User Perception of Search Task Differentiation

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
This paper examines a new approach to making the evaluation of personal search systems more feasible. Comparability and diverse coverage of personal search tasks are two main issues in evaluating these systems. To address these issues, the proposed approach relies on identifying the differences between search tasks. An experiment was conducted to measure user perceptions of such differences across pairs of typical search tasks, grouped by an underlying feature. A range of features were found to influence user perceptions of task differences. This new knowledge can be used to identify similar and different tasks, which further facilitate comparability and diverse coverage of varied personal tasks for system evaluation.

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

General Terms
Performance, Experimentation, Human Factors.

Keywords
Personal Search Task, Features, Evaluation.

1. INTRODUCTION
One of the most common search needs among individuals is *re-finding* information that has been seen previously [1], which is an example of a personal search task. Although personal search systems have been developed to support individuals when re-finding information, there is great potential for improvement in terms of retrieval accuracy [2]. This improvement requires the evaluation of such systems.

A standard approach in evaluating information retrieval systems relies on the provision of standard or example search tasks to be used for comparing search systems. Personal search tasks vary extensively, and because of privacy constraints, it is difficult to know what users search for and how they do it. These limitations make comparisons of systems difficult across different users. Acquiring the knowledge of what search tasks are common across different users and how they can be compared will facilitate personal search evaluations.

2. RELATED WORK
Studies on personal search evaluations can be categorized under two main groups: naturalistic and simulated. This categorization is based on the employed approach for either selection or creation of search tasks. In naturalistic investigations, although real user tasks are captured, they are specific and limited to a particular group of users [1, 3]. Therefore, system evaluations based on these tasks are difficult to generalize and *compare* across different users. In simulated investigations, search tasks are created in controlled laboratory-based settings to enable generalization [4]; however, it can be questioned how reliable these simulated tasks are in terms of reflecting real search needs, and how well they cover the *variety* of personal information needs. Research on personal search evaluation is hampered by the lack of comparability and adjustability to the diverse range of search tasks. To address these issues, the first step is to be able to differentiate between users' tasks.

In a study by Elsweiler and Ruthven [1], personal search tasks were differentiated based on a single common underlying *feature* of recorded tasks in a diary study. They used this way of grouping tasks to conduct balanced evaluation experiments to compare search systems by examining different tasks. Thus, although personal tasks are varied and dependent on users, it is possible to differentiate tasks by considering their common underlying features without violating the privacy of users. However, there are some questions on how comparable tasks under the same group are, and conversely how varied the tasks under different groups are. In other words, what are the other features that can impact on either the comparability or diversity of tasks?

This motivated us to examine the discriminative power of underlying features in establishing the differentiation of personal search tasks, which we refer to as the *effect* of features. Tasks with the same level of feature effect can be considered to be *similar*, while different feature effects can indicate *different* tasks. Differentiating personal tasks under these two groups, similar and different, can facilitate the comparability and coverage of tasks respectively, to be used for evaluation experiments. To this aim, we conducted an initial experiment to examine the feasibility of feature effects on differentiating tasks. The remainder of this paper is structured as follows. In sections 3.1 and 3.2, the design and methodology of our main experiment are explained, which examines the relative effect of a set of features in enabling users to differentiate search tasks. The results of this experiment are discussed in Section 4. This is followed by conclusions and future work in Section 5.

3. EXAMINING FEATURES IN DIFFERENTIATING TASKS
In this section, we explain the main contribution of this study, which is to examine a set of features and to establish how influential they are in differentiating tasks.

3.1 Design
For examining personal task features, we had to choose a particular form of personal search. We interviewed 45 users one-on-one and in focus groups, including a diverse range of people from students and academic staff, to librarians and employees of a

local company. In response to being asked to list common retrieval tasks, a high proportion of individuals mentioned re-finding. For example, on average 70% of personal search tasks of employees were described as retrieving information that they have seen previously. This dominant value is in line with the past research, where re-finding was identified as the typical retrieval request in the personal context [1]. We therefore chose to focus on re-finding searches in the experiment reported here.

