ORGANIZACIÓN AUTOMATICA DE DOCUMENTOS MEDIANTE TÉCNICAS DE ANÁLISIS DE REDES

Carlos G. Figuerola (1), José Luis Alonso Berrocal (2), Angel Zazo Rodríguez (3)

(1) Universidad de Salamanca, Instituto de Estudios en Ciencia y Tecnología, c/ Espejo, sn 37008 Salamanca (España), figue@usal.es. (2) berrocal@usal.es. (3) zazo@usal.es.

Resumen

La organización automática de documentos permite conocer la estructura temática de grandes colecciones documentales. En este trabajo se plantea modelar la colección de documentos mediante una red o grafo de nodo y enlaces y aplicar técnicas de Análisis de Redes Sociales. Se describe un experimento llevado a cabo con una colección de noticias de prensa, analizando la estructura temática resultante tras aplicar técnicas de detección de comunidades de nodos en dos niveles de agrupamiento. Los resultados parecen bastante prometedores, planteando como trabajo futuro la implementación y comparación de diferentes algoritmos de detección de comunidades.

Palabras clave: clasificación automática, redes sociales, clustering.

Abstract

Automatic organization of documents allows to know the semantic structure of broad collections of documents. This paper proposes to model a document's collection by means of a graph or network and then apply the so called Social Networks Analysis techniques. We describe a practical experiment carried out with a collection of newspress articles, and then we analyze the topic's structure resulting after applying communities discovery techniques. Results look enough promising; we envisage as future work the application and comparation of diverse communities discovery algorithms .

Keywords: Automatic classification, clustering, Social Network Analysis.

. 1 Introducción

Uno de los retos de la llamada Sociedad del Conocimiento es la organización de documentos, de forma que sea posible encontrar una estructura lógica que permita un acceso más eficaz a la información contenida en ellos. En el territorio de la archivística clásica, el Principio de Procedencia (Martín-Pozuelo Campillos, 1996) ha permitido solventar esta cuestión, buscando una organización documental basada en la estructura interna de las organizaciones productoras de documentos.

La descripción a partir de campos, más o menos explícitos, y el uso de lenguajes documentales (controlados o no) se han aplicado durante décadas en el terreno de bibliotecas y centros de documentación; están orientados, no obstante, a la aplicación con documentos con soporte físico.

El desarrollo de las Tecnologías de la Información y, especialmente, Internet ha desestabilizado por completo lo que conocíamos como producción de documentos y, por ende, todo el proceso documental.

El problema ahora no es sólo la cantidad de documentos -también- sino otra serie de factores añadidos como la volatilidad, la falta de tipología, las autorías de nuevo tipo (colaborativas, y otras), nuevas estructuras documentales, y un largo etcétera.

Una vía de abordar esta situación es la automatización de procesos; sin embargo, está lejos de poder enfrentarse de forma eficaz a estos problemas. Una de las cuestiones que tiene planteadas es la organización automática de los documentos en función de su contenido temático. Compatible con otros modelos de organización y acceso, la organización automática puede favorecer la recuperación, al colocar documentos de contenidos similares juntos; o el análisis cuantitativo de las colecciones documentales; o el desarrollo y perfeccionamiento de aplicaciones de extracción de información basados en *machine learning*, al poder ofrecer gandes cantidades de contextos similares para el entrenamiento de dichas aplicaciones.

En este trabajo se describe un sistema de clasificación automática de documentos basado en técnicas de Análisis de Redes Sociales; se describe también su aplicación en un caso concreto y se analizan los resultados. Es importante remarcar que se trata de una aplicación real en un

caso real; no se trata, por consiguiente, de un experimento de laboratorio en el cual el entorno y los parámetros de aplicación están controlados. Antes bien, al tratarse de un caso específico en una situación real, el contexto viene dado por esa situación real; esto implica, por ejemplo, que las características de los documentos con los que se ha trabajado escapa a nuestra elección; o que la evaluación de los resultados carece de puntos de referencia con los cuales establecer comparaciones.

Este trabajo está organizado de la siguiente manera: en la sección siguiente se introduce la clasificación automática de documentos y algunos de los sistemas clásicos más utilizados; se introduce también el Análisis de Redes Sociales. En la siguiente sección se detalla la metodología aplicada, la colección de documentos utilizada, el procesamiento a que ha sido sometida, así como los algoritmos aplicados. A continuación, se describen los resultados obtenidos y se evalúan, discutiendo las métricas obtenidas. Finalmente, se ofrecen unas conclusiones.

. 2 La clasificación automática de documentos

Es frecuente distinguir entre sistemas de clasificación supervisada (a veces denominada también categorización) y clasificación no supervisada o clustering (Campos Ibáñez y Romero López, 2011; Ares Brea et al., 2011). En el primero de los casos los documentos se clasifican en una estructura creada *ad-hoc* por personas; la mayor parte de los sistemas aplican técnicas de *machine learning* y consisten, de una forma u otra, en construir patrones o modelos de las diferentes categorías; y medir después el parecido o similitud de cada documento con cada uno de esos patrones (Baharudin y otros, 2010).

Existen diferentes sistemas de clasificación supervisada, como los probabilísticos (Langley y otros, 1992; McCallum y Nigam, 1998), el llamado *Nearest Neighbour* (Yang, 1999) o los *SVM* (Joachims, 2002). Estos sistemas se vienen empleando con éxito en diferentes aplicaciones; esto incluye, obviamente, el campo documental, en el que se han plicado con el fin de clasificar documentos o texto (Eyheramendy, Lewis y Madigan, 2003; Kim et al., 2006; Joachims, T., 1998). Por nuestra parte, hemos podido documentar tasas de éxito superiores al 94 % (Figuerola, 2013; Quintanilla, Figuerola y Groves, 2015; Figuerola et al. 2017).

La clasificación no supervisada o clustering carece de estructuras clasificatorias pre-definidas y es el mismo sistema el que, en función de las características de la colección documental, las crea. Se trata, pues, de un escenario más cercano a la idea de organización automática o autoorganización de documentos.

Diversos métodos pueden utilizarse para clasificar temáticamente, de manera automática, colecciones amplias de documentos; una revisión de los más importantes puede encontrarse en (Aggarwal y Zhai, 2012). Así, dentro de lo que se conoce como clasificación no supervisada, es relativamente frecuente el uso de algoritmos de clustering como *k-means* (Jain, 2010). La versión clásica de éste tiene el inconveniente de producir clusters planos, de un solo nivel, además de necesitar fijar de antemano el número de clusters deseados, lo cual no siempre es fácil y requiere, con frecuencia, de un proceso de prueba y error.

