Using Terminological Feedback for Web Search Refinement - A Log-based Study

Peter Anick
Overture
peter.anick@overture.com

ABSTRACT

Although interactive query reformulation has been actively studied in the laboratory, little is known about the actual behavior of web searchers who are offered terminological feedback along with their search results. We analyze log sessions for two groups of users interacting with variants of the AltaVista search engine – a baseline group given no terminological feedback and a feedback group to whom twelve refinement terms are offered along with the search results. We examine uptake, refinement effectiveness, conditions of use, and refinement type preferences. Although our measure of overall session "success" shows no difference between outcomes for the two groups, we find evidence that a subset of those users presented with terminological feedback do make effective use of it on a continuing basis.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Query Formulation.

General Terms

Measurement, Design, Experimentation, Human Factors

Keywords

Interactive query refinement, terminological feedback, web search, evaluation, user logs

1. INTRODUCTION

In order to assist the user in bridging the gap between an internal information need and an expression of that need in the language of the target documents, many IR researchers have proposed the use of terminological feedback mechanisms to offer search term suggestions. One of the most widely studied is the use of interactive relevance feedback, in which term suggestions are generated based on user relevance judgements of previously retrieved documents ([Harman88], [Koenemann96]) While this approach has been shown to be effective in improving the recall and precision of subsequent searches, it has been difficult to implement in practice because of the reluctance of users to make the prerequisite document relevance judgements ([Beaulieu97], [Belkin99]). One workaround for this problem has been to generate search suggestions from the top ranked documents regardless of their actual relevance, using linguistic and other heuristics to select and order the terms displayed back to the user ([Xu96], [Bruza97],[Anick99b]).

Although a number of such systems have shown promise in the laboratory setting, their incorporation into large scale web search engines has been slow, and for good reason. Space is already at a premium on the typical textually cluttered search results page. Furthermore, while the majority of end-users are familiar with the way search engines currently work, they don't really understand how or why they work and may find additional interface features clumsy or confusing. For example, a word-cluster based refinement tool offered on the AltaVista web site several years ago received scant user attention and was eventually scrapped.

Recently, several web search engines have begun offering short lists of search refinement suggestions in order to encourage the interactive narrowing of query result sets. Among these is the AltaVista *Prisma*™ tool, which allows a user to augment or replace the current query expression by clicking on feedback terms derived dynamically from an analysis of the top ranked search results. The incorporation of such refinement tools into full-scale web search engines provides IR researchers an opportunity to study user behavior both in the large and in its natural state, free of the potential confounding influences of a laboratory setting or artificial tasks.

In this paper, we will report on the results of several log-based studies of user interaction with AltaVista's Prisma assisted search tool. We were interested in the degree and nature of user uptake of the feature, as well as the conditions and effectiveness of its use within information seeking sessions. An experiment controlling the display of the interface allowed us to compare how behavior differed between users provided with the feature and those operating without it.

SIGIR'03, July 28–August 1, 2003, Toronto, Canada.

The paper is organized as follows. We begin with a brief overview of the functionality provided within the Prisma refinement interface. We then enumerate a range of issues which our study was intended to address, relating our investigation to previous work on each topic. Next we lay out the methodology we used to conduct experiments and to extract session information from activity logs. We then present and discuss our findings. Finally, we offer our conclusions and suggestions for future research.

2. PRISMA

AltaVista's Prisma refinement tool is embedded directly into the standard AltaVista search results page and is designed to be fully compatible with the user's direct manipulation of query expressions within the search box. Figure 1 illustrates how the feedback terms are displayed. In the example, the user has searched on the term "elections". Twelve feedback terms are presented just below the search box in four columns of three terms each. Terms are ordered (left to right by column) so that phrases which contain a query term are displayed first, followed by other multi-word phrases, and finally single word terms. Terms are alphabetically sorted within each category. Each term is underlined, indicating a clickable link, and followed by a pair of clickable chevrons. Clicking directly on a term adds the term to the current query (as shown in Figure 2) and immediately launches a new search for the combined (ANDed) terms. Clicking on the associated chevrons deletes the existing search expression and replaces it with the corresponding feedback term alone. Thus, for any term, the user has the option of using it as a refiner or as a completely new search. A terse one-line "explanation" of these options lies above the term listing.