With the broad type of search task selected, next we had to pick features of tasks to study. Many features of tasks have been identified in the literature, for a variety of purposes. We investigated a range of these, based on whether the feature could potentially be influential in differentiating tasks. We note that this experiment is an initial investigation to establish the feasibility of the method, and not intended to include a comprehensive list of all possible task features. Moreover, to limit the range of tasks that participants had to consider, the experiment focused on email re-finding tasks, as previous work has indicated that users frequently need to re-find e-mail messages more than other types of information items [5]. Within this setting, we now describe the features that were considered in our main experiments.

The first feature included in our experiments was the *granularity of information* to be retrieved, which has been used extensively in differentiating re-finding tasks e.g. [5, 6]. Based on this feature, three different task types have been recognized [1]: *lookup* task (looking for specific piece of information), *one-item* task (looking for one message), and *multi-item* task (looking for more than one message). Note that in this paper, we refer to the information to be retrieved as *target information*, which can be a fragment of a message, a single message, or multiple email messages. Other features that have been used for task differentiation including the elapsed time between information accesses (*recency*); and how often the target information is accessed (*frequency*) [1, 2]. Moreover, in exploring the role of memory on e-mail re-finding, some features have been found to be influential in task differences, including whether the user *remembers* the search *topic*, *sender*, or received *date* of the target information [6]. In further examination of email re-finding, message *uncertainty* was defined as a feature. This is the ratio of the number of unique viewed messages to the total number of tried messages [7], which we included in our experiment. Furthermore, we explored whether the *uniqueness* of the message (in terms of topic, or sender) can distinguish re-finding tasks.

We also considered features that have been studied in the context of establishing differences between general (rather than personal) search tasks. The *goal* of the user is an important feature that has been extensively used in proposing different types of general search tasks [8]. Therefore, we evaluated this feature for task differentiation. *Search strategy*, which considers how users get to target information, is another major feature that has been considered in the context of general search. Keyword search and browsing have been highlighted as the two main techniques [9], and we investigated these for their ability to differentiate tasks. We also incorporated *contextual* features such as *search urgency* [10, 11]. The full list of the features discussed in this section is reported in Section 4.

3.2 Method

As one way of examining the discriminative power of task features, we hypothesize that users, who have experienced different search conditions, can provide some evidence about the effect of task features. To test this hypothesis, we conducted experiments where users are asked to compare tasks which differ in one feature. Examples of such paired tasks are given in Table 1. These *feature settings* of paired tasks were designed based on their importance in past studies.

Table 1. Examples of paired tasks to be compared in terms of an underlying feature setting.

Feature	Task A	Task B
Target information granularity	Searching for information WITHIN ONE particular message	Searching for information from MORE THAN ONE message
	Searching for ONE particular message (e.g. to be forwarded or replied)	Searching for information WITHIN ONE particular message
	Searching for information from MORE THAN ONE message	Searching for ONE particular message (e.g. to be forwarded or replied)
Search strategy	Searching for a particular message when you SEARCH BY KEYWORD	Searching for a particular message when you BROWSE on inbox or folders
Remembering the sender of target information	Searching for a particular message when you REMEMBER the sender of the message	Searching for a particular message when you do NOT REMEMBER the sender of the message

