Explicit Feedback in Local Search Tasks

Dmitry Lagun[1], Avneesh Sud[2], Ryen W. White[3], Peter Bailey[2], Georg Buscher[2]
[1] Emory University, [2] Microsoft Bing, [3] Microsoft Research
dlagun@mathcs.emory.edu, {avnsud, ryenw, pbailey, georgbu}@microsoft.com

ABSTRACT
Modern search engines make extensive use of people's contextual information to finesse result rankings. Using a searcher's location provides an especially strong signal for adjusting results for certain classes of queries where people may have clear preference for local results, without explicitly specifying the location in the query directly. However, if the location estimate is inaccurate or searchers want to obtain many results from a particular location, they have limited control on the location focus in the search results returned. In this paper we describe a user study that examines the effect of offering searchers more control over how local preferences are gathered and used. We studied providing users with functionality to offer explicit relevance feedback (ERF) adjacent to results automatically identified as location-dependent (i.e., more from *this location*). They can use this functionality to indicate whether they are interested in a particular search result and desire more results from that result's location. We compared the ERF system against a baseline (NoERF) that used the same underlying mechanisms to retrieve and rank results, but did not offer ERF support. User performance was assessed across 12 experimental participants over 12 location-sensitive topics, in a fully counter-balanced design. We found that participants interacted with ERF frequently, and there were signs that ERF has the potential to improve success rates and lead to more efficient searching for location-sensitive search tasks than NoERF.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *relevance feedback*, *search process.*

Keywords
Explicit relevance feedback, Location-sensitive search.

1. INTRODUCTION
Location is an important facet of the relevance experience for a significant fraction of the search queries that modern Web search engines receive [1][9]. Search providers have constructed entity-specific local search experiences, analogous to the data contained in Yellow Pages directories (yp.com). Users may engage specifically with these results, whether within the main result listing or in a separate search vertical experience. One class of location-sensitive query is local queries, where users seek information pertaining to their current location. Search engines can support this via location awareness, allowing them to estimate the user's current location using reverse IP lookups, GPS sensors, or explicit user settings, and tailor the results accordingly. However, these estimates may be inaccurate (depending on factors such as Web proxies and internet providers), and search engines typically only add a few local results to the top 10 to mitigate the risks associated with personalization. If the searcher wants more results from a particular location of interest, they have had limited ability to signal to a search engine their interest in location-specific results beyond adding location terms to their queries (e.g., [wedding venues, *tucson az*]). Query refinement can be a cumbersome process (especially in a mobile setting where typing is difficult [3], but physical location may be particularly pertinent), and does not provide adequate control over search engine operation. In this paper we report on an investigation of an alternative method. We make the following contributions with this work:

- We define a mechanism for allowing users to provide explicit relevance feedback (ERF) reflecting an individual searcher's *location preferences*. Unlike previous work which has focused on applying relevance feedback at the term level [5][6], we focus on feedback at the geographic location level, where users explicitly request more results from a particular location.
- We describe a user study to compare a search engine with location awareness and ERF support, allowing searchers to quickly indicate their location preferences, against a comparator system with location awareness but no ERF support.
- We show that location-sensitive ERF has the potential to improve success rates, leading to more efficient searching, and that it was perceived positively by participants. This indicates that, although ERF is generally dismissed as being cumbersome [1], there are scenarios such as local search where it can be useful and users respond well to its inclusion.

In the remainder of the paper we describe relevant related work, the study that we performed and its findings, and discuss implications.

2. RELATED WORK
We focus on related work discussing the impact of offering relevance feedback (RF) (explicit or implicit) on searcher performance as well as describing location-aware search systems.

Relevance feedback is the primary post-query method for automatically improving system representations of a searcher's information need and has been studied extensively [6][8]. Explicit relevance feedback allows users to select documents or terms to be used for query expansion. Koenemann and Belkin [5] developed and studied search interfaces that varied the transparency with which users were permitted to see the terms selected for RF. A "penetrable" system variant allowed users to select the terms used for query expansion.

Despite its potential, ERF is rarely used because it requires direct interaction, placing additional cognitive load on users [4]. Explanations can be offered to help people understand the benefit of providing feedback [7]. *Implicit relevance feedback* (IRF) does not require user intervention in the feedback process – the search system automatically infers searcher preferences from user behavior such as clicks on results or query refinements. White et al. [10] investigated user and task effects for explicit and implicit relevance feedback mechanisms and found that IRF is generally preferred by participants with less search experience, while more experienced searchers preferred ERF. These studies have focused on the provision of feedback about the topical relevance of document content. However, as our understanding of the role of context improves, and as search engines exhibit more context awareness and greater ability to interpret and use context, it is useful to capture ERF on different levels and for different signals. In this paper, we target the collection and application of ERF related to location preferences.

