Paper 3343-2015

Improving SAS® Global Forum Papers

Vijay Singh, Pankush Kalgotra, Goutam Chakraborty, Oklahoma State University, OK, US

ABSTRACT

Just as research is built on existing research, the references section is an important part of a research paper. The purpose of this study is to find the differences between professionals and academicians with respect to the references section of a paper. Data is collected from SAS® Global Forum 2014 Proceedings. Two research hypotheses are supported by the data. First, the average number of references in papers by academicians is higher than those by professionals. Second, academicians follow standards for citing references more than professionals. Text mining is performed on the references to understand the actual content. This study suggests that authors of SAS Global Forum papers should include more references to increase the quality of the papers.

INTRODUCTION

Academician and practitioners both conduct research, however with different approaches. Academic research is mostly theory-driven and thus rigor. On the other hand, practitioners try to solve a particular problem and thus, focus more on relevance. Due to these differences, both communities use different approaches to conduct research. In this paper, we try to explore these differences from a different angle. The focus is on the way references are presented in the papers published in the SAS Global Forum (SAS GF) proceedings 2014 by both communities.

Papers are being published in SAS GF proceedings since 1976. These papers are presented at a meeting named SAS Global Forum that is organized by SAS community every year. More than two-third of participants are professional and hence, most of the publications in the proceedings are authored by professionals. This conference is very useful for professionals and academicians as they share ideas and try to narrow the gap between two communities.

Science cannot advance until and unless research findings are effectively and widely communicated. In our study, we try to observe the way in which academicians and professionals write a research paper. Academicians are University Professors and students whereas professionals are the people working in the industries. There are differences in the way Academicians and professionals practice the style of research writing. While academicians are more focused towards theory on the other hand professionals concentrate on the results.

Researchers have always emphasized the importance of rigor in documenting the literature search. Brocke (2009) believes that knowledge is created in the process of interpreting and combining existing knowledge which determines the quality of literature review. A comprehensively described literature helps the reader to understand the concepts presented in the paper. It helps future researchers to use past results in their own research thereby bolstering their findings. Webster and Watson (2002) emphasized on the methods of searching relevant articles using keywords and forward or backward searches which helps a researcher to find additional articles useful in their study. Keyword search gives us the previously published articles while forward search yields additional articles that have cited any article (Levy and Ellis 2006).

The focus of this study is on the references section. Many studies have tried to find the way to analyze the style of an article but analysis on the references section is rare in the literature. We try to prove two research hypothesis as stated in the next section from the data collected from SAS GF proceedings 2014 and also performed text mining for the exploratory purpose.

RESEARCH HYPOTHESES

As discussed earlier that there are differences in the research approach of authors with academic affiliation and professionals. In this section, we make two hypotheses about these differences.

Academicians follow theoretical approach in solving real world problems. This type of approach follows that new research is always built on the existing research. Hence, academicians are more likely to do adequate

literature search and include sufficient citations in their reports. In contrast, practitioner research revolves around a particular problem and the goal of the research is find a better solution to the problem. They are less likely to rely on the existing theories and do not perform an adequate literature search. Thus, number of citations in the papers are less. Based on this argument, we propose our first hypothesis that

> *H1: Average number of references by the academicians in their papers will be more than the average number of references by the professional in their papers.*

Authors with academic affiliations rely on the existing theories that are well established in the literature. As they use existing literature to motivate their research, they are more likely to use real references from the journals and conference publications. They also follow the standard style of presenting references such as APA, MLA, Chicago, etc. On the other hand, practitioners do not focus on the existing literature in the journals and conferences. To get the facts for their research, they may not use real references and get most of the information from the unauthorized websites. These differences lead to second hypothesis of this research that:

> *H2: Academicians follow the standard style of presenting references such as APA, MLA, Chicago, etc., but professional do not.*

DATA

Our dataset is a mix of Structured and Unstructured data. We collected the data from SAS online proceedings resource. Since we are trying to analyze the SAS proceedings, we tried to collect as much relevant information as we could. The list of proceedings on SAS website doesn't differentiate between the types of publications. So we collected all the types of publications which were Paper, E-poster, Quick tip or a Workshop. We extracted the name of the company with which the author is associated and used this information to create another variable. This derived variable is a binary variable having a value of 1 or 0 where 1 indicates that the author affiliation is academic while 0 shows that the author is from a company. We also collected information about the type of industry to which the paper belongs which is actually suggested by the author at the time of submission of the paper. We did not use this information because it was not reliable since we found many authors who included their paper in all the industry types while few authors didn't include any. Moving forward, we extracted references from individual proceedings and also recorded the number of references cited in that particular paper. We created another variable named Standard which shows whether the paper followed standard style of presenting references such as APA, MLA, Chicago, etc. A value of 1 indicates that the author used any one of the above mentioned standard formats and a 0 shows that no particular standard was followed.

