. *P*-query STs differ from other query types in the sense that in all WT types, it is more common that STs include *p* queries than that they do not. This may partly be because proper names are highest in the hierarchy of query types, that is, a single proper name changes the type of the query despite any other search keys (see the explanation in the analysis section). However, this is also a typical feature of the data: having proper names in most of the queries indicates quite specific information needs. *P*-query STs are least common in intellectual tasks (52 per cent). In communication tasks, *p* queries are more common than in other task types. This is perhaps caused by extensive searching for people in e-mail and instant messaging programmes, typical information sources in communication tasks.

The effects of task complexity within task types

Next, the differences between searching in simple and complex WTs are discussed in the context of WT types. In order to study task type and task complexity's mutual effects on searching, we divided each task type into two categories; simple and complex tasks. Simple ones are former complexity categories I and II (complexity 0-38.3 per cent), and complex ones categories III and IV (38.4-100 per cent). As done in the previous sections, when suitable for the variables in question, correlations were calculated with the equidistant WT complexity measure. Results of the analyses are presented in Table IV.

The number of STs seems to react differently to WT complexity depending on WT type (Table IV (a)). First, the number of STs does not change with WT complexity in

	Work task type								
	Communication		Support		Editing		Intellectual		
	Simple	Complex	Simple	Complex	Simple	Complex	Simple	Complex	Total
(a) No. of search tasks in work tasks									
Mean	1.00	1.85	0.60	2.11	1.47	1.39	1.36	1.76	1.47
Median	0.00	1.00	0.00	2.00	1.00	1.00	1.00	1.00	1.00
n	45	47	25	19	38	31	22	59	N 286
ρ sig	$\varrho = 0.21; p = 0.041$		$\varrho = 0.41; p = 0.006$		$\varrho = 0.05; p = 0.706$		$\varrho = 0.07; p = 0.53$		
(b) No. of queries in search tasks									
Mean	2.02	2.07	2.40	2.67	1.88	2.44	3.33	3.29	2.54
Median	1.00	1.00	1.00	2.00	1.00	1.00	2.00	2.00	1.00
n	45	87	15	*39*	56	43	30	104	N 419
ρ sig	$\varrho = 0.08; p = 0.372$		$\varrho = 0.16; p = 0.252$		$\varrho = 0.05; p = 0.605$		$\varrho = -0.03; p = 0.749$		
(c) Median query length of search tasks									
Mean	1.96	1.63	2.07	2.25	2.36	1.53	1.77	2.13	1.96
Median	1.00	1.00	2.00	2.00	2.00	1.00	1.50	2.00	1.00
n	45	87	15	40	56	43	30	*103*	N 419
ρ sig	$\varrho = -0.01; p = 0.882$		$\varrho = -0.33; p = 0.014$		$\varrho = -0.29; p = 0.004$		$\varrho = -0.08; p = 0.345$		
(d) No. of query types in search tasks									
Mean	1.07	1.09	1.13	1.15	1.04	1.07	1.07	1.18	1.11
Median	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
n	45	87	15	40	56	43	30	104	N 420
ρ sig	$\varrho = 0.07; p = 0.412$		$\varrho = 0.12; p = 0.381$		$\varrho = 0.06; p = 0.587$		$\varrho = 0.18; p = 0.04$		
(e) Frequency of search tasks with each query type[a]									
p	34	56	9	24	38	27	20	50	258
c	10	18	5	9	20	17	12	47	138
f	2	15	0	6	0	0	0	23	46
v	2	6	3	7	0	2	0	3	23
n	45	87	15	40	56	43	30	104	N 420

Notes: Two figures missing an outlier are marked in italic. [a]ST can include several query types

Table IV. WT complexity's effects on STs across WT types editing or in intellectual tasks. In intellectual tasks, there is a slight increase in the mean which is not supported by the correlation. In communication tasks, the number of STs grows slightly with WT complexity. However, the number of STs is clearly connected to task complexity in support tasks ($\varrho = 0.41$; $p = 0.006$). While there was an overall tendency that WT complexity increases the number of queries in STs (Table IV (b)), the weak dependency vanished in the more elaborate type-wise analysis. Support and editing tasks seem to have an increase in the number of queries in STs between the two complexity categories though not supported by the correlations.

