Optimizing academic literature review using Text Mining in R: An automated approach

Henry Osorto[1,2]
[1]Universidad Nacional Autónoma de Honduras, UNAH, Honduras, *henry.osorto@unah.edu.hn*
[2]Facultad de Postgrado, Universidad Tecnológica Centroamericana, UNITEC, Honduras, *henry.osorto@unitec.edu.hn*

***Abstract*– Literature review is crucial for research development, but the growing number of digital publications has complicated this process, increasing the risk of bias. This article aims to develop and apply an automated academic literature review approach using text mining techniques in the R programming language. A database of 86,820 articles published in scientific journals, containing the search terms ("model" AND "growth" AND "economic"), hosted in the open access database Redalyc, was retrieved. A programming syntax was developed that optimized data download, processing and analysis, allowing its replication in future literature review processes. This study demonstrates the potential of text mining tools and automated bibliometric analysis, using R, to optimize literature review in the scientific field.**

***Keywords*-- Literature review, Text Mining, R language.**

Optimización de la revisión de literatura académica mediante Text Mining en R: Un enfoque automatizado

Henry Osorto[1,2]
[1]Universidad Nacional Autónoma de Honduras, UNAH, Honduras, *henry.osorto@unah.edu.hn*
[2]Facultad de Postgrado, Universidad Tecnológica Centroamericana, UNITEC, Honduras, *henry.osorto@unitec.edu.hn*

Resumen– ***La revisión de literatura es crucial para el desarrollo de investigaciones, pero la creciente cantidad de publicaciones digitales ha complicado este proceso, incrementando el riesgo de sesgos. Este artículo tiene como objetivo desarrollar y aplicar un enfoque automatizado de revisión de literatura académica mediante técnicas de text mining en el lenguaje de programación R. Se recuperó una base de datos de 86,820 artículos publicados en revistas científicas, que contenían los términos de búsqueda ("modelo" AND "crecimiento" AND "económico"), alojados en la base de datos de acceso abierto Redalyc. Se desarrolló una sintaxis de programación que optimizó la descarga, el procesamiento y el análisis de datos, permitiendo su replicación en futuros procesos de revisión de literatura. Este estudio demuestra el potencial de las herramientas de text mining y el análisis bibliométrico automatizado, utilizando R, para optimizar la revisión de literatura en el ámbito científico.***

Palabras clave-- Revisión de literatura, Text Mining, lenguaje R.

I. Introducción

La revisión de literatura es una las tareas de mucha importancia en el proceso de investigación científica ya que proporciona un marco conceptual y teórico para el estudio [1], al tiempo permite conocer los marcos de referencia y metodologías que se han desarrollado en un campo o tema de investigación específico. Al llevar a cabo este proceso es común encontrarse con una abrumadora cantidad de artículos relacionados al tema que se está investigando, esto debido a la enorme rapidez con la que está creciendo la publicación digital de información académica [2]. A estos sistemas que contienen información de millones de autores, artículos, citas, figuras, tablas, así como redes académicas y bibliotecas digitales, se le ha denominado *Big Scholarly Data* [2], [3]. Con este contexto y dependiendo del nivel de profundidad con la que se quiere ahondar en un tema, la revisión de literatura podría llegar a tener varios inconvenientes.

Primero, debido al enorme volumen de artículos que podrían recuperarse en una sola consulta, es altamente probable que no se alcance a revisar una cantidad representativa de dichos artículos. Por lo que se podría afirmar que: existe una relación inversamente proporcional entre el número de artículos disponibles en un campo o tema específico, y la cantidad de artículos revisados en el proceso de revisión de literatura. En segundo lugar, se pueden producir sesgos en la selección de artículos, lo cual deriva de elegir solo una parte de ellos, dejando de lado información que pudo haber contribuido significativamente.

Sin embargo, gracias a los avances en el uso de métodos de aprendizaje automático y enorme capacidad de procesamiento que poseen las maquinas, es posible encomendarles tareas que requiere procesar mucha información [4], como suele ocurrir en la revisión de literatura. Por tanto, es posible reducir los problemas de representatividad y sesgos en las revisiones bibliográficas por medio del poder computacional que ofrecen muchas software hoy en día, en los que se puede elaborar datamining, text mining o crear algoritmos que realicen tareas específicas de la revisión de literatura. Así mismo, en la actualidad se han creado muchos sistemas que se usan como métodos para recomendar artículos relevantes, utilizando características relacionadas con la similitud textual, palabras clave e información estructural como la relación en una red de citas [5].

Existen varios ejemplos en el desarrollo de sistemas, uno de ellos es AKMiner (Academic Knowledge Miner) elaborado por Huang and Wan [6], el cual se encarga de extraer automáticamente conocimiento útil de los artículos en un dominio específico, en donde se extraen conjuntamente conceptos y relaciones académicas basándose en redes lógicas de Markov, para luego realizar representaciones visuales por medio de grafos de conocimiento, grafos de nubes de conceptos y grafos de relaciones de conceptos. Otro desarrollo interesante es el elaborado por Portenoy and West [2], quienes comparten por medio de **GitHub** un proyecto en Python en el cual utiliza modelos de aprendizaje supervisado para identificar artículos relevantes para su revisión, derivando características de los metadatos asociados con un artículo. En la referencia [7], profundizan aún más en la identificación de desarrollo de aplicaciones de minería de datos en bibliotecas académicas, ya que en su revisión sistemática analizaron más de cuarenta estudios elaborados entre 1998 y 2014 en los que en los que se emplean técnicas de agrupamiento, asociación, clasificación y regresión y su aplicación en los cuatro aspectos principales de la biblioteca: servicios, calidad, colección y comportamiento de uso.

