Journal of Information Technology Management

ISSN #1042-1319

A Publication of the Association of Management

IMPROVING THE ANALYSIS AND RETRIEVAL OF DIGITAL COLLECTIONS: A TOPIC-BASED VISUALIZATION MODEL

HSUANWEI MICHELLE CHEN
SAN JOSÉ STATE UNIVERSITY
hsuanwei.chen@sjsu.edu

HONGBO ZOU
QUEENSLAND UNIVERSITY OF TECHNOLOGY
hongbo.zou@hdr.qut.edu.au

ALYCE L. SCOTT
SAN JOSÉ STATE UNIVERSITY
alyce.scott@sjsu.edu

ABSTRACT

Today, the ongoing and dramatic increase in available information has motivated information professionals to seek better ways to organize and present large data sets for improved information access and retrieval. One approach to helping users find and retrieve large-scale information is information visualization. While the literature has addressed how information visualization techniques can help present data in a way that enables users to find, retrieve, and analyze large-scale information effectively, little has been done to visualize the information based on its underlying semantic meanings. In this paper, we developed a topic-based visualization model to present and analyze the Illinois Digital Archives, in which a document can now be represented as a mixture of semantic topics and the relationships between documents can be understood more clearly. The study extends the existing collection management options for librarians, archivists, and other information professionals who manage and curate large digital collections.

Keywords: digital collections, collection management, information visualization, topic modeling, visual analytics, data curation, information retrieval

INTRODUCTION

With ongoing advances in storage and archive technologies and the growing interest in sharing data, there has been a tremendous increase in the availability and presence of digital collections. Watkins et al. [29] defined digital collections as collections of "large quantities of accumulated digital 'stuff' of varying form, purpose and value". Bass et al. [4, p. 1] also argued that digital collections are "any set of documents or multimedia pieces (e.g., images, audio, files, videos, etc.) gathered and presented online for the purpose of exchanging resources and ideas". The varieties of formats, contents, and purposes of digital collections and the growing volume of digital collections poses new challenges for users who seek more information and new opportunities for information professionals who curate that information.

The organization and management of digital collections have changed greatly in the last few decades. For

years, information professionals and organizations committed to providing access to digital content have developed various methods to access, collect, generate, distribute, and preserve information to end-users. A National Research Foundation report [23] suggested that these methods are mostly based either on creating data in digital form – born digital – or transforming data into digital or digitized form. The report further highlighted that collection managers manage digital collection based on three principles: (a) digital preservation to ensure that the digital content is maintained over time in usable formats and can be made available to current and future users in meaningful ways, (b) digital stewardship that includes taking a set of actions that extend the longevity and usefulness of digital content, and (c) digital curation that includes a set of activities such as data management, archiving, and digital preservation. Digital curation is concerned with lifecycle management of a digital resource from the time it is created until it is purposely disposed of. It also involves creating, correcting, and enhancing metadata or collection descriptions so that they are appropriately described and documented and can be used and reused so that the digital content is broadly available over time to end-users without unnecessary impediments.

For years, several different algorithms based on these principles have been proposed to manage and retrieve useful information from large databases of digital collections [2]. Advances in technology have not only brought significant changes in the way librarians and information professionals create, acquire, organize, store, and distribute digital collections but has also brought changes to information retrieval systems. However, even though collection managers have made great efforts to integrate user perspectives and develop design processes that adequately accommodate end user perspectives, they have focused solely on assessing system designs [7] and user/usability interface. In fact, little attention has been paid to analyzing the impact digital collection projects have had on actual research efforts [27]. Also, no consensus exists on the best methods of managing and organizing digital collections, which may address immediate information needs of end users [12]. Horava [15] further identified major issues concerning digital collections management. These include maintaining core values such as equity of access of digital collections, scholarly communication issues such as retaining author rights and copyright issues, acquisition activities such as acquiring large amounts of on-site or offline data, budget allocation, pricing models, and licensing options, access, and delivery issues such as making on-site resources available online and disruptive innovation marked with how technology and user behavior are entwined, and learning information-seeking behavior. Apart from these, developments in information and communication technologies (ICTs) have presented other challenges for curators and digital collection managers. These include creating a networked infrastructure, purchasing hardware and software, copyright concerns and work flow and quality assurance, and other technological challenges that hinder collection, preservation, and distribution of digital content to and from the digital world [11]. As amounts of born-digital and digitized materials continue to grow, these challenges will hinder the management of these materials. As a result, digital collection managers must keep current with technical trends to develop what Horava [15] called a sustainable and forward-looking approach to collection management. One way through which digital collection managers can address these challenges is by using information technologies. Through various information technologies, collection managers can not only create and preserve a large amount of digital archives but also support, transmit, and share digital information in a way that can be easily searched and retrieved by end-users.

