

Assessment Plan

Introduction

(http://www.cdlib.org/inside/projects/melvyl_recommender/)

The goal of this project is to inform development of next-generation online public access library catalogs, through a feasibility study of strategies that have the potential to enhance the usability of traditional online library catalogs. In particular, the desirability of using varying forms of relevance ranking and recommendations will be tested.

Needs Assessment and Evaluation Goals

Needs assessment

Optimally, a development project includes sufficient time to identify target users and conduct a robust user needs assessment. This provides an opportunity to gain a rich understanding of the issues of interest and concern to the users, the information practices in which they currently engage, the barriers they encounter and associated work-arounds they employ.

However, the Melvyl Ranking and Recommender project is more aptly described as a feasibility project, as opposed to a full development project. This is reflected in the time line and scope of the grant, which do not allow sufficient time for an optimal user needs assessment if we want to feed the results back into the prototype before the end of the project. Instead, we will employ a “hybrid” model in which we will build a composite “profile” of user needs from CDL generated user personas, previously conducted CDL user needs assessment research, general literature on information use and information seeking of similar user groups, and needs assessments and evaluations of similar studies (e.g. other studies on relevance and ranking).

As we assemble these components, we will identify any serious holes. The composite profile will be validated and enriched by asking each user a small set of questions about their information needs, behaviors and academic and subject-expertise before the evaluation takes place.

Evaluation

The evaluation approach must be very constrained and streamlined. Three evaluation areas and related goals were laid out in the grant:

1. **Relevance.** Determining the optimal combination of weights for relevance will require extensive evaluation of the level of “aboutness” within a given set of results. Understanding what measures users employ to judge the accuracy of ranked results, and what factors contribute to a user being satisfied with a results set, will be assessed via interviews with both subject-expert and non-expert users.
2. **Recommending and auto-correction.** Users’ perception of the value and utility of recommending and auto-correction functions within the OPAC environment will be assessed via surveys and user interviews.
3. **User interface design.** Basic wireframe designs outlining the navigation structure, information architecture, and use of terminology will be developed for the display of relevance and recommender results. The design of wireframes will be iterative, and user task-based testing will be conducted to inform a deeper understanding of how users interact with relevance and recommender functional interfaces.

System support for Features 1 and 2, relevance and recommendations, will be evaluated independently from one another. The user interface design evaluation is a secondary goal, and will be examined as part of the relevance and recommendation systems. This approach reduces the number of separate evaluation events that must be conducted, which is important given the

short time frame available for this aspect of the project.

Target Audience

The target audience of this assessment will consist of graduate and undergraduate students. While the project was designed with any academic user in mind, it may be more difficult to gain access to faculty. Because there is significant task overlap between faculty and graduate students, we propose to work primarily with graduate students, and secondarily with undergraduate students.

The grant application specifies that both experts and novices will evaluate the various approaches to ranking. This specification will be extended to users evaluating recommendations as well. Recruiting groups of both graduate and undergraduate students can meet this requirement.

Academic Discipline

Past research on information seeking in the academic environment has tended to focus on the sciences (and indeed there is an entire area of work devoted to “science studies”) and the humanities, sometimes blended with social sciences. Scholars in the sciences tend to be more interested in journal articles and following citations. This quality makes them less desirable for evaluating our two systems, which are focused on books, conference proceedings, and other items not at the level of individual journal articles. We will therefore identify undergraduate and graduate Humanities and History students, whose information interests are more likely to be met by the collection we are working with.

Physical Access

We will need easy physical access to at participants for observed evaluations. UC Berkeley offers the most accessible pool of users, due to proximity.

Subject Selection Criteria

Broadly construed, two types of users are required for evaluating each system, subject-naive and subject-expert.

For subject recruitment purposes, a working assumption will be made that undergraduate students are more likely to have relatively little knowledge of their discipline (and therefore be subject-naive) and that graduate students are more likely to have a greater knowledge of their discipline (and therefore be subject-expert). Each of these groups will be targeted accordingly, although undergraduates who are determined to have subject expertise and graduate students determined to be subject-naive will be used as appropriate.

Two different strategies will be used to solicit users in each of these groups.

