Title
The Melvyl Recommender Project: Developing Library Recommendation Services

Permalink
https://escholarship.org/uc/item/94d0r3cx

Authors
Whitney, Colleen
Schiff, Lisa R

Publication Date
2006-12-01

Peer reviewed
The Melvyl Recommender Project

Developing Library Recommendation Services

Colleen Whitney
California Digital Library*
<whitney.colleen@gmail.com>

Lisa Schiff
California Digital Library
<Lisa.Schiff@ucop.edu>

*At the time of the study, Colleen Whitney was at the California Digital Library; however, she is now at the Mark Logic, and the email given for her is correct at the time of publication.

Popular commercial on-line 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 recommender systems. Online
library catalogs (OPACs), in contrast, have changed little and are notoriously difficult for patrons to use (University of California Libraries, 2005). Over the past year (June 2005 to the present), the Melvyl Recommender Project (California Digital Library, 2005) 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 conducted exploratory work in five topic areas: relevance ranking, auto-correction, use of a text-based discovery system, user interface strategies, and recommending. This article focuses specifically on the recommending portion of the project and potential extensions to that work.

Recommender Systems: Prior Work

At present, recommender systems are most visible in the e-commerce arena, with Amazon and E-Bay as notable examples. Likewise in the academic community, much of the research on recommender systems has stayed in this realm, and has focused on developing new and more efficient ways of providing e-commerce related recommendations. For example, important work with applicability to the private sector and the general field of recommendations has been conducted by Herlocker, Konstan, et al. (2000) and others with the MovieLens test site. This research has been targeted towards better understandings of how to generate and present recommendations, as well as how to evaluate recommender systems.

Research focusing more specifically on the needs of academic users has been less voluminous. Significant "cross-over" work can be found in the TechLens and Group Lens test sites, which were developed by Torres et al. (2004) and McNee et al. (2002) to move from recommending films to recommending peer-reviewed journal articles to faculty members who were writing papers. Houstis et al. (2000) have experimented with a system that provides recommendations to scientists who are trying to identify the appropriate software for their research needs. Sierra (2006) developed a system for recommending library subject terms to users to assist in catalog searches. Middleton et al. (2001) have explored improved recommendations of research papers.
on the World Wide Web by mining correlations between logged user behavior (profiles) and classification of URLs for papers that users have previously examined.

The Ilumina Project is a digital library of undergraduate teaching materials in the sciences for use by instructors, students and resource contributors. As described by Geisler et al. (2001), one of the original goals of the project was to provide recommendations based on document metadata, available subject expert analysis of documents, resource use as discovered in logs, and user profiles for those users who had chosen to register with the system. Inconsistency of data for generating recommendations was identified as an initial and ongoing constraint. Later work in the project turned away from the problem of providing recommendations and focused instead on the use of metadata and the creation of virtual digital libraries.

Huang et al. (2002) developed a hybrid recommender system for book selection for faculty members. Recommendations were based on a combination of textual similarity and purchasing histories, testing the effectiveness of the combined system against content and collaborative systems, respectively. Evaluations showed that the hybrid system was more effective at predicting what books individuals had purchased, while the content-based approach was more effective when subject experts were evaluating the recommendations.

Jung et al. (2004) have developed and put in place SERF – the System for Electronic Recommendation Filtering. SERF is based on collaborative filtering, and thus depends on users rating sources throughout their session with the catalog in order to later use those ratings to provide recommendations to future users. Initial analyses of the results indicate that users find the recommendations helpful.

In a project most similar to the Melvyl Recommender project, Geyer-Schulz et al. (2003) developed a recommender system for the OPAC of the University of Karlsruhe. Recommendations were generated by analyzing logs that captured when users chose to view detailed information about a certain paper or book. Such choices were operationalized as positive recommendations for those items and were collocated in order to make links between one item and another. Users
encountered recommendations in the OPAC when they were available. During a three-month window, users were asked to evaluate the quality of the recommendations using a five point Likert scale. Nearly 500 evaluations were collected, and data indicate that a majority of users were satisfied or very satisfied with the system.

