Network Modeling Methods and Metadata Extraction for Library Access Records

Henry Williams, Yi Chen, Hui Soo Chae, Gary Natriello
EdLab, Teachers College Columbia University

Abstract

The adoption of digital library services which provide users access to resources from anywhere has enabled the collection of data about the learning behavior of library patrons. Such Big Data can yield valuable insights into how learning happens and can be used to build recommendation systems for education. By their nature, such resources are interconnected by bibliometric metadata. In this paper, we develop and test methods for building graphs of research corpora accessed by patrons through a library proxy server. We provide open-source software for building and analyzing these representations and discuss the challenges of identifying and discovering metadata from sparse proxy server logs. In addition, we discuss the potential for further research in network modeling of library access records.

Keywords: Social Network Modeling, Graph Modeling, Recommendation Systems, EZProxy, Big Data

Introduction

Today, library patrons increasingly seek and access information through digital library services. Online catalogs and databases allow users to access a broad array of library resources and tools from anywhere, providing a range of library-owned materials even for those outside of the library. Such systems provide significant opportunities for learning analytics research, because they store records of all the materials accessed by users and when they were accessed. Data mining can be useful in these cases to discover insights on how users learn.

Some studies (Duderstadt, 2009; Chen, Liu, Natriello, & Hui Soo, 2019) have argued that university libraries could be the most important vector for studying how students learn. By aggregating patrons digital trails, researchers can gain an understanding of their behaviors individually and in general. Previous studies of electronic log data in the library environment (McClure, 2003; Srivastava, Cooley, Deshpande, & Tan, 2000; Jantti, 2015; Ueno, 2004; Talavera & Gaudioso, 2004; Li, Ouyang, & Zhou, 2015; Morton-Owens & Hanson, 2012; Coombs, 2005) generally focused on the management of the resource and library usage. The network structure of the online resource, standard of data process

access, log data across different scholarly publishers, and even learning analytics in the library digital environment are still underdeveloped.

Objectives

In this paper, we consider and develop methods for extracting metadata and constructing networks of user access records from a library system, as well as the various applications and challenges of these techniques. Such networks uncover useful insights into the structural relationships between library resources and user access patterns, and they can be incorporated into a multitude of graph-based analysis techniques. We make available an open-source Python program, "*biblionet*," which can be used to create graph models and interactive visualizations (like the one in Figure 1) from library proxy server data. Further, we will discuss the potential and challenges of these methods and how they can be applied by other researchers with access to similar data sets.

Data

Data Source

As a case study, we analyzed library proxy server log data from an academic library at a graduate school of education. The system, called EZProxy, is a web proxy server used at this school and many other institutions. It provides library patrons (on and off-campus) access to library databases and e-resources automatically and continuously. Every file is saved in NCSA common log format, which contains IP address, user identifier (e.g., user id), date and time, request URL, and request status (e.g., HTTP status code and size of object returned by bytes). Substantial proxy server traffic recorded over several years provides a valuable data set for learning analytics, library science research of online resource ecosystems, and even recommendation systems. This study's data come from EZProxy daily log files from March 2018 to June 2019 (over 10 million records in total).

Data Process

We filtered the records in the following processes to identify the useful records: selecting the success requests (HTTP status code in 2XX format), selecting requests whose return object has a size bigger than 0, and classifying the URL links based on different vendors' patterns. In order to gain useful metadata for network modeling, we also focused on e-resources requested using the standardized "OpenURL" request format (Walker, 2001). We then passed the information from these requests, once trimmed and cleaned, to the CrossRef Open URL API (Ramage, Rosen, Chuang, Manning, & McFarland, 2009; Rubel & Zhang, 2015; Nurse, Baker, & Gambles, 2018) which located them in the CrossRef database and returned the DOI (if available) and all attached metadata. We then processed this metadata using an open-source Python script to build graph representations using the graph tool library (Peixoto, 2014).

