Using Text Data Mining to Enhance the Literature Search Process for Novice STEM Researchers

Andres Fortino, Qitong Zhong, Luke Yeh and Sijia Fang
agf249@nyu.edu, qz676@nyu.edu, cmy286@nyu.edu and scarlett.fang@nyu.edu

Abstract - **A literature search can be an arduous process, especially for novice researchers. We have developed a tool that allows a researcher to rank order a list of references that are returned by a keyword-based search engine, based on similarity to known exemplars. This significantly accelerates literature searches by novices. Our research question was: can we produce a text-analytic tool that, when used by an inexperienced scholar, rank-orders a list of references against an exemplar, so that the time needed to find relevant literature is reduced, and the literature survey section of their paper will be superior. An experiment was set up where one course section used the tool to produce the literature review section of a thesis proposal, and the other class used traditional literature research tools. We surveyed both sections to self-report the time used for the literature search. We found some time savings by some of the students using the tool. We also provided blind, randomly selected pairs of completed proposals to SME faculty who teach that same class to assess the quality of the literature sections of the samples. We found that the tool-using section of students reported significantly less time to do the literature search, and the quality of their literature review produced had a significantly higher quality.**

Index Terms - Literature search, keywords, novice researcher, STEM, text-data mining, literature review

INTRODUCTION

Review of the pertinent literature is an essential task in research. It can be an arduous task for any researcher who is investigating an unfamiliar field. It is particularly difficult for novice university students who need to conduct research for research papers, thesis, or dissertations for the first time, especially in areas in which they are not well-versed. The search process for relevant literature can seem a daunting task. The tools commonly used in the search for relevant literature are provided by traditional repositories of reference materials: university libraries. More recently, online search engines have been harnessed for that purpose (e.g., Google Scholar.) Generally, given a set of keywords, these tools retrieve a list of relevant scholarly works and leave it up to the researcher to choose from the list which are the most relevant ones to their purpose. The subsequent process of reviewing and selecting relevant references from the list is a tedious and time-consuming task. It can be potentially very frustrating if many of the references returned by the search engine do not appear very relevant. For a novice researcher, it may even be bewildering and frustrating.

The results of studies of research processes [1] [2] [3] show that search experience affected searchers' use of many search tactics, and they suggest that subject knowledge became a factor only after searchers have had a certain amount of search and subject matter experience. The RightReference tool presented in this paper builds on this premise by improving the search tactics of novices.

Our goal was to develop a smart-search system that ranks orders research references, previously found using keywords with library and online search engines, and score them against a group of relevant references. One approach to finding research-relevant references is to start by using a search engine such as that found through most university library systems or public ones such as Google Scholar and do a first pass search with keywords. Then the researcher reads and judges which of the returned items is most appropriate to the research question. The researcher might go through the first ten to twenty such references collecting those that appear adequate to their search. Of those that appear useful, they might even ask the search engine to give us similar articles, and then they branch out in various directions looking for additional appropriate references. What we asked was: is there a machine, a text-analytic tool, that imitates this activity and shortens the process [4]?

THE SOLUTION

We developed a literature search enhancing tool, RightReference, that scores the references returned by library or other online search engines for relevancy to the research problem being studied. The researcher enters a verbose statement of the research problem into a search engine, such as Google Scholar, or use the traditional keyword search in a library search engine, and collect a list of references. These are returned ranked in some order of relevancy as dictated by the search engine algorithm. The search engine algorithm is not easy to modify or to even know what methodology is used to rank the references.

The researcher then scans the ranked list and identifies a minimal set of exemplars (3-4) and their associated abstracts that are most applicable to the research problem. The tool is then directed to use these identified exemplar abstracts to

978-1-7281-7520-1/20/$31.00 ©2020 IEEE

score all other papers in the returned list using a text data mining similarity scoring algorithm. Using the similarity score the list of returned references is rank ordered into those more likely to be useful and investigated further, and those less likely to be useful and should be ignored. The researcher now has a quickly sorted list based on his judgment of fit to the search to use for the literature review.

