This is an Author's Accepted Manuscript of an article whose final and definitive form, the Version of Record, has been published in Digital Library Perspectives [2018, vol. 34, no. 1], available online at: https://doi.org/10.1108/DLP-07-2017-0022

Account-Based Recommenders in Open Discovery Environments

Author(s): Jim Hahn, (University Libraries, University of Illinois at Urbana-Champaign, Urbana, Illinois, USA), Courtney McDonald, (Indiana University Libraries, Bloomington, Indiana, USA)

ABSTRACT

Purpose This paper introduces a machine learning based "My Account" recommender for implementation in open discovery environments such as VuFind, among others.

Design/methodology/approach The approach to implementing machine learning based personalized recommenders is undertaken as applied research leveraging data streams of transactional checkout data from discovery systems.

Findings The authors discuss the need for large data sets from which to build an algorithm; and introduce a prototype recommender service, describing the prototype's data flow pipeline and machine learning processes.

Practical implications The browse paradigm of discovery has neglected to leverage discovery system data to inform the development of personalized recommendations, with this paper, the authors show novel approaches to providing enhanced browse functionality by way of a user account.

Originality/value In the age of big data and machine learning, advances in deep learning technology and data stream processing make it possible to leverage discovery system data to inform the development of personalized recommendations.

Keywords

Discovery; recommendations; open algorithm; machine learning; research libraries; personalization

Article Classification
Research paper

Introduction

Throughout the last decade, development and enhancement of the library discovery system has seen great innovation. From federated searching to bento box style approaches, much work has focused on leveraging indexed data from journal article databases, the library catalog, and digital library projects into one unified search box and result list (Antleman *et al*., 2006; Lown, *et al*., 2013; Rochkind, 2013). Throughout this period, discovery projects have remained singularly focused on search.

Modern discovery systems, notably from commerce and entertainment, do not solely rely on the user searching a known title or subject in a database for all exploration; rather, contemporary information environments also provide recommendations relevant to a user's interests and needs, based on a user's account history - by what she has viewed or purchased in the past. Browsing experiences of YouTube (Davidson *et al*., 2010) and Amazon (Linden *et al*., 2003) rely heavily on recommendations informed by data mining. Although personalized recommendations have become an expected and helpful component in online search settings, library systems do not currently leverage data mining and machine

learning based personalization features for discovery, despite the fact that recommendations are commonly identified as a key criterion for evaluating next-generation catalogs (Moore and Greene, 2012; Chickering and Yang, 2014). As early as 2003, Geyer-Schulz et al proposed implementation of behavior-based recommender services in library systems, noting the benefits to researchers in time savings and the ability to "profit from the combined knowledge of all library users in contrast to the more restricted knowledge within their personal networks." They attributed reluctance on the part of libraries toward the development of such recommender systems utilizing patron data to concerns in the areas of privacy, budget restrictions, and data size.

The authors contend that discovery can encompass profound browse functionality by leveraging large discovery system datasets of user data and open source discovery platforms to supplement and deeply enhance the experience of discovering items relevant to a user's current interests. Research library systems hold vast stores of user data that have not been processed with machine learning and data mining for discovery purposes. Awareness and use of academic research collections can be fostered by way of unique personalization algorithms which have so profoundly impacted contemporary search.

In examining alternative paradigms, library portals for personalized learning have been prototyped and developed (Hanson, *et al.*, 2008). Researchers in information retrieval and computer science have suggested usage frequency for recommendation as well as collaborative filtering techniques (Kim and Gyo Chung, 2008; Liao *et al.*, 2010). In 2012, LibraryThing introduced the recommender tool, "BookPsychic" in order to address a Pew Internet study on Library Services in the Digital Age which found that over 64% of patrons are "interested in a library service which suggested books, audiobooks, and DVDs to them based on their own preferences" (Dibbell, 2013). These efforts underscore previous work and establish a compelling thread from which to explore the integration of personalized recommendation within modern open source discovery systems like VuFind.

