Trabajos, Comunicaciones y Conferencias

Actas del Workshop Iberoamericano de Estudios métricos de la actividad científica orientada a temas locales/regionales

Sandra Miguel
(coordinadora)

FaHCE
FACULTAD DE HUMANIDADES Y CIENCIAS DE LA EDUCACIÓN
UNIVERSIDAD NACIONAL DE LA PLATA

ACTAS DEL WORKSHOP IBEROAMERICANO DE ESTUDIOS MÉTRICOS DE LA ACTIVIDAD CIENTÍFICA ORIENTADA A TEMAS LOCALES/REGIONALES

La Plata, 21 y 22 de agosto de 2018

Sandra Miguel
(coordinadora)

FaHCE
FACULTAD DE HUMANIDADES Y CIENCIAS DE LA EDUCACIÓN
UNIVERSIDAD NACIONAL DE LA PLATA

2019

Diseño: D.C.V. Celeste Marzetti
Tapa: D.G.P. Daniela Nuesch
Editora por Prosecretaría de Gestión Editorial y Difusión: Leslie Bava

Queda hecho el depósito que marca la Ley 11.723

ISBN: 978-950-34-1742-3
Colección: Trabajos, comunicaciones y conferencias, 37

Cita sugerida: Miguel, S. (Coord.). (2019). *Actas del Workshop Iberoamericano de estudios métricos de la actividad científica orientada a temas locales/regionales* (2018 : La Plata). La Plata : Universidad Nacional de La Plata. Facultad de Humanidades y Ciencias de la Educación. (Trabajos, comunicaciones y conferencias ; 37). Recuperado de https://www.libros.fahce.unlp.edu.ar/index.php/libros/catalog/book/130

Índice

Prólogo

Este libro de actas reúne las intervenciones presentadas en el **Workshop Iberoamericano de estudios métricos de la actividad científica orientada a temas locales/regionales**, realizado en la ciudad de La Plata, Argentina, el 21 y 22 de agosto de 2018.

El evento estuvo organizado por el Instituto de Investigaciones en Humanidades y Ciencias Sociales (IdIHCS- UNLP/CONICET) en el marco del proyecto PICT 2015-2144 "La producción científica sobre los países latinoamericanos. Aproximación a su estudio desde una perspectiva bibliométrica y relación con indicadores del contexto económico y social", acreditado por la Agencia Nacional de Promoción Científica y Tecnológica del Ministerio de Educación, Cultura y Ciencia y Tecnología de la República Argentina. Para la realización de la reunión se contó con el financiamiento de este proyecto y del subsidio para reuniones científicas RC 2017-0323 otorgado por la Agencia Nacional de Promoción Científica y Tecnológica.

El Workshop reunió investigadores con diferentes perfiles formativos y trayectorias, cuyas intervenciones permitieron generar un espacio compartido de debate y un intercambio de perspectivas teóricas y metodológicas relacionadas con las políticas científicas y con la obtención de métricas y visualizaciones derivadas de la producción y su relación con indicadores del contexto económico y social. Cabe destacar que aunque el interés en los estudios sobre América Latina y España no es nueva, sí es novedoso que desde las regiones periféricas se indague con una perspectiva bibliométrica la producción científica que se realiza sobre los países de la región, sea ésta producida en el propio país o llevada a cabo desde el extranjero.

Las exposiciones se desarrollaron en los siguientes ejes temáticos:

- Presencia de temas orientados a lo local / regional en las políticas y agendas de investigación
- Aproximaciones metodológicas para el abordaje cienciométrico de la producción en temas locales / regionales
- Estudios métricos sobre la producción científica en temas locales de países iberoamericanos

Finalmente agradecer a todos los expositores y asistentes que hicieron posible el intercambio de conocimientos y experiencias en los temas abordados, y la concreción de esta publicación que recoge los principales resultados del encuentro.