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

Searcher location can be captured implicitly, from Internet provider or GPS information, or explicitly, from search queries containing location-related search terms. A significant volume of search queries have location intent, and such information can be used for a range of applications, from appropriately adapting spelling corrections to promoting search results associated with entities or businesses located near the user. Wang et al. [9] introduced an algorithm for detecting a query's most dominant location, which can be used to re-rank search results by placing results from the detected dominant location higher. Bennett et al. [1] proposed a machine-learned method for modeling and inferring query location to subsequently personalize search results for the user. For example, the intent behind the query [msg] issued in New York City may well be Madison Square Garden rather than monosodium glutamate, the dominant search intent for this query globally. Despite several additional experiments evaluating the utility of using query or user location information to improve search performance, there has been no work on helping searchers better refine their searches to target particular locations. We address that challenge in this research.

To summarize, our paper examines the use of ERF at the location level in the context of local search tasks. As we discover, ERF can improve search performance along several dimensions.

3. EXPERIMENT DETAILS

In this section we describe the experiment that we performed to understand the effect of ERF in location-sensitive tasks. The description includes the research questions that we answer, the study design, data capture, tasks, and two systems (ERF and NoERF).

3.1 Research Questions

The goal of the study was to address the following questions:

RQ1: If ERF is available, do people engage with it?
RQ2: Does search performance improve with the use of ERF?
RQ3: How do people perceive ERF?
RQ4: How do searcher and task factors affect ERF performance?

The first question addresses the availability of ERF on result pages (since it was only available for some local results). The other questions cover ERF performance and participants' perceptions of it.

3.2 Study Design

To answer the research questions we conducted a controlled user study in which participants were asked to attempt 12 pre-defined exploratory search tasks in a laboratory setting. The tasks are described in the next section. The independent variable in our study was the presence or absence of ERF functionality. Before starting to use the ERF-enabled version of the search system, each participant was presented with a short interactive tutorial introducing the ERF functionality, explaining how to use it, and describing its effect on the engine. Presence or absence of ERF functionality was counterbalanced among participants as either the first or second *block* of 6 of 12 tasks each participant carried out, and vice versa.

The study design was balanced across participants. To minimize task-related effects, tasks shown in the first and second block were fully balanced across all participants, i.e., each task appeared in the first block for half of the participants and the second block for the remaining half. Also, we controlled the position of a task within a block so that each task appeared in all positions 1 to 6 equally often.

3.3 Surveys and Data Capture

We employed questionnaires at several points during the course of the study. At the beginning of an experiment, before starting the first task, we asked participants about search, Internet usage familiarity, as well as their usual strategies for dealing with a dissatisfactory search results page when doing Web search. After the completion of each task, we asked the participants to complete a short questionnaire that inquired about their performance in attempting the task as well as the helpfulness of the search engine on multi-point scales. After finishing all tasks, participants were asked to provide their overall impression of the systems in an exit survey.

For each task, we recorded the following user interaction data: a time-stamped list of URLs visited with the browser, the full HTML contents including screenshots of the visited pages, clicks on any hyperlinks, and mouse cursor positions which let us reconstruct hover events over specific areas of interest.

3.4 Task Description and Examples

We selected 12 tasks with specific local intent, derived from query logs of the Microsoft Bing search engine. All tasks are about discovering specific types of places or services in a given city/region, e.g., airport shuttles, vineyards, courses etc. Of the 12 tasks, 9 were chosen with city-level intent, and 3 with state-level intent. To avoid local memory effects, we selected different locations for each task, all separate from the location where the user study was performed. The task description specified the location in which the participants should imagine themselves being; examples are shown in Fig. 1.

To ensure all subjects had the same entry point, the first query and first search results page were fixed for each task. The task location was not specified in the first query, but provided as part of the task description; in addition, the search engine settings were set to simulate as if the query were issued from the specified location (e.g., as though a reverse IP lookup was performed or location set explicitly). The first results page always contained at least one correct result from the local area, but never contained enough correct results to complete the task so that participants had to continue searching to successfully finish. In the ERF version of the search engine, results from the local area were decorated with an icon that could be used to provide relevance feedback (upper left of Fig. 2).