Feedback terms are generated using a pseudo-relevance feedback approach ([Xu96],[Vélez97]). The number of feedback terms displayed is limited to twelve both to conserve space on the page and to minimize cognitive load by keeping the list short enough to eyeball quickly.

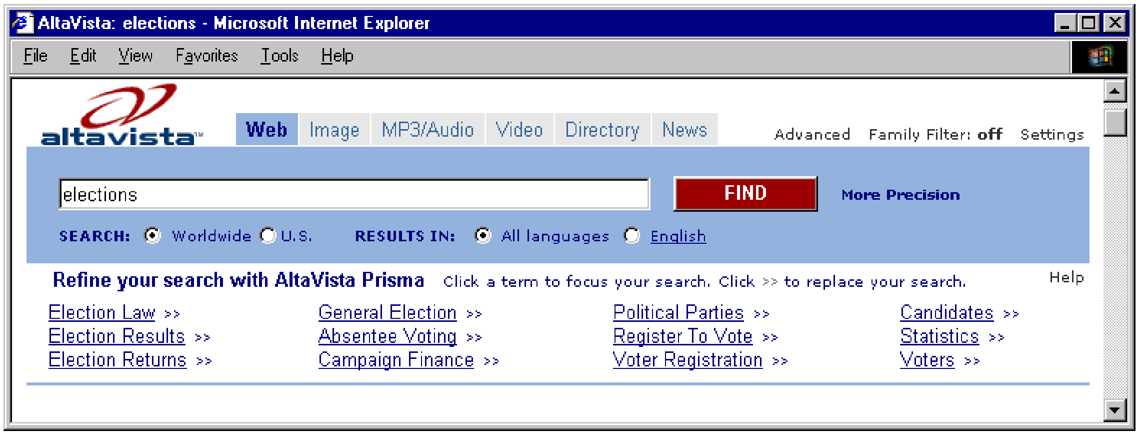
3. QUESTIONS ADDRESSED

Our study of interactive terminological feedback for web search was designed to address a number of practical questions. These included user uptake, reformulation effectiveness, term presentation order, and refinement types. In this section, we introduce each of these topics along with some of the related prior research findings.

3.1 Uptake

Will people left to their own devices adopt an interactive tool

Figure 1. Top of AltaVista search results page, showing feedback terms for the search "elections".


AltaVista: elections - Microsoft Internet Explorer

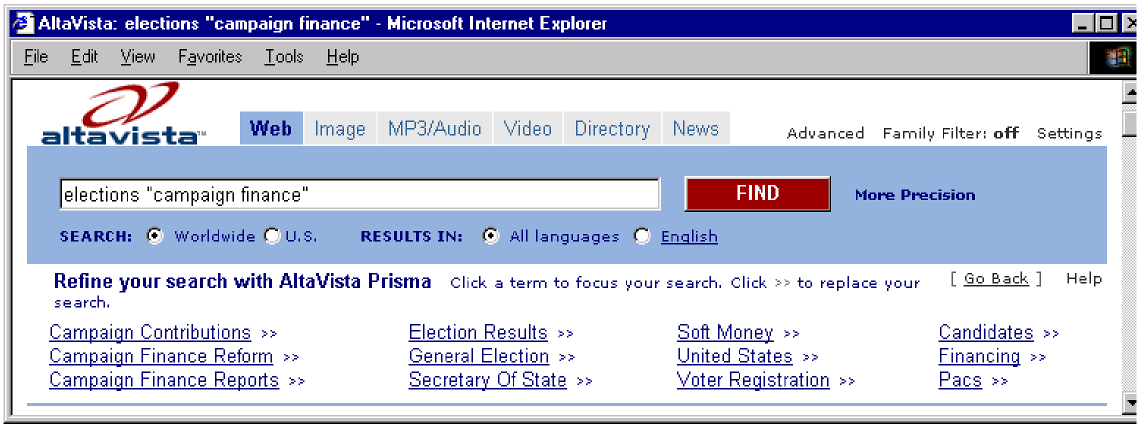
elections

SEARCH: ● Worldwide ○ U.S. RESULTS IN: ● All languages ○ English

Refine your search with AltaVista Prisma Click a term to focus your search. Click >> to replace your search.

