Using Probabilistic Topic Modeling of Library Access Records to Identify Learning Trends in Educational Research

Abstract

Advances in the architecture of digital library service infrastructure enable the collection of various types of data related to the use of library resources, tools, and services. The Big Data that is being generated provides valuable insight into library operations and has the potential to reshape the future of library work. In this paper, we describe the innovative application of topic modeling (supervised Latent Dirichlet Allocation) of research corpora accessed by patrons through a library proxy server. We found that the underlying topics of this corpus (e.g., psychology, family education, and methodology) converge with the general interests one would expect from a Graduate School of Education. In addition, we discuss the potential and challenges of utilizing library proxy log data in learning analytics research.

Keywords: Probabilistic Topic Modeling, Latent Dirichlet Allocation, EZProxy, Big Data

Introduction

Digital library services have changed how patrons seek and access information traditionally found in libraries (Nicholson, 2006). The development of specialized databases and online catalogs enable patrons to make use of library resources, tools, and services remotely. In addition to being a convenience to learners, the evolution of e-learning systems provides significant opportunities for learning analytics research. For example, educational data mining is an important research area that leverages statistical techniques for discovering new insights in order to improve the performance of the education system.

Duderstadt (2009) asserts that university libraries may be the most critical observation post for studying how students learn. Academic libraries in higher education institutions have been aggregating and analyzing patrons' information and digital trails to attain a comprehensive and in-depth understanding of their learning behaviors (Zimmerman, 1995; Tan, 1999; Zaiane, Xin, & Han, 1998; Domingos, 1999; Stolfo, 1999; Jones & Salo, 2018). It is common for researchers to apply educational data mining techniques to library data. In 2003, the term "Bibliomining," a combination of bibliometrics and data mining, was coined to explain the application of pattern recognition tools to library systems data (Nicholson & Stanton, 2003).

In this study, we analyze library proxy log data from an academic library at a graduate school of education. Years of continuous and consistent proxy server traffic from patrons provide a valuable data set to identify the learning trends in educational research. EZProxy is the web proxy server used at this school and many higher education institutions around the world. This system allows library patrons on and off-campus to gain access to databases and often e-resources (e.g., e-books). At the same time, the library system automatically saves patrons' searching and learning behaviors as detailed log files.

The goal of this study is to explore the opportunities of applying Big Data techniques in the digital library environment. As an example, we analysis the extensive EZProxy data with probabilistic topic modeling technique. These techniques uncover the distribution of patrons' learning interests. Furthermore, we will discuss the potential and challenges of Big Data applications in learning analytics research.

Literature Review

Learning Analytics and Educational Data Mining. Learning analytics is defined as the measurement, collection, analysis, and reporting of data about learners and their contexts for understanding and optimizing learning and the environments in which it occurs (Siemens & Gasevic, 2012). Learning analytics studies often use data mining techniques to explore all kinds of data sources and recognize the possible patterns from a statistical perspective. Baker and Yacef (2009) summarized four goals for educational data mining: 1) predicting students' future learning behavior, 2) discovering or improving domain models, 3) studying the effects of educational support, and 4) advancing scientific knowledge about learning and learners.

In terms of self-directed learning, academic libraries are the most critical learning environments where patrons carry out their study plans with high autonomy. Unbiased, accurate, and timely information related to learning behaviors becomes accessible with the help of modern digital technologies. Therefore, many educational data mining studies have been devoted to the library settings (Morton-Owens & Hanson, 2012a; Baikady, Jessy, & Shivananda Bhat, 2014; Coombs, 2005a; Nurse, Baker, & Gambles, 2018). Generally, these studies aim at attesting to a library's value concerning students' learning outcomes. For example, Collins and Stone (2014) from the Huddersfield University Library Impact Data Project analyzed circulation data and e-resource access and sought a correlation between library activity data and degree attainment. Similarly, Nackerud, Fransen, Peterson, and Mastel (2013) from the University of Minnesota-Twin Cities examined the correlation between library usage and students' academic performances.

Log Data and EZProxy. Ubiquitous modern academic library information systems store a significant amount of learning information generated by patrons. For example, e-learning systems document students' access in weblogs and provide the records of learners' navigation on the site (Srivastava, Cooley, Deshpande, & Tan, 2000). Jantti (2015) from the University of Wollongong did pioneering work on electronic resource tracking in the Performance Indicator Framework since 1996 to monitor and drive improvement.