There is a rich history of classification-based recommendation in library science of which this paper draws in order to advance the state of the art for account-based recommenders. Collocation objectives in library science have been leveraged to great effect by discovery systems since intellectual organization by shelf classification and already existing collocation attributes makes possible a serendipitous type of discovery for shelf browsing (Svenonius 2000, p21-22). Efforts at making virtual shelf browsing that leverage call number searching has

been integrated into modern versions of discovery systems including recent versions of VuFind. A foundational mobile discovery project leveraged shelf collocations for location-based recommendations in library book stacks and reviewed much of the foundational literature of which the prototype account-based recommender is based (Hahn, 2011).

Prototype Work

Proof of Concept Software

The necessary models to generate personalized recommendations have not yet been integrated in library practice, partly because no open algorithm exists for library practitioners to easily implement. Of the available options for recommendation software available to system designers, there are very few that are proven, maintained, and freely available. Therefore, building a custom software framework and foundations of an algorithm for library systems was a necessary first step. Proof of concept recommendation middleware was developed to provide basic personalized recommendations for research library users using VuFind accounts at a large public research-intensive institution under a Campus Research Board Grant "Research and Development of an Intelligent Personalized Recommendation Platform for Library Accounts."

As a pilot study, the recommendations software has been incorporated into the library mobile app for iOS and Android (hereafter referred to as Library Mobile App).[1] Since the pilot study gathers user data to generate personal recommendations and then analyzes user interactions with the recommendations to evaluate the software, the research team secured ethics approval to undertake human subjects research. Before securing the ethics approval, the research team also obtained approval from university library administrators to gather anonymized clusters of check out data to use for this project. The clusters of items checked out together are generated upon checkout and stored in a secure database. Personally identifiable information (or PII) are not stored. Like all Library Mobile App modules, the Recommendations module is powered by a RESTful API (Application Programming Interface).

The prototype work by the research team has mapped developmental data flows for item-based filtering using subject headings and collaborative filtering by way of user similarity. The Library Mobile App for Android and iOS 3.1 recommendation module utilizes several sources of data including user checkouts as part of a personalized recommendation research experiment.

Account Based Recommendations with Machine Learning

The basis for the account based recommendations begins with clusters of checked out items that the integrated library system records when items are checked out. Drawing on examples from "consumer data science" (e.g. Netflix) it is clear that large corpus data that receive millions of ratings daily are part of the strategy for creating compelling recommender algorithms (Amatriain, 2013). Since the prototype system doesn't yet incorporate user feedback such as ratings, the research team sought to adhere to the principles of consumer data science by collecting as much topic/subject metadata that are clustered together as possible and rapidly testing the effectiveness of personal recommendations with a pilot implementation. Topic metadata clusters, collected from transactional checkout data of items that are checked out together form the basis for generating a rule set. The prototype recommender started in October 2016 with seed data of 33,060 consequent subject association rules as the result of initial data mining and machine learning processes. At the time of writing (July 2017) there are 131,885 consequent subject association rules. After nearly a year of data stream collection the system has collected over 250,000 rows of anonymized transactions representing checkouts with topic metadata. The collection period was roughly eleven months, beginning July 28, 2016 through June 28, 2017. Note that items in

library collections often have several subject terms. A table of the transactions collected and subject associations stored since the service was developed is show in table 1.

Month and Year	**Number of Anonymized Transactions**	**Consequent Subject Association Rules**
July 2016 - October 2016	60,388	33,060
November 2017 - February 2017	145,304	86,000
March 2017 - June 2017	250,000	131,885

Table 1. Anonymized transactions (or checkouts) and consequent subject association rules since beginning data stream collection in July 2016

The research team used the data mining tool WEKA to run a machine learning process offline (Eibe, *et al.*, 2016). Once a consequent rule set for clusters of topic data are generated, this rule set is then stored in the secure library database server to be leveraged by the Library Mobile App Recommendation API when a person uses the recommendations module. The recommendation module checks the topics from the items in their Favorites and Checked-out modules against a rule set generated by clusters of checkout topic data. This is used to run a targeted search

with related topics derived by our machine learning process. The algorithm generates recommendations by filtering for candidate recommendations that are popularly circulating within the university's integrated library system reporting database. Highly circulating items are suggested to the user if the topic association is represented. Figure 1 illustrates the data sources that the Recommendation module's API relies on.

