Desarrollo de una aplicación para la construcción de mapas de conocimiento generados por un tema de investigación

Jairo Castrellón Torres

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá D.C., Colombia
2024

Desarrollo de una aplicación para la construcción de mapas de conocimiento generados por un tema de investigación

Jairo Castrellón Torres

Trabajo final de grado presentado como requisito para optar al título de:
Maestría en Ciencias - Estadística

Director:
Ph.D. Campo Elías Pardo Turriago

Línea de Investigación:
Procesamiento de Lenguaje Natural

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá D.C., Colombia
2024

Dedicatoria

...

Índice general

Agradecimientos

...

Resumen

El uso de herramientas bibliométricas para el análisis de las tendencias en investigación se hace cada vez más necesario en el mundo de la producción académica debido a la velocidad en la que se está generando nuevo conocimiento y la gran capacidad de los medios digitales para poner esta información a disposición de los interesados en las diferentes bases de datos, en este caso, de redes bibliográficas. En particular, los mapas de conocimiento se han convertido en guías importantes para los investigadores, en la medida que les permite tener un amplio panorama del flujo que presenta su tema de interés, de tal manera que visualicen las áreas y subáreas más relevantes en su investigación. Éste trabajo pretende ofrecer una alternativa a las herramientas que ya existen mediante una aplicación, haciendo un análisis más exhaustivo en la generación de palabras y conceptos clave que se puedan inferir de la información básica de un texto investigativo, a parte de los que el autor emite como palabras clave.

Palabras clave: Palabras clave, Mapas de conocimiento, Procesamiento del lenguaje natural, Aprendizaje automático.

Abstract

The use of bibliometric tools for analyzing research trends is becoming increasingly necessary in the world of academic production. This is due to the speed at which new knowledge is being generated and the significant capacity of digital media to make this information available to those interested in various bibliographic network databases. In this case, knowledge maps, in particular, have become essential guides for researchers, allowing them to have a broad overview of the flow presented in their area of interest. This enables them to visualize the most relevant areas and sub-areas in their research. This work aims to provide an alternative to existing tools through an application, conducting a more comprehensive analysis in generating key words and concepts that can be inferred from the basic information of a research text, in addition to those identified by the author as keywords.

Keywords: Keywords, Knowledge Maps, Natural Language Processing, Machine Learning

Lista de Figuras

Lista de Tablas

1. Introducción

Con el continuo y constante crecimiento de las tecnologías de la información, el acceso a las diferentes bases de datos que almacenan y disponibilizan toda la producción investigativa están cada vez más al alcance de las manos de todos aquellos que están interesados en aportar al conocimiento por medio de estos documentos. Este fenomeno, ha producido que el volúmen de artículos, libros, conferencias, entre otro tipo de escritos académicos se vea incrementado exponencialmente en los últimos anõs hasta el punto de sufrir congestiones a la hora de analizar los distintos indicadores bibliométricos como descargas, citaciones, visualizaciones y demás [Lusher et al., 2023].

En este sentido, el investigador también se ha visto afectado a la hora de consultar y analizar la información relacionada a su campo de acción, puesto que, al llegar a estos puntos de congestión informativa, se dificulta encontrar los documentos que puedan aportar acertadamente a su trabajo investigativo, lo cual, aumenta el tiempo en el proceso de revisión bibliográfica y prolonga la obtención de resultados que pueden agregar valor a la comunidad académica.

Los mapas de conocimiento se han desarrollado precisamente para dar al investigador una idea de las corrientes de conocimiento que tiene determinado tema, y ayudan a seleccionar aquellos documentos que pueden sumar a su trabajo académico [Santana and Cobo, 2020].

Distintas herramientas se han construido para tal fin, *CiteSpace* [Synnestvedt et al., 2005] por ejemplo, es una aplicación elaborada en código Java que permite realizar exploraciones visuales descubriendo el conocimiento que se tiene en las bases de datos, aunque más allá de la generación de mapas de conocimiento, hace un estudio profundo de otros incicadores bibliométricos como citaciones, co-citaciones, etc. *SciMat* [Cobo et al., 2012] y *Klink-2* [Osborne and Motta, 2015] son dos ejemplos más de aplicativos orientados a la construcción de estas redes de conocimiento.

