ASEE'S VIRTUAL CONFERENCE

At Home with Engineering Education JUNE 22 - 26, 2020 #ASEEVC

Paper ID #29079

Getting Tired of Massive Journal Usage Statistics: A Case Study on Engineering Journal Usage Analysis Using K-Means Clustering

Ms. Qianjin Zhang, University of Iowa

Qianjin (Marina) Zhang is the Engineering and Informatics Librarian at the Lichtenberger Engineering Library, the University of Iowa. As a subject librarian, she manages collection and provides instruction, reference and consultation services for the engineering faculty and students. Her work also focuses on data management education and outreach to engineering students. She holds a MA in Information Resources & Library Science from the University of Arizona, and a BS in Biotechnology from Jiangsu University of Science and Technology (China).

Getting Tired of Massive Journal Usage Statistics: A Case Study on Engineering Journal Usage Analysis Using K-Means Clustering

Abstract

In 2018-2019, due to increases in the costs of information resources and flat collection budgets, University of Iowa Libraries has experienced a large-scale journal cancellation. As part of the University Libraries system, the Engineering Library went through a difficult process of identifying a list of journals with low usage and high cost, gathering feedback from our users and finalizing a list for cancellation. Since such a difficult situation may occur again in the future, we see the importance of continuously monitoring and evaluating collections in a proactive manner.

However, it would be challenging for engineering librarians who are responsible for both collection management and public service to review massive usage statistics on a regular basis. In order to tackle this challenge, we initiated a case study of measuring engineering journal usage in an alternative approach. The dataset was extracted from a data analytics company's journal usage statistics report prepared for the University of Libraries. We decided to reuse data from their report because it would save us time in data consolidation. The dataset contained journal titles, subfields and three key indicators including the number of publications per journal by authors of our institution, the number of citations to journal made by our authors and the number of downloads. Since the downloads were only available for the most recent four years (from 2015 to 2018), we selected the same period of data for the number of publications and the number of citations. We segmented a total of 821 journal titles into four clusters using K-Means clustering technique where the first cluster of 38 titles with a high number of publications, citations and downloads; the second cluster of 142 titles with a low number of publications but a moderate number of citations and a high number of downloads; the third cluster of titles with a low number of publications and citations but a moderate number of downloads; the forth cluster of titles with a low number of publications, citations and downloads.

In conclusion, our case study of measuring engineering journal usage converted massive journal usage statistics into four clusters of journal titles in a straightforward format. The clusters of journal titles also provided us with a comprehensive view on how engineering journals had been used by both authors and users of our institution in the most recent four years. Last but not the least, this case study showed a possibility of implementing data analytics in academic libraries.

Introduction

Like many other academic libraries, University of Iowa Libraries has been experiencing flat collection budgets for a couple of years and were hit with 5%-10% collection budget cuts (10% for collections related to natural sciences, engineering and health sciences; 7% for social sciences; and 5% for humanities) for fiscal year 2020. In response to the upcoming massive collection budget cuts for fiscal year 2020, in the fall semester of 2018, the Engineering Library as part of the University Libraries system went through the difficult procedure of identifying a list of journals that met the criteria of low usage and high cost, sharing the list with our users, especially faculty, and submitted the revised list to the University Libraries. After that, the University Libraries published a finalized list of journal titles for cancellation on the Libraries' website for further feedback in the spring semester of 2019.

The campus-wide collection cancellation project makes us rethink current practices for collection management, especially the pruning practice which is primarily based on the cost-per-use model. According to Kendrick, the cost-per-use model fails to account for variability of the usage pattern, consequently overvaluing journal subscriptions [1]. Beyond the limitation of the cost-per-use model, there are four main challenges around collection management. First, we engineering librarians have many responsibilities besides collection management including instruction, reference, public services and outreach. Second, we spend increasing expenditures on “big deal” journal packages to which our libraries subscribe and have less flexibility with individual journal subscriptions as major publishers are continuously acquiring small publishers. Third, we need to constantly adjust the collections for new programs and emerging research areas at the College of Engineering. Fourth, we also need to annually show the evidence of the need of our users in order to secure the collection fund for engineering. Taking the internal and external challenges into consideration, we see the necessity of monitoring and evaluating collections in a proactive and strategic manner.

Since the University Libraries recently purchased a journal usage statistics report prepared by 1Science, a company owned by Elsevier providing subscription analysis for institutions, it would be a great opportunity for us to make good use of the data from the report and initiate a case study of measuring engineering journal usage. The purpose of this case study was to analyze and interpret the journal usage using K-Means clustering technique so that engineering librarians, especially those who were actively involved in collection management, would be able to monitor the journal usage on a regular basis.

