ASEE'S VIRTUAL CONFERENCE
At Home with Engineering Education
JUNE 22 - 26, 2020 #ASEEVC
Paper ID #28952

Using Citation Analysis as a Collections Management Tool

Mr. Paul McMonigle, Pennsylvania State University, University Park

Paul McMonigle is the Engineering Instruction Librarian at the Pennsylvania State University. He graduated from Syracuse University with a MS-LIS degree in December of 2018 and from the Pennsylvania State University with a BA degree in History in 2017. His research interests include information literacy instruction for STEM students, student engagement and outreach programs, collections development and maintenance, and the history of STEM subject libraries.

Using Citation Analysis as a Collections Development Tool for Engineering Journal Subscriptions

Abstract

Citation analysis can be a powerful tool to help librarians learn user needs, determine patterns of sources for authors using their collections, and maintain those collections based on the needs and patterns identified. In an era of decreasing collections budgets, the knowledge gained by citation analysis can be beneficial to those tasked with collections development and maintenance. Academic libraries used by graduate student researchers need to have a wide variety of materials available in the collection and citation analysis can be used to discover which materials are used most often and by which programs. A dataset of citations created from the ProQuest Dissertations & Theses database was used to determine the amount of "in-house" journals used by engineering doctoral students at a major U.S. research university compared to the number of articles that the university did not have direct access to from 2015 to 2018. The information was sorted by discipline to show which programs were well represented in the collection compared to those that could use some reinforcement. Results show that the university libraries own a large majority of the sources used by the graduate students whose dissertations were selected for this study. The author will also be able to use the dataset to learn which source titles are used most often and where to direct collections funds to ensure continued development in areas most used by the university's graduate researchers. Other academic libraries can use the methods described in this study to verify the use of their own collections and make decisions accordingly.

Introduction

The Pennsylvania State University is a Carnegie Level-1 research university located in University Park, PA. The College of Engineering, officially founded in 1894, offers bachelors', masters', and doctoral degrees in ten departments. During the period covered by this paper (2015-2018), the College's average enrollment included approximately 8000 undergraduates and 1600 graduate students. Enrollment figures for graduate students do not specify whether they are masters' or Ph.D. hopefuls.[1][2]

The University Library system at Penn State is one of the largest in North America, with several million volumes of books and materials along with several thousand journal subscriptions. The vast diversity of the collection and its strength in engineering subject areas is a selling point when the university recruits both graduate students and potential faculty members. However, even a large budget is not an unlimited one and recently, the library has been forced to tighten control of the purse strings. Collections spending, especially spending on multi-year subscriptions, has become a major concern for administration.

As a librarian new to this institution, the author wanted to learn what resources were being used by the university community in order to assist them with collection development and maintenance issues while keeping in mind the mandate to save money whenever practicable. Circulation statistics can help with much of the material, but this information is hard to come by for online databases and subscriptions. Therefore, the author decided to conduct the following citation analysis of recently published graduate doctoral dissertations to determine which parts of the collection the students are using compared to how often they needed to use sources not held by the Library.

Literature Review

The first step, of course, was to see if such a project was even feasible and whether other institutions had done such a thing in the past. The research done in 2008 by Meier and Conkling [3] proved that large-scale analysis of ten thousand or more citations is possible with the assistance of computer database software. They also used dissertations in their research, which guided the author to use the same ProQuest database to find their samples.

The citation analysis conducted by Kayongo and Helm (published in 2012) [4] became a bit of a guide for how the research initially began. Although they covered all the major subjects of their library rather than just the holdings for one college, their results showed that using citation analysis on doctoral dissertations can give a librarian a good overview of which parts of the collection get used and which subjects may need a bit of work materials-wise.

Kirkwood's study in the use of citation analysis with civil engineering dissertations [5] can almost be considered a proof of concept for the idea as usable for engineering librarians. The research and methods of Becker and Chiware, which they called a "bibliometric approach", acted as a guide for how the author could sort the data once discovered [6]. Ahmadieh, et. al used the same approach for the same reasons (assessing their collection to determine ways in which it could be improved) as the author [7] while Abeyrathne's citation analysis covered one STEM area (in this case, agriculture) with multiple departments [8]. These gave us further proof that a citation analysis was one of the best tools currently available to understand how graduate students used library collections to assist them in the creation of their theses.