LITERATURE REVIEW

Researchers that are pursuing the automation of the research process [5] break it down into five identifiable phases, each to be automated via different approaches. The first task involved the exploration phase, including developing research questions, conducting a literature review, and developing theory. Johnson and colleagues [5] see the automation of literature searches to be of great benefit and as a first fruitful step. The automation of a literature search could reduce the time spent on tasks by researchers and could save time for the scholar to focus on steps that are harder to automate and more essential, such as theory building. It could also ensure that literature reviews are more comprehensive than those developed by human researchers. The authors [5] urge software developers to prioritize automating this task, which could reduce the cost of conducting research, and that they should work to make these technologies accessible and user-friendly to researchers in all disciplines.

Searching within the full text of documents has been a standard process on the Web. With the advent of text data mining, much progress has been made on how to the search well. Full-text search of science articles is often offered on a small subset of articles by publishing groups (e.g., Nature, Science, Highwire, Science Direct). Recently Google Scholar has begun offering search over the full text of journal articles, but with no special consideration for the needs of scientists [6]. One approach undertaken by The National Institute of Health with a tool named SAPHIRE uses similarity scoring of a user's search phrase to the frequency of similar individual phrases within the search targets documents to categorize and rank the targets [7]. Our approach is similar, except we undertake the simpler task of whole text similarity scoring as the basis for analysis. It returns a less refined search and allows the researcher to provide the additional step of assessing appropriateness to the search string and provides quicker results.

In this article, we present RightReference, a Python-based application that allows novice researchers to search over abstracts of the results of a Google Scholar search, ranking, and the references by similarity to selected exemplars. This idea is based on the observation, noted by our own group of researchers, as well as many others [8] [2] [3], that when reading research literature, researchers tend to start by looking at the title, abstract, date, source and citation index for relevancy and recency. They sort the references into highly relevant and those that can safely be ignored.

The RightReference tool addresses some of the major problems encountered by novice researchers [9]: (1) overall process was extremely time-consuming; and (2) difficult to identify precisely the conditions under which a study would be of importance for the search. In a study of how to teach novices success in conducting literature searches, Lavallee [10] found that eight distinct steps are needed to be taught to novices. The RightReference tool supports expedited execution of Lavellee's fourth and fifth steps: (4) the selection process: to define inclusion and exclusion criteria (5) evaluating the strength of the evidence: to define what makes a high-quality paper. RightReference not only returns a sorted reference list by importance to the search subject but also evidence of reference quality (citation number weighted by similarity score as a quality proxy).

Another innovative approach [11] purports to maximize both the number of documents retrieved and the ratio of relevant to non-relevant documents (signal to noise ratio) during a literature search. The researcher in this case uses a test query to retrieve a relative sample of documents from a database, classifying the retrieved documents as relevant or not relevant, finding text element (phrase) frequencies and text element co-occurrences in at least the relevant documents, grouping the extracted text elements into thematic categories, and then using the thematic grouping and phrase frequency data to develop new queries and query terms.

Most semantic processing of web search results is intended for information extraction. The unstructured retrieved text is processed by some data mining algorithm to extract features that are used for classification of the document for further processing. Our contention is that a simple classification based on TF-IDF similarity scores is sufficient to improve the literature search process for the novice user. A novice researcher who needs to complete a term paper or even a master's thesis needs a basic assist in the search process when using a search engine.

RESEARCH QUESTION

Can a software search tool be built that improves the literature search process for novice researchers that (a) reduces the time needed to find relevant literature; and (b) produces high-quality literature search results leading to improved literature reviews in research activities?

Hypothesis

Relating to the primary objective of this research to test the efficacy of a literature search tool that improves the process of finding relevant literature for a literature review for research purposes by a novice two hypothesis were proposed:

Hypothesis 1: Using a text data mining literature search tool that ranks orders returned searches by search engines

reduces the time a novice researcher uses to create a literature review section of a research project.