El presente trabajo busca crear una alternativa a los aplicativos antes mencionados, de tal manera que se puedan generar mapas de conocimiento basados en la extracción de palabras (o frases) clave y así identificar áreas y subáreas de conocimiento en un corpus que contiene información como el título, resumen (o *abstract*), año de publicación y palabras clave sugeridas por el autor, de distintos artículos de investigación. Este desarrollo se realizará haciendo uso de *Python* como herramienta y lenguaje de programación para la construcción de la

aplicación.

A su vez, este trabajo se compone por los siguientes capítulos: en el capítulo 1 se hará una descripción de los diferentes trabajos y aplicativos ya desarrollados en este campo, el capítulo 2 habla del pre-procesamiento de los datos, especificando las diferentes técnicas utilizadas y también describe las diferentes metodologías consideradas para la generación de palabras clave, así como también se definirán métricas para seleccionar el método más adecuado. En el capítulo 3 se presentará la constucción de los mapas de conocimiento, tomando como base las palabras clave generadas y otras técnicas de agrupación para datos textuales y en el capítulo 4 se mostrará un caso de uso, para un conjunto de datos recopilados de manera automática, mostrando los resultados más relevantes.

2. Herramientas y trabajos relacionados

En este capítulo se presentan algunas de las herramientas ya desarrolladas, cómo funcionan, algunos ejemplos de uso, ventajas y desventajas tanto técnicas como analíticas. El objetivo de este ejercicio es poder capturar de cada una de ellas los pros y contras en la construcción de los mapas de conocimiento para, de esta manera, trasladarlo al aplicativo que se desarrollará posteriormente.

2.1. CiteSpace II

CiteSpace II es un aplicativo desarrollado en lenguaje Java que busca mezclar metodologías de visualización de la información con herramientas bibliométricas, así como también técnicas de la mineria de datos de tal manera que identifique patrones en las dinámicas de citas entre artículos de investigación.

El objetivo principal de la herramienta, como se menciona en [Synnestvedt et al., 2005], es disponer de manera más sencilla y rápida al investigador las distintas tendencias que pueden emerger de una determinada área de conocimiento. Estás áreas, o dominios de conocimiento, se modelan para posteriormente visualizarse como bifurcaciones en el tiempo de dos conceptos usualmente utilizados en la ciencia de la información: frentes de investigación y bases intelectuales. [Price, 2011] establece el concepto de frente de investigación, indicando que se refiere a un conjunto de artículos que son citados de manera activa por parte de los científicos y [Persson, 1994] establece la relación existente entre frente de investigación, refiriendose a esto como todas aquellas citaciones realizadas en diferentes artículos investigativos en un tema específico, mientras que los artículos que son citados forman una base intelectual.

CiteSpace II tiene como ejes fundamentales los siguientes conceptos:

- Adaptación al algoritmo de detección de ráfagas de Kleinberg
- Definición de medidas de centralidad
- Redes heterogéneas

Las cuales describiremos a continuación:

2.1.1. Algoritmo de detección de ráfagas de Kleinberg

En [Kleinberg, 2002] se presenta formalmente en que consiste el algoritmo de detección de ráfagas para los diferentes estados. Inicialmente, suponiendo que se tiene un flujo de documentos, por ejemplo, una carpeta con una gran cantidad de correos electrónicos, o en nuestro caso, un corpus que contiene información general de varios artículos investigativos en un campo o área amplio. Una ráfaga puede interpretarse como puntos en los cuales la intensidad en el surgimiento de temas (o subtemas) clave aumenta bruscamente, quizás pasando de una frecuencia de meses o años a una frecuencia de semanas o días.

En cualquier caso, la tasa en el surgimiento de estos nuevos temas es generalmente irregular, es decir, no tiende a aumentar suavemente para luego disminuir, sino que por el contrario, muestra frecuentes variaciones entre surgimientos inmediatos y pausas extensas. En gran medida, el objetivo es generar una estructura global a partir de una reducción de datos robusta, identificando ráfagas únicamente cuando presente suficiente intensidad.