Election Law >>	General Election >>	Political Parties >>	Candidates >>
Election Results >>	Absentee Voting >>	Register To Vote >>	Statistics >>
Election Returns >>	Campaign Finance >>	Voter Registration >>	Voters >>

Figure 2. Top of AltaVista search results page, showing the search box and feedback terms presented after the user has clicked on the feedback term "campaign finance".


AltaVista: elections "campaign finance" - Microsoft Internet Explorer

elections "campaign finance"

SEARCH: ● Worldwide ○ U.S. RESULTS IN: ● All languages ○ English

Refine your search with AltaVista Prisma Click a term to focus your search. Click >> to replace your search.

Campaign Contributions >>	Election Results >>	Soft Money >>	Candidates >>
Campaign Finance Reform >>	General Election >>	United States >>	Financing >>
Campaign Finance Reports >>	Secretary Of State >>	Voter Registration >>	Pacs >>

based on terminological feedback? Beaulieu et al's ([Beaulieu97]) studies revealed a reluctance for users to take advantage of term relevance feedback, suggesting that the additional task of judging feedback terms is itself a difficult one which users will avoid. Using a dual-task technique to measure cognitive load, Bruza et al ([Bruza00]) showed that improved result set relevance from using terminological feedback does indeed come at the cost of increased cognitive load in evaluating the feedback. The remedy adopted in Prisma of only showing twelve terms attempts to diminish the cognitive load by reducing the number of terms to peruse. However, the downside is that many potentially relevant terms cannot be displayed, reducing the chance of an intersection with the user's specific information needs.

Ominously, usability tests conducted at AltaVista prior to external release of the tool indicated that many users did not even notice feedback terms when embedded within an already textually full results page. Those that did notice them often misinterpreted them as directories or advertising. Hence, the first concerns of our log analysis were to quantify adoption, understand under what circumstances users were most likely to avail themselves of the tool, and to measure whether one instance of use implies subsequent reuse.

3.2 Effectiveness

Apart from the difficulty in getting searchers to make use of feedback interfaces, most research has shown that term relevance feedback does improve the relevance of results. The Okapi ENQUIRE Project [Beaulieu97] and Efthimiadis [Efthimiadis00] conducted detailed user-centric evaluations of searchers using feedback terms and found that system recall and user satisfaction with results generally improved when users engaged in interactive query expansion incorporating many simultaneous term selections. It should be noted, however, that these studies were done using "best-match" search engines which supported a ranked search over ORed query terms, and that they were carried out in a library environment where increased recall is usually a primary objective. As many have noted, searchers on the web tend to be more concerned with precision than recall, typically scanning only the top ten ranked results for appropriate documents. ([deLima99], [Spink02]). For precision-driven searching, phrase-based feedback systems like Hyperindex [Bruza97] and Paraphrase [Anick99b] have been shown to be effective, as they leverage the power of highly specific noun phrases to match equally specific documents. They still require users to scan a lot of terms, but the user need only choose one phrase at a time to narrow the search.

To date, the majority of user studies of terminological feedback have been conducted with very small sample sizes (with both a small number of users and small number of queries) and in a laboratory setting. Typically, users received instruction on system behavior and they performed their searches aware that they were being observed. In many cases, for good experimental reasons, the actual search topics and task guidelines were provided by the researchers, not the users themselves. Thus, while these experiments provide valuable scientific data, they are not necessarily accurate representations of user behavior in the "real world".

By analyzing a large quantity of log data from thousands of web users engaged in their own information seeking tasks and who had no explicit training in the use of feedback, we hoped to provide a new set of data points to complement prior user-centric research on the effectiveness of such assisted search techniques.

3.3 Feedback term types and ordering

As part of this study, we wished to probe the nature of the terms users preferred for refinement, as well as the effects of order of presentation. Choosing which terms to present and in what order is one of the biggest challenges for terminological feedback systems ([Efthimiadis95, McArthur00]). For any given query, the number of possible underlying information needs, and hence the number of terms that might elicit documents relevant to those needs, is potentially very large. In one experiment, Efthimiadis ([Efthimiadis00]) found that one third of the terms derived from document relevance feedback were identified by users as useful for refining their searches. This amounted to 18 terms on the average, although there was considerable variability from one query to the next. For some queries, the number of useful feedback terms was as high as 98 (out of 113 presented).