Sandra Miguel
La Plata, 2019

La recuperación de información por delimitadores geográficos y su aplicación en estudios bibliométricos de ciencia local

Claudia M. González[1], Gustavo Archuby[2] y Sandra Miguel[3]

Introducción

La vinculación entre lenguaje y espacio geográfico tiene su complejidad. El espacio físico de la tierra poblada tomada en el contexto del cosmos, puede referenciarse de diferentes maneras. A veces se usa la geografía física, otras la geografía política, pero también hay geografía en un sinnúmero de situaciones tales como la mención de áreas sanitarias, distritos de votación, zonas inundables; además de declaraciones explícitas que pueden tomar la forma de topónimos reconocidos, coordenadas geográficas, códigos postales o prefijos telefónicos.

En el discurso científico de tipo referencial -entendiendo como tal a los registros bibliográficos con resumen de las producciones científicas- se encuentran expresiones de lo geográfico muy variadas, que dependiendo del área disciplinar pueden presentar diferente grado de ambigüedad o vaguedad. Sin embargo, el esfuerzo puesto en la identificación de dichas expresiones

[1] Instituto de Investigaciones en Humanidades y Cs Sociales (UNLP/CONICET), La Plata, Argentina. cgonzalez@fahce.unlp.edu.ar

[2] Facultad de Humanidades y Cs de la Educación (UNLP), La Plata, Argentina. gustavoa@fahce.unlp.edu.ar

[3] Instituto de Investigaciones en Humanidades y Cs Sociales (UNLP/CONICET), La Plata, Argentina. smiguel@fahce.unlp.edu.ar

puede verse retribuido si a partir de ellas encontramos nuevas maneras de visualizar los resultados de la actividad científica, permitiendo demostrar el esfuerzo que un colectivo humano de investigación realiza para solucionar los problemas del propio territorio.

Para abordar esta problemática, el trabajar con los topónimos es un primer paso. En esta ponencia presentamos algunos avances a los que hemos arribado en relación a su uso para la selección de la producción bibliográfica pertinente y luego a la manera en que dicha producción puede representarse sobre un mapa.

Metodología

Los topónimos y la selección de la producción

La metodología bibliométrica centra su atención en el análisis del tamaño, crecimiento y distribución de la bibliografía científica en diferentes tipos y niveles de agregación temáticos, institucionales, geográficos, etc. (Okubo, 1997) por una parte, y en el estudio de la estructura social de los grupos que la producen y la utilizan por otra (López Piñero, 1972). Como punto de partida siempre se considera que la producción se encuentra almacenada en alguna fuente de datos que posee cierto nivel de normalización y que cuenta con interfaces de acceso a dicha información ya sea en forma de motor de búsqueda o de servicio de tipo Interfaz de Programación de Aplicaciones (*API*). La elección de las fuentes a utilizar suele ser un asunto controvertido ya que no existe una sola fuente que sea totalmente comprensiva. Las más usadas son *Web of Science* y *Scopus*, a nivel multidisciplinar mundial, *Google Scholar* y actualmente *Dimensions* para los estudios webmétricos. Para estudios relacionados con Latinoamérica, se ha propuesto el uso de *Scielo* y *Redalyc* como las más representativas de la región.

La estructura básica de los registros bibliográficos que ofrecen las fuentes nos brindan cuatro campos de datos con capacidad de contener topónimos: el título, el resumen, las palabras claves y la afiliación de los autores. Mientras la selección de registros basada en la mención de un topónimo (por ejemplo el nombre de un país) en el campo afiliación implica delimitar la producción científica "del país", es decir la que producen sus autores; seleccionar los registros de acuerdo a la misma mención en cualquiera de los otros 3 campos de datos, implica estar seleccionando la producción de investigación

que se realiza “sobre un país”. Luego, la combinación de diferentes formas de estas estrategias nos permite obtener, por ejemplo, la producción del país que menciona al país, la producción del país que no menciona al país, la producción que menciona al país escrita por extranjeros (Miguel, González y Chinchilla-Rodríguez, 2012).