Los avances en métodos de aprendizaje automático y el crecimiento exponencial del poder computacional han abierto nuevas posibilidades para abordar los desafíos de la revisión de literatura en la investigación científica. En este contexto, el objetivo de este estudio es desarrollar y aplicar un enfoque automatizado de revisión de literatura académica utilizando técnicas de text mining en el lenguaje de programación R. La elección de R como herramienta principal para este análisis se debe a su robusto ecosistema de paquetes especializados, como tidyverse, tm, tidytext, igraph, entre otros, que facilitan el procesamiento, análisis y visualización de grandes volúmenes

de datos tipo texto. R es ampliamente utilizado en la comunidad científica debido a su flexibilidad y accesibilidad, permitiendo una implementación eficiente y replicable de las técnicas de text mining. Este enfoque tiene como objetivo optimizar el proceso de búsqueda, selección y análisis de artículos científicos, permitiendo la extracción y análisis de información relevante de grandes conjuntos de datos académicos. Con esto, se busca facilitar la identificación de tendencias, patrones y relaciones entre los diferentes trabajos de investigación en un campo específico, contribuyendo así al avance del conocimiento en la disciplina.

II. Materiales y Métodos

A. *Datos del estudio*

Para llevar a cabo el estudio se recuperó una base de datos de 86,820 artículos publicados en revistas científicas que contenían los términos de búsqueda ("modelo" AND "crecimiento" AND "económico"). La consulta y recuperación de la base de datos fue realizada el 2 de abril de 2024. Los campos o variables que contiene la base de datos analizada son: año de publicación, nombre de los autores, idioma del artículo, nombre de la institución a la cual está adscrita la revista, nombre de la revista, palabras clave, resumen, título de la investigación, país de la institución.

B. *Sistema de consulta*

La consulta de los términos de búsqueda se realizó en el Sistema de Información Científica de la Red de Revistas Científicas de América Latina y el Caribe, España y Portugal, **Redalyc**. Se seleccionó este sistema debido a la flexibilidad en de la descarga de los resultados de la consulta. A pesar de que el procedimiento no resulta tan intuitivo y manual, es posible tener acceso todos los resultados con mucha más facilidad que otros sistemas; Google Scholar, por ejemplo, ya que este no posee una API pública y se debe recurrir al Web Scraping. Redalyc en cambio, permite acceder una base de datos de los resultados mostrado en cada página, y gracias a las opciones disponibles, permite mostrar desde 10 a 100 artículos por página, lo cual favorece el proceso de recuperación de la base de datos.

Para acceder a la base de datos se recurrió a las herramientas de desarrollo del navegador web (presionando F12). Luego se ingresó a la herramienta Network, la cual permite visualizar las solicitudes de red de la página web, seleccionando puntualmente las solicitudes Fetch/XHR ya que muestran únicamente la solicitud a una API Pública, que devuelve una base de datos estructurada en JSON.

Considerando que la url de la consulta hace referencia a los cien títulos de la página que se visualiza, se debían generar 869 url, de las cuales la única variante se produce en el número de página. Por tanto, el proceso de obtención de las url fue posible de automatizar mediante un ciclo for que va desde 1 (página inicial) a 869 (página final)

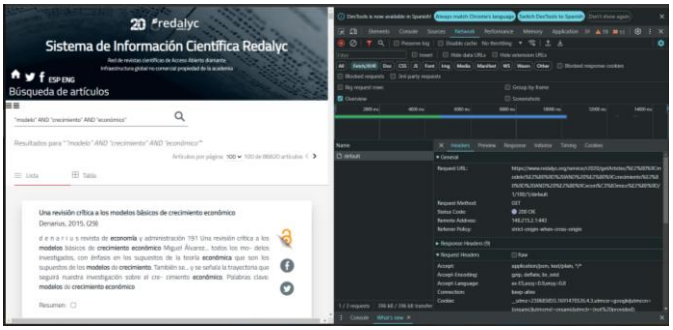

Fig. 1 Consulta de términos de búsqueda en Redalyc y ventana de Herramientas de Desarrollo

C. *Procesamiento y análisis de datos*

Teniendo en cuenta que el propósito de este estudio es dar a conocer el proceso de optimización de la revisión de literatura en R [8], el procesamiento y análisis de la información se llevó a cabo por medio de diversas funciones de este lenguaje de programación. En ese sentido se ira mostrando el código empleado en cada etapa del procesamiento y creación de resultados.

Con relación a la obtención de la base de datos, el punto de partida fue identificar si el sistema de consulta de los términos de búsqueda posee una API Pública ya que por medio de ella se puede acceder a una base de datos estructurada. Como se mencionó antes, se detectó Redalyc cuenta con una API lo cual permitió acceder a los datos en un contenedor JSON, por tanto, se utilizó el paquete jsonlite [9], para la descarga de los datos.

```
# Obtener Datos de la Consulta
# 1. Cargar Librerías ----
R> library(tidyverse)
R> library(jsonlite)

# 2. Importar Bases de datos ----
R> url <-
'https://www.redalyc.org/service/r2020/getArticles/%22modelo%22%20AND%20
%22crecimiento%22%20AND%20%22econ%C3%B3mico%22/1/100/1/default'

R> lista.data <- fromJSON(url)
R> data <- lista.data$resultados

R> n <- 86820 / 100
R> i <- 86820 %% 100
R> n <- trunc(n)

R> if(i > 0){ n <- n + 1}

R> for(i in 2:n) {
   url <-
paste0('https://www.redalyc.org/service/r2020/getArticles/%22modelo%22%20AN
D%20%22crecimiento%22%20AND%20%22econ%C3%B3mico%22/', i,
'/100/0/default')

lista.data <- fromJSON(url)
data2 <- lista.data$resultados
data <- rbind(data, data2) }

R> data <- data %>%
        select(anioArticulo, autores, idiomaArticulo, nomInstitucionRev,
               nomRevista, palabras, resumen, titulo, paisInstitucion)

# 3. Guardar datos en CSV ----
R> setwd('~/')
R> write.csv(data, 'Data Revisión Bibliográfica.csv', row.names = F)
```