Adomavicius et al. (2005) have explored generating recommendations that are more contextually precise, matching specific types of users with specific needs at distinct moments of activity. Such work is useful in considering the differing needs of novices and experts within the academy.

The Melvyl Recommender project differed from earlier efforts in a number of ways. First, it differed in its use of historical circulation data for generating recommendations. Second, unlike the majority of prior efforts, which were limited to specific subject areas (often science and technology), our project explored making recommendations in a union catalog setting. And third, in addition to focusing on the needs of an academic audience, our protocol also probed the question of how expertise level contributes to user satisfaction with recommender systems.

**Recommending Methods Explored**

The Melvyl Recommender Project team explored two methods of generating recommendations. The first method used circulation data from the University of California, Los Angeles (UCLA) to determine linkages between items ("patrons who checked this out also checked out..."). A second, content-based, strategy used terms from the bibliographic records to develop queries for similar items ("more like this...").

**Circulation-based method**

The circulation-based method hinged on the availability of two substantial sets of circulation data from UCLA. The first consisted of about 7.7 million circulation transaction records spanning almost five years, from September 1999 through July 2004. The second consisted of about 1.6 million circulation transactions from July 2004 through May 2005. The data were broken into two sets due
to a system conversion from the Taos system to the Voyager system in the summer of 2004.

These data sets were appropriate for this exploration for three reasons. First, they 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, and 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 in which we were interested. 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 Taos 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 concerns (Poe, 2005). 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 decision to check out 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 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.

![Diagram](image)

**Figure 1.** Recommendations for item A, in rank order, are items D, C and B because five people have also checked out item D, three people have also checked out item C, and one 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 because it is a fundamental work in the discipline, or 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 rough first 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 also beginning with "P" were recommended. This filtering method resulted in more cohesiveness, but at a cost. Although this method did eliminate the most jarringly bad recommendations in most cases, it also had the effect of reducing the
possibility of making 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. The groupings, based on work done by the Columbia University Digital Library Projects (Columbia, 2002), adjusted for the UCLA records where there were gaps, and supplemented the Library of Congress data with mappings from National Library of Medicine call number classes to the same scheme of general subject areas.

These mappings 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.

Content-based method

The group also experimented with a second method of producing recommendations. This method 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.

This method yielded recommendations that differ significantly in character from the circulation-based recommendations. They tended to be much more homogeneous, both within the recommendation set and with the target item.

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.

Given the accelerated timeline of the project, we opted to focus user testing on recommendations generated using the circulation-based method, forgoing testing of the content-based recommendations at present. This decision allowed us to eliminate a great deal of complexity in labeling recommendations from multiple sources and analyzing the results.

Results: Key Observations

We undertook a small-scale evaluation of the circulation-based recommending method. The questions that we were asking in the course of the evaluation were:

- Do recommendations help academic users find items relevant to a research-related task?
- How do academic users determine if a recommendation is useful?
  - Which components of the bibliographic record do users identify as helpful in deciding whether or not a recommended item is useful?
  - Is subject expertise level important a factor in this?
- What is the quality of a given recommended item?
  - Is a given recommendation interesting but not useful (i.e., is not relevant enough)?
  - Is a given recommendation relevant but not useful (i.e., the person already knows about it)?
  - Is a given recommendation new, but not surprising (i.e., novel but not serendipitous)?
  - Is a given recommendation relevant and unexpected (i.e.,
serendipitous)?

- What is the quality of a given set of recommendations?
  - How useful was the set in accomplishing the defined task?
  - Were there missing items?
  - How many relevant, unknown items were there?

Working with a liaison at the UC Berkeley library, we recruited ten undergraduate and graduate students in the Humanities and History. The protocol was task-based: participants were given typical academic tasks tailored to their subject areas, distinguishing between subject-naïve and subject-expert users. For example:

**Comparative Literature, subject-naïve level:** "You are an undergraduate majoring in Comparative Literature. You are taking a course on The European Avant-garde: From Futurism to Surrealism and need to start thinking about a paper topic. Find some items in the class subject area that would help you with this."