Diferentes sistemas de modelado de temas (*topic modelling*) (Hidayat et al., 2015; Griffiths and Steyvers, 2004) se utilizan también para descubrir de forma automática los temas tratados en una colección de documentos. Obviamente, delimitados los temas, es posible agrupar o clasificar los documentos en función de esos temas. El procedimiento más conocido es el denominado *Latent Dirichlet Allocation* (*LDA*) (Blei, Ng and Jordan, 2003), que goza de cierto prestigio en el terreno de las *Digital Humanities* (Shawn and Milligan, 2012). Aunque existen implementaciones fáciles de utilizar (de ahí, tal vez, su relativa popularidad), tiene el inconveniente de que es preciso elegir de antemano el número de temas o topics deseados, lo cual no siempre es sencillo (Arun et al., 2010). Además, el etiquetado o identificación de los temas detectados suele requerir una fase de análisis por parte de expertos, no exenta de subjetividad e inconsistencias. Este problema del etiquetado es, en realidad, común a todos los sistemas de clustering; el sistema agrupa documentos afines pero es reposnsabilidad de los utilizadores etiquetar cada uno de tales agrupamientos con una expresión realmente significativa sobre el contenido de ese agrupamiento.

En este trabajo se plantea abordar la clasificación automática de documentos mediante la aplicación de técnicas de *Análisis de Redes* (Otte and Rousseau, 2002), efectuando una aplicación sobre una colección de documentos y analizando sus resultados. La aplicación de estas técnicas no requiere prefijar el número de *clusters* deseados, puede producir estructuras jerárquicas y, como se verá, los resultados de su aplicación parecen bastante prometedores.

. 3 Metodología

El Análisis de Redes tiene su origen en la teoría matemática de Redes o Grafos; una de sus apli-

caciones en Ciencias Sociales más tempranas es la efectuada en los años 60 del siglo pasado, por sociólogos intentado modelar las relaciones entre personas y grupos sociales (Scott, 2013). Brevemente, una red es un conjunto de nodos o vértices conectados por arcos o enlaces. Los nodos pueden tener una serie de características o atributos arbitrarios, definidos por quien aplica este artefacto. Los enlaces o arcos conectan dos nodos entre sí; los enlaces pueden tener dirección (parten de un nodo y apuntan a otro) o no (simplemente conectan dos nodos en una relación ambivalente); pueden existir también arcos reflexivos (parten y llegan al mismo nodo). Los arcos o enlaces también pueden tener atributos arbitrarios a gusto del usuario, pero uno de los más habituales es el peso: un valor numérico que intenta expresar la fortaleza de la relación que representa ese arco.

Lo interesante de las Redes es que se han desarrollado métodos y procedimientos para analizar la estructura interna de una red. De manera que, si conseguimos modelar un determinado fenómeno mediante una red, podemos utilizar esas técnicas de análisis para estudiar la estructura interna de ese fenómeno. Una de las cosas que es posible hacer con una red es descubrir o detectar las posibles comunidades de nodos que haya en ella (Plantié and Crampes, 2013).

Es posible modelar una colección de documentos como una red, en la cual los documentos pueden ser representados por nodos. Dos documentos o nodos pueden estar conectados entre sí por un arco o enlace si ambos tienen un contenido parecido o similar; y el peso de ese arco podría ser al grado de similitud entre esos dos documentos.

Desde finales de los años 60 del pasado siglo disponemos de formas para medir la similitud entre dos documentos, gracias a la formulación del modelo vectorial de recuperación de la información por G. Salton (Salton, 1983). Naturalmente, ha habido desde entonces nuevos modelos y nuevas propuestas y podemos aplicar cualquiera de ellas; pero las más utilizadas hoy día siguen estando basadas en el modelo vectorial de Salton, aún cuando no haya estado exento de polémicas (Leydesdorff, 2008). Básicamente, la similitud entre dos documentos se mide en función de la cantidad de palabras que éstos tienen en común; sucede que no todas las palabras tienen la misma importancia o peso y, en consecuencia, no cuentan lo mismo. Y sucede también que los documentos tienen tamaños diferentes (en número de palabras) y esto requiere aplicar algún sistema de normalización.

Así que, de un modo u otro, es posible calcular la similitud entre cada pareja de documentos de nuestra colección, trazando una red de nodos y arcos.

A esta red es posible aplicar alguno de los varios sistemas disponibles de detección de comunidades de nodos. Una comunidad de nodos es un conjunto de éstos que enlazan fuertemente entre sí, y débilmente con los no pertenecientes a esa comunidad. Dado que los enlaces de nuestra red están basados en las similitudes de contenido entre los nodos, una comunidad debería agrupar a documentos de temática similar (Lee and Cunningham, 2014; Pons and Latapy, 2005).

. 4 La colección de documentos

Hemos aplicado estas ideas a una colección de noticias de prensa extraídas de algunos periódicos. Esta colección está formada por 50.000 noticias, todas sobre Ciencia y Tecnología y es conocida como Spanish Corpus of Scientific Culture (SCSC) (Figuerola, Quintanilla et al., 2017). Se trata de noticias de tres diarios españoles de ámbito nacional en su versión digital, publicadas entre 2002 y 2011, a texto completo. A la totalidad de noticias extraídas de la hemeroteca digital de cada uno de los periódicos se aplicó un sistema de categorización para filtrar u obtener todas las noticias relacionadas con la Ciencia y/o la Tecnología.

El sistema de categorización automática aplicado, basado en SVM (*Support Vector Machine*), consiguió, en este caso, una precisión bastante elevada (94.5 % de aciertos), dejando la colección con una cantidad muy pequeña de ruido (Groves, Figuerola y Quintanilla, 2015). Las noticias se convirtieron a texto plano, desde su formato web original.

Pese a ser noticias sobre Ciencia y/o Tecnología, es obvio que, dentro de ese amplio campo, es posible preguntarse por su estructura temática: qué campos o disciplinas científicas están presentes, y en qué medida; aquí es donde entra la aplicación de las técnicas de Análisis de Redes.

Se calculó la similitud entre cada par de noticias o documentos aplicando un sistema clásico de *tf-idf* para estimar los pesos de cada palabra y el bien conocido coeficiente del coseno (como el que aplican muchos sistemas de recuperación de la información) para medir la similitud. El resultado es una semi-matriz cuadrada de 50.730 por 50.730 elementos. Esta matriz es bastante densa, puesto que casi todos los documentos tienen alguna palabra en común, aunque sean palabras muy poco significativas, de peso muy bajo.

