Affinity Strings: Enterprise Data for Resource Recommendations

The University of Minnesota Libraries have created a MyLibrary portal, with databases and e-journals targeted to users, based on their affiliations. The University’s enterprise authentication system provides an “affinity string”; now used to personalize the MyLibrary portal. This affinity string automates discovery of a user’s relationship to the University—describing a user’s academic department and degree program or position at the University. Affinity strings also provide the Libraries with an anonymized view of resource usage, allowing data collection that respects users’ privacy and lays the groundwork for automated recommendation of relevant resources based on the practices and habits of their peers.

by Cody Hanson, Shane Nackerud, and Kristi Jensen

The MyU Portal

The MyU Portal is a campus-wide web portal running on the open-source Metadot Portal Server. Created by the University’s Office of Information Technology and Office of the Vice Provost for Distributed Education and Instructional Technology, the site is intended as “the first stop for University of Minnesota information and applications” for students, staff, and faculty [1]. By bringing together feeds of data from many enterprise applications and filtering them based on a user’s identity, the MyU Portal provides a personalized dashboard.

The interface to the MyU Portal presents several web pages, each accessible through a tabbed interface. The My Courses tab lists details of academic courses a user may be taking or teaching. My Work Life provides payroll and benefits information for full-time or student employees. Access to this sensitive personal information is provided on the basis of user authentication with the University’s enterprise Central Authentication Hub.

While the specific information on a tab can be personalized for a user, large parts of the MyU Portal interface can also be customized for categories of users. For example, incoming students accepted for fall 2008 enrollment have had access to the MyU Portal since spring. When they log in, they see a “view” of the MyU Portal that is heavily customized for the Class of 2012. Once those students choose a major, their portal view will be customized to their specific degree program. Undergraduates who log in to the Portal see information relevant to undergrads. Graduate students see information relevant to grads. More specifically, grad students from the Carlson School of Management see information relevant to their studies at that particular school. There are many different views available on the Portal, and all of them are geared towards specific audiences. This works through a unique piece of metadata, called an affinity string, which is assigned to each University of Minnesota user.

What is an affinity string?

Affinity strings are created by the Office of Information Technology using data from the University’s PeopleSoft system. Everybody at the University of Minnesota is assigned one or more affinity strings based on her area of study or work.

For example, a Ph.D. student in Psychology may receive the affinity string: “tc.grad.gs.psy.phd.”
The segment “tc” indicates that the student is on the Twin Cities campus of the University. “grad” denotes that she is a graduate student, “gs” that she is enrolled in the Graduate School, “psy” that she is in the Psychology Department, and “phd” that she is in the Doctoral degree program.

As another example, a faculty member in the Carlson School of Management may receive the affinity string “tc.fac.csom.327A.9403,” in which the last two segments correspond to PeopleSoft job codes that indicate the faculty member’s department and rank, in this case Assistant Professor of Accounting.

There are over 15,000 affinity strings covering all the disciplines and roles at the University. The MyU Portal managers have mapped each string to a particular Portal view, assuring that each user gets information relevant to her department, area of study, or affiliation with the University.

MyLibrary

As acolytes of Lorcan Dempsey’s notion of placing library services “in the flow” of users, the University’s librarians recognized the traffic on the MyU Portal as a rich workflow in which it was natural to embed our resources. The Libraries lobbied the Portal managers for, and were granted, a dedicated tab, which we labeled “MyLibrary.” Responsibility for planning and development of the MyLibrary tab was given to a group of Librarians and Library IT professionals and was overseen by the director (and later the interim director) of the Libraries’ Digital Library Development Lab.