Methods

The initial phase of the project began with selecting a sample of dissertations that would be large enough to show patterns that could be generalized to the entire population, yet small enough that the data could be handled by one researcher in a relatively short period of time. The author used the ProQuest Dissertations and Thesis A&I (PDTAI) database because it is the official dissertation repository of the U.S. Library of Congress [9] and the doctoral students of the college are required to send a completed copy to the site for storage. The university had deposited 606 engineering dissertations between the years 2015 and 2018. After consulting with both the Head Engineering Librarian and the Head of the STEM Libraries, it was decided that a sample of 20% of all the dissertations could be used as a population for this study. This would return a large enough sample that could show trends within each discipline while remaining a project that can be completed by one person. 119 dissertations were chosen and out of those, 17,329 distinct citations were entered into the dataset. The subjects were divided into eleven categories that mirrored the academic departments of the university's College of Engineering (see Table 1 and Figure 1). The dataset included 20% of the dissertations in each subject rather than random dissertations across all subjects so that any gaps in the collection can be seen in each subject.

Program	**Total Number of Dissertations**	**Number of Dissertations Used**	**Number of Total Citations**
Acoustics	28	6	433
Aerospace Engineering	43	8	1087
Biomedical Engineering	32	6	855
Chemical Engineering	63	12	3145
Civil Engineering	32	6	958
Computer Engineering	102	20	2777
Electrical Engineering	108	21	2932
Environmental Engineering	9	2	370
Industrial Engineering	43	8	936
Mechanical Engineering	125	25	3391
Nuclear Engineering	21	5	445
Totals	606	119	17329

