# Table of Contents

Table of Contents .............................................................................................................................................. 2

1 Executive summary ............................................................................................................................................. 4

2 Project overview ................................................................................................................................................ 5
  2.1 Activities ...................................................................................................................................................... 5
  2.2 Data ............................................................................................................................................................ 5
  2.3 Assessment .................................................................................................................................................. 6

3 Text-based discovery system ............................................................................................................................... 6
  3.1 Technical approach ....................................................................................................................................... 7
  3.2 Support for experimentation ......................................................................................................................... 8
  3.3 Performance ................................................................................................................................................ 8
    3.3.1 Hardware and software configuration .................................................................................................. 8
    3.3.2 Indexing ................................................................................................................................................ 9
    3.3.3 Queries ................................................................................................................................................. 9
  3.4 For further exploration .................................................................................................................................. 10

4 Spelling Correction ............................................................................................................................................. 10
  4.1 Performance goal ......................................................................................................................................... 11
  4.2 Methods ......................................................................................................................................................... 11
    4.2.1 N-gram speller ..................................................................................................................................... 11
    4.2.2 Transpositions and doubled letters .................................................................................................... 11
    4.2.3 Double metaphone ............................................................................................................................... 11
    4.2.4 Word frequencies .................................................................................................................................. 12
  4.3 Testing .......................................................................................................................................................... 12
  4.4 Results .......................................................................................................................................................... 12
  4.5 For further exploration .................................................................................................................................. 13

5 UI Strategies ....................................................................................................................................................... 13
  5.1 Faceted browse ............................................................................................................................................. 14
  5.2 FRBR grouping .......................................................................................................................................... 14
  5.3 For further exploration ................................................................................................................................ 15

6 Enhanced relevance ranking .................................................................................................................................. 15
  6.1 Data sources ................................................................................................................................................ 16
    6.1.1 Circulation data ..................................................................................................................................... 16
    6.1.2 Holdings data ....................................................................................................................................... 16
  6.2 Generating and applying boost factors ......................................................................................................... 17
    6.2.1 Circulation boosts ................................................................................................................................. 17
    6.2.2 Holdings boosts .................................................................................................................................... 18
  6.3 Assessment ................................................................................................................................................... 18
    6.3.1 Brief literature review .......................................................................................................................... 18
    6.3.2 Objectives .......................................................................................................................................... 19
    6.3.3 Ranking methods evaluated ............................................................................................................... 20
    6.3.4 Methods ............................................................................................................................................. 20
  6.4 Findings ......................................................................................................................................................... 22
    6.4.1 Participants .......................................................................................................................................... 22
    6.4.2 Experience with relevance ranking ..................................................................................................... 23
1 Executive summary

Popular commercial services such as Google, e-Bay, Amazon, and Netflix have evolved quickly over the last decade to help people find what they want, developing information retrieval strategies such as usefully ranked results, spelling correction, and recommendations. Library catalogs, in contrast, have changed little and are not well equipped to meet changing needs and expectations. The Melvyl Recommender Project has been exploring methods and feasibility of closing this gap.

Over the course of a year (June 2005 – June 2006), the project team conducted exploratory development work in five topic areas: use of a text-based discovery system, spelling correction, user interface strategies, relevance ranking, and recommending.

The use of a text-based discovery system, XTF\(^1\), with its built-in relevance ranking capability, proved to be a promising approach. Performance on a series of simple load tests suggests that the system is capable of scaling to support millions of records and hundreds of concurrent users.

Experiments with index-based spelling correction were similarly positive. Starting with an existing index-based spelling correction algorithm and applying a number of optimizations, we met the goal of producing the right correction for a misspelled word (on the first try) 90% of the time.

Although they were not a central focus of the project, we conducted a shallow initial investigation of two strategies for improving navigation through large record sets: faceted browsing, and grouping results based on functional requirements for bibliographic records (FRBR). In both cases, initial experiments suggest that delving more deeply into these areas will result in better service to patrons.

Our investigation of enhanced relevance ranking considered whether returning result sets using content-based relevance ranking, optionally boosted by weights based on circulation and holdings data, would improve the ability of patrons to complete typical academic tasks. A task-based user assessment showed that in general, academic users do prefer relevance ranked result sets to those that are unranked (current catalogs are typically unranked); preferences differed by level of subject area expertise. Limitations due to the design of the study prevented us from making a strong statement as to which of the three ranked methods that we tested will best serve the greatest number of patrons.

We explored two major strategies for generating recommendations: an approach based on the mining of circulation data (ie “patrons who checked this out also checked out…”), and an approach based on similarities in the content of bibliographic records (“more like this…”). A task-based user assessment of the former method showed that patrons are enthusiastic about using an online library catalog with a recommendation service; testing confirmed that recommendations were successful in supporting academic tasks.

---

1 See the California Digital Library Web site for information XTF.
http://www.cdlib.org/inside/projects/xtf/
Moreover, the recommendation service was useful as a query expansion tool, suggesting alternative search strategies when users were boxed in by small or single result sets.

Plans for future work consist of a mix of shorter- and longer-term initiatives that extend the work done to date. Shorter-term, more discrete activities include support for multi-word spelling correction; incorporating persistent personalization into the prototype as a building block for additional recommending work; and an exploratory effort to identify potential applications and stumbling blocks associated with retrieval in a mixed metadata/full text environment. Longer-term tasks include additional work on automated strategies for grouping and clustering to better support search and presentation of very large data sets; extended work on recommending techniques; and investment in user-centered design and integration of new services.

2 Project overview

Popular commercial services such as Google, e-Bay, Amazon, and Netflix have evolved quickly over the last decade to help people find what they want, developing information retrieval strategies such as usefully ranked results, spelling correction, and recommendations. Online library catalogs (often called online public access catalogs, or OPACs), in contrast, have changed little and are notoriously difficult for patrons to use.

2.1 Activities

Over the past year (June 2005 to the present), the Melvyl Recommender Project has been exploring methods and feasibility of closing the gap between features that library patrons want and have come to expect from information retrieval systems and what libraries are currently equipped to deliver. The project team has conducted exploratory development work in five topic areas: use of a text-based discovery system, spelling correction, user interface strategies, relevance ranking, and recommending. Work in each of these areas is described in detail in this report. (See appendices A and B for a project timeline and a roster of project team members and other contributors).

2.2 Data

Over the course of the project, the development team built a prototype to demonstrate the most promising approaches (see Appendix C for a description and screen shots). The test bed of bibliographic data underlying the prototype consists of approximately 4.5 million bibliographic records extracted from the Melvyl union catalog, the entire UCLA collection. We considered and experimented with a number of permutations. For example, we considered an expanded bibliographic record set (about 9 million total records, combining UCLA and UC Berkeley collections), or a more focused bibliographic record set (including only records with ISBN, or only records in a certain topic area). In the end we opted to maintain an environment with a mix of records similar

---

2 Many of the key issues for University of California library patrons are detailed in the Bibliographic Services Task Force Report, drafted in late 2005. [http://libraries.universityofcalifornia.edu/sopag/BSTF/interim081405.pdf](http://libraries.universityofcalifornia.edu/sopag/BSTF/interim081405.pdf)

3 The Melvyl union catalog serves the 10 campus University of California system. [http://melvyl.cdlib.org](http://melvyl.cdlib.org)
to those found in the Melvyl union catalog, but without the complexities introduced by merging records for multiple campuses.

Additional data on holdings and circulation, used in relevance ranking and recommending explorations, was obtained from several sources and evaluated. Features in the prototype are based on UC-wide holdings data from the Online Computer Library Center (OCLC), and historic circulation data from UCLA (September 1999 – May 2005). We also considered but did not use data from several sources: holdings data from a consortium of research libraries, provided by the Research Libraries Group (RLG); WorldCat-wide holdings data from OCLC; and additional circulation data from UC Berkeley. (These data sets are discussed in more detail in Appendix D, as well as in “Enhanced Relevance Ranking” and “Recommending” sections of the report.)

2.3 Assessment
The team also conducted a series of user assessment activities. They were small-scale and limited in scope: they focused only on the relevance ranking and recommending aspects of the project, and were limited to users in the humanities and history.

We chose to limit the scope for several reasons:

- The abbreviated project timeline left a very short time window for assessment activities.
- The need to submit a project plan to UC Berkeley months in advance in order to gain approval for research on human subjects imposed an additional time constraint.
- Given available time and resources, we had to choose between investing our time in a small number of observed sessions (richer qualitative data), or in developing a robust user interface for a larger number of remote sessions (more quantitative data). We opted for the former.

We chose to focus on humanities and history participants in part because there is voluminous background research available on academic information-seeking behavior and needs in these domains; this compensated to some degree for the fact that we lacked adequate time for a thorough needs assessment. In addition, these domains are well covered by the bibliographic data in the test bed.

Appendices G – J contain the assessment plan and instruments; in addition, specific assessment activities and findings are discussed in detail in appropriate sections of the report (see “Enhanced Relevance Ranking”, p. 15 and “Recommending”, p. 36).

3 Text-based discovery system
The prototype is built in a very different way than a typical library catalog. Rather than a relational database, it utilizes a text-based discovery system: the eXtensible Text Framework (XTF). Written in Java and XSLT 2.0, it is based on robust open source

4 See the California Digital Library Web site for information about the XTF system.
http://www.cdlib.org/inside/projects/xtf/
software (Lucene, Saxon) and is itself freely available. (See Appendix N for notes on XTF modules that were written or modified as a result of this project.)

There were several reasons for selecting this approach. First, it avoids the overhead of a relational database system. Second, XTF has built-in search and formatting capabilities; little development was necessary to provide basic search and display immediately. Finally, and most important for the purposes of this project, unlike a relational database this system includes built-in content ranking capability. Each of the bibliographic records in the system is treated as a document, and can be returned in ranked order by relevance to the query.

### 3.1 Technical approach

Our initial approach was to transform the bibliographic metadata records that were extracted from the Melvyl system in the MARC\(^5\) format, first to MARCXML\(^6\) and then to a more flexible MODS\(^7\) representation, finally storing the MODS records on the file system. We indexed the MODS records, and retained them on the system as individual records for display.

We encountered a series of challenges related to this approach.

- The XTF system is optimized for handling large documents, not millions of small ones. This introduced a number of technical limitations when we attempted to index up to 9 million records: for example, we ran out of inodes\(^8\), and encountered memory problems during indexing.
- The initial transformation strategy of MARC to MARCXML to MODS did not support some key user needs. For example, the metadata retained in the transformation to MODS did not adequately describe the physical format (e.g. book, CD) of items in the collection.

After some experimentation, we opted to try a different strategy. Instead of working with millions of individual MODS records on the file system, we indexed records in batches from large documents containing many MARC records, transforming them to MARCXML on the fly during the indexing process (abandoning MODS). And we built the display directly from the indexes, rather than from stored documents in the file system.

This approach was much more scaleable. But there were still a number of technical adjustments and challenges to be addressed in order to accommodate large quantities of bibliographic records.

- We encountered encoding problems with MARC records, requiring logic for skipping bad Unicode characters.

---


\(^6\) MARCXML is an XML representation of MARC data. More information is available from the Library of Congress. [http://www.loc.gov/standards/marcxml/](http://www.loc.gov/standards/marcxml/)

\(^7\) The Metadata Object Description Schema (MODS) schema includes a subset of MARC fields and uses language-based tags rather than numeric ones, as well as regrouping some MARC elements.

\(^8\) An inode is a data structure that stores basic information about objects in a file system. See [http://en.wikipedia.org/wiki/Inode](http://en.wikipedia.org/wiki/Inode).
• We had to invest significant effort to work with bibliographic records: deciding what elements and attributes to index, investigating which MARC fields contained essential data, and implementing code expansions.

• We had to modify existing index utilities for removing documents, running indexing statistics, and dumping indexes because they did not scale to very large record sets.

• We had to develop work-arounds in order to develop a more powerful keyword query to operate across multiple fields. In particular, we had to develop a way to highlight keywords used in the query in the results display (often called “keyword in context markup”) on these multi-field queries.

3.2 Support for experimentation

Some additional modifications to the system were also required in order to provide flexible support for experimental activities. For example, some development was required in order to support the application of document-level boost factors to the scoring routine at run-time, rather than at index time (see “Enhanced relevance ranking”, p. 15). Applying the boosts at run-time allowed us to iteratively generate and examine the effects of different boost factors on result sets without the enormous overhead of re-indexing. For similar reasons, we developed a better display of score explanation logic so that we could easily examine result sets with an explanation of each document’s score.

In addition, we added support for asynchronous calls to the server using AJAX (Asynchronous Javascript and XML). This strategy allowed us to rapidly integrate additional services not provided by the core retrieval system. We applied this technique to integrate recommendations that were produced outside of the XTF system into the XTF display (see “Recommending”, p. 36).

3.3 Performance

Current instances of XTF used in production settings serve thousands to hundreds of thousands of documents. One of our investigation goals was to experiment with a much larger data set (millions of documents) to see how well XTF performs on tasks like indexing and query handling. We conducted a series of performance tests to consider whether this system could realistically form the basis of a production catalog. (See Appendix E for a description of methods and performance graphs.)

3.3.1 Hardware and software configuration

Tests of performance on indexing and queries were done on a system with the following configuration:

• Sun Fire v40z server
• Two dual-core AMD Opteron 875 processors at 2.2 GHz
• 16 GB RAM
• 5 Seagate SCSI hard drives (2 are used), total capacity 146.8 GB
• ReiserFS filesystem
• SuSE Linux Enterprise Server 9 operating system

The software environment included:
3.3.2 Indexing

We monitored indexing speed on the test bed collection of approximately 4.5 million records. We considered three factors: the indexing process; the optimization process, which merges multiple index files to improve efficiency for searching; and the additional indexing time required to support spelling correction (see “Spelling correction”, p. 10).

Indexing, Spelling, and Optimization: 29 hours, 27 minutes, 54 seconds.
Indexing and Optimization: 22 hours, 37 minutes, 25 seconds.
Indexing: 21 hours, 13 minutes, 23 seconds.
Optimization Only: 1 hour, 5 minutes, 19 seconds.

3.3.3 Queries

The current Melvyl system must be able to handle queries from a maximum of 600 simultaneous users, with response times remaining under 1 second for 95% of queries. So our testing attempted to determine the feasibility of XTF serving up to 600 simultaneous users at that level, in order to assess what hardware or software improvements might be required to provide that level of service.

Running a series of simple load tests on our prototype (see Appendix E) provided data suggesting that the system could reasonably handle 500-700 simultaneous users on the existing hardware. This is encouraging, because it appears to meet the basic concurrency requirement of the existing Melvyl system.

But one significant difference between the test system and a real one is the number of records. The test system contains approximately 4.5 million records; the full Melvyl catalog is approximately 30 million, almost seven times that size. How would XTF scale to a much larger data set?

Based on study of the Lucene code base and index format, we hypothesize that performance could degrade linearly with the number of data records. If this is the case, to serve seven times the number of records, a production OPAC might need to split the load over seven machines similar to that used for the Recommender prototype.

---

* See Lucene documentation for an explanation of how indexes are constructed. 
http://lucene.apache.org/java/docs/fileformats.html
Also, memory requirements might also increase sevenfold, but this doesn’t seem to present a problem because we artificially limited the prototype system to two gigabytes, and ended up using just one gigabyte during these tests. The machine used for testing contains 16 gigabytes of RAM.

In addition, a second significant difference between our prototype and the production Melvyl system is that Melvyl is a union catalog, whereas the prototype test data comes from a single source. In a union catalog setting, load testing must take into account a combination of load, index and merge programs. The "merging" aspect, which handles duplication of records from different sources, has the potential to impact the user experience of the system if there is a lag time between indexing and merging activities, and can also significantly impact response time if merging takes place on the fly in display. Our testing set aside this consideration, but future work would need to take merging into account.

3.4 For further exploration

We know that this retrieval system performs well on our test bed of 4.5 million bibliographic records, but there are many additional questions for further study. In order to consider deploying this approach in a production setting we would need to test the system on much larger record sets, ideally the full record set, in order to confirm our analysis that performance will degrade in a linear fashion with increasing numbers of records. And we would need to access better log data for testing purposes, given missing timestamps in the data we had available.

Given a full data set, would query performance degrade given a data set and user traffic equivalent to or in excess of the current Melvyl system? How long would an indexing run take? How would merging affect performance? How would the system perform on incremental additions, deletions and updates? In a more realistic union catalog setting, what additional modifications to the system might be required in order to meet or exceed baseline service requirements? Would parallelization be required to meet service requirements, and if so, how might that be accomplished? If we collect and graph CPU and IO usage during the tests, could we improve performance by identifying bottlenecks?

4 Spelling Correction

We considered two strategies for spelling correction. The first strategy was to use available software, such as GNU Aspell, that makes spelling corrections based on one or more pre-constructed dictionaries. The second, strategy was to make alternative spelling suggestions using an algorithm that draws on terms and term frequency data from the Lucene indexes. Before starting on spelling correction, we considered the data within the system. Both the bibliographic data in our test bed and a set of sample queries from the live Melvyl catalog were multilingual and contained a substantial proportion of proper nouns. Given this environment, we opted to pursue the second, index-based, approach.