Several studies report on the process of applying educational data mining techniques to electronic log data. For example, McClure (2003) pointed out the statistics, measures, and quality standards for assessing digital library services. Ueno (2004) designed an outlier detection system for learning time data and its evolution. Talavera and Gaudioso (2004) mined patron data to classify similar behavior groups in unstructured collaboration space. X. Li, Ouyang, and Zhou (2015) designed a recommendation system for evolving e-learning systems.

As the concept of a virtual library becomes a reality, academic libraries continually explore new technologies of accessibility and delivery to all users (Bower & Mee, 2010). The library Proxy server becomes one of the most potent applications to provide more details about library e-resource usage. Coombs (2005b) from SUNY Cortland implemented a system to track the usage of the library's databases based on EZProxy data. Libraries can gather comparable statistics about e-resources usage. Morton-Owens and Hanson (2012b) from New York University Health Science Library adopted a management dashboard of library statistics. These applications allow decisions and trade-offs from the libraries to be more data-driven.

However, previous studies using proxy server log data focused on resource management and library usage. Researchers did not employ educational data mining and learning analytics techniques to examine patron learning behaviors.

Text Mining and Research Trend Analysis. Text mining is one of the educational data mining techniques for analyzing unstructured or semi-structured text data (Fan, Wallace, Rich, & Zhang, 2006). For example, Blei and Lafferty (2006b) developed dynamic topic models and examined the OCR archives of the Journal *Science* from 1880 through 2000.

Using text mining for identifying research trends has already been discussed in fields beyond education. In management science, Delen and Crossland (2008) proposed the application of text mining for identifying research trends based on three management information system journals. In business, Moro, Cortez, and Rita (2015) used the topic modeling method to analyze business intelligence in banking based on 219 articles published between 2002 and 2013. In biology, L. L. Li, Ding, Feng, Wang, and Ho (2009) analyzed the trends in global stem cell research from 1991 to 2006 (Barde & Bainwad, 2018).

Generally, these studies picked several representative journals under one specific field, collected their published corpora, and discovered the evolution of topic distributions over the publication time. This approach entails two risks: the lag effect and authoritarianism. On the one hand, the publication date is not a reliable metric for identifying learning trends, because it can take years to publish academic research. By the time an article is published, researchers may have already shifted to other research topics. On the other hand, editors and professional researchers cannot fully represent the learning interests of a much broader learner population. In this study, we focus on the corpora that have been reached by library patrons .

Topic Modeling

Topic modeling is a statistical approach to identify the probabilistic latent semantic structure in a collection of text documents. A general introduction of topic modeling is beyond the scope of this article, but Blei (2011) provided the introduction of topic modeling from a statistics perspective, and Liu, Tang, Dong, Yao, and Zhou (2016) provided a review of applications of topic modeling techniques.

The advantage of using topic modeling lies in automating latent topics detection across large scale corpora of documents. Topic modeling assumes that each document is a mixture of a small number of topics, and each word's presence is attribution to one of the document's topics. Therefore, each topic is a distribution of words in the corpora. The topic is a recurring pattern of co-occurring words.

Latent Dirichlet Allocation. LDA is the most widely used topic modeling technique. Words collected into documents are observations, and each document is supposed to be a mixture of a small number of latent topics.

LDA models cluster the observed words into different groups (topics) with a specific probability (i.e., "likelihood" belongs to topic "statistics" with 0.9 probability and topic education with 0.05 probability). Summing over the words' topic distributions within each document, we can get the topic distribution of the document. For example, an educational technology article may have a topic distribution of 50% education topic, 20% computer science topic, and 30% statistics topic. Similarly, summing over the documents' topic distributions, we can identify the topic distribution of patrons' learning interests as a whole.

