# Evolución temática de publicaciones en español. Una estrategia posible para el diseño de situaciones didácticas.

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

[1]*Grupo de Investigación en Sistemas de Información UNLa (GISI-UNLa)*

[2] *Instituto de Investigación en Informática LIDI (UNLP-CIC)*

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

## Resumen

La evolución temática es un tema relevante a la hora de procesar documentos de textos de un tema específico pero provenientes de distintas épocas de tiempo. Identificar los cambios en la terminología y la evolución en los temas de estudio resulta de sumo interés para disciplinas como la bibliometría y la cienciometría. En el caso de docentes y estudiantes, tanto el diseño y puesta a punto de clases como la preparación de actividades de investigación específicas (investigación en sitios de internet a través de Webquest, por ejemplo) implica la investigación cuidadosa y la recolección de información previamente por parte de los docentes.

Este trabajo detalla la metodología utilizada para analizar la totalidad de los artículos en español de la Revista *TE&ET* desde el momento de su creación en diciembre de 2006 y hasta la fecha, con el objetivo de observar focos temáticos junto con su evolución a lo largo de los años. Dicha metodología hace uso de estrategias pertenecientes a la Inteligencia de Datos combinadas con distintas técnicas de visualización que, si bien puede ser aplicada en distintos contextos, en este trabajo se focaliza en el análisis de la producción científica relacionada con Educación en Tecnología. Por lo tanto, como caso de estudio se analizarán artículos de una revista indexada correspondientes al dominio educativo con el objetivo de tener un panorama de los temas de interés tratados en dicha revista y cómo fueron cambiando a través de los años. Esta herramienta podría ser un insumo para el diseño de propuestas didácticas que apelen a la investigación por parte de los estudiantes.

**Palabras Clave:** Mapeo de la ciencia, Análisis de co-palabras, Estudios bibliométricos, Evolución temática.

## 1. Introducción

El análisis de documentos de textos provenientes de contextos específicos es un tema de interés para distintas áreas tales como recuperación de información, clasificación de documentos, análisis bibliométricos y cienciométricos, entre otros.

Cuando se procesan documentos de textos escritos en períodos de tiempo diferentes, la evolución temática es un aspecto que debe tenerse en cuenta. Reconocer los cambios, que se han ido produciendo a lo largo del tiempo, en la denominación de los distintos tópicos dentro a una misma disciplina o área de discurso es una herramienta sumamente útil a la hora de querer aplicar estrategias pertenecientes a la Minería de Textos.

En particular, quienes deseen realizar una tarea de investigación, ya sean docentes o estudiantes, deberán efectuar una revisión del estado del arte correspondiente. En esta dirección, suele ser necesario identificar cuáles son los posibles temas de interés dentro de un dominio en particular. Este proceso de búsqueda bibliográfica generalmente consume mucho tiem-

po y si no se orienta correctamente puede conducir a bloqueos y frustración para el investigador. Sería interesante entonces contar con métodos y herramientas para simplificar la búsqueda y el análisis de bibliografía o publicaciones de cualquier tipo que faciliten estos procesos.

La Bibliometría se conoce como una disciplina capaz de describir un conjunto de publicaciones aplicando técnicas de análisis estadístico, identificando focos temáticos relevantes, redes de colaboración de autores, información sobre citas y demás. La Cienciometría es una subdisciplina de la Bibliometría que se enfoca particularmente en publicaciones científicas.

Generalmente, estos enfoques posibilitan análisis cuantitativos como por ejemplo una lista de autores más citados, instituciones que más publicaron, temas sobre los cuáles se escribió más, entre otros. Sería interesante, por lo tanto, poder realizar un análisis cualitativo más profundo en conjunto con los métodos tradicionales de Bibliometría, aplicando técnicas de Minería de Datos y de visualización como mapas temáticos.

Los mapas temáticos son una forma de representar diferentes temas tratados en un campo de una disciplina científica en un determinado momento en el tiempo. Distinto tipo de información bibliométrica puede ser usada para armar estos gráficos, siendo el análisis y la correlación entre palabras clave una de ellas.

De los mapas temáticos se deriva una técnica de análisis denominada evolución temática. La misma consiste en mostrar en una línea de tiempo la "evolución" de la relevancia de un tema en particular. Por ejemplo, se podría mostrar que en el 2010 hubo un foco temático dedicado a la investigación en Redes Neuronales y el mismo grupo de gente que trabajaba en ese tema fue inclinando su investigación hacia otro distinto como puede ser la interpretabilidad de modelos de caja negra. La idea es mostrar que el primer tema mutó o evolucionó hacia el otro. Cabe destacar que en este contexto la palabra evolución denota cambio y transformación, y no necesariamente decir que "el Tema A evolucionó al Tema B" significa que el tema B es mejor en algún aspecto con respecto al Tema A.

