Organización automática del conocimiento: la geografía en la Wikipedia

Automatic knowledge organization: Geography in Wikipedia

Carlos G. FIGUEROLA, Angel ZAZO RODRÍGUEZ, José Luis ALONSO BERROCAL

Instituto de Estudios en Ciencia y Tecnología, Universidad de Salamanca. c/ Espejo s/n, 37002 Salamanca, España.
figue@usal.es. zazo@usal.es. berrocal@usal.es

Resumen

Las tecnologías de la información propician un crecimiento sin precedentes de la información, lo cual plantea el problema de la organización de ésta. Al tratarse de información digital es posible abordar su organización mediante procedimientos automatizados. De otro lado, las Técnicas de Análisis de Redes son un poderoso instrumento que permite modelar diferentes fenómenos y aplicar después técnicas automáticas. En este trabajo se describe la aplicación de estas Técnicas de Análisis de Redes para modelar y procesar una masa importante de documentos, como la constituida por los artículos de la Wikipedia. La aplicación posterior de algoritmos de detección de comunidades permite agrupar los artículos en función de sus hiperenlaces y su afinidad temática. Este trabajo se centra, después de haber aplicado estas técnicas, en la relación geográfica de los artículos, en sus comunidades de red y las conexiones entre ellas.

Palabras clave: Organización automática del conocimiento. Wikipedia. Geografía. Análisis de redes.

Abstract

The Information Technologies drive an unprecedented growth of information, which raises the problem of the organization of it. As it is digital information, it is possible to approach her organization through automated procedures. On the other hand, Network Analysis Techniques are a powerful tool that allows us to model different phenomena and then apply automatic techniques. In this paper we describe the application of these Network Analysis Techniques to model and process an important number of documents, such as the one constituted by Wikipedia articles. The subsequent application of community detection algorithms allows grouping the articles based on their hyperlinks and their thematic affinity. This work focuses, after having applied these techniques, on the geographical relationship of the articles, on their network communities and the connections between them.

Keywords: Automatic knowledge organization. Wikipedia. Geography. Social networks analysis techniques.

1. Introducción

El desarrollo de las Tecnologías de la Información y la generalización de su uso han propiciado la generación y publicación a través de Internet de una cantidad ingente de información. Este hecho pone de relieve la necesidad de aplicar formas de organización de dicha información. Sin embargo, el uso de procedimientos manuales presenta importantes dificultades, siendo uno de ellos, aunque no el único, el de la mera incapacidad de procesar una cantidad tan grande de información.

De otro lado, una característica importante de toda esa información propiciada por el desarrollo de las Tecnologías de la Información es que es digital, tanto por su origen o producción como por su circulación y consumo. Esto facilita la aplicación de métodos automáticos; estos métodos automáticos son motivo de investigación desde diversos enfoques y en diversos estados de desarrollo. Algunas propuestas novedosas intentan aplicar las Técnicas de Análisis de Redes y la detección de comunidades a esta tarea.

1.1. La Wikipedia

La Wikipedia es un fenómeno bien conocido; desde su nacimiento en 2001 (O'Sullivan, 2016) no ha dejado de crecer, y actualmente es utilizada diariamente por millones de usuarios; y, lo que es tanto o más interesante, millones de personas se han implicado de un modo u otro en su elaboración y puesta al día. Sus artículos, en más de 300 lenguas, alcanzan la cifra de 50 millones.

En efecto, la Wikipedia se publica en diferentes lenguas; normalmente, el origen de los autores/editores de la versión en una lengua específica coincide con los países autóctonos de esa lengua, con algunos matices. Por ejemplo, la Wikipedia en español, según datos obtenidos en 2015, tenía aproximadamente 1,1 millones de artículos y casi 6 millones de editores; de ellos, una parte importante de España, pero también de Ar-

gentina, México, Chile, Colombia, etc. (Zazo Rodríguez et al., 2015) (Figura 1). La Wikipedia en portugués, por su parte, tenía en esa misma fecha alrededor de 800.000 artículos y 2.735.806 editores, prácticamente todos de Brasil y Portugal.

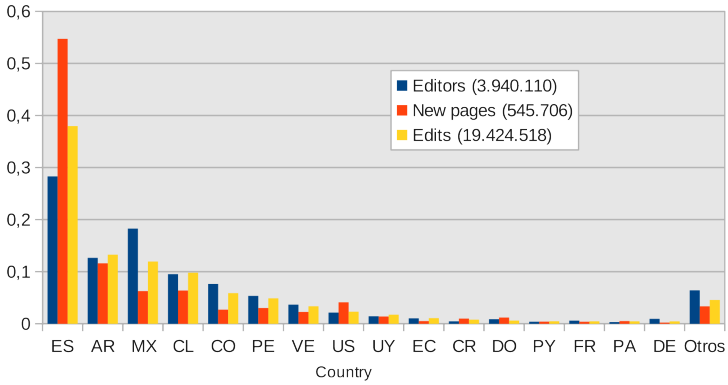

Figura 1. Procedencia de los editores de Wikipedia en español (2015) (Zazo Rodríguez et al., 2015)

Sin embargo, la Wikipedia en inglés parece ser un caso especial. Con sus más de 5 millones de artículos en 2018, una buena parte de sus editores son de países anglófonos (Zazchte, 2019). Pero también una parte considerable de ellos son de otros países no anglófonos de manera que constituye una suerte de Wikipedia internacional (Figura 2).