Methods

We reused a small portion of the data from the report because 1Science aggregated the number of publications per journal published by the authors of our institution, the number of citations to journal made by our authors from Scopus and 1Science’s own indexing database, and the number of downloads loaded from publishers’ COUNTER data which were provided by the University of Iowa Libraries. Dating back to 1970s, Pan [2] suggested that citation counts should be considered as a reliable predictor for the potential use of journals in libraries. A recent study conducted by Pastva etc. [3] also demonstrated the importance of incorporating publication and citation data into journal usage analysis. Therefore, the journal usage would be measured by three indicators including the number of publications by the authors of our institution, the number of citations to journal made by our authors and the number of downloads by our users. Since the download data were only available for the most recent four years (from 2015 to 2018), we selected the same period of years for the other two key indicators.

After inspecting the raw dataset carefully, we would preprocess the raw dataset to obtain a clean dataset because the raw dataset had three problems including title duplication, missing values and right-skewed distribution (Figure 1). The first step of the data preprocessing was to deduplicate journal titles. The second step was to take a mean of available years’ downloads for a journal and have it to replace the missing values of that journal because missing values only occurred in the number of downloads (84 titles had missing values for one year; 93 titles had missing values for two years; 168 titles had missing values for three years; and only 3 titles had missing values for four years). To solve the problem of skewed and wide distributions, we performed the log-transformation and then the 0-1 rescaling on the data. Histograms in Figure 2 and boxplots in Figure 3, 4 and 5 showed that the log-transformation greatly reduced the problem of skewed distributions. A complete procedure of the data preprocessing in Python could be found in Appendix 1.

Figure 1. Distributions for Average Number of Publications, Citations and Downloads for Four Years

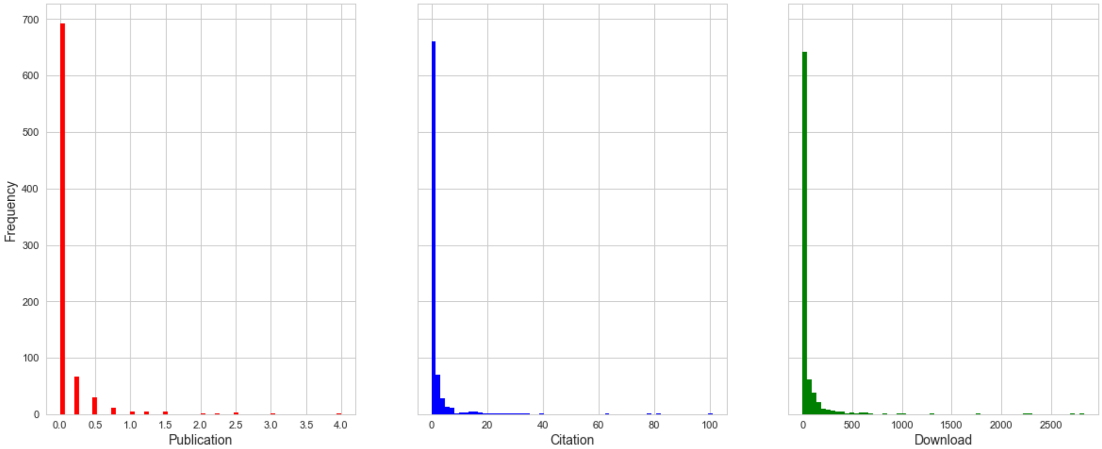


Figure 2. Distributions for Publications, Citations and Downloads After Log-Transformation

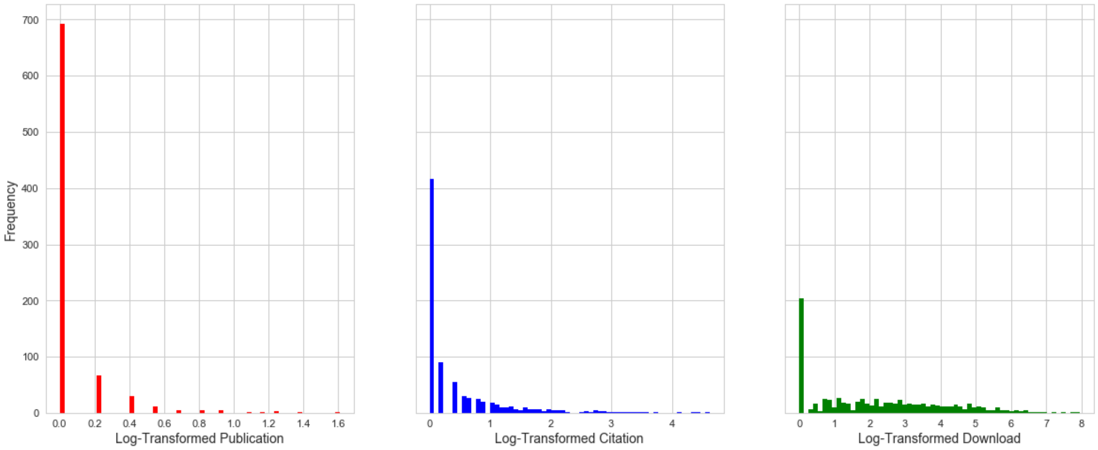


Figure 3. Distributions for Publications within Subfields After Log-Transformation

