Data Mining, Visualizing, and Analyzing Faculty Thematic Relationships for Research Support and Collection Analysis

*Wenli Gao and Loretta Wallace**

Introduction

An academic library's ability to make informed collection decisions and provide customized research support is dependent in large part on how well the research interests of the university's faculty are mapped out. Liaison librarians need to understand as thoroughly as possible the research interests and the corresponding research history of the faculty within the departments they serve. Traditionally, liaison librarians have relied upon departmental webpages, the curriculum vita of faculty, and personal interactions with faculty and researchers in order to gain a greater understanding of their research interests. This process is not particularly efficient since it is highly dependent on how often a department updates the information used for internal and external consumption. The strength of those relationships with faculty may be impeded by the number of faculty a liaison may have for each subject area they serve. One method some librarians have used to address this issue has been to utilize citation analysis to understand faculty information use behavior.[1] However, this kind of analysis explores journal title level information instead of individual article level information. As digital scholarship increasingly becomes one of top trends in academic libraries,[2] applying new technologies and tools such as text analysis, together with data visualization, has gained popularity within the research library environment.

This study analyzes faculty article publications across the entire campus, involving the innovative use of similarity analysis, topic generating algorithms, and visualization techniques. The overall goal of the project is to develop analytical tools to improve collection development and to subsequently enhance research support. This paper highlights the process of using a data mining approach to more fully understand the research focus of faculty and how to illustrate their research interests through visualization.

Literature Review

Data mining is a technique that uses statistical algorithms to identify patterns and relationships within a data set. It is utilized to analyze large information repositories and to discover implicit information.[3] The use of data mining to track library user behavior has grown with the advancement of digital libraries, and with these developments there is a growing need to understand how users interact with the technologies used.[4] Text mining, which

* *Wenli Gao is Communication, Sociology, and Anthropology Librarian, University of Houston, e-mail: wgao5@uh.edu; Loretta Wallace is Business Librarian, University of Texas at Austin, e-mail: l.wallace@austin.utexas.edu.*

is often also referred to as text data mining, includes information extraction, topic tracking, summarization, categorization, clustering, concept linkage, information visualization, and question answering.[5] The process has been used in many areas, for example, Kaplan used lexical analysis, a form of text mining, to assess student writing in statistics, incorporating Light Summarization Integrated Development Environment (LightSIDE) and IMB/SPSS Text Analytics for Surveys (TAS) software to analyze the student papers. The results produced by the two software programs were compared with each other, and also compared with results obtained from hand coding the same data sets.[6] Tirunillai and Tellis used text mining to analyze chat data on product reviews across fifteen firms in several different markets, over a specified period in order to obtain information about consumer satisfaction.[7] In a research paper by Zhang and Gu, the authors outlined how text mining could benefit academic libraries, by using existing text data to support decision-making processes.[8] Analyzing library social media content such as Twitter feeds, has led to the discovery of distinct patterns of communication, as well as the interactions between libraries and their users.[9] Another article discussing using similar processes to analyze library Twitter feeds expanded its focus by exploring the role of word frequency in understanding user behavior.[10]

A topic model, which is a type of statistical model used to discover abstract "topics" that occur within a collection of documents, is often used to analyze the co-occurrence counts of words or phrases, generating a set of topics or themes.[11] Latent Dirichlet Allocation (LDA) is a generative model that allows sets of observations to be interpreted by unobserved groups that explain why some parts of the data are similar.[12] Recently, there were several articles published, focusing on the humanities, that used topic modeling.[13] All of them discussed the potential of expanding the use of this technique in other research areas.