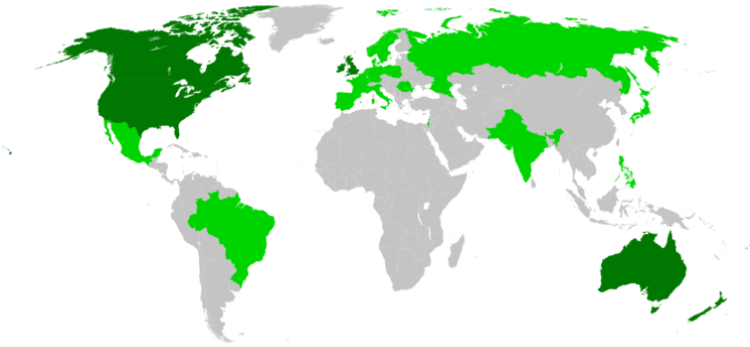

Figura 2. Procedencia geográfica de editores de la Wikipedia en inglés (las zonas más oscuras representan mayor número de editores) (Fuente: Wikipedia)

La calidad y fiabilidad de los contenidos de la Wikipedia ha sido en el pasado un asunto polémico. Muy discutida a causa del carácter anónimo de los editores de los artículos y de su carácter abierto, ha sido sometida a numerosos y abundantes escrutinios. Este asunto ha generado una cantidad importante de literatura científica (Okoli et al., 2014). El resultado de análisis tan exhaustivos resulta claramente favorable para la fiabilidad de los contenidos de la Wikipedia (Okoli et al., 2012). Algunos de los elementos que ofrecían más dudas, como su carácter abierto y la posibilidad de que cualquiera, anónimamente, pudiese alterar o introducir contenidos no fiables, o incluso abiertamente vandálicos (Shachaf y Hara, 2010) ha resultado ser su mejor defensa. Es precisamente ese carácter abierto y masivo el que hace que cualquier posible error sea detectado y corregido con rapidez.

Sea como fuere, la Wikipedia es producida y utilizada de forma masivamente en buena parte del planeta y, de alguna manera, puede considerarse un buen exponente de los conocimientos, pero también de las creencias, concepciones e incluso prejuicios de la sociedad que la produce y consume.

De otro lado, es llamativo el hecho de que toda esta masa formidable de conocimiento carece prácticamente de estructura interna. Básicamente se trata de fragmentos heterogéneos de conocimiento (los artículos) conectados exclusivamente a través de links o hipervínculos. Es cierto que esos artículos se ajustan en mayor o menor medida a alguna de las numerosas plantillas puestas a disposición de los autores, pero se trata de una estructura poco explícita para el usuario de la información (Palmero Aprosio et al., 2013). Y es cierto también que existe una estructura de categorías de artículos, pero éstas funcionan más bien como una colección interminable de descriptores libres de utilidad dudosa (Chernov et al., 2006; Weale et al., 2006).

Por este motivo, la Wikipedia es un terreno que permite probar la eficacia de técnicas de organización automática del conocimiento y experimentar con ellas, constituyendo un buen banco de pruebas para probar la eficacia de esas técnicas automáticas.

1.2. Técnicas de Análisis de Redes

El Análisis de Redes Sociales se basa en la teoría matemática de Redes o Grafos; a mediados de los años 50 del pasado siglo, algunos sociólogos que aplicaban elementos de esta teoría para modelar las relaciones entre personas y grupos sociales (Scott, 2013) acuñaron el término de Análisis de Redes Sociales.

Brevemente, una red es un conjunto de nodos o vértices conectados por arcos o enlaces. Los nodos poseen un conjunto de características o atributos arbitrarios, definidos por quien aplica este artefacto. Los enlaces o arcos conectan dos nodos entre sí; los enlaces pueden ser dirigidos (parten de un nodo y apuntan a otro) o no dirigidos (conectan dos nodos en una relación bidireccio-

nal); algunas redes tienen también arcos reflexivos (parten y llegan al mismo nodo). Los arcos o enlaces también pueden tener atributos arbitrarios definidos por el usuario; uno de los más habituales es el peso: un valor numérico que intenta expresar la intensidad de la relación que representa ese arco. Se han desarrollado métodos y procedimientos para analizar la estructura interna de una red. De esta forma, una vez modelado un determinado fenómeno mediante una red, es posible aplicar estas técnicas de análisis para estudiar la estructura interna de ese fenómeno.

Uno de los elementos del análisis de redes es el conocido como detección de comunidades de nodos. Es habitual que en una red encontremos grupos de nodos fuertemente interconectados entre sí, al tiempo que sus conexiones con nodos ajenos a ese grupo son escasas y débiles. Estos grupos son lo que denominamos comunidades de red y tienen numerosas aplicaciones (Plantie et al., 2013).

2. Metodología

2.1. Wikipedia como red de documentos

Los artículos de Wikipedia contienen, como es sabido, hipervínculos a otros artículos; también a páginas externas o a secciones administrativas de la propia Wikipedia, aunque estos últimos casos no serán tenidos en cuenta en este trabajo. Si se representan los artículos como nodos de una red, resulta obvio que esos hipervínculos pueden verse como arcos dirigidos que conectan los nodos, lo cual permite representar Wikipedia como una red. Este enfoque no es nuevo y ha sido utilizado en numerosos trabajos; de hecho, permite aplicar diversos elementos del análisis de redes (Zlatic et al., 2006; Brandes et al., 2009).

