Qualitative and Quantitative Methods in Libraries (QQML) 1: 191 –198, 2014

aiSelections: Computational Techniques for Matching Faculty Research Profiles to Library Acquisitions

Peter M. Broadwell[1] and Timothy R. Tangherlini[2]

[1]Digital Initiatives and Information Technology, U.C.L.A. Library
[2]Scandinavian Section and Department of Asian Languages and Cultures, U.C.L.A.

Abstract. As institutional libraries of all sizes adopt more focused acquisitions policies, subject librarians and other selectors will benefit from sophisticated computational approaches that help to identify the monographs, serials, and electronic resources that are likely to receive the most use, thereby reducing interlibrary loan requests, special orders, and unused materials. We describe a pilot study in which data-mining software tools and algorithms were used to summarize faculty biographies, publications, and departmental curricula and to match the resulting profiles to potential monograph selections. We evaluate the effectiveness of these tools by examining the circulation records, interlibrary loan requests, and purchasing receipts from the past several years, noting the computational techniques that are most likely to improve selection accuracy.
Keywords: selections, computation, matching, data mining, research profiles

1. Introduction

This project explores the promising intersection of two continuing trends in computational support for academic scholarship: the online aggregation of scholarly publishing data and the use of advanced statistical analyses to streamline library collections development. By demonstrating how these technologies can interact productively in the environment of a university library, this project supports an emerging conception of the 21st-century academic library as a clearinghouse of digitized institutional research data and a centralized facilitator of novel collaborative research. See CLIR (2010) for an overview of these ideas.

The online aggregation of scholarly publishing data has promoted the field of "bibliometrics" to the forefront of discussions of scholarly reputation, credit, and authority, encouraging the use of bibliographic statistics such as citation counts and journal "impact factors" when evaluating a scholar's contribution to a field. Other, more collective pursuits, including scholarly social networks such as academia.edu and databases like Google Scholar, now enable the aggregation

Received: 21.4.2013 / Accepted: 4.3.2014 ISSN 2241-1925

of a wealth of information about researchers' scholarly interests. Academic institutions also have begun to prioritize the creation of local databases of information about their faculty members' research interests and publications in order to streamline evaluation processes and to encourage interdisciplinary collaboration. The university library has an opportunity to play a major role in this new development, particularly as academic libraries begin to build and manage centralized digital "institutional repositories" of the collected research work and, in some cases, research data produced by the institution's faculty.

This study inhabits the point at which the trends in bibliometrics discussed above meet "bibliomining," the notion of using statistical data to guide library collections development. This idea has a long history, beginning as early as Greaves's (1974) foundational report advocating the use of circulation statistics in acquisitions decisions. Practical studies in this field, such as Ward et al. (2003) and Knievel et al. (2006), have proliferated in recent years. Nicholson (2003) also has been active in promulgating the term "bibliomining" to describe this growing body of technology-driven collection development practices.

In contrast, the status of the University of California, Los Angeles as a major research institution has led to a library collection development philosophy built around a large staff of subject specialists who personally select the materials in their area of expertise. Although subject selectors now utilize approval plan lists provided by vendors, selections has remained a primarily manual process. Yet in recent years, financial pressures, advances in collections management technology, and the changing roles of university libraries also have led the UCLA Library to evaluate more technologically sophisticated approaches to collections development.

2. Experimental Design and Data Collection

For this pilot study, we selected a well-bounded library subject area and time period, gathering sufficient data to build a rudimentary statistical profile of faculty members in relevant departments and to evaluate computational methods for matching this profile to potential monograph acquisitions. We chose a library subject area at UCLA that primarily serves two humanities departments of approximately three dozen faculty members and a similar number of graduate students, most of whom still consult print monographs for much of their primary- and secondary-source research. These departments provide instruction to several hundred undergraduate students per year; the subject library's holdings are also available to the university as a whole.

To construct a digital faculty research profile, we downloaded the texts of faculty biographies and publication lists from departmental web pages into text files. We then used the open-source RapidMiner data-mining environment to distill these text files into a term array that associates words with counts of their overall corpus frequency and the number of documents in which each term appeared. We employed text filtering and analysis to remove irrelevant terms from the final array, using both common and customized "stopword" lists. We also applied linguistic "stemming" algorithms and used RapidMiner's integration into the WordNet lexical database to reduce equivalent terms to their

root forms. As a final step, we performed a limited expansion of the term set by importing synonyms from WordNet, and computed the resulting frequencies of single terms as well as bigrams (phrases consisting of two consecutive words) in the combined text.

