JSCIRES RESEARCH ARTICLE

A new model to identify the productivity of theses in terms of articles using co-word analysis

Mery Piedad Zamudio Igami*, José Carlos Bressiani, Rogerio Mugnaini[1]

Department of Library, Nuclear and Energy Research Institute, (IPEN–CNEN/SP), Superintendence, Nuclear and Energy Research Institute, (IPEN-CNEN/SP), São Paulo, Brazil, [1]Department of Librarianship and Documentation School of Communication and Arts (ECA), University of São Paulo, São Paulo – SP, Brazil

ABSTRACT

A thesis defense should be considered as not the end but the starting point for scientific communication flow. How many articles truly extend doctoral research? This article proposes a new model to automatically identify the productivity of theses in terms of article publications. We evaluate the use of the co-word analysis technique to establish relationships among 401 doctoral theses and 2,211 articles journal articles published by students in a graduate program at a Brazilian National Nuclear Research Institution (IPEN-CNEN/SP).To identify the relationship between a thesis and an article published by the same author, we used co-descriptor pairs from a controlled vocabulary. To validate the proposed model, a survey was applied to a random sample of theses authors (n = 128, response rate of 79%), thus establishing a minimum threshold of three coincident co-descriptors to identify the relationship between theses and articles. The agreement level between an author's opinion and the automatic method was 86.9%, with a sampling error of 7.36%, which indicates an acceptable level of accuracy. Differences between the related or nonrelated distributions of articles were also demonstrated, as was a reduction in the median lag time to publication and the supervisor's influence on student productivity.

Keywords: Articles, co-descriptors, co-word analysis, scientometrics, theses

INTRODUCTION

A national agency called the Coordination for the Improvement of Higher Level Personnel coordinates Brazilian graduate programs, which are subjected to a thorough evaluation process every 3 years. Scientific output, measured as articles published during thesis preparation, is among the prerequisites for a positive program evaluation, although each program sets annual publication goals. Graduate students are encouraged to report the results of their research as quickly as possible, regardless of their graduation year, to contribute to scientific progress. Several graduate programs in Brazil have begun to require that candidates publish two or three journal articles to earn a doctoral degree.

Theses are considered nonconventional literature because they are generally kept out of the commercial circuit. These works of literature are considered to be among the first sources of knowledge produced under the aegis of rigorous scientific standards. Thesis elaboration demands considerable effort that helps the scholar construct knowledge and become familiar with the logistics of scientific work. This process should continue, and thesis finalization should be considered not the end but the beginning of increasing scientific flow (Igami, 2011).[1]

Although theses are recognized by the scientific community as official and validated scientific documents, theses are seldom read by more than a few scientists because they are, usually, available only in libraries, which are institutional repositories. Even though, they are indexed in international

***Address for correspondence:**
E-mail: mery@ipen.br

Access this article online

Quick Response Code:

Website: www.jscires.org

DOI: 10.4103/2320-0057.143660

databases, they are written in the scholar's native language, which presents a barrier. Research results, which are published in journal articles, are more readable and more easily accessible to all scientists through secondary publication.

In this context, the current value of a doctoral thesis is related to the articles it might produce. The thesis could be developed into many articles, but how many of these articles actually extend the doctoral research?

RELATED WORK

Doctoral dissertations have been examined in many previous studies with different focuses, but they are not a common object of analysis. One of the main studies of doctoral dissertations was conducted by Solla Price (1963)[2] and presented in his seminal book. He used PhD dissertations as one of the several indicators of scientific growth and alongside other measurements, as an indicator of the addition of a new researcher to a scientific field.

Andersen and Hammarfelt (2011)[3] discussed the general applicability of dissertations as an indicator of scholarly growth and provided a historical view on the growth of research. They demonstrated that the number of dissertations produced can be seen as an important indicator in addition to other types of publications, such as patent articles and citations, which can be used to measure research output.

In Turkish, Yaman and Atay (2007)[4] used theses presented between 1988 and 2002 to study and characterize the area of sports science. They found that most theses were prepared at the Institute of Health Sciences, although their content primarily involved training and movement – that is, these theses were more closely related to the humanities and to natural and social sciences. This finding provides a clear indication that because sports science is a multidisciplinary area, there may be a need for new programs at appropriate institutes that are devoted to and experienced with extra-medical fields of science. They also concluded that a low percentage of published papers based on theses were published in peer-reviewed indexed journals. Thus, these works remain invisible to the scientific community.

Breimer (1996)[5] conducted a study that analyzed the authorship and usage of published papers in current Swedish publication-based biomedical doctoral dissertations. He examined these dissertations in 1992 and compared them with a sample from 1968 to 1992. He proposed that three papers should form the basis of a common European PhD if it is to be completed (including an examination) within 3 years or four papers in 4 years.

Varshney (2012)[6] presented an examination of a longitudinal study (2004-2011) of citation behavior in 728 doctoral theses at the Massachusetts Institute of Technology's Department of Electrical Engineering and Computer Science. He found that the number of references cited have increased over the years. The age of documents that are cited has become more randomized over time, suggesting a greater transactive role for scientific literature. These scientometric findings suggest a Google effect that is indicative of a cognitive change in research students. When researchers remember where information is stored rather than remembering the information itself, the nature of their research may change.