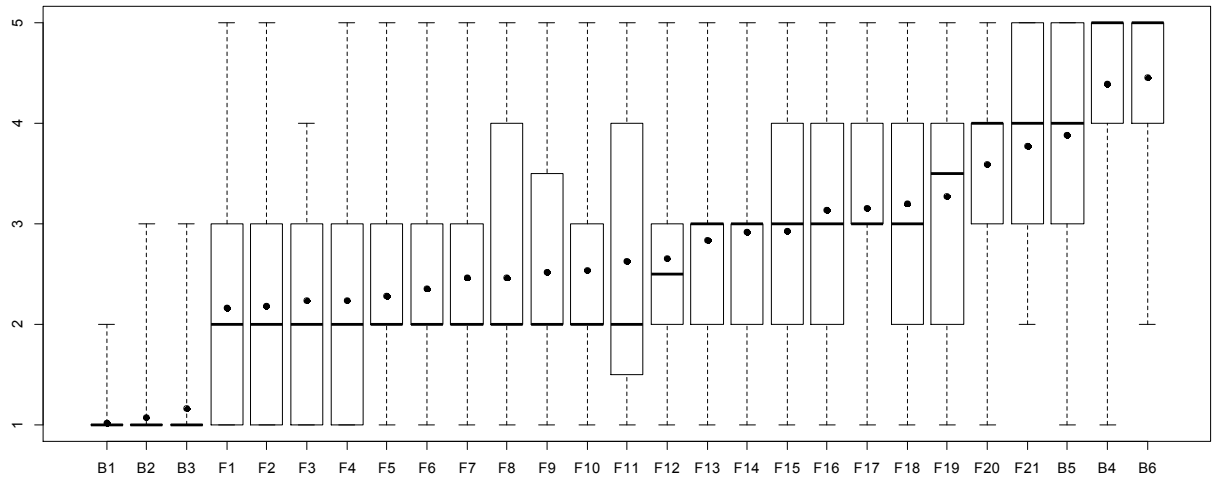
In this experiment, participants were asked to rate the differences between each task pair, answering the question: *"To what extent do you think that the difference between Task A and B will affect the way you search for the information described in the tasks?"*. They were instructed to compare the difference only in terms of the single specified feature. Responses were indicated using a 5-level ordinal scale with the categories *"Not at all"*, *"Slightly"*, *"Moderately"*, *"Very"*, and *"Extremely"*. In total, 21 paired tasks were compared by each participant. The full list of paired tasks is presented in Section 4.

We carried out the experiment using crowd-sourcing via the *CrowdFlower*[1] *platform*. For quality control in crowd-sourcing experiments, it is common to use *gold data*. Gold data are questions with known answers that a diligent participant should be able to answer correctly, if they are paying attention and completing jobs seriously. It has been suggested that a minimum of 5-10% of the total number of jobs should be designed as gold data. In our experiment, we used 12.5%. An example of such data is as follows: *"Task A: Searching for an email message that you received MANY YEARS AGO"*, *"Task B: Searching for an email message that you received NOW"*. In comparing the difference of these two tasks, if a participant was to answer that the tasks are *"Not at all"* or *"Slightly"* different, then the judgments from that participant are less likely to be trusted. Assessment continued until 25 trustable judgments were gathered for each pair of tasks; totally 525 trusted judgments were gathered, which is consistent with the level of user samples employed in previous studies on assessing the effect of factors, for example by Elsweiler et al. [5]. The order in which a pair of tasks was presented to a user was fixed in this preliminary study.

4. Results

Before analyzing gathered data, we need to examine the reliability of data. Examining the distribution of the five possible responses for each pair of tasks, we calculated a Chi square test to assess whether they were significantly different from a flat distribution, which would indicate random selection by participants. The tests

[1] http://crowdflower.com/



F1: Role of the user (receiver vs. sender)
F2: Access recency (week vs. month)
F3: Information repetition (single vs. duplicated)
F4: Temporal search context (urgent vs. not-urgent)
F5: Access recency (day vs. week)
F6: Sender frequency (frequent-sender vs. rare-sender)
F7: Access recency (month vs. day)
F8: Thread of target information (conversation vs. single-message)
F9: Information granularity (multi-item vs. one-item)
F10: Number of viewed messages (certainty vs. uncertainty)
F11: Remembering other recipients (remembered vs. not-remembered)
F12: Information granularity (one-item vs. lookup)
F13: Access frequency (rare vs. frequent)
F14: Information location (body vs. attachment)
F15: Remembering received date (not-remembering vs. remembering)
F16: Information granularity (lookup vs. multi-item)
F17: Search strategy (search vs. browse)
F18: Search goal (forwarding vs. collecting)
F19: Uniqueness of the topic of target information (not-unique vs. unique)
F20: Remembering search topic (not-remembered vs. remembered)
F21: Remembering sender (remembered vs. not-remembered)

Figure 1. Comparisons between mean categories of feature's effect in differentiating tasks (B_i: Bounds, F_i: Feature settings).

for each of the paired tasks indicated that the responses were not randomly selected by participants ($p<0.05$).