To obtain "ground truth" data about the monographs that were actually acquired and circulated during the period of the study, we queried the database records in UCLA's Voyager integrated library system and stored the results in a custom MySQL database. Table 1 summarizes the acquisitions and circulation records we obtained. Because this is a study of research-related monographs, we excluded general reference works that were classified as non-circulating. And in order to keep the scope of the project manageable, we used the Library of Congress Classification system class ranges specified in the collection development policy of the target library subject area to define the boundaries of our data gathering for this and subsequent queries.

Monograph Records from Voyager	**In subject area**	**In subject area and published 2006 and later**
Acquisitions (2007-)	13,577	10,471
	Firm orders	3,573
	Approval plan orders	6,989
Circulation (2008-)	32,668 unique titles, 81,733 circulation records	4,118 unique titles (12.6% of all circulating titles)
ILL borrowing (2008-)	3,366 (764 later acquired)	764 (210 later acquired)
	Titles acquired and circulated	4,118 (39.3% of acquisitions)
	Firm orders	1,616 (45%)
	Approval plan orders	2,567 (36.7%)

Table 1: Actual monograph acquisition, circulation, and interlibrary loan figures in the targeted subject area and date range. The circulation rate is especially impressive given that the acquisitions budget was significantly reduced during this period, especially from 2009-2011. Note that approval plan orders benefit from a considerable degree of manual filtering and basic computerized screening by vendors and librarians – see Fenner (2004).

In order to measure the "miss rate" of the collection, we also obtained the interlibrary loan borrowing records for monographs in the targeted field from January 2008 to February 2013 using the VDX interlibrary loan management systems hosted by OCLC WorldCat and also (for older records) the California Digital Library.

The most crucial requirement of our proposed analysis was the ability to compare the circulation rate of the monographs that actually were acquired in the targeted subject area and time frame to the circulation rate of the monographs that would have been acquired if they had been selected according to the recommendations of our computational selections system. To carry out this comparison effectively, we needed to obtain detailed descriptions of a sizeable portion of all books available for acquisition during the time period

under consideration. We first sought to acquire this information from vendor-produced catalogs and search indexes such as the Global Online Bibliographic Index (GOBI) service, but found that these services were not able to provide us with a suitably comprehensive listing of all available monographs.

To obtain the bibliographic details of a sizeable majority of the monographs available for ordering in the targeted subject area, we ultimately used the application programming interface of the online OCLC WorldCat Search service to query book listings in the subject ranges and languages specified in the library's collection development profile. We chose to limit the publication date range of these books to the years 2006 through 2013 in order to reduce the likelihood of selecting out-of-print editions. This query returned 150,877 distinct monograph publications within the specified subject and date ranges. These records matched 87.9% of the monographs in our Voyager acquisition records, and 90% of the books in our circulation records. The remaining titles, we determined, were those that fell outside the LCC class ranges specified in the collection development policy but had been acquired regardless, likely due to a patron request. We excluded these records from our analysis, as well as titles published prior to 2006, in order to focus on evaluating the effectiveness of faculty profile-driven selection of relatively new titles, which constitute the majority of yearly acquisitions though typically not the greater proportion of circulation records.

We modified scripting software from the OCLC Developer Network to parse the MarcXML records we received in response to a search query, and used the OCLC accession number, ISBN (when available), Library of Congress call number, year published, and language information to help catalog and disambiguate the records retrieved. We stored the texts of the title, Library of Congress subject and subject place headings, table of contents listings, and publisher descriptions (when available) in an array of relevant terms for each individual record and also for the entire corpus of available books. We accomplished this task using the same sequence of text processing steps in the RapidMiner environment as those applied to the faculty biographies and publication listings described above. As a consequence, it was possible to compare the resulting text vectors directly and accurately.

3. Experimental Results

Our primary technique for experimentally evaluating the performance of a given computational selection method was to use it to rank all of the 150,877 potential acquisitions published from 2006 to the present according to some calculated measure of their likelihood of circulating, and subsequently to compare the actual observed circulation records against the monographs that would have been purchased using the ranking algorithm. Because 10,471 books actually were acquired from 2007 to 2013, we used this figure as the upper limit of the total number of books selected in each simulation.

