Modeling User Interactions for Automatic Library Search Evaluation: Preliminary Results

Qi Guo
Math & Computer Science
Emory University
qguo3@emory.edu

Selden Deemer Arthur Murphy
University Libraries
Emory University
{libssd, armurph}@emory.edu

Eugene Agichtein
Math & Computer Science
Emory University
eugene@mathcs.emory.edu

ABSTRACT

We present a preliminary model and results for automatic, behavior-based evaluation of effectiveness of a library's online and catalog search tools for locating scholarly resources. We first introduce a simplified user model to explicitly state what we consider as "success" in library search. Then we introduce quantitative behavior-based model of searchers as Markov processes, that we tune based on more than one million real user interactions of library patrons with the search tools. Our preliminary results demonstrate feasibility of automatically, and in real time, measuring the effectiveness of library search, and suggest promising areas of improvement for creating more accurate user models, better user instrumentation, and methods for speeding up development and testing of online services.

Categories and Subject Descriptors

H.3.3 [**Information Systems**]: Storage and Retrieval: Evaluation

General Terms

Design, Experimentation, Evaluation

Keywords

User behavior modeling, library search evaluation, search success

1. INTRODUCTION

As library services move increasingly online, search for scholarly resources has been growing ever more important. Libraries now are portals that combine search over catalogs, electronic journals, online resources, and, increasingly, provide meta-search functionality to query and integrate information from other resources, such as Google Scholar [1].

Figure 1 gives an example of the EUCLID library search portal, operated by the Emory University Libraries [2]. EUCLID serves as a starting point for searching the Emory library catalogues, and electronic journal databases, and other holdings.

As online library services proliferate, empirical evaluation is becoming crucial for making informed decisions about library improvements, service offerings, and development. For example, quantitative effectiveness measures should play a key role in comparing search systems, user interfaces variants, or ranking functions. While empirical evaluation is a mature area in Information Retrieval, more work is required to adapt the metrics for the library settings. In particular, we propose automatic, *user-centric* behavior-based metrics for evaluating the effectiveness of library search tools. Our metrics allow effective real-time monitoring of the state of the current library services, and provide an effective platform both for scientific research and for development and testing of improved library services.

Specifically, our contributions include:

- *A* behavioral *user model for success of search*: We introduce a preliminary behavioral user model of searcher success that is designed to be evaluated by observing user interactions alone (Section 2)
- *A robust method for computing the relevant model values*: We estimate the success metrics using Markov processes, which provides a robust and scalable methodology for comparing success across systems and time periods (Section 3).
- *Large scale study with millions of user interactions*: We present preliminary results computed over more than a million user interactions with the EUCLID search engine, collected over the period of three months in 2007. (Section 4).

Our work builds on decades of evaluation research in information retrieval and information studies, which we review in Section 5. We then discuss our conclusions and current work in Section 6, which concludes the paper.