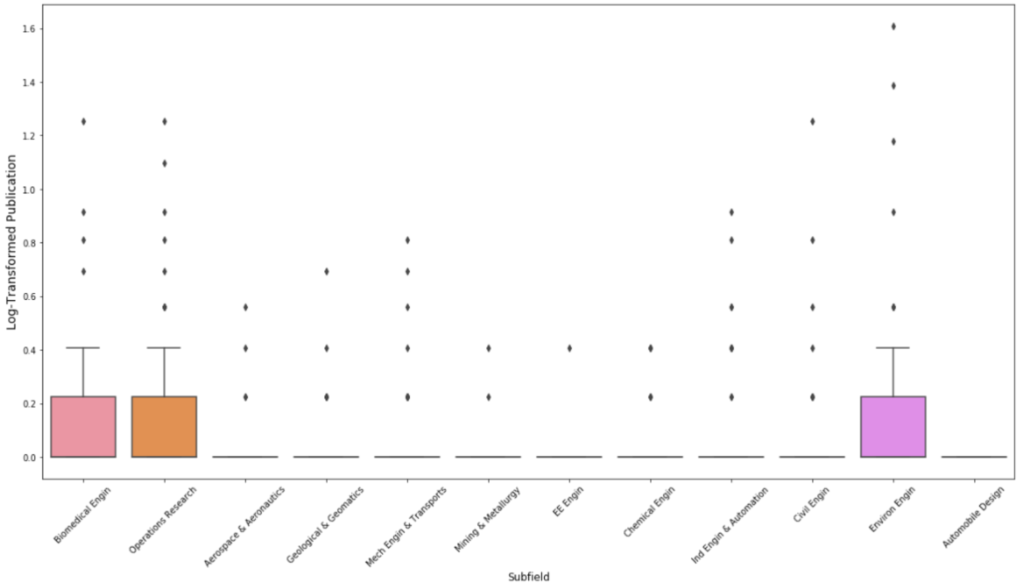


Figure 4. Distributions for Citations within Subfields After Log-Transformation

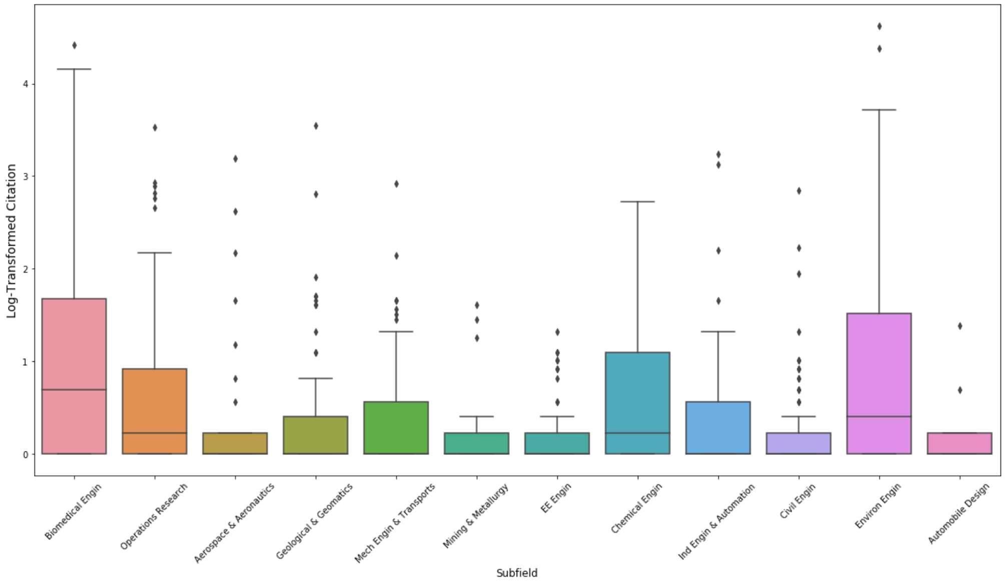
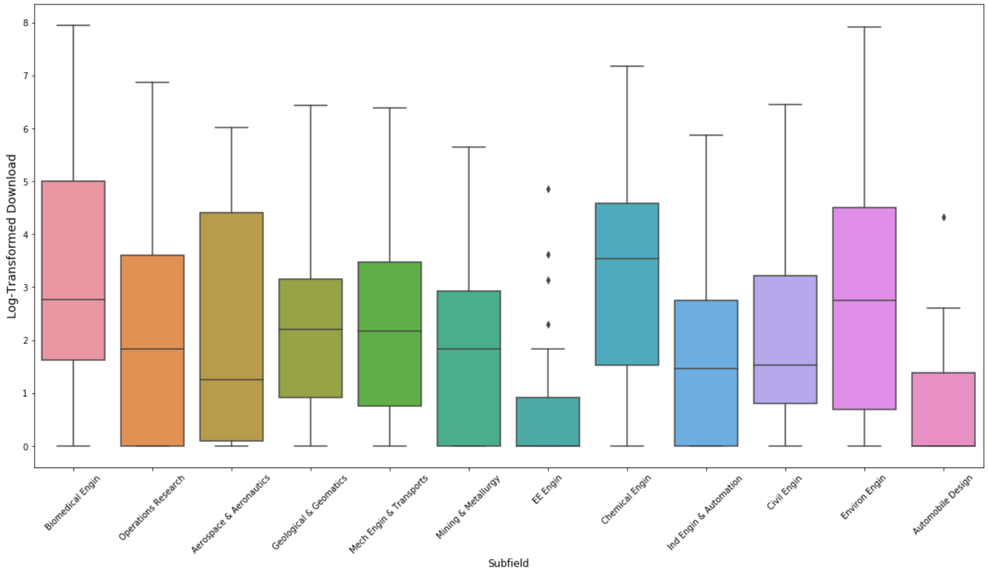
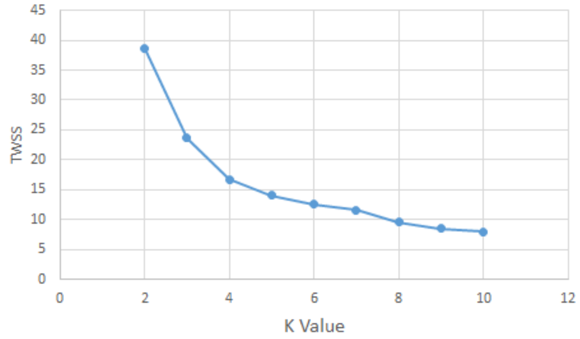