Efthimiadis ([Efthimiadis00]) and others ([Bruza97],[Joho02]) have also examined the functional relationships between user-selected feedback terms and queries. Again, through the use of log analysis, we hoped to compare the functional distribution of user selections made during user-initiated web searches to the distributions found in previous laboratory studies.

4. METHODOLOGY

This study was based solely on an analysis of anonymous user activity logs from the AltaVista search site. Timestamped logs recorded all edits to the search box and all clicks on the results pages, including selections of documents and feedback terms. The ordinal position of selected feedback terms within the Prisma tool was also recorded. Cookie ids and timestamps were used to sort records temporally for each user. Known or suspected bots and queries coming from AltaVista internal addresses were discarded from the study.

In order to evaluate user behavior within the context of multiple query information-seeking sessions, we needed a way to cluster temporally contiguous sequences of actions for each user. We used the following operational definition of a user "session": A session begins with a query and continues until there are 60 minutes of no recorded site activity. At this point, a session may be further extended if the user selects a document on a results page. However, if the first action after a 60 minute break is to enter a query (via typing into the search box), then we interpret this as initiating a new session.

Our observations of the semantic content of logged queries over time (from which we can guess whether subsequent queries are logically part of the same information seeking session), suggest that this simple automated heuristic can be quite effective. Most temporally contiguous searches tend to be about the same general information need. In addition, it is not uncommon for users to spend long periods of time (even 20 to 40 minutes) reading retrieved documents and following web links before taking a subsequent related action on a search engine.

In order to evaluate the effects of offering terminological feedback on user behavior, it was necessary to establish a baseline of user behavior in the absence of such feedback. To this end, we divided users into two experimental groups as follows. Starting at a fixed time, any cookied user arriving at the AltaVista site was assigned to one of three groups: (1) feedback, (2) no feedback, (3) non-participating. The probability of assignment into groups 1 and 2 was 5%. The remaining 90% were assigned to group 3. The assignment of arrivals to the non-participating group 3 made sure that participation in our test groups reflected the typical mix of users coming to the site. For example, very frequent users would not be disproportionately represented in the test groups by being given multiple opportunities to be selected. Assignment into groups continued for several weeks and was then terminated. At that point, the set of participants in each group remained fixed. Log records for a contiguous five day period were gathered for the two groups. The records for the feedback group consisted of 15,133 sessions representing 8006 users; those for the no-feedback group consisted of 14,595 sessions for 7857 users. About 250 users with sessions containing over 20 queries were identified as outliers and bots and removed from the study.

With sequences of user actions divided into sessions, we can extract a wide variety of statistics about user behavior. However, because the users remain anonymous throughout the experiment, there is no way to solicit feedback directly from them about the effectiveness of their search efforts. As a surrogate for a more trustworthy measure of search success, we will interpret a click on a result as an indicator of relevance to the user's information goal. This will provide us with several heuristic metrics of "success". We can ask

(1) Was a particular search action followed by a result click?

(2) Did a search session conclude with a result click?

One other experimental variable in this study was the presentation order of the feedback terms. This experiment was run independently of the twin group study outlined above and will be described later in the paper.

5. RESULTS AND DISCUSSION

5.1 Uptake

The test group, for whom feedback terms were displayed, engaged in 15,133 sessions during the five day test period. On average, 14% of a day's sessions contained at least one use of a feedback term. Given that many web searches are satisfied on the basis of a single query and don't require any refinement, we also computed the overall number of sessions in which some refinement step was made, as an indication of how often refinement steps are called for at all. Typical refinement steps would be editing an existing query or replacing it with a completely different query. On average, 56% of sessions involved some degree of refinement. Within this subset of sessions, we find that uptake of feedback increases to 25%. This suggests that a fair number of the users who engage in refining their own search expressions are at least aware that feedback is available. Explicitly measuring uptake by user id, we found that some 16% of users applied feedback at least once on any given day.

