ISBN 978 -987-633-574-4

Thematic Evolution of Scientific Publications in Spanish

Santiago Bianco[1], Laura Lanzarini[2], and Alejandra Zangara[2]

[1] *Information Systems Research Group, UNLa (GISI-UNLa)*

[2] *Computer Science Research Institute LIDI (III-LIDI), UNLP-CICPBA*

sabianco@unla.edu.ar, {laural,azangara}@lidiinfo.unlp.edu.ar

Abstract. Thematic evolution is a relevant technique when processing text documents about a specific topic but from different periods of time. Identifying changes in terminology and the evolution of research fields is of great interest to disciplines such as bibliometrics and scientometrics. In this article, strategies are proposed to improve the analysis of thematic evolution in scientific publications in Spanish, together with visualization techniques that allow highlighting the most relevant results. In principle, this methodology can be used in different contexts; however, here we apply it to the analysis of the scientific production related to technology in education and technology education, as an expansion of the work carried out in [1]. The tests implemented allow us stating that the inclusion index is an adequate metric to select the most relevant topic relationships, facilitating the understanding and visualization of the results obtained.

Keywords: Bibliometric Analysis, Text Mining, Thematic Evolution

1 Introduction

The analysis of text documents from specific contexts is a topic of interest for different areas such as information retrieval, document classification, bibliometric and scientometric analysis, and so forth.

When processing text documents written in different time periods, thematic evolution is an aspect that must be taken into account. Identifying the changes that took place over time in the nomenclature used on various topics within the same discipline or discourse area is an extremely useful tool when trying to apply text mining strategies.

As a specific case, any teacher, researcher or student who needs to write an article, thesis or research work, will have to carry out a review of the corresponding state of the art. In this sense, the possible topics of interest within a particular domain have to be identified. This bibliographic search process is generally time-consuming and, if not oriented correctly, can lead to blockages and frustration for the researcher. It would be interesting then to have methods and tools available to simplify the search and analysis of bibliography or publications of any kind for these processes.

In the first instance, bibliometric tools could be used to carry out the initial analysis of the texts of interest. Bibliometrics is known as a discipline capable of describing a set of publications applying statistical analysis techniques, identifying relevant thematic focuses, author collaboration networks, information on citations, and so forth. Scientometrics is a subdiscipline of bibliometrics that focuses specifically on scientific publications.

In any case, these approaches generally allow quantitative analyzes such as a list of the most cited authors, institutions with the highest number of publications, topics most written about, and so forth. When a more in-depth qualitative analysis is required, text mining and visualization techniques should be applied, such as thematic maps together with traditional bibliometric methods.

Thematic maps are a way of representing different topics covered in a field of a scientific discipline at a certain moment. Different types of bibliometric information can be used to build these graphs, one of them being the analysis and correlation between relevant terms.

An analysis technique called thematic evolution is derived from the thematic maps. It consists of showing the ”evolution” of the relevance of a particular topic on a timeline. For example, it could be shown that in 2010 there was a thematic focus dedicated to research in neural networks and that the same group of people who worked on this topic gradually moved their research interests towards a different topic, such as black box model interpretability. The idea is to show that the first theme mutated or evolved into the second one. It should be noted that, in this context, the term ”evolution” means ”change and transformation”, and does not imply there has been an improvement. Thus, saying that ”Topic A evolved into Topic B” does not necessarily mean that Topic B is better in some ways than Topic A. In the following section, the methodology proposed for the analysis of thematic evolution in scientific publications is detailed, including all required metrics and techniques.

2 Proposed Methodology

2.1 Gathering the Documents

To carry out a thematic analysis, documents that are representative of the area of interest over a period of time long enough to be able to divide it into sub-periods (small periods of time into which a larger interval is divided) are required. Then, the thematic evolution will try to establish relationships between the central themes from different sub-periods.

When accessing the documents, we decided to work with scientific journals that would use the Open Journal System (OJS) for their digital issues, which allows consulting all available articles in a systematic and repeatable way, as they are supported on the same standard system. OJS [5] is an open source solution for managing and publishing academic journals online, thus reducing publication costs compared to print versions and other forms of dissemination. Unlike Scopus, Web of Science and others, no access codes or any other type of authentication are required to extract information from the platform. This allows automating the journal articles extraction process through a script, which can easily be modified to extract the data from any journal that is implemented on OJS.

Raw data is downloaded as plain text. Key elements such as title, year of publication, abstract, and author's address are automatically pulled from the OJS system, without the need to directly access the full article. Authors and country affiliations are identified from their addresses and available metadata.

On the other hand, document publication language must be selected in advance. In this article, emphasis is placed on scientific documents in Spanish because this is still a little studied language from the point of view of thematic evolution. Inconsistent expressions, special characters, and ambiguities are processed after their download and collection, in a separate script. This script is used to give the final format to the publications so that they can be used as an input for the algorithms applied in the analysis.