Over the years, significant investments have been made in developing and utilizing technology systems to manage and sustain born-digital and digitized collections. These include open-source software products that use metadata schemata and controlled vocabularies for subject access across distributed digital repositories and collections [15]. However, as Deal [10] noted, these systems are heavily text-based and provide little information about the scope and contents of the digital collection. According to the author, this hampers the users' ability to easily browse and explore the contents. However, information visualization or the development of visual analytics can solve the problems that users experience when accessing and navigating through rapidly growing digital data. Leung et al. [19, p. 120] defined visual analytics as "the science of analytical reasoning supported by interactive visual interfaces". The authors claimed a need to develop interactive systems that visualize data-mining results. According to the authors, visual representations of large datasets can make it easier for users to view and analyze mining results when compared to numerous algorithms that return results in textual forms. Deal [10] stated that information visualization can enhance search in digital collections and provide information about the scope as well as the context of a collection. This in turn can allow users to more easily browse and explore contents. Cybulski et al. [9] stated that visual analytics differ from other forms of digital creativity. According to the authors, information visualization may provide opportunities for users to engage with data analytics to explore a large amount of data resources. Apart from providing the desired output of archival analysis for users, visual analytics can help curators summarize large quantities of diverse

information about the collection, analyze large-scale digital collections, and even facilitate collection management decision-making to provide easier access [30]. As seen, information visualization can meet the unique needs of users of distinctive digital collections, but, as Lemieux [18] pointed out, much attention is needed to apply information visualization and visual analytics to unprocessed archival materials to make them more effective for researchers and end-users.

In this study, we developed visualizations of digital collections based on a machine learning technique called topic modeling, which looks for patterns in the use of words and is an attempt to inject semantic meaning into vocabulary; a "topic" consists of a cluster of words that frequently occur together [1]. Through topic-based visualization, a document can now be represented as a mixture of topics, rather than just a compilation of disconnected, explicit information (i.e., author, title, and subject). For example, a document on cloning can be represented as a mixture of topics that may include "cells," "immune," and "human." In this study, topic-based visualizations were developed and tested with data from the Illinois Digital Archives. The model offers new options to information professionals for curating large-scale data sets and also expands the ways in which users can retrieve and understand documents from large digital collections. The new model allows users to view digital documents at a semantic level through topic modeling, while at the same time being able to visualize the relationships between those documents quite clearly.

The rest of the paper is organized as follows. We synthesized and discussed related work in the literature review. The Illinois Digital Archives is introduced in the data section. We describe the topic modeling and visualization methods in the methodology section, with the final visualizations presented in the results section. The paper ends with the conclusions and recommendations for future work.

LITERATURE REVIEW

The analysis, management, and organization of digital collections, or digital archives, have been approached from many different perspectives. Considering the increasing development of digital collections, a National Research Foundation report [23] brought together South African practitioners in the field of digitization to find the best management techniques for digital collections. The study emphasizes various issues concerning digital collections, such as requirements for building good digital collections and sound digital collection management practices. Several prior studies have contributed to this practice. For example, Brown [5] explored methods and resources that can be used in the development of a personal digital archiving workshop and how librarians can use these methods and resources to provide custom-made services to their audience. In a similar vein, Cartolano et al. [7] showed that open-source technologies can be used to implement this and other digital library projects, help reduce the amount of locally written code, and develop new sustainable approaches to meet the needs of new and different digital collections. Digital archivists and curators, on the other hand, have interesting perspectives on how to manage digital collections. Gengenbach [12] presented views from archivists and curators applying digital forensic tools and practices to the management of born-digital content. The author suggested that digital forensic tools are beneficial in capturing and preserving born-digital content. Lee et al. [17] also argued that considering the current needs of the library and archives community, it is imperative to develop digital forensic tools with interfaces, documentation, and functionality that can support the work flows of collecting institutions.

The advances in digital technologies have, however, presented major challenges concerning collection management, such as acquisition activities, delivery issues, innovation, access, and scholarly communication, in a rapidly changing environment [15]. Park and Tosaka [25] highlighted the existing trends in metadata-creation practices in digital repositories, collections, and libraries. By using information obtained from the community of cataloging and metadata professionals, the authors also put forward challenges that exist in creating descriptive metadata elements, using controlled vocabularies for subject access, and propagating metadata and metadata guidelines beyond local environments. In addition, numerous prior studies have tackled the challenges of digital collection management from an information retrieval and knowledge management perspective [21]. For example, Angsachin et al. [2] applied the method of artificial neural networks with multilayer perception to machine learning in pattern recognition and produced better retrieval results. The authors found that the application of this method improved the pattern of recognizing word meanings and could be a potential tool to use in developing a novel system of web information retrieval. Hsu et al. [16] also mapped the trend and intellectual structure of digital archives research by deploying text-mining techniques, such as co-word and cluster analysis, and provided a valuable tool for researchers to access digital archives literature. As technology is changing the way people locate, access, and use information and the readiness of libraries to serve the community, information visualization has been playing a major role in mediating formal humanities research and the study and management of collections in recent years

[3]. Lemieux [18] argued that immediate attention is required to define the ways in which information visualization and visual analytics can help mediate digital archives. Leung et al. [19] also contended that visual representations can enhance user understanding of the inherent relationships among frequently mined data sets and stated that many of the existing visualizers lack the design necessary to visualize these frequently mined sets.

Three existing approaches to information visualization are particularly relevant to our study. The first approach, graph-based visualization, represents information as diagrams of abstract networks, which can help users digest complex information [14]. However, graph-based visualization does not necessarily take into account the strengths of the relationships among the data. A second approach, social visualization, involves visualizing data on the interactions occurring between the data. Also, social visualization has been studied in the context of email patterns [11] and newsgroup activities [24]. Social visualization can be used to capture important relationships between documents already held in digital collections; however, no social visualization model for digital collections currently exists. The third approach, text visualization, involves visualizing the text of a collection of documents to show the relationships between terms and text patterns [22]. Text visualization can also provide visual summaries of content and relationships within a larger collection, thus allowing users to navigate these summaries as they visualize their relationships [27]. However, these existing models visualize texts mostly from a syntax level. To fill the research gaps, we developed a topic-based visualization model for digital collections. With this new approach, a document can now be represented as a mixture of topics, rather than just a compilation of disconnected, explicit information. When using the topic model to query a digital collection at the semantic level of analysis, users can visualize relationships between many documents that contain terms with similar meanings or documents from different disciplines that also cover the same topics.