The design of the MyU Portal limits each tab to a narrow column on a single page. Recognizing this constraint, and that users in the MyU Portal context would expect a high degree of personalization, the MyLibrary team sought to prioritize content that would drive users efficiently to resources relevant to their research. In keeping with the modular design of the other tabs on the MyU Portal, the MyLibrary team envisioned a layout that comprised a number of smaller cells or widgets, some of which would be universal to all users and others of which would be customized to particular affinity string groups. For example, all users would be presented with the Libraries’ search box, which initially provided a box to search MNCAT (our Aleph OPAC), our SFX e-Journals catalog, and our MetaLib federated search.
Affinity Strings for Resource Recommendations

To further personalize MyLibrary, we took advantage of affinity strings. The MyU Portal publishing system does not allow live database queries, instead requiring that we publish flat, cached data. It quickly became clear that we would not be able to publish pages completely customized for each of the 15,000 affinity strings. Instead, we created several templates: one for undergraduates; one for graduate students and faculty in the Academic Health Center; one for most other graduate students and faculty; and one for users who were assigned affinity strings and were able to access the MyU Portal, but not authorized to access library databases. Each template contained elements common to all users, and spaces for elements customized by affinity string.

Figure 3: Interface for Assigning Affinity Strings to Research QuickStart Subjects in LibData

Long before the University Libraries began work on MyLibrary, we created LibData [2], an open-source web application that provides an authoring and publishing environment for subject pathfinders and other web pages. One feature of LibData is Research QuickStart, a set of database-driven subject guides maintained by subject specialists and published in both HTML and XML formats. The MyLibrary team determined that the simplest and most sustainable method for providing relevant information in the MyU Portal would be to match a user’s affinity string to one of our approximately 250 existing subject guides. For example, by mapping the affinity strings for faculty and graduate students in Psychology to our Psychology Research QuickStart guide, we were able to present those users with a widget featuring recommended databases and indexes.
The process of matching affinity strings to templates and then to existing Research QuickStart subjects required significant work on the part of the MyLibrary team. The Libraries’ Digital Library Development Lab created an interface in LibData that made mapping all “.grad.” strings to a particular template relatively simple, but mapping every affinity string to subject content proved much more work-intensive.

The earliest affinity strings available to the MyLibrary team provided extensive information about graduate students at the University. Although we analyzed almost 800 separate affinity strings assigned to more than 38,000 people, only a portion of those represented “active” graduate students. Approximately 230 affinity strings represented the 12,000 individual graduate students then enrolled at the University. (Note that this number did not represent graduate students in the Academic Health Center and therefore excluded some graduate student groups.) The remaining strings represented students who had either completed their degree or those who had not finished a degree and were not currently enrolled. Although this latter group did not have access to licensed library resources, they did have access to the Portal and could be provided with a MyLibrary tab.

Neither affinity strings, Research QuickStart guides, nor the MyU Portal were created with the use described in this article in mind. Using these pre-existing systems to recommend resources to users has required some compromises. Our mechanism for publishing content to the MyU Portal required us to build each possible page in advance and cache them on a server rather than building them dynamically when a user visited the portal. To minimize complexity and maximize the efficiency of this process, we limited ourselves to a one-to-one relationship between Research QuickStart guides and affinity strings, presenting each user with a single set of recommendations. Unfortunately, not every degree program at the University matches cleanly with a single existing Research QuickStart subject. In some cases our subject guides were too broad to adequately reflect the specific focus of an academic department. Also, some departments span more than one subject area. For example, the Italian language program no longer exists by itself but was subsumed by the French & Italian Studies Department. Students and faculty in this department now share the same department designation in their affinity string. Because we could only map a single Research QuickStart guide to this string, the Libraries had to either create a combined subject guide, or fall back on a broader set of resource recommendations.

The MyLibrary team ultimately concluded that while affinity strings often provided a sufficient basis for determining the research needs of graduate students and faculty, they were a poor basis for providing resource recommendations to undergraduate students. Until an undergraduate student has chosen a major, his affinity string reflects only his class year status, providing no insight into his areas of academic interest. Likewise, even when an undergraduate student has chosen a major, it is safe to assume that he will continue to have coursework outside of that department. We therefore overhauled the undergraduate MyLibrary template in the winter of 2008, and rather than recommendations of resources based on affinity string, MyLibrary now offers recommendations based on each course in which an
undergraduate is currently enrolled, showing on a semester-by-semester basis the resources which are most likely to be useful.

