BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **22**, No 1
Sofia • 2022 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2022-0006

# Citation and Similarity in Academic Texts: Colombian Engineering Case

*Marco Aguilera-Prado*[1], *Octavio Salcedo*[2], *Eduardo Avendaño Fernández*[3]

[1]*Vice Rectorate for Research, Universitaria Agustiniana. Bogotá, Colombia*
[2]*Faculty of Engineering, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia*
[3]*Faculty of Engineering, Universidad Pedagógica y Tecnológica de Colombia, Tunja, Colombia*
*E-mails: marco.aguilera@uniagustiniana.edu.co osalcedo@udistrital.edu.co eduardo.avendano@uptc.edu.co*

***Abstract*: *This article provides the results of a citation determinants model for a set of academic engineering texts from Colombia. The model establishes the determinants of the probability that a text receives at least one citation through the relationship among previous citations, journal characteristics, the author and the text. Through a similarity matrix constructed by Latent Semantic Analysis* (*LSA*), *a similarity variable has been constructed to capture the fact that the texts have similar titles, abstracts and keywords to the most cited texts. The results show: i) joint significance of the variables selected to characterize the text; ii) direct relationship of the citation with similarity of keywords, published in an IEEE journal, research article, more than one author; and authored by at least one foreign author; and iii) inverse relationship between the probability of citation with the similarity of abstracts, published in 2016 or 2017, and published in a Colombian journal.*

***Keywords*: *Latent semantic analysis, text similarity, citation determinants, bibliometrics.*

## 1. Introduction

Several works on the dynamics of the production and citation of Colombian scientific literature suggest that it has been concentrated among certain authors and institutions as related to articles in national journals, with intermediate international collaborations in authorship and low participation of national authors within the bulk of the academic literature worldwide [1-4].

For the specific case of engineering, between 1997 and 2009, Colombia published 419 articles (scientific and review) in Web of Science, equivalent to 18% of the 2,297 Latin America-affiliated articles in that period. The largest number of articles with Colombian origin came from National University (205; 49%), the

University of Antioquia (94; 22%), the University of Valle (58; 14%), the University of the Andes (25; 6%), and the Industrial University of Santander (19; 5%). In terms of quality, Pontifical Bolivarian University presented the highest number of citations per document, and the Pedagogical and Technological University of Colombia obtained the highest weighted and relative impact factor. Likewise, most of the articles published in Latin American journals are in English but not in Colombian journals, and there is a concentration of articles in Dyna, the Revista Facultad de Ingeniería (School of Engineering Journal) and Revista Ingeniería e Investigación (Engineering and Research Journal) [5].

Likewise, an analysis of 2,471 Colombian engineering articles published between 2008 and 2017 in Colombian engineering journals in Scopus (Dyna, Ingeniería e Investigación, Ingeniería y Universidad and Revista de la Facultad de Ingeniería) showed that: i) Dyna has contributed a large part of the articles published (42%) and cited (55%); ii) almost half of the articles are published in Spanish (47%); iii) 35% of the articles come from the National University of Antioquia, the University of Valle and the Industrial University of Santander; and iv) there is at least one citation for 42% of the articles [6].

Some of the explanations for this pattern of low international impact may lie in the topics addressed [7, 8], which for the Colombian case could be limited to local research results, applications in which the international academic literature no longer writes about or shows interest in or unconsolidated results from scattered research studies. Along these lines, this article investigates the relationship between the topics of Colombian engineering publications and the probability of being cited; a probabilistic model with traditional variables of citation determinants [7] is built, and through Latent Semantic Analysis (LSA), a similarity vector is added to measure the similarity with the most frequently cited texts as a way to investigate whether the texts are similar and whether that influences citation. This paper uses determinants modelling instead of neural networks or similar machine learning tools because the aim is to explain how different variables affect citation and calculate marginal effects on citation of those specific characteristics of the selected texts.