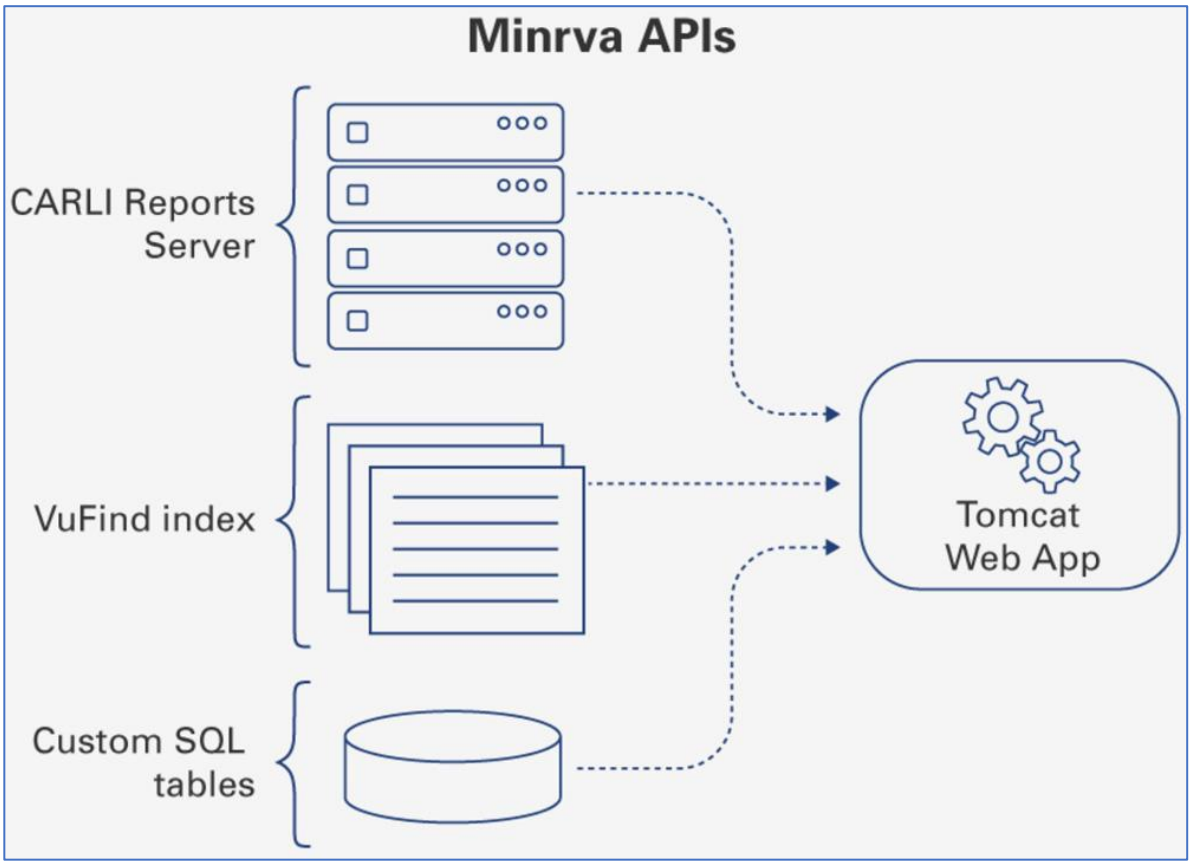

Figure 1. Topic association rules are stored in custom SQL tables. Subject filtering from the VuFind index is built into the server side business logic. Finally, a reports database server is used for deriving popularity rank

Using Splunk Enterprise analytic software over Library Mobile App weblogs, researchers found that from October 2016 – June 2017 the Recommendation APIs have recorded 5,728 events related to users browsing for recommended items based on checked out items. Table 2 below indicates monthly uses of the recommender module from within the Library Mobile App. The uses mirror typical library activity during semesters when school is in session and heavier use during the fall semester months.

