Análisis Bibliométrico de la producción científica de inteligencia artificial aplicada al análisis de videos

Carlos Schenone

6/9/2022

Contents

Resumen

Insertar acá un breve resumen (menos de 2000 palabras) del tema del TP final.

Análisis bibliométrico de la producción científica de inteligencia artificial aplicada al análisis de videos". Se plantea el uso de las técnicas de bibliometría para abordar la revisión sistemática debido a la gran cantidad de artículos.

Uno de los objetivos de la bibliometría es encontrar patrones o regularidades en un corpus o grupo de producción científica grande (áreas o tópicos de investigación relevantes, centros de investigación, autores, coautores y países de mayor producción, entre otros)

Abstract

Insertar acá un breve resumen (menos de 2000 palabras) del tema del TP final.

1) Introducción

Conceptos de IA

2) Metodología

La metodología aplicada fue de tipo descriptivo con enfoque cuantitativo. Se usaron indicadores bibliométricos para realizar el estudio. El primer paso fue desarrollar la ecuación de búsqueda y probarla en la base de datos utilizada, que fue SCOPUS; luego de obtener los resultados se aplicaron los indicadores bibliométricos; después, los resultados se analizaron y se obtuvieron las conclusiones.

2.1) Descripción del contexto

1) Descripción del Laboratorio

1) Describir la técnica de extracción a utilizar. En principio será la consulta a BD a través de APIs, para facilitar la reproducibilidad del experimento. Por lo cual se reduce el universo de candidatos a aquellos que dispongan de APIs abiertas.
2) Descripción del hardware y software a utilizar

2) Selección de las BD a consultar

1) Definición de la/las bases de datos. En este caso WebofScience y otras que dispongan de APIs abiertas.
2) Descripción de los filtros configurables en la API disponible para cada BD (armar una tabla comparativa).
3) Análisis de las características del set de datos obtenido (nombres y formato de los metadatos, entre otros). A fin de apoyar el proceso de homogenización de los distintos set de datos.

3) Definición y análisis de los metadatos de interés.

1) Definición del conjunto coincidencia entre los metadatos deseables y los filtros disponibles en las APIs.
2) Estandarización de campos (para facilitar el proceso de unificación de los sets de datos obtenidos por las consultas a las distintas BD definidas).
3) Estandarízación del formato de salida (se define el estándar csv)

2.2) Descripción de los criterios de extracción, tratamiento y limpieza de los datos

1) Construcción de ecuación de búsqueda.

Para la construcción de la ecuación de búsqueda se tomaron de base los temas de la investigación, que son: Inteligencia Artificial, Política Pública, Toma de Decisiones y Agricultura. A partir de este pundo se buscaron las palabras clave que los autores y las publicaciones utilizaban para referenciar el tema. Al final se obtuvo la ecuación, con la cual se obtuvieron 110 publicaciones. La ecuación de búsqueda se presenta a continuación.

2) Aplicación de indicadores bibliométricos

Con los datos bibliográficos de las 110 publicaciones que resultaron, se aplicaron los siguientes indicadores bibliométricos.

Histórico de publicaciones. Este indicador muestra el número de publicaciones realizadas año a año.

Autores de publicaciones. Este muestra el listado de autores y el número de publicaciones científicas de cada uno.

Afiliación de autores. Este muestra las organizaciones a las cuales pertenecen los autores y el número de publicaciones asociadas a cada organización.

País de origen. Este presenta el número de publicaciones por país, acorde al país donde se realizó la publicación.

Tipo de publicación. Este indicador muestra el número de publicaciones clasificadas por tipo de medio, por ejemplo revista especializada, memoria de conferencia, revisión, libro o capítulo de libro.

Patrocinador. Este indica los patrocinadores que financiaron las investigaciones que generaron las publicaciones. El número corresponde a las publicaciones que tiene asociadas cada patrocinador.

Palabras clave. Para las palabras clave se usó el método de fuerza de asociación. Este método busca determinar cuáles categorías se relacionan más entre ellas y se agrupan acorde a ello. Para ello se usó el software libre VosViewer, el cual es un software para análisis biblimétrico que ayuda a crear redes bibliométricas para ver gráficamente estas relaciones.