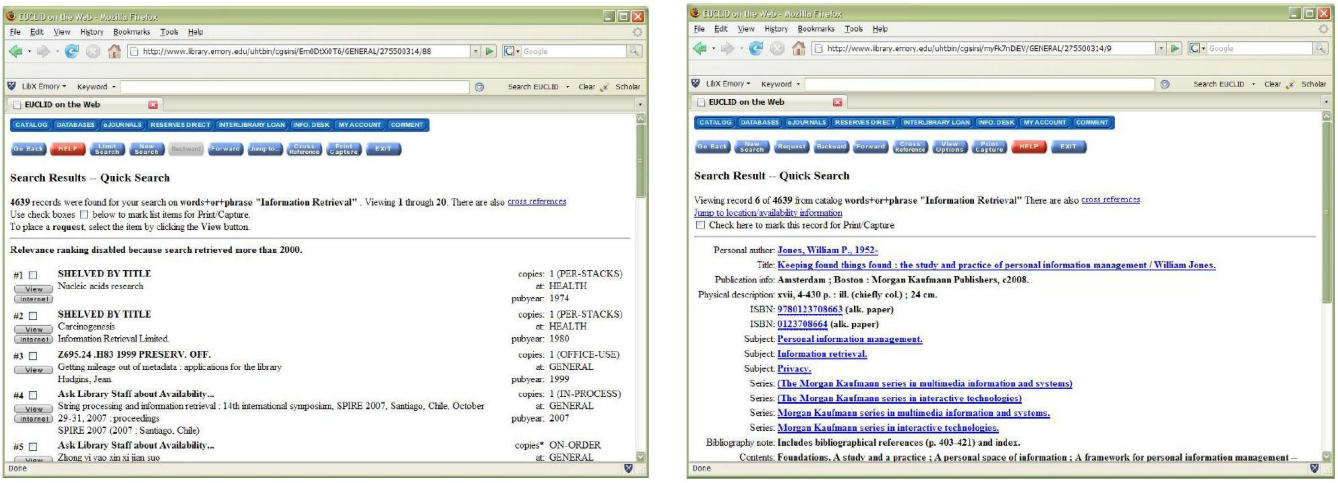
2. LIBRARY EFFECTIVENESS METRICS

Results of a typical library search are shown in Figure 1. A query (usually in free form) is submitted to the library search engine web interface (Figure 1(a)), and a number of results (usually in a list) are returned (Figure 1(b)). Our goal is to automatically determine if the searcher is successful in finding the desired scholarly resources.

We start with a simple user model of success and failure that still captures important characteristics of search effectiveness. Figure 2 illustrates the basic flow of resource discovery using EUCLID search. The flow is generic enough to be applicable to other library search systems as well. In particular, we consider primary indication of success as viewing a detailed display for an item. However, there can be other ways a searcher can be satisfied, for example, by printing the call numbers from the list of result page shown in Figure 1(a). For these preliminary results, we simplify the model to consider success as only viewing single item in detailed view. Even though we realize that not viewing a single item result may still indicate a successful search (e.g., if the user simply printed the list of results), for this preliminary study we consider any search that does not end in "Success", as "Failure". It is appropriate to point out at this point that this limitation is due to the constraints of

[1] http://scholar.google.com/
[2] http://www.library.emory.edu/

JCDL 2008



(a) (b)

Figure 1: Search result page (a) and the corresponding detailed view (b) for query "Information Retrieval"

server-side logging: we have no idea what the user is doing on her client (web browser). We are currently addressing this limitation by developing *client-side* instrumentation, which will enable much more fine-grained analysis. We discuss our current work further in Section 6.

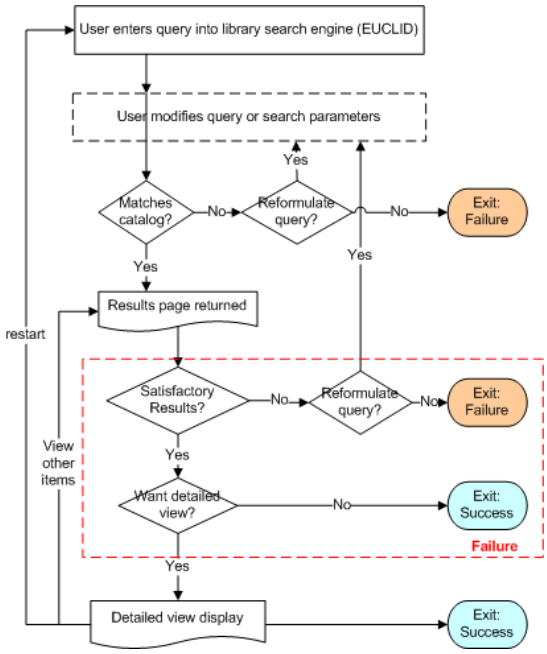


Figure 2: An illustration of the basic flow of successful resource discovery using Catalog Search (EUCLID)

Note that we distinguish *sessions* and *searches* in our analysis. A *session* is started when a user loads the EUCLID front page, and continues until either the browser is closed or there is no activity for some period of time. However, there can be multiple (successful) searches within a *session* - that can happen if a user reaches a "Success" state and then starts a new search without logging out. In contrast, there can be at most one "Failure" per session, as we consider a search to continue until either "Success" is reached or the session is ended.

3. ESTIMATING SEARCH EFFECTIVENESS

We now describe how to compute the values for our success metrics automatically based on user interactions. First, we present our model,

3.1 Modeling Searchers as Markov Processes

The states describe the different states a searcher can be in (e.g., "viewing list of results"). Transitions are actions (e.g., clicking on result or "back" button in browser). By modeling the transition and emission probabilities we can predict what an "expected" searcher is likely to do, and automatically estimate whether she was successful.

We model the following basic states in the search process, that correspond to the boxes in Figure 2.