Table 1 – Total Dataset Population

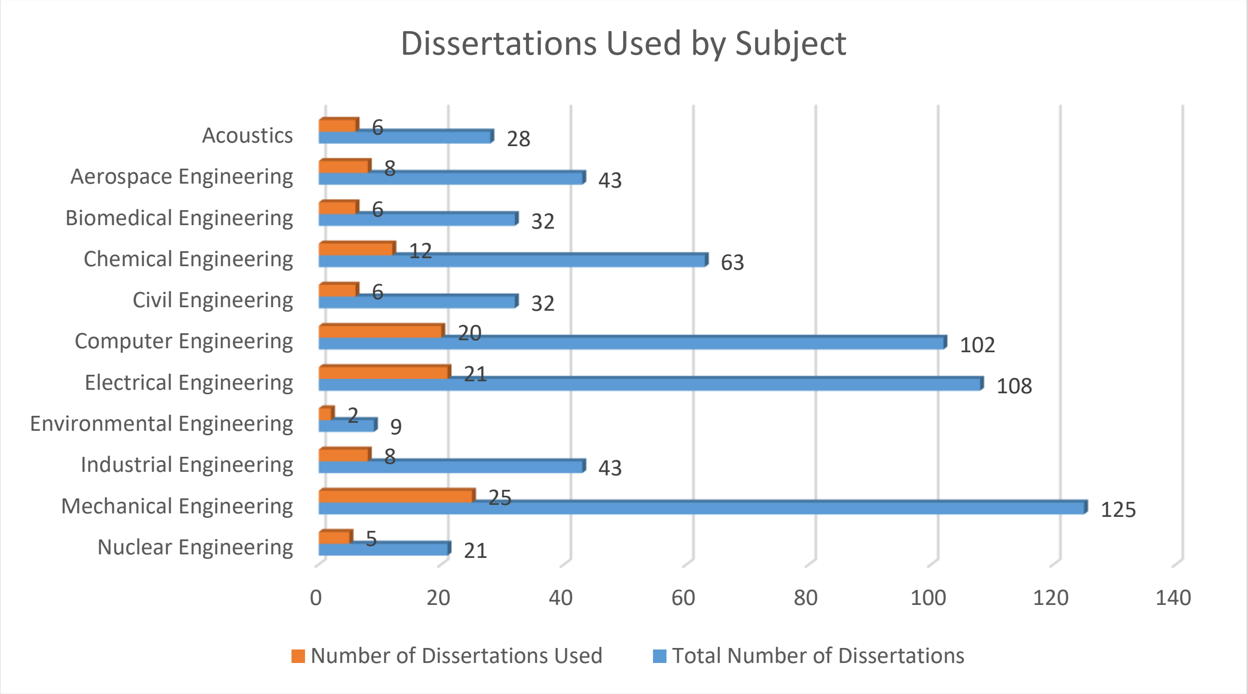

Figure 1

The original plan was to use the same program created for the research done by Cole, et. al, in 2018 [10] for their analysis. This program used separate citation lists that were found on the web page for the dissertation. Unfortunately, it was soon discovered that for many of the dissertations in the sample, there was no list and the only way to obtain the citations was to "copy and paste" directly from the document. Therefore, every reference cited had to be manually entered into a Microsoft Excel spreadsheet, with one sheet per subject. In each row, columns were created for author, title, journal or proceeding (if needed), date, and publisher (if known). The citations were then sorted into alphabetical order by source title (journal, conference proceeding, book, etc.) and any duplicates were eliminated, first by the tool within Excel and then double-checked by hand.

The results can be seen in Table 2. The entire process was rather time-consuming and took several months, since there was only one person working on the project. It was at this point in the project that the decision to only analyze journal citations was made, in order to complete the research in any sort of timely manner. All citations were transferred into the dataset anyway for possible use in future research.

Department	**Number of Journals used**	**Number of Articles Used**	**Number @ Library**	**Number not Held**	**Percentage @ Library**
Aerospace	149	521	500	21	95.97%
Electrical	290	1142	1091	51	95.53%
Industrial	243	516	490	26	94.96%
Biomedical	262	825	778	47	94.30%
Chemical	520	2441	2301	140	94.26%
Mechanical	462	1975	1833	142	92.81%
Acoustics	86	246	228	18	92.68%
Computer	276	642	594	48	92.52%
Nuclear	96	301	275	26	91.36%
Civil	200	452	401	51	88.72%
Environmental	112	333	286	47	85.89%
Total	2696	9394	8777	617	93.43%

Table 2 – Results

Once the dataset was complete, each journal title was entered into the library's catalog search engine one at a time to see if the university had a subscription. Dates of the subscription were checked as well to ensure that access was available to the dissertation author during the time of their research. This information was easily obtained by the search method mentioned above. If a search had a positive return, it was marked as "held by the library" and all relevant citations from that journal, provided the dates of the articles matched the dates of the holdings, were considered as accessible through the library. All negative returns were marked as "not owned by library".

Limitations

The first question that must be asked when conducting any citation analysis is how often did the students acquire articles through other means (Interlibrary loan, asking the author, etc.) for articles that the library already has access? Due to the way the system is currently set up, the link to a full-text article is directly next to a link proceeding to the ILL request form. How often one was chosen over the other can never truly be determined.

Another limiting factor were the citations themselves. Although most were complete and in some sort of discernable style, there were several dissertations that had incomplete citations (missing dates, journal titles, sometimes even authors). These partial citations were not used in the research. Fortunately for this project, they occurred rarely, with the vast majority in computer and electrical engineering. Even then, those partials numbered less than 100 out of over 5,800 citations between them.

