Evaluating the Success of Search Sessions in Interactive Information Retrieval

Magdalena Nikola

University of Milan Bicocca, Milan, Italy
m.nikola@campus.unimib.it

Abstract. Interactive Information Retrieval (IIR) studies include both system evaluations and users' information search behaviors, and the interaction of users with systems and information. The development and testing of appropriate measures and methodologies for evaluating IIR is central to information science. To better understand users' needs and support their interactions with information, IIR systems need to be able to understand the goals underlying users' search behaviors. This work is conceived to address some aspects of this problem. In particular, it considers how people evaluate the success of a complete search session and of the various search intentions within a search session, with respect to the task which motivated the search. In this paper a pilot study is described.

Keywords: Information Retrieval · Interactive Information Retrieval · Work Task · Evaluation · Search Session.

1 Introduction

The main goal of an Information Retrieval System (IRS) is to return to users the most relevant documents in response to their queries, thus respecting the so-called paradigm "one query-one response". However, people usually engage in longer and more complex information seeking episodes. Therefore, when people try to address a new type of problem, they need to engage in many activities other than just clicking on a search result retrieved by the system. In IIR, the crucial point is to develop systems that allow the user to easily access the information s/he needs, while also providing solutions to a series of problems that may arise during a search session. According to Cole [2], the evaluation of a system should focus on how users are able to achieve their goals, how the system helps users to identify and engage in appropriate interactions, and the relationship between the results of these interactions and the progress towards the goals. In order to understand and develop suitable measures for the evaluation of IIR systems, it is necessary to know how people evaluate the system's support for achieving the goals of an intention during a search session, and, in general, how they evaluate the success in achieving the goals of the entire search session. In order to do this, it is necessary to understand what these intentions are, and what the nature of the work tasks is since it has been shown that task's topic has an essential influence on user behavior during a search session [5]. This paper presents a methodology and a pilot study of a project undertaken during my master's thesis.[1]

2 Related Works

Several studies have shown that capturing a user's information need is one of the most critical aspects of IR. Although it is difficult to create an all-including definition of an information need, most information needs can be characterized in terms of tasks and topics: a *task* represents the goal or purpose of the search, this is what a user wants to accomplish by searching, e.g., a user wants to plan a trip; a *topic* represents the subject area that is the focus of the task, e.g., the user might plan a trip to Africa. Research has also shown that information needs evolve during the search process, as these are *dynamic information needs*. This evolution is due to the fact that during a search for information, people learn more about their needs, and consequently their pertinent behaviors change. Li & Belkin in [1] define *tasks* as "activities people attempt to accomplish in order to keep their work or life moving on". More in general, a *work task* is defined as an activity people complete in order to achieve their work's goal, e.g., writing a report, planning a vacation. Moreover, a work task is without a doubt a *motivation* for information search, and includes both *a)* information-seeking tasks and *b)* information-search tasks. With *information-search* is intended information search only through an information system. Instead, with *information-seeking* is intended the fact that users may also seek information from other sources, such as human or printed documents. One important development in IIR evaluation and experimentation has been the simulated work task that describes the situation leading to the information need. The nature of the task that leads a person to engage in the interaction with an IRS in searching for information has been shown to influence the behavior of users during the search sessions.

In recent years, the characteristics of search tasks have been studied, such as how different search tasks could be classified, what they are influenced by, and how they differ according to their attributes. A concrete example is a study conducted by Wildemuth, Freund, and Toms [4] in 2014, in which two attributes of the search tasks are studied and implemented: task complexity and task difficulty. That work provides a "detailed revision of existing practice in developing search tasks to test, observe or control" these two attributes, because as they say "it is not clear if these attributes are mutually exclusive or share some dimensions, as current definitions have tended to blur the distinction" [4].

3 A New Paradigm of User Study for Evaluating IIR

This project for the evaluation of IIR systems aims at investigating the following issues: *a)* given a search session in response to a motivating task, how would

[1] The project was undertaken under the supervision of Prof. Nicholas Belkin at Rutgers University and Prof. Gabriela Pasi at the University of Milano - Bicocca

we evaluate the system support for that search session? *b)* given an intention associated with a query segment, how would we evaluate the system support for that intention? *c)* can we discover measures for evaluating the contribution of each query segment to the success of the search session as a whole?