The median query length of STs requires careful analysis because the figures in Table IV (c) appear contradictory. The overall effect of WT complexity was, as seen earlier, to decrease query length. Editing tasks are the only WT type where this can be seen to be unequivocally true: the correlation is statistically highly significant ($p = 0.004$), and satisfactorily high ($\varrho = -0.29$) considering the data. Also, both the mean and the median decrease. In communication tasks, the query lengths keep approximately at the same level. In support tasks, median query length is connected to WT complexity ($\varrho = -0.33$; $p = 0.014$): the more complex the task, the shorter the queries. However, the mean seems to increase from 2.07 to 2.25, which must be caused by some outlier; the mean is sensitive to them.

In intellectual tasks, median query length seems to grow with task complexity. A closer review of the distributions shows that the growth is actually ambiguous. The distributions of STs between median query lengths of 1, 2, and more search keys are almost identical in simple and complex tasks. However, in simple WTs, exactly 50 per cent of STs have median query length of one key, whereas it is 49.5 per cent in the complex tasks. This causes the growth in the median. The mean is affected by the fact that the range of median query length is larger in STs of complex WTs; a few STs with long queries increase the mean in complex tasks. The correlation supports the interpretation that WT complexity does not affect the median query length of STs in intellectual WTs. Overall the median query length keeps at the same level independently of WT complexity, but when there are long queries, they tend to occur in complex intellectual tasks rather than in simple ones.

It is clear that queries become more diverse when task complexity increases only in intellectual tasks ($\varrho = 0.18$; $p = 0.004$). Otherwise, WT complexity does not affect the number of query types. Overall, *p* queries are the most common query type despite WT complexity or WT type. In editing tasks, WT complexity is not connected to the change of occurrence of any query types. In both communication and intellectual tasks, the proportion of STs with *p* queries decreases and with *f* queries increases. The proportion of *f* queries increases also in support tasks, whereas the proportion of *c* queries decreases in them. Each query type reaches its own peak as follows: *p* queries are most common in simple communication tasks (34/45 = 76 per cent), *c* and *f* queries in complex intellectual tasks, and *v* queries in simple support tasks.

Based on the frequencies of query types, it can be concluded that queries are typically quite specific. The fact that "figure only" queries typically increase with task complexity is an interesting notion, since a figure is a precise search term. Perhaps the finding is related to the information systems used or the way they are used. STs with proper name queries become clearly less common only in communication and intellectual tasks. Especially interesting is that proper name queries, though overall common, are least common in the STs of complex intellectual WTs. On the other end of the continuum there are simple communication tasks, where proper names occur in almost 80 per cent of all STs.

Discussion

This study contributes to the knowledge of IR as a real-life phenomenon. The effects of WT type and complexity on IR features (the number of STs, queries, search keys, and the types of queries) were uncharted before. In this section, two main contributions of the paper are discussed: First, the growing understanding of IR by using real-life data and how it affects designing future studies; and second, the empirical findings that show complex relations between WTs and IR.

IR as a real-life phenomenon

In the present study, the focus was, first, on carefully identifying WTs in the data. The second focus was on their subtasks containing IR, that is, STs. Other types of subtasks were left unanalysed. Some routine rules were first applied when defining an ST. As subtasks of WTs, STs cannot understandably cross the boundaries of WTs; and neither can overlap each other. This was partly a pragmatic decision, partly because an information need is a fluctuating state of mind (cf. Belkin's (1980) model of Anomalous States of Knowledge); if one IR session ends and another begins, it seems unlikely that a person would later return to and continue from exactly the same information need ("state of knowledge"). Here the information need or the ST is not a written description handed to the participant but rather a self-created and implicit phrasing of a question or a state of mind. This corresponds to the "internal generated" value of "source of task" facet in Li and Belkin's (2008) comprehensive task classification.

This view of non-overlapping STs differs from Spink *et al.* (2002) view on the popularity of multitasking. Their view is based on overlapping of abstract *topics* that were of recurring interest for the searchers. However, our interpretation of information needs is based on the fact that they were created ad hoc by the participants and immediately led to concrete steps of IR, the ST. Such quick, time-bound episodes of arising information needs and instant searching are not so likely to overlap.