Citación de autores. Este indicador presenta cómo se relacionan y son citados los autores. El método utilizado para encontrar las relaciones y los grupos es el ya nombrado fuerza de asocación, para lo cual se usó también el software VosViewer.

3) Extracción de datos

3.1) Configurar las APIs según los criterios definidos en la Etapa 3 3.2) Ejecutar el proceso de extracción

4) Depuración de los set de datos

4.1) Validar metadatos perdidos 4.2) Validar que no existan elementos repetidos (por ejemplo un autor que aparece con distintas versiones del nombre, nombre largo y corto).

5) Consolidación de los set de datos

5.1) Unificar los sets de datos obtenidos por las consultas en un solo set de datos (tibble)

6) Ajuste del set de datos de acuerdo a la capacidad de procesamiento de la infraestructura disponible.

6.1) Particionar los resultados obtenidos en rangos de años. En caso que el set de datos obtenido supere la capacidad de procesamiento del equipo disponible, se propone repetir el procedimiento dividiendo los resultados según un intervalo de tiempo, por ejemplo 10 años. En caso que el resultado aún no se pueda procesar, se deberá probar con intervalos más pequeños hasta lograr una cantidad de elementos gestionable por la infraestructura (PC) disponible.

2.3) Análisis de resultados

Los resultados obtenidos con los indicadores bibliométricos se analizaron y se obtuvieron las conclusiones. Apoyados por herramientas de visualización, por ejemplo VOSviewer.

3) Análisis de los resultados

Presentar lo encontrado: patrones o regularidades en un corpus o grupo de producción científica grande (esto puede ser red de coautores, importancia de determinados países en un determinado ámbito, entre otros)

3.1) Estructura de los datos

Describir brevemente los datos (tidy data): - Cuántas variables tienen? - De qué tipo es cada variable? - Cuántas observaciones?

