From Task-based Evaluation to Feature-based Evaluation in Personal Search

Sargol Sadeghi
School of Computer Science &
Information Technology
RMIT University
Melbourne, Australia
seyedeh.sadeghi@rmit.edu.au

Mark Sanderson
School of Computer Science &
Information Technology
RMIT University
Melbourne, Australia
mark.sanderson@rmit.edu.au

Falk Scholer
School of Computer Science &
Information Technology
RMIT University
Melbourne, Australia
falk.scholer@rmit.edu.au

ABSTRACT

Task-based evaluation has been suggested as a solution for comparing search systems in the personal context. However, as personal search tasks are broad, dependent on users, and have different levels of specificity [3], focusing on the building blocks (or characteristics) of these tasks could provide a more reliable and maintainable alternative for evaluation. Moreover, the characteristics can be used to determine to what extent evaluation results are generalizable and comparable across different users and tasks.

In this position paper, a *characteristic reference model* for personal search tasks will be introduced. Based on this model, different search systems can be compared not only in relation to task types, but also in terms of the characteristics that are most influential in search tasks, increasing the level of detail at which comparisons can be made.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Performance, Design, Experimentation, Human Factors.

Keywords

Personal Search, Task-based Evaluation, Task, Search Characteristic.

1. INTRODUCTION

Providing search solutions to retrieve information that has been seen previously is the main focus in the *personal search* context [8]. To compare the effectiveness of search systems in the personal context, identifying common search tasks is of key importance. For example, Kelly and Teevan [3] proposed building a shared collection of common tasks instead of studying tasks in separate research groups. Common tasks for evaluation purposes have also been suggested in other disciplines such as HCI (Human Computer Interaction). For an instance, Whittaker et al. [7] introduced *reference tasks* with the goal of comparing interaction techniques.

However, it is challenging to identify common search tasks, particularly in the personal context, due to the variety of search needs among different users. Controlling the variety of tasks under a set of task *types* was proposed as an approach for evaluating personal search systems by Elseweiler and Ruthven [1]. In this study, three task types were identified based on a search *characteristic* to control the evaluation experiments; and a *task-based* evaluation conducted where the search systems are compared in relation to the search tasks. However, as the task-based evaluation focuses on specific task scenarios, there is a disadvantage that the acquired results cannot be generalized [5]. This is while solving task-based evaluation problems and developing a new type of evaluation has been highlighted [9].

To overcome this problem, we propose to incorporate the underlying *characteristics* of tasks. These characteristics, being more general in nature, can support the identification of commonalities across different tasks in terms of their components. For this purpose, we introduce a characteristic reference model in the next section.

2. CHARACTERISTIC REFERENCE MODEL

With the focus on search characteristics to compare personal search systems, first we must acquire knowledge about the range of characteristics that can affect the retrieval process. Based on these characteristics, we can then identify *similar tasks*, which have common search characteristics. This notion of *explicit* similarity supports a fair comparison of search systems in relation to the user tasks.

However, it is also possible to define *implicit similarity* between tasks. Here, tasks do not necessarily share the same set of characteristics, but their characteristics have been demonstrated to have the same effect on the retrieval process. Consider the following simple example of the implicit similarity concept.

From pilot user studies that we have conducted with the aim of identifying different types of personal search tasks, the user's *level of knowledge* in relation to the target information and task has been observed as a search characteristic influential in retrieval results. Based on this characteristic, we proposed a hierarchy of personal task types for level of knowledge, as shown in Figure 1.

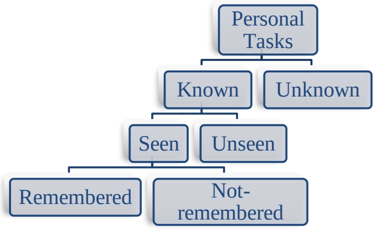

Figure 1. Personal task types and level of knowledge

In the proposed task hierarchy, for example, the user's state of knowledge might be that the target information is *unknown*, where the user does not know whether the required information item exists. Another possibility is that the user is searching for an information item that they know exists and have seen before, but is currently *not-remembered.*

In our observations of users, there are situations where user search behavior for not-remembered tasks is the same as for unknown tasks. For example, one of these situations is when the last access time to the information is prior to last month; here, the user does not know how to get to the information.