Posteriormente se procedió con el proceso de limpieza de los datos que consiste en la eliminación de signos de puntuación, pronombres, números, espacios en blanco, entre otros. En esta etapa fueron implementadas funciones de los paquetes tm [10], para la limpieza de palabras, así como los paquetes dplyr [11] y tidyr [12], para realizar algunas funciones que facilitan el procesamiento y limpieza de los datos.

Luego se procedió a la preparación de los datos para el análisis de texto, como ser la conversión del marco de datos en un objeto tibble, el cual consiste en una clase de marco de datos moderna que no convierte cadenas en factores y no utiliza nombres de filas [13]. Esto favorece el ordenamiento de los datos en formato de texto, definido como tokenización. En principio un token es una unidad de texto significativa, como una palabra de interés para el análisis, por lo que la tokenización consiste en dividir el texto en tokens [13]. Cabe mencionar que el texto se puede tokenizar por palabras individuales, oraciones o n-gramas, donde este último consiste en la separación de secuencias consecutivas de palabras. Este proceso fue posible de realizar gracias al paquete tidytext [14].

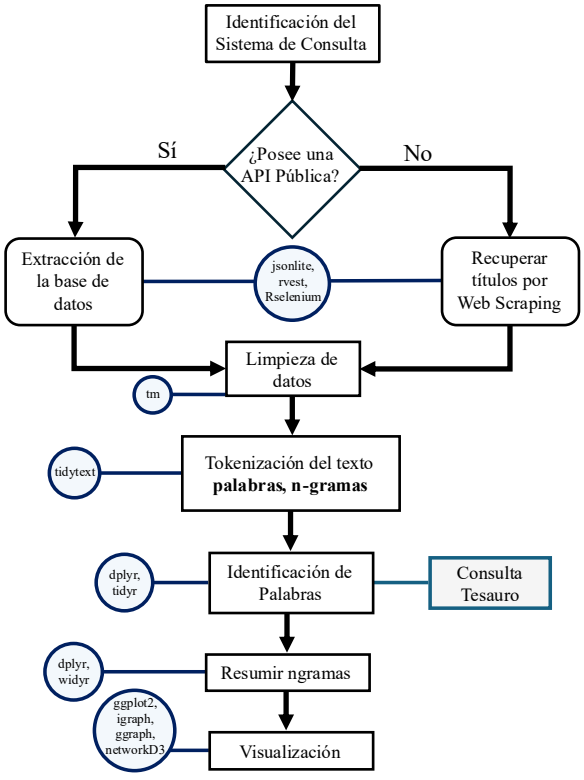

Fig. 2 Modelo de optimización de revisión de literatura mediante text mining en R

El siguiente paso consiste en la identificación de las principales palabras contenidas en las palabras clave, títulos y resumen, por lo que resulta útil la elaboración de tablas y gráficos de frecuencias de palabras, así como las nubes de palabras. En este punto es conveniente realizar una comparación entre las palabras más utilizadas por las investigaciones y las palabras o términos contenidos en un tesauro, el cual consiste en una lista de palabras o términos empleados en la referencia a ciertos conceptos. Para tales efectos se consultó el tesauro de la UNESCO, el cual proporcionó resultados para los términos (crecimiento económico), mostrando así los conceptos relacionados, conceptos específicos y el concepto genérico, así mismo, los términos usados en otras lenguas.

En el siguiente paso se procedió a resumir la información del texto tokenizado, el cual se puede realizar para analizar palabras individuales, así como las secuencias de dos palabras o bígamas. Además, en este punto se pueden realizar resúmenes agrupados por categorías de otras variables de interés, por ejemplo, por países, a fin de conocer los países de donde se han publicado investigaciones con el mayor uso de cierto término. Otro ejemplo es la agrupación de palabras por año de publicación, a fin de conocer la evolución del uso de ciertos términos en el tiempo. Por otro lado, ngramas pueden ayudar a la construcción de redes y correlaciones de palabras. Para este proceso resultó de utilidad el uso del paquete widyr [15].

Finalmente, el último paso de este modelo de optimización consiste en la representación visual del texto resumido, ya sea por palabras, n-gramas (para este caso bígamas), y palabras cruzadas o agrupadas con otras variables de interés. Para este proceso resulta útil el uso de los paquetes ggplot2 [16] dada la flexibilidad para la elaboración de representaciones visuales elegantes, así mismo, los paquetes igraph [17], ggraph [18], networkD3 [19] y svglite [20], para que el guardado de los datos sea mediante formatos vectoriales escalables, lo que brinda posibilidades de ediciones posteriores mediante programas de diseño gráfico y mejora la calidad de la imagen.

