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

Word Count: 1993

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 broad 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 this case to discover insights on how users learn.

Some studies (Duderstadt, 2009; Chen, Liu, Natriello, & Hui Soo, 2019) has 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 focus 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 insight into both the structural relationships between library resources and user access patterns, and they can be incorporated into a multitude of graph-based analysis techniques (PAPER). 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 analyze 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 ecosystem, and even recommendation system. 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

Key to processing library patron access data is understanding both what resources are being accessed and mining the associated metadata for these resources: journal, authors, subjects, publication date and more. The 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, which is 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 such that they can be interpreted by many different library databases and academic tools. Armed with either the DOI or the OpenURL parameters, our analysis pipeline incorporates the CrossRef API to match this information to the specific items they point to. CrossRef is an association of scholarly publishers that serves as a registration agency for the DOI and which hosts multiple APIs that can be used to match papers to their metadata (Pentz, 2001). Their REST API can be queried directly using a DOI, which will return all of the metadata associated with that item stored in their system. Similarly, they offer an OpenURL API which will take the parameters and search their records to match with an academic resource and its metadata if available. This process is illustrated in the first portion of the flowchart in Figure 2.

Network Modeling

With metadata gathered for the items accessed by a library proxy server, the next methodological step is to find effective techniques for analyzing this information. The high degree of inter-connectivity characterized by academic metadata points towards graphs as a logical data structure for such 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 (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, graph topology, inferring missing edges, among many others. Generating meaningful, uncluttered visualizations requires taking arbitrary subsets of the total graph in order to reduce the number of vertices to an amount 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," consistent with the fact that this data is from a graduate school of education. 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. Several journals 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.

Another kind of 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, one which interestingly has five distinct groups, which shows the splitting of subject vertices into two separate groupings. Finding such groupings allows for deeper analysis of the metastructure of these bibliometric networks.

Discussion

Applications & Potentials

The inherent hierarchical structure of academic resources implies interconnectedness: papers are published in a given journal, which discusses certain subjects; they are written by authors who usually have multiple publications; they reference each other. Given this interconnectedness, researchers have often found it useful to consider academic articles as networks, represented computationally as a graphs made up of vertices and edges. The application of graph-theoretic techniques and mathematical models has been studied for decades in the field of Social Network Analysis (SNA). Except for the techniques explored in this study, more sophisticated methods can help to uncover hidden patterns in the network, 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).

In addition, tasks of identifying their learning patterns, exploring their potential learning interests, finding encouragement to persist, searching for learning resources, and even tracking their learning process will be overwhelmingly difficult for the learner in this generation without the support from techniques like recommendation systems (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 SNA (Jamali &

Ester, 2010).

Problems & Challenges

In spite of the great success of implication 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 maybe can explain this limitation. First, Big Data methods and also elicits Big Data's band of problems (Jones & Salo, 2018). For example, the "trade-offs between patron privacy and access" to digital resources has proved challenging (Rubel & Zhang, 2015). Second, availability of the data across different publishers online content are still lack consistent, secure, and standardized frameworks. Even though, CrossRef and OpenURL provides a protential solution, a large amount of 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 implicate these open source techniques in reality.