Subject-naive users: We will work with UC Berkeley library staff to set up an evaluation room in the library and then actively recruit students in the undergraduate library. Students will be screened for their appropriateness for the study along the following criteria:

- Humanities or History student
- Junior or Senior
- Frequency of use of GLADIS or MELVYL in the last month to find a library item
- Immediately available for 1 hour

Subject-expert users: We will rely on our library users-council contacts to identify and solicit participation by such individuals. Subjects in these category will be selected based on the following criteria:

- Humanities or History student
- Year in school (prefer 3rd year+ graduate students)
- Area of focus within the user's discipline (to confirm expertise area and level)
- Available for a scheduled 1 hour appointment.

Subjects will be rewarded with a \$25 gift card at the UC Berkeley Student Store.

Evaluation Design

The evaluation will be split into two parts, one for ranking methods and one for recommendations. While evaluation of each system will be conducted independently, the overall structure will be the same:

1. **Heuristic evaluation:** conducted on an informal level, tapping CDL staff.
2. **Dry-run:** 5 users for ranking and 3 users for recommendations will engage with the system to flush out problems with the protocol and the UI. These will be observed sessions at SIMS with SIMS students (most likely targeting students in Dan Greenstein's INFOSYS 290 course) and results will be used to refine the protocol and the UI. Data from these sessions are not expected to be folded into the primary evaluation results, although we will report that the sessions were conducted in the final write-up.

Results will be recorded via tape-recorder and notes.

3. **Observed evaluation: 10 users** (5 naive, 5 expert) will use the system. Subject-naive users will engage in an orientation task and a set of predefined tasks. Subject-expert users will engage in an orientation task, a set of predefined tasks and an optionally user-defined task. One observer will be assigned to each user, and a facilitator will be present at all times. Not all evaluations will occur at the same time; the number of concurrent users will be gated by the number of available observers.

Results will be recorded via tape-recorder and notes.

Relevance Ranking

Relevance ranking evaluations will expose four ranking methods to users in order to obtain evaluations regarding the effectiveness of each method.

Users

Users will be considered either subject-naive or subject-expert. Excluding the heuristic and dry run evaluations, 10 users total are required for the evaluation of the ranking system.

Task

Evaluations will be task based and will cover all ranking methods. Each user will engage in one orientation task and 5 sets of searches associated with 5 separate scenarios, one for each ranking method and a known item search using the first ranking method of the users evaluation session.

Users will be asked to identify their expertise in regards to each scenario. The evaluation will consist of ranking the relevancy of items in result sets in terms of the predefined scenario.

Recommendations

Users

Users will be considered either subject-naive or subject-expert. Excluding the heuristic and dry-run evaluations, 10 users total are required for the evaluation of the recommendation system.

Task

Evaluation will be task based. Each user will engage in one orientation task and 4 searches associated with 4 separate scenarios, in order to exercise the system. Users will identify their expertise in regards to each scenario. Users will be asked to evaluate the quality of each recommended item and the recommendation set as a whole.

Schedule

General Project Timelines

- *Summer 2005*: Data acquisition, preparation, and analysis.
- *Late summer/fall 2005*: Development of underlying technical framework.
- *Fall/winter 2005-2006*: Development of search interfaces
- *Early spring 2006*: Assessment (preliminary results by April 2006)
- *June 2006*: Project complete

Evaluation Timeline

- December 5--Human Subjects application submitted to UCB

Ranking System

- Early January—CDL staff conducts heuristic evaluation
- January 26th and 27th—Dry Run
- February 6th - 8th—Evaluation

Recommendation System

- Early February —CDL staff conducts heuristic evaluation
- March 17th—Dry Run
- April 4th-7th—Evaluation

Objectives

This page captures the high level questions we want to answer with the evaluation of each system. The subsequent section will focus on one of these questions, enumerating the sub-questions and related evaluation tasks in which we will have users engage. Specific measures associated with each question will be also be included. Some measures may address more than one sub-question, and will be so indicated.

As a general note, the term “useful,” when used in describing either system will be operationalized to mean that through the use of the system's unique features (ranking or recommendations), the user was able to complete all or part of the assigned task. The terms “users” and “academic users” (in singular or plural) will be considered to mean the same thing, i.e. in this context, “users” are always academic users.

Ranking

- 1) Which ranking method returns the greatest number of most highly ranked relevant items, as determined by academic users?
- 2) Which ranking method was most effective at supporting the user in successfully completing the evaluation task?
- 3) How do academic users judge relevance?

Recommendations

- 1) Do recommendations help academic users find items relevant to a research-related task?
- 2) How do academic users evaluate recommendations?
- 3) What is the quality of individual recommendations?
- 4) What is the quality of a given set of recommendations?

Ranking Objective 1: Which ranking method returns the greatest number of most highly ranked relevant items for a given query, as determined by academic users?