Esta matriz de similitudes nos permite construir una red de 50.730 nodos (cada documento). Y

si enlazamos cada documento con los que mantiene alguna similitud, por baja que ésta sea, obtenemos una cantidad de enlaces cercano a 1.200 millones; una cantidad intratable debido a su tamaño. Se impone podar enlaces aplicando un umbral de similitud; tras varias pruebas, se aplicó un umbral de 0,09, eliminando todos los enlaces inferiores a este valor. En resultado es una red de 50.730 nodos y 23.612.149 enlaces.
Una representación visual de esa red puede observarse en la figura 1 (en Apéndice).
Esta representación está realizada mediante un algoritmo *force directed* conocido como *Open Ord*, (Martin et al., 2011), que intenta colocar los nodos más fuertemente enlazados más próximos entre sí. Sin entrar en las dificultades de conjugar un número tan alto de nodos y enlaces, claramente podemos observar como hay conjuntos de nodos (documentos) más o menos agrupados. Estos grupos es lo que conocemos como comunidades de nodos de una red y, dado que la relación entre los nodos, en nuestro caso, está basada en la similitud semántica; podemos pensar que esas comunidades agrupan documentos que tratan sobre los mismos o parecidos temas.
Podemos aplicar, pues, alguno de los algoritmos de detección de comunidades para obtener la estructura temática de nuestra red de documentos. Sin embargo, uno de los problemas de la detección de comunidades es que suele requerir un tiempo de procesamiento considerable (Lancichinetti and Fortunato, 2009); tras algunas pruebas, aplicamos el algoritmo de detección de comunidades conocido como *Infomap* (Rosvall, Axelsson and Bergstrom, 2009; Bohlin et al., 2014; Edler and Rosvall, 2015).
Una de sus características es su eficacia con redes grandes.

. 5 Resultados

Como resultado obtenemos 23 comunidades, muchas de ellas con un segundo nivel de subcomunidades cada una. Sin embargo, de ellas solamente 13 tienen un tamaño, en número de documentos, apreciable, abarcando el 90 % de toda la colección.
Un análisis a simple vista de las noticias componentes de cada comunidad significativa muestra la consistencia temática de las mismas; a modo de ejemplo, véanse las tablas III y IV (en Apéndice). La fig. 2 (en Apéndice) muestra las 13 comunidades más importantes (en número de documentos); cada una de ellas ha sido etiquetada manualmente tras una observación aleatoria de muestras de las noticias que las conforman.
La tabla III (en Apéndice) muestra un listado de las subcomunidades de la comunidad 3 (etiquetada como 'Energía'), junto con las palabras más significativas de cada una de ellas.
En líneas generales, parece que los grandes temas científicos de los que se ocupan las noticias analizadas tienen que ver con el medio ambiente en sus diversos aspectos (energía, conservación, recursos naturales, biodiversidad, contaminación …); así como los que tienen que ver con la Medicina en sentido amplio. También otros temas no tam abundantes en noticias, pero bien delimitados: investigación aerospacial, Astronomía, tecnologías de la información, evolución y política científica.
Una evaluación algo más rigurosa que la simple observación plantea algunas cuestiones, al no disponer de puntos de referencia· En efecto, en la evaluación de la clasificación automática suele distinguirse entre sistemas externos e internos. Los primeros se basan en contrastar los resultados obtenidos en una colección de documentos construída *ad-hoc*, de la cual se conoce previamente su estructura temática. Obviamente éste no es nuestro caso, por lo que debemos recurrir a un sistema interno de evaluación, es decir, medir la similitud entre los componentes de cada comunidad y la diferencia o distancia con los otros clusters.

Com.	Etiqueta	Silue.
1	Salud Pública	0.62
2	Biomedicina	0.52
3	Energía	0.53
4	Desarrollo Humano	0.56
5	Recursos naturales	0.55
6	Investigación Aerospacial	0.62
7	Biodiversidad	0.55
8	Astronomía y Cosmología	0.57
9	Tecnología de la Información	0.59
10	Politica Científica	0.56
11	Especies protegidas-España	0.66
12	Evolución Humana	0.58
13	Contaminación	0.56

Tabla I: Silueta media por comunidades

Una de las medidas más difundidas es la conocida como silueta (Rousseeuw, 1987; Rendón et al., 2011) de un documento. Éste es un coeficiente que se aplica a cada documento individual e intenta conjugar ambas cuestiones: la similitud con todos los integrantes de su misma comunidad y la separación con los documentos de la comunidad más próxima. Su valor oscila entre -1 y 1; los valores más alejados de -1 son

mejores. Además, cuanto más se acercan a 1 indican una buena separación entre comunidades, algo que depende mucho de los campos temáticos y de las características semánticas de los documentos.

Es importante hacer notar que la valoración que ofrece esta medida debe ser tomada con precaución; se basa exclusivamente en la misma valoración de la similitud (o distancia) entre documentos aplicada al efectuar la clasificación de éstos, lo cual introduce sesgos de base en este tipo de evaluación. Es, sin embargo, la medida que tenemos disponible.

La silueta es un valor individual para cada documento, por lo que para una valoración de conjunto tal vez es más útil calcular la silueta media. La Tabla I muestra los valores medios para cada comunidad importante (en número de documentos). Es preciso indicar que la separación entre comunidades se ha medido tomando como referencia la comunidad de primer nivel.

Los valores de silueta son bastante elevados; indican claramente una firme cohesión interna de cada comunidad o *cluster*, y una separación con el resto de los *clusters* bien definida.

La tabla V (en Apéndice) muestra las subcomunidades más significativas y sus valores de silueta. Dos observaciones son importantes para valorar esa tabla: en primer lugar no figuran todas las subcomunidades, puesto que ello daría lugar a una tabla de excesivo tamaño (más de 700 items). En su lugar figuran en ella solamente las subcomunidades de más de 100 documentos.

La tabla completa de todas las subcomunidades se ofrece en Apéndice independiente, dada su extensión.

El cálculo de la silueta, de otro lado, se ha efectuado teniendo en cuenta solamente los items dentro de cada subcomunidad. Por ello los valores de silueta a nivel de subcomunidad son relativamente bajos, incluso por debajo de 0 en algunos casos. La silueta media calculada por subcomunidades (nivel 2) es de 0.21, un valor algo bajo que puede explicarse por la relativamente baja separación entre *clusters*, dado que las subcomunidades más cercanas, que sirven como base para el cálculo de la silueta, forman parte de la misma comunidad y existe, por tanto, afinidad temática aunque de un nivel más amplio. Recordemos, de otro lado, que todas las noticias o documentos tratan sobre Ciencia y/o Tecnología, lo cual introduce ya un elemento de proximidad de contenido entre todas ellas.