LSA has been used for identifying topics, authors and networks in traditional medical journals [9] or abstracts [10], highlighting the grouping of 1,958,125 scientific texts published between 2004 and 2008 in the MEDLINE and Elsevier databases through similarity matrices that identifies groups of between 764 and 1,827 articles in 23,831 clusters of topics, based on their abstracts [11]. In engineering, the mapping of 3,207 abstracts related to the topic *Operations Management Research* between 1980 and 2015 in five representative journals is highlighted; the results have shown that: i) over the decades, the topics of greatest relevance varied and expanded in number; ii) the topics had their own rise and fall dynamics in terms of number of publications; and iii) the only topic that continuously grew throughout the period was *Supply Chain Management* [12]. For this article, LSA is used because it allows the comparison of the most cited article with the others and get a similarity measure. Others Natural Language Processing methods (Latent Dirichlet Allocation, co-words) calculate similarity comparing different sets of texts with each other through words counting.

## 2. Methods and materials

To model the number of citations, the traditional relationship expressing article citations $(Z_i)$ as the sum of three sets of variables has been used: *Paper* $(P_i)$, *Journal* $(J_i)$ and group of Authors$(A_i)$, plus a stochastic error $(\varepsilon_i)$ [7]. Formally, the relationship is as follows:

(1) $$Z_i = \beta_\text{o} + \sum_{i=1}^{m} \beta_i \cdot P_i + \sum_{i=m+1}^{h} \beta_i \cdot J_i + \sum_{i=h+1}^{w} \beta_i \cdot A_i + \varepsilon_i.$$

Information on the variables has been taken from Scopus, which provided metadata of 12,501 scientific texts of Colombian origin in engineering between 2009 and 2018. The selected working variables are shown in Table 1.

Table 1. Factors that determine the number of citations of an academic publication

| Type of variable | Variable | Form | Source |
|---|---|---|---|
| Paper | Title | Text | Metadata |
| | Author keywords | Text | Metadata |
| | Abstract | Text | Metadata |
| | Date of publication | Year of publication | Metadata |
| | Type of text | Dichotomous, *article* or *conference paper* | Metadata |
| Journal | Origin of the journal | Dichotomous, Colombian or foreign journal | Metadata |
| | Origin of the journal | Dichotomous, IEEE journal or other | Metadata |
| Author | Number of authors | Number of authors | Constructed from metadata |
| | Country of origin of the authors | Dichotomous, international collaboration or not | Constructed from metadata |
| Independent | Number of citations | Dichotomous, citation or without citation | Constructed from metadata |

For the text variables (*title, abstract,* and *keywords*), a similarity matrix has been constructed to identify the similarity through the selected texts. The idea behind this similarity is that texts with similar titles, abstracts and keywords should have similar citations; in other words, resembling the most cited texts should have positive effects on the citation of an article.

In this sense, these variables have been transformed by developing a similarity vector constructed by LSA, a computer tool that allows analyzing the relationships of meaning between large volumes of text, overcoming problems of synonymy, toponymy and repetition of terms without semantic value in a specific *corpus*[1] [13]. LSA transforms a co-occurrence matrix of terms into smaller ones that can be more easily interpreted and that functionally allow the emergence of meanings absent in the co-occurrence matrix and in the representation of each text isolated from the *corpus*.

This transformation is based on the idea that a text fragment can be represented by a linear equation where the meaning is the sum of the meanings of the words that compose it. The linear equation is constructed using Singular Value Decomposition (SVD), which recognizes that: i) the meaning of the words depends on the context; ii) there are relationships of semantic similarity in contextual use [14].

[1] The original set of texts that are reduced to matrices and that are configured in a matrix with as many rows as texts are taken from the columns; one for identification and one with the texts to be analyzed.

SVD decomposes the information contained in an initial matrix into three matrices with particular aspects of the characteristics of the terms contained in that matrix. The decomposition reduces the dimensions with which the *corpus* is described in the initial word matrix by discarding that information that contributes little to its semantic identification, resulting in a more compact and computable matrix representation of the semantic space. Formally, the rectangular matrix $X$ can be decomposed into three matrices, with $T_o$ and $S_o$ being orthonormal columns and $D_0'$ diagonal. Formally, the representation is as follows [15, 16]:

(2) $$X = T_0 S_0 D_0',$$

where the column vectors of $T_0$ are called left singular vectors, the vectors of $D_0$ are called right singular vectors and $S_0$ is the diagonal matrix of singular values. The resulting multiplication of the matrices is similar but not identical to the initial word matrix; thus [15]

(3) $$X \approx \hat{X} = TSD',$$

where $\hat{X}$ contains information not available in the first matrix, which is called latent information, and accounts for the *corpus* that describes the initial matrix. In turn, the information in $T$, $D$ and $S$ constitutes a vector space where each vector represents the meaning that the words activate within the set of texts from which they have been selected, and their interpretation depends on the relationship with the other vectors [14].