III. Resultados

A. *Coocurrencia de palabras clave*

Los resultados que se muestran a continuación, es una parte de los diversos análisis que pueden llevarse a cabo por medio del text mining en la revisión de literatura científica. Si bien es cierto, este análisis no pretende equiparar las métricas desarrolladas en un análisis bibliométrico, sin embargo, se espera que se generen nuevas contribuciones que ayuden a potenciar lo que hoy en este estudio se pudo llegar a realizar. En tal sentido, el análisis del texto se centró en el uso de las palabras clave de los 86,820 títulos recuperados.

```
# Frecuencia de Palabras
R> library(tm)
R> library(tidytext)
R> library(svglite)

# Importar Datos
R> df <- read.csv('Data Revisión Bibliográfica.csv')

# Preparar los Datos  (separar palabras clave en inglés y español)
R> df <- df %>%
      separate(palabras, into = c('Español', 'Ingles'), sep = '>>>') %>%
      select(palabras = Español)
```

```
# Limpieza del texto (Elaborar función para usar en el futuro)
R>  limpieza.texto <- function(x) {
    x <- sapply(x, removePunctuation, USE.NAMES = F)
    x <- sapply(x, tolower, USE.NAMES = F)
    x <- sapply(x, stripWhitespace, USE.NAMES = F)
    x <- sapply(x, removeNumbers, USE.NAMES = F)
    x <- sapply(x, removeWords, words = stopwords('spanish'), USE.NAMES = F) }

R>  df$palabras <- limpieza.texto(df$palabras)

# Tokenizar cadena de palabras
R>  palabras.tokens <- df %>%
        unnest_tokens(word, palabras)
R>  palabras.tokens <- palabras.tokens %>%
        count(word, sort = TRUE)
# stopwords adicionales
R>  my_stop_words <- tibble(word = c('méxico', 'argentina', 'colombia'))

# Gráfico de Frecuencia de Palabras
R>  g <- palabras.tokens %>%
        anti_join(my_stop_words) %>%
        filter(n > 1500) %>%
        ggplot(aes(y = n, x = reorder(word, n, decreasing = F))) +
        geom_col(fill = "#032B47") +
        coord_flip() +
        labs(x = 'Palabras Clave', y = 'n > 1500') +
        theme(axis.text.y = element_text(size = 14))

# Guardar Gráfico
R>  ggsave('Frecuencia de Palabras Clave.svg', g, width = 10, height = 10)
```

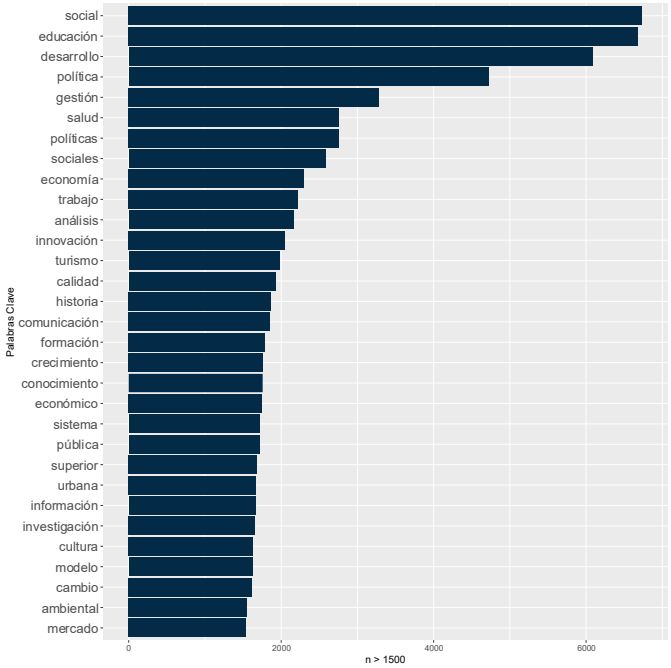

Fig. 3 Frecuencia de palabras clave con n > 1500

B. Coocurrencia de palabras clave por país

El siguiente análisis consistió en identificar la procedencia de los artículos publicados y su conexión con algunos términos particulares. Cabe mencionar que la variable país obtenida de la base, contenía valores numéricos, sin embargo, por medio del nombre de las instituciones se pudo identificar el nombre del país, por lo que se creó un conjunto de datos de países para crear la unión con la base de datos principal.

Al contar con la variable país, se realizó el resumen de los tokens mediante la agrupación de países, de modo que se pudiese obtener una frecuencia de cada uno de los términos por país. Esto permitió realizar un diagrama de Sankey el cual realiza una representación de un flujo o conexión ponderada que van de un nodo a otro. Para este caso el nodo inicial lo representan los países y el final está representado por los términos o conceptos extraídos del análisis de las palabras clave.

