ARTÍCULO ORIGINAL

Detección de comunidades a partir de redes de coautoría en grafos RDF

Community detection using co-authorship networks in RDF graphs

Detecção de comunidades a partir de redes de coautoria em grafos RDF

Ernesto Ortiz Muñoz, Yusniel Hidalgo Delgado

Universidad de las Ciencias Informáticas. La Habana, Cuba.

RESUMEN

La detección de comunidades se refiere al problema de identificar comunidades o particiones de nodos que comparten propiedades comunes en una red. Las redes de coautoría se consideran redes complejas, donde los nodos de la red son los autores, y los enlaces entre los nodos establecen la relación de coautoría en una o varias publicaciones. En los últimos años se han desarrollado investigaciones con el objetivo de publicar metadatos bibliográficos siguiendo los principios de los datos enlazados. Como resultado se obtienen grafos RDF que contienen los autores y las relaciones de coautoría que se establecen entre ellos. En este artículo se propone un método para la detección y visualización de comunidades en grafos RDF teniendo en cuenta las relaciones de coautoría como indicador para medir la colaboración científica. Con la implementación del método se pretende dotar a los especialistas en ciencias de la información de una herramienta de análisis que ayude en el proceso de toma de decisiones y la realización de estudios en esta área.

Palabras clave: autoría y coautoría en la publicación científica, bibliometría, metodologías computacionales, Grafo RDF.

ABSTRACT

Community detection refers to the problem to identify communities or partitions of nodes that shares common properties in a network. The co-authorship networks are considered complex networks, where the nodes of the network are authors and the edges between nodes provides the co-authorship relationships in one or more publications. In recent years, researches have been carried out with the aim to publish bibliographic metadata following the principles of the linked data. Resulting from this, RDF graphs are obtained, containing the authors and the co-authorship relationships among them. In this paper we propose a method for detecting and visualizing communities in RDF graphs, considering co-authorship relationships as an indicator to measure scientific collaboration. With the implementation of the method proposed, we provide an analysis tool for specialists in information sciences, which improve the process of decision making and implementation of studies in the area.

Key words: authorship and co-authorship in scientific publications, bibliometrics, computing methodologies, RDF graph.

RESUMO

A detecção de comunidades refere-se ao problema de identificar comunidades o partições de nodos que partilham propriedades comuns em uma rede. As redes de coautoria são consideradas redes complexas, onde os nodos da rede são os autores, e os enlaces entre os nodos estabelecem a relação de coautoria em uma ou várias publicações. Nos últimos anos se têm desenvolto investigações com o objetivo de publicar metadados bibliográficos seguindo os princípios dos dados ligados. Como resultado se obtêm grafos RDF que contêm os autores e as relações de coautoria que se estabelecem entre eles. Neste artigo se propõe um método para a detecção e visualização de comunidades em grafos RDF tendo em conta as relações de coautoria como indicador para medir a colaboração científica. Com a implementação do método se pretende dotar aos especialistas em ciências da informação de uma ferramenta de análise que ajude no processo de toma de decissões e a realização de estudos nesta área.

Palavras chave: autoria e coautoria na publicação científica, bibliometria, metodologias computacionais, Grafo RDF.

INTRODUCCIÓN

En siglos pasados la ciencia era considerada como una actividad solitaria; pero hace varias décadas se ha convertido en una actividad realizada en grupos. La colaboración ha sido intrínseca a la actividad científica, que va más allá de la creciente especialización descrita por *Beaver* y *Rosen* en uno de los estudios de esta materia.[1] La colaboración es considerada un desarrollo complejo, una forma de intercambiar información para trabajar juntos, utilizar los recursos de forma racional y perpetuar comunidades de científicos y tecnólogos.[2] Lo anterior evidencia que la colaboración más que una necesidad se ha convertido en una elección.