. 6 Conclusiones

La clasificación automática de documentos nos permite conocer la estructura temática y su distribución en una colección de documentos. El Análisis de Redes nos permite modelar una colección de documentos como una red de afinidades temáticas, de manera que la aplicación en esa red de sistemas de detección de comunidades descubre los temas tratados en esos documentos y la estructura temática de la colección. A diferencia de otros sistemas de organización automática, el número de temas es establecido por el propio sistema en función de las características de la colección documental; además, es posible establecer al menos un subnivel de clasificación.

Los experimentos efectuados sobre una colección de noticias de prensa muestran una gran precisión en los resultados; como trabajo futuro, se plantea la aplicación de otros sistemas de detección de comunidades y la comparación entre los resultados obtenidos.

Referencias

Aggarwal, C. C. y Zhai, C. (2012). A survey of text clustering algorithms, en Aggarwal y Zhai, eds.: Mining Text Data. Springer US: Boston MA. 77--128

Ares Brea, M.E.; Parapar López, J.; Barreiro García, A. (2011). Agrupamiento Documental // Cacheda Seijo, F.; Fernández Luna, J.M. ; Huete Guadix, J.F. Eds. (2011). Recuperación de Información. Un enfoque práctico y multidisciplinar. Madrid; Ra-Ma, 2011. 392-416.

Arun, R.; Suresh, V.; Veni Madhavan, C. E.; Narasimha Murthy, M. N.; Zaki, M. J.; Yu, J. X.; Ravindran, B.; Pudi, V. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. // Advances in Knowledge Discovery and Data Mining: 14th Pacific-Asia Conference, *PAKDD 2010*. Hyderabad, India. 391-402. http://dx.doi.org/10.1007/978-3-642-13657-3_43 (2017-01-12).

Blei, D., Ng, A.; Jordan, M. (2003). Latent dirichlet allocation. // The Journal of Machine Learning Research, *3* 993-1022.

Baharudin, B; Lee, L. H.; Khan, K. (2010). A review of machine learning algorithms for text-documents classification. // Journal of Advances in Information Technology. 1:1 4–20.

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

Campos Ibáñez, L.M.; Romero López, A.E. (2011). Clasificación documental. // Cacheda Seijo, F.; Fernández Luna, J.M. ; Huete Guadix, J.F. Eds. (2011). Recuperación de Información. Un enfoque práctico y multidisciplinar. Madrid; Ra-Ma, 2011. 359-392.

Edler, D.; Rosvall, M. (2015). The infomap software package. http://www.mapequation.org/code.html (2017-02-16).

Eyheramendy, S., Lewis, D. D., & Madigan, D. (2003). On the naive bayes model for text categorization. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.20.4949 (2017-02-16).

Figuerola, C.G. (2013). Clasificación automática de documentos. Un caso práctico. http://grulla.usal.es/figuerola2013clasificacion.pdf (2017-02-16).

Figuerola, C.G.; Quintanilla Fisac, M.A. et al. (2017): Sistema de Indicadores para el SCSC (Spanish Corpus of Scientific Culture). http://grulla.usal.es/figuerola2017-sistema.pdf (2017-03-28).

Griffiths, T. L.; Steyvers, M. (2004). Finding scientific topics. // Proceedings of the National Academy of Sciences, 101:1, 5228-5235.

Groves, T; Figuerola, C.G. y Quintanilla, M.A (2015). Ten years of science news: A longitudinal analysis of scientific culture in the Spanish digital press. Public Understanding of Science. 25:6, 691 – 705. https://gredos.usal.es/jspui/handle/10366/127539 (2017-02-16)

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. // Pattern recognition letters, 31:8, 651-666. http://www.ppgia.pucpr.br/~fabricio/ftp/Roges/Jain-Clustering_PRL10.pdf (2017-02-18)

Joachims, T. (1998, April). Text categorization with support vector machines: Learning with many relevant features. In European conference on machine learning (pp. 137-142). Springer Berlin Heidelberg. https://eldorado.tu-dortmund.de/bitstream/2003/2595/1/report23_ps.pdf (2017-02-16).

Joachims T. (2002) Learning to Classify Text Using Support Vector Machines – Methods, Theory and Algorithms. Boston, MA: Kluwer Academic Publishers.

Kim, S. B., Han, K. S., Rim, H. C., & Myaeng, S. H. (2006). Some effective techniques for naive bayes text classification. IEEE transactions on knowledge and data engineering, 18(11), 1457-1466. http://ir.kaist.ac.kr/papers/2006/some%20effective%20techniques%20for%20naive%20bayes%20text%20classification.pdf (2017-02-16).

Hidayat, E. Y.; Firdausillah, F.; Hastuti, K.; Dewi, I. N.; Azhari, A. (2015). Automatic Text Summarization Using Latent Drichlet Allocation (LDA) for Document Clustering. // International Journal of Advances in Intelligent Informatics, 1:3 132-139.

Lancichinetti, A.; Fortunato, S. (2009) Community detection algorithms: A comparative analysis.// Physical Review E. 80:5. http://arxiv.org/pdf/0908.1062v2.pdf (2017-02-18)

Langley, P.; Iba, W.; Thompson, K. (1992). An analysis of bayesian classifiers. // Proceedings of National Conference on Artificial Intelligence. San Antonio, CA: AAAI Press andMIT Press.223–228

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 (2017-02-18)

Leydesdorff, L. (2008). On the normalization and visualization of author co-citation data: Salton's Cosine versus the Jaccard index. Journal of the American Society for Information Science and Technology, 59(1), 77-85.

Martin, S.; Brown, M.W.; Klavans, R.; Boyack K.W.(2011). OpenOrd: an open-source toolbox for large graph layout. // Proc. SPIE 7868, Visualization and Data Analysis 2011. doi:10.1117/12.871402

Martin-Pozuelo Campillos, M. P. (1996). La construcción teórica en archivística: el principio de procedencia. Madrid: Universidad Carlos III de Madrid.