Doctoral researchers' productivity is also the subject of analysis in many countries, as confirmed by a number of recently published papers. Studies have been performed primarily in the health sciences in France (Salmi, Gana, and Mouillet, 2001),[7] Croatia (Frkovic, Skender, and Dojcinovic, 2003),[8] Brazil (Ramos *et al*., 2009;[9] Sacardo and Hayashi, 2011;[10] Younes, Deheinzelin, and Birolini, 2005),[11] and Peru (Arriola-Quiroz, Curioso, Cruz-Encarnacion, and Gayoso, 2010)[12] and in other scientific domains in the US (Anwar, 2004;[13] Lee, 2000;[14] Mallette, 2006)[15] and Canada (Larivière, 2010)[16] [Table 1]. Reforming doctoral training has also been discussed in Germany, particularly after the third cycle of the Bologna Process, which aimed to create a European Higher Education Area (Barrier and Musselin, 2009).[17] In this process, a PhD thesis based on published works (a reality in Sweden) can be widely adopted (Breimer, 2010),[18] reinforcing the importance of analyzing the productivity of doctoral students.

An important aspect in these studies [Table 1] is that most have used searching to match researchers' scientific manuscripts, which is sometimes followed by reading if a literal match does not occur. The most frequently used bibliographic field is the author's name, which is frequently combined with the title, abstract, supervisor's name, and in one specific case, the author's city. The author's name is a problematic field because homonyms are common. This problem is not easy to solve when conducting macro-level studies.

Different specialized databases are used to retrieve publications in each scientific area. Brazilian studies use

Table 1: Comparing earlier studies of doctoral students productivity

Citation	Matching method	Bibliographic field	Article data source	Area
Lee (2000)	Searching	Author's name	Chemical abstracts, PsycINFO, and American literature	Chemistry, psychology, and American literature
Salmi *et al.* (2001)	Searching/reading	Author's or supervisor's name, title, or abstract	MEDLINE	Medicine
Frkovic *et al.* (2003)	Searching/reading	Author's name, matching ideas (rather than a literal translation from Croatian theses) in the title	MEDLINE and current contents	Medicine
Anwar (2004)	Searching	Author's name, title, and abstract	LL, LISA, and Global Books in Print	Library and information science
Younes *et al.* (2005)	Searching	Author's name	MEDLINE and LILACS	Medicine
Mallette (2006)	Searching	Author's name	EBSCOhost meta-search engine (8 databases, including ERIC)	Education
Ramos *et al.* (2009)	Searching/reading	Author's name and title	Lattes (curriculum database)	Physical education
Arriola-Quiroz (2010)	Reading	Author and thesis supervisor's name, title, and abstract	PubMed, LILACS, LIPECS, and SciELO (Google Scholar?)	Medicine
Larivière (2011)	Searching (homographs removed manually/algorithm)	Author's name and address from Quebec	Web of Science	Health sciences, natural sciences and engineering, social sciences, and arts and humanities
Sacardo and Hayashi (2011)	Searching	Author's name	Lattes (curriculum database)	Physical and special education

LL=Library literature, LISA=Library and Information Library Science Abstracts, ERIC=The Education Resources Information Center, LILACS=Literatura Latino-americana e do Caribe em Ciências da Saúde, LIPECS=Peruvian Health Science Literature Database

the Lattes Platform, a national database of curricula maintained by the National Council for Scientific and Technological Development (CNPq, 2013).[19] This data source, which contains approximately 1.6 million curricula and is essentially the only one of its type in Latin America, has been referred as "one of the cleanest researcher databases in existence" and is widely used by the national federal agencies when making funding decisions and by universities when making tenure and promotion decisions (Lane, p. 489, 2010).[20] It has also been increasingly used in scientometric studies (Leite, Mugnaini, and Leta, 2011),[21] although data obtained from the Lattes Platform require reformatting to be used in metric studies (Mena-Chalco *et al*., 2009).[22] Although the curriculum information is self-completed, its accuracy stems from the fact that curricula must be updated with every funding request (CNPq, 2013).[19]

As shown in Table 1, some studies have restricted their analyses to articles that directly result from doctoral research. In all cases, this process consumes considerable time, indicating that it is advisable to propose an automatic method to perform this task. With this premise in mind, and in contrast to earlier studies, this article proposes a model to automatically identify thesis productivity based on the number of articles published and correlated with the same subject of the thesis.

Co-word Analysis

Co-word analysis assumes that descriptors of a given study appropriately describe its contents (i.e., a content analysis technique) because word co-occurrence enables the identification of the association levels among items found in the analyzed text. If more overlapping term pairs exist among texts in a given knowledge area, the probability of a relationship between them is higher (Van Raan, 1993).[23]

Ding, Chowdhury, and Foo (2001)[24] mapped changes in data recovery from 1987 to 1997 by combining two collection methods: Descriptors, resulting from the indexing process, and keywords, which are extracted from titles or abstracts. The authors concluded that abstracts can provide keywords more efficiently.

Most studies using co-word analysis have implemented graphical representations that show the centrality level and density of the fields studied using connection networks. The words that can be used are selected from the texts, titles, abstracts, or descriptors attributed by indexes or identified in a given context independent of their semantic content, considering the frequency of pair combinations (He, 1999).[25]

Current scientometric research focuses on several aspects using co-word analysis. Romo-Fernandez *et al.* (2013)[26] described an analysis of co-occurrence keywords that

aimed to reveal publication patterns in the field of renewable energy, including the temporal evolution of different research lines in this field over the last two decades. Zong *et al*. (2013)[27] mapped the intellectual structure of research in doctoral dissertations in library and information science in China using co-word analysis. These authors studied the internal and external structure and relationship of research fields and found that the research fields of LIS doctoral dissertations in China are varied, and many of these research fields are still immature; there are fewer well-developed and core research fields. Furthermore in China, Liu *et al*. (2012)[28] used co-word analysis to map the intellectual structure of a digital library (DL), including the relationships among keywords, the research structure, and the situation. These results provide a basis for understanding advances in the DL field in China.