The basic model assumptions of LDA model are:

$$\begin{aligned}\boldsymbol{\theta}_d|\boldsymbol{\alpha} &\sim Dirichlet(\boldsymbol{\alpha})\\ \boldsymbol{\phi_Z}|\boldsymbol{\beta} &\sim Dirichlet(\boldsymbol{\beta})\\ z_{di}|\boldsymbol{\theta}_d &\sim Multinomial(\boldsymbol{\theta}_d)\\ w_{di}|\boldsymbol{\phi_{z_{di}}} &\sim Multinomial(\boldsymbol{\phi_{z_{di}}})\end{aligned}$$

For each document d, we assume the distributions of the topics within that document is following a multinational distribution with parameter $\boldsymbol{\theta}_d$ (topic distribution for document). $\boldsymbol{\theta}_d$ is randomly sampled from a Dirichlet Distribution with the hyper-parameter $\boldsymbol{\alpha}$. The ith word in the d th document w_{di} (observed) is assumed to be generated by sampling from a topic-specific multinational distribution $\boldsymbol{\phi_{z_{di}}}$, where z_{di} the indicator of the topic that w_{di} belongs to. $\boldsymbol{\phi_{z_{di}}}$ is the topic distribution for word i in document d.

Data

This study's data comes from EZProxy daily log files from September 2015 to August 2018 (over 35 million records in total). Every file is saved in NCSA common log format:

127.0.0.1 user-identifier frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326

Each line in the file has the same syntax. It consists of seven parts: patrons' IP address (127.0.0.1), user-identifier (RFC 1413 identity), frank (userid), date, time and time zone that the request was received (10/Oct/2000:13:55:36 -0700), request URL line (GET /apache_pb.gif HTTP/1.0), HTTP status code (200) ,and the size of the object returned to the patron in bytes (2326). This record shows what e-resource a patron was reached at when and where.

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. We focused on the PDF format e-resources in this study from a representative vendor, downloaded PDFs, and converted them into text format. In addition, we concluded the following text data processing operations: word segmentation, punctuation removal, deleting of numbers, transforming the words to lowercase, updating the stop words, removing the stop words and stemming. After building the corpora, we transformed the data into a document-term matrix and removed the sparse term. Finally, we derived 19773×1402 document-term matrix.

Results

In applications of LDA and its extensions, the number of topics is a modeling choice which we need to specify *a priori*. Including more topics leads to a better fit at the expense of increasing model complexity. A more complex model has a larger number of parameters and requires more computational resources. In addition, the added complexity often leads to difficulty in interpreting the results. A common practice for choosing the number of topics is thorough cross-validation (Browne, 2000; Kohavi, 1995). The goal is to find the optimal number of topics that maximizes the model fit while penalizing the complexity. Different measures have been proposed for this purpose. One particularly popular choice is perplexity (Horgan, 1995). It has an inverse relationship with the likelihood. In other words, the model with lower perplexity fits better. In our analysis, we cross-validated three model fit measurements implemented in the "ldatuning" *R* package that are alternatives to perplexity (Murzintcev, 2016). Specifically, we considered CanJun (Cao, Xia, Li, Zhang, & Tang, 2009), Arun (Arun, Suresh, Madhavan, & Murty, 2010), and Deveaud (Deveaud, SanJuan, & Bellot, 2014). For both CanJun and Deveaud, a smaller statistic indicates a better fit. On the other hand, Arun prefers models with more extensive statistics.

Figure 1 shows the results of the cross-validation. We chose ten topics in our analysis. Figure 2 visualizes part of the posterior distribution of the words within each topic in LDA. For example, the representative root words in the first topic are psycholog(2.33%), behavior(1.87%), and sutdi(1.21%). Based on this word distribution, we can loosely summarize the true meaning of this topic as "psychology." Similarly, the representative root words in the fifth topic are: women(2.73%), health(2.50%), american(1.82%), gender(1.71%), black(1.53%), and white(1.10%). We summarized the true meaning of this topic as "Gender and Race." Finally, we interpret the 10 topics as "psychology", "management", "family", "methodology", "gender and race", "language", "school", "social", "experiment and research", and "class". Table 1 gives the posterior probabilities of representative words and papers within each topic. The representative papers give more supportive evidence about the topic meanings.

The proportion of every topic is shown in both Table 1 and Figure 3. This result is the marginal topic distribution for the patron community as a whole. The most popular topic is "experiment" and "research", "methodology", and "management", while "language", "family" and "school" are less popular. Since the patrons consist of students, faculty, and stuff in this graduate school of education, the topic distribution of their learning behaviors can widely represent their learning interests. The results from this topic distribution converge with the overall structure of institutes and departments in this school. For example, students and faculties from counseling & clinical psychology and human development departments may be interested in topics about "experiment" and "methodology."