En este contexto, se observa que si bien existen algunas herramientas que permiten estos análisis, muy pocas se encuentran disponibles para el lenguaje español y no son tan intuitivas como para que puedan ser usadas por usuarios no expertos en el área informática. Además, están orientadas a publicaciones científicas en inglés descargadas de portales tales como Web of Science o Scopus, con un formato en particular. Sería interesante poder extender estas herramientas para analizar la evolución temática en cualquier tipo de publicación y además en lenguaje español.

Teniendo en cuenta estos aspectos, este artículo detalla una metodología utilizada para analizar un conjunto de publicaciones en español provenientes de la Revista *TE&ET* desde sus comienzos en diciembre de 2006 hasta la fecha, con el objetivo de observar focos temáticos junto con su evolución entre estos años.

Es importante destacar que, si bien en este caso se procesarán artículos científicos, esta misma metodología puede utilizarse en otros contextos.

## 2. Metodología

A continuación se detalla el proceso utilizado para efectuar la evolución temática.

### 2.1. Recolección y preprocesamiento de los documentos

En este artículo se utilizaron documentos provenientes de la revista *TE&ET*. Esta revista fue seleccionada por abordar temáticas de interés para la investigación actual: tecnología aplicada en la educación y educación en tecnología. Además, contiene publicaciones en español y posee números publicados desde hace casi 15 años. Otro criterio no menor para su elección es el hecho de que utiliza el sistema Open Journal System (OJS) para sus números digitales, lo cual permite consultar todos los artículos disponibles de una manera sistemática y repetible, al estar soportadas sobre un mismo sistema estándar. OJS [5] es una solución

de código abierto para administrar y publicar revistas académicas en línea, que pueden reducir los costos de publicación en comparación con las publicaciones impresas y otros procesos de publicación. A diferencia de Scopus, Web of Science y demás, no es necesario contar con claves de acceso ni ningún otro tipo de autenticación para extraer información de la plataforma. Esto permite automatizar la extracción de todos los artículos de la revista a través de un script, que una vez implementado para $TE\&ET$, puede fácilmente ser modificado para extraer los datos de cualquier revista que esté implementada sobre OJS.

En total, se identificaron finalmente 227 artículos de la revista a analizar. Para alcanzar este número se tuvieron en cuenta solamente las publicaciones en español con resúmenes disponibles para su extracción.

Los datos en bruto se descargan como texto sin formato. Los elementos clave, como el título, el año de publicación, el resumen y la dirección del autor se extraen automáticamente del sistema OJS, sin necesidad de acceder directamente al artículo completo. Las afiliaciones de los autores y los países se identifican a partir de las direcciones y meta-datos disponibles de los mismos. Las expresiones inconsistentes, caracteres especiales y ambigüedades se procesan posteriormente a la descarga y recopilación de los mismos, en otro script en Python separado. En este mismo se le da el formato final a las publicaciones para que puedan servir como entrada para los algoritmos que se utilizan en el análisis.

En cuanto al análisis temático, además de las palabras clave del autor y las KeyWords Plus también se incluyen las palabras clave del título y el resumen utilizando algoritmos de procesamiento del lenguaje natural basado en el análisis sintáctico de árboles: 1) Las formas singulares y plurales de todas las palabras clave de autor y KeyWords Plus se almacenan primero en una base de datos; 2) Las palabras clave del texto del título y del resumen se extraen automáticamente y por separado de la base de datos; 3) Se reemplazan los caracteres especiales y se generan n-gramas para que no se pierda el significado de las palabras clave; 4) Todas las palabras clave se fusionan y se unifican como forma singular. Transformaciones tales como agrupar palabras clave que representan el mismo concepto o considerar algún peso en las palabras clave del autor, pueden mejorar los resultados del proceso.

## 2.2. Detección de temas de investigación

Para detectar los temas de investigación y/o los centros de interés que atraen la atención de los investigadores se utilizó la ocurrencia conjunta o co-ocurrencia de las palabras clave identificadas previamente [2]. Dicha co-ocurrencia se calcula como se indica en la ecuación 1 siendo $c_{ij}$ la cantidad de documentos en los que ambas palabras aparecen juntas y $c_i$ y $c_j$ la cantidad de documentos en los que aparecen las palabras individualmente.