Aspell software and documentation are available on Sourceforge. 
http://Aspell.sourceforge.net/

July 2006
4.1 Performance goal
In the post-Google world, people have come to expect a very simple "Did you mean X?" response, choosing only one term as a suggestion. Our developer set a goal of generating the right suggestion on the first try 90% of the time.

4.2 Methods

4.2.1 N-gram speller
We began with a base of existing Java spelling correction code that had been contributed to the Lucene project (written by Nicolas Maisonneuve, based on code originally contributed by David Spencer). The base Lucene algorithm first breaks up the word we're looking for into "n-grams" (for instance, the word "primer" might end up as: ~pri prim rime imer mer~ ). Next, it forms a Lucene OR query on the index (also built with n-grams), then runs that query against the index, retaining the top 100 hits (where a hit represents a correctly spelled word that shares some n-grams with the target word). Finally, it ranks the suggestions according to their "edit distance" to the original misspelled word. Those that are closest appear at the top. (“Edit distance” is used to mean the number of insert, replace, and delete operations needed to transform the misspelled word into one of the suggestions.)

The results of initial tests on this strategy were poor (see “Testing”, p. 12). Our developer looked for and applied several supplemental strategies:

4.2.2 Transpositions and doubled letters
Based on research showing that 80% of spelling errors in English are due to transposing two letters or using a single letter where a double-letter is needed (e.g. misspelling "correlation" as "corelation"), the edit-distance algorithm was modified to optimize for these errors by reducing the edit-distance “cost” for transpositions and double-letter errors.

4.2.3 Double metaphone
Originally developed by Lawrence Philips, the Double Metaphone algorithm\(^\text{11}\) is a simple phonetic mapping that transforms any English word into a 4 character code (metaphone). Words that have the same code are assumed to be phonetically similar. Relvyl’s spellchecker compares the metaphone of the original misspelled word to that of each suggestion. Words with matching metaphones receive a small score boost. Our developer found that for long words, the metaphone algorithm ignores the end of a word, sometimes resulting in spurious hits. The best results were obtained by modifying the metaphone key to include the base metaphone plus the first and last letter of the word.

\(^{11}\) Information and code samples relating to this algorithm are available on the Aspell Sourceforge site. [http://aspell.net/metaphone/](http://aspell.net/metaphone/)
4.2.4 Word frequencies
Because the indexes do contain misspelled words, we optimized further to boost the score if the suggested word is very common. More frequent words get higher boosts and less frequent get lower boosts.

4.3 Testing
For evaluation purposes, we ran a test program using two lists of commonly misspelled words and their corrections. First, Aspell developer Kevin Atkinson created a set of test words that he ran against Aspell in various modes as well as against a competitor, Ispell. Although not necessarily representative of common errors, this set offered the advantage of being able compare his test results with ours. For a larger (and hopefully more representative) set, we used a list assembled by Wikipedia for semi-automated spelling correction of their articles.

The test program iterates through a list of pairs: \((\text{misspelled} \rightarrow \text{correction})\) and runs each misspelling through our spellcheck algorithm, recording the position that the correct word occupies on the resulting list of suggestions. If the correction isn't found, it records "0". This program ran against a test index that simulates the frequency of words in real English texts, built from raw frequency data.

The test program was run on the Lucene base code alone, then on each of the modifications separately, then finally on all of the modifications together. These test runs were also compared against the existing Aspell/Ispell test runs on the Aspell dataset.

4.4 Results
The results of an initial test showed that on the Aspell test set, the basic code from Lucene was performing better than Ispell but not as well as Aspell, and that none of them met the 90% goal. It performed better on the Wikipedia set, although still nowhere near the 90% goal.

Table 1 summarizes the results of all test runs, illustrating how each of the supplemental strategies improved the result individually, and how they performed when combined. Together, the optimizations put our algorithm on par with Aspell on the Aspell data set, and met the goal of producing the right suggestion first 90% of the time on the Wikipedia set.

---

14 Frequencies are from a word frequency list constructed using the British National Corpus. [http://www.comp.lancs.ac.uk/ucrel/bncfreq/flists.html](http://www.comp.lancs.ac.uk/ucrel/bncfreq/flists.html)
Table 1. Results of Test Runs on Spelling Correction

<table>
<thead>
<tr>
<th>Data set</th>
<th>Algorithm</th>
<th>Location of Correct Suggestion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>First</td>
</tr>
<tr>
<td>Aspell</td>
<td>Ispell</td>
<td>39.3%</td>
</tr>
<tr>
<td></td>
<td>Aspell</td>
<td>59.2%</td>
</tr>
<tr>
<td></td>
<td>lucene base</td>
<td>39.6%</td>
</tr>
<tr>
<td></td>
<td>base + transpose</td>
<td>44.6%</td>
</tr>
<tr>
<td></td>
<td>base + metaphone</td>
<td>42.9%</td>
</tr>
<tr>
<td></td>
<td>base + freq boost</td>
<td>51.7%</td>
</tr>
<tr>
<td></td>
<td>base + ALL</td>
<td>61.3%</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>lucene base</td>
<td>62.7%</td>
</tr>
<tr>
<td></td>
<td>base + transpose</td>
<td>79.1%</td>
</tr>
<tr>
<td></td>
<td>base + metaphone</td>
<td>71.2%</td>
</tr>
<tr>
<td></td>
<td>base + freq boost</td>
<td>73.0%</td>
</tr>
<tr>
<td></td>
<td>base + ALL</td>
<td>90.5%</td>
</tr>
</tbody>
</table>

4.5 For further exploration

The current approach met our goal for effective correction on a word-by-word basis. But there are several avenues for further consideration. First, it may be worthwhile to step back briefly and analyze the tradeoffs of the current algorithm versus a third approach: the use of Aspell with a custom dictionary built from the index. The custom dictionary approach could offer some advantages in terms of control over misspellings that exist in the indexes, and may be more efficient in terms of indexing.

Second, queries are not always limited to single words. If one word of a multi-word query is misspelled, the current spelling correction algorithm does not produce predictable or desirable results (e.g. a query for “avant gard” rather than “avant garde” yields a perplexing suggestion: “want guard”). An investigation of natural language processing techniques for phrase-level spelling correction should yield much better results.

Finally, we would like to investigate optimal intervention strategies. When should the system offer spelling suggestions? Only on queries that yield no results? Should suggestions be offered for small result sets, and under what circumstances?

5 UI Strategies

Although not a central focus of the project, we touched lightly on the feasibility of two experimental strategies for improving the user experience of search and navigation through very large result sets.
5.1 Faceted browse

First, we investigated the feasibility of supporting faceted browse functionality in the XTF system. Faceted browse allows the user to continually narrow the data set along several defined facets, each of which is a hierarchy. Any item can be described along several facets (e.g. media, date, location). As the user selects terms from among the facets, the display keeps track of what terms have been chosen and offers options for further query refinement. The chosen terms can be removed in any order, so navigation within the faceted hierarchy is fluid.

Many of the basic elements necessary to support faceted browse were already present in the XTF system. The remaining pieces were relatively straightforward, and so we implemented them. The XTF system is now capable of supporting the functionality needed for a faceted browse interface.

We encountered a significant implementation barrier, however. The metadata structure present in bibliographic records is not always suitable for faceted browse applications. Many desirable facets could be supported within the existing structure: e.g. format and author. In contrast, a subject facet would be of primary importance to users, but it is not well supported in the existing metadata. Although Library of Congress subject headings are present, they are not appropriately structured to support faceted browse. They do provide substantial raw data that could be used to create a browsable hierarchy of subjects, however, given adequate time and resources to explore appropriate processing methods.

5.2 FRBR grouping

Result sets from a typical library catalog are difficult for the end user to navigate, in part because so many bibliographic records represent the same work in many languages or versions. This issue is addressed in the Functional Requirements for Bibliographic Records (FRBR) model developed by the International Federation of Library Associations and Institutions. This model conceptualizes bibliographic data in a series of entities at four levels: a work (e.g. Hamlet) is realized through an expression (e.g. Hamlet in French), which is embodied in a manifestation (e.g. Hamlet, published in French by Presses universitaires de France, 1987), which is exemplified in a specific copy.

Our team scanned literature, available algorithms and available tools and services (e.g. OCLC’s xISBN service), looked at existing or planned implementations (e.g. RLG’s RedLightGreen), and considered how we would apply and implement the model in our experimental setting.

15 Work on the optimal “shape” of hypertext for usable site navigation applies to the discussion of faceted browse structures. http://psychology.wichita.edu/surl/usabilitynews/42/hypertext.htm
For a first pass, we opted to do a very quick partial implementation of the OCLC FRBR Work-set algorithm. Our implementation simply creates work groups based on author/title keys (without the use of authority files, which relate variations in author names and book titles). The query parser is modified to search for 200 records rather than the normal 20, the algorithm is applied to create “FRBRized” work groups, and the top 20 results (individual items or work groups) are chosen and displayed. The primary questions we considered were:

- How will FRBR work-group clustering affect the presentation of results in a large union catalog? Will enough clusters form to make a noticeable improvement in presentation?
- Would it be worthwhile to consider some additional clustering below the work level (manifestations and expressions)? If so, how could this be accomplished?

Although grouping without the use of authority files was fairly imprecise, even the rough groupings provided a convincing case that pursuing FRBR at least at a rudimentary level is feasible. They provide noticeable improvement in the quality of results, particularly for more general queries.

### 5.3 For further exploration

There are many convincing examples of the value of faceted browse as a method of access to large and complex data sets such as those found in bibliographic systems. A logical next step would be an investigation of alternatives for automated processing of subject headings to produce a reasonable metadata structure to support browse within a subject facet. The investment necessary to produce a well-structured subject hierarchy could also potentially be leveraged to support more efficient ways of generating recommendations outside of the browsing context.

Our initial experimentation with FRBR showed that obvious next steps would be to incorporate the use of authority files, and to pursue sub-groupings in the display of results (e.g. by language). It also raised many questions for further exploration. Would it be optimal to group records at index or pre-index time rather than at query time? Are there ways to take advantage of FRBR groupings beyond the presentation of result sets (e.g. using FRBR groupings to reduce apparent duplication in recommendation sets or to provide more comprehensive measures of holdings and circulation for use in ranking and recommending)?

### 6 Enhanced relevance ranking

At present, the Melvyl union catalog (like most other library catalogs) orders result sets on a last-in-first-out basis; the most recently catalogued items are presented at the top of the list. In contrast, many popular commercial sites offer result sets ranked by relevance. Another effective approach is Google’s page rank algorithm: matching results are ranked based in part on linkages throughout the Web, capturing information about what people deem to be most important.

---

19 The OCLC FRBR Work-Set algorithm is available through a public license. [http://www.oclc.org/research/projects/frbr/algorithm.htm](http://www.oclc.org/research/projects/frbr/algorithm.htm)
Choosing a text-based discovery system enabled us to provide relevance-ranked results based on how closely the content of each bibliographic record matches the query. (We will refer to this method of ranking throughout the report as “content ranking”.) In addition we were interested in considering whether additional information about holdings and circulation might be applied to enhance this basic content ranking capability. In theory, these data represent information about what users and curators deem to be most important.

6.1 Data sources

6.1.1 Circulation data
Two UC campuses supplied circulation data for the purposes of experimentation. From UCLA, we obtained two data sets: one set of approximately 7.7 million circulation transactions recorded between September 1999 and July 2004, and a second set of approximately 1.6 million transactions recorded between July 2004 and June 2005. From UC Berkeley, we obtained a set of approximately 11 million circulation transactions recorded between July 1998 and May 2005.

There were significant differences in the structure and quality of the three data sets. The two sets from UCLA both contained anonymized, but persistent, patron identifiers. Although they contained no personally identifiable information, it was possible to see individual patterns of usage. There were some data quality problems: for example, the earlier dataset contained erroneously recorded renewal transactions (it was possible to apply a rough filter to remove most of these from the set).

The records from UC Berkeley contained no persistent patron identifiers; hence it was possible to see usage patterns for a particular item, but not to construct individual patron usage patterns. Again there were data quality problems; for example, there were several months of completely missing data. It was not possible to filter or otherwise compensate for this problem, but given the scope of the dataset the missing months were not critical.

The UC Berkeley dataset was usable for the purposes of enhanced relevance ranking (which did not require an understanding of individual usage patterns), but not for the recommending portion of the project (which hinges on individual usage patterns). When we opted to limit the test bed to a single campus, we chose to use the UCLA circulation data.

6.1.2 Holdings data
We first obtained holdings counts from the Research Libraries Group (RLG). The data were keyed on ISBN, and generated to produce counts for holdings by FRBR work group, rather than at the individual bibliographic record level. After considerable work with the data, we opted not to use it for two reasons. First, using ISBN as a key meant that we could only link holdings counts to records containing a valid ISBN (about 62% of the collection). And second, given the complexities of producing FRBR work groups, about 15% of the RLG data consisted of records with conflicting counts (same ISBN,
different holdings counts). Given the dual issues of coverage and conflicting counts, we opted to investigate other sources of holdings data.

Next, we pursued holdings data from the Online Computer Library Center (OCLC). OCLC provided holdings counts for the 10 University of California, keyed on OCLC number, as well as an aggregate count for all institutions represented in the WorldCat catalog\textsuperscript{20}. We were able to match on OCLC number for a much greater percentage of the collection (about 89%); and the counts were not grouped, eliminating the conflict present in the RLG data.

We weighed the use of WorldCat-wide holdings counts against the UC-wide holdings counts. We found that the WorldCat-wide counts appeared to be skewed away from holdings of interest to users of an academic research library; for example, one of the most widely held items in WorldCat on the topic of geology is a volume on volcanoes for the primary school level. Given this bias, likely due to the diversity of OCLC member institutions, we opted to use the UC-wide counts to generate the weights for use in our experiments.

\section*{6.2 Generating and applying boost factors}

To test our hypothesis that holdings and circulation data could improve relevance ranking, we built into our prototype the ability to apply “boost factors” generated from holdings and circulation data to augment basic the content ranking method built into Lucene. The “boost factor” does not supplant the essential content ranking or affect what is or is not retrieved; it simply increases the relevance score for items that are widely held or frequently circulated.

Boost factors are numbers in the range of 1 or more; they are multiplied with the basic content score as a final step in the calculation of relevance rank. The boost factors are stored as text on the file system, and read into memory for use at run time. This strategy allowed us to re-generate boost factor files as needed, and switch between different sets of boost factors from query to query for easy comparison.

The goal was to set the weights in such a way that the raw content scores are not completely overwhelmed by the boost factors. So, for example, consider two records returned on a query. Imagine that both items have a content score of 4. If one has a boost factor of 1.25, and the other has a boost factor of 2, the total scores are 5 and 8. The latter record is therefore returned higher on the list.

\subsection*{6.2.1 Circulation boosts}

To enable efficient generation of circulation weights, we constructed a summary database table for all circulating items and populated it by querying the transaction tables to generate summary counts: for example, total transactions, total renewals, and total patrons for each bibliographic record. We then used simple SQL queries to generate text documents containing weight factors for each of the bibliographic records in the table.

\footnotesize\textsuperscript{20} WorldCat is a worldwide union catalog, containing records for more than 9,000 institutions: http://www.oclc.org/worldcat/
The default boost factor for non-circulating items is 1 (essentially leaving the raw content score intact for that record). After some experimentation, we generated circulation boost factors using a very simple log function:

\[ 1.25 + \log_{10}(n) \],

where \( n \) is the total number of charges (check-outs) for the item. An item circulating at least once will thus have the minimum circulation boost factor of 1.25. The nature of the log function brings most values into a very narrow numeric range, with the majority of circulating items falling between 1.25 and 3. Only 80 items, out of close to 1 million, have a boost factor of 4 or more.

### 6.2.2 Holdings boosts

Another table in the database contained holdings counts for all bibliographic records that could be matched against the OCLC holdings data. We used simple SQL queries to generate text documents containing weight factors for each of the bibliographic records in the table.

The default boost factor is 1 (essentially leaving the raw content score intact for that record). We generated circulation boost factors using a very simple log function:

\[ 1 + \log(n) \],

where \( n \) is the total UC-wide holdings for the item. Records with no available holdings data were assumed to have a holdings value of 1. As with the circulation data, the log function brings most values into a very narrow numeric range between 1 and 3.

### 6.3 Assessment

#### 6.3.1 Brief literature review

The information-seeking literature examining people in academia has studied the nature of users interacting with information systems to solve problems at many different levels. On one level is the work of researchers who have focused on how people are trying to make meaning as they move through their lives and how information practices are a piece of that (Dervin, Suchman, Kulthau, etc.).

At the next level of specificity are researchers looking at the types of general tasks people engage in. In terms of academic users, this literature regularly identifies the following types of overarching goals (Friedlander, Brockman, Norman, Borgman, etc.):

- research
- teaching
- coursework
- networking and staying current

Research and teaching are the most commonly identified primary goals. In addition to being independently central to the work of academics, these goals are highly interconnected as are their subtasks, such as: reading, writing, searching and manipulation of primary resources. The literature did not mention communication between scholars, and yet this seems a logical addition, especially in support of networking and staying current.
Activities in which users engage in order to find information in support of any or all of these goals has traditionally been referred to as “Information-seeking behavior.” Information-seeking behaviors are the very precise sorts of tasks or practices that users employ, such as searching, browsing, and chaining. These tasks are used separately and together to achieve different types of results, from finding a known item to getting a sense of a subject area.