Presenting recommended resources based on course enrollment required a new mapping strategy. Like we had done with affinity strings, the MyLibrary team mapped nearly 600 course designators, e.g., ACCT for accounting or GEOG for geography, on a one-to-one basis to existing Research QuickStart pages. As with the affinity strings, we found that some of our existing Research QuickStart subjects were insufficiently specific for good matches to course designators. For example, we had no extant Adult Education Research QuickStart page to match with the ADED course designator. In some cases we have created new Research QuickStart pages, and in others, as in the image below, we have simply mapped the designator to a more general Research QuickStart page, in this case, Education. Despite these limitations, user feedback on the resources provided has been positive.

**Figure 5: Course subjects mapped to Research Quick Start subjects**

<table>
<thead>
<tr>
<th>ID</th>
<th>Course Subject</th>
<th>Subject Description</th>
<th>CIP Code</th>
<th>Preferred RGS Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>ABUS</td>
<td>Applied Business</td>
<td>52.02.01</td>
<td>Business</td>
</tr>
<tr>
<td>3</td>
<td>ACCT</td>
<td>Accounting</td>
<td>52.03.01</td>
<td>Accounting</td>
</tr>
<tr>
<td>4</td>
<td>ADED</td>
<td>Adult Education</td>
<td>13.12.01</td>
<td>Education</td>
</tr>
<tr>
<td>5</td>
<td>ADPY</td>
<td>Adult Psychiatry</td>
<td>51.14.01</td>
<td>Psychiatry</td>
</tr>
<tr>
<td>6</td>
<td>AEM</td>
<td>Aerospace Engineer</td>
<td>14.02.01</td>
<td>Engineering, Aerospace</td>
</tr>
<tr>
<td>7</td>
<td>AFEE</td>
<td>Agri, Food, and Env Educ</td>
<td>02.01.01</td>
<td>Agriculture</td>
</tr>
<tr>
<td>8</td>
<td>APRO</td>
<td>Afro-American Studies</td>
<td>05.02.01</td>
<td>African-American Studies</td>
</tr>
<tr>
<td>9</td>
<td>AGET</td>
<td>Agricultural Engineering Tech</td>
<td>15.99.01</td>
<td>Agriculture</td>
</tr>
<tr>
<td>10</td>
<td>AGRI</td>
<td>Agriculture</td>
<td>02.01.01</td>
<td></td>
</tr>
</tbody>
</table>

**Affinity Strings for Data Gathering**

As the University Libraries became better versed in the use of enterprise affinity strings as a means for providing users with a personalized view of our services, it became clear that these strings could also be used to aggregate electronic resource usage data. While we would be hesitant to track an individual’s use of databases or e-journals, we can look for her affinity string when she accesses one of our resources and generalize on that basis.

The University of Minnesota is fortunate to have a central system for online authentication. Every student, faculty member, and staff member has a unique Internet ID which is used to access a variety of systems, ranging from email to course registration to licensed library resources. Known as the Central Authentication Hub (CAH), this service has been adopted by the vast majority of campus systems requiring authentication.

The CAH features a common interface for entering a University Internet ID and password. When a user attempts to enter a system that requires authentication, he is first forwarded to a CAH web page and asked for his Internet ID and password. Upon authenticating, a secure, encrypted cookie is placed on the user’s browser. This secure cookie is a unique, session-specific hash string containing validation credentials. The user is then sent back to the original application, which uses the secure cookie to query the University’s validation service. This service decrypts the cookie and sends back any authorization data requested by the originating application. If the user has the proper validation credentials, he is allowed into the application. If the user does not close his browser or explicitly choose to log out of the system, the original CAH cookie can be used repeatedly by University applications without requiring the user to log in again.

The University Libraries have long used the CAH to control access to licensed resources. The Libraries look for unique authorization data in the form of “library flags,” which indicate that a user meets the conditions to be granted borrowing privileges and access to our indexes and databases. The CAH can also return other user-specific data, including address, status (i.e., staff, undergrad, grad, faculty), library card number, and departmental affiliation. In addition, the
CAH can return a user’s affinity string.