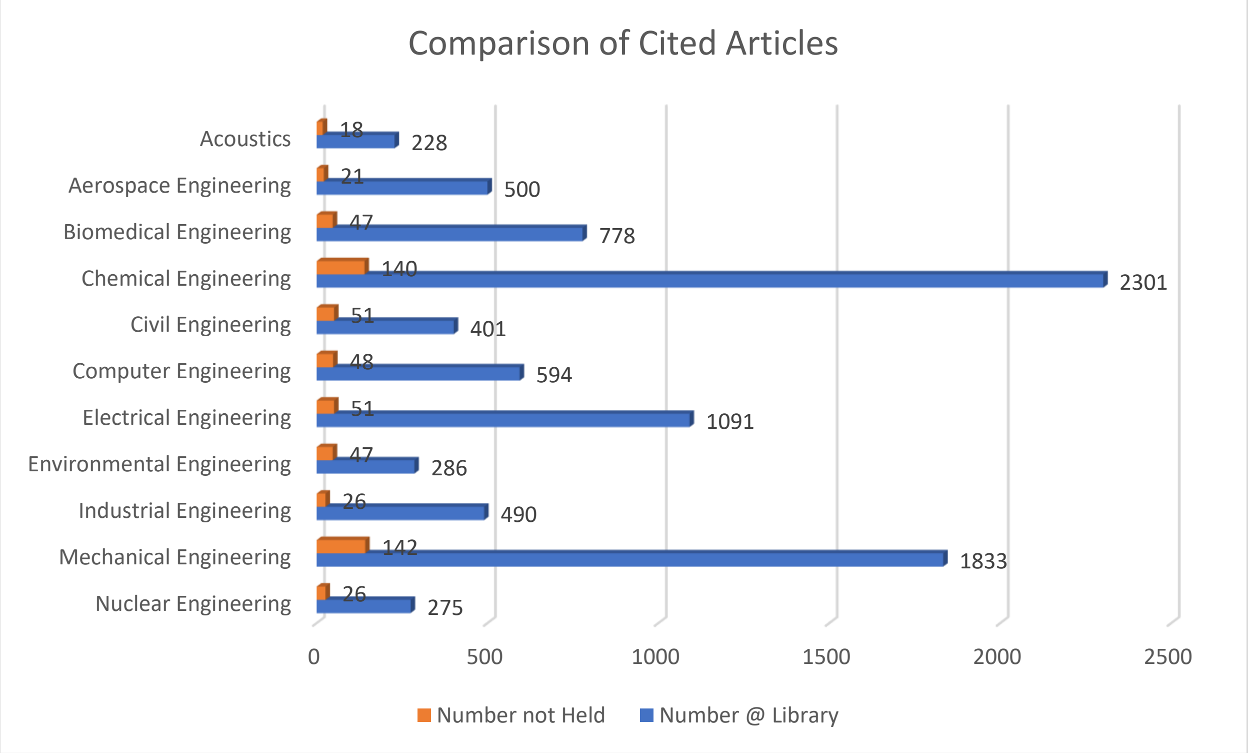

Figure 2

Results

Before the project began, it was determined that each subject should show a rate of at least 80% of citations belonging to materials held in the library's collection. This high number was chosen because of the great emphasis placed on engineering in the University's mission and vision statements as well as the overall size and breadth of the library's collections in general. It was (and still is) believed that in order to compete with peer institutions, the library needed to have a collection that meets most of the needs of the University's research community. At first, the author was concerned that 80% would be too high of a benchmark. This concern proved to be unnecessary.

The average dissertation had 145.6 citations, though this varied widely by subject, with Chemical Engineering averaging 262.1 while Acoustics averaged 72.2. Overall, the overwhelming number of citations could be found using library resources, with less than 10% of all journal articles needing interlibrary loan to find. This shows that the university has a robust array of resources available through subscription to its graduate students and most should have no trouble using what the library owns to complete their doctoral dissertations. As will be seen, the university's lowest covered percentage is in Environmental Engineering. This could be a result of the fact that the subject does not have a department of its own and is part of the Department of Civil Engineering, which had the second lowest percentage.

Aerospace Engineering had the highest covered percentage at 95.97%. In this case, however, the word “highest” does not have the weight it normally would since the overall average is only two and a half percentage points lower. With the exceptions of Civil and Environmental Engineering mentioned above, all disciplines came in over 90%. One detail that stood out to the author was how close Biomedical and Chemical Engineering were. Not only were they close in percentage, but they also shared many of the same journals (though care was taken to ensure there was no overlap in articles). It may or may not be a coincidence, then, that both department administrations are in the same building. However, the only other co-located departments in the college, Electrical and Computer Engineering, did not show this same close relationship, with Electrical above the average and Computer just below it.

Conclusion

The author was expecting that the library would have most of the resources used by students for their dissertations, but the extremely high percentages were a surprise. It was mentioned earlier under the limitations that students may have used other means to find sources that the library subscribed to (due to how the databases are set up). This research may also show the opposite: the students may not even realize that services such as Interlibrary loan are available options.

On the other hand, the research shows that the university library's engineering collection is robust enough to handle the needs of graduate students studying a wide variety of current issues in the discipline. Topics discussed in the dissertations included everything from underwater acoustics to propeller aerodynamics and from environmentally friendly lighting methods to cancer treatments. The author considers the current state of the collection to be of outstanding quality and will do their best to maintain that quality as best as possible for the foreseeable future.