In the literature, the time of last access to required information has been called the task *temperature*. For this search characteristic, three values of *hot* (accessed within the last week), *warm* (accessed within the last month), and *cold* (accessed prior to the last month) have been suggested [1]. Based on this observation and from the gathered characteristics and values, it is possible to derive a simple rule as an example of implicit task similarity, illustrated in Figure 2.

```
If:
        Task A= {<Level of knowledge: Unknown>}
        Task B= {<Level of knowledge: Not-remembered>,
                <Temperature: Cold>}
Then:
        Task A similar to Task B.
```

Figure 2: Implicit similar tasks

From Figure 2, it can be seen that if there are two task scenarios identified under two different types (e.g. unknown and not-remembered), in some situations (e.g. cold temperature) they could have a similar effect on the retrieval process. In other words, it is possible that tasks which are in fact highly similar can occur under different task types. Such relationships have not been considered in task-based evaluations, where the focus is on specific task scenarios.

The previous scenario is a simple example; more realistically, it is likely that many different characteristics affect search tasks, in terms of: user, search need, search strategy, search context, information, and the collection of information. Deriving comprehensive rules for task similarities requires extensive user studies in both qualitative and quantitative aspects. We intend to extrapolate a set of rules composed of *Characteristic: Value* settings, as a reference model for identifying similar tasks.

In building this reference model, we need to further explore:

- the key characteristics that are influential in a search task
- interdependencies between characteristics
- the importance of characteristics in affecting retrieval results

Such a model will incorporate the characteristics proposed when studying tasks in different search applications (such as the goal of the user, task complexity, and topic familiarity [2, 4, 6], in both work task and search task aspects), as these are potentially applicable in the personal context. Characteristic settings will be derived by observing real task scenarios and mapping how search characteristics affect search tasks. In this mapping, we consider the interactions of characteristics.

Based on this characteristic reference model, similar tasks can be either created from scratch, or selected from the recorded tasks in current studies where characteristic details are available. Search systems can then be compared in relation to explicitly or implicitly similar tasks. The advantage of using this model is not only limited to enriching the comparability of personal search systems, and the generalizability of comparison results, but it can also lead to a complementary evaluation approach, where assessing the effect of one characteristic on the performance of search systems is important.

3. CONCLUSION

In this paper, we proposed a characteristic reference model for evaluating personal search systems. As there are a variety of tasks in the personal context, this model is based on identifying building blocks, and how they affect search tasks. This approach will enable better control and comparability across different users and tasks, rather than focusing on specific instances of tasks as is currently done in task-based evaluation. Focusing on these characteristics not only facilitates the evaluation of search systems based on search tasks through detailed comparisons, but also provides evaluations on characteristics in affecting the effectiveness of search systems.

4. REFERENCES

[1] D. Elsweiler and I. Ruthven. Towards task-based personal information management evaluations. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 23–30. ACM, 2007.

[2] P. Ingwersen. Selected variables for ir interaction in context: Introduction to irix sigir 2005 workshop. In *Proceedings of the ACM SIGIR 2005 Workshop on Information Retrieval in Context (IRiX)*, pages 6–9. Citeseer, 2005.

[3] D. Kelly and J. Teevan. 11 understanding what works: Evaluating pim tools. *Personal information management*, page 190, 2007.

[4] S. Kim and D. Soergel. Selecting and measuring task characteristics as independent variables. *Proceedings of the American Society for Information Science and Technology*, 42(1):n–a, 2005.

[5] W. Kraaij and W. Post. Task based evaluation of exploratory search systems. In *Proc. of SIGIR 2006 Workshop, Evaluation Exploratory Search Systems, Seattle, USA*, pages 24–27, 2006.

[6] Y. Li and N. J. Belkin. An exploration of the relationships between work task and interactive information search behavior. *JASIST*, 61(9):1771–1789, 2010.

[7] S. Whittaker, L. Terveen, and B. Nardi. Let's stop pushing the envelope and start addressing it: a reference task agenda for hci. *Human-Computer Interaction*, 15(2):75–106, 2000.

[8] D. Elsweiler, D. E. Losada, J. C. Toucedo, and R. T. Fernandez. Seeding simulated queries with user-study data forpersonal search evaluation. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information*, SIGIR '11, pages 25–34. ACM, 2011.

[9] K. J¨arvelin. Ir research: systems, interaction, evaluation and theories. In *ACM SIGIR Forum*, volume 45, pages 17–31. ACM, 2012.