However, the benefit of SVD is that the decomposition results in matrices with ranges lower than the initial matrix once the range (the dimensions) is limited to that $k$ of the diagonal matrix after the lower values have been eliminated, so that

(4) $$\hat{X} = T_k S_k D_k'.$$

Thus, $\hat{X}$ can be interpreted as the set of inferred probabilities that a word or term occurs in a document and that a document contains a term because in practice, the starting point of LSA (the initial matrix) is a matrix that yields global and local weights from a co-occurrence matrix of terms, which is transformed by SVD.

Specifically, for this research, a vector has been constructed to identify the similarity of the articles with others and then crossing it with other variables of citation. As initial text, the title, keywords and abstract have been used. The decomposition and generation of the similarity vector have been performed in STATA.

The initial matrix has been constructed by creating a bag-of-words with a determined number of terms ($J$) from the reference texts that have been transformed into a matrix $X$ of relative frequency, where each input $x_{i,j}$ is calculated as follows [17]:

(5) $$x_{i,j} = \left[1 + \log\, f_{i,j}\right] \times \left[\log \tfrac{1+D}{1+d_j} + 1\right],$$

where $f_{i,j}$ is the frequency of term $j$ in document $i$, and $D$ is the total number of documents to be analyzed; $d_j$ is the number of documents in which term $j$ appears.

Applying SVD transforms the matrix $X$ of range $r$ into three matrices, as follows [17]:

(6) $$X = T_{D\times r} S_{r\times r} D_{r\times J}'.$$

The reduction in the range is based on the eigenvalues of $S_{r\times r}$; for this, rows and columns with the lowest eigenvalues are eliminated, resulting in a matrix $S_{k\times k}$ that modifies the previous approach as follows [17]:

$$\hat{X} = T_{D\times k} S_{k\times k} D'_{k\times J}. \tag{7}$$

For the calculation of the similarity between two documents $d_1, d_2$, the cosine similarity is used, taking advantage of each document input vector $\delta$ represented in $\hat{X}$. The calculation is [17]:

$$\mathrm{Sim}_{\cos} = \frac{\sum_{k=1}^{K}(\delta_{d1,k}\times\delta_{d2,k})}{\sqrt{\sum_{k=1}^{K}\delta_{d1,k}}\times\sqrt{\sum_{k=1}^{K}\delta_{d2,k}}}. \tag{8}$$

This calculation allows the construction of the vector of similarities between the articles within the workspace. This vector contains the number of articles of $D$, published after $d$, that have a similarity to a given number $\mathrm{Sim}_{\cos} > n$ (in this case 0.75).

With the variables arranged as numerical values, including text variables, a probabilistic model has been constructed to examine its significance on a text being cited and to see which variables contributed the most to citation.

## 3. Results

The set of working texts includes 10,095 – *articles* (6,040; 59.8%) and *conference paper* (4,055; 41.2%) published between 2013 and 2018. Most of the articles (6,040; 82.2%) have been published in foreign journals and to a lesser extent in national journals (1,076; 27.3%). In turn, with respect to the cited articles, 499 (53.6%) of those published in national journals received at least one citation, while 1,343 (27%) of those published in international journals were cited (Table 2).