The CDL MetaSearch team identified a compelling set of basic “information-finding tasks” that sit in the middle between an individual's larger goal (such as engaging in research) and the detailed level “unit” task (such as browsing). These tasks reflect the starting point of a user coming to a system:

- **Known Item:** “I want a copy of Hamlet”
- **Exploratory:** “I have a little idea of something but I want to get some feedback and guidance on it.”
- **Decision-Making:** “Is it better to start my business in San Francisco or Oakland?”
- **Exhaustive:** “I want to find everything I possibly can on this topic”
- **A Few Good Things:** “I have a paper due tomorrow and I need three good citations.”
- **Growing the Pearl:** “I have a good bibliography so I want to find these items and then explore their bibliographies.”

The history of evaluating relevance ranking and information-seeking grounded the assessment team's desire to have an approach that was both task and discipline based in order to approximate as much as possible the environment in which users actually conduct their academic work. Scenarios were turned to as a way to meet this goal while also drawing on the efficiency of a lab environment. Borlund (2000) found that in the evaluation of information retrieval systems, user-supplied tasks and well-crafted supplied scenarios work equally well. In addition, the CDL has a successful history of using scenarios in past evaluation efforts.21

### 6.3.2 Objectives

The goal of this study was not to determine how well different ranking methods performed against a pre-selected set of “right” answers. Instead, the purpose was to understand which of the methods best supported users in their efforts to accomplish a typical academic task that most likely would require the use of a library's online catalog.

Three basic research questions were at the center of this study, with the factor of domain expertise being addressed in each:

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

21 Prior evaluation efforts incorporating this approach are discussed in the CDL Assessment Program Toolkit. [http://www.cdlib.org/inside/assess/](http://www.cdlib.org/inside/assess/)
Data was collected for each of these questions by having users complete questionnaires on their experiences and opinions regarding relevance ranking; having users conduct searches on academic topics in their discipline and rate the usefulness of items in the top 20 positions in the result set; and by using a think-aloud protocol throughout the assessment process.

6.3.3 Ranking methods evaluated

Four different ranking methods were evaluated. Three of these are based on the content ranking capabilities built into the retrieval system, and so are closely related (see details in “Enhanced relevance ranking”, p. 15, and “Generating and applying boost factors”, p. 17):

- content ranking only (referred to as the Content Ranking method)
- content ranking, boosted by circulation data (referred to as the Circulation method)
- content ranking, boosted by holdings data (referred to as the Holdings method)

As a quasi-control, the fourth method (referred to below as the “System ID” method) produced result sets sorted by system ID only, because it most closely approximates the ordering method used in Melvyl. Although we considered methods using boost factors based on blended circulation and holdings data, we opted to focus on those factors separately in order to reduce the number of variables. Users of both levels of subject knowledge evaluated search results from each of these systems and were blind as to which ranking method was in use at any given time.

6.3.4 Methods

Task-based approach

A task-based approach was taken for user evaluation of both the relevance ranking and recommendation features of the prototype system. The goal of these features is to support academic users as they employ the library catalog to aide them in their academic work. As such, it is important to test the ability of these prototype systems to accomplish that goal. Individuals are looking for items that meet particular needs grounded in efforts to accomplish particular tasks. The knowledge, needs and interests that a person brings to the process of accomplishing a particular task extend to the process of conducting information searches and evaluating items for usefulness that are the result of such searches.

Optimally, we would want to observe individuals as they worked on an academic task from beginning to end, for example from the point at which they were assigned a paper topic to the point at which they submitted that paper. We would want to understand how the library’s catalog was used in all of the different phases of producing that paper, and how the particular prototype features played a role.

Such an approach is very resource intensive and is not geared towards the relatively rapid feasibility assessment goal of this project. A compromise approach was devised in which a small sample of ten users was recruited to conduct searches related to supplied tasks in
their fields. Discipline-based tasks were developed in History and all of the major Humanities sub-fields, fields that we knew were well-represented in our test database. Tasks were created out of recent course listings to ensure that they reflected current areas of focus and were pre-tested to ensure sufficient coverage in the test document collection (see Appendix J for the complete set of scenarios). These sessions were observed and facilitated and user evaluations of items were recorded by the prototype system. The small number of users combined with the discipline-based tasks provided the opportunity to develop a rich understanding of how academic users conducted searches related to their academic work, evaluated items, and assessed relevance.

A draft protocol was developed and tested internally. Modifications were made and a “dry run” was conducted that simulated the official evaluation environment. These practice evaluations were conducted with five individuals in separate sessions in a private room at the UC Berkeley Main Library and were focused on catching difficulties with steps in the protocol, the user interface and the recording of data. The final evaluation sessions were also conducted in a private room in UC Berkeley's Main Library.

**Target participants**

The test document collection had the richest coverage in the areas of Humanities and History, thus our recruitment efforts were focused on students in these fields. In addition, we were interested in leveraging recent user assessment work on the part of CDL in these areas, as well as the rich body of literature that exists regarding information use by humanities students. Participants were also selected on the basis of their domain knowledge, as one of the research objectives was to understand what role, if any, subject expertise played in relevance assessment. Undergraduates were considered subject-naive users and graduate students were considered subject-experts.

Recruitment was accomplished to a small degree by posting flyers in the library, but primarily through a notice on the library's website, with the added inducement of a $25 gift certificate to the student store. Ten participants were recruited for the final test and five for the dry run. A description of the participants who engaged in the final evaluation is provided in the Findings section.

**Protocol**

After being conducted into the evaluation room, participants were introduced to the observer and facilitator. They were given a brief description of the project, a $25 gift card to the student store, and a consent form which they were asked to sign if they wished to continue with the evaluation.

Once the consent form was signed, the protocol began with a set of pre-interview questions (see Appendix J) designed to elicit information about the participant's experience with systems that offer relevance ranking. These questions were asked by the facilitator and responses were recorded by the observer. Participants were then given an overview of the protocol in which they were told that they would be conducting searches related to five scenarios based on their disciplines and that they should describe their thoughts and actions throughout the evaluation. They were told that the observer would be taking notes and possibly asking clarifying questions and that the facilitator would
walk them through the entire process, starting with a warm-up scenario to orient them to the system.

Participants were instructed to ask any questions they wished, with the caveat that some might not be answerable until the session was completed. They were assured that the purpose of the study was to evaluate the system, not their search skills, that all feedback, positive and negative was welcomed, and as the assessment team had not designed the system, there would be no hard feelings regarding critical comments.

At that point, participants were asked to turn to the computer monitor, where they worked through a set of searches using the prototype. (See Appendix K for a series of screen shots and descriptions). Participants were assigned to groups, A for subject-naive and B for subject-expert, although the meaning of the letters was not revealed to them.

As touched on previously tasks were folded into a larger scenario, were based in the participant's discipline, and were differentiated by subject-expertise. For instance a subject-naive user might be asked to look for resources for a paper on a particular item, while a subject-expert user might be asked to look for items to include in a literature review. Tasks were loosely described to allow for refinement by participants in order to make them as meaningful as possible to individual users. Scenarios remained constant across the relevance ranking and recommender prototype systems.

Invisibly to users, as they progressed through the set of scenarios they were rotated through a randomly assigned order of ranking methods. These assignments were balanced to ensure that orders were distributed so that one method was not always the first or last used.

6.4 Findings

6.4.1 Participants

Ten individuals participated in the evaluation of the relevance ranking feature. Users were evenly divided across expertise level, with five undergraduate (subject-naive) and five graduate (subject-expert) students. As shown in Table 2, a small cluster of users were History majors, all from the undergraduate group. Two out of the five graduate students were studying Comparative Literature. In addition, each expertise-level group had three male and two female participants. The subject-naive group included two seniors, two juniors and one freshman. The subject-expert group included two fourth-year students, two third-year students and one second-year student.
Table 2. Participants by Expertise and Subject area

<table>
<thead>
<tr>
<th>Expertise Level</th>
<th>Discipline</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Subject-Naïve</strong></td>
<td>English</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>History</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Rhetoric</td>
<td>1</td>
</tr>
<tr>
<td><strong>Subject-Expert</strong></td>
<td>Comparative Literature</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Italian Studies</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Music</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Spanish</td>
<td>1</td>
</tr>
</tbody>
</table>

6.4.2 Experience with relevance ranking

Although a requirement of participation was experience with online library catalogs, an understanding of or familiarity with relevance ranking was not. In order to provide a context for analyzing how users assess relevance, participants were asked a few questions before the evaluation regarding their self-identified experience with systems that return ranked results. For this purpose, we described relevance ranking as a result set ordered from best to worst.

Eight of the participants, five subject-naïve and three subject-expert, said that they had used one or more online systems that seemed to return results in an order sorted from best to worst. Of these, five gave Melvyl or Pathfinder (UC Berkeley's local catalog) as examples, when in fact these systems do not employ relevance ranking. One of those users also identified Google, which of course does use relevance ranking. One user mentioned the Los Angeles Public Library’s system as one which ordered items by relevance.

Two users referenced chronological sorting of results, with one explaining that this presentation, with the most recent items appearing first, was an ordering of best to worst. This corresponds to observations from the think-aloud protocol discussed later, that publication date is very important to users in determining an item's relevancy to a particular task.

Most users who said that they had experienced a best-to-worst ordering of results felt that it was helpful, although it varied. One user said that it depended upon the topic. Another mentioned that while it was helpful, it wasn't “that great,” in that the most relevant items often did not start appearing in the result set until the third or fourth item. These differences of opinions point to varying expectations of performance.

Two users, both subject-experts, were unsure if they had used a system that ordered results from best to worst. One mentioned Amazon and Google as possibilities, but wasn't sure. Another said quite bluntly, “No, not that you can tell,” despite being familiar with Amazon and Google, which indicates that the relevance ranking these services do use is not well-mapped to that person's needs.
Amazon was given as an example of a relevance ranking system to these users, each of whom indicated that they had used Amazon in the past. One of the participants discussed not liking Amazon’s ordering because it did not reflect academic interests and needs, in part because the items that individual was interested in finding were often out of print. This user also explained that having an understanding of how Amazon’s system works is helpful in order to “get around” and identify desired items, but that “persistence” was required.

The second participant mentioned that Amazon results appeared to be ordered by popularity as opposed to relevance. This comment bears on how an item should be defined as relevant. In our study, circulation and holdings data were used to provide relevance ranking, both of which can be construed as proxies for popularity. Such a ranking method may not be effective for the academic needs of this and other individuals.

The comments of these two users indicate how search aspects of academic tasks, at least for individuals highly knowledgeable about their field, are distinct from search components of more general activities. For instance, the fact that out of print items are harder to find on Amazon is not surprising, as Amazon is geared towards a general consumer audience. An academic user is not necessarily interested in purchasing an item, but most definitely wants access to it, so the service is not really designed to meet that need.

6.4.3 Result set size

An unexpected complication in the analysis of the data was the varying length of result sets. These ranged from a total of 1163 items associated with the System ID method, to 6175 for the Content Ranking method. This variation made comparisons across ranking methods a bit difficult, as the ranking methods being studied only affecting the ordering of items, not the number of items retrieved. Instead, user queries, the topics and subject-expertise all determined the result set size, which ranged from one item to over three thousand. Subject-naive users accounted for 11 out of the 14 results sets that were greater than 100 items and all seven result sets greater than 300, including 4 at 1,051; 1,464; 1,464; and 3,108. By chance, a disproportionate number of these high-volume result sets were associated with one method, the Content Ranking method.

The wide variability of result set size influenced the number of items available to rate and made performance comparisons difficult. For example, participants rated 122 items returned using the System ID method and 207 returned using the contentcRanking method. Because of this discrepancy, comparisons will be based on the percentage of total items rated, as opposed to the number of total items returned. This approach does mean that a more precise evaluation of how effective a method was at pushing items to the top of an entire result set is lacking. Controlling for result set size is clearly a requirement for any future research efforts.

6.4.4 Performance against objectives

Objective 1: Looking at rated items only, which ranking method is associated with the greatest number of most highly rated items?
To evaluate success at producing the greatest number of most highly rated items we considered several factors:

- the highest percentage of items rated “Very Useful;”
- the lowest percentage of items rated “Not Useful;”
- the highest percentage of “Very Useful” items in the first quartile of the result set;
- and the relationship between domain expertise and ranking methods.

Each of the above mentioned dimensions will be discussed more completely below. This analysis will show that the three methods based on content ranking (Content Ranking, Circulation, and Holdings) almost always outperformed the System ID method. However this was not true for every measure, and the differences among methods are not large enough to completely reject the value of the System ID method. Further research will be required in order to confirm this difference and to assert which of the methods based on content ranking is preferable.

Table 3 provides a snapshot of the performance of each of the ranking methods.

**Table 3. Overall Performance of Ranking Methods**

<table>
<thead>
<tr>
<th>Method</th>
<th>Result Set Size</th>
<th>Usefulness Rating</th>
<th>No. Items</th>
<th>% Total Rated</th>
<th>% Total Result Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holdings</td>
<td>2050</td>
<td>Very Useful</td>
<td>59</td>
<td>29.35%</td>
<td>2.88%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Somewhat Useful</td>
<td>47</td>
<td>23.38%</td>
<td>2.29%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not Useful</td>
<td>82</td>
<td>40.80%</td>
<td>4.00%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cannot Determine</td>
<td>13</td>
<td>6.47%</td>
<td>0.63%</td>
</tr>
<tr>
<td>Circulation</td>
<td>1782</td>
<td>Very Useful</td>
<td>48</td>
<td>30.57%</td>
<td>2.69%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Somewhat Useful</td>
<td>40</td>
<td>25.48%</td>
<td>2.24%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not Useful</td>
<td>59</td>
<td>37.58%</td>
<td>3.31%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cannot Determine</td>
<td>10</td>
<td>6.37%</td>
<td>0.56%</td>
</tr>
<tr>
<td>Content Ranking</td>
<td>6175</td>
<td>Very Useful</td>
<td>60</td>
<td>28.99%</td>
<td>0.97%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Somewhat Useful</td>
<td>39</td>
<td>18.84%</td>
<td>0.63%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not Useful</td>
<td>95</td>
<td>45.89%</td>
<td>1.54%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cannot Determine</td>
<td>13</td>
<td>6.28%</td>
<td>2.11%</td>
</tr>
<tr>
<td>System ID</td>
<td>1163</td>
<td>Very Useful</td>
<td>32</td>
<td>26.23%</td>
<td>2.75%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Somewhat Useful</td>
<td>30</td>
<td>24.59%</td>
<td>2.58%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not Useful</td>
<td>40</td>
<td>32.79%</td>
<td>3.44%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cannot Determine</td>
<td>20</td>
<td>16.39%</td>
<td>1.72%</td>
</tr>
</tbody>
</table>
**Highest percentage of “very useful” items:**
The Circulation method was most effective in delivering the highest percentage of relevant items in the set of total rated items, with 30.57% of rated items considered “Very Useful.” The System ID method performed most poorly with only 26.23% of items receiving a “Very Useful” rating.

**Lowest percentage of “not useful” Items**
Looking at the converse criterion, the lowest percentage of “Not Useful” items, the System ID method performed best, in that only 32.79% of items associated with it received this rating. The Content Ranking method had the highest percentage of these non-relevant items, at 45.89% of total rated items.

**Quartile view— which method had the greatest percentage of “very useful” items nearest the top**
The goal of a ranking method is to order highly relevant items as close to the top of a result set as possible. Thus, looking at the percentage of such items in the rated set provides only a partial understanding of performance. It is also important to look at where in the result set these highly rated items are, with an understanding that placement nearer to the top is considered better.

Placement analysis was conducted by looking at the rated set by ordinal positions and by quartiles. Quartiles seemed to be a more effective way of comparing performance, as the distinction between items appearing in the first or second quartile, for instance, is clearer than between the first and second position in a result set.

Table 4 presents the percentage of total rated items rated “Very Useful” as they fall within each quartile. From this it can be seen that the Holdings method outperforms other methods, with 10.95% of items in the first quartile considered “Very Useful.” This is followed closely by the Circulation method at 10.19% and contrasts favorably against the Content Ranking method, which performed most poorly with only 8.70% of such items in the first quartile.

When disaggregated by expertise, the performance shifts. For subject-naive users, the Circulation method is strongest, with 17.24% of rated items considered “Very Useful” in the first quartile. The Holdings method is a close second, at 14.14%, while the Content Ranking and System ID methods perform quite poorly at 5.00% and 6.52% respectively.

