Use of text mining for understanding Peruvian students and faculties' perceptions on bibliometrics training

Carlos Vílchez-Román
Biblioteca Nacional del Perú / Av. De La Poesía 160, San Borja, Lima - Perú.
Universidad de San Martín de Porres / Av. Las Calandrias s/n, Santa Anita, Lima - Perú.
cvilchezr@usmp.pe
Joel Alhuay-Quispe
Universidad San Ignacio de Loyola / Av. la Fontana 550, La Molina, Lima - Perú.
Universidad Nacional Mayor de San Marcos / Ca. Germán Amézaga 375, Lima, Lima - Perú.
jalhuay@usil.edu.pe

Abstract

Background: Studies on bibliometrics and informetrics training have focused on teachers and curricular experts' opinion, only a few studies have examined undergraduate students and practitioners' perceptions. Objective: To understand how librarianship students and professionals perceive the bibliometrics and informetrics training delivered to them. Methods: For data collection, we used a survey with opened-ended questions, to know the genuine responses of the participants. After working with the automatic term extraction technique, for codifying the answers we employed a data dictionary for quantifying the frequency of occurrences. The software programs used at this stage were terMEXt and LWIC. Data analysis was carried out with statistics of mean difference and the correlation coefficient. Results: The output of statistical analysis lets us understood how students and practitioners perceive the bibliometrics and informetrics training delivered to them. Conclusion: Text mining techniques facilitates the processing of responses to opened-ended questions, and contributes with a quantitative approach to analyzing people's opinions.

1 Introduction

1.1 Bibliometrics training within librarian academic community

Consultants, analysts, as well as research managers, are increasingly using bibliometrics and informetrics based approaches and techniques. Because of the widespread adoption of bibliometrics, several times the actors mentioned earlier do not have the required training or capabilities for using those tools in a proper way. Seminal studies on informetrics training go back to the middle of the 1970s. Since then, the analysis carried out focused on opinions of Library and Information Science (LIS) faculties and experts in curricular design and, in general, those findings did not evidence any variation when it was discussed the bibliometrics training for Latin American countries.

At the beginning of the 1980s, Schader (1981) reported on a growing interest in bibliometrics teaching at higher education, particularly in the medical sciences, because this analytical approach would contribute to improving the curricular contents in information management, by considering two dimensions for teaching it: one theoretical and the other applied.

In the LIS field, Aiyepeku (1975), one of the pioneers who studied the approach used for bibliometrics training, examined the inclusion of bibliometrics teaching within the curricular plans of librarianship schools. Later, Dou, Quoniam, & Hassanaly (1988) proposed to teach bibliometric analysis by starting with the bibliographical references downloaded from the previous indexing databases. To reinforce their argument, they introduced examples based on command-line instructions of MS-DOS operating system. In the middle of the 1990s, Ungern-Sternberg (1995) stated that teaching of bibliometric methods could be delivered by presence seminars and the use of online systems; also, he proposed a curricular method and contents programming for the design of a course on bibliometrics (Ungern-Sternberg, 1998).

In brief, first studies on bibliometrics training described proposals for developed countries: United States (Schader, 1981), France (Dou, et al., 1988), and Finland (Ungern-Sternberg, 1998);

nevertheless, teaching experiences of bibliometric methods within LIS schools in Hispanic countries are recent. In Spain, studies found that statistical topics in LIS were only applied for academic programs oriented to getting the professional license (4-6 years), but this orientation was not an option for 3-year-diplomas (Jiménez-Contreras, & Pulgarín-Guerrero, 1998).

As observed, the reflections of cited authors highlight the programming of curricular contents or the acquisition of specific skills (e.g., management of bibliographic references for later analysis with statistical tools). However, the mentioned studies did not examine the opinions of beneficiaries of the educative service at the higher education level: undergraduate students and practitioners, because the last ones must select and train students in the final stage of their professional training.

1.2 Text mining for analyzing opinions

Text mining as a method for quantitative analysis –after standardization of responses to opened-ended questions– has been applied to academic and business contexts for understanding participants' views and attitudes.

Within the academic landscape, few researchers have worked with students and practitioners as subjects of the study. Sliusarenko, et al. (2013) used text mining for examining written responses in opened-composed-based evaluations. Based on this approach, authors were able to understand how responses related with the obtained scores in evaluations. Freak & Miller (2015) employed it for identifying thematic groups in the replies about perceptions of teachers specialized in physical education. Outside the academic boundaries, Yi, et al. (2015) analyzed the public perception on a Chinese touristic trademark by applying text mining approaches to deal with questionnaires responses. From another perspective, within an effort to automate the analysis of opinions, Kumar & Jain (2015) proposed a system for automatic evaluation that uses text mining for analyzing views collected in questionnaires used to measure the professional development of lecturers at a higher education institution.