An and Wu (2011)[29] analyzed the evolution of the stem cell field using co-word analysis. Articles in stem cell journals were downloaded from PubMed for analysis. These authors found that co-word analysis based on the subject heading weighting can demonstrate the trends of a specific field.

Yang, Wu, and Cui (2012)[30] explored the concept network and developmental tendency in certain fields using co-word analysis. They performed a comparison of the characteristics of three visualization methods and analyzed the development of the disciplinary structure in terms of multiple aspects by integrating three visualization methods into one readable map.

There are certain limitations when using this technique, as noted by authors concerned with the selection of descriptors. The main objection is related to the use of descriptors from a controlled vocabulary. Some authors (Callon *et al*., 1986;[31] Turner *et al*., 1988)[32] have identified this step as a limitation of the use of relationship studies with controlled vocabulary descriptors (i.e., the "indexer effect"). This effect occurs in the concept assignment step, in which indexer subjectivity may interfere with document representation and generate inconsistencies.

Indexing is a step in document representation. It is included in the macro universe and designates the document analysis and indexing languages used for indexing. A documentary reading occurs when an indexer performs a subject analysis and identifies the main concepts addressed in the documents to represent the subject (Silva and Fujita, 2004).[33] This is the point at which logical, linguistic, and cognitive aspects involved in indexing represent indexer interference factors. This aspect of indexing is especially critical for understanding this study because the co-word analysis technique uses indexer-assigned descriptors as relationship objects between articles and theses.

Whittaker *et al*. (1989)[34] and Law and Whittaker (1992)[35] conducted studies in this area using the Pascal database to compare and analyze the two descriptor extraction methods (e.g., those extracted from article titles and abstracts and those assigned by indexers). After surveying 83 experts in the subject, they concluded that concerns about indexing quality were unfounded.

Recent comparative studies on keywords assigned by authors and indexer-selected descriptors were conducted to assess the descriptor performance. The results showed that keywords are a crucial information source for enriching cataloging terms because 25% of keywords are found precisely among descriptors, and 21% can be found after some normalization, for a total of 46% of keywords that are found (Gil-Leiva and Alonso-Arroyo, 2007).[36]

Wang *et al*. (2012)[37] published a paper analyzing some limitations of keywords and indexes used in co-word analysis. They proposed a new semantic-based co-word analysis that can effectively integrate experts' knowledge into co-word analysis. These authors showed that the performance of this method proved to be very good and represent an advance in the state of the co-word analysis research, indicating future research directions.

This study aims to evaluate thesis productivity using a co-word analysis to establish a relationship between doctoral theses and journal articles published by the students in a graduate program at an IPEN over three decades. It assumes that the article is related to a thesis such that the article could not have been written without completing the thesis. The method is validated using an electronic survey sent to a sample of thesis authors. The results address the following questions:

WHAT IS THE LEVEL OF DOCTORAL RESEARCHER PRODUCTIVITY IN TERMS OF ARTICLES PUBLISHED DURING THE THESIS ELABORATION PROCESS AND SUBSEQUENT YEARS?

- Could a co-occurrence-based automatic method express the authors' opinions about the relationship between their published articles and the thesis?
- What percentage of articles is truly related to the thesis?

METHODOLOGY

Information in nuclear science is well-organized due to the availability of a database managed by the International Nuclear Information System (INIS) in Vienna. The INIS is a decentralized system in which each country member of the International Atomic Energy Agency collects and inputs its own literature. The INIS provides training courses for indexers and catalogers and distributes manuals, procedures, subject category lists, and a multilingual thesaurus, which are updated annually. However, it is not possible to locate studies that use a co-word analysis to monitor thesis productivity for nuclear energy at either the national or the regional level.

Data were collected from the local institutional database, where all of the graduate students' scientific production is deposited. This database integrates the library information system of the Institution and has a similar structure as the INIS international nuclear database. Both theses and articles, the corpus of this analysis, are indexed with descriptors extracted from a controlled vocabulary. Furthermore, both types of items are indexed by nuclear science experts and information professionals trained to perform this task, either at the INIS headquarters in Vienna or at the National Energy Commission in Rio de Janeiro, where the Brazilian liaison officer resides. This practice assures a standard application of the taxonomy.

In the graduate program, 536 theses were defended over the past three decades. However, 55 (10.26%) students had no Lattes Curriculum available, 42 (8.73%) published no articles at all, and 38 (7.90%) had no registered publications during the period evaluated. A total of 401 theses was obtained.

The respective article production was retrieved, identifying 2,211 journal articles (published between 1977 and 2009). To identify the production derived from doctoral research, a time range was fixed to retrieve articles published 5-year before and after the thesis defense year. This 5-year span covers the time lag for article preparation and publication. Searching the literature for thesis productivity indicated that papers differed in their time range for completion, although most reported a time span from 1 to 3 years (Lee, 2000;[14] Mallette, 2006).[15]

All information was corrected manually and managed in Microsoft Excel and CISIS, resulting in a relational database with theses and their respective articles, each identified with unique identifiers (i.e., the IPEN-Doc code used for internal library control), author information, thesis supervisor, defense year, complete bibliographical information about the thesis or article, and subject information from INIS descriptors [Appendix 1].

Combining descriptors in pairs allows the identification of coincident co-descriptors for each document. The automatic method is set by the minimum number of co-descriptors to identify the relationship between a thesis and an article published by the same author.

A binomial number can determine the total number of descriptor combination pairs for a document. Table 2 presents the descriptors for a thesis and an article by the same author, illustrating how co-descriptors are identified.

Six coincident descriptors can be identified to compare the thesis and article [highlighted in bold with a gray background in Table 2]. The number of combined descriptor pairs can be calculated using the binomial number:

$$\binom{6!}{2!} = \frac{6!}{2!4!} = 15.$$

For six coincident descriptors, 15 co-descriptors can be identified.