Hypothesis 2: Using such a literature search tool significantly improves the quality of the literature review section of a research project created by a novice researcher as judged by subject matter experts.

METHODS

The Tool

A software tool was developed in Python based on earlier work [12] on the use of text data mining for document comparisons. We had a number of text data mining techniques available to us, and as in the earlier work, we used Term Frequency–Inverse Document Frequency (TF–IDF). TF-IDF is a process that converts documents into a numeric matrix. The idea of TF-IDF scoring is that when a word appears more often in a document than in others, this word represents more important information to this document. The way TF-IDF converts documents into a matrix involves two calculations. TF refers to term frequency where IDF refers to Inverse Document Frequency.

TF is defined as the number of times a term appears in a document, divided by the total number of terms in the document. Therefore, the greater the TF means that the more important the word is in that document. On the other hand, IDF measures how common a term is among all documents as the formula shown below, where *n* is the number of documents in the corpus, and DF means how many documents the word "I" appears at least once [13].

After transforming the exemplar paper abstract and abstracts of the other returned references as text data into a latent space of lower dimensionality, the next step is to determine the similarity between the exemplar and each reference in the list. Each reference abstract is converted to a query vector in the same two-dimensional semantic space that we chose to perform the SVD for the abstract text data matrix [14]. Then the cosine similarity is computed to measure the distance between a given query (the exemplar text) and the reference's abstract vector [15].

An extracurricular Python coding contest was offered to students in the program to develop the tool (also as a way to develop their Python coding skills). The students were challenged to scrape the data from the results of the search engine set of references returned from a query. They then created and provided to the researcher a database of references to choose the exemplar(s). Their algorithm, implemented in Python, further ranked the references returned by a search engine using TD-IDF. The coding contest submissions were judged by a team of senior students, program alumni, and faculty. The winning entry became the search tool used in this experiment. The tool performs the following tasks:

1. Ingest verbose research problem statement or a set of keywords as text data.
2. Run a Google Scholar search on the problem statement and keywords.
3. Ingest the metadata of the first 500 papers and enter them into a dataset, including the abstracts as a separate field, and the Google search rank order.
4. Ingest the resulting dataset of paper abstracts parse it into a TF-IDF representation.
5. Present to the researcher a subset of documents returned by the search engine and create a corpus (from the first 500 papers abstracts) and allow a researcher to select 1- 5 documents and label them the exemplars. The subset would represent a group of references with a common strategic intent: the literature supporting a research problem.
6. Run a classification algorithm and implementation program that estimates whether a reference matches a set of exemplar references (TF-IDF for this experiment).
7. Return a list of the top 50 references (from the list of 500 and not including the exemplars) rank-ordered by similarity score to the exemplars. The list should be a complete reference in APA format, including the text of abstract material for each reference.

The Experiment

For the purposes of this experiment, the population consisted of university graduate students who are novice researchers tasked for the first time in their graduate studies to conduct a literature review for a proposal for original research. Their research efforts require the creation of new knowledge. Our intervention was to train a group of students (tool-using course section) in the use of the search tool and require its use, while a similar group as control (non-tool course section) did not use the tool. The two sections were ignorant of each other's activities.

For the preliminary experiment, our sample were students in the two sections of a graduate course in Research Process and Methods. They attended the course in the same semester. Each section had 15 students. We controlled for possible discrepancies in how they were taught and subject matter coverage by using the same syllabus and the same faculty for both sections. Both sections were conducted as 14-week, one meeting per week, graduate classes. Students in both sections were tested for prerequisite subject matter knowledge using a RAIKS assessment (Rapid Assessment of Individual Knowledge and Skills) [16]. No significant differences in their basic prerequisite knowledge were found.