Even though this is a fuzzy and inaccurate guess for the true topic meanings, these topics coincide with the educational studies that this institution has been focusing. With

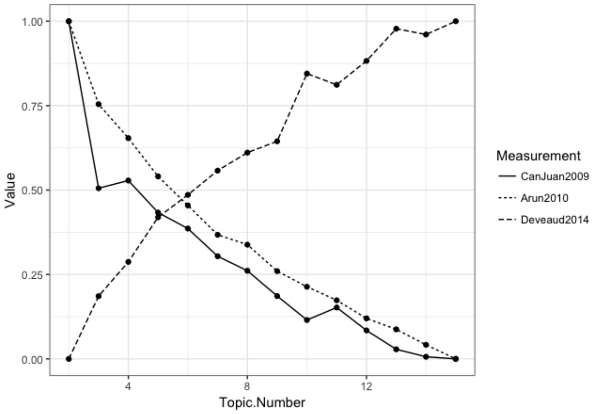

Figure 1. Cross Validation

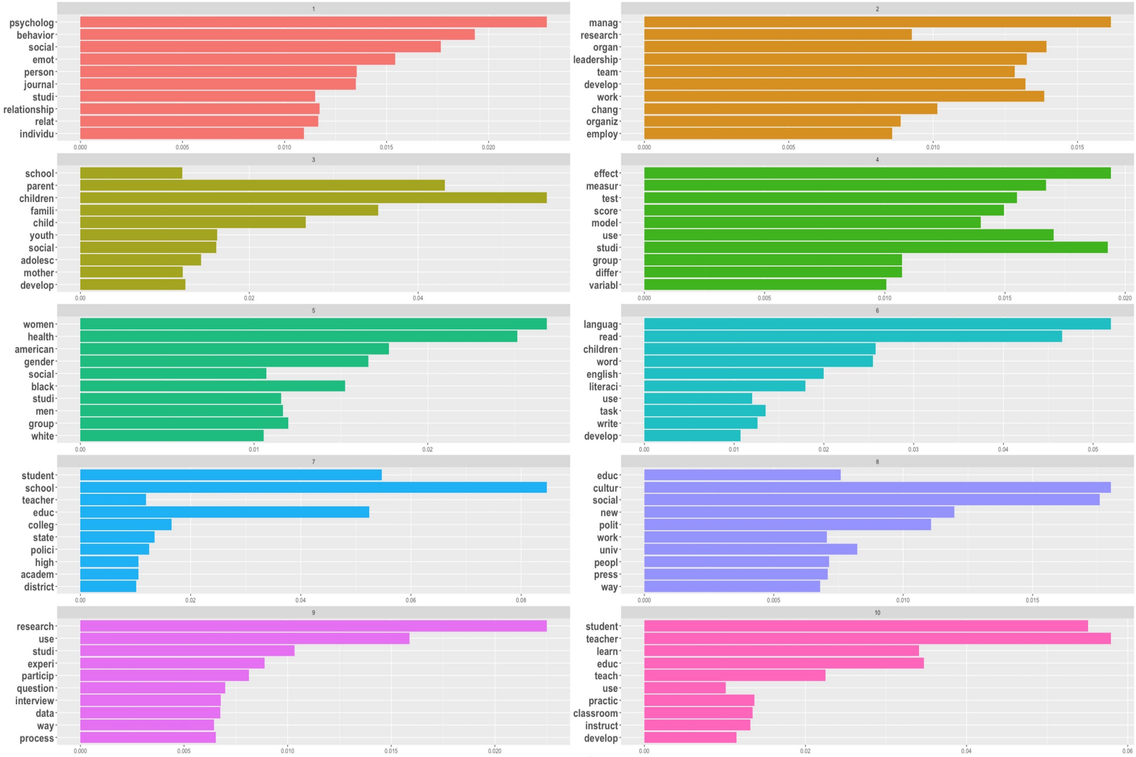

Figure 2. Posterior Word Distribution in Topics

more patron learning behaviors being collected in the EZProxy system and advances in text mining techniques, we can launch more precise, timely, and reliable learning analytics studies in the future.