Table 2: Example of thesis and article descriptors to determine co-descriptors coincident

Thesis title	Thesis descriptors	Article title	Article descriptors
"Evaluation of some essential and toxic element contents in children and elderly diet, by neutron activation analysis"	Aged adults	"Determination of various nutrients and toxic elements in different brazilian regional diets by neutron activation analysis"	Brazil
	Biological materials		Elements
	Children		Diet
	Diet		Food
	Intake		Multi-element analysis
	Multi-element analysis		Neutron activation analysis
	Neutron activation analysis		Nutrients
	Nutrients		Radiochemical separation
	Radiochemical separation		Semimetals
	Toxic elements		Trace amounts
	Trace amounts		Transition elements

To establish a minimum threshold number of coincident co-descriptors to automatically indicate that an article correlates to its thesis, we individually distributed a survey to a random sample of 128 authors by electronic mail, asking them to indicate the level of relationship of their articles and theses on a four-point scale (ranging from 1 for a "strong" relationship to 4 for "no" relationship). The questions presented were elaborated in a detailed way to minimize misunderstanding and the possibility of author bias [Appendix 2].

The authors' opinions were important for validating the automatic method. Their responses allowed us to establish the minimum number of coincident co-descriptors and to establish a relationship threshold. We obtained 100 responses (response rate of 79%) indicating the levels of relationship of the 397 articles within the established time range. We were subsequently able to make inferences about the sampling errors and estimate a confidence interval for the percentage of related articles.

RESULTS AND DISCUSSION

The results for each initial research question are presented below, including a discussion of each.

What is the level of doctoral researcher productivity during the thesis elaboration process and subsequent years?

Because a thesis must be completed within 5-year in Brazil, an interval of 5-year before and after the thesis, defense was selected for collecting articles, with the additional condition that authors had published at least one article within that time.

A total of 401 (83.37%) thesis authors published at least one article 5-year before or after the thesis defense. This result approximated the results reported by Lee (2000)[14] for chemistry (86%). However, it differed significantly from studies conducted in other areas: Psychology 59%, American literature 35% (Lee, 2000),[14] medicine 34% (Frkovic *et al*., 2003),[8] information science 66.6% (Anwar, 2004),[13] and education 36.7% (Mallette, 2006).[15]

The authors of the theses published 2,211 articles, with a mean number of articles published of 5.51 between 1977 and 2009 according to the thesis defense year. When considering unproductive authors as well, the mean number of articles published decreases to 4.13.

The percentage of productive theses [indicated by the dotted line in Figure 1] showed a stable trend beginning in the 1990s and a decrease in the subsequent period, followed by an increase between 2001 and 2004.

The mean number of articles by a thesis author increased during the evaluated period, especially when considering only productive theses.

The research productivity of thesis authors was also analyzed. Data were grouped into 5-year period and analyzed. The final 5-year period ended in 2004, ensuring 5-year of publications [Table 3]. The number of theses defended significantly increased in each 5-year period. Only theses that were productive during the period of evaluation were included in the analysis.

The thesis distributions for the number of articles published in the 5-year period of 1990-1994 and 2000-2004 were similar; the data reflect thesis authors who published one to three articles. In the 5-year period of 1995-1999, many authors published two or three articles, whereas, in the 5-year period of 1985-1989, many authors published one or two articles.

These results differed from those reported by Frkovic *et al*. (2003)[8] in a study conducted for medicine (i.e., 96% of theses generated only one article, 3% generated two articles, and 1% generated three or more). The different fields predictably showed different trends.

The median number of articles by 5-year period had its highest value in the last 5-year period, at close to five articles per thesis [Table 3].

COULD A CO-OCCURRENCE-BASED AUTOMATIC METHOD EXPRESS THE AUTHORS' OPINIONS ABOUT THE RELATIONSHIP BETWEEN THEIR PUBLISHED ARTICLES AND THESIS?

The need to identify articles related to the doctoral research of the 401 authors who published at least one article encouraged the development of the proposed automatic method, which was validated by the survey. We sampled

Table 3: Productivity levels of thesis authors

Thesis		Number of articles published				Median number of articles
Period of defense	Frequency	1%	2%	3%	More %	
1985-1989	14	28.57	28.57	14.29	28.57	1.75
1990-1994	32	15.63	9.38	12.50	62.50	4.20
1995-1999	76	10.53	17.11	18.42	53.95	3.21
2000-2004	109	15.60	7.34	12.84	64.22	4.79

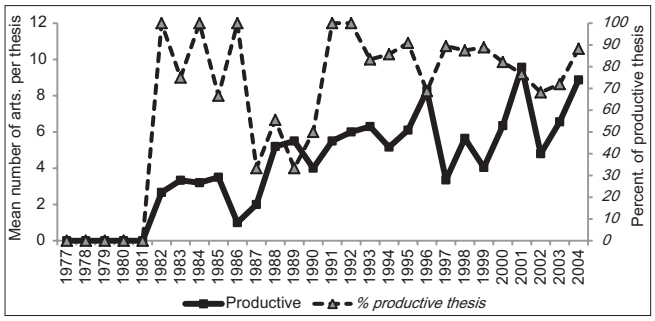

Figure 1: Mean number of articles published yearly by authors during the study period, considering only productive theses and the percentage of productive theses

128 authors and received 100 successful responses (24.9% of 401), resulting in 397 articles (17.9% of 2,211). The relationship levels stated by the authors and the number of coincident co-descriptors between the thesis and the articles were compared [Table 4].

The authors' responses that indicated a relationship (1 for "strong" and 2 for "medium") mainly concerned articles with three or more coincident co-descriptors, allowing the automatic methodology to be based on this number of co-descriptors (dotted gray line).