City-level Task: "You are temporarily living in Boston, Massachusetts (MA) and would like to learn or improve French as a foreign language. Find at least 4 Web sites that offer French courses in Boston (no online courses)."
Initial Query = [French courses]. User Location=Boston, MA.

State-Level Task: "You are on a road trip in Virginia State (VA) with some wine loving friends. Find 4 vineyards in Virginia and note down 2 wines from each vineyard."
Initial Query = [vineyard], User Location=VA.

Figure 1. Example city- and state-level tasks used in the study.

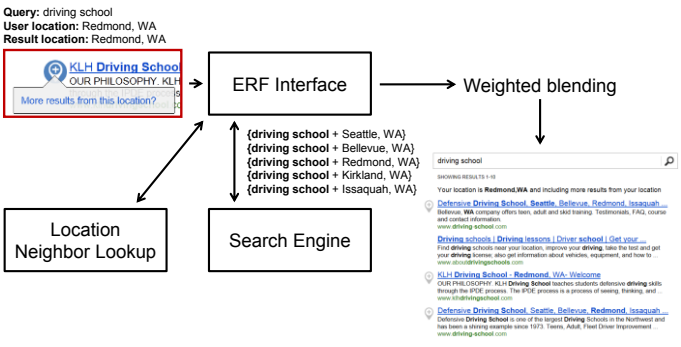

Figure 2. Provision and use of feedback in ERF system. The feedback icon and explanatory callout are shown in upper left.

As searchers attempted tasks they saved documents that they believed would answer the question. To calculate success rates on study completion, participants' answers were assigned correctness labels by experimenters, with discussion to resolve disagreements.

3.5 Systems

The ERF system consists of a simplified search interface to the Microsoft Bing search engine. The system offered modifications to highlight, capture, and add local results using ERF, allowing users to request more results from a specific location. The search results page presented on the top 10 retrieved results, with vertical answers, advertisements, and other page elements removed. Pagination was permitted to explore the results more deeply. If a caption contained a term specifying the location of the place/service covered in a result, such terms were highlighted (using bold-face font) in the same way as query terms, even if they were not in the query directly. NoERF used the same backend retrieval elements, but lacked the ERF functionality, i.e., ERF icons were not shown.

An overview of the ERF system is provided in Fig. 2. If the user requested "more results from this location" by clicking on a link next to a result of interest then the ERF system would fetch more local results by issuing one or more additional queries containing location strings, and these results would be injected into the results listing using weighted blending (CombSUM [2]). Locations were connected to each result by automatically examining its caption (title and snippet) for town and city names (e.g., in Fig. 2 we have "KLH Driving School – *Redmond, WA* – Welcome"). Any local results—inferred either from location terms in the result caption, or because they were injected from one of the additional location-specific queries—were decorated with an ERF icon. Often a local area can encompass multiple cities. To address this we pre-computed a location neighbor map, and issued multiple queries to the search engine with the neighboring locations. The weights for each additional query were adjusted based on proximity to the location of the selected result, and the number of ERF clicks for a given location.

3.6 Participants

Twelve participants were recruited from a user study pool external to Microsoft. Participants had a wide variety of backgrounds and professions. Most used computers, the Web, and Web search several times a day. Participants were generally pleased with the way that their current preferred search engine worked (M=2.18 on 1-5 scale, 1=best). They reported rarely paginating through results or selecting related searches, but reported frequently refining queries.

4. RESULTS

In this section we analyze the findings of the study. We examine a number of aspects of how ERF was used, how its performance compared to the non-ERF system (NoERF), and participants' perceptions of ERF. We also include some additional analysis of factors affecting the performance of the ERF system, specifically individual and task differences. Paired t-tests and Wilcoxon signed-rank tests are used where appropriate, with the significance level (α) set to 0.05. We begin by examining the availability and participant engagement with ERF during the tasks studied.

4.1 Availability and Engagement with ERF

We first wanted to understand the frequency with which ERF was made available to participants and quantify the extent of their engagement with it. At least one explicit feedback icon was available on 30% of the SERPs viewed on the ERF system. As described earlier, the option to provide ERF was only shown for particular results, limiting its coverage. There were on average 1.46 clicks (SD=1.36) on a feedback icon per task and 78% of the tasks on the ERF system had at least one feedback click. Overall, the findings show that ERF was available to participants fairly often and that they engaged with it frequently.