Figure 5. Distributions for Downloads within Subfields After Log-Transformation



In the modeling phase, we used an open source machine learning software called Weka to perform K-Means clustering on the clean dataset. K-Means clustering, which is a popular unsupervised machine learning algorithm, can identify k numbers of centroids and then allocate every data point to the nearest cluster. A cluster refers to a collection of data points aggregated together because of certain similarities [4]. In other words, K-Means clustering would segment journal titles with different numbers of publications, citations and downloads into different clusters which would share certain similar patterns in the number of publications, citations and downloads. Since the most well-known method for determining the optimal number of clusters (k), called the “elbow” method, suggested choosing the k for which the total within sum of square (TWSS) value stopped dropping quickly, we varied k from 2 to 10 and recorded the TWSS value for a range of values of k to obtain a plot for TWSS-versus-k shown in Figure 6. Based on the “elbow” method, we chose k of 4 as the optimal number of clusters.

Figure 6. TWSS versus k Value



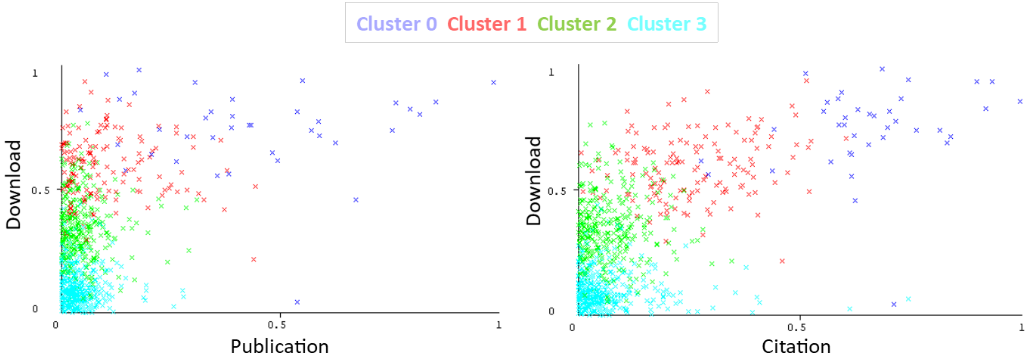
Results and Analysis

K-Means clustering segmented a total of 821 titles into four clusters. Table 1 showed the centroids of each cluster and statistics on the number of journal titles assigned to different clusters. Weka also visualized the four clusters in Figure 7 which revealed some interesting patterns of journal usage. Since the indexing in Weka started from 0 rather than 1, the first cluster would be Cluster 0, the second cluster would be Cluster 1 and so on. Cluster 0 represented journal titles with good numbers of publications, citations and downloads. Cluster 1 represented journal titles with a poor number of publications but a fair number of citations and a good number of downloads. Cluster 2 represented journal titles with poor numbers of publications and citations but a fair number of downloads. Cluster 3 represented journals titles with poor numbers of publications, citations and downloads. Journals in Cluster 0 and Cluster 1 accounting for 22% of the total number of journals were greatly used by our users. In contrast to journals in Cluster 0 and Cluster 1, journals in Cluster 2 were not heavily used but still might be of interest to our users. Although we labeled journals in Cluster 3 with relatively low usage, we would still need a further review on these journals to verify whether the low usage was true or was associated with missing data points.

Table 1. Cluster Centroids

Indicator	Full Data (821 Titles)	Cluster 0 (38 Titles, 5%)	Cluster 1 (142 Titles, 17%)	Cluster 2 (289 Titles, 35%)	Cluster 3 (352 Titles, 43%)
Publication	0.0401	0.4296	0.086	0.0091	0.0049
Citation	0.1078	0.6529	0.2668	0.0433	0.0378
Download	0.277	0.7566	0.5819	0.3406	0.0501

Figure 7. Scatter Plots of Clusters in Weka



Since it was worth knowing how the journal usage differed across subfields, we created two bar graphs in Figure 8 and 9 to show how the number of journal titles occurring in the four clusters varied across subfields. In Figure 9, Cluster 0 suggested that journals in biomedical engineering, environmental engineering and operations research were highly used. Followed by the three subfields appearing in Cluster 1, journals in chemical engineering, mechanical engineering & transports and industrial engineering & automation were moderately used. However, we noticed that most of titles in automobile design and electrical & electronic engineering (which was EE engin in the graph) did not appear in Cluster 0 and Cluster 1 with high to moderate usage. This observation seemed to be aligned with our previous question, namely whether journals with low usage was true or associated with missing data points. Later, we found two clues that might explain why most of these titles in the two subfields did not appear in the clusters with high or moderate usage as what we expected. One clue was that journals in the two subfields had a lower median than other subfields according to the boxplot of the distribution for downloads in Figure 5. The other clue was that some journals, especially IEEE journals, had missing or zero downloads in the original report.

Figure 8. Journal Titles of Subfields Assigned to Four Clusters (Group by Subfield)