While there is belief that text mining and topic modeling can help to extract useful information from large-scale data sets, it is still hard to understand the overall patterns and trends. Visualization tools transform complex data into visual representations, thus in turn helping to identify structure, patterns, trends, anomalies and relationships within the data.[14] Traditional data visualization often involves the use of charts and diagrams generated through Microsoft Excel or SPSS. Recently, libraries have begun to use more advanced data visualization tools to help interpret the meaning behind the data collected for a variety of purposes. Murphy published an article discussing the use of Tableau in supporting academic library assessment. She provided three examples from the Ohio State University Libraries where the use of data visualization highlighted three different perspectives: 1) promoting library collections, 2) informing digitization priorities, and 3) visualizing library survey results.[15] Researchers at the Stockton College of New Jersey wrote about using NodeXL to visualize their library's Twitter network in addition to visualizing the types of connections within the library's network.[16] While these visualizations provide assistance when characterizing the meaning of a large set of data, along with making numerical or text data more visually appealing, they are not intended to be used as tools to facilitate data analysis. Tukey introduced the concept of Exploratory Data Analysis (EDA), which has played a significant role in quantitative research.[17] EDA combines visualizations with quantitative analysis, serving to check basic assumptions, reveal errors in data processing, identify relationships between variables, and suggest preliminary models.[18] By incorporating the idea of EDA into this research study, we aim to fill a gap in the library literature and build a bridge between library research and humanities research.

Methods

Data Collection and Data Processing

Elsevier's Scopus was used to identify publications from University of Houston faculty. The search was conducted in February, 2016 to capture articles published in 2015. The research team searched for the University of Houston as an affiliation, and limited the data extracted to include either articles or conference proceedings published from 2006 to 2015, a ten-year time span. The results were also limited to specify English language articles only. A total of 16,771 records meet the search criteria and were downloaded from Scopus. The records included 12,952 (77.23%) articles and 3,819 (22.77%) conference papers. The records were downloaded as comma-separated values (CSV) files and the author names along with institutional affiliations were in the same column. To locate a specific department within the University of Houston, a python script was written to remove collaborating authors from other institutions. The results included only the University of Houston faculty name and department designation.

This step was necessary in order to isolate and target departmental research trends. The study selected three subject areas to analyze at the departmental level: chemistry, computer science, and psychology. Before analyzing the department information, the research team normalized the department names. The raw data from Scopus used the department name provided by the authors, which was not consistent. For instance, the department of computer science was variously listed as computer science department, department of computer science, department of CS, and CS department. To include these as one department, the research team normalized all of the department designations as computer science. Since one of the aims of the project is to discover research trends over time, the year of publication is crucial and is preserved for analysis. The team also deleted some of the information irrelevant to our research, such as author-provided keywords and Digital Object Identifier (DOI). As a result, the file used for examination contained only columns needed for our analysis and visualizations.

Word Frequencies and Topic Modeling

A customized python application was developed to analyze word frequencies, allowing users to select the number of words together. For this study, the research team explored one to four word frequencies. The more words together, the less co-occurrence of specific groupings. A topic modeling tool from Google code was downloaded and used for topic analysis. It is a graphical user interface tool using LDA topic modeling. It was chosen for the project due to its ease of use and the fact that it does not require coding by the user. The software gives users the ability to define the number of topics to be included and to import a unique stop words list. Both word frequencies and topic modeling were analyzed at the article title level. The research team analyzed topics for a ten-year span and then for each year individually. For word frequencies, departments were separated to gain a deeper understanding of the departments' research interests.