- **Search (main)**: The EUCLID search main page, the entry point where most users begin the search process.
- **View results**: List of matching items, as shown in Figure 1(a).
- **Refine search**: Results of using advanced search interface displayed at the end of the original results listing.
- **View item detail**: Detailed record information for an item (or possibly the item itself, for electronic holdings), as shown in Figure 1(b).
- **Other actions**: There are many other actions such as renewing library materials, that we omit from the subsequent discussion for clarity.
- **Exit**: End of session for the user (e.g., browser exit)

Then, for any search outcome (e.g., "Success", or "requested item") we can compute the probability of a searcher reaching that state as:

$$P(Outcome) = \sum_i P(Outcome|state_i) \cdot P(state_i)$$

Using standard Markov assumptions we calculate the probability of searcher being in state i as $P(state_i) = \sum_j P(state_i|state_j)$. The transition probabilities are computed by counting transitions in the logs. The emission probabilities $P(Outcome|state_i)$ are fixed using domain knowledge (as in Figure 2), and could be further refined with user surveys and more fine-grained instrumentation, as discussed in Section 6. Therefore, given a trained Markov model as described above we can finally compute Success Rate as:

$$P(Outcome) = \sum_i \sum_j P(Outcome|state_i) \cdot P(state_i|state_j)$$

This model allows us to make predictions about searcher behavior for any state of the search process, including two special cases

that we focus on in this paper: "Success" and "Failure" of library search.

Calculating Success Rate: As a special case of outcome we consider the rate of success or failure for search. Intuitively, for this special case, we simply compute the fraction of searches that ended in "Success". The rest, by our definition, end in "Failure".

3.2 Building The Markov Model

To collect the statistics to "train" the Markov model we parse the search logs containing interactions with the EUCLID search engine. We analyzed the EUCLID search logs, which required only minor pre-processing and data cleaning, which we omit due to lack of space. An HTTP server log contains information such as IP address, session ID, and the URL requested. Mostly, the EUCLID logs are standard HTTP request logs, but with special URL codes corresponding to particular action (e.g., "/88/" corresponds to viewing a result list). Therefore, given the definitions of states above we can compute all the necessary statistics by analyzing the search logs. In fact, the log pre-processing component is the only part of our approach that is specific to EUCLID - once the search log is processed and normalized, our techniques could be applied to almost any library search engine.

4. RESULTS AND DISCUSSION

First, we describe the datasets used for the experiments. Then, we present a simplified version of the Markov model we built over a subset of the data (for illustrative purposes). Finally, we report some statistics of searcher success that we could potentially generate in real time for any specified time period.

4.1 Datasets

For the preliminary study we used the logs from April, July, and September of 2007. The datasets are summarized in Table 1. These months were selected to compare the metrics between Spring vs. Summer vs. Fall months to demonstrate usefulness of continuous, real-time monitoring.

Month	*Total Actions*	*Sessions*	*Searches*
April	789,825	65,904	71,478
July	380,525	38,358	35,877
September	612,970	58,282	58,008
Total:	1,783,320	162,544	165,363

Table 1: Query log data from the EUCLID search for April, July, and September of 2007

4.2 Searcher Markov Model

Figure 3 illustrates users' paths to finding scholarly resources with the EUCLID search. The graph represents the simplified Markov Model for the searcher actions. Note that for clarity we do not model the "Other actions" state, and ignore the edges with transition probability of lower than 0.05. Also note the transition from the *View item detail* state to itself). The reason for this behavior is that the users use the browser button to go back to the previous page (View Results), and then click on other results to go back to detail view. We will discuss extensions and more fine-grained models in Section 6.

4.3 Success Rates

We now summarize the success rates for each time period in Figure 4. As we can see, the success rate remains nearly constant across the time periods, decreasing slightly (from 57.1% to 56.2%) from April to subsequent months.

We also report the average amount of user required effort for each time period. As we can see, the average amount of time before success decreases slightly in September (which could be attributed

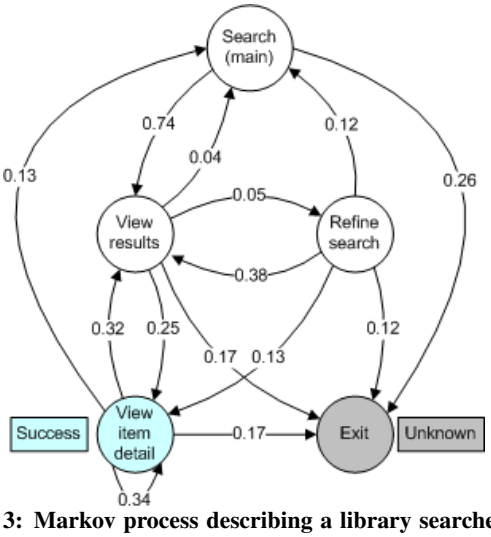


Figure 3: Markov process describing a library searcher (July 2007)