1.1 Which ranking method results in the most checked items?

1.2 How close to the top of the result set were the checked items?

Measures

1. # and % of items checked in result set by ranking (1.1)
2. Position of checked items in set by ranking (1.2)

Ranking Objective 2 – Which ranking method was most effective at supporting the user in successfully completing the evaluation task?

2.1 Which ranking method best supported the user's task for finding at least one relevant item for a given task?

2.2 Is level of subject expertise a factor in the user's ability to complete the tasks?

Measures

1. Count of relevance assignments per scenario
2. User supplied opinion
3. Position number of relevant items
4. Time spent in scenario (query logs)
5. # searches per scenario (query logs)
6. Analysis of queries (transaction logs)
7. All of the above sorted by expertise

Ranking Objective 3-How do academic users judge relevance?

3.1 Are users better able to complete their tasks with the ranked results?

3.2 What qualities of an item do users identify as being significant in determining whether or not to check an item as relevant or not relevant?

3.3 Which components of the bibliographic record, if any, do users take advantage of in deciding whether or not an item is relevant?

3.4 How do users describe the quality of the result sets returned by different ranking methods?

3.5 Are there any differences in questions 3.1 through 3.4 between subject-naive and subject-expert users?

Measures

1. Number and position of relevant item(s) found.
2. Time to complete task.
3. Qualitative feedback regarding why an item was marked relevant or not.
4. Qualitative feedback about which component of the bibliographic entry was used to determine relevance.
5. Qualitative feedback regarding how satisfied a user was with the performance of the tasks and the support of the system.
6. Qualitative feedback regarding the quality of each result set as defined by the user.
7. Comparison of data from 1-7 between subject-naive and subject-expert users.

Recommendation Objective 1: Do recommendations help academic users find items relevant to a research-related task?

1.1 Do recommendations help users find relevant items not in the original result set?

1.2 Do recommendations lead users to create new searches?

Measures

1. # relevant items identified in recommended sets (1.1)
2. Observation of new searches based on viewing of recommendations (1.2)

Recommendation Objective 2: How do academic users determine if a recommendation is useful?

2.1 Which components of the bibliographic record do users identify as helpful in deciding whether or not a recommended item is useful?

2.2 Is subject expertise level important a factor in this?

Measures

1. # items marked as relevant in sets of recommendations
2. Qualitative feedback from the think-aloud-protocol about the use of the bibliographic entry.
3. Relevant item(s) found (user supplied)
4. Qualitative feedback regarding why an item was marked relevant or not.
5. Qualitative feedback regarding how satisfied a user was with the support of the system in being able to accomplish the tasks.
6. Qualitative feedback regarding the quality of each result set as defined by the user.
7. Comparison of data from 1-6 by expertise level.

Recommendation Objective 3: What is the quality of a given recommended item?

3.1 Is a given recommendation interesting but not useful? (i.e. is not relevant enough)

3.2 Is a given recommendation relevant but not useful? (i.e. the person already knows about it)

3.3 Is a given recommendation new, but not surprising (i.e. novel but not serendipitous)

3.4 Is a given recommendation relevant and unexpected (i.e. serendipitous)

Measures

Questions in the system

1. "Based on the scenario, this recommendation is useful." (check one)
 - "no answer"
 - "strongly agree"
 - "agree"
 - "maybe or unsure"
 - "disagree"
 - "strongly disagree"
2. "How familiar are you with this recommendation?" (check all that apply)
 - "I wrote it."
 - "I have cited it."
 - "I have read it."
 - "I have heard of it."
 - "I'm familiar with author(s)."
 - "I don't know this paper at all."
3. "How would you describe this recommended item?" (check all that apply)
 - "novel"
 - "authoritative"
 - "introductory"
 - "specialized"
 - "survey/overview"
 - "I don't know"
 - empty text box for user provided response

Recommendation Objective 4: What is the quality of a given set of recommendations?

4.1 How useful was the set in accomplishing the defined task?

4.2 Were there missing items?

4.3 How many relevant, unknown items were there?

Measures

1. # relevant items checked in each recommended set
2. # new, relevant items
3. Questions about the recommended set:

(In the system)

1. Overall, these recommendations were useful.
 - “strongly agree”
 - “agree”
 - “maybe or unsure”
 - “disagree”
 - “strongly disagree”
 - user supplied description

(Asked by the facilitator)

2. Are there any items that you are surprised are not on this list of recommendations?