For subject-experts, the System ID method surprisingly performs the best, with 20% of rated items considered “Very Useful” in the first quartile. The Holdings method performs relatively poorly for these users, at only 7.84%.
# Table 4. Percentage of Rated Items Rated “Very Useful” by Quartile, Method, Expertise

<table>
<thead>
<tr>
<th>Method</th>
<th>Result Set Quartiles</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Quartile 1</td>
<td>Quartile 2</td>
<td>Quartile 3</td>
<td>Quartile 4</td>
</tr>
<tr>
<td>Holdings</td>
<td>10.95%</td>
<td>9.45%</td>
<td>5.97%</td>
<td>2.99%</td>
</tr>
<tr>
<td>subject-naive</td>
<td>14.14%</td>
<td>12.12%</td>
<td>10.10%</td>
<td>5.05%</td>
</tr>
<tr>
<td>subject-expert</td>
<td>7.84%</td>
<td>5.88%</td>
<td>1.96%</td>
<td>0.98%</td>
</tr>
<tr>
<td>Circulation</td>
<td>10.19%</td>
<td>8.92%</td>
<td>8.92%</td>
<td>2.55%</td>
</tr>
<tr>
<td>subject-naive</td>
<td>17.24%</td>
<td>13.79%</td>
<td>12.07%</td>
<td>3.45%</td>
</tr>
<tr>
<td>subject-expert</td>
<td>6.06%</td>
<td>6.06%</td>
<td>7.07%</td>
<td>2.02%</td>
</tr>
<tr>
<td>Content Ranking</td>
<td>8.70%</td>
<td>5.31%</td>
<td>8.21%</td>
<td>6.76%</td>
</tr>
<tr>
<td>subject-naive</td>
<td>5.00%</td>
<td>5.00%</td>
<td>9.17%</td>
<td>8.33%</td>
</tr>
<tr>
<td>subject-expert</td>
<td>13.79%</td>
<td>5.75%</td>
<td>6.90%</td>
<td>4.60%</td>
</tr>
<tr>
<td>System ID</td>
<td>9.84%</td>
<td>10.66%</td>
<td>3.28%</td>
<td>2.46%</td>
</tr>
<tr>
<td>subject-naive</td>
<td>6.52%</td>
<td>10.87%</td>
<td>4.35%</td>
<td>3.26%</td>
</tr>
<tr>
<td>subject-expert</td>
<td>20.00%</td>
<td>10.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
</tbody>
</table>

Distribution of the highly rated items is another dimension upon which to evaluate the ranking methods. Table 5 provides a comparison of the percentage of items rated “Very Useful” appearing in each quartile of the result set. This data shows that the Holdings and System ID methods had the highest percentages of “Very Useful” items falling in the first quartile (37.29% and 37.50% respectively). The poorest performing method, Content Ranking, still had nearly a third of these highly rated items in the first quartile, thus the difference between these is not that great.

The performance for users at different levels of domain expertise does not change when looking at distribution across quartiles. The Circulation method pushes more “Very Useful” items into the first quartile for subject-naive users (37.04%), while the System ID pushes more for subject-experts (66.67%). The performance of the System ID method is quite striking and may be a result of the limited number of total items rated.
Table 5. Distribution of Items Rated “Very Useful” Across Quartiles, by Method and Expertise

<table>
<thead>
<tr>
<th>Method</th>
<th>Result Set Quartiles</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Quartile 1</td>
</tr>
<tr>
<td>Holdings</td>
<td></td>
</tr>
<tr>
<td>subject-naive</td>
<td>37.29%</td>
</tr>
<tr>
<td>subject-expert</td>
<td>47.06%</td>
</tr>
<tr>
<td>Circulation</td>
<td></td>
</tr>
<tr>
<td>subject-naive</td>
<td>33.33%</td>
</tr>
<tr>
<td>subject-expert</td>
<td>31.58%</td>
</tr>
<tr>
<td>Content Ranking</td>
<td></td>
</tr>
<tr>
<td>subject-naive</td>
<td>30.00%</td>
</tr>
<tr>
<td>subject-expert</td>
<td>18.18%</td>
</tr>
<tr>
<td>System ID</td>
<td></td>
</tr>
<tr>
<td>subject-naive</td>
<td>26.09%</td>
</tr>
<tr>
<td>subject-expert</td>
<td>66.67%</td>
</tr>
</tbody>
</table>

Relationship between domain expertise and ranking methods

Different ranking methods appear to work better for users with different levels of domain expertise, as shown in Table 6. For subject-naive users, the Circulation method consistently proved most effective, being associated with 46.55% of “Very Useful” items and only 18.97% of “Not Useful” items. For subject-experts, the Content Ranking method is associated with the highest percentage of “Very Useful” items at 31.03%, while the System ID method is associated with the lowest percentage of “Not Useful” items at 26.67%.

Although different methods are tied to strong performances for each user group, two out of those three methods are Content Ranking based methods, which indicates that this approach is likely to be superior when tested with a larger user population.
Table 6. “Very Useful” and “Not Useful” Ratings by Method and Domain Expertise

<table>
<thead>
<tr>
<th>Method</th>
<th>No. “Very Useful” Items</th>
<th>% Total Rated</th>
<th>No. “Not Useful Items”</th>
<th>% Total Rated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holdings</td>
<td>59</td>
<td>29.35%</td>
<td>82</td>
<td>40.80%</td>
</tr>
<tr>
<td>subject-naive</td>
<td>42</td>
<td>42.42%</td>
<td>27</td>
<td>27.27%</td>
</tr>
<tr>
<td>subject-expert</td>
<td>17</td>
<td>16.67%</td>
<td>55</td>
<td>53.92%</td>
</tr>
<tr>
<td>Circulation</td>
<td>48</td>
<td>30.57%</td>
<td>59</td>
<td>37.58%</td>
</tr>
<tr>
<td>subject-naive</td>
<td>27</td>
<td>46.55%</td>
<td>11</td>
<td>18.97%</td>
</tr>
<tr>
<td>subject-expert</td>
<td>21</td>
<td>21.21%</td>
<td>48</td>
<td>48.48%</td>
</tr>
<tr>
<td>Content Ranking</td>
<td>60</td>
<td>28.99%</td>
<td>95</td>
<td>45.89%</td>
</tr>
<tr>
<td>subject-naive</td>
<td>33</td>
<td>27.50%</td>
<td>54</td>
<td>45.00%</td>
</tr>
<tr>
<td>subject-expert</td>
<td>27</td>
<td>31.03%</td>
<td>41</td>
<td>47.13%</td>
</tr>
<tr>
<td>System ID</td>
<td>32</td>
<td>26.23%</td>
<td>40</td>
<td>32.79%</td>
</tr>
<tr>
<td>subject-naive</td>
<td>23</td>
<td>25.00%</td>
<td>32</td>
<td>34.78%</td>
</tr>
<tr>
<td>subject-expert</td>
<td>9</td>
<td>30.00%</td>
<td>8</td>
<td>26.67%</td>
</tr>
</tbody>
</table>

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

In addition to looking at percentages of rated items that were considered highly relevant, ranking methods can be evaluated along other lines in terms of how effective they are in support users in completing academic tasks. Those aspects include:

- consistency across scenarios in producing highly relevant items
- time spent searching
- number of search iterations
- user preferences of methods
- performance with known item searches

As will be discussed below, the Content Ranking method stands out among the other methods in its performance against the first three of these criteria, and no method is superior for the latter two.

Methods not associated with any highly relevant items

The System ID method was associated with two scenarios with no “Very Useful” items, meaning it “failed” just over 15% of the time. The Circulation method “failed” for one scenario, about 8% of the time. These instances accounted for all of the scenarios with
no highly rated items. Overall, the Content Ranking based methods were more consistently associated with highly relevant items across scenarios.

**Time spent searching**

In general, it is preferable to spend a minimal amount of time on a search, assuming that the search leads to useful items. Ranking methods associated with the least amount of search time can be considered to be performing better. Excluding failed scenarios described above, users on average spent less time searching when the System ID method was in play (just under 5 minutes). However even with these scenarios added back in, the System ID method is still associated with shortest averaged time searching. For this measure to be used more meaningful, a means for accounting for failed searches would have to be developed.

The results change when disaggregated by domain expertise. Subject-naive users had a wide range of times associated with each method, from a low of 4:46 minutes for the System ID method to a high of 8:27 for the Holdings method.

No method stood out as significantly better for subject-experts, as they spent a consistent amount of time (just over 6 minutes) regardless of ranking method.

**Number of search iterations**

Similar to time spent searching, the number of search iterations a user feels compelled to create is another useful aspect for evaluating a ranking method. As a general statement it can be said that the fewer searches the better. Against that standard, the Content Ranking method performed best for the combined group of users, with 48 total searches and an average of four. The System ID method performed most poorly, with a total of 76 searches and an average of six.

Disaggregating by subject expertise, the Holdings method performed best for subject-naive users, with a total of 20 searches and an average of four. The Content Ranking method performed best for subject-experts, with 19 total searches and an average of three. The System ID method performed most poorly for naïve and expert users, with total searches of 30 and 35 and averages of four and seven respectively.

**Users distinguishing between ranking methods**

Users did not indicate in any way that they were aware of differences in ranking methods. They did not identify any particular patterns or qualities with any of the methods. They were however, very aware of their need to sort by date or format.

**Known item searches**

The final scenario in the protocol instructed users to locate a particular book using supplied information, which included title, author and publication date. Difficulty finding known items has been reported, so the research team was interested in whether or not ranking methods facilitated such a task. Table 7 provides a listing of the searches associated with this scenario.

Comparing results by time spent searching (less is better) and result set size (one, representing the desired item), results showed that ranking method and domain expertise were immaterial. The primary factor that went into successful searches was the quality of
the query, specifically combining terms from the title and author values, no matter which fields were searched.

**Table 7. Known Item Searches**

<table>
<thead>
<tr>
<th>Method/Expertise</th>
<th>Time Spent (Min:Sec)</th>
<th>Result Set Size</th>
<th>Final Query</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holdings/Expert</td>
<td>01:02</td>
<td>1</td>
<td>Keyword=Against All Enemies AND Author=Richard A. Clarke</td>
</tr>
<tr>
<td>Holdings/Expert</td>
<td>03:15</td>
<td>38</td>
<td>Keyword=against+all+enemies</td>
</tr>
<tr>
<td>Circulation/Naive</td>
<td>01:12</td>
<td>1</td>
<td>Keyword=enemies against clarke</td>
</tr>
<tr>
<td>Circulation/Expert</td>
<td>01:30</td>
<td>1</td>
<td>Keyword=Against All Enemies AND Author=Richard Clarke AND Year=2004&amp;Year-max=2004</td>
</tr>
<tr>
<td>Content Ranking/Expert</td>
<td>00:39</td>
<td>1</td>
<td>Keyword=against all enemies AND Author=Clarke</td>
</tr>
<tr>
<td>Content Ranking/Expert</td>
<td>02:12</td>
<td>7</td>
<td>Keyword=“against all enemies”</td>
</tr>
<tr>
<td>Content Ranking/Naive</td>
<td>04:36</td>
<td>25</td>
<td>Title-main=against all enemies</td>
</tr>
<tr>
<td>System ID/Naive</td>
<td>00:31</td>
<td>1</td>
<td>Keyword=against all enemies richard clarke</td>
</tr>
<tr>
<td>System ID/Naive</td>
<td>00:59</td>
<td>1</td>
<td>Author=Clarke AND Title-main=Against all enemies</td>
</tr>
</tbody>
</table>

**Objective 3: How do academic users judge relevance?**

**Overall conclusions**

Participants judged relevance in a variety of ways, but analysis of the data collected from the think-aloud portion of the user sessions revealed that the most highly used elements were title and publication date. As users discussed their relevance evaluations, they described “Very Useful” items as having “good” titles and recent publication dates. Titles provided a quick but important synopsis of the item and participants asserted that more recently published items were of greater value to them, which made the publication date field extremely important.

Although title and publication date were the most important fields, participants wanted to see as much information about each item as possible in order to assess relevance. Users consistently attempted to use more descriptive bibliographic elements such as subject headings and notes as substitutes for the tables of contents, indexes and excerpts they have easy access to in commercial services.
Significant bibliographic elements

From the users’ point of view, the most important elements of a bibliographic record in determining relevance were the following: title, subject heading(s), author, and date. Also potentially useful were notes fields. Users examined these fields to gain a sense of what an item was about. They began with the title, and if it was not sufficient to make a decision, then they looked at the subject headings.

Users reported looking at subject headings for the following reasons:

A. They provide guidance to users who are not sure what subjects to use.
B. They may give an indication of whether or not an item is an overview.
C. They may contain known names.
D. They may contain the user’s query terms.
E. They are good for when you want to “spin-off” and explore.
F. They can help users narrow a search.
G. They help users determine the relevance of items.

Subject headings were not a panacea, however, as many items did not include subject headings, and often those that did had subjects that were too general.

In terms of publication dates, users felt that newer books provided a better overall perspective, because there was “more scholarship to [them].” One graduate student stated that she looks at the date field to see if an item is current, “which is really important” because more recent items address contemporary issues of interest. Because users preferred newer titles, they marked them as more relevant than older titles. One user did not understand why the older of two editions of the same book was ranked higher in his search results. In a different but related vein, two users mentioned wanting to know what time period books covered. For them, publication year was not enough.

The author field was particularly important to graduate students. This was especially true for smaller disciplines where there are established authors. Recognizing these authors “is an instant ‘in’”. Two users expressed a desire for a “find similar authors” feature. One graduate student reported, “I really go through word-of-mouth more than what I get in a search.” If she sees an author she knows, that’s most helpful.

Location of query terms in the record

The location of the user’s query terms within a record was an important factor in determining relevance. To this end, users found the highlighting of search terms in the bibliographic record extremely helpful. If the terms were in the title, then the item was judged more relevant than one whose title did not include the terms. Furthermore, items with a greater number of the user’s query terms were more likely to be considered relevant. In some cases, proximity of query terms to each other was also important. “[This record] has my keywords back-to-back, which is what I wanted in this case.”

Notes field as a proxy for Amazon’s “Inside the Book”

Participants reported liking Amazon's “Inside the Book” feature and as such were noticeably pleased when a bibliographic record contained a rich notes field that included information regarding the table of contents. Once a user had encountered such a record,
he or she repeatedly looked to the full display of items hoping to find this level of descriptive information.

**Domain knowledge differences**
The primary difference between users of different levels of domain knowledge was the use of the author field. Knowledge of an author and that author's particular scholarly perspective was important in determining relevancy, both positively and negatively. For example, one subject-expert participant noted that a book was “Not Useful” even though it was on the topic area, because she did not care for the work of that author. Such discretion was not mentioned by or observed for subject-naive participants.

### 6.4.5 Summary of major findings
Evaluating performance by the number and placement of highly rated items, the number of poorly rated items, and ability to successfully complete tasks, the three methods based on content ranking, as a group, tended to outperform the System ID method. Among those three methods, the Circulation boost was generally superior to the other methods.

With regard to domain knowledge, the three methods based on content ranking were consistently better for subject-naive users. For subject-experts, the content ranking based and System ID methods were all associated with positive results along different dimensions of evaluation, prohibiting the ability to draw strong conclusions for these users.

Users drew heavily on titles and publication dates to determine relevance, tending to highly rate items with on-topic, descriptive titles and recent publication dates. Full records were used as surrogates for Amazon's "Inside-the-book" feature. Highlighted search terms within the record were particularly helpful to users in their attempts to assess the quality of an item. As distinct from subject-naive users, subject-experts also relied on knowledge of authors in their field.

### 6.5 Evaluation of assessment methods
The adopted assessment method was quite successful in some dimensions, but was limited in others. These positive and negative aspects are detailed below, followed by a discussion of suggested modifications to the assessment method.

#### 6.5.1 Strengths
The most significant strength of the assessment method was that it was grounded in the academic context of the targeted user population. Participants worked their way through a set of scenario-based searches that were drawn from the academic discipline in which they were currently studying. Each scenario included one question that was tied to a course that had been recently taught in the associated department at UC Berkeley.
Below is an example of one of the discipline-based tasks given to a subject-naive user majoring in History:

You are an undergraduate majoring in History. You are taking a course on the history of apartheid in South Africa. You are starting to work on a paper on the dismantling of apartheid. Find some items in the class subject area that would help you with this.

Appendix J includes the full set of scenarios. Feedback from participants confirmed that the questions in the task were authentic and meaningful.

In addition to the tasks being meaningful to users, the disciplines and the tasks in the scenarios were also successfully tailored to fit the test document collection. Ensuring that the disciplines and assessment tasks had a rich body of associated materials meant that users were able to iterate and refine a given search or employ very different search strategies for a given task. While not all searches resulted in result sets that users were happy with, the corpus was sufficiently broad and deep within the constrained set of domains to allow users to employ their typical search practices.

A byproduct of this assessment approach was the development of an extensive set of reusable tasks and associated queries. Because the scenarios, tasks and user-generated queries were judged by participants to be closely aligned with the academic work in which they typically engage, the original tasks and related queries can now serve as a rich collection of reusable materials for future research.

The pre- and post- questionnaires effectively generated additional contextual data to round out the understanding of how individuals think of and use relevance ranking in their work. These questionnaires were informative without being burdensome to users.

Overall, the task-based, user-observed, think-aloud assessment approach was successful, enabling the identification of trends and obvious areas for future exploration.

6.5.2 Weaknesses

While much was achieved by using the assessment method as designed, it was not strong in all areas. The small number of participants eliminated the ability to confidently identify patterns and determine statistically significant outcomes. For instance, while a strong trend was observed showing that the methods based on content ranking outperformed the System ID ranking method, initial statistical analysis did not provide confirmation of this. The population was simply too small for patterns to recur with enough frequency for this kind of analysis.

In addition to limitations in the type of analysis, the evaluation procedure was also quite time consuming for both participants and the user-assessment team. Each user session lasted one hour, and many previous hours of protocol development and logistical work in recruiting and organizing appropriate testing environments were required.