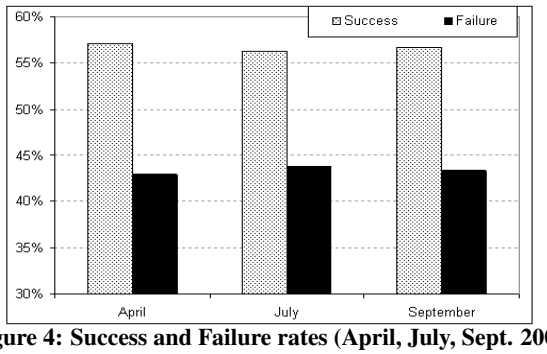


Figure 4: Success and Failure rates (April, July, Sept. 2007)

to the influx of impatient undergraduate students), while the average number of actions (e.g., queries, search refinement, clicking on a result for detailed view) remains relatively constant at nearly 5.

We now consider the distribution of effort (actions and time) across searches (Figure 5(a) and (b)). Not surprisingly, nearly 70% of searches are "easy", requiring fewer than 6 actions from a user, and nearly 80% require less than 100 seconds to find an item of interest. Nevertheless, there is a significant amount of searches in the "tail" that require many actions (or large amounts of time) to complete. Presumably, these are the searches where a librarian might be able to help, and metrics such as ours might provide automatic ways to detect such "difficult" searches. Finally, Figure 5(c) reports the comparison between April, July, and September for the distribution of searches according to the number of required actions. As we can see, there is a (statistically significant) change for the month of July, showing that in July the users were willing to expand less effort (actions).

5. RELATED WORK

Our work is related to user modeling for web search, where the goal is to predict which results will be relevant (e.g., [1, 16, 15, 7]); other uses include classifying user intent into a particular category (e.g., [14])). This work builds on the influential user model

	Period		
Metric	April	July	Sept.
Average time before success (seconds)	74.88	74.72	72.60
Average number of actions before success	4.95	4.82	4.83
Average success rate	57.1%	56.26%	56.70%

Table 2: The amount of effort required before "success" in finding needed resources (April, July, and September 2007)

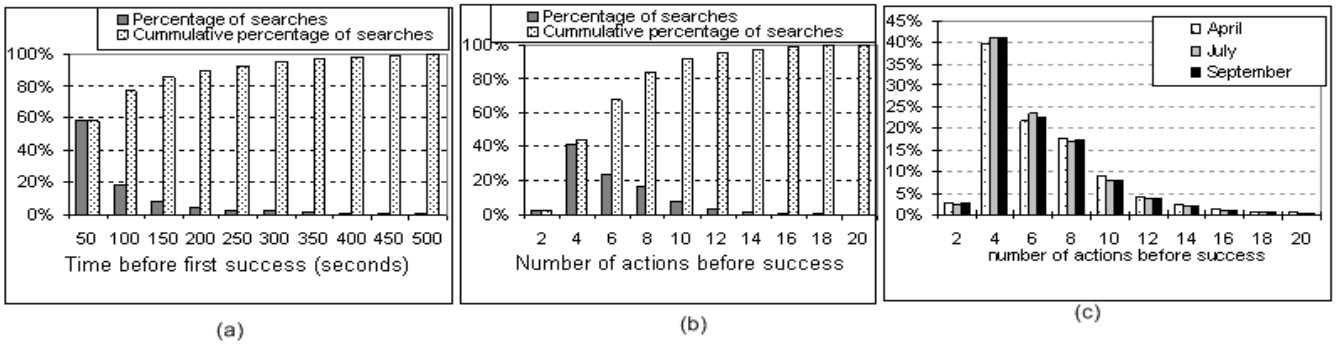


Figure 5: Distribution of searches according to number of actions (a) and the amount of time (b) before "Success" is achieved

introduced by Belkin et al. [2, 3].

More closely related to our work are previous efforts to evaluate the effectiveness of search based on user clickthrough on the results (e.g., [8, 12, 5]). Unlike these references, our focus is not to evaluate the relevance of the results, but rather the overall effectiveness of searchers measured by the amount of effort (or success) they have. Recently, eye tracking has started to emerge as a useful technology for understanding some of the mechanisms behind user behavior (e.g., [9, 6], which may provide additional insight into user satisfaction with web search results. We are currently developing precise client-side instrumentation that attempts to capture similar information.