DATA

In this study, we developed and tested our approach with data from the Illinois Digital Archives (http://www.idaillinois.org/). Established in 2000, the Illinois Digital Archives is a repository for the digital collections of the Illinois State Library as well as other libraries and cultural institutions located in Illinois. It includes manuscripts, newspapers, government documents, images, and other historical material. Figure 1 shows a screenshot of the home page of the website. As seen in Figure 1, the interface provides a button ("Browse All") to navigate through all collections, and a "Search" function to access and retrieve information based on user-defined fields in different combinations of title, subject, description, creator, date, and format.

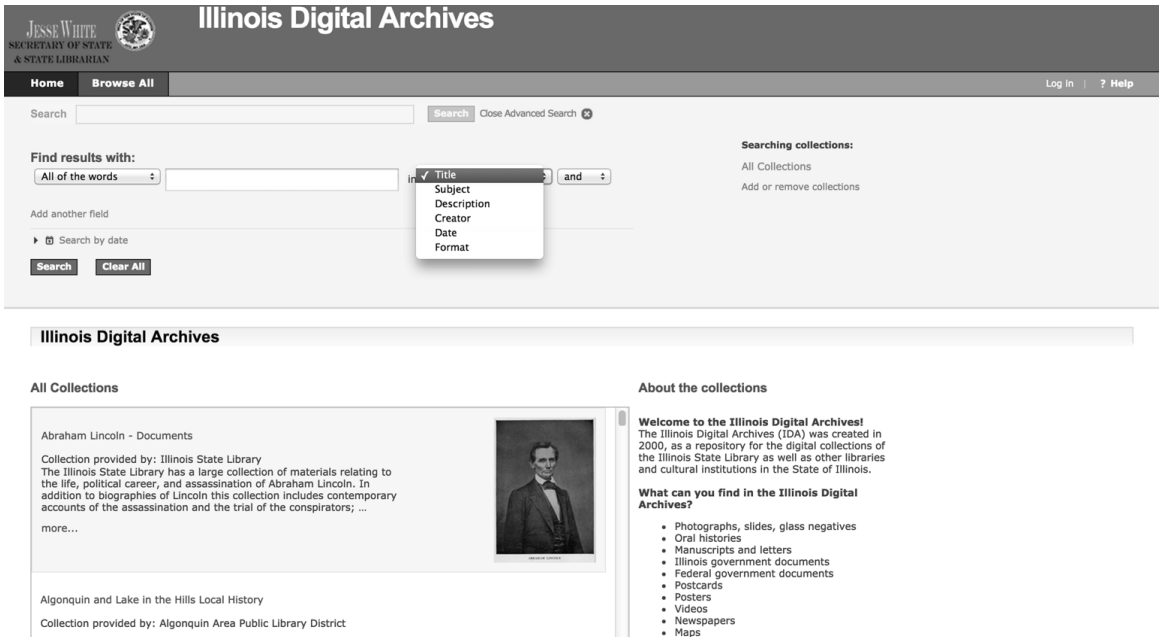

Figure 1: Illinois Digital Archives Webpage (© 2015 Illinois Digital Archives)

While the archives provide abundant digital collections, the search function is limited. The mere combination of the "explicit" data of documents such as title and subject makes it easy for users to clearly specify the search criteria, but significantly lacks the capability to provide the semantic relationships between documents. This search design and capability is commonly seen in digital collections and archives, which can make information access and retrieval inefficient, ineffective, and, moreover, inaccurate. To improve the user experience in information retrieval in large digital collections, our study focuses on providing a visualization model to be added to the web search, access, and retrieval process to better match the semantic meanings of documents with the search needs and interests. To approach the problem, our first step was to preprocess the metadata of these digital collections and extract them into a predefined, well-structured XML format. Figure 2 shows an XML example of the metadata for the "Charles Overstreet" collection, which is the photography collection of Charles Overstreet, a long-time citizen of Flora, Illinois, with a passion for photography. The collection contains 369 still images. For each piece of work, the corresponding XML file contains information such as title, creator, subject, and a description array containing publisher, contributor, date, type, format, identifier, source, language, relation, coverage, and rights. Topic-based visualization is then performed on the XML metadata to extract, identify, and present the hidden relationships between semantic topics of the documents. The methodology is detailed in the next section.