Furthermore, we need to establish the significance of responses in identifying task similarities and differences. However, there is no baseline for task differentiations to which responses can be compared. We proposed an approach, where obvious paired tasks were designed as bounds for highly similar or different tasks. The identified bounds make it possible to compare the perception of users for real paired tasks with reference to obviously similar or different bounds. Consider the case where the distribution of responses for a real paired task is significantly different from an obviously *similar* bounding case; from this it can be concluded that the two components in the real paired task cannot be similar, and it is therefore likely that the underlying feature setting of the paired task has an impact on *differences* between tasks. Conversely, if the response distribution of the real paired task is significantly different from an obviously *different* bounding case, then the paired task components cannot be different, and the feature setting of the paired task is likely to be influential in establishing task *similarities*.

Examples of bounds are: *"Searching for an email message that takes you ONE MINUTE to find vs. Searching for an email message that takes you SIXTY SECONDS to find"* for similar tasks; and *"Searching for an email message that takes you ONE SECOND to find vs. Searching for an email message that takes you MANY YEARS to find"* for different tasks. These bounds are similar to gold data in the point that they are designed in an obvious way, to establish user response rates for extreme cases. We then calculated the mean response category for all paired tasks, using the ordinal group numbers 1-5 (corresponding to the response range from *"Not at all"* to *"Extremely")*. The calculated mean categories are illustrated in Figure 1. In this figure, bounds are labeled as B_i, including three cases for obvious similarities and three for obvious differences. As expected, these appear at the extreme ends of the feature settings range (low and high bounds for task differences).

To compare the mean category values for the 21 real paired tasks, in addition to the already established *low* and *high* bounds, a medium bound is required. This is obtained by selecting three real paired tasks whose mean category value is closest to 3, the mid-point of the response scale. Based on the mean category values in Figure 1, the paired tasks with these underlying feature settings of F_{14}, F_{15}, and F_{16} were selected for the *medium* bound.

Through the identified main bounds, we compared other feature settings with two bounds that they lie between, based on their mean category values. We conducted Chi square tests to identify the significance of difference between the responses of each feature setting and the two surrounding bounds. In this comparison, three states can happen: the feature setting is significantly different from a) the left surrounding bound; b) the right surrounding bound; c) both bounds. For example, if the feature setting lies between the low and medium bounds, the following analysis was conducted:

- If the feature setting is significantly different from the low bound but not from the medium bound, then we can conclude that the feature setting is at the *medium* level of difference.

- If the feature setting is significantly different from the medium bound but not from the low bound, then we can conclude that the feature setting is at the *low* level of difference.

- If the feature setting is significantly different from both the low and then medium bounds, then the feature setting is neither at the level of low, or medium, difference. The effect of this type of feature settings can be interpreted as being *moderately low*.

A similar analysis can be applied for paired tasks that lie between the *medium* and *high* bounds; pairs that are significantly different from both the medium and high bounds can be interpreted as *moderately high*.

The Chi square results of comparing feature settings with their surrounding bounds are reported in Table 2. Although most of feature settings appeared to be at the medium level of effect, there were some features with moderately low (e.g. F_3 for *"information repetition"*), and moderately high (e.g. F_{21} for *"remembering sender*") effects. From Table 2, feature settings that have previously been used to identify personal task differences (*"information granularity",* F_9 and F_{12}) appeared at the medium

level of effect. This confirms that Elseweiler and Ruthven's chosen feature of information granularity can be important in differentiating tasks from each other. However, no significant differences were obtained at the high level of effect. In comparison to information granularity *"search goal"* and *"search strategy"* (F_{18}, and F_{17}) achieved greater mean category values of task differences (Figure 1). These features have been highlighted in past research for differentiating tasks [8, 9]. However, they could not reach a significant high effect level (Table 2). A moderately high level of effect was obtained by *"remembering sender*", F_{21}, among the feature settings with a large mean category value. Such features can be employed to identify more diverse tasks. On the other hand, features with a low effect on task differences can facilitate task comparability. As an example, *"access recency (week vs. month)"*, F_2, achieved a moderately low effect, which can be considered for identifying comparable tasks. However, the other feature settings of access recency, *"day vs. week"* and *"month vs. day"* are at the medium level of effect, and therefore less likely to lead to comparable tasks.

Table 2. Significant differences between feature settings and main bounds (F_i: Feature settings from Figure 1).