**History, subject-expert level:** "You are a doctoral student in History. You will be teaching a seminar on the history of apartheid in South Africa. You are working on the section of the curriculum having to do with the dismantling of apartheid. Identify some items that would be appropriate for the reading list."

With the task in hand, the participants searched in the Melvyl Recommender prototype (named Relvyl), with an interface modified slightly for data capture. The sessions were facilitated and observed; participants were asked to think aloud and answer questions about each recommendation as they moved through their tasks. These questions were based on a protocol developed by Torres et al. (2004), which attempted to probe some of the more intricate details: level of familiarity, appropriateness to the task (e.g., is it too specialized or too general).

Analysis of the quantitative data generated by participants as they evaluated recommendations and of the qualitative data gathered from surveys and observations revealed several key themes.
1. Users want to see recommendations in the catalog to support their academic work. Even those who are somewhat skeptical are interested in seeing and trying recommendations, until or unless they deem the quality of recommendations to be poor. When asked whether they would expect to use a recommendation feature in an online catalog for academic work, of the ten participants: seven chose "very likely", two chose "somewhat likely", and only one chose "not likely at all". (None chose neutral or somewhat unlikely.)

2. Presentation is critical. From reviewing literature on recommending, we had ample evidence that users need to understand why a recommendation is being made, and that they need to see sufficient metadata to evaluate the potential usefulness of an item. Our observational data validated this evidence. In particular, participants commented that bibliographic records do not provide cues that they find particularly helpful in other settings: book summaries or excerpts, tables of contents and indexes.

3. The preferred sources of recommendations cited by participants are faculty, bibliographies and footnotes.

4. Recommendations were successful in supporting academic tasks. Approximately one third of recommended items were rated positively by participants. 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.

5. Recommendations can serve as an effective device for query expansion. Participants who encountered single-item or very small result sets did find recommendations effective for helping to reframe the query.

6. 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. Users were unfamiliar with the majority of recommended items, but did not consider positively rated, but unknown works...
as serendipitous finds.

7. **Subject expertise influenced the way that participants evaluated recommendations.** Similar to the evaluation of any item in a traditional result set, users relied heavily on titles and publication dates to assess the usefulness of a recommended item. They additionally used the original source of the recommendations (referred to in our study as the "seed" item), as a comparison point. Subject experts were much more likely to be familiar with particular items or authors, and thus had a greater ability to evaluate items. Because of their domain exposure, canonical items were not directly useful to subject experts; however, such items added validity to recommendation sets for these users.

8. **Items in result and recommendation sets can serve different roles.** A tension exists between the construction of result sets and recommendation sets. In general, when users evaluated items in a result set, they were seeking new, unread items and would bypass known, good items. However, items known to be useful would potentially be excellent sources of recommendations, as they are already "vetted" as being relevant and presumably would have a greater chance of being associated with other useful or high quality items. An effective interface should allow users to choose items for different end purposes, in order to encourage the use of such known good items as sources for generating recommendations.

9. **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.

**Next Steps**

The small-scale evaluation made a strong case for continuing on the path of developing recommendation services for library patrons. Patrons do want recommendations, if we can generate high quality recommendations and present them appropriately.
Ongoing technical development of recommendation systems could pursue two very different strategies, which are not mutually exclusive. The first strategy, which we will call "patron-neutral," continues along a development path that requires no persistent knowledge of the patron who is viewing the recommendations. The second strategy, which can be labeled "profile-based," investigates the use of persistent patron profiles to facilitate collaborative filtering as well as popular services such as shared lists, annotations and tagging. In addition to these technical development pathways, another important next step is an investigation of privacy policy and practices in the University information environment as they relate to the development of recommender systems and similar enhancements.