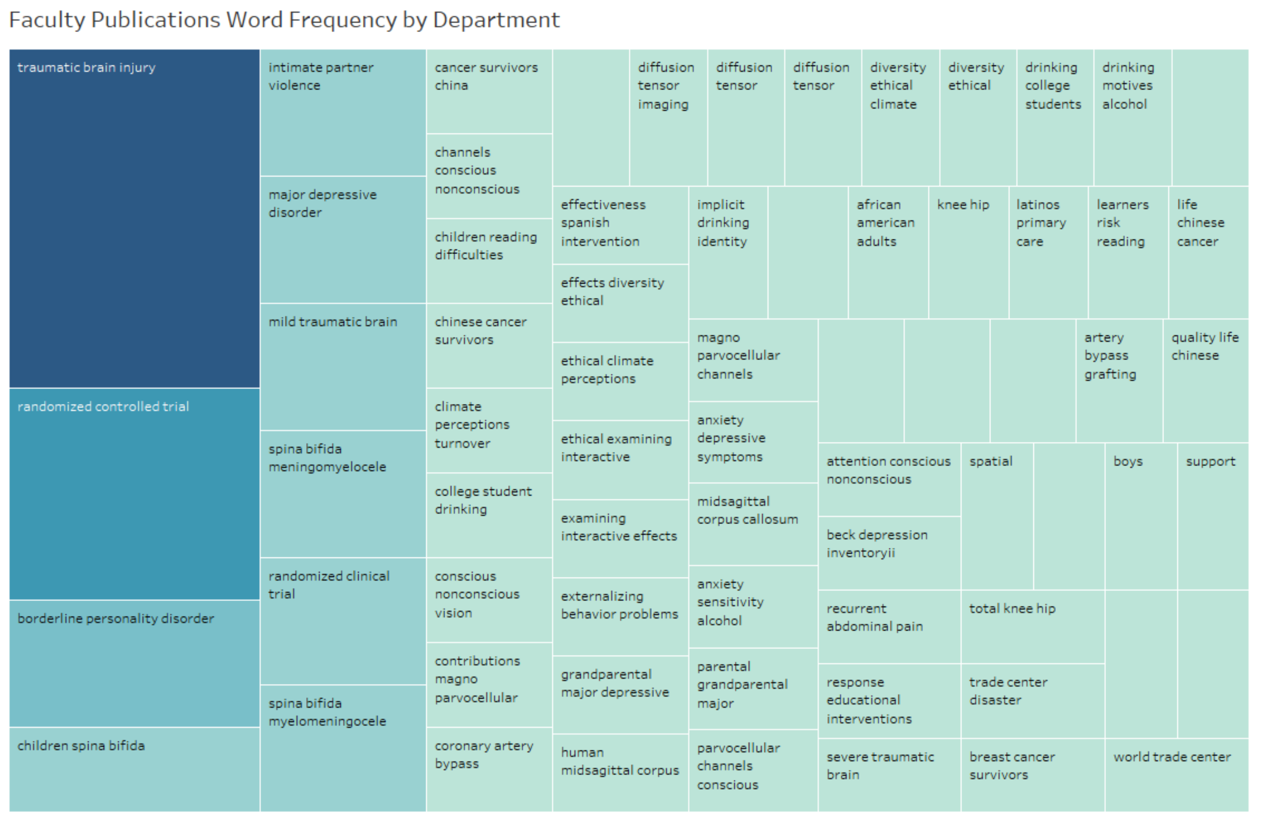
Data Visualization

The research team used Tableau 10.0 to visualize the word frequencies. Tableau enables the user to interactively select the criteria and explore the data on different levels. It is an effective way of portraying results for exploratory analysis. Most visualizations are for the most part created from numeric data. This study is an example of how Tableau can visualize text data.

Results

After reviewing the data the research team found that two to four-word combinations give ready insights as to what the campus' research focuses were. For example, in 2006 the most frequent four-word combinations were "drug diffusion cornea sclera", "influence vitro experimental conditions", "experimental conditions drug diffusion", "vitro experimental conditions drug", and "conditions drug diffusion cornea". The results allowed us to see the shifts in research focus over time. In 2015, the word combinations reflected where the focus had shifted to following categories: "using optical coherence elastography", "mild traumatic brain injury", and "African American Hispanic women". Three-word combinations also provides similar information essential in understanding the research focus on campus and how trends have developed over the years. Figure 1 is an example of the three-word combinations of 2015 faculty publication titles.

FIGURE 1
Faculty Publications Word Frequency by Department



The research focus for the psychology department from 2006 to 2015 centered upon "traumatic brain injury", "randomized controlled trial", "borderline personality disorder" and "children spina bifida". For the computer science department, the focus from 2006 to 2015 concentrated on "wireless sensor networks", "3D2D facial recognition", and "air flow patterns". Internal research shifts within a department were also observed. For instance, the psychology department in 2006 had only two co-occurrences for the three-word combinations, "effectiveness Spanish intervention" and "learners risk reading". These subjects are considered learning focused. While in 2015, there were 17 three-word combinations ranging from "anxiety sensitivity alcohol", "personality disorder features" to "severe traumatic brain". It shows the expansion of the department in its research productivity among a wide variety of topics. Figure 2 is an example of the three-word combination from 2006 to 2015 for the psychology department.