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

Utilizando estos valores de co-ocurrencia, se aplicó el algoritmo de centros simples [4] para construir redes temáticas formadas de subgrupos de palabras clave con un fuerte vínculo y que corresponden a intereses o problemas de investigación de gran importancia en el ámbito académico.

Las redes detectadas pueden representarse utilizando las medidas de densidad y centralidad definidas en [1].

La centralidad mide el grado de interacción entre las redes. Su valor se calcula como se indica en la ecuación 2 siendo $k$ una palabra clave perteneciente al tema y $h$ una palabra clave perteneciente a otros temas. Representa la fuerza de los vínculos externos con otros temas y puede considerarse una medida de la importancia de un tema en el desarrollo de todo el campo de investigación analizado.

$$c = 10 \sum e_{kh} \tag{2}$$

La densidad mide la fuerza interna de la red y su valor se calcula como se indica en la ecuación 3 siendo $i$ y $j$ las palabras clave que per-

tenecen al tema y $w$ el número de palabras clave del tema. La densidad mide la fuerza de los vínculos internos entre todas las palabras clave que describen el tema de investigación. Este valor puede entenderse como una medida del desarrollo del tema.

$$d = 100\left(\sum c_{ij}/w\right) \quad (3)$$

Sobre la base de estas dos medidas, los temas de investigación pueden representarse en un diagrama estratégico bidimensional con cuatro cuadrantes, ordenados por centralidad y densidad. Por lo general, los temas del cuadrante superior derecho, conocidos como temas motores, están bien desarrollados y son importantes para estructurar un campo de investigación. Los temas del cuadrante superior izquierdo están bien desarrollados pero tienen una importancia marginal para el campo ya que presentan mucha interrelación entre sí pero poca con el resto de los temas en los otros cuadrantes. Los temas del cuadrante inferior izquierdo están débilmente relacionados y poco desarrollados, lo que indica que son campos emergentes o en declive, en vías de desaparición. Finalmente en el cuadrante inferior derecho se encuentran los temas que son relevantes para algunos campos de investigación pero están poco desarrollados. Este último contempla aquellos temas básicos, transversales y generales. Para más detalles sobre este tema, se recomienda revisar el artículo [3].

### 2.3. Evolución temática

Un área temática es un conjunto de temas que han evolucionado a lo largo de diferentes subperíodos (pequeños períodos de tiempo en los que se divide un intervalo mayor). Supongamos que $T_t$ es el conjunto de temas detectados del subperíodo $t$, y que $U \epsilon T_t$ denota cada tema detectado. Sea $V \epsilon T_{t+1}$ cada tema detectado en el siguiente subperíodo $t + 1$. Se considera que hay una evolución temática del tema $U$ al tema $V$ si hay palabras clave presentadas en ambas redes temáticas asociadas. Las palabras clave $k \epsilon U \cap V$ se consideran un "nexo temático". Para ponderar la importancia de un nexo temático se utiliza el índice de inclusión definido en [6] calculado según la ecuación 4. Cabe señalar que un tema puede pertenecer a un área temática diferente o no pertenecer a ninguna.

$$inclusion = \frac{\#(U \cap V)}{min(\#U, \#V)} \quad (4)$$

En un mapa bibliométrico de la evolución temática, los rectángulos indican las áreas temáticas más relevantes encontradas teniendo en cuenta los límites establecidos por el conjunto de parámetros de los algoritmos empleados. El tamaño de dichos rectángulos y las líneas que los conectan dependen del índice de inclusión y el número de publicaciones asociadas a las áreas temáticas. La conexión entre dos áreas temáticas indica que hay palabras clave en los documentos que las vinculan a través de distintos subperíodos de tiempo. Si un tema de un subperíodo no tiene ningún vínculo con otro tema de un subperíodo posterior, se considera discontinuo, mientras que si hay un tema no relacionado con un subperíodo anterior se lo considera como un tema nuevo o emergente.

## 3. Resultados

En base a la aplicación de los pasos expuestos en las secciones anteriores, se obtuvieron distintas representaciones de la evolución temática de la revista $TE\&ET$. Previo al análisis de los resultados finales, además de procesar los datos se debieron ajustar distintos parámetros en los algoritmos. Los valores de los parámetros usados en este caso de estudio pueden verse en la tabla 1.