Formulación

Quizás el modelo aleatorio más simple para generar una secuencia de tiempos en los que se exhibe un surgimiento de un tema (o subtema) ráfaga, se basa en una distribución exponencial: los temas ráfaga surgen de manera probabilística, de modo que el intervalo x en el tiempo entre el surgimiento de los temas i e $i+1$ se distribuye según la función de densidad exponencial "sin memoria"$f(x) = \alpha e^{-\alpha x}$, para un parámetro $\alpha > 0$. En otras palabras, la probabilidad de que el intervalo supere x es igual $e^{-\alpha x}$). El valor esperado del intervalo en este modelo es α^{-1}, y por lo tanto se puede referir a α como la tasa de surgimiento de temas ráfaga.

FALTA FORMULACION DEL ALGORITMO

2.1.2. Medidas de centralidad

2.1.3. Redes heterogéneas

2.1.4. Funcionamiento

Se considera que un dominio de conocimiento es una función de mapeo que conecta un frente de investigación con su base intelectual. Esta función proporciona el fundamento conceptual para abordar tres problemas prácticos: 1) entender la naturaleza de un frente de investigación, 2) clasificar una especialidad y 3) identificar de manera oportuna tendencias emergentes y cambios abruptos. *CiteSpace* recopila n-gramas, es decir, palabras o frases únicas de hasta cuatro palabras, de diversas partes de los artículos en un conjunto de datos, como títulos,

resúmenes y descripciones. Los términos relacionados con el frente de investigación se determinan mediante la tasa de crecimiento pronunciado de sus frecuencias. Además, se han creado dos perspectivas, vistas de clúster y vistas de zona horaria, para analizar y visualizar las relaciones entre los artículos en un espacio bidimensional. Los nuevos métodos en *CiteSpace II* buscan mejorar la claridad y la interpretación de las visualizaciones, reduciendo la carga cognitiva del usuario al explorar tendencias y puntos pivote en la estructura del conocimiento.

2.2. SciMat

2.3. Klink

3. Generación de palabras clave

Como se vió en el capítulo anterior, un insumo primordial para la construcción de mapas de conocimiento es la generación (o extracción) de palabras (o frases) clave. En este capítulo se describirá en detalle, en primera instancia, el conjunto de datos seleccionado para avanzar en el desarrollo de la herramienta, así como todo el pre-procesamiento realizado en estos datos, el detalle formal de las metodologías contempladas para la extracción de las palabras clave y, por último, la definición de las métricas utilizadas para selecciónar la metodología más adecuada para la obtención de *keywords*.

3.1. Conjunto de datos y pre-procesamiento

Para el desarrollo del aplicativo, se ha tomado como punto de partida la selección de un conjunto de datos adecuado de tal manera que permita tener la cantidad suficiente de registros, que en nuestro caso será información general de artículos de investigación relacionados a un tema particular.

El conjunto de datos seleccionado fue obtenido de [Citation Network Dataset., 2024]. El DBLP (*database of scientific publications*) es un conjunto de datos que forma una red de citas, compuesto por información extraída de diversas fuentes. Contiene datos útiles para diversas aplicaciones, como agrupamiento con información de red, análisis de influencia en la red de citas y modelado de temas. Este conjunto de datos contiene información general de al rededor de 3.000.000 de artículos de investigación relacionados a la ciencia de datos. En la tabla **3-1** se muestra el esquema que trae este conjunto de datos.

Tabla 3-1.: Esquema del conjunto de datos [Citation Network Dataset., 2024].