Table 1

Major Topics for LDA

Topic	Representative Words	Proportions	Representative Paper

1. Psychology	psycholog(2.33%)
	behavior(1.89%)
	social(1.73%)
	emot(1.56%)
		7.88%	The What, How, Why, and Where of Self-Contractual(0.166%)
			Online Dating: A Critical Analysis From the Perspective of Psychological Science(0.152%)

2. Management	manag(1.67%)
	work (1.43%)
	develop(1.43%)
	organ(1.42%)
	leadership(1.42%)
		10.94%	Performance Adaptation: A Theoretical Integration and Review(0.118%);
			Team Effectiveness 1997-2007: A Review of Recent Advancements and a Glimpse Into the Future(0.141%)

3. Family Education	children(5.23%)
	parent(4.36%)
	famili(3.50%)
	child(2.71%)
		6.71%	Best Practice Guidelines on Prevention Practice, Research, Training, and Social Advocacy for Psychologists(0.114%);
			Evidence-based guidelines for the pharmacological treatment of schizophrenia: recommendations from the British Association for Psychopharmacology(0.105%)

4. Methodology	studi(1.92%)
	effect(1.89%)
	use(1.73%)
	measur(1.71%)
		16.09%	A Review and Synthesis of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research(0.077%);
			Comprehensive School Reform and Achievement: A Meta-Analysis(0.072%)

5. Gender and Race	women(2.72%)
	health(2.50%)
	american(1.82%)
	gender(1.71%)
		8.26%	Summarizing 25 Years of Research on Men's Gender Role Conflict Using the Gender Role Conflict Scale: New Research Paradigms and Clinical Implications(0.171%) ;
			Women in Academic Science: A Changing Landscape(0.154%)

6. Language	language(5.28%)
	read(4.73%)
	children(2.51%)
	word(2.51%)
		6.28%	Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology(0.250%)
			Teaching Reading Comprehension Strategies to Students With Learning Disabilities: A Review of Research(0.161%)

7. School	school(8.38%)
	student(6.49%)
	educ(5.32%)
	colle(1.73%)
		7.68%	Comprehensive School Reform and Achievement: A Meta-Analysis(0.148%)
			Chapter 3: Opportunities, Achievement, and Choice: Women and Minority Students in Science and Mathematics(0.115%)

8. Social	cultur(1.67%)
	social(1.50%)
	new(1.41%)
	polit(1.20%)
		9.90%	Race and Police Brutality: Roots of an Urban Dilemma(0.161%);
			The Changing Landscape of Work and Family in the American Middle Class: Reports from the Field(0.161%)

9. Experiment and Research	research(2.5%)
	use(1.52%)
	studi(1.03%)
	experi(1.09%)
		17.82%	On-line Dating: A Critical Analysis From the Perspective of Psychological Science(0.068%);
			Shared-Reality Development in Childhood(0.059%)

10. Class Education	teacher(4.83%)
	student(5.62%)
	educ(3.53%)
	learn(3.50%)
		8.43%	Mentored Learning to Teach According to Standards-Based Reform: A Critical Review(0.190%)
			Challenges New Science Teachers Face(0.133%)