Table 2. Texts cited by publication type

| Type of text/year | No citations | | | With citations | | | Grand total |
|---|---|---|---|---|---|---|---|
| | Foreign journals | National journals | Total | Foreign journals | National journals | Total | |
| **Article** | **1,343** | **499** | **1,842** | **3,621** | **577** | **4,198** | **6,040** |
| 2013 | 71 | 38 | 109 | 436 | 128 | 564 | 673 |
| 2014 | 98 | 72 | 170 | 496 | 134 | 630 | 800 |
| 2015 | 118 | 58 | 176 | 636 | 133 | 769 | 945 |
| 2016 | 197 | 79 | 276 | 690 | 84 | 774 | 1,050 |
| 2017 | 313 | 94 | 407 | 741 | 63 | 804 | 1,211 |
| 2018 | 546 | 158 | 704 | 622 | 35 | 657 | 1,361 |
| **Conference paper** | **2,531** | | **2,531** | **1,524** | | **1,524** | **4,055** |
| 2013 | 343 | | 343 | 279 | | 279 | 622 |
| 2014 | 363 | | 363 | 331 | | 331 | 694 |
| 2015 | 302 | | 302 | 312 | | 312 | 614 |
| 2016 | 297 | | 297 | 218 | | 218 | 515 |
| 2017 | 519 | | 519 | 253 | | 253 | 772 |
| 2018 | 707 | | 707 | 131 | | 131 | 838 |
| **Grand total** | **3,874** | **499** | **4,373** | **5,145** | **577** | **5,722** | **10,095** |

In that same period (2013-2018), 4,055 *conference papers* have been published in international journals that accounted for 4,611 citations in 1,524 texts. The total number of citations of these texts tended to decrease between 2013 (948) and 2018

(238), peaking in 2014 (1,107) and 2015 (1,127). In turn, the number of texts with at least one citation decreased from 279 in 2013 to 131 in 2018, with peaks in 2014 (331) and 2015 (312). Number of citations in that period are higher for *articles* (37,324) than *conference papers* (4,611) in total, as well as for the number of texts cited: 4,198 *articles* and 1,524 *conference papers* (Table 2).

The results of the logistict model being constructed for determining the relationship between the characteristics of the text (Table 1) and the probability that this text is cited establishes[2] the following:

1. Joint significance of the variables: similarity of keywords, similarity of titles, similarity of abstracts, year of publication, published in a Colombian journal, published in an IEEE journal, published in an open-access journal, research article, number of authors, and foreign author (which explains whether an engineering text with Colombian affiliation is cited and not significant at the 95% level for similarity of titles, published in 2014 or 2015 and published in an open-access journal) (Table 3).

2. Direct relationship of citation with similarity of keywords (0.5%); published in an IEEE journal (17%); research article (35%); more than one author (2%) and foreign author (12%) (Table 3).

3. Inverse relationship between the probability of the text being cited with similarity of abstracts (0.04%); published in 2016 (9.2%) or 2017 (16.7%) and published in a Colombian journal (14.7%) (Table 3).

For the constructed model, the similarity between keywords and abstract is significant (at 95%), and the relationship with citation probability is direct for the first variable and inverse for the second. The variable title is significant at 90% and directly related to the probability of being cited at least once. The combined effect of these variables on the probability of citation is less than 1%: keywords 0.57%; title 0.33%; abstract –0.04% (Table 3).

The results show that year of publication does not affect the number of citations received for the years 2014 and 2015, while the years 2016 and 2017 significantly and negatively impacted citation probability: –9.2% and –16.7%, respectively. Thus, for 2013, keeping anything else constant, articles published in a three-year window (2013, 2014 and 2015) have the same probability of receiving one or more citations, while more recent publications have a smaller probability of being cited (Table 3).

The model reveals that there is a significant difference in the citation probability for articles and conference papers; articles have a 35% higher probability of being cited. Likewise, there is a significant difference in citations between texts published in Colombian journals (Dyna, Ingeniería e investigación, Revista de la Facultad de Ingeniería, Ingeniería y Universidad) and those published in foreign journals. Text published in Colombian journals had a 14% lower possibility of being cited at least

[2] In general, the estimated coefficients of the logit models do not directly quantify the changes in probability given a unit change in the corresponding independent variable. The magnitude of the variation in probability depends on its original level and, therefore, of all and each of the regressors and their coefficients. Thus, while the sign of the coefficients perfectly indicates the direction of the change, the magnitude of the variation depends on the specific value that the density function takes on, which depends on the slope of said function at a given point.

once. In contrast, the texts published in IEEE journals are 17% more likely to be cited than those in other publications, and those with foreign authors are 12% more likely (Table 3).