The findings of the study supported the view that IR is one of many means to achieve a goal, the task outcome (Ingwersen and Järvelin, 2005; Järvelin *et al.*, 2015; Rose and Levinson, 2004; Wildemuth and Freund, 2009). This is seldom explicitly taken into account empirically. Assigned STs in test settings keep static. However in authentic WTs, both the physical retrieval actions and the mental model of the task performer are all the time in a complex interaction with all the possible contexts (cf. the shifts in intentions and information seeking strategies analysed by Xie (2000)). This is in line with the acknowledged difficulty of defining context conceptually (see e.g. Courtright, 2007; Dervin, 1997). Thus calling the authentic phenomenon information searching instead of IR seems more appropriate; people are trying to find information rather than simply retrieving it.

A typical ST in the data were in a common WT, "replying/reacting to e-mails". The e-mails received need not include extensive requests for information but even small questions and situational notices from a supplier may lead to further information needs and searching beyond the obvious matters presented in the e-mail. Finding information to reply e-mails could last most of the working day leaving little time to the so-called core of work.

Another example of information-intensive but not IR-driven WT is a situation where a manager has a task related to a decision to update the design of a product. She does not design by herself but coordinates it. During the observation, she tried to gather together the best experts within the organisation. Mostly she knows them by heart, but sometimes e-mail addresses have to be checked. This is quite trivial. However, she suddenly has the suspicion that one of the experts has been replaced by another one she does not yet know. Because she is not sure, she tries to find the information from the intranet. That being unsuccessful, she has to call to two colleagues, which requires identifying the right people and finding their contact information.

Afterwards, one can say that in the above situation, there was an ST of finding that missing expert, and the participant herself might have been able to tell beforehand, if asked in enough detail, that she has to find the relevant experts. However, the problem (the uncertainty) was unlikely to be known beforehand, and the situation led to a small-scale but important IR in view of the larger task she was responsible for, the successful design of a product. The participant was also able to choose – at least in theory – between several options how to proceed with the task.

Since IR is a natural and inseparable phase in the flow of WTs, it is difficult to integrate into simulated WTs. In the present study, the data had only two WTs where the participant explicitly stated that the WT goal was to find information. STs did not exist as systematically designed assignments in the minds of the participants but they were normally really spontaneous and thus quickly evolving. Though definitely an important part of many tasks, IR was surprisingly rare; there were STs in only 60 per cent of all WTs. STs were simple: half of them had only one query. This seems to be connected with the role of IR in WT performance. It may be crucial but the WT is not in itself an ST, the ruling type of simulated WTs where IR is something the participants are expected to do, despite the background story.

The role of IR in real life sets new questions on how to assess search performance. The relevance of documents or some other measure of success is often of interest in traditional IR studies and even in naturalistic IIR studies (Kelly, 2006). Measuring success was intentionally left out from the present study because other study interests seemed more relevant in the context of real life. As already emphasised in this paper, STs were intertwined into other work activities; STs and information needs arouse and ended naturally. There were no official assignments whose information requirements could be compared to the output or outcome of IR. The main issue was to proceed in the WT and continuously select suitable paths to pursue. Successful IR at least sped up the task process but a failing query simply had to be managed somehow. Abandoning a WT was not an option. Since information needs also evolve it would have been quite hard to evaluate success as an outsider or even to cut the STs into components to be evaluated. Should this have been done after each query? What about exploratory searching?

Empirical findings

In this study, IR was studied without stimulating it; i.e., simply by observing WTs and noting when an ST occurs. It was found that not all WTs include IR. In the WTs that did, it seemed quite simple. The STs were short in the number of queries, as well as the queries in the number of search keys. The prevalence of short queries is shown in many earlier studies (Silvestri, 2010); people do not search more than they have to. Jansen and Spink (2006) show the prevalence of single-query search sessions: on average, they comprise about a half of all sessions. This is surprisingly similar to our finding though our "sessions" were not technically but content-wise limited STs and included also other sources besides the web. Overall, queries were quite specific because most STs featured queries with proper names. This finding is in line with those of Huurnink *et al.* (2010). Spink *et al.* (2004) found that in web search engines, queries with personal names (a subset of proper names) are quite common. Of course, personal names in web searching differ from the multi-source, work-related searching in the present study: Spink *et al.* (2004) found that most personal names searched for were names of celebrities whereas in our study, a typical person name query was for finding contact information.

