The Journal of Academic Librarianship xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

The Journal of Academic Librarianship

journal homepage: www.elsevier.com/locate/jacalib

Automated Classification to Improve the Efficiency of Weeding Library Collections

Kiri L. Wagstaff*, Geoffrey Z. Liu

School of Information, San Jose State University, One Washington Square, San Jose, CA 95192-0029, United States

A B S T R A C T

Previous studies have shown that weeding a library collection benefits patrons and increases circulation rates. However, the time required to review the collection and make weeding decisions presents a formidable obstacle. This study empirically evaluated methods for automatically classifying weeding candidates. A data set containing 80,346 items from a large-scale weeding project running from 2011 to 2014 at Wesleyan University was used to train six machine learning classifiers to predict a weeding decision of either 'Keep' or 'Weed' for each candidate. The study found statistically significant agreement ($p = 0.001$) between classifier predictions and librarian judgments for all classifier types. The naive Bayes and linear support vector machine classifiers had the highest recall (fraction of items weeded by librarians that were identified by the algorithm), while the k-nearest-neighbor classifier had the highest precision (fraction of recommended candidates that librarians had chosen to weed). The variables found to be most relevant were: librarian and faculty votes for retention, item age, and the presence of copies in other libraries.

Introduction

As library collections grow and patron needs evolve, there is an ongoing need for reviewing and maintaining physical collections. A key component of this process is weeding, the selective removal of items that are outdated, physically worn, no longer relevant to patron interests and needs, and/or available in electronic form. Librarians generally agree that weeding benefits not only the library by reducing the number of items that have to be maintained, but also the user population by making desired items easier to find (Dilevko & Gottlieb, 2003). There is also a general belief that pruning the collection to remove unwanted items can increase library circulation rates, although experimental studies assessing the effect of weeding on circulation have yielded mixed results (Moore, 1982; Roy, 1987; Slote, 1997). Weeding creates space that can be used for new acquisitions or to support other library needs, such as programming, maker spaces, or study areas (Lugg, 2012; Slote, 1997). Despite these benefits, weeding is often low on the priority list for busy librarians. Only 24% of libraries weed continuously, and 39% weed at regular intervals (Dilevko & Gottlieb, 2003).

Multiple factors contribute to librarians' reluctance to weed their collections. Weeding provides no immediate observable benefit. As noted by Dilevko and Gottlieb (2003), weeding can impose a psychological strain on those tasked to implement it, and many librarians find it stressful to make the decision to discard an item. One of the largest obstacles is that weeding is extremely time-consuming. Making a decision about a single title can take several minutes (Zuber, 2012), and the amount of reviewing that can be done is limited by the number of people who can devote time to the task. Large-scale weeding projects can require reviewing tens of thousands of titles, and the weeding project can take years to complete. For example, Concordia University reviewed 25,000 books per year for two years, weeded a total of 12,172 items before deciding that this level of review "could not be maintained" and consequently reduced the review rate by 50% (Soma & Sjoberg, 2010). Monmouth University librarians took two years to review 72,500 items and select 12,800 for removal (Dubicki, 2008). Rollins College weeded 20,000 from a collection of 286,000 items over two years (Snyder, 2014). Wesleyan University weeded 46,000 of approximately 90,000 candidates over three years, from 2011 to 2014. They began by identifying 90,000 weeding candidates that were reviewed individually by the librarians and then reviewed again by interested faculty members (Tully, 2011). The project involved 17 librarians, two consultant subject specialists, and approximately 20 staff members, plus two new employees (a reference librarian and a staff member) who were hired specifically to support the project (P. Tully, personal communication, October 16, 2014).

One possible way to remove the time obstacle and reduce librarians' psychological stress is automation. Existing methods (e.g., Lugg, 2012; McHale, Egger-Sider, Fluk, & Ovadia, 2017) enable librarians to specify

* Corresponding author. Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, CA 91109, United States.
E-mail addresses: wkiri@wkiri.com (K.L. Wagstaff), geoffrey.liu@sjsu.edu (G.Z. Liu).

https://doi.org/10.1016/j.acalib.2018.02.001
Received 30 October 2017; Received in revised form 31 January 2018; Accepted 6 February 2018

a set of weeding criteria and apply them to a library's circulation records to generate the initial list of weeding candidates automatically. Each candidate is manually reviewed and marked 'Keep' or 'Weed'. However, for the purposes of weeding, each candidate that is labeled 'Keep' on the initial list represents an unproductive expenditure of the librarian's time. An ideal candidate list would be one that contains only items that the librarian would agree to weed. There is potential for significant time savings if an automated method could be employed to filter and refine the list of weedable candidates.

This paper reports on an experimental study that was designed to assess the potential of improving weeding efficiency by using a data mining approach. Specifically, using existing records from the Wesleyan University Library's weeding project, a set of automated classifiers based on different machine learning algorithms were trained to predict the librarians' weeding decisions. The study found statistically significant agreement between the automated classifiers' predictions and the librarians' weeding decisions.