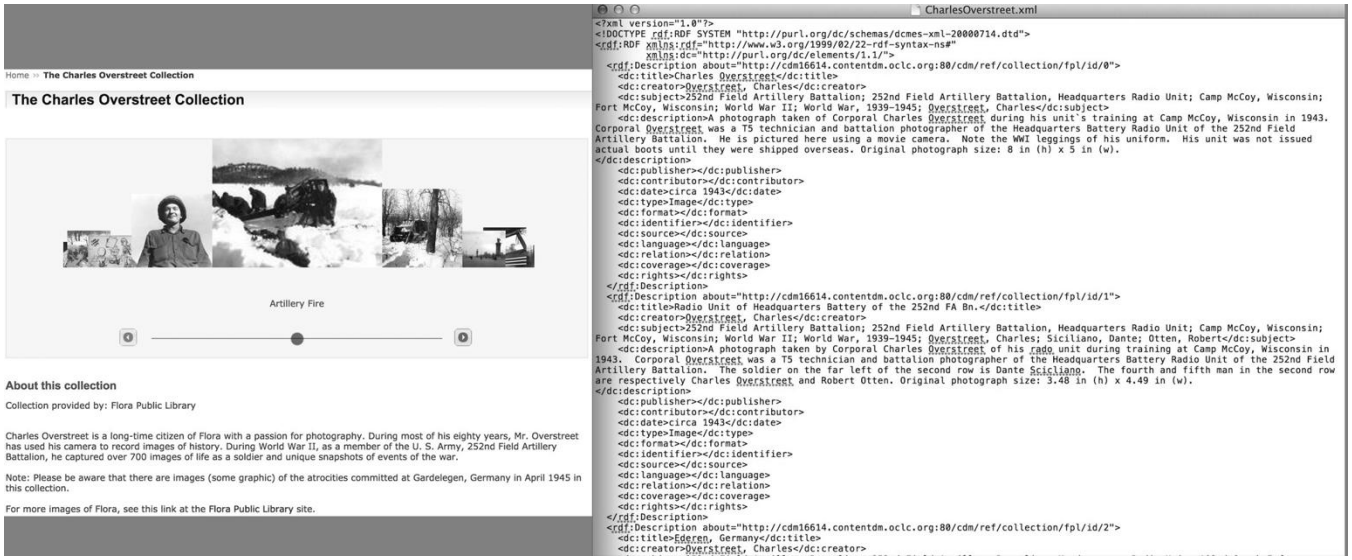

Figure 2: Screenshots of the "Charles Overstreet" Collection and XML Metadata

METHODOLOGY

In this study, we propose a visualization model to present the semantic relationships between digital collections using a machine learning method called topic modeling. Topic modeling looks for patterns in the use of words and attempts to inject semantic meaning into vocabulary. The result is a "topic" that consists of a cluster of words that frequently occur together [1]. Through topic modeling, large volumes of unlabeled text can be analyzed using contextual clues [28]. Topic modeling connects words with similar meanings and distinguishes between uses of words with multiple meanings. In this study, we adopt a machine learning toolkit, MALLET [20], to perform topic modeling. MALLET provides an efficient way to build up topic models based on Latent Dirichlet Allocation (LDA) [1], a generative probabilistic model that allows sets of observations to be understood by unobserved groups that explain why some parts of the data are similar. In this study, MALLET takes each set of XML metadata and learns patterns in the use of words by assuming that any XML metadata is composed of selected words from possible baskets of words where every basket corresponds to a specific topic.

To apply topic modeling to discovering the underlying thematic structure in the Illinois Digital Archives, we set the number of topics to be 4, 8, or 16 to allow for different options of topic granularity for users.

Note that in traditional topic modeling, choosing too few topics will produce results that are overly broad, while choosing too many will result in "over-clustering" the collected data into many small, highly similar topics [13]. However, in the context of searching and retrieving digital collections, the availability of a wider range and number of topics provides users with the interactive capability of specifying how many topics are to be presented at the same time (i.e., how "fine" the relationships between documents the users wants to explore). For space considerations, Table 1 shows the topic modeling results of the XML metadata of the "Charles Overstreet" collection, only with the number of topics equal to 4 and 8. In Table 1, for each topic cluster, there is a cluster probability, which represents the likelihood that this cluster occurs (i.e., the probability that the words in the cluster co-occur). The key words in each topic cluster that were learned and identified are also presented. We then create interactive visualizations based on the topic modeling results using *R* language. The visualization aims to provide graphical presentations of three key components for the digital collections: (1) the probability of each topic cluster, (2) the words learned and identified in each topic cluster, and (3) the relationships among the topic clusters, the probabilities, and the words. The visualization results are presented in the next section.

# of Topics = 4			
Cluster = 0 Prob.= .02132	Cluster = 1 Prob.= .01973	Cluster = 2 Prob.= .07652	Cluster = 3 Prob.= .09323
war overstreet charles quot postcards drawing ii world illinois original home back flora army charlie postcard life belt pictures	quot troops benning unit war fort georgia overstreet bodies prisoners radio artillery ii world official gardelegen point top candidates	unit germany overstreet corporal photograph war battalion field radio headquarters artillery size radioman original photographer france part berlin world	unit photograph corporal overstreet battalion artillery size headquarters original radio field charles photographer germany battery mccoy camp wisconsin photo

# of Topics = 8							
Cluster = 0 Prob.= .03315	Cluster = 1 Prob.= .06967	Cluster = 2 Prob.= .02565	Cluster = 3 Prob.= .42029	Cluster = 4 Prob.= .06756	Cluster = 5 Prob.= .06702	Cluster = 6 Prob.= .0103	Cluster = 7 Prob.= .04161
quot benning unit fort georgia artillery troops war school made member point top candidates officer west procedures correct demonstate	germany war world ii berlin radioman german capture part overstreet bid corporal pictures opportunities travelled architecture charles lanscape castle	war postcards drawing back france illinois charles overstreet home flora world charlie ii postcard life belt pictures original humorous	unit overstreet corporal photograph radio artillery headquarters battalion original size photographer field charles photo time pictured member picture larry	mccoy wisconsin camp battery training photograph technician battalion shows corporal overstreet fire march learning charles firing range photo gun	germany radioman town ederen spent winter heidelberg stopped marching photos castle city netherlands shelled heavily fellow worked throught passed	bodies prisoners troops gardelegen war overstreet survivors shot straw hide barn soldiers german escape htm www mr ii world	germany france part ninth beach army radioman omaha spent october rivers belgium units foot traveled england served bn taking