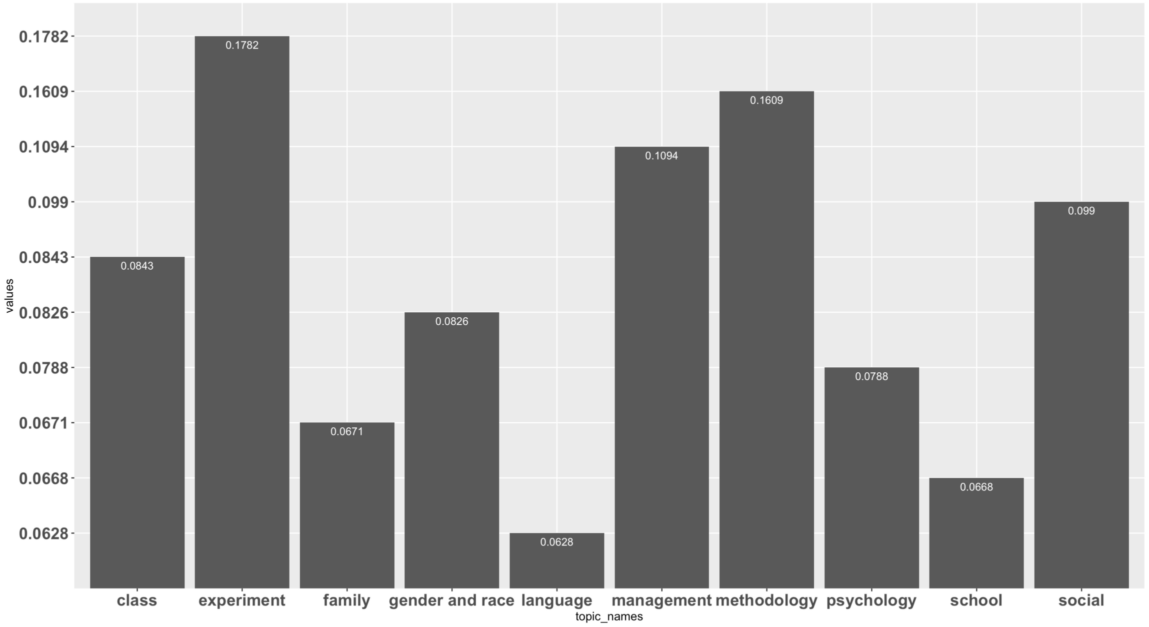

Figure 3. Topic distribution for LDA

Discussion

Problems and Challenges. Educational data mining employs Big Data methods and also elicits Big Data's band of problems (Jones & Salo, 2018). Especially in the practice of learning analytics in academic libraries, privacy, and confidentiality are still significant concerns. Showers and Stone (2014) argue that academic libraries have been struggling with privacy problems associated with emerging data and information flows.

In this study, we did not track any personal indicator or share any personal information. In addition, analyses are all on the patron community level instead of the personal level. However, there is no apparent technical difficulty in tracking more patron personal information and doing personalized data analyses. For example, the new version of EZProxy system in this graduate school library now provides the methods for tracking the patrons' Id when they are off-campus.

Rubel and Zhang (2015) note, resolving the "trade-offs between patron privacy and access" to digital resources has proved challenging. In order to cast light on the mystery of patrons' learning process, we require more personalized information about the learner. However, seeking access to patrons' data at the same time degrades patrons' privacy. Like the EZProxy, people can track a complete sequence of actions during a single visit as well as all related information, including location, time, and e-resources. It enables actors with the right privileges to keep a detailed audit of these activities and, consequently, judge students' behaviors. Moreover, many patrons are not sensitive to this issue when they are using the library service.

Another challenge is the availability of the data across different vendor-managed systems. As more and more modern libraries start using the new techniques (e.g., EZProxy), researchers need a secure, consistent, and accurate method for identifying all the resources patrons have reached. In our study, we focused on a representative vendor's database and PDF format e-resources to simplify data collection. Using the same approach for all the records in the EZProxy logs and collecting the article, information from different vendors could be extremely time-consuming and result in inconsistent outcomes. Rubel and Zhang (2015)'s investigation into 42 unique licensing agreements uncovered the broad spectrum of data collection and sharing protections in existence. As Coombs (2005b) emphasized, similar problems exist in the application of EZProxy log data. Each vendor uses its own set of usage statistics. While libraries can have multiple accounts with the same vendor, they have no direct access to the vendor usage statistics. It creates unnecessary barriers for researchers and librarians until all databases across different vendors can be connected under a consistent, secure, and standardized system.

In addition, EZProxy log data sometimes are limited. We cannot know whether the patrons are actively learning. We do not know the contribution that each article makes to their research, and how the reading habits for each individual could make a difference. We may over-count the number of times that an article has been read since we cannot distinguish several visits from different patrons with the same IP address, or an unintentional double click from the same person. The links from the same article may occur several times during a single visit for search, preview, share, and download.

Applications and potentials. In this study, we use topic modeling techniques with a critical yet largely untapped resource for learning analytics research: EZProxy log data. There is enormous potential for using digital data from EZProxy in educational research. In-depth and more personalized research about patrons' learning behavior is possible using the new version of EZProxy. A personal online library website based on every patron's learning profile is not out of reach. In addition, as the volume of patrons' learning behaviors saved in EZProxy increase, more timely and reliable personal data could be used for a more sophisticated search and recommendation system. Libraries can come to know their patrons better and encourage more patrons to use the services, resources, and environments they provide. For example, (Wang & Blei, 2011) analyzed library patrons' data for user classification and recommendation.