Sin embargo, es difícil asignar pesos a este tipo de enlaces. Una alternativa, que es la adoptada en este trabajo, es calcular la similitud semántica entre dos artículos enlazados por uno de estos hipervínculos. Por similitud semántica entendemos, en este caso, la similitud entre vectores que representan cada documento, siguiendo el bien conocido modelo del espacio vectorial (Salton, 1983). Para el cálculo de la similitud se ha seguido un esquema tf x idf clásico, aplicando la fórmula clásica del coseno. Así, el peso del arco que conecta dos documentos viene a ser, en este trabajo, la similitud entre los vectores de cada uno de los artículos (Dalton et al., 2012; Gavrilovitch et al. 2007). Una de las ventajas de este enfoque mixto (similitud vectorial aplicada a pares de artículos conectados) es que limita ese cálculo solamente a los documentos conectados, reduciendo el tiempo y capacidad de proceso necesarios.

Así pues, se construyó una red con los 5,5 millones de artículos de la Wikipedia en inglés, sus hipervínculos y las similitudes semánticas entre los artículos conectados. El número original de arcos enlaces es cercano a los 415 millones. Pero, tras diversas pruebas, se aplicó un umbral de 0,2 en la similitud entre artículos lo que redujo la cantidad de arcos a considerar en poco más de 145 millones de arcos; esto produce una red más fácil de procesar. Una representación gráfica de esa red puede verse en la Figura 3. En ella cada punto es un nodo o artículo, habiéndose omitido la representación gráfica de los arcos para mejorar la visualización; aunque, obviamente, los arcos y sus pesos se han tenido en cuenta en la colocación espacial de los nodos.

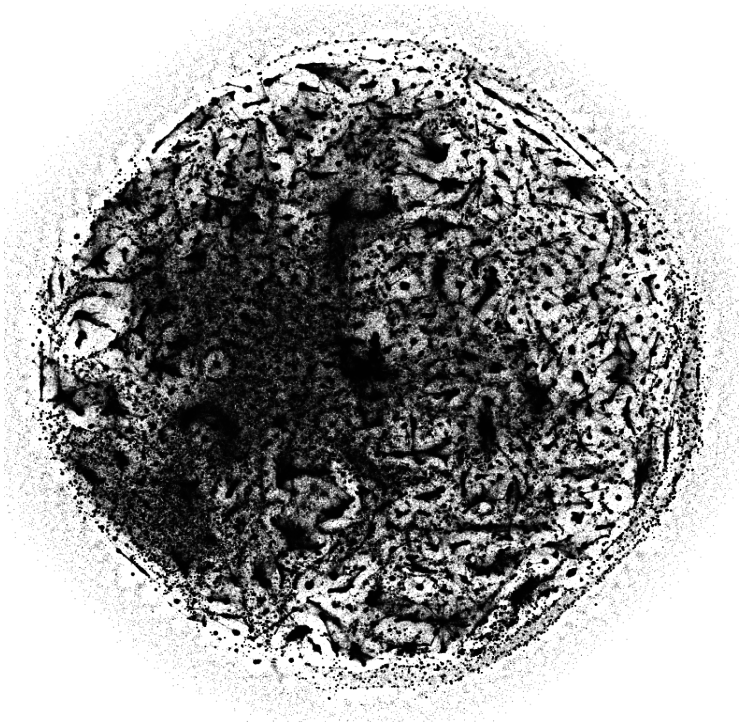

Figura 3. Red de artículos de Wikipedia

2.2. Comunidades de artículos en Wikipedia

En el caso de la Wikipedia, dado que la conexión entre nodos o artículos se basa en la existencia de hipervínculo y en una cierta similitud semántica, es posible establecer la hipótesis de que comunidades de artículos constituyen bloques temáticos de artículos afines en contenidos. De confirmarse esta hipótesis, la identificación de estos bloques o comunidades permitiría la organización temática de los artículos de una manera automatizada. Obviamente, esto tiene interés cuando se opera con grandes cantidades de documentos; por ejemplo, los 5,5 millones de artículos de la Wikipedia en inglés.

Sin embargo, en este mismo aspecto, el tamaño de la colección documental, reside uno de los principales problemas. Existen diversas técnicas de detección de comunidades de red; pero, en general, ésta es una operación que consume grandes cantidades de tiempo y recursos de ordenador. En el trabajo de Lancichinetti et al. (2009) pueden encontrarse varias de tales técnicas y una comparación entre ellas. De hecho, varias de ellas directamente no son utilizables con redes grandes (Lee et al., 2014).

Uno de los algoritmos más recientes, conocido como Infomap (Blondel et al., 2008; Rosvall et al., 2009; Edler et al., 2015), ha sido diseñado expresamente para trabajar con redes de gran tamaño. La literatura científica documenta numerosas aplicaciones prácticas y sus resultados (Bohlin et al., 2014).

3. Resultados

Tras aplicar Infomap, se ha obtenido un total de 1.225 comunidades de primer nivel; ésta parece, inicialmente, una cantidad demasiado grande de comunidades para ser analizada manualmente. Sin embargo, solamente 370 comunidades contienen más de 100 documentos; el resto son comunidades pequeñas, fragmentarias, buena parte de ellas con un solo artículo. Lo interesante es que esas 370 comunidades contienen el 99 % de todos los artículos de la Wikipedia. Hay que hacer notar que Infomap adscribe cada nodo a una única comunidad.