Since our data collection approach was manual, we performed data validation to check the accuracy and legitimacy of our data. In our dataset, we included a binary variable which shows whether the references cited by the author followed a particular standard. Since this was a subjective decision, two researchers independently coded the variable. We validated the data by matching the observations and obtained a matching agreement of 90.9%. The remaining 9.1% of the observations were further discussed and agreed upon by mutual consensus.

We extracted 439 proceedings from SAS's website which were submitted in SAS Global Forum 2014. These proceedings consisted of paper, e-poster, quick tip and workshops. E-poster, Quick tip and Workshops are mainly presented for feedback purpose and they are not rigorous and hardly contains any literature review. Moreover, our study was to analyze all the papers and therefore we excluded all the other proceedings and ended up with 330 papers. Table 1 provides a detailed distribution of the types of proceedings from SAS Global Forum 2014.

Proceeding Type	Count
Paper	330
E- Poster	71
Quick Tip	21
Workshop	17

Table 1: Summary Table of all the Proceedings in SAS Global Forum 2014.

We found that 36 papers which were included in the proceedings list were actually not submitted. We also found 7 papers which were submitted as slides which we did not include in our analysis. On a negative side, 78 authors did not include references section in their paper which led us remove them from the analysis. Finally, we were left with 209 papers which had references and these were the ones on which we performed Hypothesis testing and Text mining. We performed statistical tests on the numerical data (structured) and text mining on the textual data (unstructured) one. The structured data consists of the author's affiliation, standard of the references used and the count of references while the unstructured data had a corpus of references cited in a paper.

METHODS & ANALYSES

HYPOTHESES TESTING

To prove our hypotheses, we performed two different statistical tests to support our claim. For the first hypothesis, we tried to examine whether the mean of the number of references cited by academicians is greater than those by professionals in their papers. To do so, we performed two sample t-test on the author's affiliation and the number of references cited in that paper. We used F- statistic as our method.

Secondly, to prove our second hypothesis which was to verify whether academicians follow the standard style of presenting references, we performed chi- squared test. Chi- Squared test is statistical hypothesis test used to test whether there is a significant difference between expected frequency and observed frequency.

TEXT MINING

Using Text Miner in SAS Enterprise Miner, we studied the corpus of references that we extracted from the SAS Global Forum 2014 proceedings. As followed in any Text Mining project, we started off by parsing the document collection in order to quantify information about the terms that are contained therein. We rejected entities such as address, currency, date, time, etc. as these were not relevant in our quest to obtain meaningful clusters. Moving on, our next task was to filter out terms and we considered only those terms which appeared in at least 6 documents which is around 3% of the total number of documents. Using Interactive Filter viewer, we tried to eliminate nuisance in the documents since there were a handful of terms which did not make any sense. There were terms which we treated as synonyms for example 'sas institute' and 'sas institute inc.' because it's the name of a company named SAS. Finally, we formed clusters of the terms that we shortlisted using Text Cluster node keeping SAS's default properties in most of the cases. We tried our hands by interchangeably changing the number of descriptive terms to look for terms that were similar in its own cluster and different from terms in other cluster. Consequently, we ended up with 5 unique clusters consisting of 8 descriptive terms in each.

RESULTS

HYPOTHESES TESTING RESULTS

To test our first hypothesis, we used all the 209 papers which consisted of references, irrespective of whether the author followed a particular standard or not. The average number of references used by

academicians was 8.65 having a standard deviation of 7.32 while the average number of references used by professionals was 6.60 with a standard deviation of 5.97. During the exploratory analysis, we found 3 observations which were outliers as far as the number of references is concerned. We removed the above mentioned three observations because the outliers can potentially bias the results. We found significant differences between average numbers of references by academicians and professionals based on 95% confidence interval. The F- value for this test is 4.71 (p= 0.0312). Thus, our first hypothesis is supported.