Each time a user with an active CAH cookie accesses a library resource, we can capture his affinity string and associate it with the resource being requested. We thus gain valuable insight into which of our user populations are using (or not using) a given resource. By aggregating the activity of a number of users with the same affinity string, we can sufficiently anonymize our data to protect user privacy.

### Affinity String Tracking in Practice

The University Libraries have been capturing resource usage by affinity strings since the end of March 2008. To date, we have captured over one million rows of data. These data do not, however, represent 100% of the usage of our electronic resources. We are able to capture an affinity string only when a user accessing our resources has a live cookie from the CAH. Many users on campus are able to access our licensed resources directly via IP address authentication, which does not require authentication through the CAH.

That said, many applications at the University require CAH authentication and are not IP authenticated. We are able to read affinity strings on a CAH cookie regardless of whether that cookie was issued when a user sought to access a library resource. By comparing our affinity string tracking data with other resource traffic data, we estimate that we are capturing 60-70% of the usage of our resources.

In any analysis of the data we capture, we must consider this shortfall, as it may introduce significant bias. Perhaps, for example, undergraduate users are more likely than other populations to access resources from campus IP addresses and as such may be underrepresented in our data. This tracking project is in its infancy, and the University Libraries have not yet begun to analyze the data collected. However, we envision a broad range of potential applications:

#### Scenario 1: Purchasing Decisions

The University Libraries spend thousands of dollars licensing electronic material every year. Until this year we could only track usage of these resources through hit counts. With tracking based on affinity strings we could determine if the intended University audience is using a resource. Selectors could see at a glance which affinity strings are most heavily using a resource they have purchased. Perhaps more importantly, a selector could see if the intended audience is not using a costly resource, perhaps indicating an opportunity for user education, or that those purchasing dollars would be better spent on another resource.

#### Scenario 2: Resource Recommendations

The thousands of data points the University Libraries collect on resource usage by affinity string could form the basis for a recommendation system. Much as we currently present users with resource recommendations based on our Research QuickStart subject guides, we could recommend resources being heavily used by users’ colleagues. We could indicate to incoming graduate students the databases and e-journals most often used by the faculty and other
graduate students in their departments.

**Scenario 3: Resource Recommendations in other Fields**

The University of Minnesota, like many institutions, is focusing intently on interdisciplinary research. Scholars working at the boundaries of the discipline in which they have been trained can find the prospect of researching in a new field daunting. Our affinity string data could provide easy access to top resources used by researchers in that new field.

Affinity string data could also be used to provide undergraduates a quick look into the research ecosystem of a new subject. When an undergraduate is faced with the inevitable need to cite one or two scholarly sources in a paper, the Libraries could provide direct links to scholarly journals that are heavily used by graduate students and faculty at the University in a particular field.

**Scenario 4: Interdisciplinary connections**

Given a sufficiently large data set, affinity string data could be mined to reveal groups of affinity strings using the same or similar library resources. The University Libraries could then receive early indications that researchers in different departments are working toward similar goals and, ideally, sow the seeds of interdisciplinary collaboration. While it is unlikely that affinity string data at the database and journal level will uncover many connections among disciplines that aren’t well known to researchers in those fields, the data could be used to create an “interdisciplinary map” of research at the University.

**Scenario 5: Marketing and Promotion**

Affinity string data can show librarians which affinity string groups are using the libraries the most, as well as which groups are using the libraries the least. With this data we could target our services more effectively to heavy library users or direct outreach to groups we may have failed to reach.

**Scenario 6: Library Budget and Staffing Decisions**

Affinity string tracking could for the first time measure what percentage of electronic resource usage is by a given user population. This has the potential to inform the Libraries’ decisions on staffing and other budgeting. If, as our early data seem to indicate, use of our electronic resources is dominated by users from the University’s Academic Health Center (AHC), can we justify investment in tools or development that would benefit AHC users exclusively? Alternatively, can we presume that strong use of electronic resources by AHC users bespeaks comfort and self-sufficiency, and should we redirect staffing and instruction to other user populations in an attempt to increase their usage of our resources? Before we made any such drastic decisions based on affinity string data, we would need to establish great confidence in the validity and significance of the data.