Growing task complexity slightly increases the number of queries in STs and the number of STs in WTs. This could be expected since the more complex the task is, the more new information is needed. These findings are in agreement with Saastamoinen *et al.* (2012) and Aula *et al.* (2010). However, the correlations found here are so weak that several other factors must affect the number of STs and queries along WT complexity. High WT complexity may indicate that the WT is formed of several subtasks with different information needs, but fulfilling these information needs does not self-evidently lead to clearly increased IR. This is because task performers have several options to proceed: issuing a query in an information system is seldom precisely planned beforehand but only performed "ex tempore" if considered the optimal way to go.

Based on the correlation coefficient, queries tend to become shorter when task complexity increases. It may imply that the information needs behind queries become more unspecific (cf. Vakkari, 2001). This interpretation is supported by the fact that the share of STs with proper name queries is smallest in the most complex tasks. The connections between shorter queries and more complex tasks has been referred to in, e.g., Liu *et al.* (2015), Hansen (2011) and Vakkari (2001). Vakkari's findings also show that the more complex the task, the more vague are the search terms used. Aula *et al.* (2010) defined difficult STs as unsuccessful and easy as successful. Their findings show, contrary to the ones presented here, that the difficulty slightly increases the number of query terms per query. The difference can be caused by many factors: a complex WT does not necessarily imply complex STs (Aula *et al.* (2010) considered only STs). Also the methods and operationalisation of complexity were different.

Exploratory vs lookup STs are discussed by, e.g., Marchionini (2006), White and Marchionini (2007), and Wildemuth and Freund (2009, 2012). To some extent, explorative searches can be considered complex and lookup searches simple; and by definition exploratory STs should include more searching (White and Marchionini, 2007). However, a complex WT does not determine that the STs are complex. Exploratory searching is often defined by multiple factors (Wildemuth and Freund, 2009). The definitions seem not to differentiate between complex information needs and complex search actions which may not coincide. A factual, "simple", information need may end up in complex search actions, and on the other hand, a larger, topical information need may be quickly fulfilled. The earlier findings based on the same data set (Saastamoinen and Järvelin, in press) show that even in the complex, and in the intellectual tasks, most information needs and search processes are simple. However, topical needs and developing search processes are more common in the most complex and in the intellectual tasks than in other task types or complexity categories (Saastamoinen and Järvelin, in press).

WT types brought an interesting perspective and a contrast to task complexity. To some extent, support, editing and intellectual tasks formed a continuum equivalent to task complexity. Common nouns as query types in STs became more common along this continuum which may imply the information needs becoming vaguer (cf. the paragraph above). In intellectual WTs, the STs featured more queries than in other task types. This resembles the findings of the connections between WT complexity and STs.

It seems that task complexity's effects on IR are best elicited within task types. In support tasks, correlations clearly show that WT complexity increases the number of STs and shortens the queries. Thus WT complexity is actually an important factor affecting IR in support tasks, though they can easily be neglected in IR research as uninteresting routine. Typical assigned STs or simulated WTs often require searching information about something rather unfamiliar to the participant in order to compile some sort of aggregate of the information found. Still, these kinds of tasks seem rare in the flow of work and they would also represent quite complex tasks. Support tasks – if found complex – are often experienced especially frustrating because the task performers, too, anticipate that they are routine. It is interesting, that in editing tasks, simple WTs include more STs than complex ones. However, at the same time the share of STs with several queries increases clearly and queries become shorter with growing WT complexity.

Intellectual tasks. Intellectual tasks are the opposite of support tasks: task complexity does not affect almost at all the number of STs or the query length. This indicates that searching is built-in in tasks where something new is created. However, the types of queries change and become more variable: task complexity does not affect the quantity of searching but its quality.

Communication tasks. Communication tasks are special because they are common but varying in their content. Nearly a third of WTs belong to this category which is similar to the findings of Czerwinski *et al.* (2004). Communication tasks are important, though they actually often only support other tasks. When communication tasks become more complex, the number of STs increases and queries change. With growing task complexity, searching with people's names and other proper names decreases and using figures increases. This finding may originate especially from communication tasks in commercial companies. Figures are really specific search keys (such as product codes) that worked as starting points when the participants were trying to solve complex inquiries received through e-mail, for example. Thus seemingly specific search keys worked in the context of more muddled information needs.