Author responses of 3 or 4 indicated a weak or no relationship, respectively. Table 4 shows a dotted gray line that maximizes the agreement between authors' opinions and the number of coincident co-descriptors. For 211 articles (53.1% of the sample), there was agreement about the presence of a relationship (i.e., the author said "yes," and there were three or more coincident co-descriptors). For 134 articles (33.8%), there was agreement about the absence of a relationship (i.e., the author said, "no," and there were one or zero coincident co-descriptors). Hence, there was disagreement for a total of 52 articles (13.1% of the sample).

The automatic method, thus expresses the author's opinion with acceptable precision (86.9%). Due to the sample size (397 articles), the sampling error related to the percentage estimate of related articles must be calculated.

To obtain this information reliably, a confidence interval for the sample was calculated as shown below.

Table 4, depicting the relationship between articles and theses, verifies the agreement between an author's opinion and the automatic method when

Table 4: Comparative results between the automatic methodology and authors opinion

Number of coincident co-descriptors	Authors' relationship (survey)				Total
	Yes		No		
	1	2	3	4	
0	3	2	18	93	116
1	4	3	14	9	30
3	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

Both declare a relationship (e.g., the author answers "1" or "2" and there are at least 3 coincident co-descriptors) or

Both declare no relationship (e.g., the author answers "3" or "4" and there are fewer than 3 coincident co-descriptors).

The sampling frame consisted of the list of all 401 authors, from which a random sample of size 128 was selected, with 100 successful responses (primary sampling unit). Every article of each author sampled was considered as an element (totaling 397 articles), creating a two-stage element sampling (Bolfarine and Bussab, 2005).[38] The authors were randomly selected and asked to identify which of their articles was related to the doctoral thesis.

To calculate the proportion of agreement in the sample, we considered the following:

$a_{ji} = 1$, if the j^{th} author's i^{th} article exhibited the agreement between the author's opinion and the automatic method and

$a_{ji} = 0$ otherwise,

where $n = 100$, $j = 1, 2, \ldots, n$, $i = 1, 2, \ldots, b_j$ such that $\sum_{j=1}^{n} b_j = 397$ and b_j is the number of articles published by j^{th} author.

Then,

$c_j = \sum_{i=1}^{b_j} a_{ji}$ is the j^{th} author's number of articles that agree with the automatic method. The proportion of agreement is given as

$$p = \frac{\sum_{j=1}^{n} c_j}{\sum_{j=1}^{n} b_j} = \frac{345}{397} = 0.8690$$

The respective standard error is

$$\sigma_p = \sqrt{\frac{1}{n \times (n-1) \times \bar{b}^2} \sum_{j=1}^{n} (c_j - p \times b_j)^2}$$

$$\left(\text{where } \bar{b} = \frac{\sum_{j=1}^{n} b_j}{n} \right)$$

$$= \sqrt{\frac{1}{100 \times (100-1) \times 3.9700^2} .55.4127} = 0.0375 .$$

The 95% confidence interval will be

$$\left[p - z_{95\%} \times \sigma_p ; p + z_{95\%} \times \sigma_p \right]$$
$$= \left[0.8690 - 1.96 \times 0.0375 ; 0.8690 + 1.96 \times 0.0375 \right]$$
$$= \left[0.7954 ; 0.9426 \right]$$

The sampling error was approximately 7.36%, which resulted in a 95% confidence interval between approximately 79.5% and 94.3%. Of all samples that could be selected from the author population, 95% would lie inside this interval, according to this agreement proportion.

The degree of agreement between an author's opinion and the automatic method had a low probability of being less than 79.5%, thus validating the automatic method.

WHAT PERCENTAGE OF ARTICLES HAS A RELATIONSHIP TO THE THESIS?

After defining the threshold number of coincident co-descriptors to confirm the relationship, the automatic methodology was applied to the entire article corpus (2,211 articles). This process allowed us to analyze the percentage of related published articles compared to the total number of articles published by each author in that year. The median was used to represent the group of authors who defended their theses in the same year. Figure 2 shows that production is directly related to doctoral research in most years (i.e., the median percentage of articles related to theses, according to the defense year, is over 50% in most years). Another aspect to note is a lower oscillation over the period, which denotes stabilization from 2000, close to 75%.

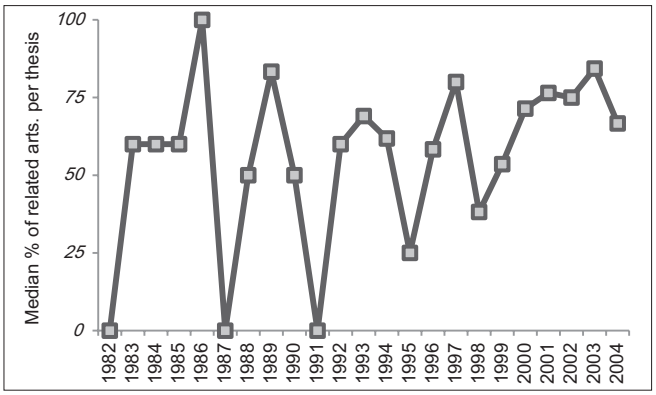

Figure 2: Distribution of the median percentage of related articles per thesis according to thesis defense year: 1982-2004

Another aspect that was analyzed was the lag time for publishing an article, considering the publication years for a thesis and its author's published articles [Figure 3]. The number of articles, which is more significant after 1987, indicates that prior to 1997, nonrelated articles took a shorter median time to be published after the thesis defense year. From 1998-2004, this relationship was inverted in almost every year, showing a strong decrease in the time required to publish related articles after 2001. It is important to mention that in this figure, both series are interrupted in the period due to a lack of articles published in these years.