2.2 Terms extraction

To analyze the documents collected, the title, abstract and keywords indicated by the authors are used. All these sections must be preprocessed and grouped so that they can be used properly by the algorithms. The process consists of the following steps:

1. The terms contained in the previously mentioned sections (title, abstract and keywords indicated by the authors) are extracted and standardized. This process consists of replacing special characters, unifying synonyms and acronyms, and writing all terms in lowercase.
2. The n-grams obtained, made up of two, three or four words, are added to the set of terms.
3. Those that exceed a certain threshold value of TD-IDF (Term Frequency — Inverse Document Frequency) are selected from the set of terms. This metric assigns high values to those terms that have a high frequency (in the given document) with a low frequency in the entire collection of documents, thus filtering common terms [7].
4. All selected terms are unified in a corpus to be analyzed by the algorithms.

2.3 Research Topics Identification

To detect the research topics and/or fields of interest for researchers, the joint occurrence or co-occurrence of previously identified terms is used [3]. This co-occurrence is calculated as indicated in equation 1 where c_{ij} is the number of documents in which both terms appear together, and c_i and c_j are the number of documents in which they appear individually.

$$e_{ij} = \frac{c_{ij}}{c_i c_j} \tag{1}$$

Using these co-occurrence values, the simple center algorithm [4]] is applied to build thematic networks made up of subgroups of strongly linked terms that correspond to interests or research issues of great importance in the academic field.

The detected networks can be represented using the density and centrality measures defined in [2].

By analyzing the relationship between the terms that make up the different thematic networks within a discipline in different sub-periods of time, the development of a given topic over the years can be analyzed, and the changes in relevant thematic focuses can be seen. This is known as thematic evolution, and is discussed in more detail in the next section.

2.4 Thematic Evolution

A thematic area is a set of topics that have evolved over different sub-periods. Each topic is made up of a set of terms. Let T_t be the set of topics detected in sub-period t and let $U \epsilon T_t$ be a topic detected in sub-period t. Let $V \epsilon T_{t+1}$ be a topic detected in the following sub-period $t + 1$. A thematic evolution from topic U to topic V is considered to have happened if there are common terms in both sets. Each $k \epsilon U \cap V$ term is considered a *thematic link*. To weight the importance of a thematic link, the inclusion index defined in [8] calculated according to equation (2) is used. This index is a simple metric that in this context is used to measure how strong the relationship between two topics is. Its value is between 0 and 1; a higher value corresponds to a stronger relationship.

$$inclusion = \frac{\#(U \cap V)}{min(\#U, \#V)} \tag{2}$$

If a topic from a sub-period has no thematic link with another topic from a later sub-period, it is considered to be discontinuous, whereas if there is a topic unrelated to a previous sub-period, it is considered as a new or emerging topic.

Understanding that the inclusion index is important when it comes to identifying the degree of relationship between the topics, in this work it is used as a metric to simplify thematic evolution visualization.

3 Results

To measure the performance of the methodology proposed in this article, documents from the $EDUTEC$ journal were used. This journal was selected because it addresses two specific topics; namely, technology applied to education and technology education. This latter aspect is relevant because all articles published use specific common vocabulary. In addition, it has publications in Spanish and has numbers published for more than 25 years.

During the article collection phase, only publications in Spanish with abstracts available for extraction and keywords or abstracts uploaded were taken into account. As a result, 392 documents were identified.

Subsequently, the set of documents obtained was divided into three sub-periods of similar duration, as follows:

- Sub-period 1: 1995-2005
- Sub-period 2: 2006-2013
- Sub-period 2: 2014-2020

For each sub-period, the most relevant topics were identified using two different strategies – first, using only the keywords and second, adding abstracts and titles to these keywords. The values of the parameters used in both cases are indicated in Table 1. For each case, the relationships identified were considered and the most relevant ones were selected according to their inclusion level value.

Parameter description	**Value**
Number of words to use in each topic (set of terms)	250
Minimum frequency for a term to be considered a member of a topic	20

Table 1. Parameters used in the analysis

As a result of the first strategy, that is, the evolution of topics from sets of terms selected taking into account only keywords, the graph in Figure 1 was obtained. In this figure, the thickness of the bands that join the topics in the different sub-periods is proportional to their inclusion index. In other words, the wider the band that joins two topics, the greater the value of the inclusion index between the two. As it can be seen, despite the fact that the number of terms involved is scarce, it is somewhat complex to identify the most relevant ones. This is where the use of the inclusion index can be very useful.