A Need for Cognition assessment [17] was also administered to both groups as a baseline comparison of the two sections, and was used as an indicator of their desire to do research. This level-set their capacity to undertake a research activity. We used the shorter six-item NCG survey

[18]. No significant differences were found between the intervention and control groups (tool-using section score n=15, M=22.1, SD= 3.1, non-tool section score n=15, M= 22.6, SD=3.5 , p =.71, α=.05).

The tool-using group was taught how to install and use the search tool as part of the class presentations on how to conduct a literature search. Both groups were given exactly the same assignment dealing with literature reviews and the same amount of time to complete the assignment. All students turned in acceptable (B or better and on-time) deliverables, which eventually were folded into the final course deliverable, which was a proposal for original research. Both sections were taught by the same experienced faculty member who had taught the class previously.

A survey of both sections was conducted after the literature review assignment was completed to determine how long it took the students to complete the assignment. Both sections were asked what their impression of their experience was in creating the literature review, and the length of time it took them to complete the assignment. The tool-using section was asked additional questions on their experience in the use of the tool

We measured the quality of the literature review by engaging faculty subject matter experts to judge the results. The final deliverable, the research proposal, is intended as the basis for the program's final capstone class, the research thesis. The thirty research proposals from both sections were assigned random numbers and anonymized by replacing student's names with their assigned random numbers. Fifteen faculty from the program were engaged to review a pair of papers, randomly selected, one from the tool-using section, and one for the control section. These faculty all teach this particular course as well as other capstone courses based on this proposal. A thorough literature survey a requirement in the research proposal, as well as in the final capstone work. These faculty are subject matter experts in evaluating research proposals and literature reviews. They were tasked with returning which of the two research proposals, in their opinion, had a better literature review. To form that opinion, they were asked to base it on which of the two papers had: (a) most appropriate references to the research question; (b) a better summary of the reference's appropriateness to the research question and student's theory of the case; (c) more appropriate to the background of the proposed research question. We engaged 15 such faculty SMEs and had 100% participation in the experiment.

RESULTS

In our preliminary measurements, we found no statistically significant difference (tool-using section n=15, M=5.35 hours, SD=.63, non-tool section n=15, M=5.43 hours, SD=.63, p=.468, α=.05) in the time it took the tool and the control sections to perform the literature review. Therefore, we had to accept the null hypothesis. One of the reasons for this negative result is that students reported that the tool took a long time to install (in some cases, up to half the time needed to create the literature review). This may account for the finding of no significant difference. In half the cases, tool using students reported finding the tool to be very useful in narrowing down their search and making it easier to find appropriate references. Figure 1 shows a definite skew to the shorter time frames for the tool-using section. Tool-using students who took an excessive time (more than eight hours) to complete the assignment also reported difficulties with installing the tool and the awkward tool interface.

Descriptive Statistics of the References

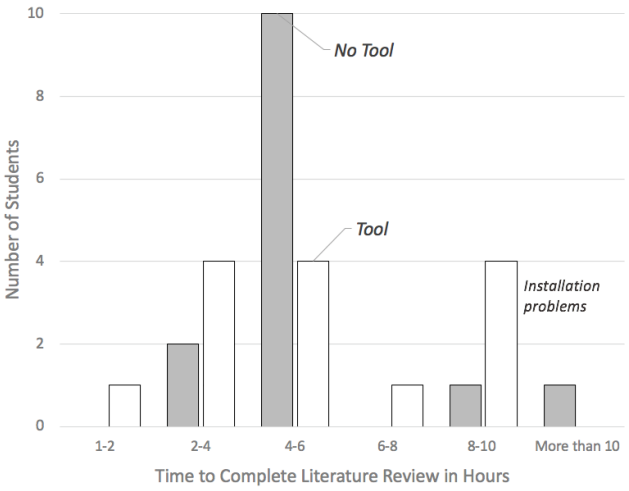

FIGURE 1
REDUCTION IN TIME TO COMPLETE THE LITERATURE REVIEW FOR TOOL-USING STUDENTS VS. NO TOOL USE, SHOWING A REDUCTION IN TIME TO COMPLETE AFTER ACCOUNTING FOR INSTALLATION PROBLEMS.