Finally, we examined the publication of articles related to the thesis defense year, looking for differences between related and nonrelated articles and considering doctoral research as well as the influence of a supervisor's presence on the related article distribution [Figure 4]. The frequency distribution of the articles indicates that related articles were published mainly after the thesis defense. The results showed that 31.4% of related articles and 37.6% of nonrelated articles were published before the thesis defense, indicating a statistically significant difference (P = 0.0013). Most related articles (65.1%) were published in a period between 1-year before (−1) and three years after (+3) the thesis defense, indicating a clear concentration of the results near the end of the doctoral research, with a peak in the first year after the defense (+1) and decreasing subsequently. This concentration was not observed in the distribution of the nonrelated articles, reinforcing the validity of the automatic method and showing that 51.2% of the articles were published between these years (−1 and + 3).

Figure 4 shows data for the way a supervisor's presence affected article publication. Considering all 2,211 articles,

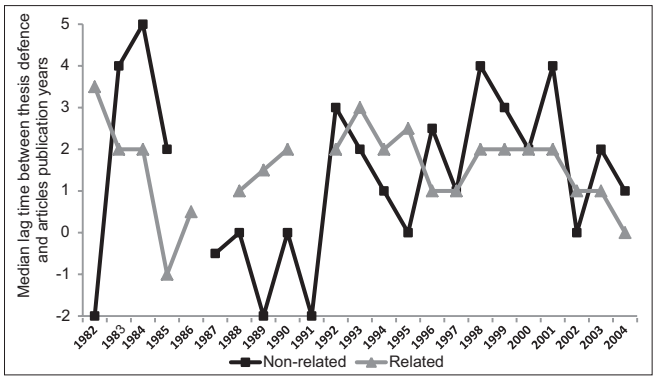

Figure 3: Distribution of articles according to thesis defense year: Comparing the relationship of the articles to the doctoral research in the median lag time between the thesis defense and article publication years (1982-2004)

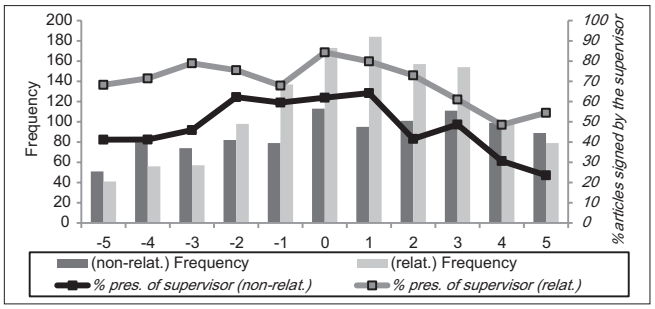

Figure 4: Distribution of articles according to the lag time between the thesis defense and article publication years: Comparing the relationship of the articles to the doctoral research using article frequency and the rate of the supervisor's presence in article authorship

the supervisor was present in 70.60% of the related articles and 47.69% of the nonrelated articles.

Another aspect that was considered was that the supervisor's presence was prevalent in related articles published in the same year as the thesis defense, decreasing uniformly afterward. During doctoral studies (−5 to − 1), the supervisor's presence was much more frequent, especially three years before the defense.

CONCLUSIONS

One main aspect addressed in this paper is the possibility of using an automatic methodology to identify thesis productivity compared to publishing. A co-word analysis (in this case, co-descriptors) is demonstrated as a satisfactory methodology to identify relationships between theses and articles, because the percentage of agreement between the automatic method and authors' opinions was 86.9%. Moreover, the methodology represents an alternative to sequential and similar studies avoiding manual matching. It is possible to identify the thesis productivity level for a specific knowledge area, and these results answer our second research question: Could a co-occurrence-based automatic method express authors' opinions about the relationship between their published articles and theses? The authors' opinions are important to validate the proposed model because authors are appropriate people to evaluate their production. These opinions allowed us to establish a minimum threshold to consider this relationship and to extend it to the entire article corpus.

Much of the method's success depends on the descriptors' quality, database availability, and data confidence level. In this study, it was observed that the quality of the indexing process, or the descriptors assigned to the documents, contributed to identifying the relationship between theses and articles.

The results presented in this study should be used with caution. The interpretation and validation processes conducted by experts who are familiar with the studied fields' structures and dynamics should be retained, thus avoiding indiscriminate data use. We note the applicability of the method to information units with a similar structure, such as those in which the data organization and processing standards use indexing with a controlled vocabulary. The medical field, which uses Medical Subject Headings (MeSH, elaborated by the National Library of Medicine), is an example.

The strict use of descriptors may be considered as a limitation of the proposed model, although further studies could use author-assigned keywords, natural words extracted from the abstracts, or paper titles to automatically identify the relationship between articles and doctoral research.

The results of the data analysis indicate that the median lag time between the thesis defense and article publication has been shortened. This can be considered as a positive development because it shows a decreasing trend in the lag between knowledge production and publication.

The characteristics of the distribution of the related articles differ after applying the proposed automatic method, and they indicate differences with nonrelated article distribution, demonstrating an expectedly higher number of publications after thesis writing.

The characteristics found in this analysis are related to a specific document corpus produced in the highly specialized field of nuclear science. Similar analyses in other fields would contribute to comparing and consolidating the proposed model. Results such as those presented here are useful for science policy managers at fellow institutions because this information can support and confirm the decisions adopted, determine course corrections, or even contribute to establishing guidelines that reflect the reality of graduate programs.