To prove our second hypothesis that academicians follow standard style of presenting references and professionals do not, we used all the 209 observations where the authors cited references. The Chi-Square value for this test is 12.6489 (p= 0.0004) which means that academicians are more likely to follow a standard format for writing references as compared to professionals. Table 2 lists out the distribution of standard and non-standard references between academicians and professionals as well as the chi-square vale for the test.

	Academicians	Professionals	Total
Standard	34 (79.07%)	81 (48.80%)	115
Non-Standard	9 (20.93%)	85 (51.20%)	94
TOTAL	43	166	209
Chi- Square = 12.6489 (p-value= 0.0004)			

Table 2: Distribution of Academicians and Professionals by Standard and Non-standard reference

TEXT MINING RESULTS

After performing the traditional Text Mining process, we ended up with 5 clusters with 8 descriptive in each. The clusters were fairly spread apart and the distance between clusters was significant enough to come to a conclusion that the terms in its own cluster were similar to each other and different from the terms in other clusters.

TEXT CLUSTERS

Table 3 shows all the clusters formed as a result of the text mining on the references with their respective frequency and percentage.

ID	Descriptive Terms	Frequency	%
1	https intelligence analytics software visual web guide users	40	19.14
2	+model model +test pp journal +analysis +statistics press	41	19.62
3	+report health care 'et al.' quality clinical research university	33	15.79
4	+cary +http://support.sas.com/documentation +'global forum' +'sas institute' +http://support.sas.com/resources/papers/ proceedings09/043-2009.pdf guide +proceeding conference	72	34.45
5	+american models series +'john wiley' society statistical +association +statistics	23	11.00

Table 3: Text clusters of selected terms from the references.

Definition of the Clusters

- Cluster ID#1: Cluster 1 contains references related to visual analytics, intelligence as well as web references from user guides.
- Cluster ID#2: Cluster 2 contains references related to modeling, testing, analysis and statistics. Terms such as 'pp' (page number) and 'journal' suggests citations from journals and books.
- Cluster ID#3: Cluster 3 contains references related to healthcare and clinical research paper.
- Cluster ID#4: Cluster 4 which has the highest frequency contains terms which suggests us that the authors have used references to SAS support website as well as SAS proceedings as citations in their paper.
- Cluster ID#5: Cluster 5 contains references from leading publications as well as about statistics. Its frequency is the lowest among all the clusters.

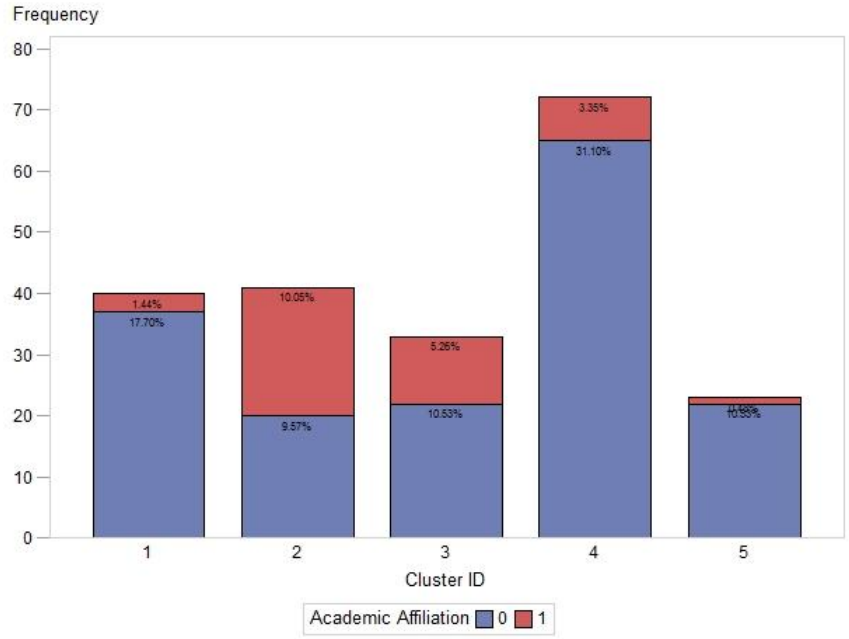

Figure 1. Text Cluster vs. Author Affiliation: Percentage of academic and professional affiliated papers within each cluster.