4.2 ERF Task Performance

We now analyze the performance of the ERF system compared to NoERF. We focus our analysis on two aspects: (i) *task outcomes* (levels of success in completing tasks correctly), and (ii) *task efficiency* (time taken and actions performed during searching). We begin our analysis by focusing on task outcomes.

4.2.1 Task Outcomes

Task success rate on each of the two systems was calculated on a per-task basis and reflected the fraction of the required number of correct answers for each task that participants found. Answer correctness was determined by human judgments, as described earlier. To provide estimated answer volume on each system, we also computed the average number of answers found (including those exceeding the number required for the task), and the fraction of all found answers that were correct. Table 1 summarizes the findings.

Table 1. Measures of task outcomes (averaged per task).

Measure	*ERF*	*NoERF*	*Significance*
Task success rate	88.33%	83.66%	t(71)=1.57, p=0.12
Num. answers found	4.41	4.18	t(71)=1.74, p=0.09
Answer success rate	87.81%	84.83%	t(71)=1.23, p=0.22

The table shows ERF performs slightly better than NoERF: ERF has a higher task success rate, ERF helped participants find more answers, and a greater fraction of the answers found on ERF were correct. Although the differences are not significant at α=0.05 the findings are consistent and the p-values are fairly small.

4.2.2 Task Efficiency

In addition to studying task outcomes, we also examined the *process* by which people attempted the tasks. We computed a number of temporal and action-based features to better understand the efficiency with which users found information on ERF and NoERF. Table 2 presents the results. Statistical significance at α=0.05 is denoted by star (*). Task termination was determined by participants.

Table 2. Measures of task efficiency (averaged per task).

Measure	*ERF*	*NoERF*	*Significance*
Time to first answer	37.3s	45.8s	t(71)=1.75, p=0.08
Time to task finish	146.9s	157.3s	t(71)=1.16, p=0.25
Time to required ans.	120.2s	141.6s	t(71)=3.01, p=0.001*
Num. unique queries	2.03	2.64	t(71)=3.13, p=0.003*
Query length (words)	2.60	3.11	t(71)=2.68, p=0.009*
Num. paginations	0.32	1.21	t(71)=3.57, p<0.001*

The findings summarized in Table 2 show that there are fewer queries, shorter queries, and less pagination through the search results on ERF. Although there may be additional clicks associated with the use of ERF, the reduction of query length and quantity is important as entering queries directly is challenging in mobile settings [3], an important emerging area where local ERF could be helpful.

The findings suggest that in many key aspects of the search process, ERF helped searchers be more efficient. However, if we consider the time to first answer and the total task time, the times are not significantly lower for ERF than NoERF (p-values=0.08-0.25). One explanation for this, other than the small sample size, is that in ERF participants found more answers than were actually required for the task (on average 16% extra for ERF versus 6% extra for NoERF). While this improved coverage—and the answers found were still more accurate (see last row of Table 1)—processing the

additional answers in ERF also took participants longer. If we focus only on time to at most the number of answers required for the task, we can see from the third row of Table 2 that ERF is much faster.

4.3 Perceptions of ERF

We were also interested in participants' perceptions of ERF. After each task, a survey asked the following about the system used:

Q1: How good was the search engine for this task? (Scale: *very good*, *good*, *neither*, *bad*, *very bad*) (*SearchEngineGood*).

Q2: I think there was better information available (that the system did not help me find) (*agree* to *disagree*) (*BetterInfoAvailable*).

Q3: I believe that I have succeeded in my performance of this task (*agree* to *disagree*) (*PerceivedSuccess*).

All ratings had a five-point answer scale. Q2 and Q3 were presented as Likert scales, with options spanning *agree* to *disagree*. Table 3 shows the average answer ratings for each question.

Table 3. Participant perceptions of the system used. Ratings range from 1 to 5. Higher is more positive (Q1) and more agreement with the attitude statement (Q2 and Q3).

Measure	*ERF*	*NoERF*	*Significance*
SearchEngineGood	4.04	3.67	$Z = -2.33$, p=0.009*
BetterInfoAvailable	3.03	3.21	$Z = -0.84$, p=0.201
PerceivedSuccess	4.38	4.39	$Z = -0.10$, p=0.464

Table 3 shows that when ERF is available in the search engine, participants perceived that search engine more positively. There were also some indications that they believed that the engine had helped them find more complete information (*BetterInfoAvailable*), although not significantly. Interestingly, although participants appeared slightly more successful on ERF (Table 1), they reported task success as highly similar on each system (4.38 vs. 4.39).