REFERENCES

1. Igami MZ. Construction of scientific production indicators based on scientometric analysis of IPEN dissertations ans theses. PhD thesis, IPEN/CNEN-SP, São Paulo; 2011. Available from: http://www.teses.usp.br. [Last accessed on 2011 Apr 15].
2. de Solla Price DJ. Little Science, Big Science. New York: Columbia University Press; 1963.
3. Andersen JP, Hammarfelt B. Price revisited: On the growth of dissertations eight research fields Scientometrics 2011;88:371-83.
4. Yaman H, Atay E. PhD theses in Turkish sports sciences: A study covering the years 1988-2002. Scientometrics 2007;71:415-21.
5 Breimer LH, Nilsson TK. A longitudinal and cross-sectional study of Swedish biomedical PhD processes 1991-2009 with emphasis on international and gender aspects. Scientometrics 1996;85:401-14.
6. Varshney LR The Google effect in doctoral theses. Scientometrics 2012;92:785-93.
7. Salmi LR, Gana S, Mouillet E. Publication pattern of medical theses, France, 1993-98. Med Educ 2001;35:18-21.
8. Frkovic V, Skender T, Dojcinovic B, Bilic-Zulle L. Publishing scientific papers based on Master's and Ph.D. theses from a small scientific community: Case study of Croatian medical schools. Croat Med J 2003;44:107-11.
9. Ramos PS, Furtado EC, Carvalho ER, Campos MO, de Souza DV, de Almeida LD, et al. Do dissertations and theses generate scientific articles? An analysis based on three physical education programs. Braz J Biomotricity 2009;3:315-24.
10. Sacardo MS, Hayashi MC. Bibliometric balance of scientific production in the areas of physical education and special education deriving from dissertations and theses. RBPG Rev Bras Pós-Graduação 2011;8:111-35.
11. Younes RN, Deheinzelin D, Birolini D. Graduate education at the faculty of medicine of the University of Sao Paulo: Quo vadis? Clinics (Sao Paulo) 2005;60:6-8.
12. Arriola-Quiroz I, Curioso WH, Cruz-Encarnacion M, Gayoso O. Characteristics and publication patterns of theses from a Peruvian medical school. Health Info Libr J 2010;27:148-54.
13. Anwar MA. From doctoral dissertation to publication. A study of 1995 American graduates in library and information science. J Libr Inf Sci 2004;36:151-7.
14. Lee WM. Publication trends of doctoral students in three fields from 1965-1995. J Am Soc Inf Sci 2000;51:139-44.
15. Mallette LA. Publishing Rates of Graduates Education Ph.D. and Ed.D. Students: A Longitudinal Study of University of California Schools. PhD thesis, Pepperdine University; 2006. Available from: http://www.gradworks.umi.com/32/39/3239922.html. [Last accessed on 2012 Set 18].
16. Larivière V. On the shoulders of students? The contribution of PhD students to the advancement of knowledge. In: 11th International Conference on STI "Creating Value for Users." Leiden, Netherlands; 2010. Available from: socialsciences.leiden.edu/cwts/news/11th-international-conference-on-sti-cwts.html. [Last accessed on 2010 Dec 15].
17. Barrier J, Musselin C. The rationalization of academic work and careers: Ongoing transformations 'of the profession and policy challenges. In: Kehm B, Huisman J, Stensaker B, editors. The European Higher Education Area: Perspectives on a Moving Target. Rotterdam: Sense Publishers; 2009. p. 2003-221.
18. Breimer LH. Swedish biomedical PhD examination: An international forum and a proposed procedure for Europe. Scientometrics 2010;83:583-7.
19. Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). Plataforma Lattes. Brasília: CNPq; 1990. Available from: http://www.plataformalattes.com.br. [Last accessed on 2013 Aug 15].
20. Lane J. Let's make science metrics more scientific. Nature 2010;464:488-9.
21. Leite P, Mugnaini R, Leta J. A new indicator for international visibility: Exploring Brazilian scientific community. Scientometrics 2011;88:311-9.
22. Mena-Chalco JP, Cesar RM Jr. ScriptLattes: An open-source knowledge extraction system from the Lattes platform. J Braz Comput Soc 2009;15:31-9. Available from: http://www.scielo.br. [Last accessed on 2013 Mar 14].
23. van Raan AF. Advanced bibliometric methods to assess research performance and scientific development: Basic principles and recent applications. Res Eval 1993;3:151-66.
24. Ding Y, Chowdhury GG, Foo S. Bibliometric cartography of information retrieval research by using co-word analysis. Inf Process Manag 2001;37:817-42.
25. He Q. Knowledge discovery through co-word analysis. Libr Trends 1999;48:133-59.
26. Fernandez LM, Guerrero-Bote VP, Moya-Anegon F. Co-word based thematic analysis of renewable energy. Scientometrics 2013. [April 2nd published on-line].
27. Zong YW, Shen HZ, Yuan QJ, Hu XW, Hou ZP, Deng SG. Doctoral dissertations of library and information science in China. Scientometrics 2013;94:781-99.
28. Lui GY, Hu JM, Wang HL. A co-word analysis of digital library in China. Scientometrics 2012;91:203-17.
29. An Y, Wu QQ. Co-word analysis of the trends in stem cells field based on subject heading weighting. Scientometrics 2011;88:133-44.
30. Yang Y, Wu M, Cui L. Integration of three visualizations methods based on co-words analysis. Scientometrics 2012;90:659-67.
31. Callon M, Law J, Rip A. Mapping the Dynamics of Science and Technology. London: McMillan Press; 1986.
32. Turner WA, Chartron F, Michelet B. Packaging information for peer review: New co-word analysis technique. In: van Raan AF, editor. Handbook of Quantitative Studies of Science and Technology. North-Holland: Elsevier; 1988. p. 291-323.
33. Silva M dos R, Fujita MS. Indexing practice : Analysis of the theoretical and methodological evolution trends. Transinformação 2011;16:133-61. Available from: http://www.sicelo.org. [Last accessed on 2012 Aug 15].
34. Whittaker J, Courtial JP, Law J. Creativity and conformity in science: Titles,keywords and co-words analysis. Soc Stud Sci 1989;19:473-96.
35. Law J, Whittaker J. Mapping acidification research: A test of the co-word method. Scientometrics 1992;23:417-61.
36. Gil-Leiva I, Alonso-Arroyo A. Keywords given by authors of scientific articles in database descriptors. J Am Soc Inf Sci Technol 2007;58:175-1187.
37. Wang YZ, Li G, Li CY, Li A. Research on the semantic based co-word analysis. Scientometrics 2012;90:855-75.
38. Bolfarine H, Bussab WO. Elementos de Amostragem. São Paulo: Edgard Blücher; 2005.