Como resultado del análisis de la evolución temática se obtuvo el gráfico de la figura 2, junto con tres mapas temáticos correspondientes a cada subperíodo evaluado: 2006-2012, 2013-2016 y 2017-2020. En la figura 1 se muestra el mapa correspondiente al subperíodo 2013-2016, en el cual pueden observarse las siguientes agrupaciones:

- Temas motores: acceso abierto, redes sociales, simulación, web y TIC.

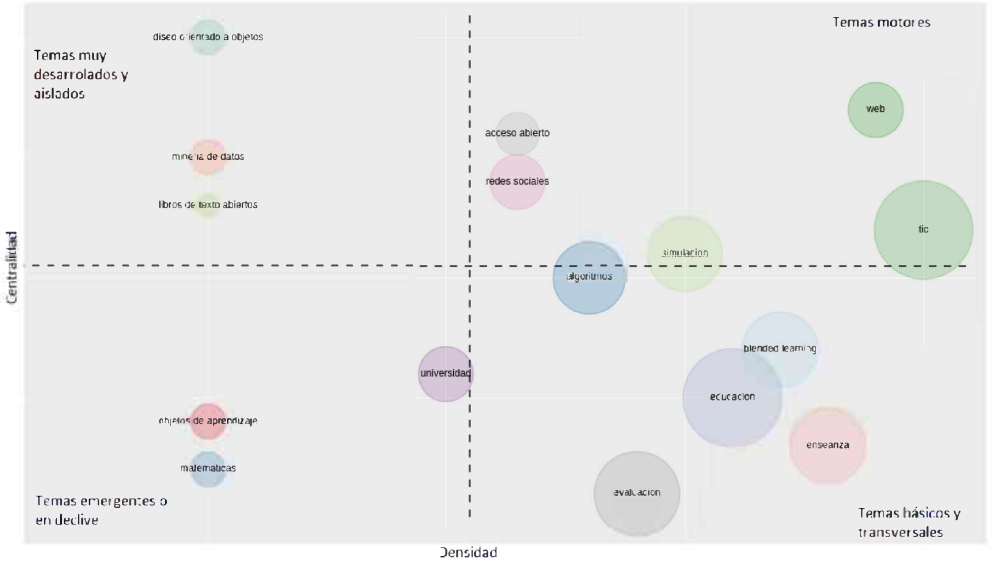


Figura 1: Mapa temático para la revista $TE\&ET$ en el subperíodo 2013-2016

| Descripción del parámetro | Valor |
|---|---|
| Cantidad de palabras a usar en cada red temática | 250 |
| Frecuencia mínima de cada palabra para ser considerada en un cluster | 20 |
| Umbral de tamaño para el radio de los clusters | 0.35 |

Tabla 1: Valores de los parámetros usados en el análisis

- Temas bien desarrollados pero aislados: diseño orientado a objetos, minería de datos y libros de texto.
- Temas emergentes o en declive: universidad, objetos de aprendizaje y matemáticas.
- Temas básicos y transversales: algoritmos, educación, blended learning, enseñanza y evaluación.

Teniendo en cuenta todos los mapas temáticos en el período y las relaciones entre ellos, se genera el análisis de la evolución temática.

Como se mencionó anteriormente, los gráficos que representan la evolución temática indican cómo el interés en un área en un determinado subperíodo de tiempo, puede volcarse hacia otra área temática en un subperíodo distinto. La figura 2 muestra la evolución de los temas detectada en los artículos analizados para el período 2006-2020. Allí puede observarse que las líneas que conectan los rectángulos representan este proceso de transformación, mientras que el grosor de las mismas indica el peso que tiene esa área temática en esa etapa de análisis. Por ejemplo, puede verse la aparición entre 2006 y 2016 de un área temática marcada como "webquest", la cual está conectada con el nodo "programación 2013-2016". Esto quiere decir que las personas o los grupos de investigación que en el primer período publicaron acerca de webquest o utilizando dicho término, por algún motivo comenzaron a usar el término programación en 2013 en sus trabajos.