Comunidad	*Artículo*
86:1	Beetle
86:1	Longhorn_beetle
86:1	Polyphaga
86:1	Cucujoidea
86:1	Leaf_beetle
86:1	Tenebrionoidea
86:1	Cucujiformia
86:1	Dermestidae
86:1	Byrrhoidea
86:1	Scarabaeoidea
86:1	Bostrichoidea
86:1	Elateroidea
86:1	Chrysomeloidea

Tabla I. Comunidad de insectos

Comunidad	*Artículo*
2:3	Rely_Pas-de-Calais
2:3	La_Couture,_Pas-de-Calais
2:3	Campagne-lès-Guines
2:3	Communes_of_the_Pas-de-Calais
2:3	Saint-Floris
2:3	Lignereuil
2:3	Nielles-lès-Calais
2:3	Le_Wast
2:3	Louches
2:3	Flers,_Pas-de-Calais

Tabla II. Comunidad de lugares de Francia

Com.	*Artículo*
21:1	Solar_eclipse
21:1	Solar_eclipse_of_March_7,_1970
21:1	List_of_solar_eclipses_visible_from_the_United_Kingdom
21:1	List_of_solar_eclipses_visible_from_the_Philippines
21:1	Solar_eclipse_of_July_1,_2011
21:1	List_of_solar_eclipses_visible_from_Australia
21:1	List_of_solar_eclipses_visible_from_Ukraine
21:1	Solar_eclipse_of_December_24,_1992
21:1	Solar_eclipse_of_December_23,_1908
21:1	Solar_eclipse_of_January_25,_1982

Tabla III. Comunidad de eclipses

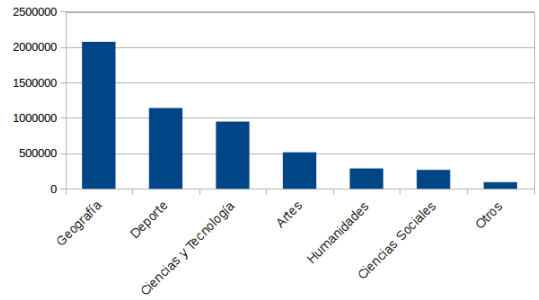

Figura 4. Artículos y adscripción a bloques temáticos

Las Tablas I, II y III presentan, a modo de ejemplo, fragmentos de varias comunidades, mostrando los títulos de los artículos que las componen. La coherencia temática es evidente; y es posible efectuar manualmente un etiquetado rápido en grandes bloques temáticos.

La Figura 4 presenta esos bloques temáticos y la cantidad de artículos en cada uno de ellos. El más numeroso, que se ha denominado como de artículos geográficos, merece una explicación. Este bloque está compuesto por los artículos típicamente enciclopédicos dedicados a países, regiones, ciudades, etc.; pero también a artículos dedicados a otros temas, como historia, arte, literatura. Lo interesante es que los artículos de esta temática tienen, en las comunidades obtenidas, un hilo conductor que los aglutina y que los hace aparecer en la misma comunidad: el país. Así, es posible encontrar artículos puramente geográficos sobre lugares de Francia, por ejemplo; pero junto con ellos, en la misma comunidad o en comunidades muy próximas, aparecen artículos sobre pintores franceses, o sobre militares franceses, o sobre historia de Francia, etc.

La Figura 5 muestra la representación gráfica de la Wikipedia con los bloques temáticos en colores diferentes.