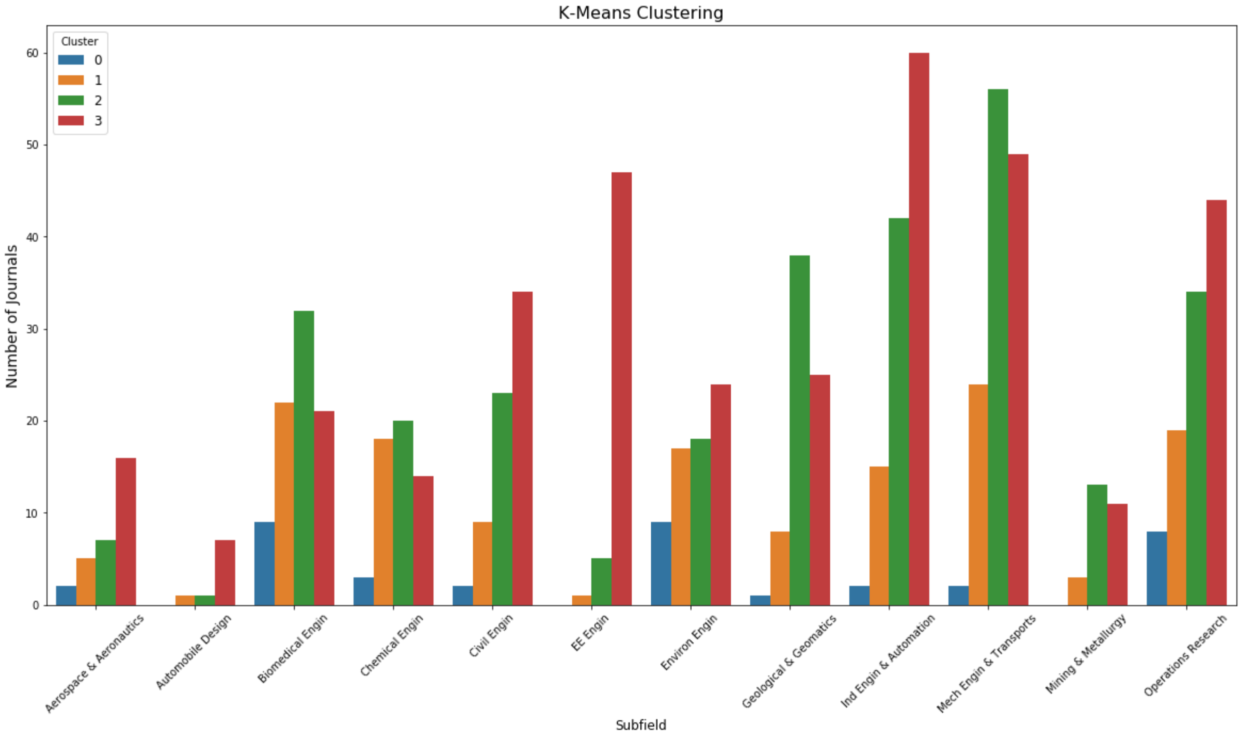
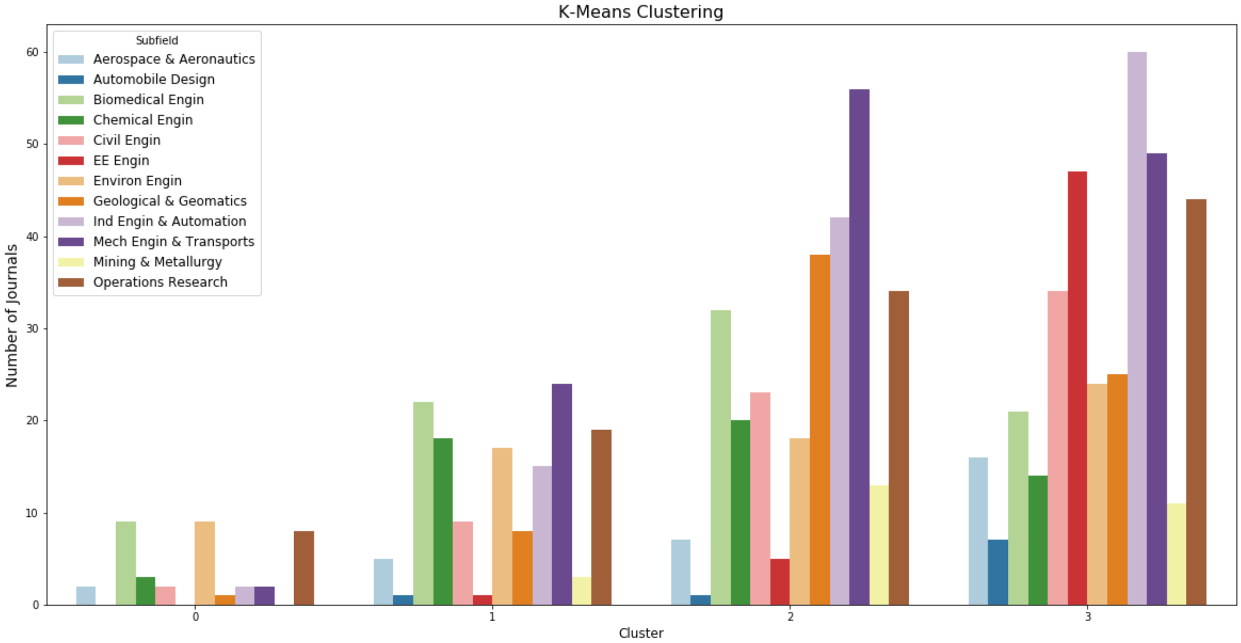


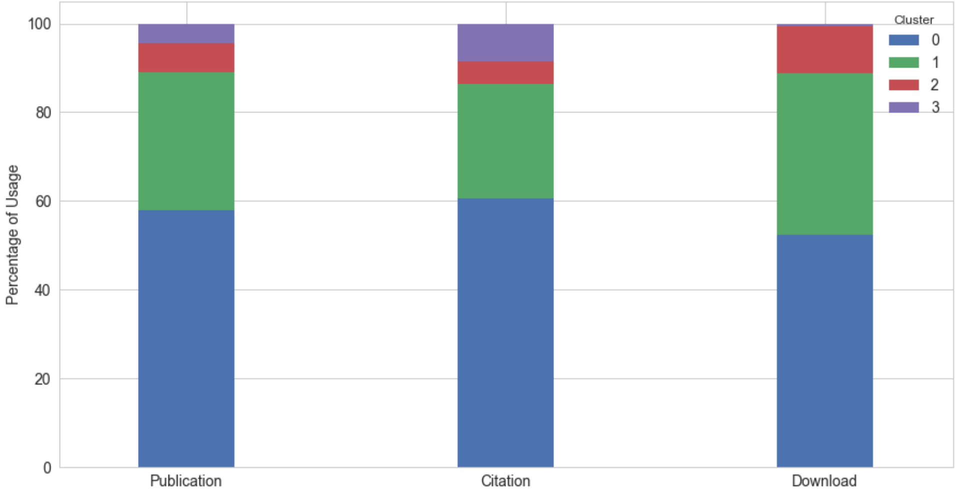
Figure 9. Journal Titles of Subfields Assigned to Four Clusters (Group by Cluster)