La selección del/los topónimos a utilizar para la recuperación de registros tiene al menos dos aspectos problemáticos. Por un lado se debe decidir hasta que nivel de subagregados se utilizarán. Así, si el nivel de agregación elegido es claramente administrativo (caso país), es fácil de resolver ya que se piensa en términos de los subagregados que lo integran: provincias, estados, etc., solo conlleva decidir hasta que nivel de profundidad se desea llegar. El otro aspecto es que hay que ser cuidadoso al momento de elegir la forma completa o truncada del topónimo garantizando la mayor exhaustividad sin pérdida de precisión, ya que las designaciones ambiguas producen recuperación de registros no deseados que luego deberán revisarse manualmente para su descarte. (Miguel, González e Hidalgo, 2013)

Reconocimiento automático de topónimos

Una vez que se ha realizado la selección de los registros de la fuente de acuerdo a alguno de los criterios presentados en el apartado anterior, tenemos conformado un corpus textual susceptible de ser trabajado con técnicas de Procesamiento del Lenguaje Natural (NPL). Se pueden usar para identificar los temas de investigación, o para el caso que nos interesa aquí que consiste en la extracción de información que hace referencia al espacio o territorio.

Dentro de las técnicas que se pueden utilizar están las denominadas NERC (Named Entity Recognition and Classification), que se emplean para reconocer automáticamente en textos los nombres de entidades, entendiendo como tales los nombres propios de organizaciones, personas, localizaciones, y entidades numéricas tales como fechas, tiempo, montos, expresiones porcentuales. Existen diferentes aproximaciones heurísticas para su implementación. Algunas formas se encuadran dentro de lo que se denomina Aprendizaje Supervisado, tales como los Modelos Ocultos de Markov (HMM), los Árboles de Decisión, Modelos de Máxima Entropía, Máquinas de Vectores de Soporte (SVM) y Campo Aleatorio Condicional (CRF) (Nadeau y Sekine, 2007). Son todas variaciones que consisten en leer un corpus anotado, memorizar las listas

de entidades y crear reglas de desambiguación basadas en las características discriminativas. Un método de referencia que a menudo se propone consiste en etiquetar palabras de un corpus de prueba cuando están anotadas como entidades en el corpus de entrenamiento.

En este trabajo se realizó una prueba de evaluación utilizando una aplicación desarrollada por el Grupo de procesamiento del Lenguaje Natural de la Universidad de Stanford denominada StanfordNE (Stanford NLP, 2018), que es una implementación en Java basada en CRF. La aplicación cuenta con tres modelos que fueron entrenados con una mixtura de los corpus anotados para nombres de entidades denominados CoNLL, MUC-6, MUC-7 y ACE. El modelo Clase 3 reconoce localizaciones, personas y organizaciones; el modelo clase 4 reconoce localizaciones, personas, organizaciones y misceláneas y el modelo Clase 7 reconoce localizaciones, personas, organizaciones, monedas, porcentajes, fechas y tiempos. El objetivo en este trabajo es evaluar el desempeño de los tres modelos sobre un corpus muestral de registros, de tal manera que nos permita decidir cuál de los tres utilizar para la etapa de entrenamiento.

Las medidas utilizadas para la evaluación son la Exhaustividad (*Recall*) y Precisión, medidas clásicas utilizadas para la evaluación de la recuperación de información desde el pionero Proyecto Cranfield (Cleverdon, 1966). También utilizamos la medida conocida como F1, o score F1, que permite dar cuenta del balance que existe entre las otras dos:

Precisión: Verdaderos positivos / Verdaderos Positivos + Falsos positivos
Exhaustividad (*Recall*): Verdaderos positivos / Verdaderos positivos + Falsos negativos
F1: 2 * (Precisión * Exhaustividad / Precisión + Exhaustividad)

De esta manera se puede estimar de manera comparativa el desempeño de cada uno de los modelos en la identificación y anotación automática de los nombres de localidades.

Resolución de topónimos y elaboración de mapas

Una vez que se tienen marcados los topónimos, la función de resolución es mapear esos topónimos del documento con los topónimos contenidos en otro documento de referencia (Gazetteer), que además tiene resueltas las coordenadas geográficas. El sistema de coordenadas es prácticamente universal,

donde un punto $P = \varphi, \lambda$, en el que φ es la latitud, es decir el ángulo entre el punto y el ecuador, medido en grados entre -90 y + 90; y λ es la longitud, el ángulo este oeste del punto 0 por acuerdo internacional que es el meridiano de Grenweech, medido en grados entre -180 + 180. Si bien la latitud y la longitud son siempre relativas a un sistema geodésico de referencia, es decir a un modelo matemático que aproxima a la forma de la tierra (como esfera, elipsoide o geoide), el sistema WG84 (World Geodesic System) es el más usado.

