Effectiveness of Real-Time Query Expansion in a Library Catalog

Xi Niu
University of North Carolina at Chapel Hill
216 Lenoir Dr., Chapel Hill, NC 27599
xiniu@email.unc.edu

Bradley M. Hemminger
University of North Carolina at Chapel Hill
216 Lenoir Dr., Chapel Hill, NC 27599
bmh@ils.unc.edu

ABSTRACT
Query expansion (QE) is a potentially effective technique to help searchers formulate or improve their search. In January 2010, the University of North Carolina at Chapel Hill (UNC) Library made an effort to improve user interactions with authority data in its online catalog. It introduced a real-time query expansion feature for the catalog interface that would provide real-time suggestions based on what the user had typed in the search box. QE is not a new technique, and has been widely used in web search interfaces. It has rarely been adopted by library catalogs though. From the libraries perspective, populating the query expansion feature with authority data provides a balance between traditional library use of authority records and the expectations users have from their use of Web search engines. By analyzing the search logs from the library server, we investigate the impact of this feature on users’ search behavior. Two sampling periods of logs were chosen: April 2009 (before the feature’s launch) and April 2010 (after the feature’s launch). Totally, we have analyzed approximately 200M log data with approximately 800,000 useful log records. Results showed that users take the suggestions 20.26% of all searches. There was a significant increase in the rate of subject searches. The library catalog interface supporting dynamic query expansion also leads to longer queries, less query iterations per search, and fewer result items viewed per search. In addition, search failure rate has seen a great drop. However, results also imply some issues may exist with the use of query expansion, necessitating proper handling of the search expansion interface. Our findings have implications for how query expansions should be offered to library catalogs to increase the use of library’s collection.

Keywords
real-time query expansion, library catalog, search behavior, query formulation, query reformulation, log analysis

INTRODUCTION
In the summer of 2008, the University of North Carolina at Chapel Hill (UNC) switched from a traditional library catalog interface to a new interface, supported by Endeca1. This interface incorporates faceted search using the existing MARC (MAchine-Readable Cataloging) records and metadata in the library collections. The Endeca platform has been used in a number of e-commerce websites and is seen as an improvement over the searching capabilities of the traditional search interfaces. Faceted search incorporated in library catalog interfaces allows users to click the metadata to refine the search or browse the whole collection. While the overall feedback on the new interface was positive, one of the most common criticisms the catalog implementation team received from both patrons and staff concerned the interface’s lack of support for authority searching and browsing (Pennell & Sexton, 2010). Without such support, users may have difficulty in choosing suitable search terms to represent their subject interests. Inadequate subject access was one of the reasons why many items in large academic libraries were hardly, if ever, checked out (Knutson, 1991).

Before implementing the faceted interface, UNC’s traditional catalogs, supported, to some degree, authority browsing. The authority browsing functionality took a user’s query and returned an alphabetical list of adjacent authority headings based on a “begin with” match. It was not always intuitive for users because it forced users to follow certain rules when entering query terms. For instance, the searchers needed to enter the last name first and first name last if they wanted to find items by an author; for another instance, they needed to strip the initial article before the title, i.e., using “bible” instead of “the bible”. The traditional catalogs also needed users to have some knowledge about the Library of Congress Subject Heading (LCSH) structure in order to perform successful subject searches. Usability studies, as well as searchers’ queries gleaned from log data, revealed that many searchers did not adhere to these conventions and knew nothing about

---

1 Endeca is a software company that provides the faceted search platform.
The "anywhere" was chosen. About the person Charles Dickens that would be included if only items written by Charles Dickens in the search box, the user limits the results to. By selecting the field "author" and typing "Charles title, journal title, author, and subject. For example, a user may be seeking all records authored by Charles Dickens. By selecting the field "author" and typing "Charles Dickens" in the search box, the user limits the results to only items written by Charles Dickens, excluding the works about the person Charles Dickens that would be included if "anywhere" was chosen.

The real-time automatic query suggestion feature presents relevant authority-controlled suggestions from MARC records to searchers. These suggestions are presented as a list, containing 15 most relevant ones, very shortly (in milliseconds) after the searcher finish typing the first three letters of a word, and update after each additional letter is typed. Searchers may either select a suggested query or ignore the suggestions, and completing their query. This approach integrates query suggestions into query formulation, giving help at a stage in the search when it could positively affect query quality most likely. That is why this feature is called “real-time”.