To address these issues, the main practical goal of the work is to develop a framework for the evaluation of IIR Systems. To do this, the following research questions have to be answered:

1) **RQ1** How do people judge the success of a search session?
2) **RQ2** How useful was each intention/query segment in accomplishing the goal of the search session?

RQ1 concerns the ability to learn how satisfied are users in carrying out the search task, or how successful, according to them, was their search session. Specifically, it wants to investigate the kind of measures that users adopt when they evaluate the whole search session: what is/are appropriate measure(s) for evaluating the system support of the search session? Do different types of motivating tasks require different evaluating measures?

RQ2 aims to learn about the usefulness of each intention of the search session and the usefulness of each query segment of the same search session in accomplishing the goal of the search task. Furthermore, it aims at understanding what are the appropriate measures for evaluating the contribution of each intention/query segment in accomplishing the goal of the search session.

4 Research methodology

In the performed pilot study, users were required to follow a specific procedure, whose steps are summarized in Table 1.

Table 1. Summary procedure.

	Procedure	**Time**
1	Read and sign the consent form	3 min
2	Initial questionnaire	2 min
3	Shown the tutorial about the system	10 min
4	Shown the task and the topic of the search	3 min
5	Second questionnaire	2 min
6	Search, all behaviors are logged	20 min
7	Replay the search, by query segment & annotation of query segments	40 min
8	Search session evaluation and comparison	12 min

As shown in Table 1, prior to conducting their searches, subjects were asked to read and sign a consent form in which each of them was informed about the experiment. Then, searchers completed a brief questionnaire about their demographic characteristics and their normal searching behaviors. Next, searchers were given a video tutorial which was designed to interactively guide them through the workings of the experimental system. In the next step, to the users were shown the tasks and the topics of the search. Before doing their search, subjects were asked to take familiarity with the topic and the motivating task and to anticipate their supposed difficulty in completing the assignment. While doing the search they had the possibility of saving/unsaving pages they considered useful/not useful for accomplishing the task. The search ended when the time required for the search was expired or when users have felt that the task was accomplished. After the search was completed, participants were required to fill a questionnaire, whose focus was understanding their intentions in each query segment and the successes related to them. At the end of the entire searching experience, subjects participated in a structured post-search interview which was designed to elicit confidence, attitudes, strategies, and behaviors directly related to the success or unsuccessful of their search session.

4.1 Study Motivating Tasks

The task type classification framework proposed by Li & Belkin [1] was used to construct two motivating tasks for this study. The specific intention in task construction was to design motivating tasks that differed systematically on several of the facets of the task that were shown to affect search behavior. In particular, two task types were chosen because they have shown, in previous work, to lead to significant differences in search behaviors, including frequency of search intentions. We hypothesize that the understanding of success in the two tasks is different.

The motivating tasks used in the study are based on the following **Task Scenario:** You are about to plan a vacation with your partner to improve your personal relationship between you and him/ her. You want to do this after the end of Spring semester, when you have 18-26 May when you'll both be free, and can book for a week somewhere, including travel time. The considered two tasks to be executed by participants are **Task 1** or **Task 2**, summarized in Table 2 below.

Table 2. Description of the tasks.

No.	Your Task
1	Find at least three resorts in different countries that you think will be good for the purpose of the trip to show your partner. If you book flights now to the general area for that period, you'll be able to afford a nice resort. Be sure that the region safety won't be a problem for the places you find. Please save up to three pages for each resort, that you think will be useful in helping you and your partner to decide which one is best.
2	Please find and save the page(s) for each of the resorts named below, which give you information about the best available weekly rate for two at that resort, during the period March 16-24. Available resorts: Indonesia - The Santai, Malaysia - The Banjaran Hotsprings Retreat, Singapore - Capella Singapore, Thailand - Pimalai Resort & Spa, Vietnam - Fusion Maia Da Nang.

5 Results

Undergraduate students were recruited from Rutgers University to participate in the study. The age of participants ranged between 18 and 21, and the average number of years the participants have been conducting online searches was 11 years. All participants rated themselves as an experienced searcher in using search engines (e.g., Google, Bing). Some of them indicated that they are also experts in searching through social media (e.g., Facebook, Twitter, YouTube), or marked that they are also experts in searching within community-based forums (e.g., Quora, Stack Overflow). However, only one participant rated himself as an experienced searcher in using other search tools, such as a library database. In general, on average, participants were experienced with online information searching, because they usually search for information online for their every-day needs (e.g., homework, studies).