As described, there is evidence showing that text mining tools facilitate the processing of opened-ended questions and make it possible a quantitative approach for analyzing peoples' opinions.

1.3 Research purpose

It is an exploratory study (do not include hypotheses) that analyzes students and practitioners' opinions, graduated from Peruvian LIS schools, on bibliometrics and informetrics training. For that purpose, researchers worked with text mining-based techniques and tools.

2 Methods

2.1 Sample

We worked with a nonprobabilistic sample, and contacted participants attending academic events, joining mailing lists or sending invitations through email. We considered the two Peruvian LIS schools: Pontificia Universidad Católica del Perú (PUCP) and Universidad Nacional Mayor de San Marcos (UNMSM).

2.2 Data collection

We designed a questionnaire (see Annex) that had two sections: background data and three opened-ended questions for exploring librarians' opinions in three primary areas for bibliometrics and informetrics training: shared meanings, contribution to the career, and required conditions for its teaching.

1. How do you define bibliometrics and informetrics?
2. What role can play bibliometrics and informetrics within the profession?
3. What knowledge and skills are required to deal with the bibliometrics approaches and techniques?

Researchers applied a print-based and an online version (Google Forms). Before giving their answers, participants signed an informed consent form. The time for filling out the questionnaire was 15 minutes.

2.3 Text processing and coding

We transcribed responses into simple text files. Before creating the data dictionaries for carrying out the quantitative analysis of content, we used a text mining technique known as automatic term extraction, for which we worked with the software program terMEXt. This application, developed by Barrón Cedeño (2008), is based on the online service TerMine, created by Frantzi, Ananiadou, &

Mima (2000) from the National Centre for Text Mining at Manchester University.

Given that the responses to the first question were redundant with the other two, this item was omitted from text processing. We parsed out the answers to the second and third questions by using the default dictionary of terMEXt, and computed the NC-value for multi-word terms with the higher semantic value. Algorithms for automatic term extraction use the NC-value to identify multiword terms with a meaning that makes it possible to differentiate from other candidate terms. Based on these words, we build dictionaries for the two questions considered for later analysis.

After creating both dictionaries, we cleaned out text files for processing them with the program Linguistic Word Inquiry Count (LWIC), that counts the number of occurrences, stems and conditional structures associated with each dictionary category. This procedure allowed us to compute the score used in correlation analysis for all the variables. Data obtained with LWIC program were entered into the Statistical Program for Social Sciences (SPSS) for later analysis.

2.4 Data analysis

For the description of results, we used measures of central tendency and for dispersion only the standard deviation. To examine whether variables were associated, we computed a correlation matrix, based on the product moment Pearson's coefficient or Spearman's rho, if variables had a high dispersion.

3 Results

Most of the variables showed a high data dispersion (Table 1). In average, participants had 30 years and their educations lasted five years (the difference between the year they entered university [academic age] and the year they finished their undergraduate studies [professional age] from their center of studies; nevertheless, both variables showed the highest dispersion of all analyzed variables). Word counting –of those variables detailed in the data dictionary, as described in the Methods section–for categories of questions two and three also exhibited a high dispersion. The large values of the standard deviation of the variables were taken into account when carrying out the correlation analysis.

Regarding the quantitative analysis of the content, in question two, the categories evaluator and promoter showed a small correlation ($\rho = 0.364$, p = 0.004). We observed a similar pattern for question three: categories theoretical and analytical skills also exhibited a low correlation ($\rho = 0.388$, p = 0.002), but a moderated one for the categories analytical skills and information search expertise ($\rho = 0.432$, p = 0.001). Table 2 details the correlation matrix for variables considered for the study.