Prior work on weeding library collections

There are two primary approaches to weeding a collection: inclusive and exclusive. The inclusive approach considers each item in turn to decide whether it should be weeded or kept (Soma & Sjoberg, 2010; Tully, 2011), as exemplified by the widely employed Continuous Review, Evaluation, and Weeding (CREW) method (Larson, 2012). In contrast, the exclusive approach first identifies the "core collection" for the library and then weeds items that fall outside of this subset (Trueswell, 1966). In either approach, the weeding process ultimately comes down to deciding if a given item ought to be removed or not based on criteria typically formulated in terms of some conditional factors and instilled in the library's collection development/weeding policy.

Factors in weeding decisions

According to Dilevko and Gottlieb (2003), the criteria most often used by librarians to make weeding decisions were circulation statistics, the physical condition of the item, and the accuracy of its information. This is consistent with the CREW method's advice to weed items with low circulation, poor appearance, or poor content (Larson, 2012) and the methods used by previous weeding projects such as those of the University of Toledo (Crosetto, Kinner, & Duhon, 2008), Concordia University (Soma & Sjoberg, 2010), and Rollins College (Snyder, 2014). A summary of the criteria used by several weeding projects to identify candidates is given in Appendix A. The following sections examine each of the key factors in weeding decisions identified from the literature.

Circulation records

Many weeding efforts are motivated by the empirical observation that a large fraction of the library collection never circulates (Kent et al., 1979; Silverstein & Shieber, 1996; Soma & Sjoberg, 2010). Non-circulating items can be a liability for libraries in that they consume shelf space and resources but do not directly benefit patrons. They also reduce the library's overall circulation rate. High circulation is valued by librarians because it contributes to a feeling that the library is "serving its community well" (Dilevko & Gottlieb, 2003, p. 93).

Slote (1997) surveyed the literature on weeding and found that past use of items consistently emerged as the best single criterion for making weeding decisions. One way to characterize past use is the measure of an item's "shelf-time", i.e., the length of time that has elapsed since the item last circulated. Slote advocated shelf-time as the most reliable criterion for determining objectively which books could be weeded with the least impact on patron needs. In his own 1969 study of five libraries, he found that "past use patterns, as described by shelf-time period, are highly predictive of the future use, and can be used to create meaningful weeding criteria" (p. 63). However, Goldstein (1981) studied eleven libraries and found that none of them took shelf-time into consideration when making weeding decisions, although they did employ use statistics (e.g., number of checkouts) as a weeding criterion. Others have argued that demand (number of checkouts per year) may be more informative than shelf-time (Snyder, 2014).

Circulation records alone may be insufficient for making informed weeding decisions. Some studies found that in-house use mirrors that of circulation, while others found that they could be quite different. Selth, Koller, and Briscoe (1992) found that 11% of the books in their library had in-house use but zero circulation. Weeding based only on circulation records could potentially remove these items despite their evident popularity and utility for visiting patrons (Slote, 1997). The process of weeding requires the examination of these and other additional factors, which increases the time required to evaluate weeding candidates accurately.

Physical condition

Libraries seek to provide materials that are in a useful state. Items that have been damaged (e.g., food spills, ripped pages, water damage, weakened spines, missing pages) are less valuable to patrons and may even become unusable. As items age, they become more vulnerable to physical decay and damage. Sometimes items can be repaired. If they are deemed unusable, the library must decide whether to simply discard the item or to replace it based on the value of its content to the user community.

Quality of content

The CREW manual identifies six negative factors that relate to the quality of an item's content and summarizes them with the acronym "MUSTIE" (Larson, 2012, p. 57). These negative factors are: Misleading (or factually inaccurate), Ugly (worn), Superseded, Trivial (no longer of literary or scientific merit), Irrelevant (to the user community), or the same information can be easily obtained Elsewhere (e.g., interlibrary loan or electronic format).

Additional factors

Several other factors may be used by librarians in making weeding decisions. They may consider whether the item is a duplicate of other items in the same collection and whether it is held by other libraries or available in digital form (Metz & Gray, 2005). They may consult book reviews or canonical bibliographies, assess local relevance, track in-house use of the item, and consider unique features of the book. Soma and Sjoberg (2010) developed a standard checklist to be used by all librarians participating in a collaborative weeding effort. The checklist included circulation and browse statistics as well as an indication of whether the item appeared in *Resources for College Libraries* and how many copies were held by other libraries.

Faculty input

Weeding is not always viewed favorably by library patrons, and involving them in the process is helpful. For academic libraries, some faculty members may oppose the entire project and refuse to sanction the removal of any titles. Some are concerned about the loss of the scholarly record or institutional prestige (Dubicki, 2008). In a psycholinguistic analysis of relevant literature, Agee (2017) identified several negative emotions expressed in faculty responses to weeding projects, which include anger, sadness, and anxiety, in decreasing order of occurrence. Public library patrons may disapprove of discarding items purchased with tax dollars. To overcome opposition to weeding, librarians often devote time to educating and involving patrons. For example, Wesleyan University librarians attended several faculty meetings and set up a website for interested faculty to review the candidates and vote on which ones should be retained (Tully, 2012). Olin Library at Rollins College also invited patrons to participate in the weeding process. Weeding candidates were flagged but remained on the shelf for two months, during which time faculty members were