Considerando que la cantidad de palabras puede ser abrumadoramente enorme, se consideró para este análisis, la elección de términos en función de los términos o conceptos del tesauro de la UNESCO, así como la elección de algunos términos que se consideraron importantes por visualizar, por ejemplo, el nombre de Solow (derivado de las teorías de crecimiento económico) la palabra innovación (asociado al significado de un choque de una variable endógena en un el análisis impulso-respuesta). Por otro lado, las líneas que se conectan entre nodos indican la cantidad de flujo. De este modo, se puede observar qué países han realizado contribuciones en las cuales han utilizado los términos o palabras claves mostradas en el diagrama.

```
# Palabras Clave por País - Diagrama de Sankey

R> library(networkD3)

# Crear data frame de países
R> países <- data.frame(País = c("Alemania", "Angola", "Argentina", "Bolivia",
                                 "Brasil", "Canadá", "Chile", "Colombia",  "Costa
                                 Rica", "Cuba", "Dinamarca", "Ecuador",
                                 "España", "Estados Unidos", "Eslovenia",
                                 "México", "Panamá", "Paraguay", "Perú",
                                 "Polonia", "Portugal", "Puerto Rico", "Suiza",
                                 "Venezuela", "Uruguay"),
                        paisInstitucion = c(2, 4, 9, 19, 21, 25, 26, 30, 33, 35, 36,
                                            37, 42, 43, 67, 73, 79, 80, 81, 82, 83,
                                            85, 89, 94, 93))

# Unir Datos
R> df <-  df %>%
        left_join(países)

# Limpiar los Datos
R> df <-  df %>%
        separate(palabras, into = c('Español', 'Ingles'), sep = '>>>') %>%
        select(palabras = Español, País)

R> df$palabras <- limpieza.texto(df$palabras)

# Tokenizar cadena de palabras
R>  palabras.tokens <- df %>%
        unnest_tokens(word, palabras) %>%
        group_by(País) %>%
        count(word, sort = TRUE)

# Diagrama de Sankey
R> palabras.tokens <- palabras.tokens %>%
        filter(word %in% c('crecimiento', 'desarrollo', 'económico', 'económica',
                           'innovación', 'modelo', 'solow', 'reactivación')) %>%
        drop_na()

R>  nodes <- data.frame(name=c(as.character(palabras.tokens$País),
                               as.character(palabras.tokens$word)) %>%
        unique())

R>  palabras.tokens$IDsource <- match(palabras.tokens$País, nodes$name)-1

R> palabras.tokens$IDtarget <- match(palabras.tokens$word, nodes$name)-1
```

```
R> g <- sankeyNetwork(Links = palabras.tokens, Nodes = nodes,
                      Source = "IDsource", Target = "IDtarget", Value = "n",
                      NodeID = "name", sinksRight = FALSE, fontSize = 12)
# Guardar Diagrama
R> saveNetwork(p, 'Diagrama de Sankey.html')
```

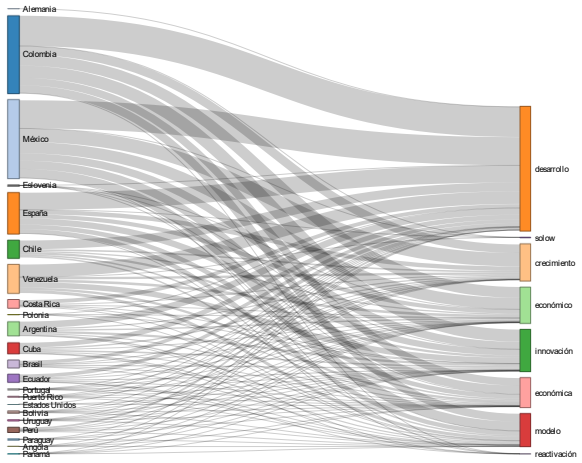

Fig. 4 Diagrama de Sankey entre países y uso de términos o palabras claves en artículos científicos

C. *Coocurrencia de palabras en el tiempo*

En el siguiente análisis se hizo uso del texto contenido en el resumen de los artículos científicos, el cual permite tener una mayor cantidad de palabras que pueden evidenciar objetivos, las metodologías empleadas y los hallazgos de las investigaciones. En tal sentido, el análisis exploratorio de los resúmenes podría ayudar a responder preguntas importantes como ¿cuáles han sido los objetivos desarrollados con mayor frecuencia en los artículos recuperados?, ¿cuáles han sido los métodos, modelos, fuentes de información, y análisis empleados con mayor frecuencia en las investigaciones?, así como ¿cuáles son los principales hallazgos o conclusiones a las que han llegado los estudios consultados?. Fue así como basado en este enfoque se eligieron algunos términos identificados en el tesauro y en ciertos términos relacionados con aspectos metodológicos. Luego se hizo la agrupación de estos términos por la variable año de publicación, a fin de conocer el uso que se les ha dado a estos a lo largo del tiempo.

Las palabras que se eligieron obedecen principalmente a los aspectos métodos de aplicados en las investigaciones del tema. Por tanto, se seleccionó la palabra impulso, haciendo referencia a la cuantificación del impulso-respuesta de un choque de varianza en los modelos de vectores autorregresivos; consecuentemente se eligió la palabra var, que son las siglas en ingles de estos modelos. También se escogieron las palabras modelo y regresión que hacen referencia a las herramientas estadísticas implementadas en los estudios recuperados. La palabra educación se seleccionó debido al uso de esta variable en los modelos de crecimiento económico.

```
# Frecuencia de Palabras en el Tiempo
R> library(tibble) # Ya cargado en tidyverse

R>  df <- df %>%
      filter(idiomaArticulo == 'Español') %>%
      select(anioArticulo, resumen) %>%
      mutate(resumen = gsub('en:', '{', resumen)) %>%
      separate(resumen, into = c('resumen', 'abstract'), sep = '{')

R>  palabras <- tibble(line = 1:nrow(df), año = df$anioArticulo,
                       resumen =df$resumen)

# Limpieza del texto
R>  df$resumen <- limpieza.texto(df$resumen)

# Tokenizar cadenas de textos
R>  palabras.tokens <- palabras %>%
      unnest_tokens(word, resumen) %>%
      group_by(año) %>%
      count(word, sort = TRUE)

# Gráfico de uso de palabras en el tiempo
R>  g <- palabras.tokens %>%
      filter(word %in% c('crecimiento', 'socioeconómicas', 'modelo', 'regresión',
                         'desarrollo', 'var', 'impulso', 'solow', 'educación')) %>%
      ggplot(aes(x = año, y = n)) +
      geom_point() +
      geom_smooth() +
      facet_wrap(~ word, scales = "free_y") +
      labs(y = 'Frecuencia de Palabras', x = 'Año') +
      theme(strip.text = element_text(size = 20),
            axis.title.x = element_text(size = 18),
            axis.text.x = element_text(size = 18),
            axis.title.y = element_text(size = 18),
            axis.text.y = element_text(size = 18))

# Guardar Gráfico ----
R>  ggsave('Uso de palabras en el tiempo.svg', g, width = 15, height = 10)
```