Nombre campo	**Tipo campo**	**Descripción**	**Ejemplo**
id	string	Paper ID	013ea675-bb58-42f8-a423-f5534546b2b1
title	string	Paper title	Prediction of consensus binding mode geometries for related chemical series of positive allosteric modulators of adenosine and muscarinic acetylcholine receptors
authors	list of strings	Paper authors	["Leon A. Sakkal", "Kyle Z. Rajkowski", Roger S. Armen"]
venue	string	Paper venue	Journal of Computational Chemistry
year	int	Published year	2017
n_citation	int	Citation number	0
references	list of strings	Citing papers' ID	["4f4f200c-0764-4fef-9718-b8bccf303dba", .ª699fbf-fabe-40e4-bd68-46eaf333f7b1"]
abstract	string	Abstract	This paper studies ...

Una vez seleccionado el conjunto de datos para realizar el desarrollo, lo que procede es la fase relacionada al pre-procesamiento de este, ya que es necesario hacer algunas transformaciones de tal manera que se pueda extraer información más precisa. Para lo anterior, existen diferentes etapas en el procesamiento de los datos que se van a describir a continuación y que se explican con más profundidad en [Manning and Schutze, 1999].

3.1.1. Eliminación de palabra vacías

Dentro del vocabulario que se trabajará a lo largo del trabajo y que se manejarán en nuestro corpus lingüistico, existe un conjunto de palabras con las cuales es necesario tratar, y son aquellas conocidas como palabras vacías (o *stopwords* en inglés). Las *stopwords* son aquellas que tienen una alta frecuencia dentro de todo el corpus, pero que no representan un significado como tal, algunos ejemplos de estas son algunas figuras gramaticales como lo son los conectores en español (pero, aunque, sobre, etc.).

En el desarrollo propuesto en este trabajo, al tener un corpus completamente en ingles, se deberá tratar estas palabras vacías con las diferentes herramientas con las que se cuentan.

En [Blanchard, 2007] se menciona históricamente como se le ha dado tratamiento a este tipo de palabras y en la tabla **3-2** se muestran algunos ejemplos de la usualmente utilizada lista de palabras vacías de van Rijsbergen [Van Rijsbergen, 1977].

Tabla 3-2.: Ejemplos de la lista clásica de *stopwords* de van Rijsbergen [Van Rijsbergen, 1977].

a	about	above	across
after	afterwards	again	against
all	almost	alone	along
(...)			
such	than	that	the
their	them	themselves	then
thence	there	thereafter	(...)

El proceso de eliminación de estas palabras permitirá hacer un análisis más exhaustivo sobre las palabras que si agregan significado y contexto al los textos evaluados y de esta manera generar palabras clave más contundentes. *Python* cuenta con multiples librerias que tienen predeterminadamente la lista de palabras vacías mencionadas anteriormente y de esta forma permite hacer su eliminación, sin embargo, existen palabras que no hacen parte de esta lista y que posiblemente no agreguen valor al estudio realizado, por ejemplo, al ser este un corpus con temas relacionados a la ciencia de datos, se puede observar que la palabra «datos» es una palabra muy frecuente, pero por contexto, no es una palabra que pueda generar palabras clave relevantes al saber previamente que esta es una palabra que se ve frecuentemente.

3.1.2. Tokenización

La tokenización es un paso crítico en el procesamiento de lenguaje natural que implica dividir un texto en unidades más pequeñas llamadas"tokens". Estos tokens pueden ser palabras individuales, frases o incluso caracteres, dependiendo de la granularidad deseada. Un ejemplo sencillo de tokenización basada en palabras es la siguiente oración: "La tokenización es esencial para procesar texto". Al aplicar la tokenización, obtendríamos una lista de tokens como ["La", "tokenización", "es", "esencial", "para", "procesar", "texto"].

3.1.3. Stemming

El *stemming* es un proceso clave en el preprocesamiento de datos textuales que implica reducir las palabras a sus raíces o formas base. Este método es especialmente útil para normalizar

las palabras y agrupar las variaciones léxicas de una misma palabra. Un ejemplo común es el uso de algoritmos de stemming para reducir palabras a su raíz, eliminando prefijos y sufijos. Por ejemplo, las palabras "*running*", "*runner*" y "*ran*"se reducirían a la misma raíz "*run*"mediante el proceso de *stemming*. Esto facilita la tarea de análisis de texto al agrupar palabras relacionadas y reducir la complejidad léxica.