Figure 1 shows the distribution of academic and professional affiliated papers within each cluster. Cluster number 2 & 3 consists of maximum authors having academic affiliation. On the other hand, all the other 3 clusters contains relatively high number of authors with professional affiliation.

CONCLUSION

We found statistical evidence that academic authors use more references than professionals in SAS proceedings. Does it mean that more references in the paper makes it of better quality? Not really, using too many references can decrease the readability of the article and may also lose reader's interest. However, lack of citations may indicate the ignorance of author/s about the topic of interest. So, there is a need of sufficient number of references to position the study in the given stream of research. In some conferences and journals, instructions about the number of references are given. But in our context, SAS proceedings does not have any limitation on them.

By proving second hypothesis, we found that academic authors followed a particular style for presenting references but professional did not. If references are not properly mentioned, it can become hard for the readers to reach them. Thus, not following any standard can decrease the quality of the paper. SAS proceedings suggests to use APA format (but not restricted to it) in the template provided. So, we suggest both professionals and academicians to follow a particular style those are not following. On the other hand, we also suggest the conference chairs to be strict about formatting of references and other sections of the paper while reviewing for increasing the quality of proceedings.

An exploratory text cluster analysis of the actual content of references showed that there are differences in the choice of references by professionals and academicians. Academicians are more likely to use authentic references, on the other hand, professionals rely on website links that may not be authentic and not validated. Based on these findings, we suggest professionals to use more authentic references to increase the authenticity of the study.

Besides useful insights, this study has some limitations. First is the sample size. We only focused on the papers published in the SAS Proceedings 2014. So, including more data can even strengthen the results. Second, results may not be generalizable as we only used data from one conference. The same results may not hold true with the other similar conferences. This can be an interesting study for the future to see that professionals participating in the other similar conferences show same behavior or not. The audience of this study may be very limited (only SAS authors), however, we show an interesting application of text mining to improve the quality of research.

Notwithstanding the limitations illustrated above, this research can be very helpful for the SAS authors and SAS conferences chairs.

REFERENCES

Brocke, J. V., Simons, A., Niehaves, B., Niehaves, B., Reimer, K., Plattfaut, R., & Cleven, A. (2009). Reconstructing the giant: on the importance of rigour in documenting the literature search process.

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *Management Information Systems Quarterly*, *26*(2), 3.

Levy, Y., & Ellis, T. J. (2006). A systems approach to conduct an effective literature review in support of information systems research. *Informing Science: International Journal of an Emerging Transdiscipline*, *9*(1), 181-212.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Vijay Singh, Oklahoma State University, Stillwater, OK, Email: vijayms@okstate.edu

Vijay Singh is a second year graduate student majoring in Management Information Systems at Oklahoma State University. He has two years of experience of using SAS® tools for Data Mining, Text Mining, and Sentiment Analysis projects. He is a SAS® certified Base Programmer and Business Analyst. In May 2014, he received his SAS® and OSU Data Mining Certificate and he has been awarded SAS Global Forum 2015 student scholarship award.

Pankush Kalgotra, Oklahoma State University, Stillwater OK, Email: pankush@okstate.edu

Pankush Kalgotra is a second year doctoral student majoring in Management Information Systems at Oklahoma State University. He has four years of experience of using SAS® tools for Data Mining, Texting Mining, and Sentiment Analysis projects. He is a SAS® certified Predictive Modeler using SAS® Enterprise Miner 6.1. In December 2012, he received his SAS® and OSU Data Mining Certificate. He was recognized as SAS Student Ambassador in SAS Global Forum 2014. His team stood third in the SAS Analytics Shootout competition 2014.

Goutam Chakraborty, Oklahoma State University, Stillwater OK, Email:goutam.chakraborty@okstate.edu

Dr. Goutam Chakraborty is Ralph A. and Peggy A. Brenneman professor of marketing and founder of SAS® and OSU data mining certificate and SAS® and OSU marketing analytics certificate at Oklahoma State University. He has published in many journals such as Journal of Interactive Marketing, Journal of Advertising Research, Journal of Advertising, Journal of Business Research, etc. He has chaired the national conference for direct marketing educators for 2004 and 2005 and co-chaired M2007 data mining conference. He has over 25 years of experience in using SAS® for data analysis. He is also a Business Knowledge Series instructor for SAS®.