Privacy and personalization

Some of the alternatives for development of recommendation services in library systems hinge on difficult questions around library policy designed to protect patron privacy (American Library Association, 2006). We know that we could develop richer and more personalized services if we, like Amazon and Netflix, could do sophisticated mining of browsing and purchasing habits based on personal profiles.

But the tenet of intellectual freedom is deeply rooted in library culture and in the culture at large, for good reason. This is not just a matter of arbitrary practice, but has been encoded in law and policy at many levels of the system (Coyle, 2002).

A thoughtful investigation of University privacy policies designed to protect library patrons, the risks to patrons and to University posed by personalization, potential alternatives for mitigating those risks (e.g., an opt-in approach for personal profiles), and an analysis of user perceptions of those alternatives should accompany development of profile-based recommending strategies. Do the existing policies allow for an appropriate balance between privacy protection and service provision? If patrons are offered alternatives, how could they be clearly communicated such that patrons can adequately assess the risks and potential benefits? If patrons are presented with the option of receiving enhanced services in return for some loss of privacy, what will they choose?
Will the trade-off yield enough of a benefit to justify the risks?

**Continued development of patron-neutral strategies**

The existing circulation-based approach is patron-neutral: although it incorporates information about prior usage patterns, it is devoid of specific data to allow tracking back to individual patrons. Although we were able to generate some useful recommendations, there is clearly ample room for improvement. So, we could consider continued refinement of the current strategy, including potential refinements such as:

- **Application of additional grouping methods, including FRBR.** This could be applied in a number of different ways: such as aggregating circulation statistics at the work level for recommending, or reducing the appearance of duplication in recommendation sets.
- **Experimentation with presenting multiple "bins" of recommendations in several content areas, to maximize the possibility of serendipitous recommendation across disciplinary boundaries.**
- **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, 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.

Instead of relying on the same sources of data, we could alternatively apply what we have learned to potential sources of new data. These could include course reading lists freely available on the
Internet, bibliographies from full-text articles and books, and online usage statistics, each of which presents a different set of challenges.

Another patron-neutral strategy to consider would be anonymous, session-based personalization of recommendation sets. Rather than requiring the use of persistent profiles, this strategy could allow for some light-weight, temporary customization that could improve recommendations. Offering a session-based "book bag" functionality could also contribute data for recommendations without requiring personally identifiable information.

**Profile-based strategies**

Finally, we could consider approaches based on the maintenance of persistent user profiles. One strategy would include the application of true collaborative filtering (although data distribution and sparsity issues are still likely to exist). A second area of exploration is the use of stored profiles to 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. A third strategy would be to leverage personal and shared resource lists, tags and annotations.

Some key questions in this realm 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?
Conclusion

Early results of the Melvyl Recommender Project show strong evidence that University of California (UC) library patrons are interested in receiving recommendations to support both academic and personal information needs. Our first attempt to produce recommendations using circulation data met with mixed results. Only about a third of the recommendations our system generated were helpful to participants in user testing. Nevertheless, participants were almost unanimous in their support for development of such services.

Possibilities for further development can be grouped roughly into two strands: patron-neutral (requiring no storage of persistent information about the patron who is accessing the system), and profile-based (requiring some level of persistent knowledge about user demographics or patterns). Because the long-term availability and utility of circulation data is uncertain, further development of patron-neutral strategies will likely require applying what we have learned to new sources of data. These might include on-line reading lists, bibliographies from full-text articles and books, or log data.

The investigation of profile-based strategies would necessarily need to include analysis of library policy and user perceptions surrounding the trade-off between enhanced services and library patron privacy. Technical explorations could encompass collaborative filtering techniques, patron control of customized recommendation settings, or storage and mining of personal and shared resource lists, tags and annotations.

Underpinning any of these alternatives, but particularly those requiring persistent user profiles, is a need for a thorough examination of policies protecting patron privacy, patron needs and attitudes surrounding privacy, and how these affect development of effective recommending services.

References


Copyright © 2006 Colleen Whitney and Lisa Schiff