Table 1: Topic Modeling Results (# of Topics = 4 and 8) for "Charles Overstreet Collection"

RESULTS

In this section, we present topic-based visualizations for the Illinois Digital Archives to explore the hidden semantic relationships between documents through the learned, thematic topics, as described in the previous section. For demonstration purposes, we will continue using the "Charles Overstreet" collection as an example. Figures 3, 4, and 5 show the document visualizations for 4, 8, and 16 topics, respectively. The x-axis represents the cluster probability of the words in each cluster co-occurring, and the y-axis represents the number of the topic cluster (e.g., for 4 topics, there will be topic cluster 0, 1, 2, and 3). The bubbles are placed on the x-y plan based on the probability distribution. In addition, the size of the bubbles is also proportional to the probability, which gives users an immediate understanding of the relative relationships of the cluster probabilities. In each bub-

ble, the words within each cluster are presented, showing the hidden semantic topics of the selected document. The visualizations can be added to regular search tasks (i.e., search results given in texts) to provide more semantic information about the documents and their relationships.

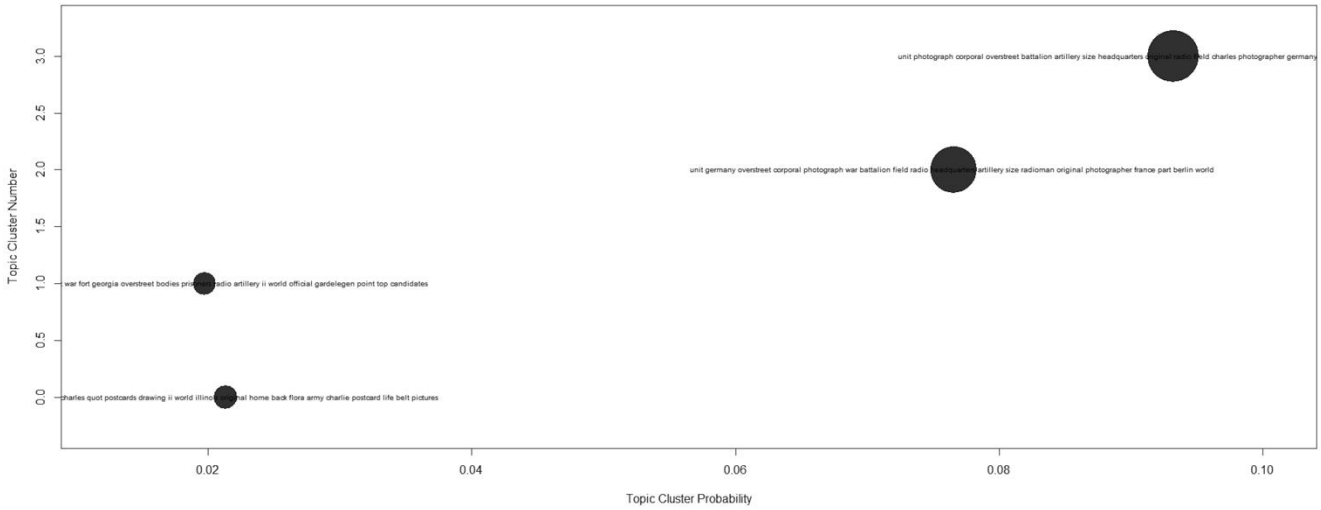

Figure 3: Topic-Based Visualization for "Charles Overstreet Collection": Four Topic Clusters