As Table 1 indicates, 4,118 of the acquired books from the total WorldCat search results actually circulated in the time period under consideration, and an additional 606 were borrowed through interlibrary loan, so only 4,724 / 150,887

= 3.1% of the new titles available for purchase at this time actually were ordered and subsequently checked out by patrons. Therefore, if we were to select 10,471 titles completely at random, we would expect to see only 10,471 * 3.1% = 325 of these circulate. Of course, we would prefer that the computational methods at least approach the benchmark of 4,118 circulating books (39.3% of acquisitions) achieved by the library selectors using approval plan lists and firm orders. We had also hoped to calculate the total purchase costs and funds spent per circulating acquisition in the automated scenario, but unfortunately the pricing information in the WorldCat records is not sufficient to facilitate such an evaluation.

Our most effective computational method for ranking the suitability of potential acquisitions used the text-mining technique of **cosine similarity** calculation to quantify the degree of similarity between the sets of words appearing in the faculty research profile and the array of descriptive terms derived from the WorldCat catalog records for each book. See Hastie et al. (2009) for a more in-depth description of the cosine similarity algorithm; it essentially summarizes a given document as an arrow (technically a vector) surrounded by words in some high-dimensional space; the arrow points towards the words that appear more frequently in the document, and away from those that appear less frequently (or not at all). The "cosine similarity" between two documents, then, is a measurement of the *angle* between their two vectors; similar documents will have a smaller angle between them.

Considering all of the faculty profiles as a single document vector and ranking the potential monograph acquisitions in the WorldCat search results in order of decreasing cosine similarity to this document yielded 1,873 acquisitions that would have circulated according to the actual circulation and interlibrary loan records, producing a circulation rate of 17.9% of the 10,471 titles ordered in the simulation. An additional 3,193 of the recommended acquisitions were actually selected by the subject librarians but did not circulate.

To determine the upper limit on the performance of the automated selection algorithms, we used the target data – in this case, the records of the monographs that actually circulated – in place of the experimental faculty research profile. Re-running the cosine similarity-based simulation with these inputs generated 1,899 circulating new acquisitions, for a precision of 18.1%, indicating that the performance of the faculty research profile matching is nearly optimal. This does not mean that it is impossible to improve upon these results, but rather that the cosine similarity ranking would need more complete and sophisticated data about the research profiles and candidate monographs to improve its performance significantly.

Another evaluation method we pursued was to compare the performance of automated selections driven by faculty research profiles to the selection of materials for acquisition based upon the characteristics of books that have circulated recently – a technique that libraries increasingly have been adopting. One of the most popular computational approaches to the latter task is to "train" a statistical model called a naïve Bayes classifier on the acquisition and circulation records from previous years, and then to use the trained model to

classify potential acquisitions according to the likelihood that they will circulate – a probability calculated based upon features they share with circulating and non-circulating items from the past, as well as the proportion of recent acquisitions that actually circulated. This type of machine learning classifier is also a popular technique for filtering unwanted "spam" email – see Sahami et al. (1998). The equivalence here is apt but perhaps not very charitable to the monographs rejected by the algorithm.

We chose to train the Bayesian classifier on a subset of our historical data from the Voyager catalog, specifically the acquisition and circulation records from January 2007 to December 2009. We then compared the titles the classifier recommended from January 2010 to February 2013 to the actual acquisition and circulation records in this period. Given a limit of 5,077 simulated acquisitions from that time, the Bayesian classifier selected 640 monographs that subsequently circulated (12.6% of the total acquisitions). Running the cosine similarity recommendation algorithm based on the faculty profiles produced slightly better performance for this period: 721 "hits" (14.2%). This result suggests that collection development based upon previous circulation records may be less effective than a technology-assisted policy that takes into account faculty research interests. It is very difficult, however, to predict how either computational model would improve if provided with more detailed data; Bayesian classifiers in particular grow increasingly accurate as they amass incrementally greater volumes of training data.