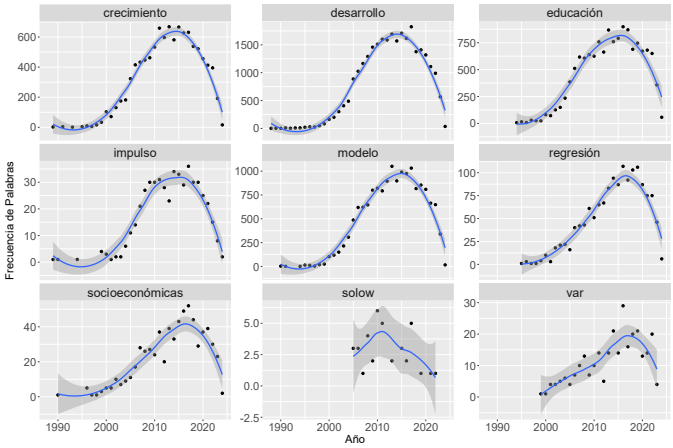

Fig. 5 Análisis del texto del resumen de la investigación por año de publicación

D. *Red de coocurrencia*

Ahora el siguiente análisis de texto consiste en las relaciones entre palabras por medio de los n-gramas, el cual brinda la utilidad de determinar qué palabras tienden a seguir a otras palabras inmediatamente, o bien, qué palabras tienden a coexistir en diferentes documentos [10], lo cual permite la creación de diagramas de red y en este caso diagrama de red de palabras. Para este análisis se utilizaron las palabras clave de los artículos científicos, sin embargo, no se debe desechar la

idea de hacer uso del resumen en el análisis de red de palabras, ya que pueden aportar mucha más información acerca de los objetivos, metodología y hallazgos de los artículos recuperados.

```
# Diagrama de Red de Palabras Clave
R> library(widyr)
R> library(igraph)
R> library(ggraph)

R>  df <- df %>%
      mutate(id = 1:nrow(df)) %>%
      separate(palabras, into = c('Español', 'Ingles'), sep = '>>>') %>%
      select(palabras = Español, id)

# Limpieza del texto
R>  df$palabras <- limpieza.texto(df$palabras)
R>  my_stop_words <- tibble(word = c('méxico', 'américa', 'colombia', 'argentina',
                                     'costa', 'buenos', 'europea', 'palabras', 'clave',
                                     'estados', 'unidos', 'jel', 'siglo', 'tesauro'))
# Tokenizar cadenas de textos
R>  tokens.palabras <- df %>%
      unnest_tokens(word, palabras) %>%
      anti_join(my_stop_words) %>%
      count(word, sort = TRUE)

# Crear n-gramas (bigramas)
R>  palabras_word_pairs <- tokens.palabras %>%
      pairwise_count(word, id, sort = TRUE, upper = FALSE)

# Crear Diagrama de Red ----
R>  subt <- palabras_word_pairs %>%
      filter(n >= 200)

R>  vert <- subt %>%
      gather(item, word, item1, item2) %>%
      group_by(word) %>% summarise(n = sum(n))

R>  g <- subt %>%
      graph_from_data_frame(vertices = vert) %>%
      ggraph(layout = "fr") +
      geom_edge_arc(strength = 0.8, width = 0.5, alpha = 0.5) +
      geom_node_point(aes(size = n, color = n), show.legend = FALSE) +
      scale_size(range = c(2,20)) +
      geom_node_text(aes(label = name, size = n/2), repel = TRUE,
                     point.padding = unit(0.2, "lines"), show.legend = FALSE) +
      theme_void()

# Guardar Gráfico ----
R>  ggsave('Red de Palabras Clave.svg', g, width = 15, height = 10)
```

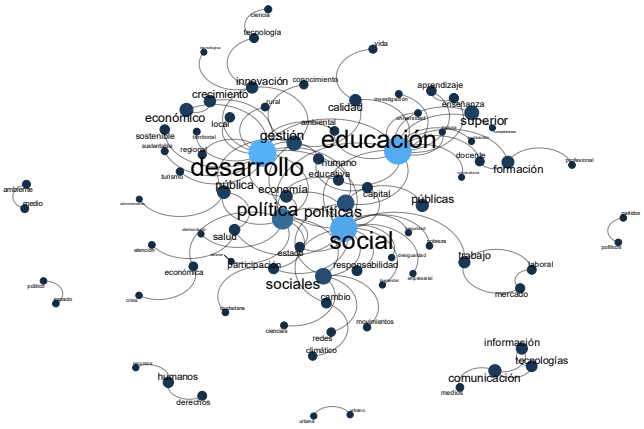

Fig. 6 Diagrama de red de palabras clave en investigaciones de modelo de crecimiento económico

Una de las principales utilidades que brinda el uso de los n-gramas y los diagramas de red en esta optimización de la revisión de literatura, es que puede ayudar a identificar las principales áreas temáticas o conceptuales que están siendo investigadas con relación al tema del cual se realiza la revisión, así mismo, puede ayudar a identificar tendencias emergentes o áreas de investigación interconectadas en el campo. En la figura 6 se puede observar cuales son las palabras que más se han utilizado y cómo estas se conectan con otras palabras que denotan una importante relación.