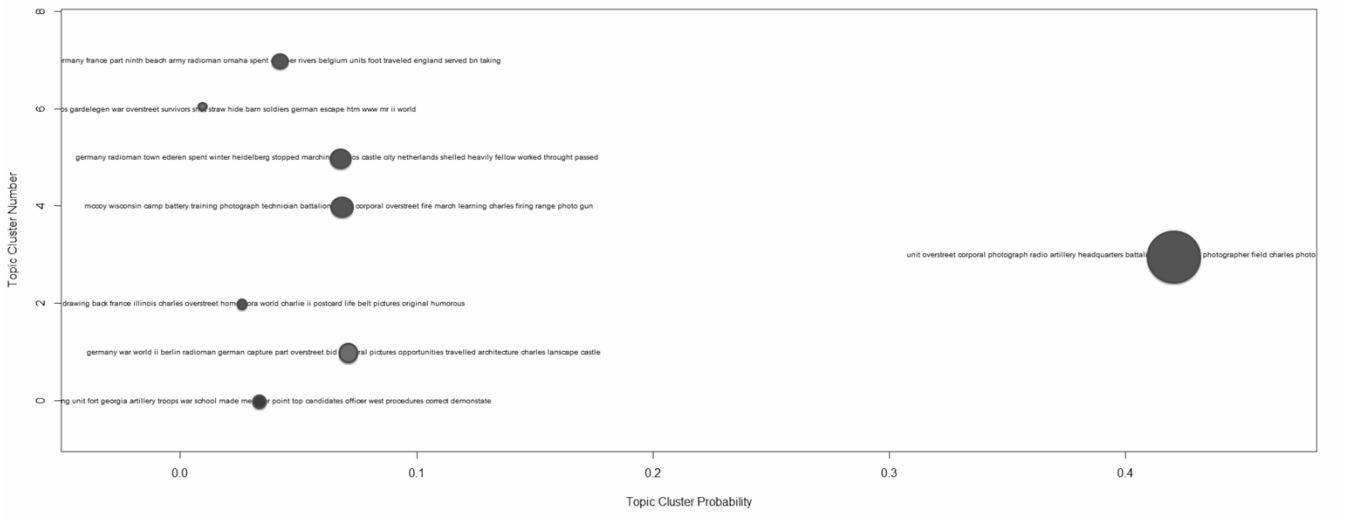

Figure 4: Topic-Based Visualization for "Charles Overstreet Collection": Eight Topic Clusters

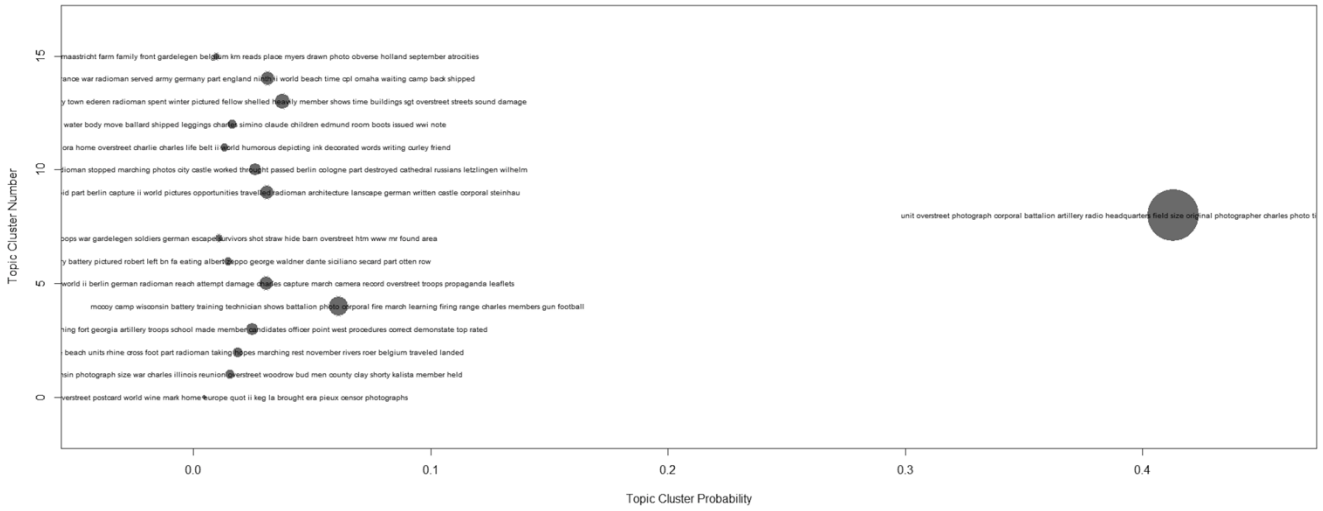

Figure 5: Topic-Based Visualization for "Charles Overstreet Collection": Sixteen Topic Clusters

When hovering over each bubble, the complete set of words within that topic cluster will show up. This interactive feature provides users with a "focus+context" capability, enabling the viewers to see words of the topic of primary interest presented in full detail while at the same time obtaining an impression of all the surrounding information available [6]. Figure 6 shows the visualization result. The design of the topic-based visualization also allows users to select among 4, 8, or 16 topics to view the different granularities of the document and topic relationships, as shown in Figure 7. This feature enables users to also decide how much deeper, detailed, and fine-tuned they want exploration of the hidden thematic topics of the selected document to be.