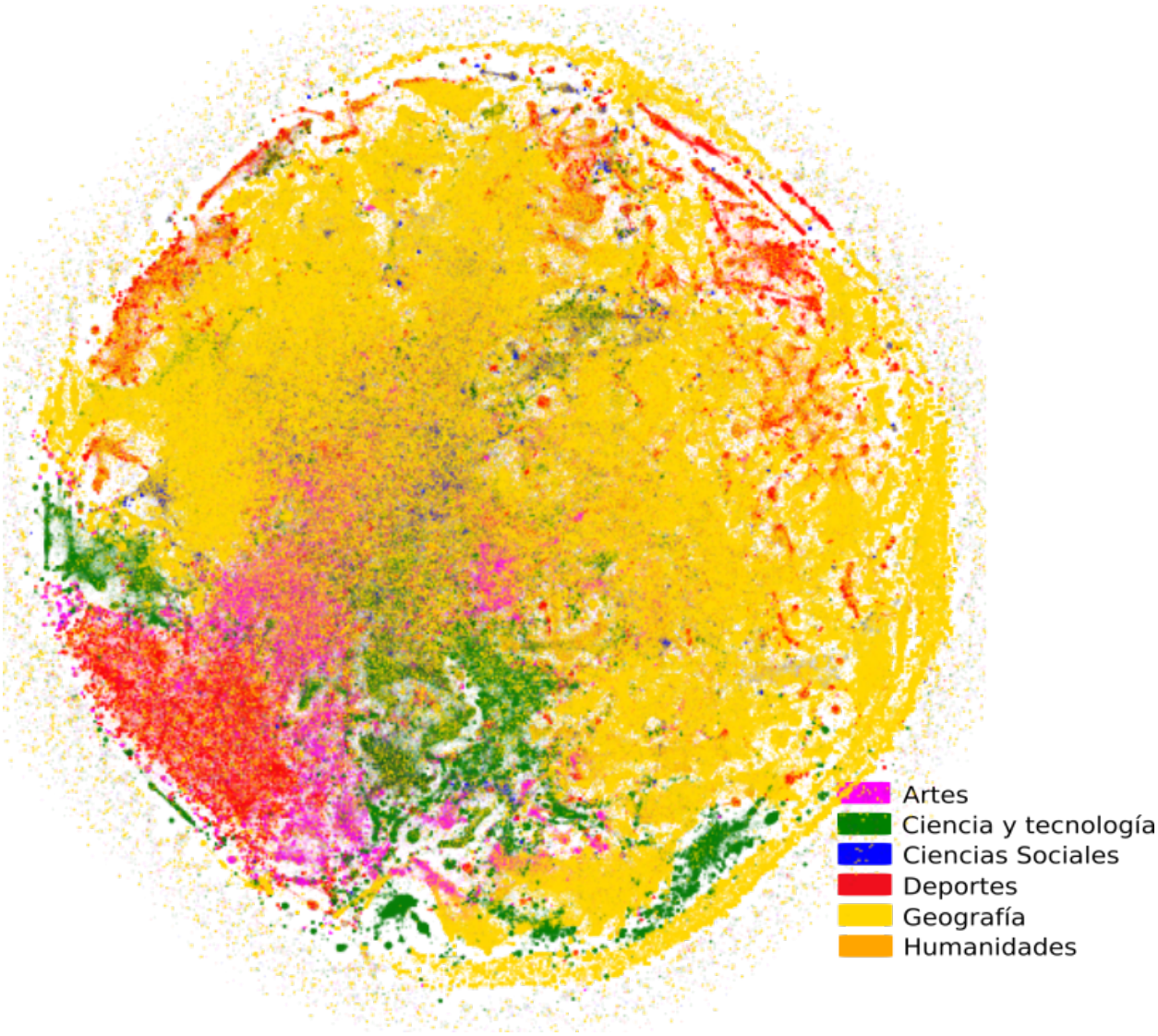

Figura 5. Artículos y bloques temáticos

Observando sólo este bloque de tipo geográfico, en sentido amplio, es posible considerarlo como una subred, constituida solamente por los nodos de este tipo y exclusivamente por los arcos o conexiones entre ellos. Si a esta subred se la somete al mismo proceso de detección de comunidades, se obtiene un total de 18 comunidades de artículos geográficos, fácilmente etiquetables de manera manual.

El resultado, los bloques de artículos geográficos obtenidos puede apreciarse en la Figura 6; y, nuevamente, en la representación gráfica de la Wikipedia, con cada bloque marcado en un color. Aquí se pueden apreciar con facilidad esos grandes

bloques geopolíticos y su distribución espacial en la red. La distribución espacial es importante porque refleja no solamente el tamaño de cada bloque, sino también la relación de proximidad o lejanía entre ellos. Proximidad o lejanía no se refieren aquí a la distancia física, sino a cómo los centenares de miles de editores de Wikipedia reflejan su visión de la distancia cultural, en sentido muy amplio, de los diferentes bloques geopolíticos.