The topic modeling tool was used to analyze similarities between publication titles and observe if there were additional trends besides topics identified by word frequencies. Combining ten years of data, the topics generated were: cells receptor cancer human activity production analysis expression b quality. Within each year, additional topics are identified. Figure 3 lists the topics by year.

FIGURE 2
Faculty Publications Word Frequency by Department

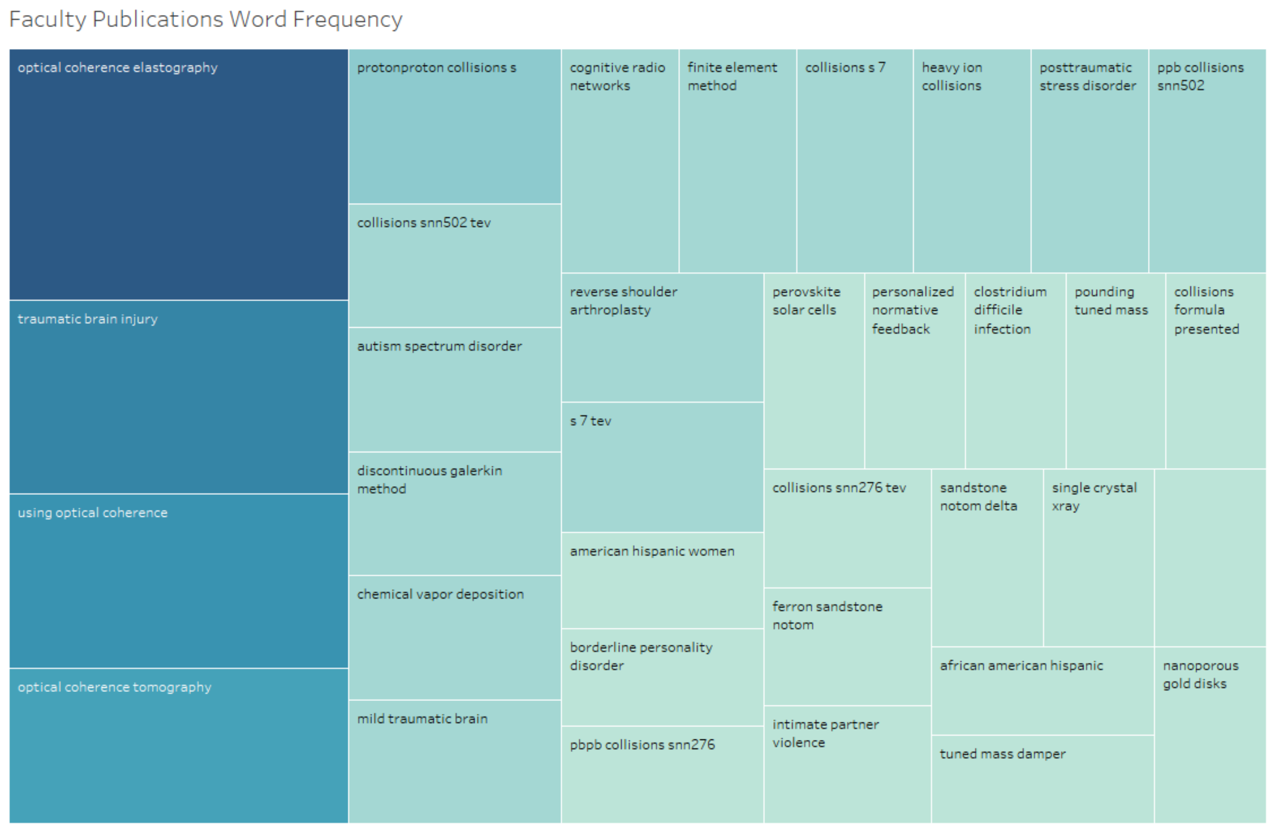


TABLE 1
Faculty Publications Topic by Year

Year	Topic
2006	risk applications reading research methods intervention dimensional force thermal states
2007	self method response flow image blood therapy group problems vibration
2008	data magnetic receptor activity large power polymer energy Houston interface
2009	data cells media applications multi pressure scattering formation problem receptor
2010	study control modeling memory new behavior development activity role
2011	based networks detection study near induced performance long human
2012	analysis data high performance human network study layer power injury
2013	data treatment applications films cell type nanoparticles development growth stress
2014	collisions pb performance children p au x monitoring gas
2015	based disorder impact risk low traumatic college carbon nanoparticles