The query suggestions are populated with the selected MARC fields: 100 for authors, 245 for titles, and 6** for subjects. Subject headings are only loaded in their complete forms, rather than as individual subcomponents. By providing the MARC suggestions to searchers, the library catalog promotes the authority data access, without users having to understand the LCSH and MARC metadata.

To some degree, the real-time query expansion feature is complicated by the search field option. If a user preselects the search field before entering a query in the search box, the user limits the query suggestions to only that type. For instance, if a user chooses “subject” as the search field and as (s)he types the query in the search box, all the suggestions (s)he receives are subject heading suggestions. If the user does not preselect the search field and keeps the default setting as “anywhere”, the query suggestions would be of all the four possible types. Figure 1 illustrates an example with “bra” in the search box and “anywhere” in the search field, the available suggestions are of author type, keyword (generic) type, and subject type, as indicated by the label on the right side of the suggestions (no label for keyword type).

The result is a real-time automatic query expansion feature that presents relevant authority-controlled data of author, subject, title, and generic suggestions to users as they are typing queries. The auto-suggest feature is a query expansion technique, which has been an established technique in information retrieval (IR) field. In January 2010, the UNC Library implemented this real-time expansion feature on its catalog with the hope of providing a balance between the authority data access and the expectations that users have gained from searching the Web.

To determine the effectiveness of the real-time query expansion feature on the UNC library catalog, we have performed periodic log analysis before and after the feature’s launch in January 2010. Usage of query expansion and the impact it has on users’ search behavior are reported in this paper.

**METHOD**

**UNC Library Catalog with Real-Time Query Expansion**

UNC Library catalog (http://search.lib.unc.edu) is a representative of the Online Public Access Catalogs (OPACs) in leading research universities with large collection and comprehensive resources. It implemented Endeca faceted search on top of its traditional catalog systems in July 2008 to enhance the search capabilities. It launched the real-time suggestion feature in January 2010.

Like most library catalogs, the UNC catalog interface allows users to specify the search field against which the query searches. Available fields are anywhere (keyword), title, journal title, author, and subject. For example, a user may be seeking all records authored by Charles Dickens. By selecting the field “author” and typing “Charles Dickens” in the search box, the user limits the results to only items written by Charles Dickens, excluding the works about the person Charles Dickens that would be included if “anywhere” was chosen.

The real-time automatic query suggestion feature presents relevant authority-controlled suggestions from MARC records to searchers. These suggestions are presented as a list, containing 15 most relevant ones, very shortly (in milliseconds) after the searcher finish typing the first three letters of a word, and update after each additional letter is typed. Searchers may either select a suggested query or ignore the suggestions, and completing their query. This approach integrates query suggestions into query formulation, giving help at a stage in the search when it could positively affect query quality most likely. That is why this feature is called “real-time”.

The query suggestions are populated with the selected MARC fields: 100 for authors, 245 for titles, and 6** for subjects. Subject headings are only loaded in their complete forms, rather than as individual subcomponents. By providing the MARC suggestions to searchers, the library catalog promotes the authority data access, without users having to understand the LCSH and MARC metadata.

To some degree, the real-time query expansion feature is complicated by the search field option. If a user preselects the search field before entering a query in the search box, the user limits the query suggestions to only that type. For instance, if a user chooses “subject” as the search field and as (s)he types the query in the search box, all the suggestions (s)he receives are subject heading suggestions. If the user does not preselect the search field and keeps the default setting as “anywhere”, the query suggestions would be of all the four possible types. Figure 1 illustrates an example with “bra” in the search box and “anywhere” in the search field, the available suggestions are of author type, keyword (generic) type, and subject type, as indicated by the label on the right side of the suggestions (no label for keyword type).