Conclusions

This study provided an in-depth view on authentic IR. IR is seldom studied with its natural context that is underlying WTs. The main contributions include the following:

The basic principle through all data collection and analysis phases was to stay true to the authentic phenomena of IR. It means reconsidering many preconceptions and prevailing practices. One of the main points was that the researcher identified the endogenous, spontaneous STs rather than setting an experiment. IR is an integral part of the flow of WTs. People do not normally start working search-minded. It may be that STs could be afterwards clearly identified by researchers or even by the participants themselves; but IR is seldom a part of a work plan. Without a doubt, this is a feature dependent on work role, profession and task. However, the search actions in the present data were mainly quick (often nearly inadvertent) choices made between, for example, browsing, calling a colleague, looking up in a book, delegating, etc. Why IR was selected among the alternatives would be an interesting research question.

To maintain validity of IR experiments, attempts should be made to retain the work flow context in the more controlled experimental environments. It would be important to study whether the claim that simulated WT situations are actually rare in practice also holds for other settings than the ones analysed here. The present findings can be exploited when designing simulated WTs that would better correspond to authentic WTs. The scope of IR studies should exceed search-intensive tasks and include also tasks where IR is brief while still important. Otherwise the development and evaluation of IR methods are not based on the whole range, or even the common types, of STs. In reality, a single ST may be straightforward, but together all STs form a versatile whole, considering, for instance the information resources. In a typical (I)IR experiment, a major deficiency is that the participants cannot choose where to search.

The findings suggest that simulated WTs could be more problem-driven rather than directly stating to the participant that the idea is to find certain kinds of documents, which is an unusual type of WT in the present data. This would require more tailoring of the simulated WTs. The participants should be directed in a situation where they decide to search rather than letting them read the ready-made information (or document) need. An option would be to ask the participant first to do something that is not actually in focus, but that leads to some sort of "problematic situation" where she has to find further information to proceed in the initial task. This problem solving stage would represent a naturally formed ST, the key interest for the researchers.

A data-driven classification for WT types was formed. It proved useful in providing further insight into IR actions. The WT types do not depend on organisation type but represent abstract features of tasks. They might serve as a starting point when classifying tasks or designing simulated WTs in later studies. The task type classification seems applicable to other organisational settings beyond the ones studied. However, its applicability may depend on how information about the WTs is collected. In the present paper, participants' free-form, written task descriptions were used. The WTs may appear differently if described by the researchers. Since this was the first time this classification was applied, it may need revisions in the future. Perhaps some work roles include several WTs that build around information seeking, which then could be an additional category.

The authors' opinion is that naturalistic IR studies should be regularly conducted to ensure the relevance of state-of-the-art IR experiments for the end-users and authentic use situations. An interesting future study object could be information workers in small companies. Typically, their work is really variable (compared to large companies where the work roles are potentially more differentiated), and they may not have expensive and advanced information systems and databases in use. This kind of environment places high demands on the available systems and their ways of use.

It would be optimal to conduct naturalistic studies in a larger research group to increase the number of participants without losing the quality, the depth of the study. With a larger data set, a more thorough statistical analysis would be possible. This would foster theory growth in the field. The present study is exploratory, but in the future, even naturalistic studies can be more clearly focused without having to artificially control the participants. For these purposes, the paper presented easily applicable operational definitions of WTs, STs and their features. Instead of focusing on the development and evaluation of individual information systems, it is time to find out how people act in their information environments, what the role of various information systems is in their activities, and when, why and how they ask questions to be answered by which systems.

References

Aula, A., Khan, R.M. and Guan, Z. (2010), "How does search behavior change as search becomes more difficult?", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM*, New York, NY, pp. 35-44.

Belkin, N.J. (1980), "Anomalous states of knowledge as a basis for information retrieval", *Canadian Journal of Information Science*, Vol. 5, May, pp. 133-143.

Bell, D.J. and Ruthven, I. (2004), "Searcher's assessments of task complexity for web searching", in McDonald, S. and Tait, J. (Eds), *Advances in Information Retrieval*, Springer, Berlin, pp. 57-71.

Borlund, P. (2003), "The IIR evaluation model: a framework for evaluation of interactive information retrieval systems", *Information research*, Vol. 8 No. 3, available at: www.informationr.net/ir/8-3/paper152.html (accessed 6 July 2015).