Discussion

The results of this study can be used by librarians to restructure their collection development practices. By assessing multiple years of data, librarians can monitor research trends and observe how research topics evolve over time. Reviewing information of this type has become necessary as more universities strive to develop interdisciplinary programming to align with their overall mission. Librarians need to adapt to these rapid changes in curriculum and be adept at recognizing when attention should be paid to various subjects or developing academic themes, all of which may allow for the reallocation of funds to focus upon major or new research areas. Another advantage is the ability to identify university wide thematic research clusters within an individual college, singular and/or multiple subject areas within departments, or at the faculty level, thus enabling librarians from different liaison areas to work together to make collaborative collection development decisions.

An additional benefit of the study is the ability to enhance outreach efforts derived by studying faculty research output. Librarians will also have definitive material to use when targeting researchers with customized research support. In addition, liaisons will be in the position to assist faculty, by demonstrating how these tools can be used to find potential collaborators among their peers and bolster research partnerships within the university.

By incorporating a graphical user interface tool for LDA topic modeling, the research team explored large and what may be perceived as complicated data sets. Moreover, interactive visualizations such as Tableau make the results of this study easy to understand and visually appealing, creating a more comprehensive picture of the overall research which can be accomplished by using the methods included within the study.

Limitations and Direction for Future Research

There are some limitations with this research. First, the faculty publication information was retrieved solely from Scopus, thus including only publications indexed in Scopus. Also, in some disciplines, especially in the humanities, faculty publish more books than articles. It would be useful to include book publications to reflect the research output for faculty from those disciplines and show a more complete picture of the research trends. Moreover, faculty within the sciences have more articles and conference proceeding than their peers within the humanities and the social sciences. When counting word frequencies, science topics have more co-occurrences than topics in the humanities and the social sciences. Librarians could have a better understanding of those subject areas if these three subject areas were analyzed separately. During the study the research team encountered a problem with some of the data due to the inconsistent use of departmental names. The issue was alleviated manually and future studies will most likely experience the same problem, having to normalize department names so that every subject can be analyzed individually. In addition, faculty publications usually reflect faculty research interests, but it does not necessarily indicate what their teaching interests may be. While librarians can use the tools developed in this research for collection development, they should also seek other methods to determine what faculty interests may be in the way of instruction support.

When analyzing departmental information, it would be useful to identify similar research focuses across different units. Librarians could then use this information to collaborate with other librarians when developing collections. It would also be useful to connect the research topics of faculty to help build research partnerships. An exploratory visualization tool can be developed for faculty to identify other researchers conducting similar research studies, thus promoting collaboration across disciplines within the university. Moreover, researchers can include more faculty publication information as well as include more institutions into the analysis and add geographic information and other information as desired. For example, a faculty can select a research focus, find others with similar research interest, and limit by location or institution type. This will be a great tool for people to find potential collaborators.

Conclusion

Overall, the results of the study showed how the pairing of text mining and topic modeling software, along with data visualization tools, can provide insight into the nature of the types of research being conducted across a university's campus. The text mining and topic modeling tools used enabled the research team to examine data which was originally in text format only. The data used for the project was extracted directly from the citations of works published by researchers at the University of Houston. This information allowed the group to pursue the process of analyzing the placement of words and the ensuing subject combinations. By using a broad range of citations covering a ten-year span, the research team could effectively examine the growth of research subjects and the diverse nature of those themes, over time. Though the results of the analysis revealed a variety of interests among the faculty across campus, the work performed by the research team only scratched the surface of what could be accomplished. It should be noted that without the use of visualization tools to create clear and concise images from the research, the results of the study would not have been as useful or compelling.

There is an assortment of benefits arising from the study, one being how the process could enrich collection development practices by exploring the research ecosystem of a university, thereby, setting the foundation for monitoring trends evolving within research clusters. Finally, the research project helped to provide a clearer picture into the research behavior of the faculty at the University of Houston, allowing the group to investigate more fully faculty interests and research partnerships. This has opened new pathways to discovering the interdisciplinary nature of the research being conducted across campus.