Mining Metadata from OpenURL. A problem presented by library proxy servers like EZProxy is that the stored logs only include an "address" field with whatever URL the user was redirected to by the server. These URLs are obtuse and vary widely

depending on the database or library resource the user was linked to, each often using different standards and identifiers. As a result, a major hurdle to analysis of proxy server logs is finding some way of matching these URLs to the items they direct to and mining the associated metadata, none of which is recorded by the server. However, many of these links use OpenURL, a framework designed to facilitate open linking for libraries trying to direct to scholarly research (Walker, 2001). It is a standardized method of formatting requests so they can be interpreted by many different library databases and academic tools. These parameters (or a DOI) can be then passed to the CrossRef REST API, which will return all of the metadata associated with that item stored in their system (Pentz, 2001). This process is illustrated in the first portion of the flowchart in Figure 2.

Network Modeling

The acquired academic metadata is characterized by a high degree of interconnectivity, pointing towards graphs as a logical data structure for analysis. Previous work at the Network Lab of the University of Waterloo, particularly the Python library "Metaknowledge," has considered building such graph representations for bibliometric data, but their work was confined to pre-prepared files from databases like SCOPUS and Web of Science, and did not consider patron access records for libraries as this study does (McIlroy-Young & McLevey, 2015).

In building graph representations, the two primary considerations are which metadata to include as vertices in the graph and which edges to draw between them. For this study we considered papers (identified by DOI), journals (identified by ISSN), authors (identified by name or ORCiD if available), subjects (identified by ASJC code), and users (identified by username or IP address) as discrete vertices and drew nodes based on authorship, being published in a given journal, a journal being tagged with a given subject, a paper citing another, and a user accessing a paper. Vertices were then be tagged with other metadata including unique identifier, title, times cited, journal impact factor, and more, while edges were tagged based on whether an author was first or supporting, and given weights based on the number of times a user accessed a given paper.

Results

For our analysis, we developed an open-source Python program, *biblionet*, which mines metadata and builds graphs using server logs. Using the high-efficiency library graph-tool, which is built in C++, and the associated ".gt" file format, we can build graphs with tens of thousands of vertices in minutes and have built in analysis features for calculating centrality and graph topology, inferring missing edges, and many other factors. Generating meaningful, uncluttered visualizations requires taking arbitrary subsets of the total graph in order to reduce the number of vertices to a number which can be shown in a single image, as well as trimming the relatively small number of "orphan" vertices unconnected to the main graph, which can be done by isolating the largest connected component.

In Figure 3 we have drawn a graph showing the overall structure and inter-relatedness of a subset of 3000 academic papers from our EZProxy dataset, with their authors, journals, and subjects included. Immediately, several key points are clear. The largest and most central nodes are the most popular subjects, with the very largest being "education." Given

that this dataset covers papers accessed by students at a graduate school of education, this is a sensible result, and implies that these graphs can tell us something not just about the research interests of our userbase, but critically all the interconnected subjects which characterize their behavior (most of the other large nodes are for education-related subjects, e.g. psychology and linguistics). Isolating this node and examining a network with only its descendants shows that it spans nearly the entire graph, and graph centrality measures similarly identify it as most central, meaning nearly every accessed resources was in some way related to it. Several journals (e.g., Review of Educational Research and Educational Researcher) can be seen to be key in influencing this subject's central position as the large edges between them and it indicate a high betweeness centrality, meaning these are the most pivotal relationships in shaping the structure of the access records here displayed. Since this data is drawn from real user traffic, it tells us not just the most popular journals but also the subject relationships those journal share and whether a journal is popular for only one major paper or for many less-major papers.

Another graph we can generate with *biblionet* is a hierarchical block partition, which minimizes the description length of the network according to the nested (degree-corrected) stochastic blockmodel, essentially, the minimum number of groups needed to describe the hierarchical relationships of the graph (Holten, 2006). Figure 4 is one such visualization, which interestingly has five distinct groups, with the splitting of subject vertices into two separate groupings. These groupings imply for us that the overall meta-structure of the educational resources being accessed by users can be characterized by five core groups, and that the subjects accessed fall into two groups (one centered on education and the other a combination of many other topics less connected to education). These results can contribute to describing user behavior and can find the connections between resources which form the basis of recommendation systems.

