Social Media Article Visualizer Project

Project Investigator Contact Information:

Stephen Carradini, Assistant Professor, Technical Communication program
College of Integrative Sciences and Arts, Polytechnic Campus, Arizona State University

7271 E. Sonoran Arroyo Mall, 251F
Mesa, Arizona 85212-2780
Stephen.Carradini@asu.edu
480-727-3687

Administrative Contact Information:

Ms. Virginia E. DeSanto, Treasurer
Arizona State University Foundation for a New American University

P.O. Box 2260
Tempe, AZ 85280-2260
OCFRgrants@asu.edu
480-965-3759

Project Description

This project's goal is to create a visualizer for social media articles, offering technical communication researchers a new way to map, understand, and use interdisciplinary articles on social media.

Verzosa Hurley and Kimme Hea (2014) argue that "we see continued critical study of social media as integral to [technical communication]." Technical communicators have taken up this call and investigated social media from many angles. However, technical communication is a naturally interdisciplinary field, and it does not seem that we are reading very interdisciplinarily when it comes to social media. Verzosa Hurley and Kimme Hea's article cites no journals outside writing studies when talking about the social media topic of "reach," despite prior articles in other fields on the topic, including two in *Journal of Medical Internet Research* (Mackey & Liang, 2013; Ramo & Prochaska, 2012). In a *CDQ* article, Hopton and Parry (2016) discuss the use of social media in a nonprofit. They cite *Public Relations Review* (2 articles), *Journal of Computer Mediated Communication* (2 articles), and *Journal of Social Marketing* (1 article) as journal articles of other disciplines out of their 31 sources. Articles from specially tailored journals such as *Nonprofit and Voluntary Sector Quarterly* or large journals like *New Media and Society* could connect TC with the most up-to-date research on nonprofit use of social media.

I do not see this state of affairs as a failure of the authors, but as a difficulty of technological/academic systems: I argue that many articles on social media are published in journals spread across fields that are not indexed in our disciplinary journal databases. I expect that other fields do not cite our work in social media for the same reason. Being able to quickly find cutting-edge research from multiple fields could allow researchers in technical communication and all fields to stay up-to-date on research about social media topics and not reinvent the wheel through similar studies. Furthermore, current databases often return results in list format, which makes understanding the whole shape and geography of social media research on a topic difficult. Understanding visually where certain types of social media research are published could allow researchers to make different types of connections between research than a list format may allow. I am proposing a tool that aims to accomplish those goals.

Benefits to SIGDOC

This project will benefit SIGDOC and SIGDOC members via a useful tool and meta-research. The tool will benefit SIGDOC members by allowing researchers to search for terms within the 4,000,000+ social media articles of the tool and visualize in map form what fields are creating research about that topic. This is an alternative information design from the list format that journal databases use. While the journal article metadata information will be drawn from metadatabase Microsoft Academic, it will be presented as a network map that will allow researchers to more quickly understand insights such as "It seems that engineering journals have a great amount of research about (social media concept)" or "Medical journals have much more work on (social media property) than agricultural journals." These are questions that can be understood in a map format potentially faster than in a list format. This front end of this tool will be published/hosted on my website: StephenCarradini.com. This will allow anyone with a browser to access the tool and use it for research.

SIGDOC will also benefit from meta-analysis articles on technical communication social media research using the tool. This tool will allow me as a researcher to see which fields cite *Communication Design Quarterly* social media research and who researchers in *Communication Design Quarterly* are citing. I propose to present on this project at SIGDOC 2020 and publish the article about *CDQ* citations in *CDQ*. This tool also allows for meta-research on social media in technical communication: I will write an article that will look at all of the articles regarding social media in major technical communication journals and assess what we know about social media as a field. Other future articles could focus on interdisciplinary integrative literature reviews developed through the tool: what do social media researchers know about Twitter influencers? What do faculty outside business schools know about Kickstarter?

Project Methodology / Research Tasks

The tool is already being designed. I am working with a team of six undergraduate computer science students as their capstone project coordinator. They have broken up the tasks into scraping, backend development, frontend/UI development, and visualizations. A different group of students created a pilot version of the software in 2018-2019. The current students will complete a live beta of the project by May 2020. (If more development is needed after May 2020, I have the ability to work with a third group of students to develop the tool further.) The tool will be a web-based project hosted on DigitalOcean's cloud service. IRB approval is not needed for this project, as human subjects are not involved in any capacity.

We are scraping from Microsoft Academic because it has an enormous amount of information. It indexes 48,838 journals and 229,591,865 articles from those journals (Microsoft, 2019). Disciplinary databases familiar to technical communication are smaller than this. Communication and Mass Media Complete indexes 1300 journals (EBSCO Industries Inc., 2019), while JSTOR promises "more than 12 million journal articles, books, and primary sources in 75 disciplines" (ITHAKA, 2019). Microsoft Academic is larger than disciplinary and even some interdisciplinary databases. This immense size makes it a good fit for this interdisciplinary project. We have already scraped more than 3 million citations referencing social media. The process for creating this tool is:

1. Scrape selected metadata (article title, abstract, author names, journal name, doi, reference list, who cited the article) from articles referencing social media in the title or abstract from Microsoft Academic and import these into a database. (We have decided not to use keywords because not all journals use keywords; those that use them do so in different ways.)
2. Categorize these citations into fields that have published these social media articles. This is done by using SCOPUS Journal Rankings' pre-existing journal categories; SCOPUS has 40,503 journals categorized into 27 fields and more than 100 sub-fields.
3. Visualize these articles in a map.
   a. The flow of the software starts with the user searching a term within the program's UI (Twitter, Facebook, Reddit, reach, nonprofit, etc.) and have only the nodes that feature work with that term appear; all other nodes would disappear.
   b. The program will generate a map showing how many articles use that term in the article abstract within each of the 27 subject areas. This number will be represented as a numeral superimposed on a circle with a title that represents an individual subject area. The map will list how many articles are in each of the 27 subject areas (e.g. 1000 in agriculture, 20000 in communication, etc.). This map shows users in a visual format what fields house work about a particular subject, which is an important outcome of the program.
   c. Users can click on one of the 27 circles to see the amount of articles including that original searched term in the field's attendant subfields.
   d. Users can click on a subcategory and see all journals within that subfield that include articles on the topic and the number of articles within each journal.
   e. Users can then click on a journal to see each article in that journal on the topic. From there, the user can look at the metadata for each article or move to the citation view.
   f. In the citation view, the subject article will be represented in the center of the screen as a circle. The circle will have lines leading to the left of the circle, where all of the articles that the subject article cited will be shown and clustered together in relation to which of the 27 categories the cited article is published in. The circle will have lines coming out of it to the right, showing all of the articles that have cited the subject article, again clustered by the category they are published in. Clicking on any of these articles will allow the user to move to the citation view of that article. The user will be able to turn off "cited by" or "has cited this article" so that only one of those types of lines appears in the citation view.
4. Develop an exporting function for the maps
5. Create a scheduled scraper to add new articles as they are found

Limitations
Using any database means the contents of the tool are subject to the contents of the database; this is why we have used the largest available database we can find. (Google Scholar may be bigger, but it is not easily scrapable due to a Google-imposed limit on the amount of pages we can view. It has no API that we know of.) Second: there may be journals or proceedings that are not indexed in SCOPUS' provided categories. These will be included in a 28th category, sub-clustered by journal. Third: SCOPUS categorization does not extend to books, so books are not included. Fourth: the references of journal articles using Chicago footnote style may be indexed imperfectly or not at all by Microsoft Academic. This limitation will be foregrounded in the program as a note attached to all citation-view maps.

Budget
The backend of the tool will be hosted in DigitalOcean. Some of the funds requested in this grant are requested to purchase hosting space for the tool. Some of the resources will be planned for current and future scraping. We are planning to develop a way to continuously add newly-published articles about social media into the database by continuously searching Microsoft Academic for new content added in the current year; for instance, in 2020, we will scrape for articles with 2020 publish dates and add those to the database. SIGDOC funds will allow us to do this for 2020, 2021, and 2022. We intend to do this in a rolling process, but we may have to do it once a year at the end of the year. However, we intend for this to be a dynamic project, just as study of social media is dynamic.

Cost Categories	Year 1 2/1/2020 1/31/2021	Year 2 2/1/2021 1/31/2022	Year 3 2/1/2022 1/31/2023	Cumulative
Other Direct Costs:	**$480**	**$480**	**$230**	**$1190**
Data Collection: DigitalOcean Droplets	$120	$120	$80	$320
Data Storage: DigitalOcean Spaces	$360	$360	$150	$870
Direct Costs:	**$480**	**$480**	**$230**	**$1190**
Indirect Costs:	**$0**	**$0**	**$0**	**$0**
Total Direct and Indirect Costs:	**$480**	**$480**	**$230**	**$1190**

This budget proposes to use SIGDOC support for the purchasing of space in the online software platform DigitalOcean for assistance in data collection and then data storage for the project.

Data Collection (DigitalOcean Droplets): The cost for data collection is $40/month for 3 months ($120) in the first two years and $40/month for 2 months in the third year through a service called DigitalOcean Droplets. The Droplets will be used to run the scraping tool that collects articles and citational information on social media articles.

Data Storage (DigitalOcean Spaces): The cost for storage of data through DigitalOcean Spaces would be $30/month for 29 months ($870). The database, backend, and frontend code would be stored in Spaces procured by this grant for 29 months; after this approximately 2.5 years, another funding method will be procured to keep the tool running for the next three years.

The total budget costs would be $1,190 of the $1,200 maximum.

References

EBSCO Industries Inc. (2019). *Communication & Mass Media Complete*. Retrieved from
https://www.ebsco.com/products/research-databases/communication-mass-media-complete

Hopton, S.-B., & Parry, M. R. (2016). Saving the sea, socially: Measuring the relationship between
content and gesture on Facebook. *Communication Design Quarterly, 4*, 32-43.

ITHAKA. (2019). *About JSTOR*. Retrieved from https://about.jstor.org/

Mackey, T. K., & Liang, B. A. (2013). Global reach of direct-to-consumer advertising using social media
for illicit online drug sales. *Journal of Medical Internet Research, 15*, e105.

Microsoft. (2019). *Microsoft Academic*. Retrieved from https://academic.microsoft.com/home

Ramo, D. E., & Prochaska, J. J. (2012). Broad reach and targeted recruitment using Facebook for an
online survey of young adult substance use. *Journal of Medical Internet Research, 14*, e28.

Verzosa Hurley, E., & Kimme Hea, A. C. (2014). The rhetoric of reach: Preparing students for technical
communication in the age of social media. *Technical Communication Quarterly, 23*, 55–68.
10.1080/10572252.2014.850854