Future Research

This project covering journal use is only the first step in determining overall collection use. The database created for this research includes information on books, conference proceedings, websites, technical reports, and even other dissertations that were used by engineering graduate students to complete their work. These will be looked at in an upcoming project.

Another interesting line of inquiry, although undoubtably larger than this one, would be to look at faculty publications. It would be interesting to see how the two datasets match up and where they would diverge. Comparisons could be made within disciplines as well as between them, and even between types of material used (books, journal articles, etc.).

The current research will be used as a baseline and the author will return to it in five years to compare new usage statistics to the existing database. Any trends that develop over that time can be discovered and it will be interesting to see if the planned physical and fiscal changes to the library have an effect on where the students find their sources.

When returning to this project, it would be better to use Microsoft Access instead of Microsoft Excel to create the datasets. The ability to easily sort and recall the data will speed up the process immensely and create less of a headache for researchers. Also, the dataset that was created for this paper includes a wealth of information that was not used, including books, conference proceedings, websites, and even other dissertations that were cited in the selected dissertations. These statistics can be used much like the journals to determine if and where there are any collection deficiencies. Future projects will go into more depth on these sources.

Another future research project could explore the other side of this question: which journals are not being used by our thesis-writing graduate students? As mentioned previously, we do need to be conscientious of the demands on our budget; it would be just as useful to determine which subscriptions are no longer needed (or not needed as much). The Library is a member of multiple state and academic library alliances and usually has no difficulty finding articles not held directly by it though Interlibrary loan and other means.

Acknowledgements

The author would like to thank John Meier, Beth Thomsett-Scott, and Paige Andrew for their assistance and encouragement during both the research and communication phases of this project.

The author received no financial support for the research, authorship, or publication of this article.

References

[1] The Pennsylvania State University, “Undergraduate Enrollment,” Penn State Engineering, 2019. [Online]. Available: https://www.engr.psu.edu/facts/undergrad-enrollment.aspx. (Accessed: 22-Dec-2019).

[2] The Pennsylvania State University, “Graduate Enrollment,” Penn State Engineering, 2019. [Online]. Available: https://www.engr.psu.edu/facts/grad-enrollment.aspx. (Accessed: 22-Dec-2019).

[3] J. J. Meier and T. W. Conkling, "Google Scholar's coverage of the engineering literature: An empirical study," *J. of Acad. Librarianship*, vol. 34, no. 3, 196-201, May 2008, doi: 10.1016/j.acalib.2008.03.002.

[4] J. Kayongo and C. Helm, "Relevance of library collections for graduate student research: A citation analysis study of doctoral dissertations at Notre Dame," *College and Res. Libraries,* vol. 73, no. 1, 47-67, Jan 2012, doi: 10.5860/crl-211.

[5] P. Kirkwood, "Using engineering theses and dissertations to inform collection development decisions especially in civil engineering," *Proc. of the Amer. Soc. of Eng. Educ. Annu. Conf.*, Austin, TX, June 2009.

[6] D. A. Becker and E. R. T. Chiware, "Citation analysis on masters' theses and doctoral dissertations: Balancing library collections with students' research information needs," *J. of Acad. Librarianship*, vol. 41, no. 5, 613-620, Sept. 2015, doi: 10.1016/j.acalib.2015.06.022.

[7] D. Ahmadieh, S. Nalbandian, and K. Noubani, "A comparative citation analysis study of masters' theses at the American University of Beirut, Lebanon," *Collection Build.*, vol. 35, no. 4, 103-113, Oct. 2016, doi: 10.1108/CB-06-2016-0013.

[8] D. K. Abeyrathne, "Citation analysis of dissertations for collection development," *Collection Build.*, vol. 34, no. 2, 30-40, Apr. 2015, doi: 10.1108/CB-11-2014-0055.

[9] ProQuest dissertations & theses global brochure. https://www.proquest.com/documents/ProQuest_Dissertations_Theses_Global_Brochure.html.

[10] C. Cole, A. R. Davis, V. Eyer, and J. J. Meier, "Google Scholar's coverage of the engineering literature 10 years later," *J. of Acad. Librarianship*, vol. 44, no. 2, 419-425, Mar. 2018, doi: 10.1016/j.acalib.2018.02.013.