El problema que se presenta al tratar de realizar mapas basados en topónimos y en información bibliométrica -que cuenta la cantidad de publicaciones- es resolver visualmente de manera óptima el anidamiento geográfico y la densidad de publicación por áreas temáticas.

Resultados y discusión

Los topónimos y la selección de la producción

Se muestran los resultados de aplicar diferentes estrategias de búsqueda para el caso de la producción Argentina que contiene como topónimo el nombre del país, de alguna de sus provincias o de una región (Bonaerense, Cuyo, Patagónica, Pampeana) en el título, resumen o palabras claves para el periodo 2007-2016.

Tabla 1. Resultados de diferentes estrategias de búsqueda en Scopus

1	argentin*	27938	91,31 %
	[Provs] NOT argentin*	1549	5,06 %
	[Region] NOT argentin*	1222	3,99 %
		30709	100 %
2	argentin* AND [Provs]	13026	42,42 %
	argentin* AND [Region]	3710	12,08 %
3	argentin* OR [Provs]	29548	96,22 %
4	argentin* OR [Provs] OR [Region]	30598	99,64 %

Cuando se piensan las estrategias de búsqueda lo importante es poner en consideración que los nombres de los topónimos sean únicos de la región. Por ejemplo, las provincias como Córdoba o Santa Cruz pueden recuperar registros que tienen que ver con otras regiones del mundo, España y Bolivia, en este caso. Cuando se tiene conocimiento de casos tan evidentes, será conveniente ajustar la expresión excluyendo dichos países.

Reconocimiento automático de topónimos

Luego de aplicar la estrategia de búsqueda 4 de la tabla 1, la muestra de registros que se utilizó para la evaluación de los 3 modelos de StanfordNER sin entrenamiento fueron 372 casos de un total de 12379 correspondientes al área de Agricultura y Ciencias Biológicas.

Después de ejecutar los 3 algoritmos se revisaron manualmente los registros de la muestra para realizar la anotación y el conteo de los casos positivos y negativos.

Tabla2. Resultado de las medidas de evaluación para los 3 modelos

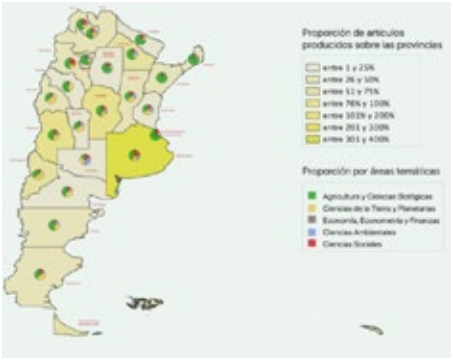
	VP	FP	FN	PRECISION	RECALL	F1
Mod 4	843	808	232	0,5106	0,7842	0,6185
Mod 3	766	257	326	0,7488	0,7015	0,7243
Mod 7	733	63	394	0,9209	0,6504	0,7624

Como se puede observar, el mejor desempeño se obtuvo con el Modelo 7 (F1=0,7624). Recordemos que cuanto más precisos se vuelven los sistemas, por lo general pierden exhaustividad y a la inversa. En este caso, se observa que los modelos 4 y 7 tienen valores menos armónicos que el modelo 3. Mientras el Modelo 4 es más exhaustivo y menos preciso, el Modelo 7 es más preciso y menos exhaustivo.

Resolución de topónimos

Se muestra aquí un mapa que muestra la producción científica obtenida con una estrategia similar a la de la búsqueda 4 (ver Tabla 1) para el periodo 2008-2012.

Figura 1. Producción asociada a temas de investigación locales (Argentina 2008-2012)



En el mismo se puede observar el resultado de asignar a la capa de las provincias Argentinas los atributos de datos totales de producción y los mismos discriminados por áreas temáticas. Los primeros se manifiestan en la diferente intensidad de colores que toma cada provincia en el mapa, los segundos en los gráficos circulares. Así, tenemos provincias con gráficos con preponderancia del área Agricultura y Ciencias Biológicas,como es el caso de Misiones y Corrientes (Esteros del Iberá), o un gráfico con preponderancia de las Cs de la Tierra y Planetarias como es el caso de la zona petrolera de Neuquén.