5.2 Reuse

Another way to probe uptake is to examine the reuse of a feature over time. As mentioned earlier, usability tests indicated that users may be slow to recognize that feedback terms are even being offered by the interface. Therefore, we decided to measure the likelihood that a user, having used Prisma in the course of one session, made use of it again in subsequent sessions. To this end, we captured log data for the feedback group over a two-week interval and divided it as before into sessions. For each user that used Prisma at all during that time period, we extracted all sessions that occurred subsequent to the first use. This gave us a set of 8,157 sessions restricted to users known to have used feedback in the near past.

Of the 2,318 users falling into this subset, 47% used Prisma again within the two-week window. The percentage of refined sessions in which they employed feedback increased from 25% to 38%. These results are suggestive that usage does increase somewhat with familiarity, although a much longer longitudinal study would be required to verify this.

5.3 Effectiveness

Table 1 compares the baseline (no feedback) group with the feedback group on a number of criteria. The values shown are averages across each of the five days from which the log data was gathered. T-tests were run over the mean values for each day of the evaluation; the probability scores are reported in the last column, when appropriate. No difference was found between the two groups on the percent of sessions ending in a click, our primary heuristic measure of search satisfaction. This suggests that, overall, users achieved the same level of search success regardless of whether feedback was offered. However, the feedback group was slightly more likely to do at least one refinement within a session and less likely to click on a result immediately after the initial query.

These latter differences may be accounted for in part by the fact that 6% of initial queries were followed directly by a feedback refinement in the feedback group's sessions. This may have had the effect of lengthening sessions (via adding a search move) for some sessions in which a relevant result may have already been available near the top of the initial result list.

Table 1. Comparison of baseline and feedback group search behaviors.

	Baseline group	Feedback group	t-test p
% sessions containing refinement	53	56	.01
% sessions ending in a click	77	76	.6
% non initial queries	62	65	.001
% initial queries followed by click	65	62	.01
% non-initial (excluding feedback-based) queries followed by click	56	52	.001
% feedback selections followed by click	<NA>	59	
% initial queries followed by feedback selection	<NA>	6	

Feedback term selections constituted 12.5% of all search refinements. The percentage of feedback selections followed directly by a click was slightly higher than for other manual query refinements, suggesting that those users who availed themselves of the feedback were using it effectively.

5.4 Refinement types and conditions

In this section, we look into the conditions under which feedback terms are selected by users and categorize the selections that users make.

5.4.1 Conditions of usage

Figure 3 breaks down all the instances of feedback use into the four contexts in which they can occur. As the chart indicates, terminological feedback was most often used directly after the user had engaged in some other reformulation step (38%). 28% of feedback refinements occurred directly after the initial query, 21% after a result click. The relatively low percentage of refinements directly following a previous feedback term selection (13%) suggests that there is less need for further refinement at this point.

The "new search" option, in which clicking on the ">>" symbol alongside a term launches a search for the feedback term alone, was used in fewer than 2% of all Prisma interactions. This low uptake was most likely due to a combination of three factors:

- the nature of the interface, in which this feature is much less prominent than the term link itself
- the fact that users are more likely to need to narrow an existing search than to launch a new (even if related) search.
- the likelihood that users don't have a full mental model of the tool and its two options

We'll return to this issue below when we consider the functional relationships between the refinement terms users choose and the queries they are refining.

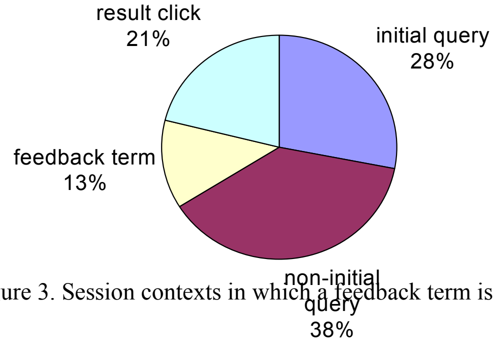


Figure 3. Session contexts in which a feedback term is selected.

5.4.2 Term category preferences

As mentioned earlier, the Prisma interface divides up feedback terms into three superficial categories: phrases containing a query term, other multi-word phrases, and single word terms. By default, the interface presents terms in category order on the assumption that phrases containing query terms would be the most useful, followed by phrases and finally single word terms. In order to test this hypothesis, however, we needed to remove the confounding effect of the presentation order. We therefore conducted a short experiment to compare user term type selection under two presentation conditions – (1) the category ordering described above and (2) a purely alphabetical sort, which should effectively randomize the positions of terms from various categories.