How to cite this article: Igami MP, Bressiani JC, Mugnaini R. A new model to identify the productivity of theses in terms of articles using co-word analysis. J Sci Res 2014;3:3-14.

Source of Support: Nil, **Conflict of Interest:** None declared

Appendix 1: One author's information as an example of the database used

Thesis of one author						
IPEN-doc code	**Author**	**Title**	**Descriptors**	**Year**	**Supervisor**	**Subject category**
11132	Calvo, W.A.P.	Development of an Irradiation System for a Small Size Continuous Run Multipurpose Gamma Irradiator	Irradiation Plants; Irradiation Devices; Gamma Radiation; Cobalt 60; Gamma Sources; Modular Structures; Radiation Sources; Industrial Plants	2005	Andrade E Silva, L.G.	S43
Articles by the same author						
IPEN-doc code	**Author (s)**	**Title**	**Descriptors**	**Year**	**Journal title**	
14922	Calvo, W.A.P.; Hamada, M.M.; Sprenger, F.E.; Vasquez, P.A.S.; Rela, P.R.; Martins, J.F.T.; Pereira, J.C.S.M.; Omi, N.M.; Mesquita, C.H.	Gamma-Ray Computed Tomography Scanners For Applications In Multiphase System Columns	Porosity; Gamma Radiation; Cobalt 60; Radiation Sources; Performance; Materials Testing; Radiation Detectors; Computerized Tomography	2009	Nukleonika	
14929	Calvo, W.A.P.; Rela, P.R.; Napolitano, C.M.; Kodama, Y.; Omi, N.M.; Costa, F.E. Da; Andrade E Silva, L.G.	Development Of An Irradiation System For A Small Size Continuous Run Multipurpose Gamma Irradiator	Irradiation Plants; Irradiation Devices; Gamma Radiation; Cobalt 60; Gamma Sources; Modular Structures; Brazil; Radiation Sources; Industrial Plants	2009	Nukleonika	
9450	Rela, P.R; Calvo, W.A.P.; Springer, F.E; Omi, N.M; Costa, F.E; Vieira, J.M; Andrade E Silva, L.G.	Desenvolvimento E Implantação De Um Irradiador Multipropósito De Cobalto-60 Tipo Compacto	Irradiation Plants; Cobalt 60; Gamma Radiation; Ionizing Radiations; Modular Structures	2003	Revista Brasileira De Pesquisa E Desenvolvimento	
12830	Gonçalves, J.A.C.; Botelho, S.; Pascholati, P.R.; Ridenti, M.A.; Fraga, M.M.F.R.; Camara, J.R.; Calvo, W.A.P.; Bueno, C.C.	Activity Measurements Of Sup (192) Ir Solid Sources Using A Well-Type Ionization Chamber	Iridium 192; Technetium 99; Ionization Chambers; Activity Levels; Decay; Sealed Sources	2007	Nuclear Instruments And Methods In Physics Research. Section A	

APPENDIX 2: QUESTIONNAIRE SENT TO THE AUTHORS

Dear author:

I am working in a search about the thesis productivity, concerning the number of articles published, and presented in the Graduation course of this institution, from 1977 to 2009. This makes part of my PhD work, oriented by Dr. XXX.

For this reason I ask your collaboration in order to indicate if the following articles have relation with your thesis "Neutronic evaluation of metallic fueled and lead cooled nuclear reactor" presented in year 2000.

To answer this survey please, click first in "answer" and after mark with X in the brackets space the article relationship, after this click in "sent".

Note that the search comprehension time of articles is limited to five years after or before de thesis presentation.

Thanks for your collaboration.

Article 1 of 2

Santos A, Nascimento JA. An integral lead reactor concept for developing countries. Nucl Technol 2002;140:1-22.

1. (X) the article publishes results obtained along or after the thesis elaboration, this is to say it has a high relationship with the thesis.

2. ( ) the article publishes a great deal of the results obtained with the thesis, although, applied or complemented with the results of other researches, this is to say, it has a medium relationship with the thesis.

3. ( ) the article publishes results of other researches and only uses the thesis experience , this is to say, it has a small relationship with the thesis.

4. ( ) the article publishes results of other researches with no relationship with the thesis.

Article 2 of 2

Santos A, Nascimento JA. An integral metallic-fueled and lead-cooled reactor concept for the 4th generation reactor. J Nucl Sci Technol 2002;Suppl 2:1081-4.

1. (X) the article publishes results obtained along or after the thesis elaboration, this is to say it has a high relationship with the thesis.

2. ( ) the article publishes a great deal of the results obtained with the thesis, although, applied or complemented with the results of other researches, this is to say, it has a medium relationship with the thesis.

3. ( ) the article publishes results of other researches and only uses the thesis experience this is to say, it has a small relationship with the thesis.

4. ( ) the article publishes results of other researches with no relationship with the thesis.