4. Conclusions and Future Work

This paper describes a pilot study conducted at the University of California, Los Angeles to evaluate the effectiveness of computationally matching faculty research profiles to potential subject-area acquisitions using text data-mining techniques. We limited our initial study to monograph circulation statistics and acquisitions records from the past six years in a well-bounded library subject area in the humanities. These data enabled us to evaluate the effectiveness of different approaches to building statistical summaries of faculty biographies, publications, and course descriptions, and ranking potential acquisitions based upon their semantic similarities to this faculty-based digital selection profile.

Our research determined that document matching via cosine similarity calculations provided the most accurate recommendations, achieving an estimated circulation rate of 17.9% for materials acquired in the last 6 years. This was roughly half the performance of the actual selections process during this period, which involved a human selector, vendor approval plans, and firm order requests. Our initial analysis suggests that data-mining techniques are unlikely to outperform a knowledgeable subject area specialist unless additional data is available to build more sophisticated profiles of the target patrons as well as the materials to be evaluated for acquisition. In the absence of such data, the technological approaches described here are best used to facilitate the work of a knowledgeable subject specialist, reducing the volume of potential acquisitions under consideration by emphasizing high matches and excluding those with very low scores.

We believe that the aggregation of data about faculty research interests and the use of advanced statistical tools in collections development will continue to be of great interest to university libraries, and will help these libraries to establish a central position in the 21st-century academic landscape. More raw data about faculty interests will become available as librarians archive the full texts, data sets, and primary sources of their research in digital institutional repositories. Data from online research interest surveys and academic social network profiles also would enable statistics-driven library selection tools to take advantage of the sophisticated methods developed in the computational field of collaborative filtering, which now drive the ubiquitous online "recommendation engines" that match users to products and entertainment options based upon reviews and previous behavior.

As a final consideration, it is possible that e-books, next-day document delivery and on-demand, "just in time" acquisitions soon will render the predictive monograph acquisition processes described in this paper largely unnecessary. But it is also probable that the faculty interest-based ranking approaches discussed above can be applied just as effectively to the selection of costly limited-access electronic resources, such as online database and journal subscriptions. Tools like "aiSelections" may become indispensible in helping librarians to make discerning, well-informed choices regarding the online subscriptions that will provide the greatest value to their patrons.

5. Acknowledgments

The initial idea for this project emerged from discussions between the authors and Todd Grappone, Associate University Librarian for Digital Initiatives and Information Technology at UCLA. The project could not have been completed without his assistance. The authors also thank UCLA University Librarian Gary Strong and Associate University Librarians Sharon Farb and Kevin Mulroy for their support, as well as the many library technology specialists and subject librarians whose assistance enabled this study become a reality.

References

Council on Library and Information Resources (2010). *The Idea of Order: Transforming Research Collections for 21st Century Scholarship*. Washington, D.C. http://www.clir.org/pubs/abstract/pub147abst.html

Fenner, Audrey (2004). The Approval Plan. *The Acquisitions Librarian*, Vol. 16.

Greaves, F. L., Jr. (1974). The Allocation Formula as a Form of Book Fund Management in Selected State-Supported Academic Libraries. Florida State University, unpublished doctoral dissertation.

Hastie, T., R. Tibshirani, and J. H. Friedman (2009). *The Elements of Statistical Learning: Data mining, Inference, and Prediction*. Springer Series in Statistics. Springer, New York, NY.

Knievel, Jennifer E., Heather Wicht and Lynn Silipigni Connaway (2006). Use of Circulation Statistics and Interlibrary Loan Data in Collection Management. *College & Research Libraries*, Vol. 67, No. 1 (January 2006), 35-49.

Nicholson, Scott, Stanton, Jeffrey M. (2003). Gaining Strategic Advantage through Bibliomining: Data Mining for Management Decisions in Corporate, Special, Digital,

and Traditional Libraries. *Organizational Data Mining: Leveraging Enterprise Data Resources for Optimal Performance*. Idea Group Publishing, Hershey, PA, 247-262.

Sahami, Mehran, Susan Dumais, David Heckerman, Eric Horvitz (1998). A Bayesian Approach to Filtering Junk E-Mail. Microsoft Research, Redmond, WA.

Ward, Suzanne M., Tanner Wray, Karl E. Debus-López (2003). Collection Development Based on Patron Requests: Collaboration between Interlibrary Loan and Acquisitions. *Library Collections, Acquisitions, and Technical Services*, Vol. 27, Issue 2 (Summer 2003), 203-213.