Many qualitative studies have been done in information and library studies. Closely related to our work is the study of Nygren et al. [13] which compared the effectiveness of MetaLib with that of Google Scholar[3]. More recently, Wrubel and Schimdt [17] evaluated the usability of a metasearch interface to understand student perceptions of metasearchs usefulness and to learn if students could effectively complete research tasks. Most recently, Jung et al. [10] reported an evaluation and usability testing experiments for a library-based metasearch system. For a broad review of usability studies of web-based tools done in the library setting, see [4]. In contrast to previous work our goal is to develop automatic, quantitative methods based on user behavior and interactions with the search tools – allowing A/B testing (e.g., similar to the controlled experiments for web services [11]) comparison between systems, and general understanding how well the library is performing.

6. CONCLUSIONS AND FUTURE WORK

We presented a preliminary model for automatic behavioral success metrics for evaluating library search. With our definition of success, we evaluated the library's web services on some reasonable parameters (e.g., average amount of effort required to successfully complete a search). Our techniques could be used for comparing effectiveness across different search systems (e.g., to compare products from different vendors; for comparing the user interface variants; and for evaluating, and improving, the ranking functions and other core search functionality.

A significant limitation of this preliminary study is that all our metrics were collected on *server-side*: meaning, we have no access to information what the user may be doing in her web browser. We are currently experimenting with *client-side* instrumentation, which would tell us, for example, if the user has printed a page (hence, likely to be successful) before exiting the browser session. We are also working on developing more precise success metrics (that go beyond single item display and printing a page). In both cases, we plan to use the current framework with minimal modification. In summary, despite the current limitations, our preliminary results demonstrate the feasibility of our approach, and are already serving as a valuable building block for the more fine-grained analysis.

[3] http://scholar.google.com

7. REFERENCES

[1] E. Agichtein, E. Brill, S. Dumais, and R. Ragno. Learning user interaction models for predicting web search result preferences. In *Proc. of SIGIR*, 2006.

[2] N. Belkin, R. N. Oddy, and H. M. Brooks. Information retrieval: Part ii. results of a design study. *Journal of Documentation*, 38(3):145–164, 1982.

[3] N. J. Belkin. User modeling in information retrieval. *Tutorial presented at the Sixth International Conference on User Modelling (UM97)*.

[4] B. A. Blummer. *A Literature Review of Academic Library Web Page Studies*. The Haworth Press, Inc, 2007.

[5] B. Carterette and R. Jones. Evaluating search engines by modeling the relationship between relevance and clicks. In *Proceedings of the Advances in Neural Information Processing Systems (NIPS)*, 2007.

[6] E. Cutrell and Z. Guan. Eye tracking in msn search: Investigating snippet length, target position and task types, 2007.

[7] D. Downey, S. T. Dumais, and E. Horvitz. Models of searching and browsing: Languages, studies, and applications. In *Proc. of IJCAI*, 2007.

[8] T. Joachims. Evaluating retrieval performance using clickthrough data.

[9] T. Joachims, L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. Evaluating the accuracy of implicit feedback from clicks and query reformulations in web search. *ACM Trans. Inf. Syst.*, 25(2), 2007.

[10] S. Jung, J. L. Herlocker, J. Webster, M. Mellinger, and J. Frumkin. Libraryfind: System design and usability testing of academic metasearch system. *Journal of the American Society for Information Science and Technology*, 59:375–389, 2008.

[11] R. Kohavi, R. M. Henne, and D. Sommerfield. Practical guide to controlled experiments on the web: listen to your customers not to the hippo. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, pages 959–967, 2007.

[12] Y. Liu, Y. Fu, M. Zhang, S. Ma, and L. Ru. Automatic search engine performance evaluation with click-through data analysis. In *Proceedings of the 16th international conference on World Wide Web (WWW)*, 2007.

[13] E. Nygren, G. Haya, and W. Widmark. Students experience of metalib and google scholar. *Report to BIBSAM*, 2006.

[14] D. E. Rose and D. Levinson. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web (WWW)*, 2004.

[15] R. White, M. Bilenko, and S. Cucerzan. Studying the use of popular destinations to enhance web search interaction. In *Proc. of SIGIR*, 2007.

[16] R. W. White and S. M. Drucker. Investigating behavioral variability in web search. In *Proc. of WWW*, 2007.

[17] L. Wrubel and K. Schmidt. Usability testing of a metasearch interface: A case study. *College and Research Libraries*, 68, 2007.