The results are shown in Figure 4. On the left are the average percentages of each term category that were actually displayed, based on an analysis of the categories of terms returned for a random sampling of 100 queries in the log. To the right are the distributions of actual user selections when feedback terms were presented in the two different sort orders. What is clear is that users prefer to choose phrases containing query terms. Relative to the number of terms actually displayed within each category, users choose considerably more container phrases, a roughly proportional number of phrases, and a smaller than proportional number of single word terms.

The effect of presentation order is evident in Figure 4 as well. When container phrases and other phrases are presented before single word terms, their likelihood of selection is increased. This influence of presentation order on term selection is even more obvious in Figure 5, which plots the number of user selections for each term position when terms are presented in alphabetical and category sorted orders. Selection is more or less evenly distributed for the alphabetical sort, whereas the category ordered sort reveals a large primacy effect with a subsequent tapering off. Note the slight increases at positions 4 and 10. These positions are both in the top row of the display area, suggesting that some users are scanning terms from left to right row by row, rather than column by column.

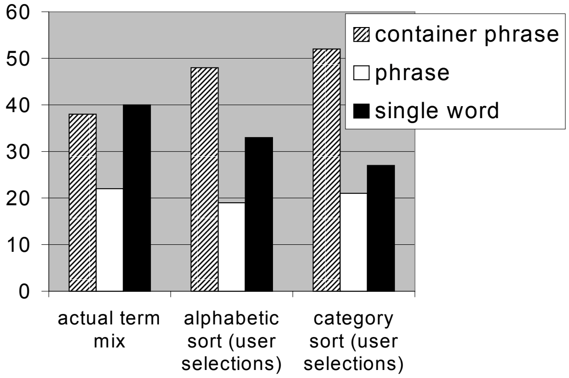


Figure 4. Percent of user selections of terms, broken down by term category types

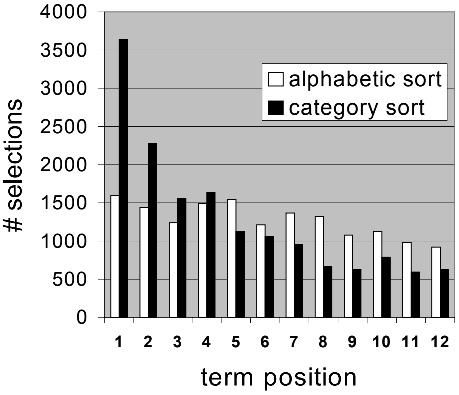


Figure 5. Number of user selections of feedback terms by position for alphabetic and category sorted presentation orders.

5.4.3 Functional classification of refinements

In order to better understand the nature of the refinements that users were making using the feedback tool, we extracted a random sample of 100 refinements and classified them by hand into eleven categories:

Head – a phrase which adds a linguistic head term to a term in the original query

Modifier – a phrase which adds a linguistic modifier to a term in the original query

Elaboration – a term which restricts or adds further context to the concept expressed in the original query

Location – a term which adds location information to the query

Alternative – another way of expressing the same concept as in the original query

Hyponym – a term which is more specific than some term in the query, but which is not expressed linguistically as a phrase containing the query term

Morphological variant – a term which is a morphological variant, such as a plural or adjectival form, of a query term

Syntactic variant – a phrasal rearrangement of the query terms

Acronym – an expansion of an acronym in the query (or vice versa)

Spelling – a spelling correction or variant

Change – a new topic from the original query

Examples of these categories are displayed in Table 2.

Table 2. Examples of functional categories of query refinements.

Refinement category	Examples (query / refinement)
Head	triassic / triassic period
Modifier	buckets wholesale / plastic buckets
Elaboration	jackson pollack / museum of modern art
Location	vietnam / ho chi minh city
Alternative	job listings / job openings
Hyponym	birds of prey / falcons
Morphological variant	norse myth / norse myths
Syntactic variant	map of sudan / sudan map
Acronym	usa maps / united states of america
Spelling	stationary catalog / stationery
Change	skateboards / mountainboards