Conclusion

We hope that the present 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

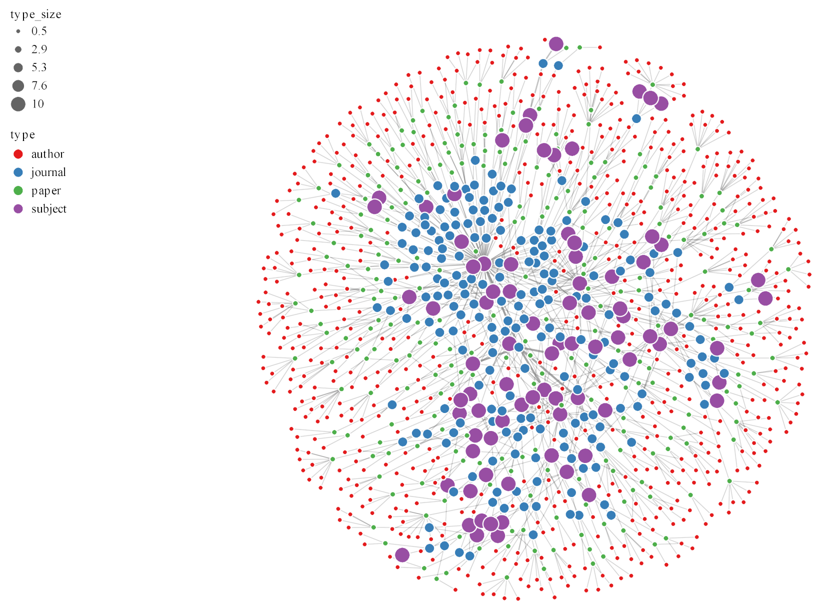

Figure 1. Graph of 300 Randomly-Selected Items from EZProxy Access Records

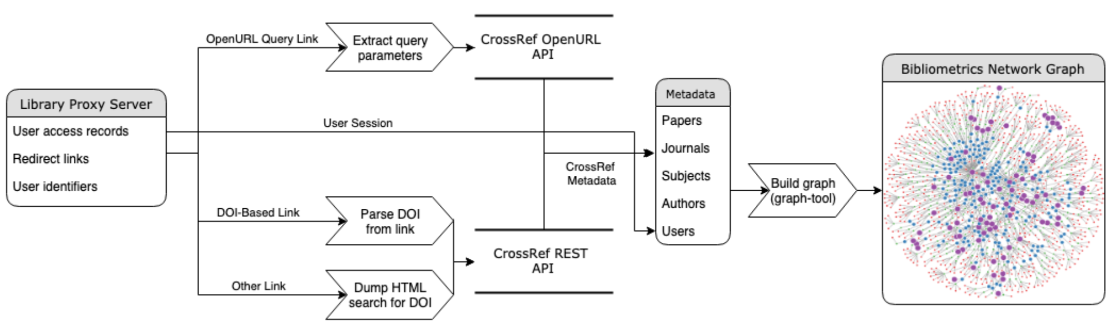

Figure 2. Flowchart of Network Modeling Pipeline

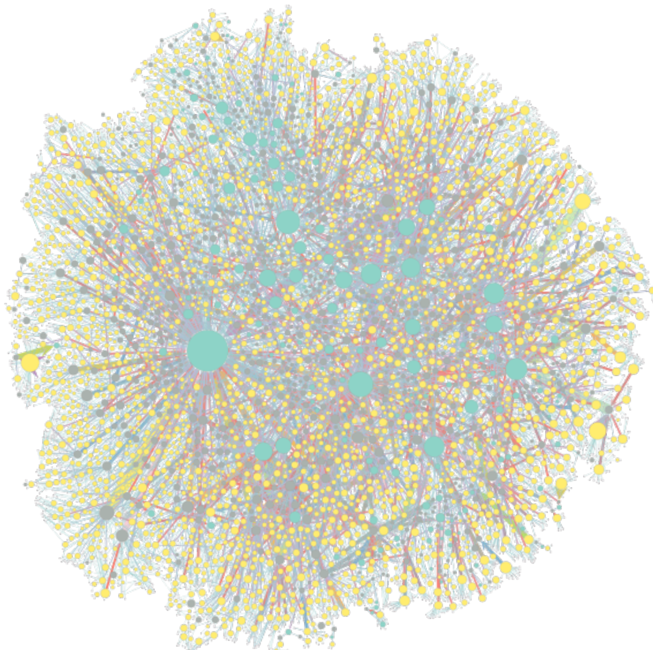

Figure 3. Directed Graph of 3000 Randomly-Selected Items From EZProxy Access Records (Nodes Scaled by In-Degree, Colored by Type and Edges Scaled and Colored by Betweeness Centrality)