Al igual que con los mapas temáticos, tener una perspectiva visual de la evolución temática de una disciplina científica ayuda a relevar el

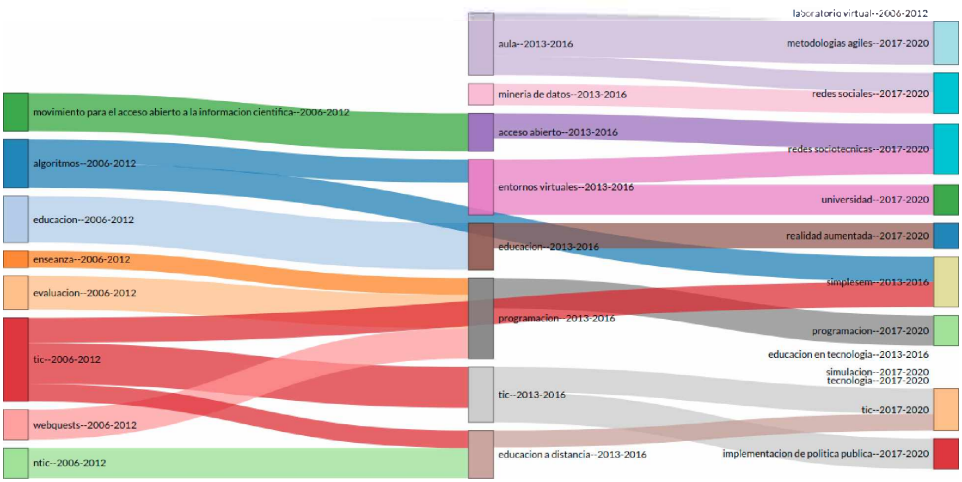


Figura 2: Resultados para el análisis de la evolución temática de la revista $TE\&ET$ en el período 2006-2020

cambio del estado del arte e identificar los posibles campos de investigación de interés en la actualidad. Nuevamente los resultados obtenidos y el propósito final de los mismos dependerá de la naturaleza de los textos que se estén analizando con la metodología. Puede aplicarse para analizar el contexto histórico de una disciplina en particular, orientar los inicios de una investigación científica o diseñar propuestas didácticas que apelen justamente a la investigación sobre un área temática en particular, por mencionar algunos ejemplos.

## 4. Conclusiones y líneas de trabajo futuras

En el presente artículo se describieron los conceptos de Bibliometría y Cientometría junto con una metodología de estudio de la evolución temática basada en el análisis de textos aplicable en diferentes contextos. Nuevamente cabe destacar que la utilidad de dicha metodología y los resultados que se puedan obtener son variables y dependerán de la naturaleza de los textos analizados. Es así como se puede utilizar la evolución temática como herramienta en contextos muy diferentes tales como: el diseño propuestas didácticas que incluyan actividades de investigación en algún período de tiempo, el análisis de los temas discutidos en una red social ante determinado acontecimiento, la obtención de un resumen de la evolución del estado del arte de una disciplina, entre otros.

Particularmente en este trabajo el objetivo fue generar una descripción de los temas abordados durante todo el período de publicación de la revista $TE\&ET$ (2006 a 2020) y cómo fueron cambiando los focos de interés en la misma, para corroborar qué tan efectiva es la metodología propuesta en [3] para el análisis de textos en español. Durante el proceso se pudo verificar que, tras el procesamiento de los datos de entrada y el refinamiento de los parámetros de los algoritmos utilizados, se pueden conseguir resultados interesantes que permiten describir la evolución de distintos temas en un período de tiempo definido, independientemente del tipo de fuente de información. Es difícil, sin embargo, encontrar la mejor configuración de parámetros para un conjunto de datos sin contar con un experto en el área en la cuál se esté investigando. Se observó también que la aplicación de todo el proceso, desde la obtención de la información hasta

la evaluación de los resultados es difícil de realizar para alguien ajeno a las áreas de sistemas y estadística.

El proceso utilizado para el análisis de la evolución temática es prometedor y se detectaron distintas formas de extender este trabajo. A raíz de los resultados obtenidos y la experiencia realizada se espera a futuro trabajar en una metodología replicable para realizar este mismo análisis en cualquier dominio, la cual esté validada a través de la comparación de los resultados que se obtengan con un grupo de expertos. Sería interesante también poder facilitar la ejecución de dicha metodología a través de una herramienta accesible y sencilla para usuarios no expertos en informática o análisis de datos.

Estas dos líneas de trabajo futuras son complementarias y van a requerir, además, el diseño de dispositivos de evaluación tanto para los expertos que validen la metodología como para los posibles usuarios de la herramienta, de manera que se pueda verificar la eficiencia del proceso a desarrollar junto con su usabilidad.

## Referencias

[1] 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. 22:155–205, 1991.

[2] 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.

[3] M. Cobo, A. López-Herrera, E. Herrera-Viedma, and F. Herrera. An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field. *Journal of Informetrics*, 5(1):146–166, 2011.

[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.

[6] 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.