Developing Recommendation Services for a Digital Library with Uncertain and Changing Data

Gary Geisler
Interaction Design Laboratory
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3360, USA
+1 919 489 2759
geisg@ils.unc.edu

David McArthur, Sarah Giersch
Eduprise
2000 Perimeter Park Drive
Morrisville, NC 27560, USA
+1 919 376 3424
{dmcarthur, sgiersch}@eduprise.com

ABSTRACT

In developing recommendation services for a new digital library called iLumina (www.ilumina-project.org), we are faced with several challenges related to the nature of the data we have available. The availability and consistency of data associated with iLumina is likely to be highly variable. Any recommendation strategy we develop must be able to cope with this fact, while also being robust enough to adapt to additional types of data available over time as the digital library develops. In this paper we describe the challenges we are faced with in developing a system that can provide our users with good, consistent recommendations under changing and uncertain conditions.

Keywords

Digital Library, Recommender System, User Services.

1. INTRODUCTION

Digital libraries—and Web-based resource collections in general—have traditionally enabled their users to locate resources through search and browse services. Over the past decade there has been growing use of recommendation systems as a way to suggest new items of potential interest to people [3]. In designing a new digital library (DL) called iLumina, we are exploring the potential of recommendation services as a way for users to find resources of interest in a digital library. iLumina is a DL of undergraduate teaching materials for science, mathematics, engineering, and technology (SMET) education [1], now being developed by Eduprise, The University of North Carolina at Wilmington, Georgia State University, Grand Valley State, and Virginia Tech. Because the DL is new and non-commercial, the data available to use as input into a recommendation system will be uneven and will change over time. We are particularly interested, therefore, in how to create robust recommendation services for a DL that contains changing and uncertain data.

JCDL '01, June 24-28, 2001, Roanoke, Virginia.

2. INFORMATION AVAILABLE FOR RECOMMENDATIONS

The iLumina digital library contains content across a wide range of science, math and engineering disciplines. Our user community includes instructors, students, and resource contributors; some—but not all—will register and provide profile information. From our server logs we expect to have usage data related to both resources and users. These relatively standard characteristics provide us with four classes of data potentially useful for recommendation services:

Resource characteristics: All resources will be described by basic, IMS derived [4] metadata that identifies elements such as title, description, format, and requirements. This metadata will be rich and consistent.

Resource quality judgements: Subject experts will formally review many (but not all) of library resources. All of the content will, however, be available to be reviewed informally by users; similarly, all materials will have passed minimal acceptability standards. But the written reviews, where available, can provide richer and more subjective information about the potential usefulness of the resource for specific situations.

Resource use: Data describing how often a given resource has been downloaded, reviewed, or used as a component in larger resources is available from server logs and the resource database. Data about resources can also be associated with individual users to develop patterns of usage by user characteristics.

User profile: Our database contains descriptive information about registered users, such as status, affiliation, areas of interest, explicit resource ratings, and service preferences. This information can be used to group users based on various similarity characteristics, track resource usage data for a given user, and adjust recommendations based on expressed preferences.

All of this data can be very useful for generating recommendations, but they vary in quality and likelihood of existence. For example, we do not require users to register, so profile information will vary. Reviews will range from expert-level, instructor judgements to student comments. Some registered users will rate resources, others will not. Table 1 summarizes the quality and existence characteristics for each of the basic types of data we expect to use as input for generating recommendations.

Table 1. Characteristics of the types of data available for generating recommendations

Type of data	Quality	Existence
Resource characteristics	High	Always
Resource quality	Variable	Variable
Resource use	High	Variable
User profile	Variable	Variable

3. GENERATING RECOMMENDATIONS

Given the characteristics of the types of data we have available, several schemes for providing recommendations are possible. At one extreme, we could use only data that we know is high quality and always available, such as resource characteristics and resource use. One example of this strategy would be to recommend to users resources that are similar, on specific metadata fields, to ones they have previously downloaded. The obvious downside of this strategy is that other potentially rich sources of data are ignored.

At the other extreme, we can include all potentially useful sources of data in our recommendation scheme, and attempt to compensate for missing inputs and inputs of variable quality. The downside of this strategy is that it is hard to combine multiple information sources in a principled way, and the variable quality of sources threatens the usefulness of recommendations based on them.

Different rules could be used to, in effect, define various points along this continuum of recommendation strategies. These might include (information types used noted):

- If the user suggests a resource she likes, we can suggest structurally similar resources (resource characteristics)
- If profile information for the user exists, we can use it to suggest resources (resource characteristics, user profile)
- If profile information exists and the user has previously downloaded resources, we can suggest resources based on the previously downloaded resources (resource characteristics, user profile, resource use)
- If the user suggests a resource she likes or profile information exists, we can suggest resources based on all available data (resource characteristics, user profile, resource use, resource quality judgments)

4. CHALLENGES

There is evidence that recommendations can be improved by combining methods, such as collaborative and content-based approaches [2]. In iLumina we will collect a range of data that supports such a multiple-source recommendation strategy. However, we can only be certain of the existence and reliability of the basic resource descriptive data; the availability of other data such as reviews, ratings, and user profiles is less predictable.

This uncertainty brings into question the value of a DL recommendation scheme that attempts to use all data sources and adapts when some of the data does not exist or is unreliable. In such cases, will the recommendations still be useful? More generally, do DL recommendations improve as the number of information sources they use increases? Is a strategy that uses all available information, some of which may not be available, more effective than one that uses the high-quality, always available resource characteristics? If so, given that gathering each type of data comes at some cost, is the difference in effectiveness worth the cost? Can we evaluate the cost/benefit ratio of using the different types of data?

Because the iLumina DL is new we will be in a position to address these kinds of questions. We expect to phase-in user services that provide us with relevant data over time; the data available to use in the second year of operation will be much broader than that available in the first year. As iLumina grows, then, we will not only be able to devise recommendation schemes that incorporate new types of information as they become available, but will also be able to employ user opinions to judge how the perceived value of different schemes improves (or not) as the data sources become richer.

5. CONCLUSION

Recommendations can be a valuable service for users of a DL such as iLumina, which will eventually contain many resources. The variety of data associated with both the resources and users of a DL represent a potentially rich source of input for recommendation services. As the iLumina DL evolves, we will be interested in developing both strategies that provide useful recommendations to users even if the data available is uneven and changes over time and methods for evaluating these methods. We expect that the results of our efforts will be informative to developers of many types of DLs.

6. REFERENCES

[1] McArthur, D., Giersch, S., Graves, B., Ward, C.R., Dillaman, R., Herman, R., Lugo, G., Reeves, J., Vetter, R., Knox, D., & Owen, S. (in press). Towards a Sharable Digital Library of Reusable Teaching Resources: Roles for Rich Metadata. *Journal of Educational Resources in Computing*.

[2] Mooney, R. J. and Roy, L. (2000). Content-Based Book Recommending Using Learning for Text Categorization. In *Proceedings of the 5th ACM Conference on Digital Libraries* (San Antonio, TX, June 2000).

[3] Recommender Systems [theme issue]. (1997). *Communications of the ACM* 40(3).

[4] The IMS Global Learning Consortium, http://www.imsproject.org