The characteristics of the 30 research proposals were analyzed to see if the tool made a significant difference in the number as well as the quality of the references collected. Although the average number of references by use or non-use of tool was quite different (tool-using proposals had an average of 23 references vs. non-tool which had 20 references, an 8% difference, see Figure 2A), the difference is not statistically significant (n=15, tool-using M=22.8, SD=2.5, non-tool n=15, M=19.4, SD=2.6, p=.36, α=.05). Using the percent of peer-reviewed references was used as the metric of reference quality, an increase in quality was also observed for the tool-using versus non-tool proposals. For the proposals where the tool was used, 68% of the references were peer-reviewed, where the proposals where the tool was not used 54% references were peer-reviewed, a 14% difference, see Figure 2B. Again, a large difference but not statistically significant (n=15, tool-using M=.683, SD=.096, n=15, non-tool M=.57, SD=.053, p=.187, α=.05) probably due to the small sample size.

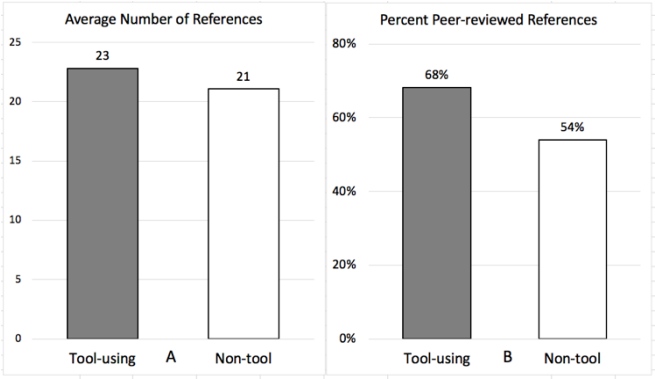

FIGURE 2

TOOL-USING VS. NON-TOOL SECTION RESEARCH PROPOSALS CHARACTERISTICS SHOWING: (A) AVERAGE NUMBER OF REFERENCES AND (B) PERCENT OF THOSE REFERENCES THAT WERE PEER-REVIEWED SHOWING IMPROVEMENTS WHEN USING THE SEARCH TOOL.IN TIME TO COMPLETE THE LITERATURE REVIEW FOR TOOL-USING STUDENTS VS.

Measures of Reference Quality

Fifteen faculty were engaged in reviewing the resulting research proposals and the corresponding literature surveys. The tool-using section literature student surveys received 25% more positive reviews than the non-tool section. Figure 3 shows the results for the 15 pairs of papers reviewed. There is a significant difference (χ^2=3.75, p=.001, α=.05) between the tool-using section papers and the non-tool section papers.

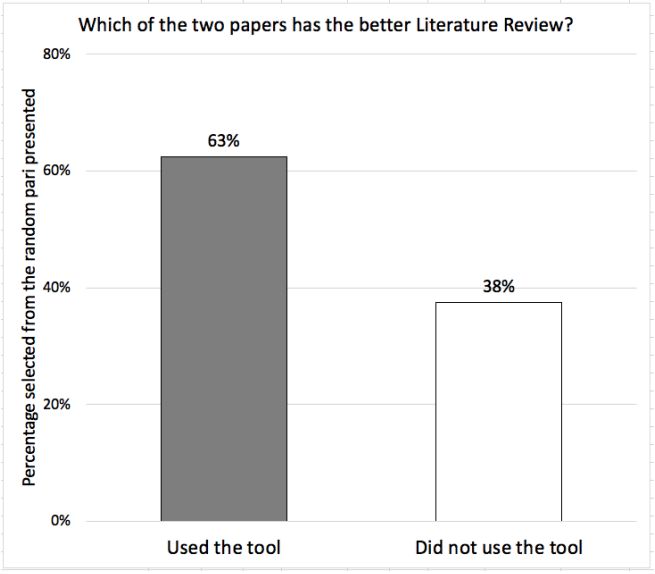

FIGURE 3

TOOL-USING VS. NON-TOOL SECTION REVIEWS BY SME FACULTY THAT, IN THEIR OPINION, SHOW QUALITY IMPROVEMENT WHEN USING THE SEARCH TOOL.