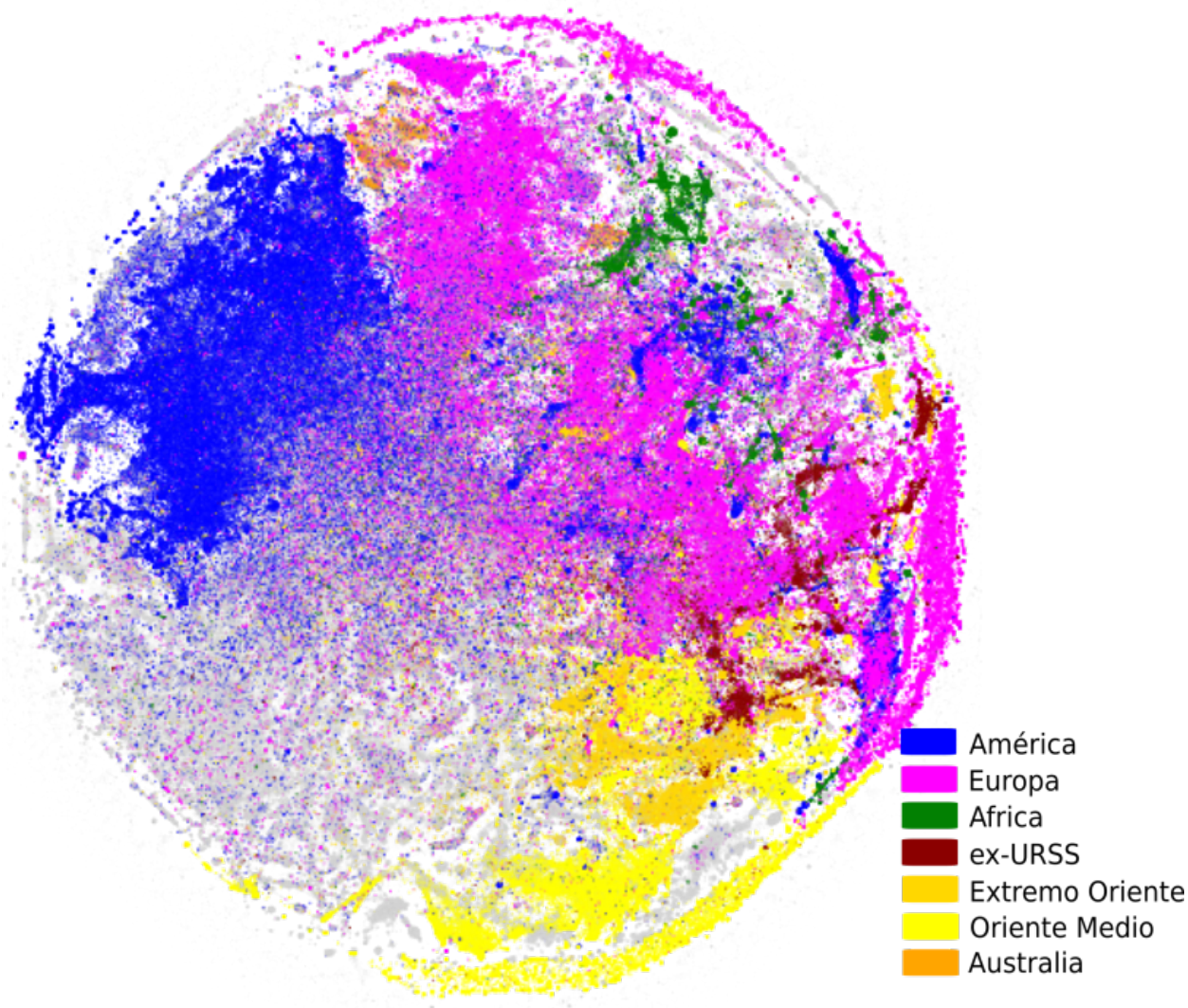

Figura 6. Artículos geográficos y adscripción a continentes

Como es esperable, los artículos más numerosos son los referidos a Europa y América. Los referidos a África, aparte de ser minoría, se encuentran próximos a sub-bloques europeos, pero también a algunos americanos claramente desgajados del núcleo central americano. Los artículos referentes a Medio y Extremo Oriente se muestran periféricos y, salvo excepciones, relativamente compactos entre sí. Australia parece mucho más relacionada con los bloques angloparlantes de América y Europa, físicamente mucho más distantes. Y las repúblicas de lo que en su día fuera la URSS aparecen, algunas, vinculadas a Europa, pero otras más próximas a Oriente.

En cuanto a Europa, claramente muestra varios focos; uno de ellos, especialmente abundante y compacto, es el de los artículos referidos a zonas anglosajonas. Aparecen muy próximos a América del Norte y Australia. Otros bloques, como el de España, parece especialmente relacionado con países de Latinoamérica, como era de esperar. Sorprende, tal vez, la posición periférica de parte de los artículos acerca de Francia (Figura 7).

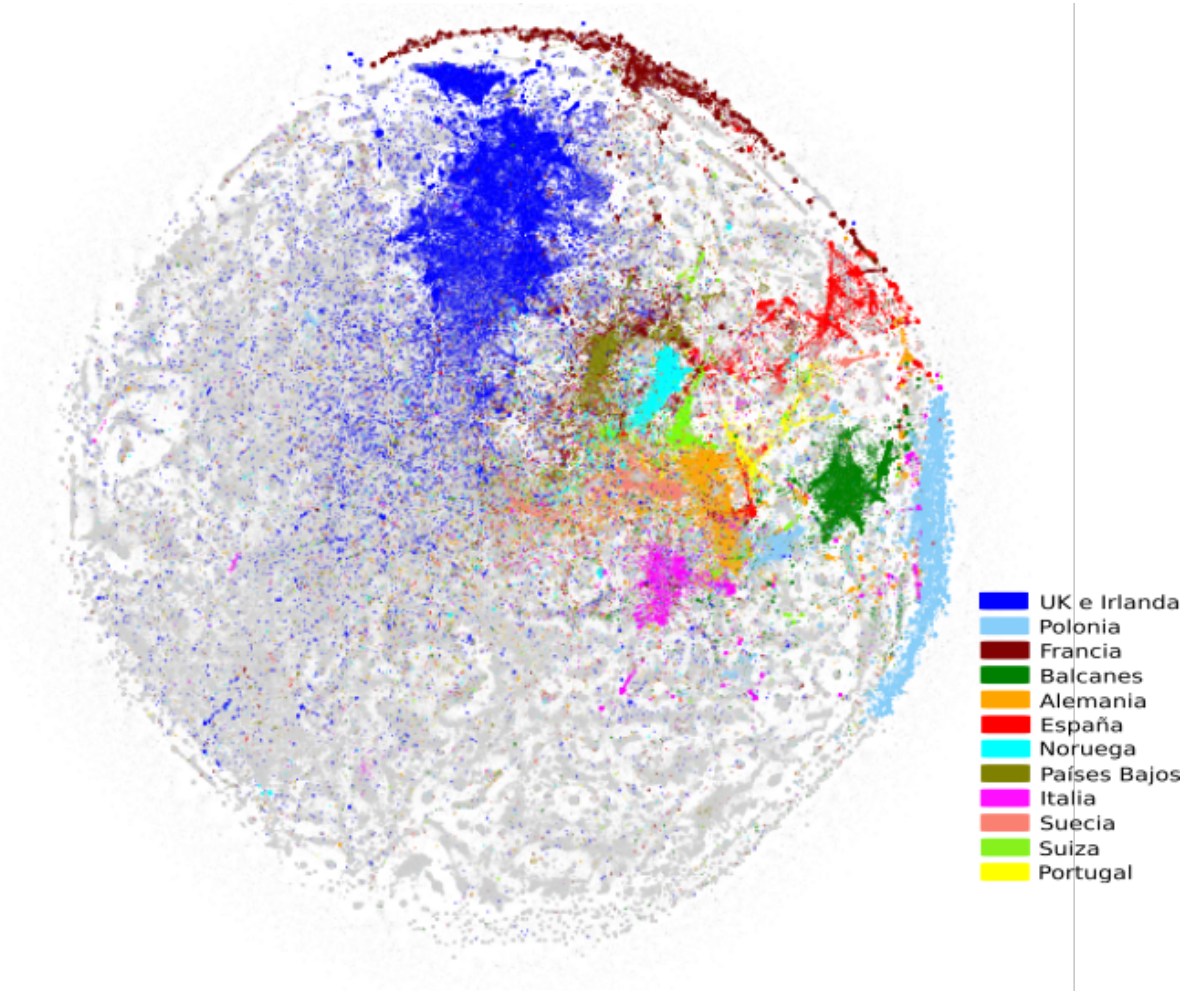

Figura 7. Artículos sobre Europa

La Figura 8, en la página siguiente, muestra los artículos americanos y sus diferentes países. El más abundante es Estados Unidos, y muy cerca Canadá; el resto aparecen muy lejanos. Dentro de esta lejanía, México, Argentina y Cuba están algo más próximas. Brasil por su parte, queda totalmente en la periferia, al igual que sucede con Perú y Bolivia.