Discussion

K-Means clustering helped us identify the most frequently used journals and relatively less frequently used journals with respect to the publications, citations and downloads. When we assigned the cluster labels to the journal titles and computed the percentage of usage respectively for publications, citations and downloads, we found a few interesting observations in Figure 10. One observation was that the 38 most frequently used journals in Cluster 0 accounted for 58% of the publications, 60% of the citations to journals and 52% of the downloads; the 142 second most frequently used journals in Cluster 1 accounted for 30% of the publications, 25% of the citations and 38% of the downloads. The other observation was that about 22% of the journals accounted for 88% of the publications, 85% of the citations and 90% of downloads if we summed up the number of journals in Cluster 0 and Cluster 1. This observation seemed to be aligned with the well-known "80/20 rule" stating that about 20% of the collection account for about 80% of the circulations or use in libraries [5]. So, we suggested defining the list of 38 journals in Cluster 0 as the core journal list that we must include in our collection (Appendix 2), the list of 142 journals in Cluster 1 as the important list that we should include in the collection, and the list of the journals in Cluster 2 and Cluster 3 as a potential list for revisit if we would encounter a budget cut in the future. The results reflected variability of the usage pattern while the cost-per-use model failed to do so, suggesting additional criteria for the current pruning practice.

Figure 10. Percentage of Journal Usage Respectively for Publications, Citations and Downloads



Next, K-Means clustering had two advantages in journal usage analysis compared to a synthesis method used for the original journal usage report. K-Means clustering was easy to apply because it would simply find journals with a similar pattern of how a journal had been used with respect to publications, citations and downloads and then put them together as clusters. However, the synthesis method would construct a complex weighting scheme to add a download-publication ratio in front of the number of publications and a download-citation ratio in front of the number of citations respectively. The other advantage was that K-Means clustering would avoid favoring one indictor over the other. On the contrary, the synthesis method might inappropriately either undervalue or overvalue a journal because not only the synthetic usage was highly vulnerable to the number of downloads but also the download-citation ratio differed substantially among disciplines [6].

However, we observed three challenges in applying K-Means. The quality of data was the first challenge because library usage statistics were unfortunately not tidy. As we mentioned above, the dataset we obtained in this case study had three major problems including title duplicates, missing data points in the downloads and right-skewed distributions. A common way of handling missing data points is to replace the missing data points with the mean or median of the feature which they occur (a feature is the same as an indicator described in this case study). But simply adopting this method would not be appropriate for our dataset because the data distribution for each indicator was right skewed rather than normal. Given the data points for four years, we could use the available data points of a few years to estimate the missing data points of a particular year for a journal. For example, if we had the number of downloads in 2015 missing, we could replace it with a mean of the downloads occurring in 2016, 2017 and 2018. For the problem of a right-skewed distribution, we performed log transformation on the dataset to correct the distribution and improved the K-Means clustering results to some extent. As an evidence of improvement, the data points (shown as dots) of each scatter plot in Figure 7 were much widely spread. In contrast, if we did not perform log transformation, the data points would be concentrated within a smaller area in the lower-left corner of each plot.

The next challenge was how to make K-Means clustering results interpretable. Although K-Means clustering did a good job of grouping journal titles with similar usage patterns in this case study, domain knowledge would still be needed in interpreting the results and even justifying the performance of K-Means clustering. Based on the cluster centroids in Table 1, we interpreted that journals in Cluster 0 and Cluster 1 were defined as more frequently used journals than those in Cluster 2 and Cluster 3. When comparing the clusters with the original journal usage report, the journals with the actual number of publications, citations and downloads aligned with our interpretation. The other example of the importance of domain knowledge in the interpretation of K-Means clustering results was that we demystified why many journals in automobile design and electrical & electronic engineering, especially IEEE journals, did not appear in Cluster 0 and Cluster 1 but fell into Cluster 3. As we mentioned previously, this occurrence was caused by missing or zero downloads in the original journal usage report while we had full subscriptions to the IEEE Xplore Digital Library.

In addition, a challenge for K-Means itself was choosing the initial centroids because it could affect the clustering results. To address this problem, we ran K-Means multiple times with different initial centroids and evaluated the results. To be specific, we changed the seed value of a random number generator from 1 throughout 10 (default) in Weka because the random number was used to initialize centroids [7]. Since we received consistent clustering results at different seed values, we were confident that the clustering results were locally optimal.