McCallum, A.; Nigam, K. (1998) A comparison of event models for naive bayes text classification. // AAAI-98 workshop on learning for text categorization. 41-48. http://www.kamalnigam.com/papers/multinomial-aaaiws98.pdf (2016-12-14)

Otte, E.; Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. // Journal of information Science. 28:6 441-453. http://www.academia.edu/download/42254790/Social_Network_Analysis_A_Powerful_Strat20160206-25456-1pc1lcl.pdf (2017-02-18)

Plantié, M. ; Crampes, M. (2013) Survey on social community detection. // Social media retrieval, 65–85. http://hal.archives-ouvertes.fr/docs/00/80/42/34/PDF/Survey-on-Social-Community-Detection-V2.pdf (2017-02-18)

Pons, P.; Latapy, M. (2005). Computing communities in large networks using random walks. // Computer and information sciences (ISCIS) 284–293. http://arxiv.org/abs/physics/0512106 (2017-02-18)

Rendón, E.; Abundez, I.; Arizmendi, A.; Quiroz, E. (2011). Internal versus external cluster validation indexes. // International Journal of computers and communications. 5:1 27-34.

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

Rousseeuw, P. J. (1987). Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. Computational and Applied Mathematics. 20 53–65. doi:10.1016/0377-0427(87)90125-7.

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

Shawn, G.; Milligan, I. (2012).Review of MALLET, produced by Andrew Kachites McCallum. // Journal of Digital Humanities, 2:1, http://journalofdigitalhumanities.org/2-1/review-mallet-by-ian-milligan-and-shawn-graham/ (2017-03-15)

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. // Information retrieval. 1:1-2 69-90.