On the other hand, other topic modeling techniques also have great potential to solve learning analytics problems with EZProxy data. Algorithmic improvements in text mining provide the ability to fit more complicated models and handle massive data like EZProxy logs. For example, many alternative methodologies are available to explore different research interests in topic modeling. The Correlated Topic Model provides a "map" that tells how the topics related as well as a better fit for text data (Blei & Lafferty, 2006a). The continuous-time dynamic topic model and the Dynamic Topic Model analyze how the topics drift in a time sequence as the time-corrected similarity between articles (Blei & Lafferty, 2006b; Wang, Blei, & Heckerman, 2008). Labeled LDA incorporates credit attribution in multi-labeled corpora into the LDA framework (Ramage, Hall, Nallapati, & Manning, 2009). The Document Inference Model measures the scholarly impact with a sequence of texts and provides a retrospective estimate of articles that influence (Chang & Blei, 2012).

In addition to LDA, other methods also used in topic modeling for uncovering hidden structures of large scale corpora. The Vector space model is a representative solution for keywords search and has been involved in large part of information retrieval research (Salton, Wong, & Yang, 1975). Latent Semantic Indexing is used for identifying relevant documents from search words (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). Barde and Bainwad (2018) discussed these methods with their features and limitations. In addition, tools for processing extensive collections of the document using topic modeling are being developed, including MALLET, Gensim, BigARTM, and Standford topic modeling toolbox (McCallum, 2002; Khosrovian, Pfahl, & Garousi, 2008; Vorontsov, Frei, Apishev, Romov, & Dudarenko, 2015; Topic & Toolbox, 2012).

We hope that the present article will encourage researchers and librarians to use EZProxy data and topic modeling in their studies. The result will be a more in-depth and more informative analysis of patrons' learning behaviors.

References

Arun, R., Suresh, V., Madhavan, C. E., & Murty, M. N. (2010). On finding the natural number of topics with Latent Dirichlet Allocation: Some observations. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics).* doi: 10.1007/978-3-642-13657-3_43

Baikady, M. R., Jessy, A., & Shivananda Bhat, K. (2014). Off campus access to licensed e-resources of library: A case study. *DESIDOC Journal of Library and Information Technology.* doi: 10.14429/djlit.34.6.7509

Baker, R. S. J. D., & Yacef, K. (2009). The State of Educational Data Mining in 2009 : A Review and Future Visions. *Journal of Educational Data Mining.* doi: http://doi.ieeecomputersociety.org/10.1109/ASE.2003.1240314

Barde, B. V., & Bainwad, A. M. (2018). An overview of topic modeling methods and tools. In *Proceedings of the 2017 international conference on intelligent computing and control systems, iciccs 2017.* doi: 10.1109/ICCONS.2017.8250563

Blei, D. M. (2011). Introduction to Probabilistic Topic Modeling. *Communications of the ACM.* doi: 10.1145/2133806.2133826

Blei, D. M., & Lafferty, J. D. (2006a). Correlated Topic Models. *Advances in Neural Information Processing Systems 18.* doi: 10.1145/1143844.1143859

Blei, D. M., & Lafferty, J. D. (2006b). Dynamic topic models. In *Proceedings of the 23rd international conference on machine learning - icml '06.* doi: 10.1145/1143844.1143859

Bower, S. L., & Mee, S. A. (2010). Virtual delivery of electronic resources and services to off-campus users: A multifaceted approach. *Journal of Library Administration.* doi: 10.1080/01930826.2010.488593

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology.* doi: 10.1006/jmps.1999.1279

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing.* doi: 10.1016/j.neucom.2008.06.011

Chang, J., & Blei, D. M. (2012). Hierarchical relational models for document networks. *Annals of Applied Statistics.* doi: 10.1214/09-AOAS309

Collins, E., & Stone, G. (2014). Understanding patterns of library use among undergraduate students from different disciplines. *Evidence Based Library and Information Practice.* doi: 10.18438/B8930K