A side-effect of the small number of participants combined with the task-based, user-generated searches was the wide discrepancy between result set sizes associated with different ranking methods. Although the ranking methods themselves had no impact whatsoever on the set of documents retrieved, by chance two of the methods were used in
scenarios that produced very different numbers of search results, making it difficult to compare performance of the ranking methods.

Despite the small number of users, analysis of the data, both the qualitative observation notes and the quantitative item evaluations recorded during sessions, required a substantial amount of time. The time requirements for the quantitative analysis would quite likely remain the same if the number of participants was increased, but it would be difficult to expand the number of participants in observed sessions, because of the time required for the actual observation and for the analysis of the observation-notes afterwards.

6.5.3 Recommendations for future assessment efforts

A few basic elements stand out as offering the most improvement if they were changed. Increasing the number of participants is the first suggested modification. A follow-on effort at examining ranking methods would need to include a large enough sample of participants to allow the team to confidently conduct statistical analysis of results. Increasing the number of users would involve redesigning the protocol and the user interface so that users could participate unobserved and at their own convenience during an established evaluation period. A small number of observed sessions could be conducted in order to provide a measure of triangulation.

In addition to increasing the number of participants, the protocol would have to be modified to control for varying result set sizes while remaining true to the discipline- and task-based nature of the original assessment method. The collection of tasks and queries that were developed from the original assessment provide the potential for creating “canned” searches that users would evaluate.

Finally, substantially increasing the number of users would quite likely require reducing the total evaluation time. Pre-constructed searches, which could address the problem of varying result set sizes, would potentially eliminate a significant amount of time spent revising search queries. Additionally, cutting out the last scenario, focused on finding a “known” item, would reduce the total time spent in an evaluation session.

6.6 For further exploration

Although limitations of the assessment methods prevented us from making a strong statement as to which of the proposed ranking methods was the strongest overall, it was clear that for supporting typical academic tasks, relevance-ranked result sets are preferable to unranked result sets.

A reasonable next step would be an analysis of the availability and costs of circulation and holdings datasets over the long term. If the costs of incorporating the holdings and/or circulation data are very high, then the basic Content Ranking method with no boosts may provide enough of an improvement over unranked sets to satisfy the needs of most users. If the costs are very low, then larger-scale research to differentiate between the methods would be worthwhile.
7 Recommending

The team explored two methods of generating recommendations. The first was circulation-based, using UCLA circulation data (see “Project Overview”, p. 5) to determine linkages among items in the catalog. The second was content-based, using terms from the bibliographic records to develop queries for similar items (“more like this…”). See Felicia Poe’s excellent introduction to recommender systems22 for more background information on approaches to recommending.

7.1 Circulation-based recommendations

The data we used were appropriate for this exploration for three reasons. First, the transactions retained anonymous but persistent patron identification numbers. We could, thus, see linkages between items checked out by individuals over time, although we could not identify each individual as a specific patron. Second, the volume of transactions was very large. This is important because relatively few items are used very frequently and most others form a “long tail” of rare use; relatively few patrons are extremely active, with most others forming a “long tail” of infrequent activity. A very large volume of data amassed over time made it more likely that we could observe the patterns we were interested in. Finally, we were able to relate the circulation records to bibliographic records in our test bed of records extracted from the UC union catalog (Melvyl).

It is important to note that there were some weaknesses in the data that we could not correct. The patron identification numbers were persistent within each of the two data sets, but discontinuous between the two sets. For the purposes of our work, it is as though the second data set is populated by a different group of patrons. In addition, UCLA alerted us to a system flaw in the earlier of the two sets of circulation data that resulted in thousands of improperly recorded renewal transactions.

There are a variety of problems with applying standard Collaborative Filtering techniques to an OPAC using circulation data, including data distribution, data sparsity and patron privacy concerns23. Moreover, circulation data are not a good proxy for patron ratings. We cannot infer that a checkout is a positive rating; we do not know whether the item was a compromise based on availability; and the circulation data only reflect physical circulation activity, omitting information about use of readily available digital alternatives.

The approach we took, therefore, was not a Collaborative Filtering approach. Instead we explored a very simple approach based on a weighted graph model, with the books as nodes, and the edges formed by patrons who have checked out the books in common. The more often the books have been checked out in common, the heavier the weight on that edge of the graph. Recommendations are generated for any node in the graph by

23 Ibid p. 2 - 3.
following the edges to other items that have been checked out by the same patrons. The recommendations can quickly be ordered by sorting on the weights of the edges.

Figure 1 illustrates this approach, modeling the relationships between items A, B, C and D. The numbers on the lines connecting A with the other three items represent the number of people who have checked out both of those items. Recommendations for item A, in rank order, are items D, C and B: five people have also checked out item D, three people have also checked out item C, and 1 person has also checked out item B.

This approach yielded mixed results. There were some intriguingly good recommendations, and some wildly off-topic recommendations. A few items, probably required for large undergraduate courses, were recommended constantly and inappropriately.

We considered, briefly, eliminating items at the extremely high end of the circulation frequency distribution. But adopting a numerical cut-off ignored the fact that there may be many reasons why an item might circulate often: it may be required, or it may be popular because it is very useful within a particular domain. We opted instead to pursue a strategy of filtering, by restricting recommendations to items within the same general content area.

The initial pass at filtering used the first letter of call number class. If the item on view had a call number beginning with “P”, only items labeled with a class beginning with “P” were recommended. This approach did yield more cohesiveness, but at a cost. The most jarringly bad recommendations were eliminated in most cases, but so were many interesting cross-disciplinary linkages. Moreover, the groupings were too crude in some areas of the call number class range, and too fine in others. Interesting recommendations were being lost in some topic areas, and poor recommendations were not being filtered out in others.

In a second pass, we created groupings by general subject area using the entire call number class. They were based on Library of Congress subject groupings created by the Columbia University Digital Library Project, adjusted for UCLA records where there were gaps, and supplemented by mappings from National Library of Medicine call number range.

---

number classes to the same scheme of general subject areas. (See Appendix F for more details on subject area groupings.)

These groupings produced a content filter that resulted in more balanced recommendations: more permissive where the earlier groupings were excessively fine-grained, and more restrictive where the earlier groupings were too crude. This approach is the one we applied during user testing.

### 7.2 Similarity-based recommendations

We also experimented with a second recommending method that exploits vector-based retrieval strategies. The algorithm analyzes the content of the bibliographic metadata for the target item, chooses the most important terms in the record, and formulates a new query. Top-ranking items resulting from the new query are presented as recommendations.

While simple in theory, the number of permutations and complications to this approach are vast. There are many methods for choosing and ordering the top terms, and many approaches to formulating the new query. Moreover, bibliographic records are inconsistent. Some records are catalogued exhaustively, others are sparse. Particularly in sparse records, the choice of a single subject heading can significantly affect the choices and weights of terms. In extreme cases, this can result in unexpected results: two versions of a book, sparsely catalogued and with slight differences in subject headings, can yield very different recommendations. We experimented with various approaches to try to balance these complexities.

In our final iteration, each term of each metadata field in the source document is considered in turn. The number of occurrences of that term in the field, $tf$, is computed. Also, the total number of documents containing that term in that field, $df$, is fetched from the inverted index. Terms are filtered out if they occur in too few or too many documents (the limits are adjustable.) Next, a score is calculated for the term by multiplying $tf \times idf$, where $idf$ is the standard $\log(numDocs / df) + 1$. Finally, the score for each term is totaled across all fields it occurs in. The resulting term list is ranked by score and the top-scoring 25 terms are chosen (also adjustable.)

The chosen terms are turned into what we call an "Or-Near" query. Each term is searched in each field and document, increasing the score of documents it is found in. Documents with more terms appearing in a single field receive an extra boost. In this way, a score is calculated for each matching document, and the top 5 scoring documents are displayed to the user.

This method yields recommendations that differ significantly in character from the circulation-based recommendations. They tend to be much more homogeneous than those produced using circulation data, both within the recommendation set and with the target item.
7.3 Assessment

7.3.1 Brief literature review
Evaluating the efficacy of a recommendation feature is similar, yet distinct from the traditional methods of evaluating information retrieval systems, most basically represented by precision and recall statistics. Developers of recommendation systems want to present users with items that are topical (i.e. relevant), but are also hoping to get at more elusive aspects, such as serendipity and novelty.

While the literature on the task of evaluating recommendation systems is growing, much of the attention has been on systems outside academia, which focus on assisting users in finding movies, music, commodities or books of general interest. The typical goal of such a system is to help an individual who knows that he or she likes a certain movie or band to find other movies or bands that were previously unknown to that individual but that would also be pleasing.

Herlocker et al. (2004) provides an excellent overview and example of evaluating recommender systems, looking at questions of evaluation settings (in the lab or the field) the quality of the recommended items (the degree of novelty), and how well the user interface works (explaining recommendations). A movie recommendation service (MovieLens) was established as a test bed for this research. This work draws on the work of others, notably Herlocker et al. (2000) and Swearingen and Sinha (2001, 2002) who have closely studied the presentation of recommender systems. Swearingen and Sinha look at sites supporting commodities, such as Amazon.com and MovieCritic and identify aspects critical to the success of recommendation services, such as transparency of recommendation sources and trust in recommendations.

Literature on recommendation systems serving academic users is much sparser. The seminal work to date is from Torres et al. (2004), a team that has cross-over with the MovieLens project team. Instead of looking at movie recommendations, the TechLens group examined recommendations for academic papers, in support of faculty looking for sources to draw on in the development of publications. Comparing several types of recommendation algorithms, the study was based on a set of very precisely targeted evaluation questions well-mapped to the perspectives and needs of academic users.

This work significantly influenced the development of our assessment protocol because the work was shaped by the research from Herlocker et al. and Swearingen and Sinha, and was also focused on academic tasks. The TechLens team asked three basic questions about each recommendation, which we adopted as three major poles in our analysis: 1) usefulness of the recommendation for a particular academic task; 2) familiarity with the recommendation; and 3) the nature or quality of a recommendation.

These data were collected in partial support of four research questions:

- Do recommendations help academic users find items relevant to a research-related task?
- How do academic users determine if a recommendation is useful?
- What is the quality of a given recommended item?
- What is the quality of a given set of recommendations?
An additional question running through each of these was whether or not differences existed among individuals who have different degrees of domain expertise.

### 7.3.2 Method

**Task-based approach**

The assessment approach for the relevance ranking prototype system was also used for the recommender prototype system. All sessions were observed, and internal testing of the protocol was conducted, as well as a small “dry run” session with three individuals who ran through the process in a private room on the UC Berkeley campus.

**Target participants**

Fewer participants were used for the “dry run” since much of the protocol and process was duplicated from the previous relevance ranking sessions. Participants for the final session were recruited in the same way as they had been for the relevance ranking assessment, via a notice on the UC Berkeley library's website and the offer of a $25 gift card at the student store.

**Protocol**

The assessment protocol followed the same structure as had been used for the relevance ranking assessment, except that questions were oriented around issues regarding recommendation systems. There were only four scenarios to work through, because each one took longer to complete than with the relevance ranking assessment. Additionally, the warmup task was changed from a search regarding Benjamin Franklin's library to one about Cesar Chavez, as the items associated with Benjamin Franklin had unusually long and confusing bibliographic records that were not representative of the majority of records in the system.

The pre-evaluation session questions were modified to elicit information regarding familiarity with current online systems with recommendation services, opinions and use of those services (see Appendix J). The post-session questions aimed to get at user opinions regarding how useful a recommender service in the library's online catalog would be for helping with academic tasks (see Appendix J). Similarly, while the evaluation tasks were the same scenario and discipline-based tasks that had been used for the relevance ranking sessions, users were asked to evaluate recommended items rather than result sets.

After having the process explained and signing the consent form as described earlier, participants were directed to the computer screen and asked to step through the on-line testing protocol (see screen shots and descriptions in Appendix L). The computer session ended with a thank you screen, and users were asked to fill out the post-session questionnaire.
7.4 Findings

7.4.1 Participants
Ten individuals participated in the user assessment effort of the recommendation system. A network failure caused the system to be unavailable after the warm-up exercise for one participant, so only observational and questionnaire data exist for this individual. Additionally, one participant was reclassified from a subject-naive user to a subject-expert. Although this individual was an undergraduate student and therefore technically fell in to our category of a subject-naive user, previous to entering academia he had spent twelve years in the military as a translator in one of the languages in his domain, which positioned him simultaneously as a subject-naive user in terms of his experience working within the field academically and as a subject-expert in terms of advanced familiarity with a portion of his field. Because of this subject's background and because of domain knowledge expressed during the evaluation session, he was re-coded as a subject-expert.

Taking into account the issues described above, the participant group comprised four subject-naive users (three women and one man) and six subject-expert users (two women and four men). The user group was slightly weighted towards subject-experts.

As shown in Table 8, three participants, all subject-naive users, chose History as their domain in which to work. Three other participants, two subject-experts and one subject-naive user, chose Art History. Other disciplines had only one user each.

<table>
<thead>
<tr>
<th>Expertise Level</th>
<th>Discipline</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject-Naive</td>
<td>Art History</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>History</td>
<td>3</td>
</tr>
<tr>
<td>Subject-Expert</td>
<td>Art History</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Classics</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>English</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Near Eastern Studies</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Slavic Languages and Literatures</td>
<td>1</td>
</tr>
</tbody>
</table>

Experience with recommendation systems
All users were familiar with online systems that provided recommendations, with all but one indicating that they had used such a system. Nine of the ten users gave Amazon as an example of a system that they had previously used that they knew provided recommendations. The other participant was aware of Amazon's recommendation service even though she did not use that service. On participant had also used Amazon's wiki.

In addition to Amazon, participants described experiences with other online systems that provide various sorts of recommendations, from other online booksellers that provide...
book recommendations based on purchases, such as Half.com, to Google's search recommendations for logged in users.

**Purposes of using online systems**

All participants used online systems such as Amazon to locate books in order to buy books or to find more information about them. One user said that she doesn't use Amazon because of the cost, and instead types in “cheap books” and uses sites that compare prices. Most but not all of this use was for academic purposes. A graduate student participant mentioned using Amazon to do quick scans of subject overviews when not using Melvyl, and also as a source of more “user-friendly” materials for undergraduate courses he helps with as a research assistant.

Apart from academic needs, several individuals indicated that they also looked for music CDs and one participant indicated that he tried to buy everything he could online as opposed to in-person and therefore had received recommendations for a variety of items, such as cameras. One participant described only using Amazon for personal interests and discussed how he maintains a list of DVD recommendations to rent but not buy. One person mentioned that buying books online had previously been his only means for obtaining books until moving physically closer to a bookstore.

In addition to traditional search features, one participant brought up her use of Amazon's Wish List as a way to keep track of all items that were even remotely interesting to her, including items she had seen in recommendation sets.

**Looking at and evaluating recommendations**

Many participants mentioned that they tended to look for specific items when they went to an online commercial service. For some this rendered the recommendation feature irrelevant—they knew what they wanted and did not need recommendations for items they did not want.

Most users had not bought a recommended item because they already knew what they wanted, did not have enough time to thoroughly look at the items, or more frequently, lacked sufficient money to purchase spur-of-the-moment books.

Despite this, all but one user considered recommendation services interesting and of good quality. A few participants commented that the recommendations were well-targeted to their interests. One participant described regularly submitting item ratings even though he would not buy any items online. This was a way to influence what other people looked at or purchased.

The single user who felt that the recommendations were not good was inspired by this lack of quality to spend a significant amount of time de-selecting the list of items Amazon used to generate recommendations for her in order to “disable it.”

**Determining why an item was recommended**

Most participants felt they were able to tell why items were recommended to them by Amazon or other services. Users mentioned the fact that the item was related to another item they had purchased, that it was in the same genre or category, and that other users had purchased similar items. One individual remarked that he felt confident in the
quality of these recommendations as he had spent time testing Amazon's system by putting in books he had read and discovered that the recommended items were related to materials that he was currently reading.

**Preferences regarding recommendations**

Most users do not mind getting recommendations and some like getting them. Users cited the fact that recommendations can provide access to otherwise unknown items. One user remarked that recommendations occasionally take him someplace he had not thought of, while another user mentioned recommendations as a way to be introduced to other material by a known author or artist. This latter point is interesting in that the same result could be achieved by simply searching on an author's or artist's name, but that would presume an active interest in finding such information, whereas recommendations support a passive approach to knowledge expansion.

One user commented that although he doesn't mind getting recommendations, he is already at capacity for what he can read, so an effective recommendation system would be “too good” because he has no time to pursue new items. Several participants independently mentioned that rather than using Amazon or Amazon-type recommendations, they get suggestions from professors and bibliographies and footnotes. Friends and fellow students were brought up as either useful or not-sought-after sources of recommendation.

Two participants mentioned that recommendations would be more closely allied to their interests, of higher quality and therefore more useful if they were generated based on more than just one item. These same two participants prioritized better intellectual access to anthologies and collections over more robust recommendation services. Related to this, at different points throughout the protocol most users expressed a desire to have a table of contents, index, and book excerpts available to them in the academic library catalog.