Borlund, P. and Ingwersen, P. (1997), "The development of a method for the evaluation of interactive information retrieval systems", *Journal of Documentation*, Vol. 53 No. 3, pp. 225-250.

Byström, K. (1999), "Task complexity, information types and information sources", Acta Universitatis Tamperensis 688, University of Tampere, Tampere, available at: www.adm.hb.se/~KBM/kby-diss.pdf (accessed 18 April 2016).

Byström, K. and Hansen, P. (2005), "Conceptual framework for tasks in information studies", *Journal of the American Society for Information Science and Technology*, Vol. 56 No. 10, pp. 1050-1061.

Byström, K. and Järvelin, K. (1995), "Task complexity affects information seeking and use", *Information Processing & Management*, Vol. 31 No. 2, pp. 191-213.

Courtright, C. (2007), "Context in information behavior research", *Annual Review of Information Science and Technology*, Vol. 41 No. 1, pp. 273-306.

Czerwinski, M., Horvitz, E. and Wilhite, S. (2004), "A diary study of task switching and interruptions", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, New York, NY*, p. 175.

Dervin, B. (1997), "Given a context by any other name: methodological tools for taming the unruly beast", *Proceedings of an International Conference on Information Seeking in Context, Taylor Graham Publishing, London*, pp. 13-38.

Du, J.T. (2012), "Information use and information sharing in marketing: a diary study", *Proceedings of the ASIST Annual Meeting*, Vol. 49 No. 1, pp. 1-4.

Du, J.T. (2014), "The information journey of marketing professionals: incorporating work task-driven information seeking, information judgments, information use, and information sharing", *Journal of the Association for Information Science and Technology*, Vol. 65 No. 9, pp. 1850-1859.

Haerem, T. and Rau, D. (2007), "The influence of degree of expertise and objective task complexity on perceived task complexity and performance", *Journal of Applied Psychology*, Vol. 92 No. 5, pp. 1320-1331.

Hansen, P. (2011), "Task-based information seeking and retrieval in the patent domain", *Processes and Relationships*, academic dissertation, University of Tampere, Tampere.

Huurnink, B., Hollink, L., van den Heuvel, W. and de Rijke, M. (2010), "Search behavior of media professionals at an audiovisual archive: a transaction log analysis", *Journal of the American Society for Information Science and Technology*, Vol. 61 No. 6, pp. 1180-1197.

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

Jansen, B.J. (2009), "The methodology of search log analysis", in Jansen, B.J., Spink, A. and Taksa, I. (Eds), *Handbook of Research on Web Log Analysis*, Information Science Reference, Hershey, pp. 100-123, available at: https://faculty.ist.psu.edu/jjansen/academic/jansen_search_log_analysis.pdf (accessed 29 February 2016).

Jansen, B.J. and Spink, A. (2006), "How are we searching the World Wide Web? A comparison of nine search engine transaction logs", *Information Processing & Management*, Vol. 42 No. 1, pp. 248-263.

Järvelin, K. (1986), "Information technology and information use: towards a unified view of information and information technology", in Ingwersen, P., Kajberg, L. and Pejtersen, A.M. (Eds), *On Information, Information Technology and the Development of Society: An Information Science Perspective*, Taylor Graham Publishing, London, pp. 35-55.

Järvelin, K., Vakkari, P., Arvola, P., Baskaya, F., Järvelin, A., Kekäläinen, J., Keskustalo, H., Kumpulainen, S., Saastamoinen, M., Savolainen, R. and Sormunen, E. (2015), "Task-based information interaction evaluation: the viewpoint of program theory", *ACM Transactions on Information Systems*, Vol. 33 No. 1, pp. 3:1-3:30.

Kellar, M., Watters, C. and Shepherd, M. (2006), "The impact of task on the usage of web browser navigation mechanisms", *Proceedings of Graphics Interface 2006, Canadian Information Processing Society, Toronto*, p. 235.

Kelly, D. (2006), "Measuring online information seeking context, Part 1: background and method", *Journal of the American Society for Information Science and Technology*, Vol. 57 No. 13, pp. 1729-1739.

Kelly, D. (2009), "Methods for evaluating interactive information retrieval systems with users", *Foundations and Trends in Information Retrieval*, Vol. 3 Nos 1-2, pp. 1-224.