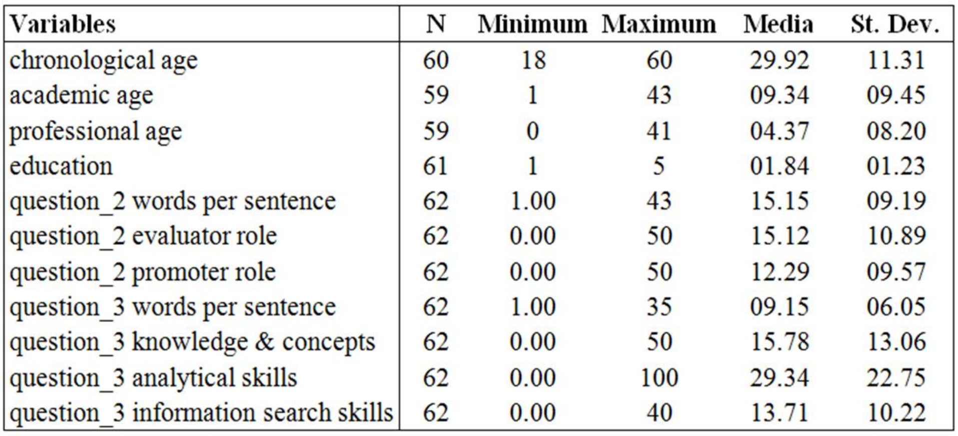
4 Discussion and conclusions

4.1 Text mining and bibliometrics training

Question two of the questionnaire asked about the role of bibliometrics and informetrics in the development of LIS as a discipline. Text mining techniques led us to identify two categories that helped us to understand this function: one dimension oriented to evaluation and the other one to promotion, which also showed a moderate correlation.

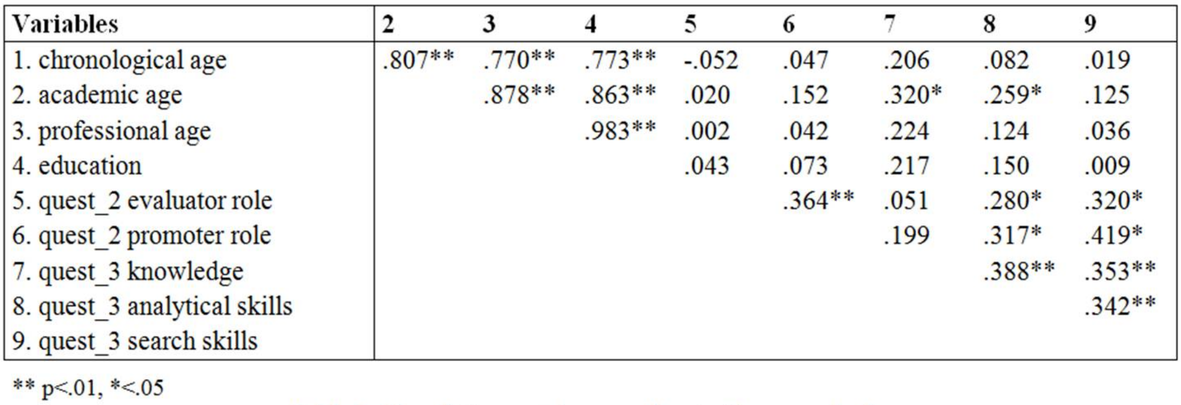
Effectively, the application of bibliometrics tools contributes to disseminating and promoting one the few contributions originated from LIS: metric studies on information (informetrics). It is true that, because of its applied nature, librarianship receives influences from several disciplines: psychology, sociology, history, and, in recent years, computer science, particularly a specialized field known as computer-human interaction, given the current significance of the development of search interfaces and the implementation of institutional repositories. However, those contributions come from outside the LIS profession. For that reason, it is revealing that practitioners highlight the promoter role of bibliometrics as a driver for renewal within this discipline. This promoting-oriented dimension complements the other role: evaluator, in the sense that bibliometrics becomes a sort of thermometer to measure the internal development of the discipline. This way, it will be possible to determine whether the library profession is ready to face the challenges set out by academic community; otherwise, the demand for professionals-experts-in-evaluating-scientific-production will be met by practitioners educated in other careers.

On the other side, question three focused on knowledge and skills required to deal appropriately with bibliometric approaches and techniques. About this issue, text mining let us identify three primary areas for advising the necessary professional training: knowledge and concepts, an-



Variables	N	Minimum	Maximum	Media	St. Dev.
chronological age	60	18	60	29.92	11.31
academic age	59	1	43	09.34	09.45
professional age	59	0	41	04.37	08.20
education	61	1	5	01.84	01.23
question_2 words per sentence	62	1.00	43	15.15	09.19
question_2 evaluator role	62	0.00	50	15.12	10.89
question_2 promoter role	62	0.00	50	12.29	09.57
question_3 words per sentence	62	1.00	35	09.15	06.05
question_3 knowledge & concepts	62	0.00	50	15.78	13.06
question_3 analytical skills	62	0.00	100	29.34	22.75
question_3 information search skills	62	0.00	40	13.71	10.22