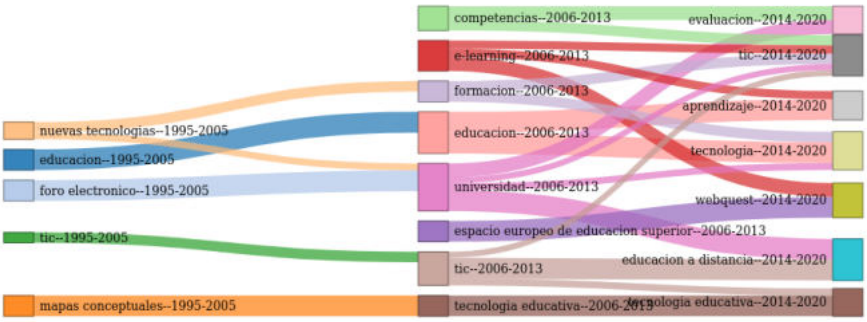

Fig. 1. Thematic evolution results using just keywords

ISBN 978 -987-633-574-4

For example, a simpler and more readable visualization applying a filter with a threshold of 0.5 for the inclusion index can be observed in Figure 2. By eliminating the topics with lower inclusion, charts become easier to read and the possibility of considering unreliable results in the analysis is reduced. Thus, it is easier to visualize those terms whose relationship between sub-periods is supported by a higher level of inclusion.

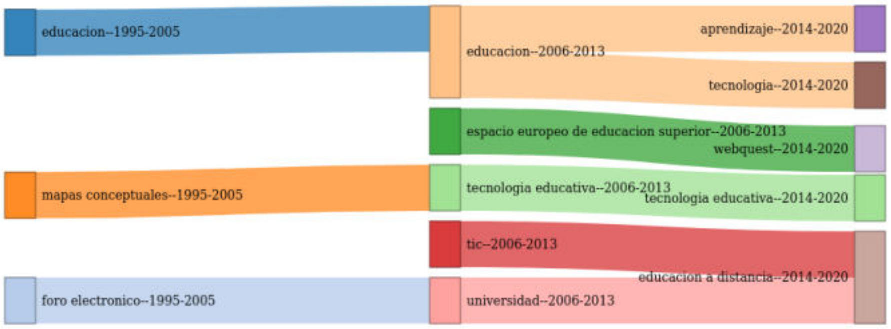

Fig. 2. Thematic evolution results using just keywords, filtered by inclusion index

Applying the second strategy, as expected, the topics identified were more closely related to the documents. This is reflected in how topics are linked, shown in Figure 3 This figure shows the relationships in a clearer way in the absence of an excessive number of crosses between connections, as it was the case in Figure 1. Regardless of this, the value of the inclusion index for each pair of terms is still directly proportional to the importance of their relationship.

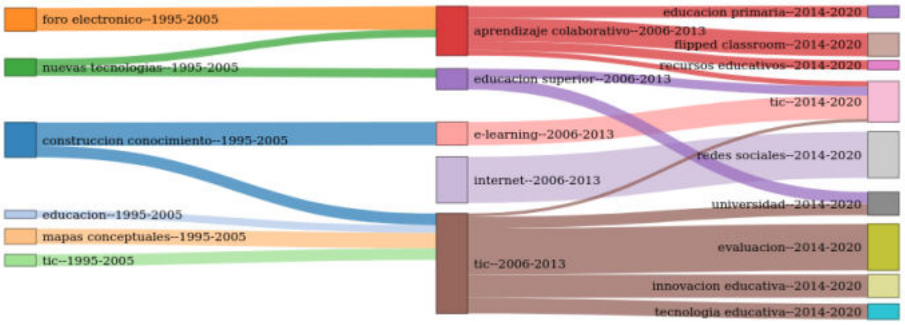

Fig. 3. Thematic evolution results using keyword terms, abstracts and titles

Table 2 shows the highest inclusion level values obtained as a result of the first procedure. These values correspond to the most interrelated topics, which were graphically joined with the thickest bands in Figure 1.

Topic A (Sub-period)	**Topic B (Sub-period)**	**Inclusion**
Education (1995-2005)	Education (2006-2013)	0.5
Electronic Forum (1995-2005)	University (2006-2013)	0.5
Conceptual Maps (1995-2005)	Educational Technology (2006-2013)	0.5
Education (2006-2013)	Learning (2014-2020)	0.5
Education (2006-2013)	Technology (2014-2020)	0.5
European Higher Education Area (2006-2013)	Webquest (2014-2020)	0.5
Educational Technology (2006-2013)	Educational Technology (2014-2020)	0.5
ICTs (2006-2013)	Distance Education (2014-2020)	0.5
University (2006-2013)	Distance Education (2014-2020)	0.5
Competencies (2006-2013)	Evaluation (2014-2020)	0.33

Table 2. Summary of results when using keywords.