Month Year	No. of recommendation module events by users
Oct-2016	1,716
Nov-2016	1,100
Dec-2016	525
Jan-2017	545
Feb-2017	352
Mar-2017	331
Apr-2017	476
May-2017	437
Jun-2017	246

Table 2. Number of recommendation module events by users since the service became available in the mobile app on October 2016

The machine learning workflow described above utilizes data streams of transactional data which are mapped to subject metadata, which is then tied to server side subject based searches with topic metadata from the user's VuFind account. It is desirable and necessary to extend this basic personalization service

into an open algorithm; to do so, larger sets of data and additional test environments will be required. Furthermore, a future goal for the recommendation is to design a more versatile user response system in order to encompass user ratings from the provided recommendations. A more responsive recommender system would result from the incorporation of personal feedback extending the inputs of the system, beyond what is currently checked out, and into more immediate areas of interest to the user.

Big Data

Personalized portals within libraries have not yet made use of ensemble methods of data mining and drawing on individual information (current checkouts, enterprise affiliations, department information, course registration history, and curriculum vitae in community generation). User and system data are profoundly crucial to informing the production of useful and relevant recommendation results, a key strategy for which will be building and integrating quality data corpuses. For open personalized recommendations, the authors hypothesize that large heterogeneous data sets will boost performance dramatically. Gathering data streams from complementary systems will be instrumental in testing and shaping a personalized recommendation algorithm.

The open algorithm will learn by several methods. The first level of recommendations, covered in covered in the previous section is derived from topics modeled from checkout streams which the integrated library system has been continuously collecting since July 2016 for the purposes of this project. These streams are topically valuable since they include sets of items that are checked out together. The second method is by looking at user actions such as favorite items and metadata within their account. A user's current interests are informed by items checked out to the user and by chronological data mining sourced from when an item in a user account is renewed.

Further improvement of the recommender's performance can be achieved by making a basic rating feature available from the app itself to rate generated recommendations, supplemented by click stream data provided by mobile analytics software. This feedback will factor back into the personalization filter. Other data points that a user's personalized filter could encompass would afford the user the ability to exclude subject areas they are specifically not interested in receiving additional recommendations. If the user has the functionality to curate their subject targets, this would help the algorithm learn items of more immediate interest to the user. In keeping with the aims of open discovery innovation, the authors propose directly integrating the recommendation algorithm and

implementation into a future version of the VuFind discovery tool. Direct integration would offer several advantages such as speeding up the service by using the native index topic search, rather than relying on custom API overlays to perform searching from a web based API service.

Privacy

There are several privacy considerations and risks that have been addressed with a privacy policy authored by the university library covering the usage of user provided data in a recommender from the Library Mobile App.[2] Considerations for the protection of human subjects related to data mining patron data include subject privacy, data confidentiality, and consent. The researchers are interested in relatedness among collections or what collections should be recommended from a local user account.

Future research will explore a continuum of data points for providing recommendations so that within user communities, clusters of users who belong to similar communities by department or major may be useful to use as baseline comparisons in development of personalized recommendations. In terms of the data mining component the research team will work to completely de-identify

data. There are risks involved in re-identification after the dataset is constructed, therefore the team will systematically study the best way to ensure user privacy and maintain security of user data.

Building on the privacy policy established in the pilot test of basic recommenders, the project team plans to update and revise the existing policy in order to make general recommendations and policy guidelines for user data protection relevant to the larger scale of libraries nationally. This could help research libraries begin to successfully navigate the policy questions introduced by large scale data mining of library systems.

Conclusion and Next Steps

Most users of academic libraries who log in to their user accounts will never know to search the collections of other academic libraries nationally, nor are they likely to be aware of all the potentially relevant resources within their own library consortia, university system libraries, or regional networks. Personalized recommendations can increase access, use, and impact of the investments in digital content and research collections globally.

A truly useful recommender will result if heretofore untapped novel datasets extracted from university library data stores are utilized in providing future

recommendations for users. Intelligently mined recommendations offer new insights into information needs and providing the best digital library resources available. At the same time, leveraging user interactions with the provided recommendations will help the algorithm filter for individual preferences.