Table 3. Marginal effects on the probability of being cited one or more times

| | | | | | | |
|---|---|---|---|---|---|---|
| Logistic regression | | | | Number of obs | = | 7.896 |
| Log likelihood = | –4401.5101 | | | LR chi$^2$(13) | = | 1645.41 |
| | | | | Prob> chi$^2$ | = | 0.0000 |
| | | | | Pseudo $R^2$ | = | 0.1575 |
| Variables | *dy/dx* | Stdandard Error | *z* | *P> z* | 95% confidence | Interval] |
| keywords | 0.0057 | 0.0022 | 2.6400 | 0.0080 | 0.0015 | 0.0099 |
| titles | 0.0033 | 0.0019 | 1.7500 | 0.0800 | –0.0004 | 0.0069 |
| abstracts | –0.0004 | 0.0002 | –2.3700 | 0.0180 | –0.0008 | –0.0001 |
| pub_year (2013) | | | | | | |
| 2014 | –0.0218 | 0.0159 | –1.3800 | 0.1690 | –0.0530 | 0.0093 |
| 2015 | –0.0097 | 0.0160 | –0.6100 | 0.5440 | –0.0411 | 0.0217 |
| 2016 | –0.0926 | 0.0166 | –5.5700 | 0.0000 | –0.1252 | –0.0600 |
| 2017 | –0.1668 | 0.0160 | –10.4500 | 0.0000 | –0.1981 | –0.1355 |
| 2018 | . | (N. E) | | | | |
| jour_col (foreign journal) | | | | | | |
| Colombian journal | –0.1462 | 0.0174 | –8.4200 | 0.0000 | –0.1802 | –0.1122 |
| IEEE (Not IEEE journal) | | | | | | |
| IEEE journal | 0.1680 | 0.0125 | 13.4600 | 0.0000 | 0.1436 | 0.1925 |
| open_access (No) | | | | | | |
| Open-access Yes | 0.0263 | 0.0144 | 1.8300 | 0.0680 | –0.0019 | 0.0546 |
| Article (*conference paper*) | | | | | | |
| Article | 0.3508 | 0.0108 | 32.3400 | 0.0000 | 0.3295 | 0.3720 |
| n_authors | 0.0198 | 0.0033 | 5.9500 | 0.0000 | 0.0133 | 0.0264 |
| auth_foreign (No) | | | | | | |
| Foreign author | 0.1160 | 0.0111 | 10.4300 | 0.0000 | 0.0942 | 0.1379 |

Regarding authors, the model has confirmed that the probability of being cited depends not so much on the number of authors but on whether authors have foreign affiliations. In the results, both variables are significant, but the contribution of the first variable to the citation rate is 2.0% per additional author, while for the second, it is 11.6% (Table 3).

## 4. Conclusion

Thus far, the results indicate that the relationship (similarity) between the texts, calculated using a similarity vector for the oldest article with the highest citation, is significant at 95% for abstracts and keywords and at 90% for titles; therefore, the hypothesis cannot be rejected. Additionally, resembling the most cited texts positively affects the probability of citation when similarity is measured by the similarity of key words and abstracts.

Notably, although these effects are significant, their contributions do not seem very relevant. The effect is less than 1%, hence the need to explore interactions between variables, for example the effect of similarity by year, understanding that the academic literature has periods of boom and bust in relation to citations [12], or the differentiated effect of the most cited text by year, by type (article, conference paper) and other types of models, such as neural networks, for which the inputs are the variables identified as significant and the layers allow controlling the effects for each relevant set in this first model: year of publication, origin of the journal, type of publication.

Importantly, the indirect effect of the abstract may be due to variations between years. That is, due to a possible quadratic effect (increasing in one period, decreasing in another), the impact of abstract similarity on citation rates may not be negative every year as a consequence of the fact that the texts have periods of citation booms in response to topics they address, and other periods when they are not cited [12]. Likewise, the relationship may be negative because of several topics that are cited differently from the one that has been taken as reference (dispersion). Then, it would be necessary to identify different reference texts to make the comparison and measure effects on different texts of various topics.