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

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

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science.* doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Delen, D., & Crossland, M. D. (2008). Seeding the survey and analysis of research literature with text mining. *Expert Systems with Applications*. doi: 10.1016/j.eswa.2007.01.035

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*. doi: 10.3166/dn.17.1.61-84

Domingos, P. (1999). The role of Occam's Razor in knowledge discovery. *Data Mining and Knowledge Discovery*. doi: 10.1023/A:1009868929893

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

Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. *Communications of the ACM*. doi: 10.1145/1151030.1151032

Horgan, J. (1995). From Complexity to Perplexity. *Scientific American*. doi: 10.1038/scientificamerican0695-104

Jantti, M. (2015). One score on – 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

Khosrovian, K., Pfahl, D., & Garousi, V. (2008). GENSIM 2.0: A customizable process simulation model for software process evaluation. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. doi: 10.1007/978-3-540-79588-9_26

Kohavi, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Appears in the international joint conference on articial intelligence (ijcai)*. doi: 10.1067/mod.2000.109031

Li, L. L., Ding, G., Feng, N., Wang, M. H., & Ho, Y. S. (2009). Global stem cell research trend: Bibliometric analysis as a tool for mapping of trends from 1991 to 2006. *Scientometrics*. doi: 10.1007/s11192-008-1939-5

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

Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). *An overview of topic modeling and its current applications in bioinformatics.* doi: 10.1186/s40064-016-3252-8

McCallum, A. (2002). MALLET: A Machine Learning for Language Toolkit. *http://mallet.cs.umass.edu.*

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

Moro, S., Cortez, P., & Rita, P. (2015). Business intelligence in banking: A literature analysis from 2002 to 2013 using text mining and latent Dirichlet allocation. *Expert Systems with Applications*. doi: 10.1016/j.eswa.2014.09.024

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

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

Murzintcev, N. (2016). *R: Package 'ldatuning'.*

Nackerud, S., Fransen, J., Peterson, K., & Mastel, K. (2013). Analyzing Demographics: Assessing Library Use Across the Institution. *portal: Libraries and the Academy*. doi: 10.1353/pla.2013.0017

Nicholson, S. (2006). The basis for bibliomining: Frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services. *Information Processing and Management*. doi: 10.1016/j.ipm.2005.05.008

Nicholson, S., & Stanton, J. M. (2003). Gaining strategic advantage through bibliomining: Data mining for management decisions in corporate, special, digital, and traditional libraries. *Or-*

ganizational data mining: Leveraging enterprise data resources for optimal performance.

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

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing.* doi: 10.3115/1699510.1699543

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

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM.* doi: 10.1145/361219.361220

Showers, B., & Stone, G. (2014). Safety in numbers: Developing a shared analytics service for academic libraries. *Performance Measurement and Metrics.* doi: 10.1108/PMM-03-2014-0008

Siemens, G., & Gasevic, D. (2012). Guest editorial - learning and knowledge analytics. *Educational Technology and Society.* doi: 10.1207/s15327752jpa8502

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

Stolfo, S. J. (1999). KDD cup 1999 dataset. *UCI KDD repository. http://kdd.ics.uci.edu.*

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.*

Tan, A.-H. (1999). Text Mining: The state of the art and the challenges. In *Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases.* doi: 10.1.1.38.7672

Topic, S., & Toolbox, M. (2012). The Stanford Natural Language Processing Group. *Learning.*

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.*

Vorontsov, K., Frei, O., Apishev, M., Romov, P., & Dudarenko, M. (2015). Bigartm: Open source library for regularized multimodal topic modeling of large collections. In *Communications in computer and information science.* doi: 10.1007/978-3-319-26123-2_36

Wang, C., Blei, D., & Heckerman, D. (2008). Continuous Time Dynamic Topic Models. *Proc of UAI.*

Wang, C., & Blei, D. M. (2011). Collaborative topic modeling for recommending scientific articles.. doi: 10.1145/2020408.2020480

Zaiane, O. R., Xin, M. X. M., & Han, J. H. J. (1998). Discovering Web access patterns and trends by applying OLAP and data mining technology on Web logs. *Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries -ADL'98.* doi: 10.1109/ADL.1998.670376

Zimmerman, G. (1995). Library of Congress Cataloging in Publication Program. *Publishing Research Quarterly.* doi: 10.1007/BF02680460