The example described in this paper provides a useful test case in loosely-coupled recommendation overlays for library systems. As the middleware solution developed for this project shows, the system could extend to other discovery environments such as VuFind without extensive re-engineering. Such an approach to open algorithms would be important to making this work extensible to a broad audience of research libraries, and useful for open discovery environments worldwide.

This work is valuable to library discovery generally, in part because such an approach helps to support researchers and scholars in ways that have previously been overlooked and underdeveloped within the research library community. It is the aspiration of the authors to increase discovery of unique digital content as well as of research library holdings which have heretofore been overlooked by users of academic library discovery environments. Further work on account based recommenders will focus on focus groups of users of this service in order to better understand what actual users of the service are looking for and if there are data

points which should be considered that have not already been integrated into the service. Further refinement of the account-based recommender will necessarily require a mixed method research approach combining elements of the quantitative use and qualitative inputs.

References

Amatriain, X., 2013. Mining Large Streams of User Data for Personalized Recommendations. *ACM SIGKDD Explorations Newsletter* 14, 37–48. doi:10.1145/2481244.2481250

Antelman, K., Lynema, E., Pace, A.K., 2006. Toward a Twenty-First Century Catalog. *Information Technology and Libraries* 25, 128–139. doi:10.6017/ital.v25i3.3342

Chickering, F.W., Yang, S.Q., 2014. Evaluation and Comparison of Discovery Tools: An Update. *Information Technology and Libraries* 33, 5–30. doi:10.6017/ital.v33i2.3471

Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., Sampath, D., 2010. The YouTube Video Recommendation System, in: *Proceedings of the Fourth ACM Conference on*

Recommender Systems, RecSys '10. ACM, New York, NY, USA, pp. 293–296. doi:10.1145/1864708.1864770

Dibbell, J., 2013. "Pew study: Library patrons want personalized recommendations," *The Thingology Blog*, available at http://blog.librarything.com/thingology/2013/01/pew-study-library-patrons-want-personalized-recommendations/ (accessed January 31, 2017).

Eibe, F., Hall, M.A., Witten, I.H., 2016. *The WEKA Workbench*, in: Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques." Morgan Kaufmann.

Geyer-Schulz, A., Neumann, A., Thede, A., 2003. An Architecture for Behavior-Based Library Recommender Systems. *Information Technology & Libraries* 22, 165–174.

Hahn, J. (2011), "Location-based recommendation services in library book stacks", *Reference Services Review,* Vol. 39, No. 4 pp. 654-672.

Hanson, C., Nackerud, S., Jensen, K., 2008. Affinity Strings: Enterprise Data for Resource Recommendations. *The Code4Lib Journal*. 5.

Kim, Y., Gyo Chung, M., 2008. Personalised information services using a hybrid recommendation method based on usage frequency. *Program* 42, 436–447. doi:10.1108/00330330810912106

Liao, I., Hsu, W., Cheng, M., Chen, L., 2010. A library recommender system based on a personal ontology model and collaborative filtering technique for English collections. *The Electronic Library* 28, 386–400. doi:10.1108/02640471011051972

Linden, G., Smith, B., York, J., 2003. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing* 7, 76–80. doi:10.1109/MIC.2003.1167344

Lown, C. et al. 2013. How Users Search the Library from a Single Search Box. *College & Research Libraries*. 74, 3 (May 2013), 227–241.

Moore, K.B. and Greene, C. 2012. Choosing Discovery: A Literature Review on the Selection and Evaluation of Discovery Layers. *Journal of Web Librarianship*. 6, 3 (Jul. 2012), 145–163.

Rochkind, J. 2013. A Comparison of Article Search APIs via Blinded Experiment and Developer Review. *The Code4Lib Journal*. 19.

Svenonius, E. 2000. *The Intellectual Foundation of Information Organization*, MIT Press, Cambridge, MA.

[1] https://minrvaproject.org/download.php

[2] https://sif.library.illinois.edu/prototyping/RecPrivacyPolicy.html