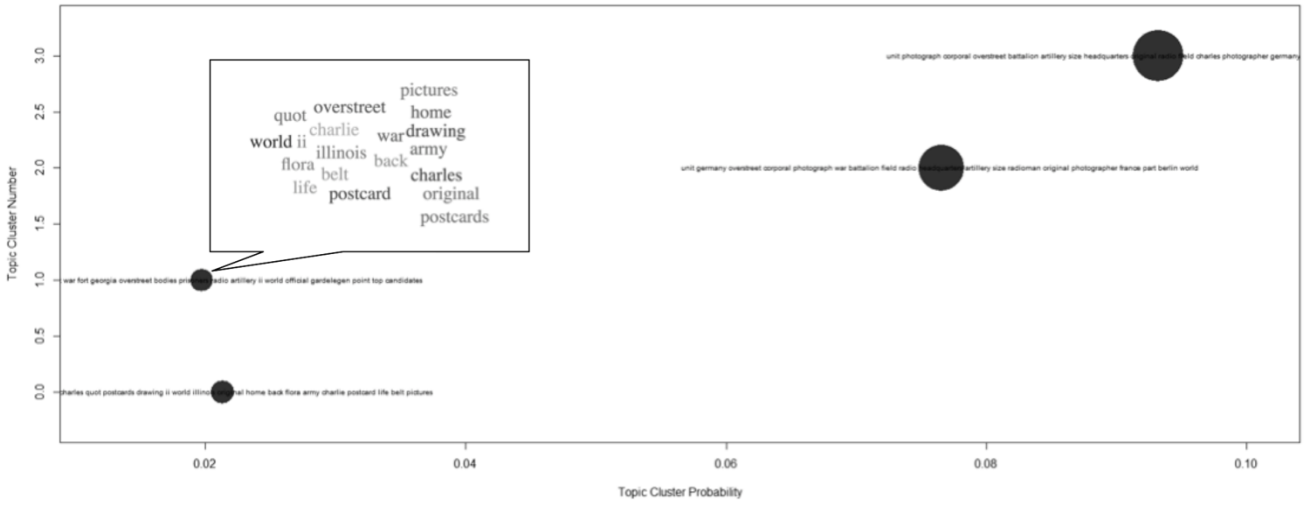

Figure 6: Focus+Context Topic-Based Visualization for "Charles Overstreet Collection": Example of Four Topic Clusters

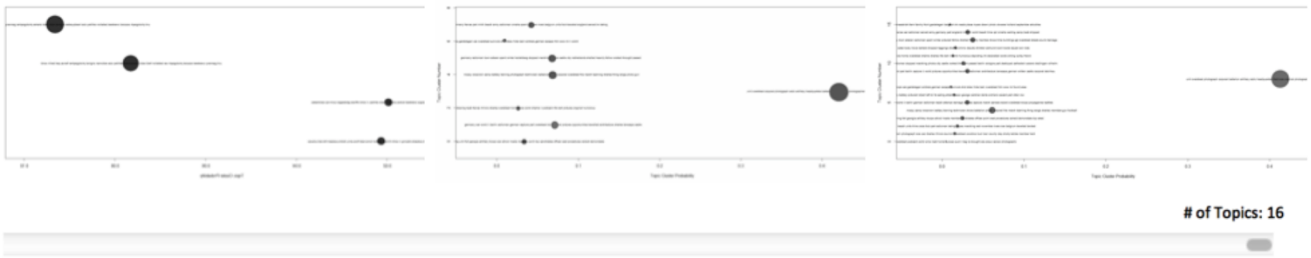

Figure 7: Interactive Topic-Based Visualization with Three Levels of Granularities: Number of Topics = 4, 8, or 16

CONCLUSIONS AND FUTURE WORK

In this study, we developed a topic-based visualization model to be added to an information search, retrieval, and access interface for large digital collections, enabling users to perform more efficient, effective, and accurate web search tasks based on the semantic relationships of documents. The hidden, thematic topics of documents are first extracted with topic modeling. The topic clusters are then visualized to present the different cluster probabilities, the words co-occurring within a specific topic cluster, and the probability distributions of topic clusters for a selected document. This new model offers advantageous opportunities to manage large digital collections and improve web search and retrieval capabilities for users of those collections. More specifically, this topic-based visualization model can (1) improve the users' experiences as they interact with large digital collections, thus offering a semantic-level approach to querying and retrieving data that allows users to visualize even more precise relationships between documents, (2) extend and advance the existing collection management options for librarians, archivists, and other information professionals who manage large digital collections, and (3) provide valued insights to scholars on the new research stream of visual data mining and its ability to undertake digital collection analysis.

While the proposed visualization model offers basic features to interact with users, more interactivity functions can be added in future work. Cybulski et al. [9] suggested that interactive data visualization provides opportunities for people to creatively engage with data analytics. More advanced interactive features, such as zoom in/out to display topic hierarchies, linkages between different words across different topic clusters, and the relevant positions of the topic clusters with respect to the whole set of collections, would significantly improve the capability and practical value of the model.

REFERENCES

[1] Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. *Proceedings of the Workshop on Languages in Social Media*. Portland, Oregon, USA, 23 June (pp. 30-38).

[2] Angsachin, A., Lursinap, C., & Sriborisutsakul, S. (2014). The application of artificial neural networks with the multilayer Perceptron classification for discovering webpages. *Proceedings of the International Conference on Intelligent Systems, Data Mining and Information Technology*. Bangkok, Thailand.

[3] Bailey, J. (2014). Speak to the eyes: The history and practice of information visualization. *Jefferson Bailey*. Retrieved from http://www.jeffersonbailey.com/speak-to-the-eyes-the-history-and-practice-of-information-visualization/

[4] Bass, K. M., Puckett, C., & Rockman, S. (2008). Models of digital collection use in a university community. *Educational Technology*, *48*(1), 44-49.

[5] Brown, N. (2015, May/June). Helping members of the community manage their digital archives: Developing a personal digital archiving workshop. *D-Lib Magazine*, *21*(5/6).

[6] Card, S. K., Mackinlay, J., & Shneiderman, B. (1999). *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann.

[7] Cartolano, R. T., Davis, S. P., & Fabian, C. A. (2015). Less code, more product: Leveraging open source technologies to develop digital library collections. *Columbia University Academic Commons*.

[8] Clough, P., & Sanderson, M. (2013). Evaluating the prformance of information retrieval systems using test collections. *Information Research*, *18*(2).

[9] Cybulski, J. C., Keller, S., Nguyen L., & Saundage, D. (2015). Creative problem solving in digital space using analytics. *Computers in Human Behavior*, *42*, 20-35.

[10] Deal, L. (2014). Visualizing digital collections. *Technical Services Quarterly*, *32*(1).