Conclusiones

Si bien la frase "ciencia local" puede presentar interpretaciones ambiguas, en este trabajo se le ha dado la connotación de ser la ciencia que produce un país sobre cuestionesque de alguna manera involucran a su propio territorio. Si bien los estudios de producción científica realizados con este recorte no cubren la totalidad de la ciencia de interés local, se considera que es una aproximación a su delimitación.

Trabajar con la representación espacial que aparece mencionada en el discurso científico se presenta como una oportunidad poco explorada en los estudios Bibliométricos. La capacidad de aportar información valiosa implica encontrar técnicas automáticas o semi-automáticas de tratamiento del lenguaje natural que permitan explotar corpus referenciales grandes.

De un tiempo a esta parte, la cienciometría ha puesto empeño en encontrar formas de visualización con mayor potencia explicativa. Los Sistemas de Información Geográfica pueden ser un aporte para reflejar distintos aspectos de la relación ciencia/territorio. En el caso particular de los temas de investigación, un primer paso para realizar su georeferenciamiento proviene de la identificación de los topónimos mencionados en el texto.

Referencias bibliográficas

Cleverdon, C.W., Mills, J.Y Keen, E.M. (1966). *ASLIB Crandfield proyect: Factors Determining the performance indexing Systems*. Cranfield: ASLIB.

López Piñero, J. M. (1972). *El análisis estadístico y sociométrico de la literatura científica*. Valencia: Facultad de Medicina. Centro de Docum. e Informática Médica.

Miguel, S, González, C.M. y Chinchilla-Rodríguez, Z. (2012). Lo local y lo global en la producción científica argentina con visibilidad en Scopus, 2008-2012. *Información, Cultura y Sociedad*, *32*, 59-78. Recuperado de http://revistascientificas.filo.uba.ar/index.php/ICS/article/view/1375/1352

Miguel, S, González, C.M. e Hidalgo, M. (2013). Argentina como objeto de investigación. Reflexiones conceptuales y aproximaciones metodológicas para el abordaje bibliométrico de la producción científica sobre temas de alcance nacional. Trabajo presentado en Actas de las 3ras Jornadas de Intercambio y Reflexiones de la Investigación en Bibliotecología. La Plata: UNLP. Recuperado de http://jornadabibliotecologia.fahce.unlp.edu.ar/jornadas-2013/actas-2013/miguel_gonzalez_hidalgo.pdf/view

Nadeau, D. y Sekine, S. (2007). A survey of named entity recognition and classification. *Linguistic e Investigationes*, *30*(1), 3-26.

Okubo, Y. (1997). *Bibliometric Indicators and Analysis of Research Systems: Methods and Examples.* (Report No. STI Working Papers 1997/1). Paris: OECD. https://doi.org/10.1787/18151965

Stanford N.L.P. (2018). Recuperado de https://nlp.stanford.edu/software/CRF-NER.shtml.

En un contexto de creciente incorporación del acontecer local y nacional en contextos globales, el impulso de políticas que orientan la investigación hacia temas locales y a la resolución de problemas sociales, productivos y medioambientales, plantea nuevos desafíos en la valoración de los resultados e impacto que esa investigación produce, así como también de la transferencia e innovación. Este libro de actas reúne las ponencias presentadas en el Workshop Iberoamericano de estudios métricos de la actividad científica y tecnológica en temas locales/regionales, con aportes de autores de diferentes perfiles formativos y trayectorias, en un intento por contribuir a los debates teóricos y metodológicos para la obtención de métricas y visualizaciones derivadas de la producción científica de países iberoamericanos y la relación con indicadores del contexto económico y social. Los contenidos están organizados en tres ejes temáticos. El primero se enfoca en cuestiones relativas a las políticas y agendas de investigación; el segundo presenta diferentes aproximaciones metodológicas para el abordaje cienciométrico de la producción científica en temas locales/regionales, y el tercero recoge estudios de caso de un grupo de países de la región.

Trabajos, Comunicaciones y Conferencias, 37

ISBN 978-950-34-1742-3