4. Conclusiones

Las técnicas de análisis de redes pueden aplicarse a colecciones documentales, complementando los hipervínculos entre documentos, insertados por los propios autores de los documentos, con la similitud semántica entre los documentos enlazados, calculándose ésta con cualquiera de los métodos utilizados habitualmente en la Recuperación de Información.

A modo de experimento, se ha utilizado la Wikipedia en inglés como colección documental; una característica importante es su considerable tamaño, tanto en documentos (5,5 millones) como en hiperenlaces (145 millones). El objetivo era comprobar la capacidad de organización automática de una masa documental a través de las técnicas de análisis de redes, en función de sus contenidos.

La aplicación de algoritmos de detección de comunidades de red ha permitido detectar grandes bloques temáticos; una característica importante de estos algoritmos es la adecuación a redes de gran tamaño.

Pero la aplicación de forma recursiva de tales métodos de detección de comunidades a los bloques temáticos recurridos ha permitido obtener bloques más específicos y precisos. A través de la representación de la colección documental como una red de nodos y enlaces, además de organizar los documentos por bloques temáticos, se ha podido observar las relaciones de proximidad o lejanía entre tales bloques, añadiendo así más información a la estructura obtenida.

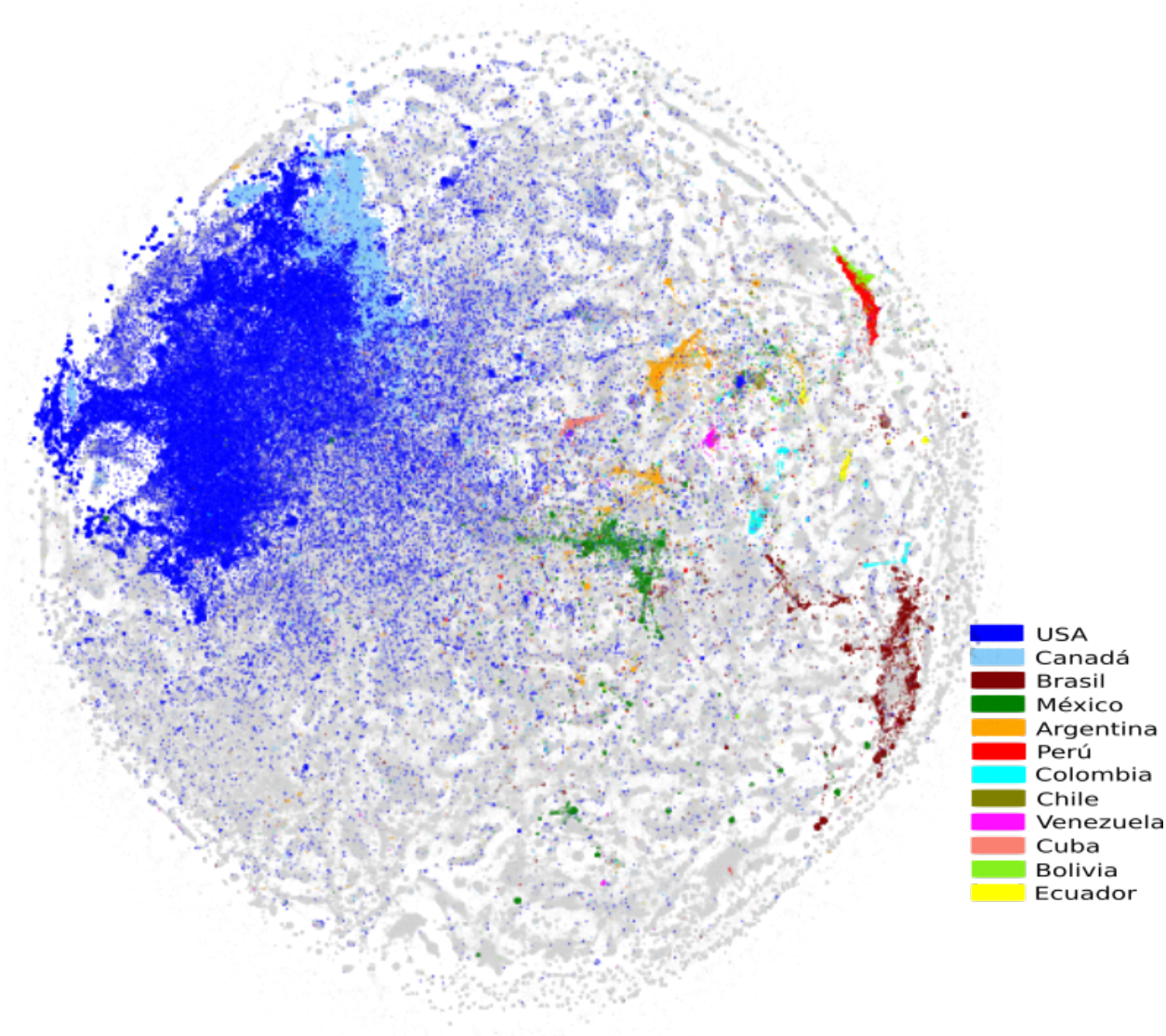

Figura 8. Artículos sobre América

Agradecimientos

La investigación objeto de esta comunicación se está financiando con fondos de FEDER de la Unión Europea a través del Programa Nacional del Plan de Investigación Científica, Desarrollo e Innovación Tecnológica (I+D+i) del Ministerio de Economía y Competitividad (CSO2013-49278-EXP) y del Programa Estatal de Generación de Conocimiento y Fortalecimiento Científico y Tecnológico del Sistema de I+D+i (PGC2018-093755-B-I00).