Kuhlthau, C.C. (1993), *Seeking Meaning*, Ablex, Norwood, NJ, available at: http://faculty.arts.ubc.ca/emeyers/LIBR535/readings/Kuhlthau_2004.pdf (accessed 25 February 2016).

Kuhlthau, C.C. (2005), "Towards collaboration between information seeking and information retrieval", *Information Research*, Vol. 10 No. 2, available at: http://InformationR.net/ir/10-2/paper225.html (accessed 8 February 2016).

Kumpulainen, S. (2014), "Trails across the heterogeneous information environment", *Journal of Documentation*, Vol. 70 No. 5, pp. 856-877.

Li, Y. (2009), "Exploring the relationships between work task and search task in information search", *Journal of the American Society for Information Science and Technology*, Vol. 60 No. 2, pp. 275-291.

Li, Y. and Belkin, N.J. (2008), "A faceted approach to conceptualizing tasks in information seeking", *Information Processing & Management*, Vol. 44 No. 6, pp. 1822-1837.

Li, Y. and Hu, D. (2013), "Interactive retrieval using simulated versus real work task situations: differences in sub-facets of tasks and interaction performance", *Proceedings of the American Society for Information Science and Technology*, Vol. 50 No. 1, pp. 1-10.

Lingel, J. (2015), "Information practices of urban newcomers: an analysis of habits and wandering", *Journal of the Association for Information Science and Technology*, Vol. 66 No. 6, pp. 1239-1251.

Liu, J., Kim, C.S. and Creel, C. (2015), "Exploring search task difficulty reasons in different task types and user knowledge groups", *Information Processing & Management*, Vol. 51 No. 3, pp. 273-285.

Liu, P. and Li, Z. (2012), "Task complexity: a review and conceptualization framework", *International Journal of Industrial Ergonomics*, Vol. 42 No. 6, pp. 553-568.

Marchionini, G. (2006), "Exploratory search: from finding to understanding", *Communications of the ACM*, Vol. 49 No. 4, pp. 41-46.

Nicholas, D., Clark, D., Rowlands, I. and Jamali, H.R. (2009), "Online use and information seeking behaviour: institutional and subject comparisons of UK researchers", *Journal of Information Science*, Vol. 35 No. 6, pp. 660-676.

Nicholas, D., Williams, P., Rowlands, I. and Jamali, H.R. (2010), "Researchers' e-journal use and information seeking behaviour", *Journal of Information Science*, Vol. 36 No. 4, pp. 494-516.

Pharo, N. (2002), *The SST Method Schema: A Tool for Analysing Work Task-Based Web Information Search Processes*, Tampere University Press, Tampere.

Pharo, N. (2004), "A new model of information behaviour based on the search situation transition schema", *Information Research*, Vol. 10 No. 1, available at: www.informationr.net/ir/10-1/paper203.html (accessed 6 July 2015).

Pyöriä, P. (2005), "The concept of knowledge work revisited", *Journal of Knowledge Management*, Vol. 9 No. 3, pp. 116-127.

Robinson, J. and Yerbury, H. (2015), "Re-enactment and its information practices; tensions between the individual and the collective", *Journal of Documentation*, Vol. 71 No. 3, pp. 591-608.

Rose, D.E. and Levinson, D. (2004), "Understanding user goals in web search", *Proceedings of the 13th International Conference on World Wide Web, ACM, New York, NY*, pp. 13-19.

Ruotsalo, T., Athukorala, K., Glowacka, D., Konyushkova, K., Oulasvirta, A., Kaipiainen, S., Kaski, S. and Jacucci, G. (2013), "Supporting exploratory search tasks with interactive user modeling", *Proceedings of the American Society for Information Science and Technology*, Vol. 50 No. 1, pp. 1-10.

Saastamoinen, M. and Järvelin, K. (in press), "Search task features in work tasks of varying types and complexity", *Journal of the Association for Information Science and Technology*.

Saastamoinen, M., Kumpulainen, S. and Järvelin, K. (2012), "Task complexity and information searching in administrative tasks revisited", *Proceedings of the 4th Information Interaction in Context Symposium, Nijmegen*, pp. 204-213.