Finally, two users mentioned concerns about privacy and government tracking, described by one as the “big brother” aspect of recommender systems. He elaborated that he knows he is being tracked and so does not try to keep anonymity, but that it was still disconcerting to go to a site and be recognized, presumably without logging in. One participant was worried that Amazon was selling lists with user preferences. Both expressed alarm at potential government intrusion and use of this information against online users.

**Sources of recommendations**

The most frequent source of recommendations across both naïve and expert users was bibliographies and footnotes from a known good item in the subject area. Three naïve users and two subject-experts identified this as their top source. Two more subject-experts identified this as their second-most common source, after faculty members.

Both subject-naïve and subject-expert users identified faculty members as the source from which they were most likely to follow through on getting a recommended item. The next highest rated source was a combination of bibliographies and footnotes from books. The source from which users were least likely to follow through on recommendations
was Amazon or other online services, followed by friends. This was true of both undergraduates and graduates, although graduate students ordered friends overall more highly than undergraduates. Interestingly, “friends” was a separate choice from “students in the same department,” which may point to more interdisciplinary needs and interests on the part of graduate students as opposed to undergraduates.

**Attitudes regarding recommendations in the library's online catalog**

As shown in Table 9, the majority of participants would use a recommendation feature for academic work and a slightly smaller but still significant number would use it for personal interests. Subject-experts indicated that they would be equally inclined to use this service for both personal and academic work, while subject-naive users were more inclined to use it for academic work, and a smaller number would use it for personal interests or for helping out friends and family.

Interestingly, no one indicated that they would not use the feature, although this was qualified by a few participants saying that they would stop if the feature did not seem to be performing well. Even the participant who did not think that the feature would be helpful or useful opted not to chose the “Would Not Use” option, indicating at least a degree of curiosity about the feature.

**Table 9. Potential Uses of a Recommendation Feature in an Online Library Catalog**

<table>
<thead>
<tr>
<th>Expertise</th>
<th>Personal Interests</th>
<th>Academic Work</th>
<th>Finding Information for Friends and Family</th>
<th>Other</th>
<th>Would Not Use</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Expert</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>8</td>
<td>10</td>
<td>6</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 10 clearly shows that participants at all levels feel a recommender feature would be useful, with seven out of ten individuals choosing “Somewhat Helpful.” The guarded enthusiasm is best explained by users' consistent statements expressing that they would be interested in this feature only if it performs satisfactorily.

The one user who marked that the feature would not be helpful at all pointed out that there were already a significant number of useful items showing up in result sets and that the more important challenge for her was developing better searching techniques and obtaining a better command of the existing system.
Table 10. Evaluations of Usefulness of Recommendations in an Online Library Catalog

<table>
<thead>
<tr>
<th>Expertise</th>
<th>Very Helpful</th>
<th>Somewhat Helpful</th>
<th>Neutral</th>
<th>Somewhat Unhelpful</th>
<th>Not Helpful At All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Expert</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>2</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Most users indicated that they would be very likely to use a recommendation feature in the catalog, although one did say that that use would discontinue if the feature did not work well. Two other users of differing subject expertise noted that they would only be somewhat likely to use such a feature. This information is captured in Table 11. One subject-naïve user indicated that a recommender service would be very helpful in working her way through the enormous amount of academic resources that are available. As she described it, sometimes she might have “some good items, but need more and there is so much to wade through that it's hard.”

Only one participant (a subject-expert), the same individual who indicated that the system would not be useful at all, indicated that she would not at all be likely to use it for academic work. However, in an earlier question, she chose personal, academic and friends/family as her choices for potential uses for a recommender feature instead of choosing the “Would Not Use” option, implying that she might in fact use the system.

Table 11. Expectations of Using a Recommendation Feature in an Online Library Catalog for Academic Work

<table>
<thead>
<tr>
<th>Expertise</th>
<th>Very Likely</th>
<th>Somewhat Likely</th>
<th>Neutral</th>
<th>Somewhat Unlikely</th>
<th>Not Likely At All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Expert</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

7.4.2 Performance against objectives

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

Recommendations are a beneficial aide for academic users attempting to find items relevant to academic tasks. Analysis of participant evaluations of recommendations indicates that despite a large number of items receiving negative evaluations, a very large percentage were highly rated. Table 12 presents the results by counts and percentages of total rated items. Users agreed or strongly agreed that 65 out of the total 222 items, just
under 30%, were useful. All users strongly agreed that 30 of the total 222 items, approximately 14%, were useful.

Subject-expert users more frequently chose the second highest favorable evaluation, “Agree,” than did subject-naive users. They also chose this more frequently then they did the “Strongly Agree” option. This is the reverse of subject-naive users, which may indicate either stricter criteria for an item being considered very useful or the need to examine items in more detail before evaluating them so positively. This trend carried through when items were negatively rated as well. Interestingly, subject-experts appeared more willing to declare uncertainty about an item, choosing “Maybe/Unsure” about 6% of the time, compared with subject-naive participants who only chose it about 2% of the time.

While a large number of items were rated positively, significantly more items were evaluated negatively by both groups. All users disagreed or strongly disagreed that 147 out of 222 of the items, just over 66%, were useful. For almost half of the recommended items, users strongly disagreed that the items were useful for accomplishing the task described in the scenario. This level of disagreement held true for subject-experts as well, with 66 of the 132 rated items being evaluated this way. Subject-naive users placed a significantly smaller number in this category, with exactly one third of the total recommended items rated by this group (30 out of 90) falling into this category.

Overall, subject-experts seemed more satisfied with the recommendations than participants with less domain knowledge, choosing to “Strongly Agree” or “Agree” that items were useful for about 31% of the recommended items. For subject-naive users, the rate was only about 26% of the items. While the majority of ratings were negative, these percentages seem very promising.

### Table 12. Usefulness Ratings By Subject Expertise

<table>
<thead>
<tr>
<th>Expertise</th>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
<th>Maybe/Unsure</th>
<th>Total Rated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve</td>
<td>16 (17.78%)</td>
<td>8 (8.89%)</td>
<td>34 (37.78%)</td>
<td>30 (33.33%)</td>
<td>2 (2.22%)</td>
<td>90 (40.54%)</td>
</tr>
<tr>
<td>Expert</td>
<td>14 (10.61%)</td>
<td>27 (20.46%)</td>
<td>17 (12.88%)</td>
<td>66 (50.00%)</td>
<td>8 (6.06%)</td>
<td>132 (59.46%)</td>
</tr>
<tr>
<td>Combined</td>
<td>30 (13.51%)</td>
<td>35 (15.77%)</td>
<td>51 (22.97%)</td>
<td>96 (43.24%)</td>
<td>10 (4.50%)</td>
<td>222</td>
</tr>
</tbody>
</table>

**Effect of recommendations on searches and tasks**

Recommendations had no observed effects on searches. Only one user mentioned that the recommendations would prompt him to create a new search. The recommended items were not in the specific area in which he was interested, causing him to believe that his original query was not well-aligned with the topic he wanted to investigate.

Recommendations did have an influence on tasks. For instance, users occasionally categorized recommendations as useful because they could serve as a means to
accomplish pieces of an overall task, such as serving as a source for other items or for helping to refine the task.

One subject-expert participant experienced one of the ultimate goals of a recommendation system—encountering a serendipitous recommendation that suggested a new dimension to the task. In this case, the user was looking for works for undergraduates to use in writing a paper on terrorism. One of the recommended items was a seminal work on culture, known and used by many disciplines. The participant mentioned that he might not have considered that aspect, and if he were truly undertaking this task, would look for an easier to understand “descendant” of this particular work, as he still felt it was too advanced for undergraduates.

**Domain expertise and role of canonical items**

Canonical items played different roles for users with different levels of domain knowledge. For subject-experts, those roles were conflicting. These users were already familiar with central items in their fields, so recommendations of such items were superfluous. However, the appearance of such items extended validity to the recommendation service, and increased the confidence subject-experts had in such a system. One participant commented that he did not need a recommendation to one of these works, but that it would be a “crime” if it were not there.

Participants with less domain expertise had a different experience, in part because they lack the expertise to recognize highly relevant recommendations. Canonical items are mostly unknown to such users, so their appearance in a recommendation set is highly useful, even though they may not be aware of the value of these items in the result set at the time. Similarly, a subject-expert pointed out that for students new to the discipline, seeing works and authors collocated in a recommendation set will reinforce reference to these sources that are made later in other academic settings, thus deepening their knowledge of the field.

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

Users evaluated recommended items in much the same way they evaluated items in result sets, but with the added touchstone of the “seed” item from which the recommendations were generated. Users wanted as much detail as possible in order to determine if a recommended item was useful, and looked at sources within and beyond the bibliographic record.

**Bibliographic record**

Within the bibliographic record, certain fields were of particular importance, specifically title, subject headings and publication dates. These fields were looked to to determine topicality and the quality and nature of the item. Generic terms in the title were used to help describe recommended items. For instance, the term “Guide” was seen as an indicator of a “Survey/Overview” type of item.

Subject headings were drawn on heavily to evaluate items, especially recommended items as the “More Details” view was unavailable for recommendation sets. Participants identified particular types of words in the subject headings in order to characterize items. For instance, the term “description” was used by some to denote a “Survey/Overview”
item. Other participants noted that a high number and diversity of subject headings indicated that an item was “Introductory” in nature. Very specific or narrow subject-headings indicated that an item was “Specialized.” Similarly, subject-naive users in particular used subject-headings to determine if items were broad enough to meet their desire for material providing overviews of the task topic.

One subject-expert user noted that the subject headings were important in evaluating the usefulness of a recommended item because he really trusted catalogers' sense of what an item was about.

Publication date was a key element for users in identifying an item's relevance, with users of all types generally preferring more recently published items. In addition to publication date, users looked to publishers to determine the authoritativeness of an item, both in describing the item as “Authoritative” and in determining how useful an item would be. University presses were almost universally taken as authority sources. Museums often were as well.

The appearance of search terms in any of the fields was used to determine how well the item covered the topic at hand. The more search terms in the record the better.

Finally, the language of an item was important to users in determining usefulness. Users would like to be able to select more than one language to filter on. Participants of both levels of domain knowledge expressed this.

**Non-bibliographic record sources**

Although an item's bibliographic record was the primary means for evaluating a recommended item, participants drew on other sources as well, including the seed item from which recommendations were generated, the sense of authority of the information or entire library system, and other recommendations.

Some participants used the seed item as a proxy for their tasks, reviewing the more detailed display of the seed item's bibliographic record before and during the evaluation of the recommendations. The goal of this close examination was to try to establish connections between the recommended item and the seed item.

Similar to the trust a user expressed in catalogers' assignments of subject headings, for one subject-expert user, the very fact of an item being recommended was an a priori assertion of quality. Such an acceptance of the authority of the information system led her initially to doubt her own expertise. This user was surprised to find a recommendation that appeared to be irrelevant, but she indicated that it must have something to do with her topic because the system had recommended it. It was not until she got further down the set of recommendations that continued to be poorly associated with her topic that she revised her judgments based on her own knowledge as opposed to the authority she had ascribed to the system.

The perceived quality in a given set of recommendations can affect assessments of recommendations in that same set or in other sets. For instance, one user commented that a previous set of recommendations that she had considered quite poor made her suspicious about the quality of the set of items associated with a different seed item. In
this case, the second set turned out to be superior, but the user went into the evaluation of those items with low expectations.

In another instance, the strong evaluation of one recommended item affected the evaluation of other items. When the user was unsure about most of the recommended items, finally encountering an item that was strongly identified as not useful made the user go back and negatively revise assessments of previously evaluated recommendations.

The position of items in the recommendation set may be significant. One user noted that the set of recommendations contained only one useful item, but that because it was at the top, that made the entire set useful.

Finally, one participant commented several times that very poor recommendations were distracting.

When recommendations were not obviously related to the seed item or the topic, users often spent significant time trying to understand why specific items were recommended. The most common strategies were to try to find similarities between the subject headings and the title terms of the seed and recommended items. Participants also drew on their own knowledge in their attempts to make logical connections, until, if none could be made, they determined that the recommendation was simply poor. One participant, a subject-expert, developed a coherent story of why the recommended items were generated, but because they were on an aspect of his topic he wasn't interested in, said that he would be prompted to revise his search in order to get more appropriate items.

Users' struggles with determining the logic of poor recommendations contrast with their reported experience of typically understanding why Amazon recommendations are generated. Not all users reported being able to make sense of Amazon recommendations, however, although most did.

Several users mentioned that they would trust recommendations more if multiple sources converged to produce a given recommendation.

**Domain knowledge makes a difference**

Unsurprisingly, domain expertise was a noticeable factor in how participants approached and accomplished the scenario-based tasks. First, knowledge of the scenario's topic area allowed participants to conceptually refine a broadly written task's scope (for example by constructing a query that would bring up more than one artist in the Cubist school of painting). Similarly, subject-expert users expressed more precise desires, for instance not wanting canonical items, because those had already been studied, but recognizing that such books would be important for subject-naive users.

Second, participants used knowledge of specific titles and authors as a starting point or a tool for narrowing down a topic. This strategy occurred more frequently with subject-experts, but occasionally transpired with participants of all sorts when they happened to have some background familiarity with the topic.

Knowledge of the field allowed users to more easily evaluate items using title and author information, as there was a greater likelihood of the user being familiar with items in the search results. Additionally, such knowledge allowed users to be more discriminating in
the items they chose for different tasks, distinguishing for example between known items that would be useful for an undergraduate introductory class as opposed to a seminar. One step removed from this was being able to filter out items that were new even to the subject-expert, but that could be asserted to be topically relevant and appropriate for different audience levels.

When users had insufficient or no familiarity with a topic, their strategy was to create broad searches and to look for items that provided general overviews of the topic. This was often successful, but occasionally subject-naive users expressed great difficulty in evaluating and describing recommended items due to their lack of subject knowledge

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

One goal of recommender systems is to suggest items that the user would not have been likely to find through traditional search processes and that the user is confident are at least worth further investigation. To measure the success of our prototype system in accomplishing this goal, we asked three questions about individual recommendations, each trying to get at different aspects of this goal—the usefulness, familiarity and nature of an item.

The first question concerned how useful the item appeared to be for addressing the task. Results from this question were discussed earlier and are captured in Table 5. Results from the answers regarding familiarity and the nature of the item are discussed below.

**Familiarity**

Users indicated their knowledge of an item along a spectrum of familiarity, from having cited the item to not being familiar with it at all. Table 13 shows the user responses to the set of possible answers, broken down by subject expertise.

Not surprisingly, subject-experts were more likely to be familiar with recommended items, choosing one of the first three responses showing familiarity almost 21% of the time as opposed to just over 3% for subject-naive users. On the whole, participants were completely unfamiliar with most of the recommended items. For subject-naive users, this was true for approximately 90% of the items and for subject-experts, this was the case for just under 70% of recommended items. This relatively high degree of unfamiliarity on the part of subject-experts is explained in part by the fact that recommendations were not necessarily within a given subject area, since they were generated from circulation data and were not restricted to specific call number classes or subject headings.
Table 13. “How Familiar Are You With This Recommendation?”

<table>
<thead>
<tr>
<th>Expertise</th>
<th>I have cited it</th>
<th>I have read it</th>
<th>I have heard of it</th>
<th>I’m familiar with the author(s)</th>
<th>I don’t know this at all</th>
<th>Total Rated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve</td>
<td>0 (0.00%)</td>
<td>2 (2.22%)</td>
<td>1 (1.11%)</td>
<td>5 (5.55%)</td>
<td>82 (91.11%)</td>
<td>90 (40.54%)</td>
</tr>
<tr>
<td>Expert</td>
<td>6 (4.55%)</td>
<td>6 (4.55%)</td>
<td>15 (11.36%)</td>
<td>13 (9.85%)</td>
<td>92 (69.70%)</td>
<td>132 (59.46%)</td>
</tr>
<tr>
<td>Combined</td>
<td>6 (2.70%)</td>
<td>8 (3.60%)</td>
<td>16 (7.21%)</td>
<td>18 (8.11%)</td>
<td>174 (78.38%)</td>
<td>222</td>
</tr>
</tbody>
</table>

The nature of an item

The third question in the protocol asked participants to describe the nature of a given recommended item choosing among such options as “Authoritative,” “Introductory” and “Survey/Overview.” The complete set of categories is shown in Table 14.

Due to system constraints users could only pick one descriptor from the set of choices. However, as more than one category could logically be assigned to a recommendation, the facilitator instructed participants to verbally indicate if there was more than one appropriate category for the item. Participants only occasionally wanted to choose more than one descriptor (for 17 items out of a total of 222 recommended items evaluated), and when they did, the observer recorded any additional choice in the session notes. Because of these multiple assignments, the total number of ratings is higher than the total number of items rated. User choices are summarized in Table 7.

Table 14. “How Would You Describe This Recommended Item?”