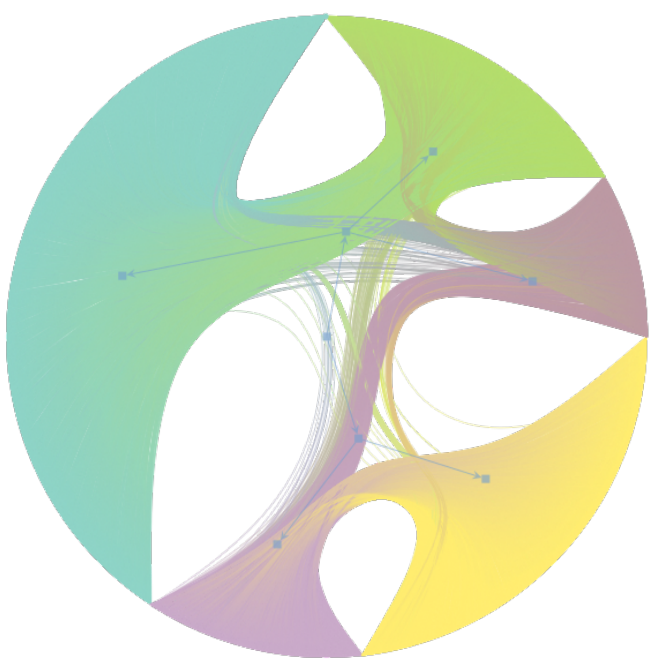

Figure 4. Hierarchical block partition of 3000 Randomly-Selected Items From EZProxy Access Records

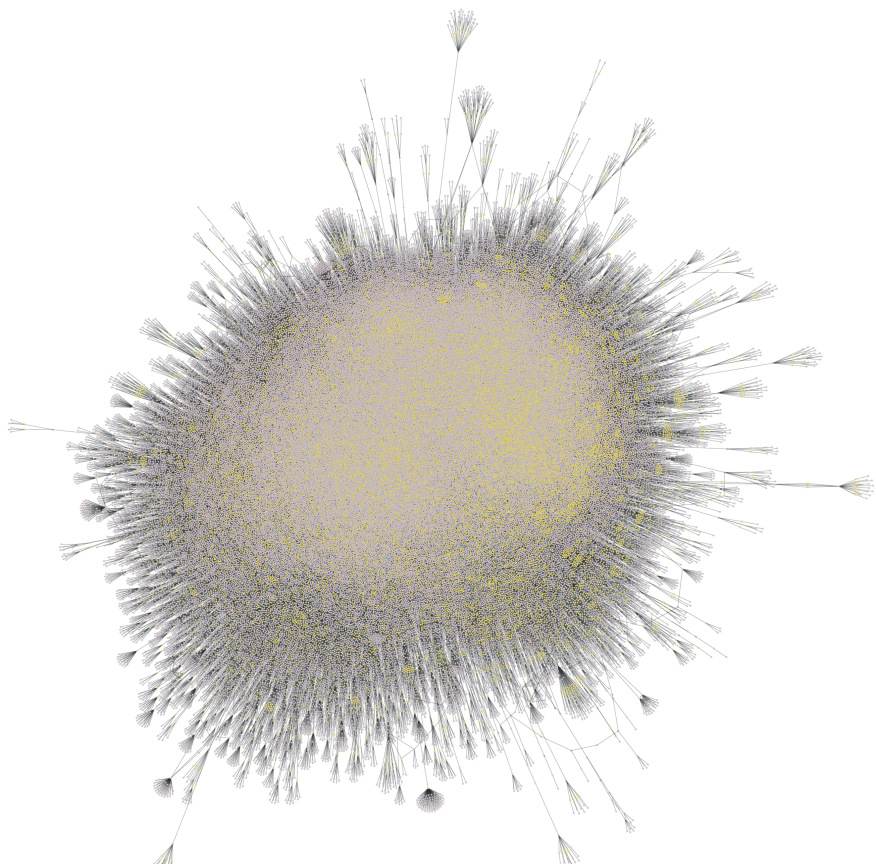

Figure 5. Largest Connected Component of All Items From EZProxy Access Records