Apéndice

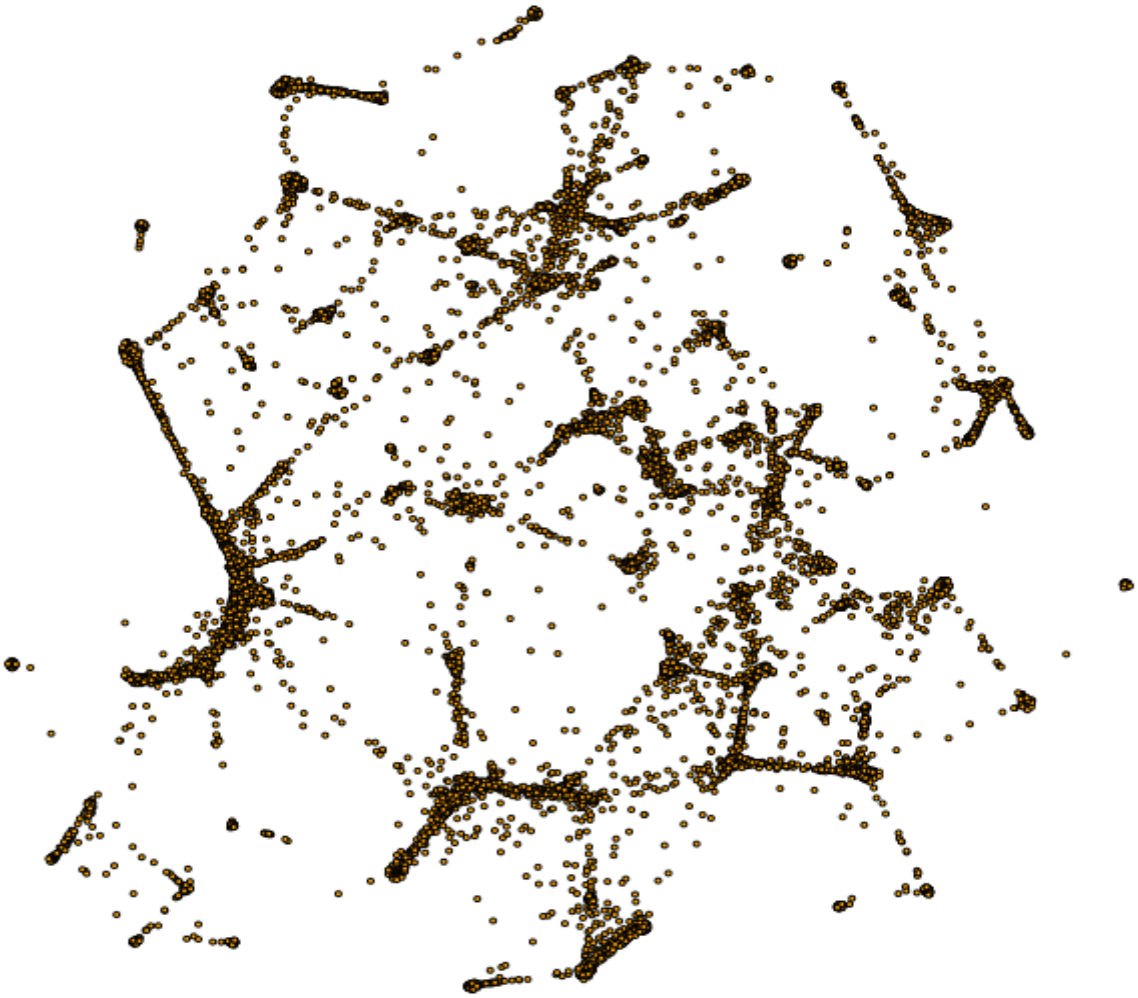

Figura 1. Red de documentos

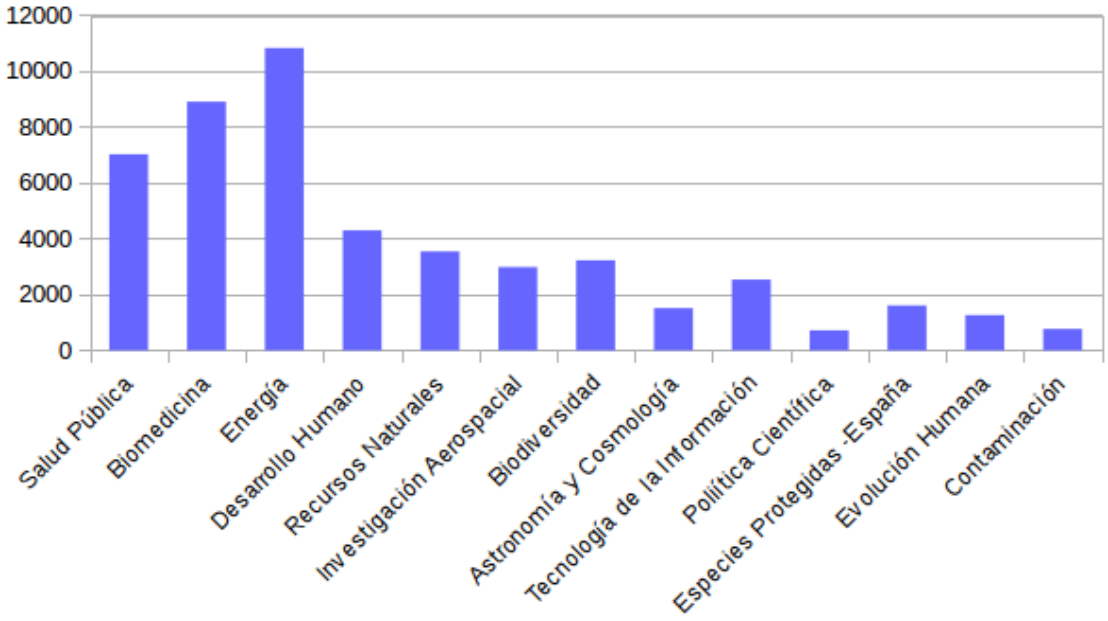

Figura 2 Comunidades etiquetadas y su tamaño

Subc.	Palabras clave	Subc.	Palabras clave
3.1	energía, nuclear, solar	3.10	gas natural
3.2	cambio climático, emisiones CO2, kyoto	3.11	seísmos, volcanes
3.3	energía nuclear, Fukushima	3.12	nuclear, reactor fusión
3.4	energía nuclear, centrales nucleares en España	3.13	medio ambiente, capa de ozono
3.5	cambio climático, calentamiento global, ecologis-mo	3.14	energía, petróleo
3.6	seísmos, tsunamis	3.15	automóvil, motor, motor y medio ambiente
3.7	medio ambiente, residuos, contaminación	3.16	automóvil, tráfico, seguridad vial
3.8	medio ambiente, contaminación	3.18	arquitectura, rascacielos
3.9	medio ambiente, cambio climático edificación	3.19	casas solares

Tabla II: Subcomunidades de la comunidad 3

Comunidad	Titular
1.1	¿Cambia algo la situación en España tras la primera muerte por la nueva gripe A/H1N1?
1.1	'La gripe A no es agresiva. Ha habido suerte'
1.1	El rebrote de la gripe del pollo en Asia reactiva el riesgo de epidemia mundial
1.1	China detecta 'gripe del pollo' en cerdos, el paso para que el virus salte a humanos
1.1	El riesgo de pandemia es inminente
1.1	Lecciones del pasado

1.1	La gripe, de vacuna en vacuna
1.1	Guía informativa de la gripe aviar
1.1	La OMS espera que el virus de la gripe se propague hasta en 'tres oleadas' distintas
1.1	¿Y qué pasaría si todo fuera mal?
1.1	La primera pandemia del siglo
1.1	Cerca de 400.000 personas han padecido la gripe A en todo el mundo
1.1	Sanidad refuerza la campaña de la gripe por el riesgo de una epidemia a gran escala
1.1	La OMS incluye la cepa de la gripe A en la nueva vacuna de gripe estacional
1.1	Así será la pandemia de gripe aviar
1.1	La OMS pronostica una "explosión" de nueva gripe en invierno
1.1	Zafarrancho contra la gripe aviar - Público.es
1.1	Alarma en México y EE UU por una rara gripe porcina
1.1	Hora de armarse contra la pandemia

Tabla III: Comunidad 1, ejemplo de subcomunidad 1.1

Comunidad	**Titular**
1.6	Mosquitos transgénicos contra la malaria
1.6	Malaria, un recorrido desde la quinina hasta la esperada vacuna
1.6	Acabar con la malaria costará 4.400 millones durante 50 años - Público.es
1.6	Males importados pero no exóticos
1.6	Los más ricos, entre los más pobres
1.6	Un tratamiento de 0,20 euros al año por niño evita un 20% de la malaria infantil
1.6	Optimismo ante un tratamiento para prevenir la malaria en niños
1.6	La malaria regresa a Europa
1.6	Noventa millones de niños africanos siguen sin dormir bajo una mosquitera
1.6	La terapia contra la malaria sólo alcanza al 16% de los enfermos
1.6	La malaria mata cada año a más de un millón de personas en África
1.6	Una vacuna ejemplar
1.6	Pedro Alonso: 'La vacuna contra la malaria estará lista en 2011'
1.6	Bill y Melinda Gates: "No pararemos hasta la erradicación"
1.6	Las iniciativas antimalaria salvaron a 750.000 niños

Tabla IV: Comunidad 1, ejemplo de subcomunidad 1.6

Com.	Subcom.	silueta	Com.	Subcom.	silueta
1	1	0,22	3	11	0,44
1	2	0,19	4	1	0,10
1	3	0,16	4	2	0,10
1	4	0,16	4	3	-0,02
1	6	0,34	4	5	0,37
1	7	0,57	5	1	0,05
2	1	-0,07	5	2	0,11
2	2	0,03	6	1	0,09
2	3	-0,39	6	2	0,38
2	4	-0,05	6	6	0,33
2	5	0,37	7	1	-0,11
2	6	-0,07	7	2	0,36
2	7	0,38	7	3	0,32
2	8	0,20	8	1	0,34
2	14	-0,21	8	2	0,24
2	18	-0,08	8	3	0,51
3	1	-0,11	9	1	0,01
3	2	0,10	9	2	0,00
3	3	0,34	9	3	0,44
3	4	0,53	10	1	0,02
3	5	0,18	11	1	0,42
3	6	0,32	12	1	0,04
3	7	0,20	13	1	0,15
3	9	0,01			

Tabla V: Silueta de subcomunidades (sólo mayores de 200 documentos)

Valores de silueta por comunidads y subcomunidades

Comunidad 1

Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta
1	2268	0,22	12	107	0,63	23	9	0,78	34	2	0,86
2	1681	0,19	13	133	0,54	24	18	0,62	35	2	0,92
3	373	0,16	14	51	0,69	25	9	0,64	36	3	0,77
4	725	0,16	15	112	0,53	26	6	0,82	37	3	0,80
5	161	0,46	16	40	0,69	27	5	0,83	38	3	0,56
6	296	0,34	17	32	0,62	28	3	0,81	39	4	0,81
7	268	0,57	18	38	0,37	29	4	0,82	40	2	0,55
8	138	0,58	19	21	0,81	30	7	0,55	41	2	0,72
9	101	0,62	20	29	0,65	31	5	0,79	42	2	0,92
10	158	0,31	21	10	0,79	32	2	0,80			
11	167	0,39	22	21	0,76	33	3	0,79			