<table>
<thead>
<tr>
<th>Expertise</th>
<th>New to Me</th>
<th>Authoritative</th>
<th>Introductory</th>
<th>Specialized</th>
<th>Survey/Overview</th>
<th>I don’t know</th>
<th>Total Ratings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve</td>
<td>0 (13.40%)</td>
<td>9 (9.23%)</td>
<td>47 (48.45%)</td>
<td>23 (23.71%)</td>
<td>5 (5.15%)</td>
<td>97 (40.75%)</td>
<td></td>
</tr>
<tr>
<td>Expert</td>
<td>2 (1.42%)</td>
<td>25 (17.73%)</td>
<td>60 (42.55%)</td>
<td>29 (20.57%)</td>
<td>9 (6.38%)</td>
<td>141 (59.24%)</td>
<td></td>
</tr>
<tr>
<td>Combined</td>
<td>2 (0.84%)</td>
<td>38 (15.97%)</td>
<td>25 (10.50%)</td>
<td>107 (44.96%)</td>
<td>52 (21.85%)</td>
<td>14 (5.88%)</td>
<td>238</td>
</tr>
</tbody>
</table>

More items were categorized as “Specialized” (44.96% for both groups combined) by both subject-experts and subject-naïve users. For both levels of domain knowledge, the next most frequently assigned descriptor was “Survey/Overview” (21.85% for both groups combined). Users were rarely unable to assign a category, choosing the “I don’t know” option for only 5.88% of items.

Because most recommended items were unknown to participants, users drew heavily on titles, subject headings and publisher information in their efforts to describe the nature of
an item. For instance, certain types of words in the title field were indicators of certain categories, such as “Guide” being associated with the “Survey/Overview” descriptor.

Participants attempted to mine as much information as possible from subject-headings. This was especially important since only a brief display of the recommended item was available to them. As with titles, users looked for certain types of words in the subject headings, which they felt indicated a certain descriptor. For example, the term “Description” in the subject headings was used by some to indicate that an item was likely to be a “Survey/Overview” of a topic.

In addition to analyzing the words in a given subject heading, other participants evaluated the entire set of subject headings. A high number and greater diversity of headings indicated that an item was “Introductory,” while very specific subject headings indicated that the item was “Specialized.”

University presses and museums were almost universally taken as authority sources, prompting participants to describe an item as “Authoritative.”

Identifying and distinguishing between primary and secondary sources was important for some users. Indeed, one user noted the need for an additional category “Primary Source” for items that were themselves an object of study in the task described in the scenario. This participant indicated three items for which this would have been an appropriate additional descriptor. Another participant described a recommended item as a primary source and commented that it had become a “cultural artifact” in the field.

Users identified items as primary or secondary by drawing on their own knowledge and by examining the item’s title and publisher. One subject-expert commented that the recommended items he was looking at were mostly primary sources and that unless he was being exhaustively thorough, he would not be interested in these items. Since he was already aware of these items, the recommendations provided no new or supportive information for accomplishing the task.

In a slightly different use of dates, the indicated birth and death dates of an item’s author were used by one subject-expert as a way of determining whether a source was a critical evaluation or not. Ultimately the user decided that the item was a primary source, given his knowledge of the way individuals who lived in that era wrote about the topic under consideration.

**Useful, unknown items were successfully generated**

Success at achieving the goal of presenting users with items likely to be both useful and unknown can be measured by taking a cross-section of the item ratings for the two questions which asked about item usefulness and familiarity. From looking at data regarding items that users rated highly but that they did not know (shown in Table 15), it can be seen that the recommender feature was able to present useful, previously unknown items to users.

For the most positively rated items, we can see that for users of both levels of domain expertise a significant percentage of items were completely unknown (75% for subject-naive users and 57.14% for subject-experts). Almost 67% of highly rated items were entirely new to both sets of participants.
Table 15. Unknown Items with "Strongly Agree" Usefulness Evaluation

<table>
<thead>
<tr>
<th>Expertise</th>
<th>Item Familiarity</th>
<th>Total Rated “Strongly Agree”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I know the author</td>
<td>I don’t know this item</td>
</tr>
<tr>
<td>Naive</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>(18.75%)</td>
<td>(75.00%)</td>
</tr>
<tr>
<td>Expert</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>(21.43%)</td>
<td>(57.14%)</td>
</tr>
<tr>
<td>Total</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>(20.00%)</td>
<td>(66.67%)</td>
</tr>
</tbody>
</table>

Table 16 captures the nature of these highly rated, unknown items. For “Strongly Agree” items, participants were inclined to describe items as “Authoritative” (36.67% of items) or “Specialized” (40% of items). This trend was also true for subject-experts when taken as a separate group, as 50% of items were described as “Authoritative” and 42.86% as “Specialized.” Subject-naive users also identified a great number of items as “Specialized” (37.50%), but identified items as “Authoritative” or as a “Survey/Overview” with equal frequency (25% of items).

In sum, subject-experts appeared to prefer items that they deemed authoritative or specialized in nature. Subject-naive users were also interested in items with these qualities, but in addition were looking for items that provided topical overviews. Domain knowledge does seem to effect the type of material a user finds helpful.

Table 16. The Nature of Items with "Strongly Agree" Usefulness Evaluations

<table>
<thead>
<tr>
<th>Expertise</th>
<th>New To Me</th>
<th>Authoritative</th>
<th>Introductory</th>
<th>Specialized</th>
<th>Survey/Overview</th>
<th>I Don’t Know</th>
<th>Total Rated “Strongly Agree”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve</td>
<td>0 (0.00%)</td>
<td>4 (25%)</td>
<td>3 (18.75%)</td>
<td>6 (37.50%)</td>
<td>4 (25.00%)</td>
<td>0 (0.00%)</td>
<td>16 (53.33%)</td>
</tr>
<tr>
<td>Expert</td>
<td>0 (0.00%)</td>
<td>7 (50.00%)</td>
<td>3 (21.43%)</td>
<td>6 (42.85%)</td>
<td>2 (14.29%)</td>
<td>0 (0.00%)</td>
<td>14 (46.67%)</td>
</tr>
<tr>
<td>Total</td>
<td>0 (0.00%)</td>
<td>11 (36.67%)</td>
<td>6 (20.00%)</td>
<td>12 (40.00%)</td>
<td>6 (20.00%)</td>
<td>0 (0.00%)</td>
<td>30</td>
</tr>
</tbody>
</table>

The pattern observed for the most highly rated items re-occurs for the slightly less positively rated items. Table 17 provides data about the items for which users chose “Agree” that the item was useful. All users taken together were totally unfamiliar with almost 72% of these items. For subject-experts, the rate was just under 63% of the items and for subject-naive users, the rate was 100%.
Table 17. Unknown Items with ”Agree” Usefulness Evaluation

<table>
<thead>
<tr>
<th>Expertise</th>
<th>Item Familiarity</th>
<th>Total Items Rated “Agree”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I know the author</td>
<td>I don’t know this item</td>
</tr>
<tr>
<td>Naïve</td>
<td>0 (0.00%)</td>
<td>8 (100.00%)</td>
</tr>
<tr>
<td>Expert</td>
<td>4 (14.81%)</td>
<td>17 (62.96%)</td>
</tr>
<tr>
<td>Total</td>
<td>4 (11.43%)</td>
<td>25 (71.43%)</td>
</tr>
</tbody>
</table>

A change can be observed in the nature of the items for which users chose “Agree” as opposed to “Strongly Agree,” as show in Table 18. Most notable is a drop in items considered “Authoritative.” For these relatively less useful items, the “Authoritative” descriptor was only chosen 20% of the time, compared to 36.67% of the time for the more highly rated items. The more common categories were “Specialized” and “Survey/Overview,” which were chosen with equal frequency (37.14%). These choice patterns held true for subject-experts and subject-naïve users alike.

Table 18. The Nature of Items with “Agree” Usefulness Evaluations

<table>
<thead>
<tr>
<th>Expertise</th>
<th>Nature of the Item</th>
<th>Total Items Rated “Agree”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>New To Me</td>
<td>Authoritative</td>
</tr>
<tr>
<td>Naïve</td>
<td>0 (0.00%)</td>
<td>1 (12.50%)</td>
</tr>
<tr>
<td>Expert</td>
<td>0 (0.00%)</td>
<td>6 (22.22%)</td>
</tr>
<tr>
<td>Total</td>
<td>0 (0.00%)</td>
<td>7 (20.00%)</td>
</tr>
</tbody>
</table>

These results provide strong indications that the recommender feature was successful in its goal of presenting items that the user was not aware of but that seemed potentially useful.

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

Recommendation sets can be evaluated in comparison to the result sets from which they were initially generated, user evaluations of the entire set, and the evaluations of individual items within a given set.
**Result sets compared to recommendation sets**

Recommendation and result sets are overlapping, but distinct types of information constructs, both of which are designed to help individuals accomplish tasks. They can be compared for redundancy and for their respective quality in terms of supporting the user in accomplishing a task.

Out of the 222 total recommended items, only eight were actually in both the result and recommendation sets. Users commented and were occasionally frustrated when recommendation sets included items that were in the initial result set. Some individuals saw this as a positive sign, indicating the quality of the recommendation set, while others saw it as unnecessary redundancy.

Because of the way result sets were generated, the overlapping of the two sets may not be significant. Only about a quarter of items in the test collection circulate, and hence have recommendations associated with them. In order to ensure that result sets contained items with associated recommendations, for purposes of testing we used a ranking algorithm that limited result sets to circulating items. Since the recommendations were also generated based on circulation data, an overlap was to be expected. A similar artifact of the ranking method may have been that while users did not comment on expected items that were missing from recommendation sets, they did occasionally comment on missing items they thought should have appeared in the original result set.

On occasion, the initial result set seemed much better in quality than the resulting recommendation sets. This prompted one subject-expert user to comment that the recommendations were not useful in general because she would have found the same items by pursuing her usual thorough review of her result set. Given the manner in which result sets were constructed however, this may in fact not be true.

In contrast, there were several occasions when users had a result set with only one item. That item was on topic, but on its own would have been insufficient for accomplishing an academic task. Users commented on this as being something that occurs in typical searches. When encountering this in the test scenario, users expressed the hope that the recommendations generated from the single good item would lead to other items. In this situation, the recommender feature is performing a query-expansion function in a form that might be meaningful to some groups of users. In the best case, users would find additional useful items when their search strategies were exhausted. In the second-best case, recommendations might suggest new search possibilities.

Although the goal of presenting useful items is the same for result and recommendation sets, the paths to such items may sometimes conflict. One subject-expert user identified a book in a result set that was excellent for the topic, but he had already read it. Within the context of a result set it was not useful to him, so he chose to keep other items in the result set as part of the evaluation process. However, since the goal was to generate recommendations, it would have made more sense to choose the excellent but already known item, which perhaps would have had other good but unknown material associated with it. The challenge for an information system would be to allow users to distinguish easily between search result items that were inherently interesting and those that were of interest only to the degree that they could generate recommendations for other items.
Quality of recommendation sets

Table 19 displays the evaluations users made of the recommendation sets. Participants were more likely to rate a recommendation set negatively than positively, with subject-experts disagreeing or strongly disagreeing that a set was useful about 55% of the time and subject-naive users making the same evaluation about 61% of the time.

In contrast, subject-experts agreed or strongly agreed about 40% of the time that a set was useful, while subject-naive users chose these evaluations about 38% of the time. While the majority of sets were not considered useful, well over one-third were rated positively for users of both levels of domain knowledge.

### Table 19. Evaluation of Recommendation Sets
(“Overall these recommendations were useful.”)

<table>
<thead>
<tr>
<th>Expertise Level</th>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
<th>Maybe/Unsure</th>
<th>Total Rated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve</td>
<td>2 (11.11%)</td>
<td>5 (27.78%)</td>
<td>9 (50.00%)</td>
<td>2 (11.11%)</td>
<td>0 (0.00%)</td>
<td>18</td>
</tr>
<tr>
<td>Expert</td>
<td>3 (18.52%)</td>
<td>8 (29.63%)</td>
<td>6 (22.22%)</td>
<td>9 (33.33%)</td>
<td>1 (3.70%)</td>
<td>27</td>
</tr>
<tr>
<td>Total</td>
<td>5 (11.11%)</td>
<td>13 (28.89%)</td>
<td>15 (33.33%)</td>
<td>11 (24.44%)</td>
<td>1 (2.22%)</td>
<td>45</td>
</tr>
</tbody>
</table>

Interestingly, both subject-experts and subject-naive users evaluated recommendation sets more positively than they did individual items. Table 20 shows this contrast by combining choices of “Strongly Agree” and “Agree” into a single positive assessment and “Strongly Disagree” and “Disagree” into a single negative assessment. For example, while subject-naive users rated individual recommendations positively 26.67% of the time, they rated entire sets positively 38.88% of the time, a significant difference. Similarly, subject-experts rated individual items negatively 62.88% of the time, but rated entire sets negatively only 55.55% of the time. This indicates that users are more tolerant of poor quality items within a given recommendation set as long as that set also contains some solid recommendations.

### Table 20. Evaluations of Recommendation Sets and Recommended Items

<table>
<thead>
<tr>
<th>Expertise</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Item</td>
<td>Set</td>
</tr>
<tr>
<td>Naïve</td>
<td>26.67%</td>
<td>38.88%</td>
</tr>
<tr>
<td>Expert</td>
<td>31.06%</td>
<td>40.74%</td>
</tr>
<tr>
<td>Total</td>
<td>29.28%</td>
<td>40.00%</td>
</tr>
</tbody>
</table>

Quality of Items in Recommendation Sets

Tolerance for poor quality recommendations within an entire recommendation set means that some flexibility exists in the construction of a recommendation set. How much
flexibility is the question that immediately follows—how many very good or good individual items are required for a set to still be considered useful? Tables 14 through 17 provide some initial answers to this question, with breakdowns of the ratings of individual recommendations in sets of each evaluation level. Ratings are separated out by expertise.

Table 21 contains the ratings of individual items within highly rated sets. Five participants, two subject-expert and three subject-naïve, each “Strongly Agreed” that one recommendation set was useful, accounting for almost 65% if items in these sets. Of those, the minimum number of items for which users chose “Strongly Agree” was two. The maximum number of items associated with the “Strongly Disagree” evaluation was also two. For one of these highly rated sets, there were no items associated with the “Strongly Agree” evaluation, however it contained the serendipitous recommendation described on p. 46 (“Effect of recommendations on searches and tasks”). Subject-experts had a more even distribution of item ratings, whereas subject-naive users rated most items positively.

<table>
<thead>
<tr>
<th>Expertise</th>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
<th>Maybe/Unsure</th>
<th>Total Rated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>6 (54.55%)</td>
<td>2 (18.18%)</td>
<td>2 (18.18%)</td>
<td>1 (9.09%)</td>
<td>0 (0.00%)</td>
<td>11</td>
</tr>
<tr>
<td>Expert</td>
<td>5 (33.33%)</td>
<td>3 (20.00%)</td>
<td>3 (20.00%)</td>
<td>3 (20.00%)</td>
<td>1 (0.67%)</td>
<td>15</td>
</tr>
<tr>
<td>Total</td>
<td>11 (44.00%)</td>
<td>5 (20.00%)</td>
<td>4 (16.00%)</td>
<td>4 (16.00%)</td>
<td>1 (0.40%)</td>
<td>25</td>
</tr>
</tbody>
</table>

Participants chose the “Agree” option for thirteen recommendation sets. Five of those ratings came from subject-naïve users and eight came from subject-experts. Table 22 shows how the individual items within those sets were rated. For these slightly less positively rated sets, only about 40% of items were rated positively across both groups of expertise.

At a more fine-grained level, participants varied in their ratings depending upon domain knowledge. Subject-naïve users and subject-experts differed greatly in the number of items for which they chose “Strongly Agree,” with users with less domain knowledge making this choice more frequently (36% compared to 17.5%). Subject-experts chose “Agree” more frequently (22.5% in contrast to 4.00%). Subject-experts also chose “Strongly Disagree” more frequently (37.5%) than did subject-naïve users (24%).
### Table 22. Individual Recommendation Ratings for Agree Sets (13 total)

<table>
<thead>
<tr>
<th>Expertise</th>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
<th>Maybe/Unsure</th>
<th>Total Rated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve</td>
<td>9 (36.00%)</td>
<td>1 (4.00%)</td>
<td>8 (32.00%)</td>
<td>6 (24.00%)</td>
<td>1 (4.00%)</td>
<td>25</td>
</tr>
<tr>
<td>Expert</td>
<td>7 (17.5%)</td>
<td>9 (22.50%)</td>
<td>4 (10.00%)</td>
<td>15 (37.50%)</td>
<td>5 (12.50%)</td>
<td>40</td>
</tr>
<tr>
<td>Total</td>
<td>16 (25.62%)</td>
<td>10 (15.39%)</td>
<td>12 (18.46%)</td>
<td>21 (32.31%)</td>
<td>6 (9.23%)</td>
<td>65</td>
</tr>
</tbody>
</table>

Tables 23 and 24 show how the ratings of individual items are distributed for sets for which users chose “Disagree” or “Strongly Disagree” when asked how useful the set was. As would be expected, all users had an increase in the number of items negatively rated. For “Strongly Disagree” sets, however, the majority of items, 80% or above for both categories of users, are rated “Strongly Disagree.” This is in contrast to the “Disagree” sets, for which there is more of a distribution between “Disagree” (45.33%) and “Strongly Disagree” (34.67%). Additionally, the most negatively rated sets have no individual items rated most positively, while the “Disagree” sets have two such items.