As a learner with access to these graph representations, you could take the papers you have recently accessed, view just the portions of the network directly connected to those papers (those which share a subject, journal, author, or citation), and then explore outward from there. The graph structure reveals surprising new connections between information in a clear and concise way which appeals to visual learners and could speed up the research process.

Discussion

Applications & Potentials

Because of the vastness of digital information, the task for learners of identifying their learning patterns, exploring their interests, finding encouragement to persist, searching for resources, and even tracking their learning process will be overwhelmingly difficult. Without the support from techniques like recommendation systems, learners will be navigating an ocean of information without a Reccomendation System (RS; Chen, Natriello, & Hui Soo, 2019. The network of library access records provide rich and accurate information for content-based RS (Lops, de Gemmis, & Semeraro, 2011), collaborative filtering (Ricci, Rokach, & Shapira, 2011), and even link prediction in Social Network Analysis. (Jamali & Ester, 2010).

Future researchers could apply the analysis pipeline described here to any similar

library access records dataset, and could utilize *biblionet* to generate graphs with similar insights into the structure of their user's behavior. A future recommendation engine based on these techniques could use the connections between resources (in the graphs generated by *biblionet*) to rate how related they are and thus recommend resources which are above a certain relatedness threshold to resources a user has already found. Another application could be a classroom management system which charts students research activity in *biblionet* graphs, helping teachers understand the interconnection between different students' research interests for the purposes of group projects or guiding their research approach. Yet another application could allow an individual learner to visualize their own behavior in a graph and then compare it to the usage of their entire school or other students to find resources they might otherwise miss.

In addition to the techniques explored in this study, more sophisticated methods can help uncover hidden patterns in the *biblionet*-generated networks, including probability based models (e.g., Friendship-interest propagation model; Yang et al., 2011), machine learning models (e.g., graphic kernel model; Li & Chen, 2013), and latent factor models (e.g., Friend of a Friend model; Golbeck & Hendler, 2006).

Problems & Challenges

In spite of the great success of the implementation of RS and SNA with electronic log data in business, social media, and entertainment, only a small amount of attention has been paid to recommendation systems in educational contexts. Two big challenges can explain this limitation. First, Big Data methods elicit Big Data's band of problems (Jones & Salo, 2018). For example, the "trade-offs between patron privacy and access" to digital resources have proved challenging (Rubel & Zhang, 2015). Second, availability of the data across different publishers' online content still lacks consistent, secure, and standardized frameworks. Even though, CrossRef and OpenURL provide a potential solution, a large portion of the corpus (in particular the content beyond English or not machine readable) are still underdeveloped. In addition, there is a gap for the modern libraries to implement these open source techniques in reality.

Conclusion

We hope this article will encourage researchers and engineers to study and apply network modeling and EZProxy log data in education. As richer metadata, more research, and improved techniques and methodologies are provided, the scope of what EZProxy data can do and support will continue to grown. Consequently, modern libraries will have more opportunities to enhance discovery, learning, and services for the next generalization of learners.

References

Chen, Y., Liu, X., Natriello, G., & Hui Soo, C. (2019). Using Probabilistic Topic Modeling of User Search Behavior to Identify Learning Trends in Educational Research. In *Annual meeting of the american educational research association,*. Tonronto, CN.

Chen, Y., Natriello, G., & Hui Soo, C. (2019). Challenges and Opportunities of Using Recommendation System in Self-directed Learning. Unpublished manuscript.

Coombs, K. A. (2005). Lessons learned from analyzing library database usage data. *Library Hi Tech*. doi: 10.1108/07378830510636373

Duderstadt, J. J. (2009). Possible futures for the research library in the 21st century. In *Journal of library administration.* doi: 10.1080/01930820902784770

Golbeck, J., & Hendler, J. (2006). FilmTrust: Movie recommendations using trust in Web-based social networks. In *2006 3rd ieee consumer communications and networking conference, ccnc 2006.* doi: 10.1109/CCNC.2006.1593032

Holten, D. (2006). Hierarchical edge bundles: Visualization of adjacency relations in hierarchical data. *IEEE Transactions on visualization and computer graphics*, *12*(5), 741–748.