Comunidad 2

Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta
1	1415	-0,07	50	28	0,81	99	10	0,91	148	6	0,74
2	1641	0,03	51	40	0,61	100	6	0,83	149	5	0,36
3	1377	-0,39	52	27	0,45	101	7	0,58	150	3	0,49
4	483	-0,05	53	31	0,41	102	5	0,65	151	3	0,44
5	311	0,37	54	16	0,56	103	10	0,78	152	3	0,79
6	374	-0,07	55	38	0,39	104	11	0,25	153	5	0,70
7	275	0,38	56	23	0,09	105	9	0,37	154	5	0,38
8	266	0,20	57	16	0,41	106	6	0,54	155	4	0,85
9	188	-0,04	58	45	0,12	107	7	0,54	156	3	0,71
10	146	0,25	59	66	0,26	108	7	0,63	157	4	0,38
11	97	0,22	60	38	0,64	109	3	0,48	158	3	0,71
12	176	0,10	61	33	0,52	110	8	0,79	159	2	0,85
13	100	0,42	62	25	0,57	111	9	0,62	160	2	0,55
14	361	-0,21	63	20	0,84	112	6	0,81	161	3	0,76
15	107	0,49	64	23	0,76	113	5	0,25	162	2	0,85
16	179	0,33	65	16	0,57	114	4	0,59	163	4	0,58
17	94	0,15	66	27	0,58	115	6	0,81	164	2	0,53
18	211	-0,08	67	11	0,67	116	7	0,65	165	3	0,58

19	95	0,46	68	24	0,58	117	5	0,43	166	3	0,32
20	121	0,11	69	23	0,62	118	8	0,54	167	3	0,43
21	110	0,34	70	23	0,80	119	4	0,72	168	3	0,48
22	54	0,05	71	18	0,52	120	5	0,65	169	2	0,89
23	83	-0,12	72	9	0,66	121	5	0,73	170	6	0,79
24	75	0,09	73	12	0,20	122	6	0,44	171	2	0,69
25	60	0,48	74	13	0,53	123	3	0,82	172	2	0,79
26	48	0,58	75	12	0,26	124	5	0,53	173	3	0,64
27	84	0,21	76	7	0,48	125	5	0,84	174	2	0,88
28	50	0,47	77	14	0,76	126	5	0,72	175	4	0,33
29	44	0,48	78	9	0,61	127	4	0,77	176	3	0,78
30	53	0,47	79	17	0,42	128	3	0,70	177	2	0,80
31	48	0,37	80	11	0,40	129	6	0,49	178	2	0,84
32	50	0,39	81	9	0,67	130	4	0,68	179	3	0,59
33	30	0,55	82	7	0,85	131	3	0,85	180	5	0,77
34	48	0,66	83	11	0,61	132	3	0,77	181	3	0,72
35	41	0,69	84	12	0,73	133	6	0,26	182	2	0,64
36	42	0,27	85	13	0,63	134	7	0,37	183	2	0,81
37	35	0,39	86	11	0,33	135	4	0,80	184	2	0,37
38	60	-0,05	87	11	0,31	136	7	0,87	185	3	0,55
39	78	0,38	88	13	0,37	137	2	0,91	186	2	0,62

40	61	0,15	89	11	0,55	138	4	0,48	187	2	0,49
41	32	0,44	90	13	0,42	139	3	0,61	188	2	0,91
42	49	0,46	91	12	0,75	140	5	0,45	189	2	0,57
43	17	0,70	92	22	0,66	141	5	0,77	190	2	0,82
44	32	0,59	93	12	0,58	142	6	0,52	191	2	0,43
45	36	0,66	94	6	0,56	143	3	0,73	192	2	0,72
46	31	0,53	95	15	0,52	144	4	0,78	193	2	0,39
47	31	0,65	96	9	0,51	145	2	0,91	194	2	0,83
48	30	0,58	97	10	0,79	146	3	0,57	195	2	0,91
49	67	-0,08	98	12	0,60	147	3	0,69			

Comunidad 3

Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta
1	1861	-0,11	19	42	0,66	37	17	0,55	55	5	0,86
2	2621	0,10	20	57	0,35	38	13	0,64	56	5	0,70
3	710	0,34	21	87	0,41	39	13	0,68	57	4	0,67
4	435	0,53	22	44	0,39	40	9	0,59	58	4	0,68
5	587	0,18	23	46	0,54	41	16	0,63	59	2	0,55
6	225	0,32	24	35	0,64	42	9	0,68	60	5	0,90
7	277	0,20	25	37	0,64	43	8	0,42	61	3	0,77
8	169	0,35	26	29	0,80	44	6	0,78	62	3	0,81
9	222	0,01	27	43	0,46	45	7	0,56	63	3	0,46
10	140	0,07	28	39	0,49	46	10	0,62	64	2	0,58
11	273	0,44	29	15	0,61	47	2	0,60	65	2	0,37
12	62	0,74	30	23	0,58	48	4	0,78	66	2	0,85
13	78	0,75	31	16	0,81	49	3	0,84	67	4	0,60
14	41	0,63	32	14	0,48	50	4	0,29	68	2	0,47
15	145	0,25	33	28	0,61	51	5	0,74	69	3	0,62
16	88	0,26	34	23	0,60	52	8	0,43	70	2	0,44
17	59	0,57	35	21	0,50	53	5	0,61	71	2	0,80
18	94	0,32	36	20	0,71	54	5	0,46			

Comunidad 4

Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta
1	930	0,10	12	70	0,54	23	12	0,67	34	3	0,69
2	563	0,10	13	46	0,48	24	16	0,81	35	2	0,44
3	434	-0,02	14	49	0,41	25	9	0,45	36	2	0,50
4	175	0,56	15	40	0,59	26	11	0,56	37	2	0,45
5	213	0,37	16	28	0,55	27	7	0,70	38	2	0,82
6	107	0,38	17	14	0,52	28	2	0,91	39	2	0,61
7	110	0,50	18	23	0,67	29	5	0,71	40	2	0,58
8	68	0,71	19	13	0,70	30	4	0,82	41	2	0,79
9	95	0,30	20	13	0,48	31	3	0,92			
10	51	0,56	21	17	0,67	32	3	0,77			
11	50	0,46	22	23	0,67	33	3	0,85			

Comunidad 5

Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta
1	1496	0,05	16	16	0,62	31	10	0,52	46	6	0,47
2	238	0,11	17	18	0,58	32	5	0,63	47	2	0,90
3	112	0,36	18	24	0,60	33	11	0,62	48	4	0,36
4	58	0,37	19	28	0,74	34	8	0,46	49	4	0,70