Because the LSA on which this work is based can be used for the construction of other types of models that aim to identify words (not semantic groups) [15, 13], on which the citation can be contrasted, it would be possible to construct a particular set of words that refer to specific topics on which a geographical or temporal space is assigned and evaluate how much having or not having those words in the title, keywords or abstract affects being cited, i.e., looking at the citation probabilities of texts that deal with certain previously established topics.

## References

1. Gregorio, O. Análisis bibliométrico y de calidad de la revista Signo y Pensamiento. – Signo y Pensamiento, Vol. **50**, 2007, No xxvi, pp. 22-32.
2. Rivera-Garzón, D. M. Caracterización de la comunidad científica de Psicología que publica en la revista Universitas Psychologica (2002-2008). – Universitas Psychologica, Vol. **7**, 2008, No 3, pp. 917-932.
3. Suárez, J. O. Análisis bibliométrico de la revista Infectio, 1995 a 2011. – Infectio, Vol. **16**, 2012, No 3, pp. 166-172.
4. Rincon, H., G. Valencia, J. Cárdernas. Colombia’s Contribution to the Use of Biogas from Solid Waste: A Bibliometric Analysis. – Contemporary Engineering Sciences, Vol. **11**, 2018, No 78, pp. 3873-3882.
5. Rojas-Sola, J. I., C. De San-Antonio-Gómez. Análisis bibliométrico de las publicaciones científicas colombiana en la categoría engineering, multidisciplinary de la base de datos Web of Science (1997-2009). – Dyna, Vol. **77**, 2010, No 164, pp. 9-17.
6. Aguilera-Prado, M., C. Aguirre, O. Salcedo. Approach to Citation Determinants of Articles from Colombian Engineering Journals in Scopus. – Contemporary Engineering Sciences, Vol. **10**, 2017, No 26, pp. 1279-1286.
7. Tahamtan, I., A. Safipour Afshar, K. Ahamdzadeh. Factors Affecting Number of Citations: A Comprehensive Review of the Literature. – Scientometrics, Vol. **107**, 2016, pp. 1195-1225.
8. Mingers, J., L. Leysderdoff. A Review of Theory and Practice in Scientometrics. – European Journal of Operational Research, Vol. **246**, 2015, pp. 1-19.
9. Chen, L., A. Baird, D. Straub. An Analysis of the Evolving Intellectual Structure of Health Information Systems Research in the Information Systems Discipline. – Journal of the Association for Information Systems, Vol. **20**, 2019, No 8, pp. 1023-1074.

10. Yang, J., J. Ward, E. Gharavi, J. Dawson, R. Alvarado. Bi-Directional Relevance Matching between Medical Corpora. – In: Proc. of 2019 Systems and Information Engineering Design Symposium (SIEDS'19), 2019, pp. 1-6.
11. Boyack, K. W., D. Newman, R. J. Duhon, R. Klavans, M. Patek, J. R. Biberstine, B. Schijvenaars, A. Skupin, N. Ma, K. Börner. Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-Based Similarity Approaches. – PLOS One, Vol. **6**, 2011, No 3, pp. 1-11.
12. Kulkarni, S., U. Apte, N. Evangelopoulos. The Use of Latent Semantic Analysis in Operations Management Research. – Decision Sciences, Vol. **45**, 2014, No 5, pp. 971-994.
13. Landauer, T. K., S. T. Dumais. A Solution to Plato's Problem: The Latent Semantic Analysis Theory. – Psychological Review, Vol. **104**, 1997, No 2, pp. 211-240.
14. Gutiérrez, M. R. Análisis Semántico Latente: ¿Teoría psicológica del significado?. – Signos, Vol. **38**, 2005, No 9, pp. 303-323.
15. Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman. Indexing by Latent Semantic Analysis. – Journal of American Society for Information Science, Vol. **41**, 1990, No 6, pp. 391-407.
16. Berry, M. W., S. T. Dumais, G. W. O'Brien. Using Linear Algebra for Intelligent Information Retrieval. – SIAM Review, Vol. **37**, No 4, 1995, pp. 573-595.
17. Schwartz, C. Lsemantica: A Command for Text Similarity Based on Latent Semantic Analysis. – The Stata Journal, Vol. **19**, 2019, No 1, pp. 129-142.

*Received: 19.06.2021; Second Version: 01.12.2021; Accepted: 23.12.2021*