In the exit survey we asked participants about whether they would like ERF to be included in their favorite search engine. Nine of the 12 respondents (75%) reported positively. Those who liked ERF cited its utility (e.g., "makes search easier", "very useful") and that it removed the need to specify the location (e.g., "saves time when dealing with location specific queries"). Those who disliked it felt that it should not be needed ("the engine should know my location and provide results according to my location if my search is location specific"). No participants mentioned any additional burden from providing explicit feedback. Overall, participants perceived ERF favorably and desired it in search systems.

4.4 Factors Affecting ERF Performance

Finally, we wanted to understand the extent to which individual and task factors impacted ERF performance. We computed the task success rate (introduced in Section 4.1) for each participant and for each task. We then identified participants and tasks where each of the systems performed best and dropped cases where the performance of ERF and NoERF was equal. For each participant group, we computed the average search frequency based on their responses to the entry questionnaire (on a scale of one to four, four being most frequent), and for each task group, we computed the average number of required answers (ranging from one to five). We chose these variables since they reflected important aspects of the search process that could affect search outcomes, but were also independent of both system and task effects.

Table 4 shows that less expert users derive more value from ERF (yet not significantly; p=0.07), and that ERF performs significantly better for tasks where more answers are required. Less frequent searchers may be less expert and benefit from additional support. We also showed that ERF was helpful for tasks with more required answers, likely because it helped participants locate similar results.

Table 4. Effect of search frequency and number of required answerers on task success on ERF and NoERF.

Measure	*ERF best*	*NoERF best*	*Significance*
Search frequency	3.50	4.00	$Z = -1.51$, p=0.065
Num. answers needed	4.67	3.00	t(9)=3.06, p=0.014*

5. DISCUSSION AND CONCLUSION

Location is important in a substantial fraction of search queries and an increasing volume of search traffic originate from smartphones and other mobile devices. Therefore providing more efficient methods for users to both type less and get more relevant local results will continue to grow in importance in the coming years. The user study we performed was limited in terms of the number of participants and carefully contrived to focus on location-sensitive tasks that required multiple answers. Participants were also asked to search in multiple unfamiliar locations. Such tasks do occur, for example with vacation planning, but may not be fully reflective of common usage patterns. Note that even though ERF was available for only 30% of SERPs on the ERF system, our findings show that participants interacted with ERF frequently, and that ERF has the potential to improve success rates and result in more efficient searching. Participants perceived the engine more favorably when ERF was included, and most participants wanted to see it added to search engines. There were also weak indications that ERF performed better for less frequent searchers and a significant signal that it performed best for tasks with more required answers.

Our findings are promising, but more work is needed to validate them and improve the ERF technology. Future directions include experiments with larger numbers of participants; improving the relevance quality of results when ERF is provided by the searcher; and improving the discoverability and explanation of ERF.

REFERENCES

[1] Bennett, P., Radlinski, F., White, R., and Yilmaz, E. (2011). Inferring and using location metadata to personalize web search. *SIGIR*, 134−144.

[2] Fox, E. and Shaw, J. (1994). Combination of multiple searches. *NIST Special Publication SP*, 243−243.

[3] Henze, N., Rukzio, E. and Boll, S. (2012). Observational and experimental investigation of typing behavior using virtual keyboards on mobile devices. *SIGCHI*, 2659−2668.

[4] Jansen, B., Spink, A., and Saracevic, T. (2000). Real life, real users, and real needs: A study and analysis of user queries on the web. *IP&M*, 36(2): 207−227.

[5] Koenemann, J. and Belkin, N.J. (1996). A case for interaction: A study of interactive information retrieval behavior and effectiveness. *SIGCHI*, 205−212.

[6] Rocchio, J.J. (1971). Relevance feedback in information retrieval. *The SMART Retrieval System: Experiments in Automatic Document Processing*, 313−323.

[7] Ruthven, I. (2002). On the use of explanations as mediating device for relevance feedback. *ECDL*, 338−345.

[8] Salton, G. and Buckley, C. (1997). Improving retrieval performance by relevance feedback. *Readings in IR*, vol. 24.

[9] Wang, L., Wang, C., Xie, X., Forman, J., Lu, Y., Ma, W.-Y., and Li, Y. (2005). Detecting dominant locations from search queries. *SIGIR*, 424−431.

[10] White, R., Ruthven, I., and Jose, J.M. (2005). A study of factors affecting the utility of implicit relevance feedback. *SIGIR*, 35−42.