Notes

1. Wenli Gao, "Information Use in Communication Research: A Citation Analysis of Faculty Publication at the University of Houston," *Behavioral & Social Sciences Librarian* 34, no. 3 (2015): 116–128; Cory Tucker, "Analyzing Faculty Citations for Effective Collection Management Decisions" *Library Collections, Acquisitions, and Technical Services* 37, no. 1–2 (2013): 19–33.
2. ACRL Research Planning and Review Committee, "2016 Top Trends in Academic Libraries." *College & Research Libraries News* 77, no. 6 (June, 2016): 274–281.
3. Jiawei Han, Micheline Kamber, and Jian Pei, *Data Mining: Concepts and Techniques.* (Elsevier, 2011), 625.
4. Christos Papatheodorou, Sarantos Kapidakis, and Michalis Sfakakis, "Mining User Communities in Digital Libraries," *Information Technology & Libraries* 22, no. 4 (December 2003): 152–157.
5. Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, "Tapping the Power of Text Mining," *Communications of the ACM* 49, no. 9 (September, 2006): 76–82.
6. Jennifer Kaplan, Kevin Haudek, Minsu Ha, Neal Rogness, and Diane Fisher, "Using Lexical Analysis Software to Assess Student Writing in Statistics," *Technology Innovations in Statistics Education* 8, no. 1 (2014).
7. Seshadri Tirunillai and Gerard J. Tellis, "Mining Marketing Meaning from Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation," *Journal of Marketing Research* 51, no. 4 (2014): 463–479.
8. Yan Zhang and Haiming Gu, "Text Mining with Application to Academic Libraries," In *Computer Science for Environmental Engineering and Ecoinformatics*, (Springer, 2011), 200–205.
9. Sultan M. Al-Daihani and Suha A. AlAwadhi, "Exploring Academic Libraries use of Twitter: A Content Analysis," *Electronic Library* 33, no. 6 (2015): 1002–1015.
10. Sultan M. Al-Daihani and Alan Abrahams, "A Text Mining Analysis of Academic Libraries' Tweets," *The Journal of Academic Librarianship* 42, no. 2 (March, 2016): 135–143.
11. David Blei, "Probabilistic Topic Models," *Communications of the ACM 55,* no. 5 (April, 2012):77–84.
12. David Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research* 3, (2003): 993–1022.
13. David J. Newman and Sharon Block, "Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper," *Journal of the American Society for Information Science and Technology* 57, no. 6 (2006): 753–67; Matthew L. Jockers, *Macroanalysis: Digital Methods and Literary History*. (University of Illinois Press, 2013); Matthew L. Jockers and David Mimno, "Significant Themes in 19th-Century Literature," *Poetics* 41, no. 6 (2013): 750–769; Andrew Goldstone and Ted Underwood, "The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us," *New Literary History* 45, no. 3 (2014): 359–384; Lauren F. Klein, Jacob Eisenstein, and Iris Sun, "Exploratory Thematic Analysis for Digitized Archival Collections," *Digital Scholarship in the Humanities*, no. 30 (2015): 130–141.
14. David P. Tegarden, "Business Information Visualization," *Communications of the Association for Information Systems* 1, (1999):4.

15. Sarah Anne Murphy, "How Data Visualization Supports Academic Library Assessment Three Examples from the Ohio State University Libraries Using Tableau," *College & Research Libraries News* 76, no. 9 (2015): 482–486.
16. Jewelry Yep and Jason Shulman, "Analyzing the Library's Twitter Network Using Nodexl to Visualize Impact," *College & Research Libraries News* 75, no. 4 (2014): 177–186; Jason Shulman, Jewelry Yep, and Daniel Tomé, "Leveraging the Power of a Twitter Network for Library Promotion." *Journal of Academic Librarianship* 41, no. 2 (2015): 178–185.
17. John Tukey, *Exploratory Data Analysis.* (Addison-Wesley Publishing Company, 1977), 3.
18. Andrew Gelman, "Exploratory Data Analysis for Complex Models," *Journal of Computational and Graphical Statistics* 13, no. 4 (2012):755–779.