Table 3 summarizes the results obtained with the second procedure; i.e., using n-grams of abstracts and titles in the analysis. As it can be seen, the first terms have high inclusion level values, indicating a consolidated relationship between both periods. Additionally, because new terms are added to represent document content, new topics appear, such as social media, flipped classroom and e-learning.

Topic A (Sub-period)	**Topic B (Sub-period)**	**Inclusion**
Internet(2006-2013)	Social Media (2014-2020)	1
ICTs (2006-2013)	Evaluation (2014-2020)	1
Knowledge-Building (1995-2005)	E-learning (2006-2013)	0.5
Electronic Forum (1995-2005)	Collaborative Learning (2006-2013)	0.5
Collaborative Learning (2006-2013)	Flipped Classroom(2014-2020)	0.5
E-Learning(2006-2013)	ICTs (2014-2020)	0.5
ICTs (2006-2013)	Educational Innovation (2014-2020)	0.5
Conceptual Maps (1995-2005)	ICTs (2006-2013)	0.33
ICTs (2006-2013)	Educational Technology (2014-2020)	0.33
Knowledge-Building (1995-2005)	ICTs (2006-2013)	0.25

Table 3. Summary of results when using keywords, abstracts and titles.

Through the expert consultation method, it was determined that the results obtained by this method are related to the development of ideas about teaching and ICTs in the field of research.

For example, the topics related to ”Electronic Forums” may have become ”Collaborative Work,” since the ideas of collaborative work and learning went through an automation process when tools became available to carry out tasks. Thus, it would make sense for articles that covered tools to start working on conceptual models. This may also explain the transition between ”Collaborative Learning” and ”Flipped Classroom”.

The transition between ”ICTs” and ”Educational Technology” could also be understood if we consider the discipline that builds conceptual models that include the use of technological tools.

4 Conclusions and future lines of work

This article describes a methodology capable of analyzing thematic evolution in scientific documents. Even though the results obtained come from the analysis of a journal in the domain of computer technology applied to education published in Spain, due to the nature of the text mining and bibliometric methods used, this can be replicated in other domains.

Two different procedures were carried out when creating the sets of terms on which co-occurrence would be measured; this is the metric used at the beginning of the topic identification process. We were able to corroborate that, by adding n-grams previously filtered by their TD-IDF value in the set of documents analyzed, an improvement is observed in the value of the metrics obtained and the identification of the terms that appear as most relevant in the results. Considering both keywords and the terms present in the title and the abstract of each document proved to yield relationships with a higher level of inclusion than when considering only keywords. This is because topic building is enriched and intersections with greater cardinality are obtained. On the other hand, both procedures proved that filtering the relationships by level of inclusion is effective in simplifying visualization. This is an extension of the work presented in [1] on that occasion, the articles (taken from the TEyET journal) were analyzed using only keywords. Term co-occurrence was represented in greater detail using the ”density” and ”centrality” metrics, but inclusion level was not used.

Once again, our conclusion is that the process used to analyze thematic evolution in Spanish is promising, and different methodologies for extending this work were identified. Firstly, the same analysis could be carried out using metrics other than TF-IDF to select the n-grams generated. This is useful because TF-IDF has certain limitations when applied to large sets of documents [6]. Other result visualization methods were also devised, so that the most relevant results can be highlighted automatically, modifying the threshold or the metric used.

Further extensions to this work are also planned - based on the results obtained, an analysis methodology will be built and validated by comparing its results with those produced by a group of experts. It would also be interesting to build an accessible and simple tool that would allow users who are not experts in computing or data analysis to take advantage of this methodology. These two future lines of work are complementary of each other and will also require the

ISBN 978 -987-633-574-4

design of evaluation devices, both for the experts who validate the methodology and for the potential users of the tool, so that process efficiency and usability can be verified.

References

1. S. Bianco, L. L., and Z. A. Evolución temática de publicaciones en español. una estrategia posible para el diseño de situaciones didácticas. In *XVI Congreso de Tecnología en Educación y Educación en Tecnología (TEyET 2021)*. RedUNCI, 2021.
2. M. Callon, J.-P. Courtial, and F. Laville. Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry. *Scientometrics*, 22:155–205, 1991.
3. M. Callon, J.-P. Courtial, W. A. Turner, and S. Bauin. From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, 22(2):191–235, 1983.
4. N. Coulter, I. Monarch, and S. Konda. Software engineering as seen through its research literature: A study in co-word analysis. *Journal of the American Society for Information Science*, 49(13):1206–1223, 1998.
5. P. K. Project. Open journal system, 2001. http://pkp.sfu.ca/ojs.
6. J. Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer, 2003.
7. S. Robertson. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of documentation*, 2004.
8. C. Sternitzke and I. Bergmann. Similarity measures for document mapping: A comparative study on the level of an individual scientist. *Scientometrics*, 78:113 – 130, 2009.