Silvestri, F. (2010), "Mining query logs: turning search usage data into knowledge", *Foundations and Trends in Information Retrieval*, Vol. 4 Nos 1-2, pp. 1-174.

Spink, A., Jansen, B.J. and Pedersen, J. (2004), "Searching for people on web search engines", *Journal of Documentation*, Vol. 60 No. 3, pp. 266-278.

Spink, A., Ozmutlu, H.C. and Ozmutlu, S. (2002), "Multitasking information seeking and searching processes", *Journal of the American Society for Information Science and Technology*, Vol. 53 No. 8, pp. 639-652.

Vakkari, P. (1999), "Task complexity, information types, search strategies and relevance: integrating studies on information seeking and retrieval", in Wilson, T.D. and Allen, D.K. (Eds), *Exploring the Contexts of Information Behaviour*, Taylor Graham Publishing, London, pp. 35-54, available at: www.informationr.net/isic/ISIC1998/98_Vakkari.pdf (accessed 8 February 2016).

Vakkari, P. (2000), "Relevance and contributing information types of searched documents in task performance", *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY*, pp. 2-9.

Vakkari, P. (2001), "A theory of the task-based information retrieval process: a summary and generalisation of a longitudinal study", *Journal of Documentation*, Vol. 57 No. 1, pp. 44-60.

Vakkari, P. (2003), "Task-based information searching", *Annual Review of Information Science and Technology*, Vol. 37 No. 1, pp. 413-464.

Vakkari, P., Pennanen, M. and Serola, S. (2003), "Changes of search terms and tactics while writing a research proposal a longitudinal case study", *Information Processing & Management*, Vol. 39 No. 3, pp. 445-463.

White, R.W. and Marchionini, G. (2007), "Examining the effectiveness of real-time query expansion", *Information Processing & Management*, Vol. 43 No. 3, pp. 685-704.

Wildemuth, B.M. and Freund, L. (2012), "Assigning search tasks designed to elicit exploratory search behaviors", *Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval, ACM, New York, NY*.

Wildemuth, B.M. and Freund, L. (2009), "Search tasks and their role in studies of search behaviors", in Kules, B., Tunkelang, D. and White, R. (Eds), *Third Annual Workshop on Human Computer Interaction and Information Retrieval*, ACM, Washington, DC and New York, NY, pp. 17-21.

Wildemuth, B., Freund, L. and Toms, E.G. (2014), "Untangling search task complexity and difficulty in the context of interactive information retrieval studies", *Journal of Documentation*, Vol. 70 No. 6, pp. 1118-1140.

Xie, H. (2000), "Shifts of interactive intentions and information-seeking strategies in interactive information retrieval", *Journal of the American Society for Information Science*, Vol. 51 No. 9, pp. 841-857.

Appendix

Morning questionnaire

What WTs do you have to perform today? Base your replies on your present knowledge.

Contact information

Name:

Organisation:

Tasks

[The questionnaire includes place holders for 5 tasks. Each slot is of size one line and approx. 40 characters but allowing the participants to type more. The participants fill in as many tasks as they need to. For each:]

The WTs of today. Task complexity should be given as a number 0 (really simple)-100 (really complex).

Task description:

Task complexity 0%-100%:

Task performance. You should answer with a number 0 (not at all)-100 (perfectly) to the question of how well you know the task performance process.

Programs/information systems needed:

Other information/sources of information needed:

How well do you know the task performance process?:

Other remarks (e.g. about the tasks or the course of the day):

Afternoon questionnaire

What WTs did you perform today? Think about the whole working day from the morning to the afternoon as you reply.

Contact information

Name:

Organisation:

Tasks

The WTs of today. Task complexity should be given as a number 0 (really simple)-100 (really complex).

[The questionnaire includes place holders for 5 tasks. The participants fill in as many tasks as they need to. For each:]

Task description:

Task complexity 0%-100%:

At what time (approximately) did you perform the task? (E.g. 12-13)

Task performance

Programs/information systems you used:

Other information/sources of information you used:

Did any problems turn up in the task? Please describe them.

[The questionnaire includes places for 5 tasks; the participants fill in as many as they need to.]

Other remarks (e.g. about the tasks or the course of the day):

Corresponding author

Miamaria Saastamoinen can be contacted at: miamaria.saastamoinen@staff.uta.fi