Referencias

Chernov, S.; Iofciu, T.; Nejdl, W.; Zhou, X. (2006). Extracting Semantics Relationships between Wikipedia Categories. // SemWiki'06 Buvda, Montenegro. DOI=10.1.1.73.5507

Blondel, V. D.; Guillaume, J. L.; Lambiotte, R.; Lefebvre, E. (2008). Fast unfolding of communities in large networks. // Journal of Statistical Mechanics: Theory and Experiment. 10:2008. http://arxiv.org/pdf/0803.0476.pdf (23/03/2019).

Bohlin, L.; Edler, D.; Lancichinetti, A.; Rosvall, M. (2014). Community detection and visualization of networks with the map equation framework. // Measuring Scholarly Impact. 3-34.

Brandes, U.; Kenis, P.; Lemer, J.; van Raaij, D. (2009). Network Analysis of collaboration structure in Wikipedia. // Proceedings of the 18th international conference on World Wide Web. New York.731-740. Doi: 10.1145/15226709.1526808

Dalton, J.; Dietz, L. (2012), Bi-directional Linkability From Wikipedia to Documents and Back Again: UMass at TREC 2012. // Text Retrieval Conference 2012. Knowledge Base Acceleration Track. http://trec.nist.gov/pubs/trec21/papers/umass_CIRR.kba.final.pdf (23/03/2019).

Edler, D.; Rosvall, M. (2015). The infomap software package.http://www.mapequation.org/code.html (23/03/2019).

Gabrilovich, E.; Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. // Proceedings of the 20th international joint conference on Artifical intelligence. Hyderabad, India: AAAI Press. 1606-1611.

Lancichinetti, A.; Fortunato, S. (2009). Community detection algorithms: A comparative analysis.// Physical Review E. 80:5. http://arxiv.org/pdf/0908.1062v2.pdf (23/03/2019).

Lee, C.; Cunningham, P. (2014). Community detection: Effective on large social networks. // Journal of Complex Networks. 2:1 19–37. http://comnet.oxfordjournals.org/content/2/1/19.full.pdf+html (23/03/2019).

Okoli, C; Mehdi, M.; Nielsen, F.Å.; Lanamäki, A. (2012). The People's Encyclopedia Under the Gaze of the Sages: A Systematic Review of Scholarly Research on Wikipedia. https://ssrn.com/abstract=2021326, http://dx.doi.org/10.2139/ssrn.2021326 (23/03/2019).

Okoli, C.; Mehdi, M.; Mesgari, M.; Nielsen, F. Å.; Lanamäki, A. (2014). Wikipedia in the eyes of its beholders: A systematic review of scholarly research on Wikipedia readers and readership. // Journal of the Association for Information Science and Technology. 65:12, 2381-2403.

O'Sullivan, D. (2016). Wikipedia: a new community of practice?. London: Routledge.https://doi.org/10.4324/978

1315547183 (23/03/2019).

Palmero Aprosio, A.; Giuliano, C.; Lavelli, A.(2013) Automatic Mapping of Wikipedia Templates for Fast Deployment of Localised DBpedia Datasets. // Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies (i-Know '13). New York:ACM. DOI: https://doi.org/10.1145/2494188.2494196 (23/03/2019).

Plantié, M.; Crampes, M. (2013). Survey on social community detection. // Social media retrieval, 65–85. http://hal.archivesouvertes.fr/docs/00/80/42/34/PDF/Survey-on-Social-Community-Detection-V2.pdf (23/03/2019)

Rosvall, M.; Axelsson, D.; Bergstrom, C. (2009). The map equation. // European Physical Journal Special Topics. 178 13–23.

Salton, G.; McGill, M.J. (1983). Introduction to Modern Information Retrieval. New York, NY: McGraw-Hill.

Scott, J. (2013). Social network analysis. Thousand Oaks, CA, US: Sage Publications, Inc

Shachaf, P.; Hara, N. (2010). Beyond vandalism: Wikipedia trolls. // Journal of Information Science. 36:3. 357-370.

Weale, T. (2006). Utilizing Wikipedia categories for document classification. ftp://ftp.cse.ohio-state.edu/pub/tech-report/2008/TR14.pdf (23/03/2019).

Zachte, Erik (2019). Wikimedia Traffic Analysis Report – Page Edits per Wikipedia Language – Breakdown https://stats.wikimedia.org/wikimedia/squids/SquidReportPageViewsPerLanguageBreakdown.htm (23/03/2019).

Zazo, A. F.; Figuerola, C. G.; Alonso Berrocal, J. L. (2015). Edición de contenidos en un entorno colaborativo: el caso de la Wikipedia en español. // Scire: representación y organización del conocimiento, 21:2 57-67.

Zlatic, V.; Bocicevic, M.; Tefancic, H.; Domazet, M. (2006). Wikipedias: collaborative web-based encyclopedias as complex networks. // Physical Review E, 77:1 doi 10.1103/PhisRevE.74.016115.

Enviado: 2019-04-27. Segunda versión: 2019-07-02.
Aceptado: 2019-07-31.