[11] Donath, J., Karahalios, K., & Viegas, F. (1999). Visualizing conversations. *Proceedings of Hawaii International Conference on System Sciences*. Maui, Hawaii, 5-8 January 5-8.

[12] Gengenbach, M. J. (2012). The way we do it here: Mapping digital forensics workflows in collecting institutions. Retrieved from http://digitalcurationexchange.org/system/files/gengenbach-forensic-workflows-2012.pdf

[13] Greene, D., O'Callaghan, D., & Cunningham, P. (2014). How many topics? Stability analysis for topic models, machine learning and knowledge discovery in databases. *Lecture Notes in Computer Science*, *8724*, 498-513.

[14] Herman, I., Melancon, G., & Marshall, M. S. (2000). Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, *6*(1), 24-43.

[15] Horava, T. (2010). Challenges and possibilities for collection management in a digital age. *Library Resources & Technical Services*, *54*(3), 142-152.

[16] Hsu, F., Lin, C., & Fang, C. (2015). The trend of researches in digital archives. *Proceedings of the IC-ITS 2015 International Conference on Information Technology & Society*. Kuala Lumpur, Malaysia.

[17] Lee, C. A. Kirschenbaum, M. G., Chassanoff, A., Olsen, P., & Woods, K. (2012). BitCurator: Tools and techniques for digital forensics in collecting institutions. *D-Lib Magazine*, *18*(5-6).

[18] Lemieux, V. (2012). Using information visualization and visual analytics to achieve a more sustainable future for future archives: A survey and critical analysis of some developments. *Comma*, *2*.

[19] Leung, C. S. K., Carmichael, C. L., & Johnstone, P., Yuen, D. S. (2013). Interactive visual analytics of databases and frequent sets. *International Journal of Information Retrieval Research*, *3*(4), 120-140.

[20] Loudon, L., & Hall, H. (2010). From triviality to business tool: The case of Twitter in library and information services delivery. *Business Information Review*, 27(4), 236-241.

[21] Naik, N. R., & Rao, A. M. (2011). Information search and retrieval system in libraries. *Proceedings of the 8th International CALIBER Conference*. Goa, India.

[22] Nasukawa, T., & Yi, J. (2003). Sentiment analysis: Capturing favorability using natural language processing. *Proceeding of the 2nd International Conference on Knowledge Capture*, Sanibel Island, FL, USA, 23-25 October (pp. 70-77).

[23] National Research Foundation [NRF]. (2010). Managing digital collections: A *collaborative initiative on the South African Framework.* Retrieved from http://digi.nrf.ac.za/publ/Managing%20Digial%20Collections.pdf

[24] Paolillo, J. (2000). Visualizing Usenet: A factor-analysis approach. Proceedings of Hawaii International Conference on System *Sciences*. Maui, Hawaii, 4-7 January.

[25] Park, J., & Tosaka, Y. (2010). Metadata creation practices in digital repositories and collections: Schemata, selection criteria, and interoperability. *Information Technology and Libraries*, *29*(3), 104-116.

[26] Sinn, D. (2012). Impact of digital archival collections on historical research [Abstract]. *Journal of the American Society for Information Science and Technology*, *63*(8), 1521-1537.

[27] Viegas, F. B., & Wattenberg, M. (2008). Tag clouds and the case for vernacular visualization. *Interactions*, *15*(4), 49-52.

[28] Wang, D., Irani, D., & Pu. C. (2014). Is email business dying?: A study on evaluation of email spam over fifteen years. *EAI Endorsed Transactions on Collaborative Computing*, *14*(1).

[29] Watkins, R. D., Sellen, A., & Lindley, S. E. (2015 April). Digital collections and digital collecting practices. Paper presented at the CHI 2015, Seoul. Retrieved from http://orca.cf.ac.uk/69328/1/CHI%20Collecting%20Final% 2013%20Jan.pdf

[30] Xu, W., Esteva, M., Jain, S. D., & Jain, V. (2014). Interactive visualization for curatorial analysis of large digital collection. *Information Visualization*, *13*(2), 159-183.

ACKNOWLEDGEMENT

The authors wish to express their gratitude to the Illinois State Library and the Illinois Digital Archives for providing the metadata sets used in this study. This research was also supported by a Research, Scholarship and Creative Activity (RSCA) grant and a College of Applied

Sciences & Arts RSCA Infusion grant at San José State University.

AUTHOR BIOGRAPHIES

Hsuanwei Michelle Chen is currently an assistant professor at San José State University School of Information. Her research and teaching interests include data mining, social media, virtual communities, and online user behavior. In particular, she is interested in investigating the range of motives that drive online users to seek and exchange information and the interaction dynamics shaped by the networked environments. Chen received her Ph.D. in Information Systems from the University of Texas at Austin and her M.S. and B.S. in Computer Science and Information Engineering from National Taiwan University.

Hongbo Zou is a doctoral student in the San José Gateway PhD program, an international doctoral degree program offered in partnership between the San José State University School of Information and Queensland University of Technology. His primary areas of research interests include big data mining and online social network analysis. Currently, he is working on a research project that studies how social media mining can be utilized to help libraries better engage their users and provide improved services accordingly.

Alyce L. Scott is currently a lecturer at San José State University School of Information. Her research and teaching interests include digital collections, metadata, and data mining. In particular, she is interested in investigating how users search, explore, and use digital collections. Scott received her M.S. in Library and information Science from the University of Illinois at Urbana-Champaign.