Jamali, M., & Ester, M. (2010). A matrix factorization technique with trust propagation for recommendation in social networks.. doi: 10.1145/1864708.1864736

Jantti, M. (2015). One score on âĂŞ the past, present and future of measurement at UOW library. *Library Management*. doi: 10.1108/LM-09-2014-0103

Jones, K., & Salo, D. (2018). Learning Analytics and the Academic Library: Professional Ethics Commitments at a Crossroads. *College & Research Libraries*. doi: 10.5860/crl.79.3.304

Li, X., & Chen, H. (2013). Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach. *Decision Support Systems*. doi: 10.1016/j.dss.2012.09.019

Li, X., Ouyang, J., & Zhou, X. (2015). Supervised topic models for multi-label classification. *Neurocomputing*. doi: 10.1016/j.neucom.2014.07.053

Lops, P., de Gemmis, M., & Semeraro, G. (2011). Content-based Recommender Systems: State of the Art and Trends. In *Recommender systems handbook.* doi: 10.1007/978-0-387-85820-3_3

McClure, J. (2003). Statistics, Measures and Quality Standards for Assessing Digital Reference Library Services: Guidelines and Procedures (review). *portal: Libraries and the Academy*. doi: 10.1353/pla.2003.0093

McIlroy-Young, R., & McLevey, J. (2015). metaknowledge: Open source software for social networks, bibliometrics, and sociology of knowledge research. *ON: Waterloo*.

Morton-Owens, E. G., & Hanson, K. L. (2012). Trends at a Glance: A Management Dashboard of Library Statistics. *Information Technology and Libraries*. doi: 10.6017/ital.v31i3.1919

Nurse, R., Baker, K., & Gambles, A. (2018). Library resources, student success and the distance-learning university. *Information and Learning Science*. doi: 10.1108/ILS-03-2017-0022

Peixoto, T. P. (2014). The graph-tool python library. *figshare*. Retrieved 2014-09-10, from `http://figshare.com/articles/graph_tool/1164194` doi: 10.6084/m9.figshare.1164194

Pentz, E. (2001). Crossref: a collaborative linking network. *Issues in science and technology librarianship*, *10*, F4CR5RBK.

Ramage, D., Rosen, E., Chuang, J., Manning, C. D., & McFarland, D. a. (2009). Topic Modeling for the Social Sciences. In *Advances in neural information processing systems.*

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to Recommender Systems Handbook. In *Recommender systems handbook.* doi: 10.1007/978-0-387-85820-3_1

Rubel, A., & Zhang, M. (2015). Four Facets of Privacy and Intellectual Freedom in Licensing Contracts for Electronic Journals. *College & Research Libraries*. doi: 10.5860/crl.76.4.427

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. N. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns fromWeb Data. *SIGKDD Explorations*. doi: 10.1145/846183.846188

Talavera, L., & Gaudioso, E. (2004). Mining student data to characterize similar behavior groups in unstructured collaboration spaces. In *Workshop on artificial intelligence in cscl. 16th european conference on artificial intelligence.*

Ueno, M. (2004). Online outlier detection system for learning time data in e-learning and its evaluation. In *Proceedings of the seventh iasted international conference on computers and advanced technology in education.*

Walker, J. (2001). Open linking for libraries: the openurl framework. *New Library World*, *102*(4/5), 127–134.

Yang, S. H., Long, B., Smola, A., Sadagopan, N., Zheng, Z., & Zha, H. (2011). Like like alike: joint friendship and interest propagation in social networks. *WWW*. doi: 10.1145/1963405.1963481