5	102	0,17	20	21	0,56	35	7	0,90	50	3	0,58
6	62	0,42	21	12	0,57	36	6	0,59	51	2	0,87
7	86	0,12	22	22	0,79	37	6	0,56	52	4	0,84
8	90	0,41	23	12	0,49	38	5	0,68	53	2	0,38
9	43	0,70	24	24	0,54	39	6	0,60	54	2	0,86
10	76	0,22	25	16	0,34	40	4	0,80	55	2	0,67
11	70	0,46	26	15	0,74	41	3	0,68	56	2	0,64
12	31	0,53	27	12	0,71	42	4	0,61			
13	50	0,53	28	8	0,63	43	4	0,62			
14	37	0,63	29	12	0,76	44	3	0,75			
15	64	0,26	30	6	0,40	45	4	0,53			

Comunidad 6

Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta
1	2232	0,09	10	166	0,39	19	4	0,58	28	3	0,79
2	633	0,38	11	61	0,76	20	11	0,43	29	2	0,88
3	160	0,65	12	44	0,73	21	4	0,53	30	3	0,80
4	127	0,56	13	64	0,43	22	9	0,68	31	2	0,84
5	88	0,61	14	7	0,45	23	3	0,80	32	2	0,76
6	298	0,33	15	21	0,39	24	4	0,87	33	2	0,84
7	63	0,71	16	27	0,76	25	3	0,84	34	3	0,80
8	90	0,66	17	21	0,54	26	2	0,59	35	3	0,60
9	121	0,59	18	8	0,35	27	2	0,91	36	2	0,51

Comunidad 7

Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta
1	719	-0,11	15	31	0,50	29	16	0,63	43	6	0,23
2	304	0,36	16	35	0,40	30	21	0,59	44	3	0,86
3	209	0,32	17	34	0,18	31	13	0,77	45	4	0,58
4	79	0,51	18	21	0,64	32	15	0,76	46	6	0,44
5	84	0,50	19	39	0,28	33	13	0,60	47	4	0,68
6	93	0,71	20	21	0,67	34	7	0,86	48	4	0,75

7	108	0,38	21	22	0,53	35	12	0,76	49	3	0,74
8	68	0,50	22	27	0,53	36	10	0,58	50	3	0,51
9	68	0,39	23	29	0,52	37	6	0,83	51	4	0,46
10	57	0,50	24	43	0,55	38	6	0,53	52	2	0,72
11	55	0,61	25	30	0,67	39	6	0,83	53	3	0,51
12	33	0,39	26	27	0,56	40	10	0,38	54	4	0,66
13	34	0,65	27	12	0,45	41	11	0,46	55	3	0,65
14	35	0,51	28	13	0,56	42	4	0,79	56	2	0,82

Comunidad 8

Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta
1	448	0,34	5	28	0,44	9	9	0,54	13	2	0,64
2	708	0,24	6	26	0,50	10	3	0,56	14	2	0,79
3	215	0,51	7	6	0,84	11	5	0,74			
4	52	0,62	8	7	0,69	12	2	0,54			

Comunidad 9

Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta
1	2594	0,01	14	15	0,68	27	5	0,59	40	2	0,57
2	240	0,00	15	18	0,60	28	4	0,75	41	2	0,77
3	241	0,44	16	10	0,54	29	3	0,53	42	2	0,69
4	30	0,94	17	31	0,16	30	4	0,76	43	3	0,67
5	32	0,68	18	11	0,41	31	4	0,72	44	3	0,48
6	49	0,45	19	20	0,24	32	4	0,80	45	2	0,82
7	22	0,69	20	9	0,63	33	5	0,53	46	3	0,45
8	25	0,53	21	6	0,73	34	3	0,77	47	2	0,81
9	18	0,43	22	5	0,59	35	4	0,41	48	2	0,86
10	25	0,58	23	3	0,91	36	2	0,79	49	2	0,49
11	25	0,16	24	7	0,39	37	3	0,75	50	3	0,86

12	10	0,80	25	3	0,82	38	3	0,56	51	2	0,72
13	15	0,41	26	3	0,61	39	2	0,78			

Comunidad 10

Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta
1	802	0,02	11	19	0,78	21	4	0,63	31	2	0,46
2	161	0,11	12	15	0,76	22	4	0,58	32	4	0,63
3	152	0,13	13	25	0,61	23	3	0,67	33	3	0,43
4	105	0,20	14	9	0,52	24	3	0,58	34	2	0,62
5	85	0,33	15	10	0,48	25	6	0,37	35	2	0,87
6	38	0,41	16	7	0,73	26	3	0,85	36	2	0,32
7	26	0,76	17	7	0,71	27	3	0,83	37	2	0,59
8	29	0,50	18	7	0,76	28	2	0,88	38	2	0,52
9	19	0,63	19	4	0,88	29	4	0,48	39	2	0,74
10	23	0,48	20	5	0,82	30	3	0,28	40	2	0,64

Comunidad 11

Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta
1	369	0,42	5	18	0,71	9	4	0,89	13	2	0,71
2	141	0,37	6	6	0,74	10	2	0,83	14	2	0,84
3	80	0,32	7	5	0,70	11	2	0,87			
4	79	0,56	8	5	0,67	12	2	0,74			

Comunidad 12

Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta
1	509	0,04	10	29	0,63	19	7	0,40	28	3	0,35
2	156	0,50	11	26	0,34	20	9	0,67	29	2	0,73
3	156	0,31	12	15	0,48	21	4	0,52	30	3	0,85
4	41	0,63	13	12	0,56	22	4	0,54	31	2	0,63
5	44	0,51	14	8	0,52	23	5	0,74	32	2	0,80
6	39	0,67	15	9	0,39	24	3	0,73	33	2	0,84
7	27	0,62	16	20	0,71	25	2	0,82	34	2	0,88
8	65	0,45	17	6	0,70	26	4	0,88			
9	36	0,27	18	4	0,77	27	5	0,48			

Comunidad 13

Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta	Sub-com.	Num.Do-cs.	Silueta
1	409	0,15	6	9	0,63	11	5	0,65	16	4	0,67
2	139	0,46	7	16	0,39	12	7	0,57	17	3	0,70
3	83	0,39	8	11	0,58	13	5	0,65	18	4	0,55
4	15	0,62	9	10	0,75	14	3	0,83			
5	16	0,60	10	22	0,50	15	7	0,58			