Por ejemplo, se puede observar una fuerte asociación entre "desarrollo" y "económico", lo que sugiere un enfoque centrado en el crecimiento económico como motor del desarrollo. Asimismo, se identifica una conexión cercana entre "desarrollo" y "sostenible", lo que indica una creciente atención a la sostenibilidad ambiental en el contexto del desarrollo económico. Además, se observan conexiones entre términos relacionados con políticas públicas y sociales, como "política económica", "política social" y "política pública". Estas asociaciones sugieren una intersección entre las políticas económicas y sociales en el contexto del desarrollo, destacando la importancia de abordar los aspectos económicos y sociales de manera integral en las políticas de desarrollo.

También se encuentran conexiones entre "desarrollo" y términos relacionados con aspectos territoriales y locales, como "desarrollo territorial" y "desarrollo local". Estas asociaciones apuntan a la importancia de considerar las dimensiones territoriales y locales en las estrategias de desarrollo, reconociendo la diversidad de contextos y realidades regionales. Por último, se destaca la conexión entre "política educativa" y "educación superior", lo que indica un interés en el papel de la educación en el desarrollo económico y social. Estas asociaciones subrayan la importancia de las políticas educativas orientadas hacia la promoción del acceso a la educación superior como medio para impulsar el desarrollo humano y económico. En conjunto, estos hallazgos resaltan la complejidad y la interconexión de los distintos aspectos involucrados en el desarrollo económico y político, enfatizando la necesidad de enfoques integrales y multidisciplinarios para abordar los desafíos del desarrollo en la actualidad.

IV. Discusión y Conclusiones

Los análisis que se han generado brindaron un demostración del alcance que se puede lograr en la revisión de literatura mediante el uso de herramientas del text mining y en este caso mediante la programación en R. Aunque se ha logrado optimizar la revisión de gran volumen de información y automatizar algunos de los procesos, este enfoque puede requerir de ciertas tareas manuales diferentes a las aquí mostradas, en relación con la programación de otros análisis en R.

Al comparar esta metodología de revisión de literatura, se puede clasificar como un enfoque no estructurado pero automatizado, al igual que otros ya mencionados. Por otro lado, se encuentran las revisiones sistemáticas que dependiendo de su

objetivo pueden ser clasificadas en resumen del conocimiento, integración de datos, construcción de explicaciones y evaluaciones criticas [21], [22], [23], [24]; lo cual las caracteriza en metodologías estructuradas pero automatizadas, ya que las tareas son manuales. Luego se encuentran los análisis bibliométricos los cuales utilizan métodos cuantitativos y herramientas digitales que aprovechan el poder computacional y la capacidad de procesamiento de datos para identificar patrones, tendencias y relaciones en la literatura científica de manera automatizada [25], [26], [27]. Debido al desarrollo que ha existido en el campo de la bibliometría, esta se puede caracterizar como una metodología estructurada y automatizada.

Además de su eficacia metodológica, este estudio ofrece una contribución significativa al destacar la importancia de la literatura en español en el ámbito de la investigación científica. En un contexto donde muchas veces los análisis bibliométricos se basan en bases de datos dominadas por artículos en inglés (Scopus, Web of Science, etc.), este enfoque ofrece una perspectiva particular al incluir una amplia gama de literatura en español recuperada de Redalyc. Esto no solo enriquece la diversidad lingüística en la investigación, sino que también asegura una representación más equitativa de las contribuciones científicas a nivel global.

Además, el uso de literatura en español ofrece la oportunidad de descubrir contribuciones importantes a la ciencia que podrían pasar desapercibidas en contextos dominados por el inglés. Así, este estudio no solo presenta una metodología innovadora, sino que también promueve la relevancia y visibilidad de la literatura en español en el panorama científico internacional.

En conclusión, este estudio demuestra el potencial del uso de herramientas de text mining y análisis bibliométrico automatizado, utilizando el lenguaje de programación R, para optimizar el proceso de revisión de literatura en el ámbito científico. La aplicación de estas técnicas ha permitido abordar eficazmente el desafío de manejar grandes volúmenes de información académica y ha proporcionado una visión amplia y detallada de los patrones, tendencias y relaciones en la literatura científica, especialmente en el contexto de la producción académica en español.

Al integrar la literatura en español recuperada de Redalyc, este enfoque no solo enriquece la diversidad lingüística en la investigación, sino que también asegura una representación más equitativa de las contribuciones científicas a nivel global. Además, el estudio destaca la importancia de considerar la literatura en español para descubrir contribuciones importantes a la ciencia que podrían pasar desapercibidas en contextos dominados por el inglés. En última instancia, este trabajo ofrece una metodología valiosa y una perspectiva innovadora que promueve la relevancia y visibilidad de la literatura en español en el panorama científico internacional, contribuyendo así al avance del conocimiento y la investigación en el ámbito científico.

REFERENCES

[1] D. N. Boote and P. Beile, "Scholars Before Researchers: On the Centrality of the Dissertation Literature Review in Research Preparation," *Educational Researcher*, vol. 34, no. 6, pp. 3–15, 2005, doi: 10.3102/0013189X034006003.