**Table 1. Description of the logs**

<table>
<thead>
<tr>
<th>Log Dataset</th>
<th>Time Frame</th>
<th>Size</th>
<th>Available Fields</th>
</tr>
</thead>
<tbody>
<tr>
<td>April 2009</td>
<td>30 days</td>
<td>90M raw data</td>
<td>IP address /data/time/URL/reference URL/user agent</td>
</tr>
<tr>
<td></td>
<td>4/1/2009—</td>
<td>378,454 useful records</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4/30/2009</td>
<td></td>
<td></td>
</tr>
<tr>
<td>April 2010</td>
<td>30 days</td>
<td>109.3M raw data</td>
<td>IP address /data/time/URL/reference URL/user agent</td>
</tr>
<tr>
<td></td>
<td>4/1/2010—</td>
<td>412,483 useful records</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4/30/2010</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 1. UNC Library interface with real-time query expansion**
month in two consecutive years was chosen: April 2009 and April 2010. We have analyzed a total of approximately 200M log data with 790,937 useful log records. Description of the two datasets is summarized in Table 1. The data were extracted and processed using Perl scripts and MySQL database. Details about analyzing the data have been elaborated elsewhere (Niu, 2009; Lown, 2008).

RESULTS

Usage of Query Expansion

In April 2010, users submitted a total of 204,115 queries (approximately 6804 queries per day) to the UNC catalog. Out of these queries, 46164 were the real-time query suggestions, accounting for 22.6% (See “Take Suggestion” in Figure 2). Users’ self-created queries accounted for 77.4% (see “No Suggestion” in Figure 2). If we further break down the suggestions, 88.5% of them were without a search field pre-selected beforehand; the other 11.5% were with a particular search field pre-specified. We examined the overlap between the queries searchers entered manually and the suggested queries to see how many duplicates occurred. That is, how many times a manually entered query matched exactly a suggested query. Although there is no way to ascertain that searchers did not naturally type the duplicate queries on their own, these results suggest that some searchers took suggestions by manually entering the queries rather than clicking on them. Thus, using a simple click metric to measure uptake may not tell the whole story. However, in all analyses reported in this paper, only clicks were used to measure usage. Another point worth mentioning is that the “Non-preselected” suggestions did not necessarily mean that the suggestions were of the “anywhere” or “keyword” type. They might be an author, subject, or title suggestion. The “non-preselected” suggestions only meant the user did not pre-select any search field before (s)he took a particular suggestion. But the suggestion itself might have the type other than “anywhere”.

As shown in Figure 3, for the suggestions taken without a search field preselected (non-preselected suggestions), most (62.1%) were the “anywhere” suggestions, followed by subject suggestions (18.8%) and author suggestions (14.75%). On the other hand, of the taken suggestions with a preselected field (preselected suggestions), most (46.44%) were the title suggestions, followed by author suggestions (34.55%). “Anywhere” suggestions were the least (0.79%). The height differences of the clustered bars in Figure 3 suggest if conducting a title or an author search, searchers were likely to preselect the particular field beforehand. In contrast, if performing a subject search or a generic search, they tended not to select a search field in advance, relying on where the type of suggestions led them.

Influence of Real-Time Query Expansion on Search Behavior

One of our primary interests is to determine what effect the real-time query expansion could have on users’ search behavior. Thus we focused much of our attention on statistics gathering on tracing the changes in search field, query length, and number of query reformulation, and most importantly, the failure rate, in April 2009 and April 2010.

Search Field

Figure 4 compares the change in number of search requests by search field before and after implementing the suggestion feature. As shown in Figure 4, the largest change was the rate of subject search. Although subject searches were a relatively minor portion of the overall queries submitted to the catalog, just 1.88%. They have seen the largest percentage increase to 8.99% in April 2010, of which 3.91% increase was from the suggestions. The other 5.80% increase was from the user typed subject searches, which might include the cases where users manually duplicated the subject suggestions rather than clicking them. Before enabling the auto-suggest feature, there were on average 108 subject searches each day. After
implementing the feature, the average number increased to 612 per day, a 466.5% increase.

Over the same period, title searches saw a minor decrease. This is a little surprising since we expected the query suggestion feature would also promote title searches, which belongs to the authority search as well. The reason might be that the availability of subject suggestions facilitated the subject search to a larger degree. Many subject searches could have been a title search had the subject suggestions not been available. Keyword and author search stayed approximately the same across two periods.

**Query Length, Query Submissions, and Viewing Items**
Query length is one of the parameters to evaluate searchers’ query formulation tactics. In April 2009, an average query had 3.1 terms. After implementing the query suggestion feature, queries were a bit longer than before, averaging 3.7 terms. This is highly likely due to the standard format of the authority data as suggested queries, which are typically longer than user-generated queries. Longer queries were generally desirable according to a research suggesting that longer query led to higher satisfaction and less iteration (Belkin et al., 2003).