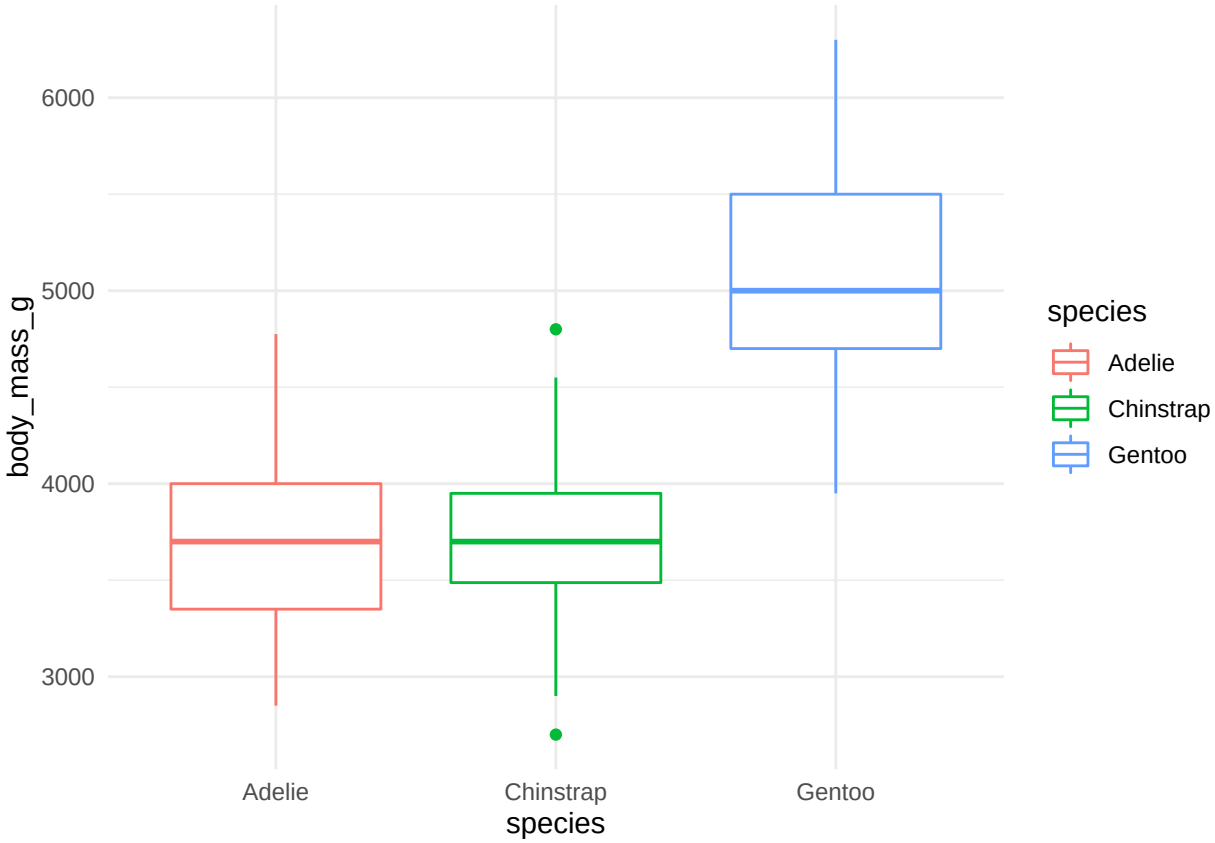
En caso de que ya tengan los datos en formato tidy como para cargar en **R**, se puede acompañar con algún breve análisis exploratorio. Por ejemplo:

```
library(palmerpenguins)
str(penguins)
```

```
## tibble [344 x 8] (S3: tbl_df/tbl/data.frame)
##  $ species          : Factor w/ 3 levels "Adelie","Chinstrap",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ island           : Factor w/ 3 levels "Biscoe","Dream",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ bill_length_mm   : num [1:344] 39.1 39.5 40.3 NA 36.7 39.3 38.9 39.2 34.1 42 ...
##  $ bill_depth_mm    : num [1:344] 18.7 17.4 18 NA 19.3 20.6 17.8 19.6 18.1 20.2 ...
##  $ flipper_length_mm: int [1:344] 181 186 195 NA 193 190 181 195 193 190 ...
##  $ body_mass_g      : int [1:344] 3750 3800 3250 NA 3450 3650 3625 4675 3475 4250 ...
##  $ sex              : Factor w/ 2 levels "female","male": 2 1 1 NA 1 2 1 2 NA NA ...
##  $ year             : int [1:344] 2007 2007 2007 2007 2007 2007 2007 2007 2007 2007 ...
```

```
penguins %>% ggplot(aes(x = species,
                        y = body_mass_g,
                        color = species)) +
  geom_boxplot() +
  theme_minimal()
```

```
## Warning: Removed 2 rows containing non-finite values (stat_boxplot).
```



3.2) Identificar las Áreas de investigación estudiadas.

A fin de enfocarse en las áreas que producen la mayor cantidad de artículos. (Tabla 1)

3.3) Identificar los años de explosión de artículos (Yearly growth rate)

3.4) Identificar el origen de la producción (los países)

3.5) Identificar las revistas y libros más importantes sobre el tema.

Es importante identificar las fuentes de producción cuando se está haciendo un artículo bibliométrico, porque sería un potencial espacio donde publicar el artículo que se está realizando)

3.6) Detectar los artículos más citados y los autores más productivos (con mayor cantidad de artículos)

3.7) Potencialmente: Utilizar el índice de concentración de autores, como una especie de entropía, definido en el artículo publicado en IDEAS.

3.8) Análisis de las visualizaciones

1) Enfocados en los artículos. Representando los artículos por nodos, junto con el tamaño del nodo representando la cantidad de citas que recibe ese nodo. El nodo se etiqueta con la palabra clave más característica. Análisis de las visualizaciones (por ejemplo, en el artículo de bitcoin se observó que hay tres cluster importantes, uno más vinculado a business economics, otra que era mas computer science pero divididos en una parte que se dedicaba a blockchain duro (más teórico) y una parte más aplicada a los protocolos que se aplicaban a criptomonedas en concreto y bitcoin en particular.
2) Enfocados en los autores. Donde se destaquen los autores y coautores mediante citas.

4) Conclusiones

Resultados destacados y Próximos pasos.