[2] J. Portenoy and J. D. West, "Constructing and evaluating automated literature review systems," *Scientometrics*, vol. 125, no. 3, pp. 3233–3251, 2020, doi: 10.1007/s11192-020-03490-w.

[3] F. Xia, W. Wang, T. M. Bekele, and H. Liu, "Big Scholarly Data: A Survey," *IEEE Trans. Big Data*, vol. 3, no. 1, pp. 18–35, 2017, doi: 10.1109/TBDATA.2016.2641460.

[4] F. N. Silva, D. R. Amancio, M. Bardosova, L. F. Da Costa, and O. N. Oliveira, "Using network science and text analytics to produce surveys in a scientific topic," *Journal of Informetrics*, vol. 10, no. 2, pp. 487–502, 2016, doi: 10.1016/j.joi.2016.03.008.

[5] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "Research-paper recommender systems: a literature survey," *Int J Digit Libr*, vol. 17, no. 4, pp. 305–338, 2016, doi: 10.1007/s00799-015-0156-0.

[6] S. Huang and X. Wan, "AKMiner: Domain-Specific Knowledge Graph Mining from Academic Literatures," in *Lecture Notes in Computer Science, Web Information Systems Engineering – WISE 2013*, D. Hutchison et al., Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 241–255.

[7] L. Siguenza-Guzman, V. Saquicela, E. Avila-Ordóñez, J. Vandewalle, and D. Cattrysse, "Literature Review of Data Mining Applications in Academic Libraries," *The Journal of Academic Librarianship*, vol. 41, no. 4, pp. 499–510, 2015, doi: 10.1016/j.acalib.2015.06.007.

[8] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria. [Online]. Available: https://www.r-project.org/

[9] J. Ooms, "The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects," 2014. [Online]. Available: http://arxiv.org/pdf/1403.2805.pdf

[10] I. Feinerer, K. Hornik, and D. Meyer, "Text Mining Infrastructure in R," *J. Stat. Soft.*, vol. 25, no. 5, 2008, doi: 10.18637/jss.v025.i05.

[11] H. Wickham, R. François, L. Henry, K. Müller, and D. Vaughan, *dplyr: A Grammar of Data Manipulation.* [Online]. Available: https://cran.r-project.org/package=dplyr

[12] H. Wickham, D. Vaughan, and M. Girlich, *tidyr: Tidy Messy Data.* [Online]. Available: https://cran.r-project.org/package=tidyr

[13] J. Silge and D. Robinson, *Text mining with R: A tidy approach*. Sebastopol, CA: O'Reilly Media, 2017.

[14] J. Silge and D. Robinson, "tidytext: Text Mining and Analysis Using Tidy Data Principles in R," *JOSS*, vol. 1, no. 3, p. 37, 2016, doi: 10.21105/joss.00037.

[15] D. Robinson and J. Silge, *widyr: Widen, process, then re-tidy data (R package version 0.1. 5)*, 2022.

[16] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. [Online]. Available: https://ggplot2.tidyverse.org/

[17] G. Csardi and T. Nepusz, "The igraph software package for complex network research," Complex Systems, 2006. [Online]. Available: https://igraph.org/

[18] T. L. Pedersen, *ggraph: An Implementation of Grammar of Graphics for Graphs and Networks.* [Online]. Available: https://cran.r-project.org/package=ggraph

[19] J. Allaire, C. Gandrud, K. Russell, and C. Yetman, *networkD3: D3 JavaScript Network Graphs from R.* [Online]. Available: https://cran.r-project.org/package=networkD3

[20] H. Wickham, L. Henry, T. Lin Pedersen, T. J. Luciani, M. Decorde, and V. Lise, *svglite: An 'SVG' Graphics Device.* [Online]. Available: https://cran.r-project.org/package=svglite

[21] F. J. García-Peñalvo, "Desarrollo de estados de la cuestión robustos: Revisiones Sistemáticas de Literatura," *Educ. Knowl. Soc.*, vol. 23, 487-502, 2022, doi: 10.14201/eks.28600.

[22] D. Papaioannou, A. Sutton, and A. Booth, "Systematic approaches to a successful literature review," *Systematic approaches to a successful literature review*, pp. 1–336, 2016.

[23] G. Paré, M.-C. Trudel, M. Jaana, and S. Kitsiou, "Synthesizing information systems knowledge: A typology of literature reviews," *Information & Management*, vol. 52, no. 2, pp. 183–199, 2015, doi: 10.1016/j.im.2014.08.008.
[24] R. Whittemore, A. Chao, M. Jang, K. E. Minges, and C. Park, "Methods for knowledge synthesis: an overview," *Heart & lung : the journal of critical care*, vol. 43, no. 5, pp. 453–461, 2014, doi: 10.1016/j.hrtlng.2014.05.014.
[25] W. Blockmans, L. Engwall, and D. Weaire, *Bibliometrics: Use and abuse in the review of research performance*. London: Portland Press, 2014.
[26] L. Bredahl, "Introduction to bibliometrics and current data sources," *Library Technology Reports*, vol. 58, no. 8, pp. 5–11, 2022.
[27] A. Żarczyńska, "Nicola De Bellis: Bibliometrics And Citation Analysis, from the Science Citation Index to Cybermetrics, Lanham, Toronto, Plymouth 2009," *TSB*, vol. 5, 1 (8), 2012, doi: 10.12775/TSB.2012.009.