As mentioned before, query reformulation is the process the initial query is adjusted. In April 2009, people submitted 4.43 queries on average per search whereas in April 2010, the number was decreased to 2.67, a roughly 40% decrease. It seems that query expansion feature increases the search efficacy by reducing the number of search iterations.

In both periods, approximate 50% of searches had the initial query reformulated. That means another 50% searches had only the one query submission before the search was finished up. We also observed that many of the queries beyond the first iteration were simple deviations from the initial one. This observation is in line with White and Marchionini’s finding (2007) that many further queries were just “syntactic variants” of the initial one. The initial query appears important in determining search success. When checking with the initial query, we found the average length in 2009 is 2.8 terms, and in 2010 3.8 terms. We saw larger difference in the initial query length than that in all query length. In addition, 32% of the initial queries were from query expansions, larger than the proportion for all the queries (22.6%). This finding agrees with White and Marchionini (2007)’s study that query suggestions were generally most needed at an early stage during a search.

**Failure Rate**

One of the purposes for implementing the query expansion is to promote search success. Many OPAC studies have included failure analysis, in which a failed search was typically defined as a search that matched no documents in the collection (Jones et al., 2000). There was no agreement, however, on the definition of searching failure among researchers. Large and Beheshti (1997) stated that not all zero hits represent failures and not all hits represent successes. Some researchers also defined an upper number of results for a successful search (e.g., Cochrane & Markey, 1983). In this study, we define a success search as a search that returns from zero to an upper number of search results. Based on our experience processing with the logs and the user experiments we have conducted, the upper number is chosen to be 30 items (i.e., 3 pages of result). One challenge for tagging those failed searches is that log data did not capture the number of results returned by a particular query. It only captured the request the searcher sent to the server and the total amount of bytes the server sent back to the searcher. While this is not perfectly reliable since the response size varies depending on the length of the users search string, but generally speaking, zero hit pages would be much smaller than result pages that get results, and large number of returned results would be much larger than typical result pages. Thus, the amount of bytes sent from the server may serve as a rough estimate of the number of results returned by a query. Through testing, we figured out a rough threshold—which searches with data flow between 15,000 bytes and 70,000 bytes, roughly equal to zero to 30 results, would be a successful search. Figure 5 below demonstrates the failure rate before and after the query expansion’s launch. As we see, this failure has greatly dropped in 2010.

We admit that this failure definition is not perfect. Particularly it is only at the single text search action level, not at the search session level. It is really challenging to define failure with only logs, because logs did not record any context information within which the search was conducted. Lacking this information, it is hard to judge what is needed by the searcher and if (s)he finds the ones (s)he needed. The disadvantage of the log studies could be

<table>
<thead>
<tr>
<th></th>
<th>April 2009</th>
<th>April 2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query Length</td>
<td>3.1</td>
<td>3.7</td>
</tr>
<tr>
<td>Number of Query Submissions per Session</td>
<td>4.43</td>
<td>2.67</td>
</tr>
<tr>
<td>Number of Items Viewed per Session</td>
<td>2.31</td>
<td>1.36</td>
</tr>
</tbody>
</table>

Table 2. Query length, number of query submissions, and number of items viewed
complimented by user experiments, where we recruit subjects to the lab and give them the pre-selected tasks. Under such a controlled environment with such controlled search tasks, it is much easier to pre-define the search success. The user experiment is ongoing in our lab.

CONCLUSION
Offering users with real-time query suggestions offers opportunities for presenting users with catalog authority data in a familiar and helpful manner. Users were able to take advantage of this feature, adopting it 22.6% of all searches. The generic keyword suggestions were taken most, followed by subject suggestions and author suggestions. The catalog has seen a significant increase in the rate of subject search, part of which were selected directly from the subject suggestions. There has also been an increase in query length, from 3.1 terms to 3.7 terms per query. It is found that the auto-suggest feature has reduced the query reformulation based on the decreased number of queries submitted per search. The failure rate has also been dropped dramatically. The study has given us insight into the circumstances under which the real-time suggestions are more needed, how searchers use them, how the usage change searchers’ behavior, and potential enhancements for the feature.

REFERENCES