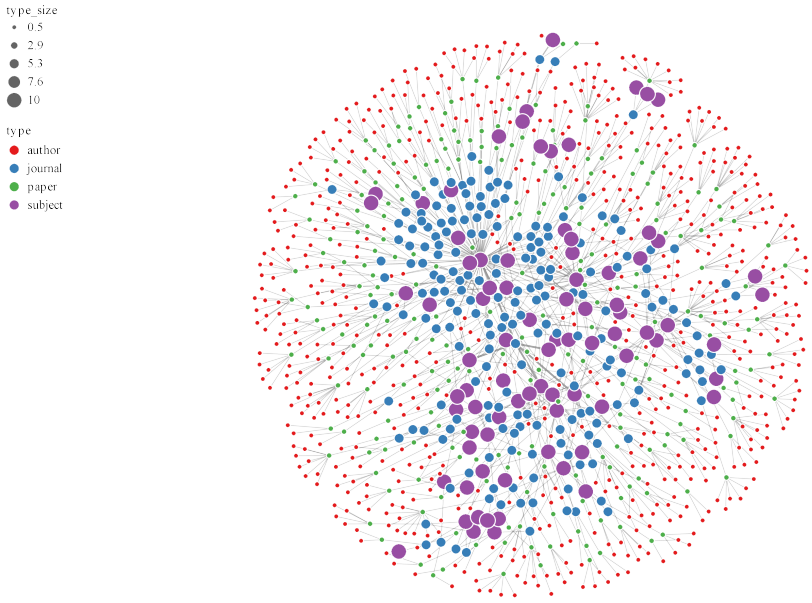

Figure 1. Graph of 300 Randomly-Selected Items from EZProxy Access Records

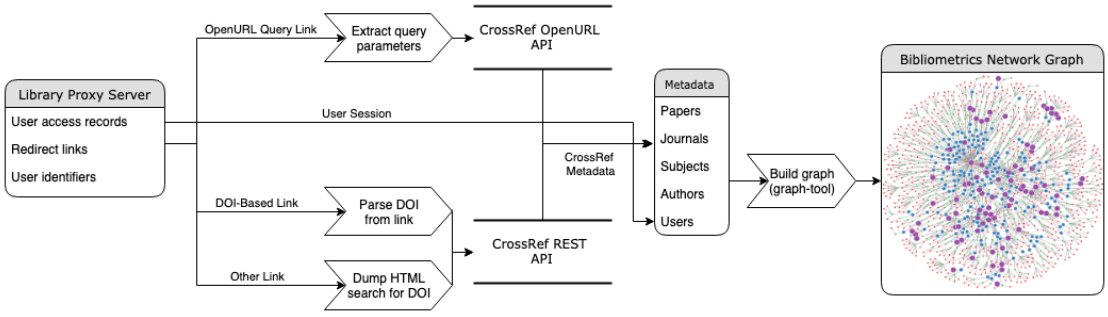

Figure 2. Flowchart of Network Modeling Pipeline

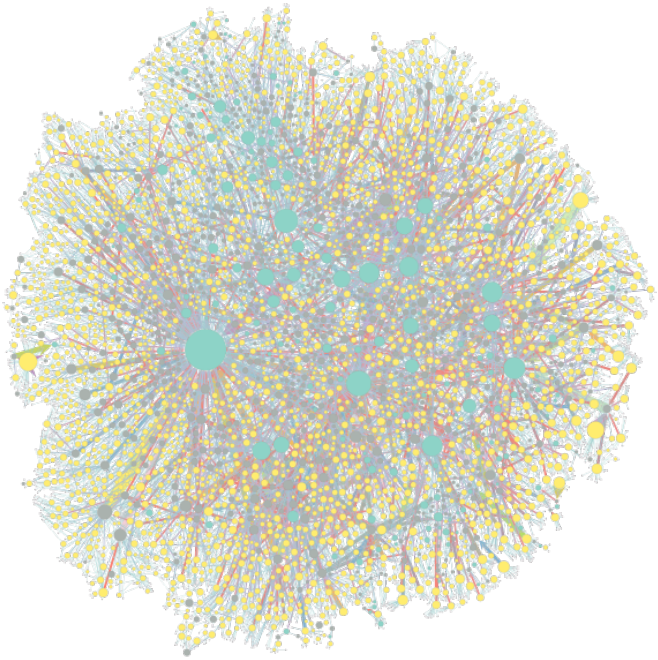

Figure 3. Directed Graph of 3000 Randomly-Selected Items From EZProxy Access Records (Nodes Scaled by In-Degree, Colored by Type and Edges Scaled and Colored by Betweeness Centrality)