Uno de los algoritmos de *stemming* más conocidos es el algoritmo de Porter [Willett, 2006], desarrollado por Martin Porter. Este algoritmo se ha convertido en un estándar en la industria del procesamiento de lenguaje natural. El *stemming* es esencial para mejorar la eficacia de las tareas de recuperación de información, análisis de sentimientos y otros procesos de minería de texto al simplificar la representación de las palabras y aumentar la coherencia entre términos similares.

3.1.4. Lematización

La lematización es un proceso de normalización en el procesamiento de lenguaje natural que consiste en reducir una palabra a su forma base o lema. A diferencia del *stemming*, que simplemente corta los sufijos para llegar a una forma base, la lematización considera la morfología de las palabras y las reduce a su forma más general o canónica. Por ejemplo, en inglés, la lematización convertiría "*running*" a "*run*", "*better*" a "*good*", y "*swimming*" a "*swim*".

Este proceso es especialmente útil cuando se desea reducir las palabras a su forma más fundamental, facilitando la identificación de similitudes semánticas. Un ejemplo práctico sería en la tarea de búsqueda de información, donde la lematización podría agrupar todas las formas conjugadas de un verbo bajo un mismo lema, mejorando la relevancia y coherencia de los resultados obtenidos. Además, la lematización ayuda a eliminar la redundancia y simplificar el análisis de texto al reducir las palabras a su forma más básica y comprensible.

3.2. Generación de palabras clave

4. Construcción mapas de conocimiento

5. Ejemplo práctico y resultados

6. Conclusiones y recomendaciones

6.1. Conclusiones

6.2. Recomendaciones

A. Anexo: Nombrar el anexo A de acuerdo con su contenido

B. Anexo: Nombrar el anexo B de acuerdo con su contenido

Bibliografía

[Blanchard, 2007] Blanchard, A. (2007). Understanding and customizing stopword lists for enhanced patent mapping. *World Patent Information*, 29(4):308–316.

[Citation Network Dataset., 2024] Citation Network Dataset. (2024). Dblp-citation-network v10 `https://paperswithcode.com/dataset/dblp`, 22 de Enero de 2024.

[Cobo et al., 2012] Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E., and Herrera, F. (2012). Scimat: A new science mapping analysis software tool. *Journal of the American Society for information Science and Technology*, 63(8):1609–1630.

[Kleinberg, 2002] Kleinberg, J. (2002). Bursty and hierarchical structure in streams. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 91–101.

[Lusher et al., 2023] Lusher, L., Yang, W., and Carrell, S. E. (2023). Congestion on the information superhighway: Inefficiencies in economics working papers. *Journal of Public Economics*, 225:104978.

[Manning and Schutze, 1999] Manning, C. and Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

[Osborne and Motta, 2015] Osborne, F. and Motta, E. (2015). Klink-2: integrating multiple web sources to generate semantic topic networks. In *The Semantic Web-ISWC 2015: 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part I 14*, pages 408–424. Springer.

[Persson, 1994] Persson, O. (1994). The intellectual base and research fronts of jasis 1986–1990. *Journal of the American society for information science*, 45(1):31–38.

[Price, 2011] Price, D. d. S. (2011). Networks of scientific papers. In *The Structure and Dynamics of Networks*, pages 149–154. Princeton University Press.

[Santana and Cobo, 2020] Santana, M. and Cobo, M. J. (2020). What is the future of work? a science mapping analysis. *European Management Journal*, 38(6):846–862.

[Synnestvedt et al., 2005] Synnestvedt, M. B., Chen, C., and Holmes, J. H. (2005). Citespace ii: visualization and knowledge discovery in bibliographic databases. In *AMIA annual symposium proceedings*, volume 2005, page 724. American Medical Informatics Association.

[Van Rijsbergen, 1977] Van Rijsbergen, C. J. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of documentation*, 33(2):106–119.

[Willett, 2006] Willett, P. (2006). The porter stemming algorithm: then and now. *Program*, 40(3):219–223.