Table 3 shows the percentage of each refinement type within the 100-query sample we analyzed. By far, the largest categories are head, modifier, and elaboration. Both head and modifier are phrasal refinements that further specify a term in the original query. The attraction of users to such refinements is consistent with findings from Bruza ([Bruza97]), Gutwin ([Gutwin98]) and Anick ([Anick99a]). Elaborations, while not containing a query term, also serve to restrict the query to a more narrowly prescribed set. Locations can be considered a form of elaboration. In total, these four refinement types account for 68% of the user term selections. Given that the system ANDs refinement terms with the original query, one would expect such refinements to narrow the set of results in a manner consistent with the user's presumed intentions. ANDing is less appropriate, however, for some of the other refinement types, such as spelling and change. Narrowing results to documents containing both a correct and incorrect spelling of a term is likely to be much more restrictive than desired. Similarly, if the intent of the user is to change the search to an altogether new topic, then ANDing with the previous search terms is also overly constraining. These refinement types account for 6% of the total. The remaining refinement types fall in a gray area. Documents that contain alternatives, hyponyms, morphological and syntactic variants in addition to the query terms may often be good matches for the user's search topic, although the process of ANDing the terms will inevitably eliminate some relevant documents from the result set.[1] These latter cases amount to 26% of the mix.

Table 3. Percent frequency of use of feedback terms for refinement and new search, broken down by functional refinement type.

Refinement type	**Percent frequency of use as refinement**	**Percent frequency of use for new search**
Head	15	12
Modifier	25	10
Elaboration	24	27
Location	4	5
Alternative	12	30
Hyponym	5	0
Morphological variant	3	2
Syntactic variant	4	2
Acronym	2	0
Spelling	4	0
Change	2	12

In theory, the savvy user could utilize the "new search" option (clicking on the ">>" symbol located after each term to launch a feedback term as a new search) when a term's use as a refiner would be overly restrictive. It is instructive to observe whether users who do click on the "new search" option are utilizing it as intended. That is, do their choices reflect a conscious deliberation between launching a new search and refining an existing one?

We selected a random sample of 40 feedback-based "new searches" from the logs and analyzed them according to the refinement categories described above. Results are shown in the third column of Table 3. The higher relative percentages found for *alternative* and *change* refinements are consistent with the intended interpretation of the "new search" option, suggesting that some users, at least, have formed a proper mental model. Similarly, most of the head and modifier cases in the sample subsume the original query, so that performing a new search is logically equivalent to adding the phrase as a refinement (modulo some effects on ranking due to inclusion of the initial query terms as separate items in the refinement case.) However, nearly all of the elaboration and location refinements we identified in the sample were likely to produce inferior result sets when issued as new searches, since the users dropped valuable context terms that were present in the initial query. Several examples are shown in Table 4.

Table 4. Examples of the inappropriate use of the "new search" feedback option.

Initial query	**New query**
famous hispanics "hispanic people"	biographies
airlines	fares
borderline personality	treatment

It seems reasonable to assume in such cases that the user intended to use the feedback term as a refinement, not as a new search. Whether these represent "typos" on the part of the user versus a real misunderstanding of the UI options requires further study. However, the low uptake of the "new search" feature strongly suggests that many, if not most users are unaware of the distinction the UI is attempting to offer.

6. CONCLUSIONS AND FUTURE RESEARCH

We have employed a fine-grained analysis of users' session logs to investigate their web search behavior when offered a simple form of interactive query reformulation using terminological feedback. Using document clicks as a heuristic measure of reformulation effectiveness, we found that feedback-based refinements, *when applied*, were as effective as the average manual reformulation at locating relevant documents. Nevertheless, the vast majority of reformulations were still done manually. This may have been for many reasons, such as users' habits, impatience with scanning term lists, ignorance of the tool, or simply that it is difficult to capture all the possible refinement needs of a diverse set of users in twelve terms. Laboratory tests or questionnaires will be needed to answer such questions.

For those users who did avail themselves of term suggestions, the distribution of the items selected was consistent with a need for greater search precision, preferring phrases that modified terms in the original query. The evidence of repeated uptake among this population of users was encouraging.