DISCUSSION

The results of this experiment are an imperfect indication of the usefulness of the tool. Clearly, the students struggled with the installation at first, which has caused us to engage developers to streamline the installation and improve the user interface. Unfortunately, the improved interface was not available until the end of the semester and could not be used in these experiments. In this second version, we added a more informative ranking, the citation index and an "importance" index composed of citation index multiplied by the similarity score. For a novice researcher, this becomes an imperfect indicator of the relative importance of the references.

We have shown that a simple text analytic tool may be built that improves the literature search process for novice researchers. It significantly reduces the time for a literature search, supporting alternative hypothesis 1. It also allows the novice researcher to produce a better literature survey for their research papers as attested by the SME faculty reviewers, supporting our alternative hypothesis 2, but the results do not allow us to reject the null hypothesis.

Additionally, 15% of the students in the tool section had positive mentions of the search tool in their end-of-course student evaluations. It is not conclusive evidence but, given how little information students share in these evaluations, it was deemed significant.

LIMITATIONS

The tool, as it now exists, needs a better interface. RightReference requires the researcher to install Python on their computer, then run a loading program to make sure all the proper Python add-on functions are loaded, and then run the code as a Python script. It is not a tool to be used by people unfamiliar with Python. Students with a facility for technology and eager to be familiar with a popular program such as Python embrace the opportunity. A better user interface will be needed to make it more widely available. That is the intention of future work. It requires rewriting the program in a different language that has better interface capabilities, especially one that allows the creation of executable code. Another alternative is to host the tool on a Haruko server and provide a browser interface. That would make the program more user-friendly, and it is being contemplated as an extension of this work.

CONCLUSIONS

We have shown that a simple text analytic tool may be built that improves the literature search process for novice researchers. It significantly reduces the time for a literature search. It also allows the novice researcher to produce a better literature survey for their research papers as attested

by the review of SME faculty reviewers. We will continue to improve this search tool and actively use it in our teaching.

REFERENCES

[1] Hsieh-Yee, I. (1993). Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *Journal of the American Society for Information science*, *44*(3), 161-174.

[2] Wildemuth, B. M. (2004). The effects of domain knowledge on search tactic formulation. *Journal of the American Society for Information Science and Technology*, *55*(3), 246-258.

[3] Vakkari, P., Pennanen, M., & Serola, S. (2003). Changes of search terms and tactics while writing a research proposal: A longitudinal case study. *Information processing & management*, *39*(3), 445-463.

[4] Beel, J., & Gipp, B. (2009, July). Google Scholar's ranking algorithm: an introductory overview. In *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)* (Vol. 1, pp. 230-241).

[5] Johnson, C. D., Bauer, B. C., & Niederman, F. (2017) The Automation of Management and Business Science. *Academy of Management Perspectives* (ja).

[6] Haddaway, N. R., Collins, A. M., Coughlin, D., & Kirk, S. (2015). The role of Google Scholar in evidence reviews and its applicability to grey literature searching. *PloS one*, *10*(9), e0138237.

[7] Hersh, W., Hickam, D. H., Haynes, R. B., & McKibbon, K. A. (1991). Evaluation of SAPHIRE: an automated approach to indexing and retrieving medical literature. In *Proceedings of the Annual Symposium on Computer Application in Medical Care* (p. 808). American Medical Informatics Association.