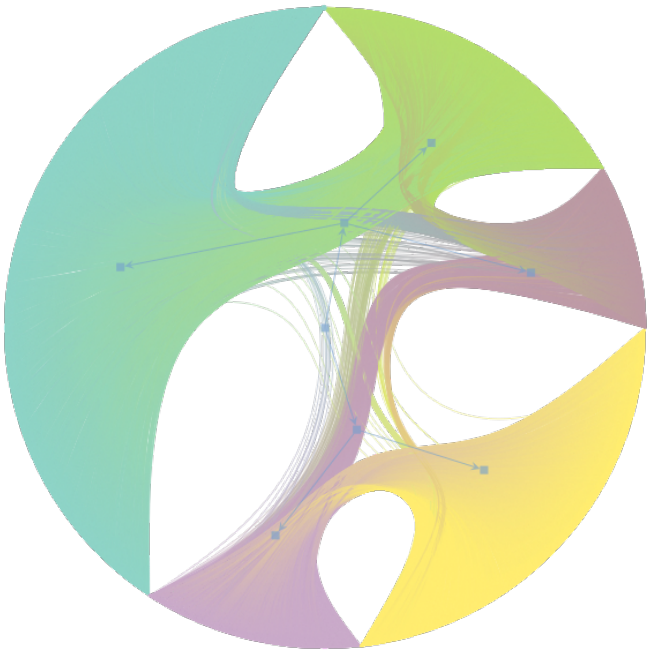

Figure 4. Hierarchical block partition of 3000 Randomly-Selected Items From EZProxy Access Records, which minimizes the description length of the network according to the nested (degree-corrected) stochastic blockmodel

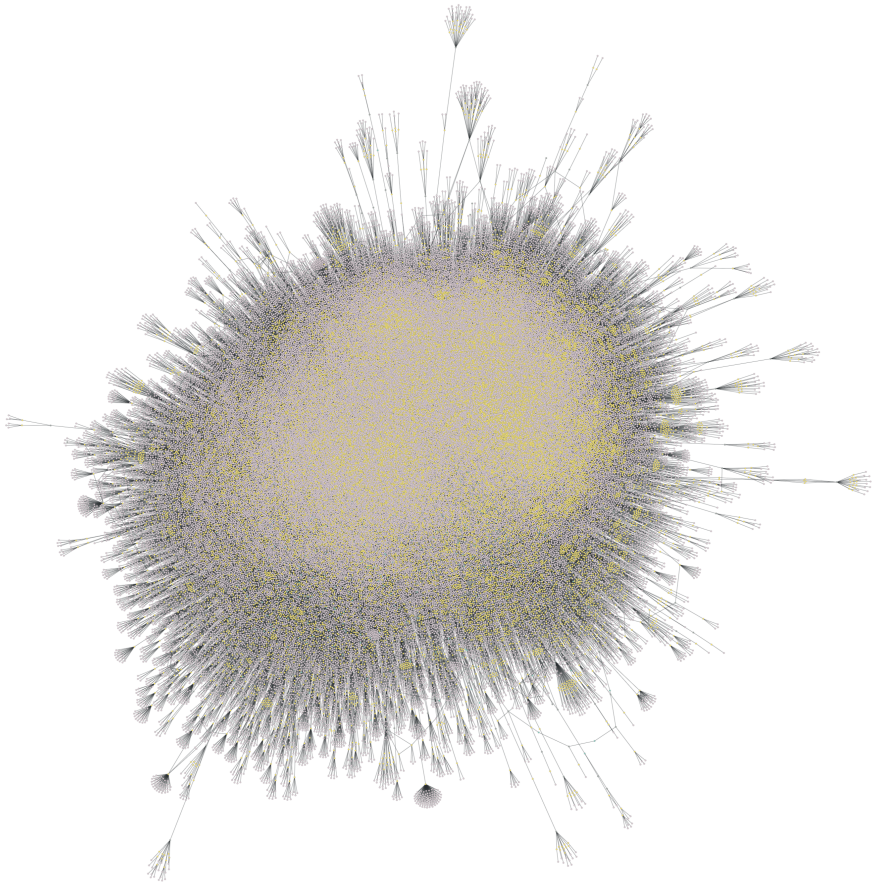

Figure 5. Largest Connected Component of All Items From EZProxy Access Records