Table 1: Central tendency and dispersion of variables



Variables	2	3	4	5	6	7	8	9
1. chronological age	.807**	.770**	.773**	-.052	.047	.206	.082	.019
2. academic age		.878**	.863**	.020	.152	.320*	.259*	.125
3. professional age			.983**	.002	.042	.224	.124	.036
4. education				.043	.073	.217	.150	.009
5. quest_2 evaluator role					.364**	.051	.280*	.320*
6. quest_2 promoter role						.199	.317*	.419*
7. quest_3 knowledge							.388**	.353**
8. quest_3 analytical skills								.342**
9. quest_3 search skills								

** p<.01, *<.05

Table 2: Correlation matrix according to Spearman's rho

alytical skills, and information search expertise. Correlation between the last ones was higher than between the first two variables; nevertheless, all of them got statistical significance. Those relationships exemplify that the most valued skills of librarians (information search) is associated with the ability to analyze bibliometric studies and indicators, but this link is weak when it deals with the acquisition of theoretical foundations of bibliometrics.

We can verify this fact by looking at the popularity gained by bibliometric indicators –at least the first generation ones– within the Peruvian LIS community. However, this widespread adoption has not translated into studies or investigations featuring sound theoretical foundations, but until now Peruvian librarians have preferred the applied dimension, rather than the integration of theoretical and applied perspectives, which would lead us to bibliometrics studies published in specialized journals.

4.2 Challenges for bibliometrics training

Bibliometrics education, as an element of a specialized training program or as part of a curricular plan, needs to be oriented toward a target population with a previous background in issues related to information management or scientific publications. A practical way to carry out an educative program in bibliometrics would be targeting to managers of information services at university libraries, with courses necessarily lectured by recognized experts (Laitinen, 2015). Also, the methodological proposal for bibliometrics education must emphasize the use of open access software programs (Sanz-Casado, et al., 2002), because of its widespread adoption in Latin American countries and the cost-effectiveness of using these tools, compared with the benefits obtained by working with licensed programs and platforms.

References

Aiyepeku, W. O. 1975. Bibliometrics in Information-

Science curricula. *Information Scientist*, 9(1), 29-34.

Barrón Cedeño, L. (2008). *Manual para el extractor de término terMEXt*. Retrieved from http://goo.gl/jQjlk2

Dou, H., Quoniam, L., and Hassanaly, P. 1988. Teaching bibliometric analysis and MS/DOS commands. *Education for Information*, 6(4), 411-423.

Frantzi, K., Ananiadou, S. y Mima, H. 2000. Automatic recognition of multi-word terms. *International Journal of Digital Libraries*, 3(2), pp.117-132.

Freak, A., & Miller, J. 2015. Magnifying pre-service generalist teachers' perceptions of preparedness to teach primary school physical education. *Physical Education and Sport Pedagogy*, 1-20.

Jiménez-Contreras, E., y Pulgarín-Guerrero, A. 1998. Bibliometrics-Informetrics and other quantitative subjects in Library and Information Science curricula in Spain. *Education for Information*, 16(4), 341-355.

Kumar, A., & Jain, R. 2015. Sentiment analysis and Feedback Evaluation. En *MOOCs, Innovation and Technology in Education (MITE), 2015 IEEE 3rd International Conference* (pp. 433-436)

Laitinen, M. 2015. The Benefits of Learning Bibliometrics on the Teaching Librarians' Information Literacy. En *The Third European Conference on Information Literacy (ECIL)* (p. 159).

Sanz-Casado, E., Suarez-Balseiro, C., García-Zorita, C., Martín-Moreno, C., and Lascurain-Sánchez, M. L. 2002. Metric studies of information: An Approach towards a Practical Teaching Method. *Education for Information*, 20(2), 133-144.

Schader, A. M. 1981. Teaching bibliometrics. *Library Trends*, 30(1), 151-172.

Sliusarenko, T., Clemmensen, L. K. H., and Ersbøll, B. K. 2013. Text mining in students' course evaluations: Relationships between open-ended comments and quantitative scores. En *CSEDU 2013-Proceedings of the 5th International Conference on Computer Supported Education*. SciTePress.

von Ungern-Sternberg, S. 1995. Applications in teaching bibliometrics. En *61st IFLA General Conference - Conference Proceedings*. Retrieved from http://hdl.handle.net/10150/106138

von Ungern-Sternberg, S. 1998. Teaching Bibliometrics. *Journal of Education for Library and Information Science*, 39(1), 76-80. http://doi.org/10.2307/40324182

Yi, C., Yang, Y., and Hong Mei, Y. 2015. Research on audience's perception of tourism brand of Guiyang based on the text mining of ROST. *Journal of Chongqing Normal University*, 32(1), 126-134.

A Annex

A.1 Informed consent protocol and survey form [In Spanish]

Avaliable at http://dx.doi.org/10.6084/m9.figshare.3817155