Productivity and doctoral research in a Brazilian Nuclear Research institution: validating co-word analysis technique

Rogério Mugnaini[1] , Mery P. Zamudio Igami,[2] José Carlos Bressiani[2]

mugnaini@usp.br
[1]University of São Paulo, Av. Arlindo Bettio, 1.000, Prédio A-1 – sala 54, CEP 03828-000, São Paulo (Brazil)

[2] mery@ipen.br [2] jbressia@ipen.br
[2]Instituto de Pesquisas Energéticas e Nucleares, IPEN-CNEN/SP, São Paulo (Brazil)

Introduction

Nowadays, the value of doctoral thesis has been related to the articles it might produce. In some areas, two or three journal articles are required to achieve the PhD degree, but in areas where the obligation to produce a thesis is maintained, the article is the real final publication.

These items of literature are considered as one of the first knowledge sources, produced under rigorous scientific patrons. Thesis elaboration demands considerable effort that strengthens the scholar in the paths of knowledge construction to get familiar with the logistics of the scientific work. With that in mind, it is expected that the process should continue, increasing the flux of science.

Co-word analysis is a content analysis technique that uses patterns of co-occurrences of pairs of items in the corpus of a text to identify the relationship between them. The technique was initially proposed to map the dynamics of science (Law & Whittaker, 1992; Courtial, 1994; Ding & Chowdhury, 2001) and has been applied to specific purposes such as identifying research topics and clustering them in different scientific areas (Stegmann & Grohmann, 2003; Neff & Corley, 2009) a vv v vnd supporting scientific policy needs (Yoon et al., 2010).

This study aimed at evaluating the use of co-word analysis to establish a relationship between doctoral thesis and journal articles published by the students of a Post-graduation program in a Brazilian National Nuclear Research institution, in São Paulo, Brazil, for three decades of operation. The method was validated through application of survey to the authors.

Methods

Data were collected from a local institutional database, where the scientific production of the students is indexed. Both thesis and articles are indexed with descriptors extracted from a controlled vocabulary used in the nuclear area from the International Nuclear Information System (INIS). The corpus of analysis was composed by the doctoral theses of 400 students in the period of 1982-2009 and respective articles they have signed. A time range of five years before and after the thesis presentation was established, resulting in 2,209 articles in the period of 1977-2010.

It is known that each thesis could be unfolded in many articles, but do all of them have contents really correlated to the thesis? That's what has been observed through the descriptors co-occurrence that acts as a filter to determine the correlation. For this purpose, the co-word analysis was applied in order to determine if an article is really a product of the same research carried out in the doctoral research. But, after establishing the number of co-words, a survey was applied to validate the methodology.A random sample of 128 authors was contacted by an electronic survey and asked to indicate the correlation degree of their articles in a scale of 4 (ranging from 1 for "strong" correlation to

4 for "weak"). So, the results between the automatic methodology and the author's opinion were compared to establish a minimum level of co-words to determine correlation. Finally, the automatic methodology was applied to the total corpus.

Results

From the 128 authors of the sample, 100 (79%) answered, resulting a total of 397 articles (18% from the total of 2,209). So, the level of correlation declared by the authors and the number of coincident co-words were crossed (Table 1).

Authors' responses that indicate correlation (1 for "strong" and 2 for "medium") are concerning mainly to articles with three or more coincident co-words, allowing the delimitation of the automatic methodology to be based in this amount of co-words (dotted gray line). Authors' responses 3 and 4 indicate weak or no correlation, respectively. From table 1, one can observe that the dotted gray line maximizes the agreement between author's opinion and number of coincident co-words: for 211 articles (53.1% of the sample) the agreement regards the presence of correlation (author says "yes" and three or more coincident co-words); and for 134 articles (33.8%), the agreement regards the absence of correlation (author says "no" and one or less coincident co-words). There is disagreement in 52 articles or 13.1% of the sample.

Table 1 - Comparative results between automatic methodology and author's opinion.