In the first part of the study, it may be said that most of the participants were successful with their intentions: in fact, in most cases, users have managed to complete the intentions of query segments, so these intentions have been marked as successful. During the search, however, there were cases in which the participants failed to positively conclude some intentions, which is why they have been labeled as non-successful intentions. Summarizing, 77% of the intentions chosen during the search sessions were marked as successful, and 15% as non-successful. Moreover, some intentions have not been reported either as successful or as not successful, this number covers 8% of the total intentions. Instead, the reasons for which users have reformulated their queries can be grouped as follows: *a)* the user entered the new query because s/he was able to find the best-rated resort from what TripAdvisor stated, *b)* the user was still trying to find information from each of the websites, *c)* the user wanted to find another review of the resort besides TripAdvisor, *d)* the participant found the top resort in Vietnam and was looking for more information about the resort, *e)* the user was trying to load the website for another resort but it would not load, so s/he moved onto the next resort which also would not load, *f)* the user wanted to obtain details about the best luxury resorts in Malaysia.

It can be said that the most important part of this project was to understand what the users meant by the success of a task, what it means to achieve the goal of the task and positively conclude the search session. For this reason, all users were asked to provide us with their own and personal definition of successful. To the question “What do you mean by successful?”, we have received several answers, which vary from the simplest answer in which the user says that s/he was able to find three websites/resort in three countries, to the most reformulated ones in which the users explain that s/he found what s/he was looking for to the best of his/her ability, or that s/he did not find package pricing, rather the nightly pricing for each of the resorts and their amenities.

6 Discussion of the Results

The most important outcomes from the study are: *a)* even with this small sample, participants made use of almost all the available intentions and they seem to have been sufficient to describe what the participants wanted to accomplish; *b)* the reasons for judgments of success or unsuccess of the different intentions depend on the considered intention, thus indicating that they would require different measures for evaluating the system support. What such measures would be could not be determined, given the small number of participants, but with more data, it seems to be possible to infer categories of different measures; *c)* the reasons for the success of the search session have to do with the accomplishment of the task, which means that any possible measure for evaluating the search session as a whole should be directly related to the type of motivating task. Since there are two task types in the study, with more data it should be possible to identify, based on both the reasons given and the reasons for changes of search strategy, some general evaluation measures for the different tasks; *d)* the descriptions of plans or search strategies and the reasons for changing can clearly be sources for identifying criteria, and possible measures, for evaluation of the search session as a whole.

7 Conclusions and Future Developments

In the field of Interactive Information Retrieval (IIR), the main goal of this work was to understand the reasons why people change their queries, what is successful to them and why, and, more precisely, to understand how people evaluate the success of a search episode. The few data obtained in this pilot study and described in this paper, indicate that we are in a promising direction to arrive at defining standard methods and metrics for the evaluation of IIR systems. In order to validate in a more complete way our hypothesis and results, it will be necessary to wait for the conclsion of the project, and for the global collection of data relative to all the participants expected for this project.

References

1. Li, Yuelin and Belkin, Nicholas J: A faceted approach to conceptualizing tasks in information seeking., vol. 44, pp. 1822–1837. Information Processing & Management (2008).
2. Cole, Michael and Liu, Jingjing and Belkin, Nicholas and Bierig, R and Gwizdka, J and Liu, C and Zhang, J and Zhang, X: Usefulness as the criterion for evaluation of interactive information retrieval., vol. 44, pp. 1–4. Proc. HCIR (2009).
3. Lin, Shinjeng and Xie, Iris: Behavioral changes in transmuting multisession successive searches over the web., vol. 64, pp. 1259–1283. Journal of the American Society for Information Science and Technology (2013).
4. Wildemuth, Barbara and Freund, Luanne and G. Toms, Elaine: Untangling search task complexity and difficulty in the context of interactive information retrieval studies., vol. 44, pp. 1118–1140. Journal of Documentation (2014).
5. Hienert, Daniel and Mitsui, Matthew and Mayr, Philipp and Shah, Chirag and Belkin, Nicholas J: The role of the task topic in web search of different task types., pp. 72–81., Proceedings of the 2018 Conference on Human Information Interaction & Retrieval, ACM (2018).