**Privacy Considerations**

Monitoring library user behavior and the long-term storage of resulting data strike fear into the hearts of most librarians. We believe, however, that by capturing only a user’s affinity string and aggregating the usage data of all users with that string, we can ensure privacy.

Members of the MyLibrary team and the University Libraries’ Associate University Librarian for Information Technology met recently with a representative from the University’s Office of General Counsel to discuss the legal and privacy implications of affinity string tracking, and of the publication of affinity string data back to users. We were told that we could be justified in collecting affinity string data and in publishing back to users any data aggregated from an affinity string with at least five members.

We are inclined at this point to use a significantly higher threshold, restricting our use of affinity string usage data to those strings that have perhaps fifteen or twenty members. We intend to make our decision based on our perception of the value we can derive from those strings with small user populations.
Future Directions

The University Libraries have yet to extensively analyze the data we have collected using affinity strings. In addition to the possible uses outlined above, we expect that there are greater opportunities for deriving value from these data.

Though we’ve captured hundreds of thousands of data points, experience has taught us that we can never have too much data, especially as we look to automated recommendation services. In the future we hope to collaborate with other institutions in a position to collect similar information, so that we can standardize and pool our data.

To date we have limited our data capture to the database and e-Journal level. However, there are few, if any, technical barriers to extending our tracking to the OPAC and article level, where that data is available. Tracking usage patterns for individual works would provide far richer data and fodder for recommendations than our current system.

We’re aware that affinity strings are clumsy, if convenient, indicators of an individual’s true area of interest. Research evolves far more rapidly than does a University’s administrative structure, and being paid through a given department is not necessarily a good indicator that a researcher’s work lies completely within the traditional bounds of that department. To provide truly useful recommendations we may need to collect more detailed data.

We have discussed the idea of creating a system through which users could opt in to individualized tracking. We could then answer some of the questions we get from our users along the lines of “Can you tell me the name of that book I checked out last June?” or, “I was reading an article last week; I think the authors name was something like…” The data we could collect through such a system would increase the fidelity of our recommendation systems, allowing us to base recommendations not on institutional affiliation, but on an identity derived from interaction with the Libraries’ resources.

Conclusion

We have just begun to explore the possibilities for data gathering and data mining by tracking usage of our electronic resources by affinity strings. We expect that aspects of our efforts will prove controversial with some of our users and colleagues. However, we believe that we have unprecedented opportunities to create a host of new data-driven library practices. Affinity strings are a remarkable tool for segmenting our user base. By limiting our tracking to strings with a sufficient population, we can gather valuable data while protecting the privacy of our users.

Acknowledgements

The tools, systems, and processes described in this article are the products of the hard work of many of our colleagues on the Libraries’ Portal team. We would especially like to acknowledge Paul Bramscher, John Butler, and Kevin O’Rourke and our graduate student programmers Tariq Islam and Ankur Sharma.

Notes

[1] https://umconnect.umn.edu/myuorient/
[2] LibData is a library oriented web based application which provides authoring environments for subject pathfinders (Research QuickStart), course related pages (CourseLib) and general purpose web pages (PageScribe). For more information, see the libdata.sourceforge.net.

About the Authors

Cody Hanson is Technology Librarian in the Coordinated Educational Services Department of the University of Minnesota Libraries. He serves on the University Libraries’ Web Services Steering Committee and is a member of the MyLibrary team.
Shane Nackerud is Web Coordinator and Interim Co-Director of the Digital Library Development Lab at the University of Minnesota Libraries. He co-chairs the University Libraries’ Web Services Steering Committee and chairs the MyLibrary team.

Kristi Jensen is the Map Librarian and head of the John R. Borchert Map Library at the University of Minnesota Libraries. She is a member of the MyLibrary team.