Conclusion

The case study on engineering journal usage analysis succeed in converting massive journal usage statistics into the four clusters of journal titles in a straightforward format. The four clusters of journal titles not only provided us with a comprehensive view of how engineering journals had been used by both authors and users of our institution in the most recent four years but also might be used as additional criteria for the current pruning practice in the library. In addition, the frequently used journals seemed to reflect the “80/20 rule” for collection management. Last but not the least, this case study showed a possibility of implementing data analytics in academic libraries.

Acknowledgements

The authors thank ELD reviewers for their invaluable feedback on the manuscript, thank 1Science for the permission of reusing the data of the journal usage report, thank Professor Hong Cui and Professor Nick Street for their constructive feedback on the project/manuscript, and Data Analyst Sherry Xie for her input in the early stage of the project.

References

1. Kendrick, C., Guest Post: Cost per Use Overvalues Journal Subscriptions. Scholarly Kitchen, 2019. Retrieved from https://scholarlykitchen.sspnet.org/2019/09/05/guest-post-cost-per-use-overvalues-journal-subscriptions/
2. Pan, E., Journal citation as a predictor of journal usage in libraries. Collection Management, 1978. 2(1).
3. Pastva, J., et al., Capturing and Analyzing Publication, Citation, and Usage Data for Contextual Collection Development. Serials Librarian, 2018. 74(1-4): p. 102-110. DOI: 10.1080/0361526X.2018.1427996
4. Garbade, M. J., Understanding K-means Clustering in Machine Learning. Towards Data Science. Retrieved from https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1
5. Nisonger, T.E., The "80/20 rule" and core journals. The Serials Librarian Serials Librarian, 2008. 55(1-2): p. 62-84.
6. Wood-Doughty, A., T. Bergstrom, and D.G. Steigerwald, Do Download Reports Reliably Measure Journal Usage? Trusting the Fox to Count Your Hens? 2019, 2019. 80(5). DOI: 10.5860/crl.80.5.694
7. Mobasher, B., K-Means Clustering in WEKA. Retrieved from http://facweb.cs.depaul.edu/mobasher/classes/ect584/WEKA/k-means.html

Appendix 1. Data Preprocessing Procedure

1: Extract 1371 engineering journal titles from the journal usage report through selecting Applied Sciences and Engineering in the subject areas.

2: Receive a total of 821 journal titles after duplicates removed. 460 titles occur twice, 41 titles occur three times and 2 titles occur five times.

3: Handle missing data points which only occur in the number of downloads. Replace missing data points with a mean of the number of downloads across four years for each journal.

4: Calculate a mean of each indicator across four years for a journal.

5: Perform log transformation on the data points for each indicator.

6: Perform the 0-1 rescaling on the data points for each indicator.

Appendix 2. The Most Frequently Used Journal Titles

Journal	Subfield
Acta Biomaterialia	Biomedical Engineering
Advances in Space Research	Aerospace & Aeronautics
Advances in Water Resources	Environmental Engineering
AIAA Journal	Aerospace & Aeronautics
Annals of Biomedical Engineering	Biomedical Engineering
Applied Ocean Research	Civil Engineering
Automatica	Industrial Engineering & Automation
Biomaterials	Biomedical Engineering
Biomechanics and Modeling in Mechanobiology	Biomedical Engineering
Chemical Engineering Journal	Chemical Engineering
Computers & Industrial Engineering	Operations Research
Computers & Operations Research	Operations Research
Ecological Engineering	Environmental Engineering
Environmental Modelling & Software	Environmental Engineering
European Journal of Operational Research	Operations Research

Hydrological Processes	Environmental Engineering
Hydrological Sciences Journal / Journal des Sciences, Hydrologiques	Environmental Engineering
IEEE Transactions on Automatic Control	Industrial Engineering & Automation
Industrial & Engineering Chemistry Research	Chemical Engineering
International Journal of Heat and Mass Transfer	Mechanical Engineering & Transports
International Journal of Multiphase Flow	Mechanical Engineering & Transports
International Journal of Production Economics	Operations Research
International Journal of Production Research	Operations Research
JAWRA Journal of the American Water Resources Association	Environmental Engineering
Journal of Biomechanics	Biomedical Engineering
Journal of Biomedical Materials Research, Part A	Biomedical Engineering
Journal of Biomedical Materials Research, Part B: Applied Biomaterials	Biomedical Engineering
Journal of Hydrology	Environmental Engineering
Journal of Membrane Science	Chemical Engineering
Journal of the Mechanical Behavior of Biomedical Materials	Biomedical Engineering
Management Science	Operations Research
Mathematical Programming	Operations Research
Ocean Engineering	Civil Engineering
Remote Sensing of Environment	Geological & Geomatics Engineering
SIAM Journal on Optimization	Operations Research
Tissue Engineering, Part A	Biomedical Engineering
Water Research	Environmental Engineering
Water Resources Research	Environmental Engineering