[8] Adeva, J. G., Atxa, J. P., Carrillo, M. U., & Zengotitabengoa, E. A. (2014). Automatic text classification to support systematic reviews in medicine. *Expert Systems with Applications*, *41*(4), 1498-1508.

[9] Riaz, M., Sulayman, M., Salleh, N., & Mendes, E. (2010, April). Experiences Conducting Systematic Reviews from Novices' Perspective. In *EASE*.

[10] Lavallee, M., Robillard, P. N., & Mirsalari, R. (2013). Performing systematic literature reviews with novices: An iterative approach. *IEEE Transactions on Education*, *57*(3), 175-181.

[11] Kostoff, Ronald N. "Method for data and text mining and literature-based discovery." U.S. Patent 6,886,010, issued April 26, 2005.

[12] Fortino, A., Zhong, Q., Huang, W., Lowrance, R. (2019). Application of Text Data Mining To STEM Curriculum Selection and Development, IEEE ISEC'19 Conference, Princeton University, NJ, March 2019.

[13] Kim, D., Seo, D., Cho, S., & Kang, P. (2019). Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. *Information Sciences*, *477*, 15-29.

[14] Karakatsanis, I., AlKhader, W., MacCrory, F., Alibasic, A., Omar, M. A., Aung, Z., & Woon, W. L. (2017). Data mining approach to monitoring the requirements of the job market: A case study. *Information Systems*, *65*, 1-6.

[15] Huang, A. (2008). Similarity measures for text document clustering. In *Proceedings of the sixth New Zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand* (pp. 49-56).

[16] Fortino, A. and Lowrance, R. (2019). Practice Makes Perfect: Memory Retrieval Strategies to Improve Student Academic Performance, Proceedings of the Academy of Management Annual Meeting, Boston, MA August 13, 2019.

[17] Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of personality assessment*, *48*(3), 306-307.

[18] de Holanda Coelho, G. L., Hanel, P. H., & Wolf, L. J. (2018). *The very efficient assessment of need for cognition: developing a six-item version*. *Assessment*, *1*, 16.

AUTHOR INFORMATION

Andres Fortino is a Senior Member of IEEE. He is also the Chief Learning Officer at Autonomous Professional Development and a Clinical Assistant Professor of Management and Systems at New York University School of Professional Studies and Academic Community of Practice Leader in the Masters in Management and Systems program. He received his PhD in Electrical Engineering at the City University of New York. He is a member of the Academy of Management and INFORMS. His main area of research is evidence-based education, business analytics and data visualization, text data mining and its applications to higher education.

Qitong Zhong is a Senior Data Analyst, working in Omnicom group, and is an alumnus of New York University (2018). She obtained her Master of Science Degree in Management and Systems, with the specialization in database technologies and business intelligence. Her research interest is E-commerce consumer behaviors. She obtained her Bachelor degree in Sun Yat-sen University, Guangzhou, China. She published her undergraduate graduation thesis, Factors Affecting Trust Formation in the Context of Social Commerce, in a Chinese journal, Information Science.

Luke Yeh is a Master of Science Candidate in Management and systems at New York University. He is a data analyst with two years of experience in product operations and CRM in the technology and public sectors. During his master's studies, he served internships at the United Nations and in the New York City Fire Department. His research interest is the application of emerging technologies. He obtained his Bachelor of Science Degree in Transportation Technology and Logistic Management at National Chiao Tung University.

Sijia Fang was born in Luoyang, China. She is currently a graduate student at New York University's School of Professional Studies majoring in Management and Systems. She received her B.S degree in Financial Management from Shanghai University of Finance and Economics, Shanghai, China, in 2015. After graduation, she became a full-time auditor at KPMG China for two years, where she gained professional skills and insight into the finance-related area, which is experienced in qualitative and quantitative analysis and forecasting. She is proficient in Python, SQL, HTML, Tableau, and different data analysis and visualization tools. She is aspiring to be a data analyst.