Number of co-words coincidents	Author's correlation (survey)				Total
	Yes		No		
	1	2	3	4	
0	3	2	18	53	116
1	4	3	14	9	30
5	29	12	12	9	62
6	34	18	10	4	66
9	1				1
10	45	15	2	1	63
14	1				1
15	25	10	1		36
21	12	2		1	15
27	1				1
28	4	2			6
Total	159	64	57	117	397

Defined the number of coincident co-words, the automatic methodology was applied to the whole corpus of articles. Fig. 1 permits the analysis of percentage of published correlated articles in relation to the total of articles published by each author in that year. One can observe that in the majority of the years the production is directly related to the doctoral research. The polynomial tendency line shows a small variation in the period.

Another aspect analysed was the time for publishing an article by comparing the year of publication of the thesis and the author's article (fig. 2). Correlated articles (light gray square) were presented separated from not correlated ones (dark gray triangles), showing that until 1997, not correlated articles take less average time to be published in comparison to the year of the thesis lecture; from 1998 to 2004 this relation is inverted; and after 2004, the period of 5 years from publication of articles is not complete, producing an unreal decrease in the average time for publishing them (for this reason the bullets are signed in white colour).

Final remarks

Co-word analysis demonstrated to be a quite satisfactory methodology to find close

correlation between pieces of literature, as a matter of fact, much of its success depends of the quality of the keywords, database availability and confidence data. Author's opinion was very important to validate the methodology and it allows extending to the total corpus of articles.

Although, another element to identified co-occurrence between items could be used, information systems which maintain bibliographic repositories using controlled vocabulary, have more chance to succeed in using co-word analysis. This could be considered a limitation of this approach.

Also, this kind of work is important to support managers' decisions for establishing an institutional scientific policy.

Figure 1 – Average percentage of correlated articles per thesis (1982-2009)

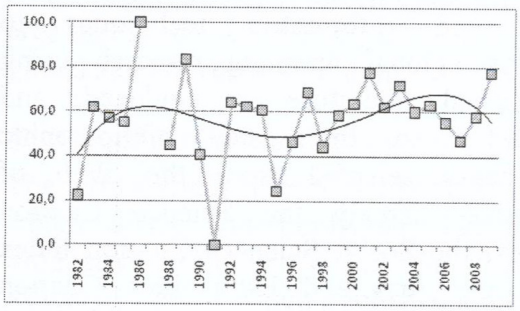
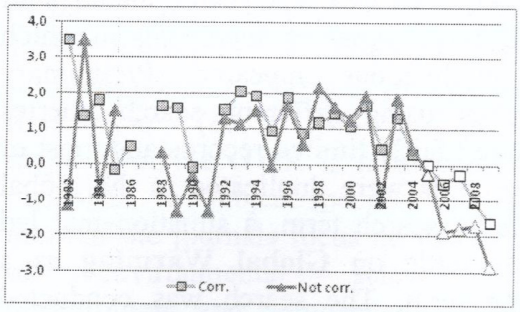


Figure 2 – Average of the lag time between the year of publication of the thesis and the articles (1982-2009)



Acknowledgment:

This study was supported by CNPq - Process number 483095/2009-5.

References

Courtial, J. P.(1994) A co- word analysis of scientometrics. ***Scientometrics***, 31.251-260.

Ding, Y. ; Chowdhury, G. G. (2001) Bibliometric cartography of information retrieval research by using co-word analysis. ***Information Processing and Management,*** v.37, 817-842.

Law, J. & Whittaker, J.(1992) Mapping acidification research: a test of the co-word method. ***Scientometrics,*** 23, 417-461.

Neff, M. W. ; Corley, E. A. (2009). 35 years and 160,000 articles: a bibliometric exploration of the evolution of ecology. ***Scientometrics***, 80, .657-682,

Stegmann, J. & Grohmann, G.(2003) Hypothesis generation guided by co-word clustering. ***Scientometrics*** 56, 111-135.

Yoon, B. & Lee, S.; Lee, G. (2010) Development and application of a keyword-based knowledge map for effective R&D planning. ***Scientometrics***, 85, 803-820.