Feature setting	Difference from low bound	Difference from medium bound	Difference from high bound	Effect level
F1	p<0.0005	p<0.005	-	moderately low
F2	p<0.0005	p<0.005	-	moderately low
F3	p<0.0005	p<0.05	-	moderately low
F4	p<0.0005	p<0.005	-	moderately low
F5	p<0.0005	p= 0.05597	-	medium
F6	p<0.0005	p= 0.08096	-	medium
F7	p<0.0005	p= 0.1509	-	medium
F8	p<0.0005	p= 0.5667	-	medium
F9	p<0.0005	p= 0.09945	-	medium
F10	p<0.0005	p= 0.4243	-	medium
F11	p<0.0005	p<0.05	-	moderately low
F12	p<0.0005	p= 0.3808	-	medium
F13	p<0.0005	p= 0.8961	-	medium
F17	-	p= 0.3083	p<0.0005	medium
F18	-	p= 0.5212	p<0.005	medium
F19	-	p= 0.3218	p<0.0005	medium
F20	-	p= 0.06847	p<0.005	medium
F21	-	p<0.05	p<0.05	moderately high

In further relative comparisons between feature settings, *"access frequency"*, F_{13}, achieved higher effect than *"access recency"* (week vs. month). This suggests that in some settings, how frequently the target information has been accessed can be more important than how recent the access was. Moreover, what users remember about the target information has different effects in differentiating search tasks. For example, *"remembering the sender of an email message"* has moderately high effects, while *"remembering other recipients"*, F_{11}, does not indicate a large difference between tasks. This knowledge about feature settings can reveal signals regarding the search experience of users and the potential improvement for systems. As an example, remembering the sender of target information highly influenced the user's experience. Personal search systems need to improve the experience of users in the absence of remembering such information. These signals can be further explored to develop more effective personal search systems.

Overall, the new knowledge on the effect of feature settings facilitates addressing two main issues, discussed in Section 2, on identifying both diverse and comparable tasks. In future work, we aim to expand this work to identify feature settings with different levels of effects, and to propose more fine-grained bounds for distinctions within settings at the medium level.

5. CONCLUSIONS

This paper presents an initial experiment investigating the influence of features in task differentiation to facilitate diverse coverage and comparability of tasks, which are two main issues in personal search evaluation. In this experiment, a series of search features was examined by testing user perceptions of the similarity or differences between tasks, as differentiated by the specified feature. A wide range of the features was perceived by users to be indicative of different search tasks. Although features that have previously been used to identify task differences were influential, other features achieved higher effects in terms of enabling task differentiation. On the other hand, features with low effect in task differences could potentially be used for identifying comparable tasks. In future work, we plan to investigate factors such as task presentation order. Moreover, we aim to identify features with different levels of effects on user and system performance, to enable better understanding of diverse coverage and comparability of personal search tasks.

6. REFERENCES

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

[2] Elsweiler, D., Losada, D. E., Toucedo, J. C. and Fernández, R. T. Seeding Simulated Queries with User-study Data for Personal Search Evaluation. *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR* 2011, 25-34.

[3] Dumais, S., Cutrell, E., Cadiz, J., Jancke, G., Sarin, R. and Robbins, D. Stuff I've seen: a system for personal information retrieval and re-use. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval* 2003, 72-79.

[4] Kim, J. and Croft, W. B. Retrieval experiments using pseudo-desktop collections. *Proceeding of the 18th ACM conference on Information and knowledge management* 2009, 1297-1306.

[5] Elsweiler, D., Baillie, M. and Ruthven, I. What makes re-finding information difficult? a study of email re-finding. *Advances in Information Retrieval* 2011, 568-579.

[6] Elsweiler, D., Baillie, M. and Ruthven, I. Exploring memory in email refinding. *ACM Transactions on Information Systems (TOIS)*, 26, 4 2008), 21.

[7] Elsweiler, D., Harvey, M. and Hacker, M. Understanding re-finding behavior in naturalistic email interaction logs 2011), 35-44.

[8] Broder, A. A taxonomy of web search. *ACM Sigir forum*, 362002), 3-10.

[9] Teevan, J., Alvarado, C., Ackerman, M. S. and Karger, D. R. The perfect search engine is not enough: a study of orienteering behavior in directed search. *Proceedings of the SIGCHI conference on Human factors in computing systems* 2004), 415-422.

[10] Li, Y. *Relationships among work tasks, search tasks, and interactive information searching behavior*. Rutgers University-Graduate School-New Brunswick, 2008.

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