### Table 23. Individual Recommendation Ratings for Disagree Sets (15 total)

<table>
<thead>
<tr>
<th>Expertise</th>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
<th>Maybe/Unsure</th>
<th>Total Rated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve</td>
<td>1 (2.22%)</td>
<td>4 (8.89%)</td>
<td>24 (53.33%)</td>
<td>15 (33.33%)</td>
<td>1 (2.22%)</td>
<td>45</td>
</tr>
<tr>
<td>Expert</td>
<td>1 (3.33%)</td>
<td>7 (23.33%)</td>
<td>10 (33.33%)</td>
<td>11 (36.67%)</td>
<td>1 (3.33%)</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>2 (2.67%)</td>
<td>11 (14.76%)</td>
<td>34 (45.33%)</td>
<td>26 (34.67%)</td>
<td>2 (2.67%)</td>
<td>75</td>
</tr>
</tbody>
</table>

### Table 24. Individual Recommendation Ratings for Strongly Disagree Sets (11 total)

<table>
<thead>
<tr>
<th>Expertise</th>
<th>Strongly Agree</th>
<th>Agree</th>
<th>Disagree</th>
<th>Strongly Disagree</th>
<th>Maybe/Unsure</th>
<th>Total Rated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naïve</td>
<td>0 (0.00)</td>
<td>1 (10.00%)</td>
<td>1 (10.00%)</td>
<td>8 (80.00%)</td>
<td>0 (0.00)</td>
<td>10</td>
</tr>
<tr>
<td>Expert</td>
<td>0 (0.00)</td>
<td>6 (14.29%)</td>
<td>0 (0.00)</td>
<td>35 (83.33%)</td>
<td>1 (2.38%)</td>
<td>42</td>
</tr>
<tr>
<td>Total</td>
<td>0 (0.00)</td>
<td>7 (13.46%)</td>
<td>1 (1.92%)</td>
<td>43 (82.69%)</td>
<td>1 (1.92%)</td>
<td>52</td>
</tr>
</tbody>
</table>
Only one participant, a subject-expert, chose “Maybe/Unsure” for an evaluation of a recommendation set.

Figure 2 shows the spread of evaluation of recommended items for variously rated recommendation sets. As one would expect, the evaluations of the individual items map to the evaluation of the recommendation sets. For example, sets for which users chose “Strongly Agree” had a greater percentage of individual recommended items for which “Strongly Agree” was also chosen. The pattern holds true for the other evaluations as well. One difference is that the proportion of items for which “Strongly Disagree” was chosen is substantially greater (82%) than the corresponding individual item/set rating for other levels (from 15.38% to 45.33%). User seem to be clearer in assigning the most negative category than any other category, perhaps indicating that it is easier to determine when an item is definitely not useful than it is to determine a particular degree of usefulness.

Figure 2. Item Ratings Associated with Set Ratings

7.4.3 Summary of major findings

Users expressed interest and enthusiasm for trying an online library catalog with a recommendation service to support their academic tasks.

Recommendations were successful in supporting academic tasks. These items were interesting to users in and of themselves and as intermediary resources in the research process. On one occasion, a recommendation helped a user consider the task from a new vantage point.

Recommended items were generally not of a novel or surprising nature. Items that users preferred tended to be described by them as "Authoritative" or "Specialized," with less knowledgeable users also preferring items that were overviews or surveys. Subject-experts were much more likely to be familiar with particular items or authors, and thus had a greater ability to evaluate items. Users evaluated the relevance of recommended
items in the same way search result items were evaluated, with the added touchstone of the "seed" or recommendation source as a comparison point. Because of their domain exposure, canonical items were not directly useful to subject-experts, however they added validity to recommendation sets.

The recommendation service was useful as a query expansion tool when users were boxed in by small or single search results.

Selecting interesting items from a list of results is subtly different from selecting an item for use in generating recommendations. In general, when participants evaluated items in a result set, they were seeking new, unread items and would by-pass known, good items. However, these known items could be excellent sources of recommendations, as they are already proven relevant and presumably would have a greater chance of being associated with other useful or high quality items. Interfaces should be designed to allow users to select items for different end purposes, in order to encourage the use of known good items as sources for generating recommendations.

Users were more likely to be satisfied with a given recommendation set than a given recommended item. Only a few items in a set of recommendations needed to be considered useful for a participant to consider the entire set useful.

### 7.5 Evaluation of assessment methods

The assessment method used with the recommender feature was very similar to the method used for evaluating the relevance ranking methods. Below are discussed the relative strengths and weaknesses of the assessment of the recommender feature, followed by some initial ideas for improvement.

#### 7.5.1 Strengths

The strengths of the recommender assessment were the same as with the relevance ranking feature. Participants felt that the discipline-based tasks were authentic and meaningful (only one participant cited a problem with the scenarios, commenting that one was too broad); the disciplines and tasks were well-covered by the test database; rich qualitative data regarding how users search and evaluate items in terms of usefulness was collected; and the pre- and post-questionnaires were a manageable length and provided information that was important to the overall analysis.

#### 7.5.2 Weaknesses

The primary weakness of the relevance ranking assessment method was also a weakness of the recommender assessment. The small number of participants eliminated the ability to confidently determine statistically significant outcomes. Trends, however, were much more easily spotted. In addition to the number of participants, there were a number of other issues that arose that were specific to the recommender assessment.

First, the number of scenarios was too great and the evaluation procedure was too long for both participants and the user-assessment team. Second, the source of
recommendations was verbally explained to participants, but as identifying this source is a primary user requirement, it needed to be explicitly labeled in the User Interface.

Third, the full bibliographic record of recommended items was not available to users. Users frequently looked at the “More Details” view of items during the search phase of the evaluation and commented on the need to see more information in order to assess a recommended item. Developing a way to display the full record without letting the screen become visually overwhelming would greatly assist users in their evaluations.

Fourth, the aspects of novelty and serendipity were not well addressed for several reasons. The question designed to focus on these aspects of an item was difficult for users to answer because describing the nature of a book with minimal bibliographic data is very challenging, especially for users with limited domain knowledge. The sense of surprise occasioned by some recommendations was almost universally due to the item being completely off-topic as opposed to its being a unique approach to the subject. Apart from one instance, there did not appear to be any recommended items that were particularly serendipitous or novel in their approach to the topic. Finally, a technical constraint forced users to select only one descriptor for this question when the protocol had been designed to support multiple answers. Participants were instructed to verbally describe any additional choices (which some did), but the descriptor that touched on novelty was rarely chosen.

Finally, dry-run participants did not sufficiently mirror the target user population. Participants used to test out the protocol and the interface were recruited from a professional program, which rarely requires traditional academic, library-based research, unlike the disciplines focused on for the actual study. This contrast did not stand out as significantly for the ranking evaluation, perhaps because these users could only approximate subject-naive users and not subject-experts. Evaluations of the recommended items were quite different between the two levels of domain knowledge, so this lack of testing may have masked the potential difficulties associated with the question regarding novelty.

7.5.3 Recommendations for future assessment efforts

As with the evaluation of the ranking feature, increasing the number of participants would be the most significant change, as it would allow the team to confidently conduct statistical analysis of results. Triangulation purposes could be served by conducting a small number of observed sessions. Similarly, the number of scenarios would need to be reduced, as the recommender feature analysis was intense for the participants and the assessment team.

Relatively straightforward changes include providing the full bibliographic record for recommended items; clearly labeling the sources of recommendations, and ensuring that the UI sufficiently supports all of the protocol requirements.

A more difficult but required change is to develop a precise understanding of how to capture information regarding dimensions other than usefulness and familiarity. Aspects of novelty were combined with other attributes such as authoritativeness, which may have muddied the categories. Determining exactly what is meant by novelty and how to convey that to users will be essential for designing and evaluating a system that supports
identifying these types of items. One solution may be to develop a very specific definition of novelty and illustrate it with several different sets of recommendations.

7.6 For further exploration
There are many options for further development of recommendation services.

Refinement of circulation-based approach. Although we were able to generate some useful recommendations with the circulation-based approach, there is clearly ample room for improvement. Refinements to the current strategy could include:

- Application of additional grouping methods, including FRBR. This could be applied in a number of different ways: aggregating circulation statistics at the work level for recommending, or grouping related versions in recommendation sets to reduce the appearance of duplication.
- Experimentation with presenting multiple bins of recommendations in several content areas, to maximize the possibility of serendipitous recommendation across disciplinary boundaries.
- Additional examination of content-area groupings to address gaps and incorrect groupings.
- Modification of the algorithm to create a subject-area “weight” on each node rather than applying a filter. This would allow more sophisticated manipulation of recommendation sets, to balance the signals from human-generated pairings against similarity in content.

However, future availability of appropriate, anonymized circulation data must be ascertained before embarking on any of these refinements. Unless there are changes in library policy and data handling practices within the UC system, it is unlikely that we will continue to have access to such substantial pools of anonymized circulation data. Moreover, as more content is available on-line, traditional physical circulation data will be progressively less representative of library usage patterns, and recommendations generated from these data will become progressively less relevant.

Exploration of other data sources. Of the many potential development alternatives that do not require storage of personal profiles, one of the most interesting would be an attempt to utilize other data sources. For example, we could attempt to harvest and mine reading lists that are freely available on the Web. Conceptually, these are valuable as a source of thematically grouped recommendations by subject matter experts. This effort would present numerous technical challenges, including identification of an appropriate and substantial enough body of reading lists, development of quality heuristics, and parsing and matching of citations.

Book bag mining. Our current system allows users to place items in a temporary book bag without logging in, so that they can keep track of interesting items found during the current session of browsing. One development option would be to explore the ability to anonymously log the contents of these session-based book bags, thus recording sets of items assembled during a single browsing session. Adding logging capability would require little technical effort. However, pursuing this option would require some
understanding of how patrons use temporary book bags; would such sets hold together thematically? And would there be sufficient volume of data in order to generate recommendations?

**More on content-based techniques.** Further development of content-based clustering techniques to identify similar items is another avenue. As an enhancement or alternative to the “more like this…” method we have already developed, further refinements could include consideration of relevance feedback methods that incorporate the original query when formulating the “more like this…” query.

While serving the primary purpose of offering better alternatives for user interface to the data, development of a structure of faceted hierarchical browse nodes could also provide a more efficient way of identifying groups of similar records to display as recommendations. Although there would likely still be a need to do some processing to narrow the set, assignments of browse node groupings could streamline the process of locating similar items within the collection.

**Persistent personalization.** Another approach entirely would be the application of persistent user profiles. Storage of personal data in a production environment would require a thorough examination of policies protecting patron privacy, patron needs and attitudes surrounding privacy, and how these affect development of effective recommending services. Some key questions to explore would include:

- How much data should be kept as part of a profile? How much demographic information? How much usage pattern data?
- Would it be possible to build modular recommending services, with different levels of service based on the amount of personal data the user chose to store?
- How would users opt in? How easily could they reverse the decision to opt in? Could they choose to “toggle” between anonymous and profile-based services depending on the task at hand?

Stored profiles could allow for richer and more persistent customizations to recommendation sets. Patrons could conceivably fine-tune services by choosing from configurable options like expressing subject areas of most interest and disallowing previously checked out items.

Another approach to personalization would be to develop a set of personal resource management tools that could sit on top of the catalog and other library resources to allow for maintenance of personal and shared resource lists, tags and annotations. Some of the questions for exploration would include:

- What levels of granularity and configurability would be necessary to support the range of academic information management tasks?
- Would such services be available only within a constrained academic community, or publicly accessible?

---

25 Flexible Searching Using Faceted, Hierarchical Metadata, from the Flamenco research project. [http://www.eecs.berkeley.edu/IPRO/Summary/03abstracts/kirstens.1.html](http://www.eecs.berkeley.edu/IPRO/Summary/03abstracts/kirstens.1.html)
• What would the incentive be for use of this service? Could faculty, teaching assistants and library subject specialists utilize it to construct course reading lists and subject-specific resource lists to share?
• How many users would opt in and share lists anonymously? How many would be willing to identify themselves, in the interest of promoting scholarly communication?
• What would constitute a critical mass of personal and shared lists and tags sufficient to build a storehouse of data large enough to drive recommendations?

Intersection of methods
The many methods of generating recommendations are not mutually exclusive; the availability and quality of recommendations available for any given item will vary with the data used to generate them. We may be able to generate few or poor recommendations for items with little circulation history, but very good content-based recommendations. For items with strong circulation history, we may be able to offer better and more interesting recommendations based on circulation. Offering multiple sets of recommendations based on different criteria may give patrons more flexibility and richer opportunities to explore.

It is not technically difficult to generate multiple sets of recommendations. But it will require careful attention to the design of the user interface. We will need to ensure that recommendations are clearly presented, that adequate metadata is available, and that users have cues that help them understand that recommendations can serve as a query expansion device, but without overwhelming them.

8 Dissemination
The project team pursued a number of strategies for disseminating information about work in progress and results throughout the course of the project. Our target audiences were peers in the digital library and academic information communities, and colleagues within the University of California library system. With these audiences in mind, we developed an informational web page within the California Digital Library site26 as well as a flyer for distribution as hard copy and PDF.27 In addition, we seized opportunities to mention the project in newsletters and articles for both internal and external audiences (e.g. the CDLINFO newsletter28, and a Library Journal column29)

Over the course of the year, three team members (Peter Brantley, Brian Tingle and Colleen Whitney) made five presentations featuring work being done in this project:


28 CDLINFO newsletter article: http://www.cdlib.org/inside/news/cdlinfo/cdlinfo101305.html
29 Library Journal article: http://libraryjournal.com/article/CA6256255.html
9 What next?

This report has identified areas for further work extending from each of the five major areas of exploration carried out in this project. Of these many possibilities, CDL plans to pursue a handful of initiatives for ongoing work.

9.1 Additional work on spelling correction

Although our investigation of index-based spelling correction yielded good performance on single words, deployment in a production system would be hobbled by the lack of support for multi-word queries. Furthermore, although this work has been undertaken specifically for the Melvyl Recommender project, a robust spelling correction algorithm could be extended to benefit many applications built on the XTF code base. A small amount of additional development time to address this deficiency is thus likely to yield large returns.

9.2 Mechanics of persistent personalization

Many attractive recommendation strategies for further exploration depend on the ability to create and store persistent user profiles. Although there are many larger-scale questions to be investigated about how the profiles should be deployed and utilized, one very short-term goal is to develop the capacity to store persistent profiles. This discrete, small-scale chunk of work would provide an essential building block for many avenues of further development.

9.3 Integration of full text resources

Growing pools of full text electronic resources are becoming available to the academic community and to the public at large, fed by streams of text from book scanning efforts, digital publishing initiatives and digital repositories. Our work to date has focused on the traditional catalog environment, in a retrieval system operating only on metadata records.
Given the likely prevalence of full text in the future, we are interested in exploring methods of integrating discovery of full text resources into the catalog, in order to enrich and streamline services to our patrons. This small-scale investigation would take a first pass at identifying potential applications and stumbling blocks associated with retrieval in a mixed metadata/full text environment.

9.4 Grouping and clustering
Working with very large sets of bibliographic records to create direct services for users presents many challenges, many of which relate to grouping or clustering records within a set. Users want to be able to browse by subject, but subject classifications are not well structured for this purpose. Users have difficulty sorting through masses of very similar records, but bibliographic data are not structured such that it is easy to group multiple versions of the same work. Our work to date focused on the mechanics of the system, developing the technical capacity to present grouped records in the interface. Effective use of this capacity will require additional attention to automated methods for analysis and manipulation of the existing metadata.

9.5 Extended work on recommending
Our assessment work in the course of this grant confirmed that patrons are eager for recommendation services. Given a multitude of options for further exploration, there are two main avenues that we wish to pursue. The first is further refinement of the current circulation-based techniques, potentially incorporating alternative data sources. The second is to experiment with the application of book bags and personal profiles (see “Mechanics of persistent personalization”, p. 65) in recommending. The latter avenue might include tailoring recommendation sets to fit expressed areas of interest, as well as mining relationships between items bagged or tagged by patrons.

9.6 User-centered service design
The many technical gains made during the course of this project are exciting, but successful deployment of new services based on this work will require careful attention to how users will interact with the system. In addition to considering how to present multiple recommendation sets, we will need to consider carefully how to integrate spelling correction, faceted browse, book bags, and other new features into a coherent and usable system. This is, perhaps, the least tractable task in front of us, but it is critical in terms of ensuring a smooth bridge from development to deployment.

10 Acknowledgments
Funding for this project was provided by the Andrew W. Mellon Foundation. Holdings and circulation data were provided by the Research Libraries Group, the Online Computer Library Center, and the libraries of the University of California, Los Angeles and the University of California, Berkeley.
This report was written by Colleen Whitney, Lisa Schiff and Martin Haye, with review and input from many members of the Melvyl Recommender Project team. Thanks to Bradburn Young for careful reading and editing of the final draft.

11 Appendices

Appendices are available as separate documents on the project Web site: http://www.cdlib.org/inside/projects/melvyl_recommender/

A. Project timeline  
B. Project team and contributors  
C. Description and screen shots of prototype  
D. Descriptions of holdings, circulation data sets  
E. Performance testing  
F. Subject area groupings  
G. Bibliography for assessment activities  
H. Assessment plan  
I. Human subjects approval  
J. Assessment instruments  
K. Screen shots from relevance ranking assessment  
L. Screen shots from recommending assessment  
M. Project flyer  
N. New and modified XTF modules