The baselines established in this study will allow us to evaluate the effects of further modifications to the interface, both to the presentation and to the content of the feedback. Would showing more/fewer feedback terms help or hurt uptake? Would the "new search" option be more widely (and correctly) used if it were

[1] This is less of an issue for web search. Since web search engines often include matches on anchor text and meta keywords, there are many opportunities for variants of concepts to be associated with a page, in addition to the content itself.

displayed more prominently? Should the underlying feedback ranking formula be altered to display relatively more container phrases? Would ORing of feedback terms be more appropriate in some cases or would it add unnecessary complexity? These are some of the questions raised by our study, questions that reflect the on-going conflict between functionality and ease-of-use that has challenged designers of assisted search systems from the beginning.

7. REFERENCES

[Anick99a] Anick, Peter G. Automatic Construction of Faceted Terminological Feedback for Context-Based Information Retrieval. PhD. Dissertation, Brandeis, 1999.

[Anick99b] Anick, Peter and Suresh Tipirneni, The Paraphrase Search Assistant: Terminological Feedback for Iterative Information Seeking. *Proceedings of SIGIR'99*,153-159, 1999.

[Beaulieu97a] Beaulieu, Micheline, Thien Do, Alex Payne and Susan Jones, "ENQUIRE Okapi Project*", British Library Research and Innovation Report 17*, Jan. 1997.

[Belkin99] Belkin, Nicholas J., Colleen Cool, J. Head, J. Jeng, Diane Kelly, Shin-jeng Lin, L. Lobash, Soyeon Park, Pamela A. Savage-Knepshield, C. Sikora.: Relevance Feedback versus Local Context Analysis as Term Suggestion Devices: Rutgers' TREC-8 Interactive Track Experience. *TREC*, 1999.

[Bruza00] Bruza, P.D., Dennis, S. and McArthur, R. Interactive internet search: keyword directory and query reformulation mechanisms compared. *Proceedings of SIGIR'2000*, pp. 280-287, 2000.

[Bruza97] Bruza, P. D. and S. Dennis, "Query Reformulation on the Internet: Empirical Data and the Hyperindex Search Engine". *Proceedings of RIAO'97*, pp. 500-509, 1997.

[deLima99] de Lima, Erika F. and Jan O. Pedersen. Phrase Recognition and Expansion for Short, Precision-biased Queries based on a Query Log. *Proceedings of SIGIR'99*, pp. 145-152, 1999.

[Efthimiadis00] Efthimiadis, Efthimis N. Interactive query expansion: A user-based evaluation in a relevance feedback environment. *JASIS* 51(11): 989-1003, 2000.

[Efthimiadis95] Efthimiadis, Efthimis N. User Choices: A New Yardstick for the Evaluation of Ranking Algorithms for Interactive Query Expansion. *Information Processing & Managment, 31* (4), 605-620, 1995.

[Gutwin98] Gutwin, C., Paynter, G., Witten, I.H., Nevill-Manning, C., and Frank, E. Improving Browsing in Digital Libraries with Keyphrase Indexes. Technical Report 98-1, Computer Science Department, University of Saskatchewan, 1998.

[Harman88] Harman, Donna, "Towards Interactive Query Expansion". *Proceedings of SIGIR'88*, pp. 321-331, 1988.

[Joho02] Hideo Joho, Claire Coverson, Mark Sanderson, Micheline Hancock-Beaulieu: Hierarchical presentation of expansion terms. *SAC 2002*: 645-649

[Koenemann96] Koenemann, Jürgen and Nicholas J. Belkin, "A Case for Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness", *Proceedings of CHI96*, 1996.

[McArthur00] McArthur, R. and Bruza, P.D. The Ranking of Query Refinements of Interactive Web-based Retrieval. *Proceedings of the Information Doors Workshop (held in conjunction with the ACM Hypertext and Digital Libraries Conferences), 2000.*

[Spink02] Spink, Amanda, B. J. Jansen, D. Wolfram & T. Saracevic. From E-Sex to E-Commerce: Web Search Changes. *IEEE Computer*, 35(3), 107-109, 2002.

[Vélez97] Vélez, Bienvenido, Ron Weiss, Mark A. Sheldon, and David K. Gifford. Fast and Effective Query Refinement. *Proceedings of SIGIR'97*, pp. 6-15, 1997.

[Xu96] Xu, Jinxi and W. B. Croft. Query Expansion Using Local and Global Document Analysis. *Proceedings of SIGIR'96*, pp. 4-11, 1996.