Revisiones bibliográficas hacen referencia al aumento de artículos científicos donde se pone en práctica la coautoría. La colaboración científica es considerada como un requisito previo para la coautoría.[3] Varios investigadores se refieren a la coautoría como elemento importante en las redes de colaboración científica (RCC). *Miquel* y otros presentan a la coautoría como el único indicador disponible para analizar la colaboración científica y que, a partir de los resultados que con ella se obtienen, surgen instrumentos estratégicos de gran valor tanto para los investigadores que participan en esos trabajos como para las entidades (universidades, organismos de investigación, ministerios, etcétera) involucradas en su elaboración.[4]

Para *Stokes* y *Hartley* las asociaciones de coautoría entre científicos reconocen tanto las deudas intelectuales como las personales y ofrecen la posibilidad de identificar y medir la actividad social y la influencia entre distintas especialidades científicas. El examen de los enlaces de coautoría entre científicos muestra a aquellos investigadores que trabajan en la misma área de conocimiento, aunque no necesariamente en conjunto. El resultado es un número de grupos colaboradores de tamaño variable, conectados o aislados los unos de los otros, dentro de los cuales algunos científicos juegan un papel principal; otros son los que sirven de nexo, de unión entre grupos y otros desempeñan ambos papeles simultáneamente.[5] Las redes de coautoría (RC) son representaciones de distintas literaturas académicas que han sido objeto de análisis cuantitativo en los últimos años. En una RC los nodos son autores y una arista no dirigida conecta dos autores si han escrito una publicación en conjunto.[6] Las RC forman estructuras de comunidades, lo cual constituye una propiedad de las redes complejas.[7] Las redes complejas son redes cuya estructura es irregular, compleja y en evolución dinámica en el tiempo.[8]

Una comunidad puede ser definida como un conjunto de nodos que están más densamente conectados entre ellos que con el resto de la red.[9] La importancia de este planteamiento radica en que se espera que los nodos que están contenidos dentro de una misma comunidad compartan atributos, características comunes o relaciones funcionales. Es necesaria la aplicación de la teoría de redes para al estudio de las comunidades presentes en las RC a partir de las publicaciones científicas.

Las comunidades detectadas a partir de RC constituyen elementos de análisis de gran impacto en estudios bibliométricos y por los propios investigadores. Permiten además identificar y cuantificar las relaciones de colaboración existentes entre los autores de diversas instituciones y áreas del conocimiento. Por otra parte, han alcanzado un notorio avance la publicación de metadatos bibliográficos como datos enlazados. Los datos enlazados se refieren a un conjunto de principios y buenas prácticas para la publicación y enlazado de datos estructurados en la web. Con la publicación de metadatos bibliográficos como datos enlazados se proporciona mayor interoperabilidad, descubrimiento y visibilidad de los recursos en el espacio de la web.

Varias tecnologías se han desarrollado por la W3C para estandarizar el proceso de publicación de datos siguiendo los principios de los datos enlazados. Algunas de estas tecnologías son RDF, SPARQL y OWL. RDF es un modelo de datos basado en grafos dirigidos para describir recursos en el contexto de la web. Por su parte, SPARQL es un lenguaje de consultas para el modelo de datos RDF y, por último, OWL es un lenguaje para la creación y mantenimiento de ontologías en el contexto de la web semántica.

En este artículo se propone un método para la detección y visualización de comunidades en las redes de coautoría existentes en los grafos RDF aplicando teoría de redes complejas. El método propuesto permite identificar y cuantificar las relaciones de colaboración que se establecen entre los autores presentes en los metadatos bibliográficos.

MÉTODOS

El método propuesto cuenta con tres fases para la detección de comunidades a partir de redes de coautoría en grafos RDF. Este método sigue un enfoque basado en tuberías o filtros donde la salida de una fase constituye la entrada a la próxima. Las fases propuestas son las siguientes:

1. Modelación de la red de coautoría.
2. Detección de las comunidades.
3. Visualización de las comunidades detectadas.

La figura 1 muestra las fases del método propuesto. Se muestran en orden lógico el lugar donde intervienen los diferentes artefactos, algoritmos y herramientas en cada una de las fases.

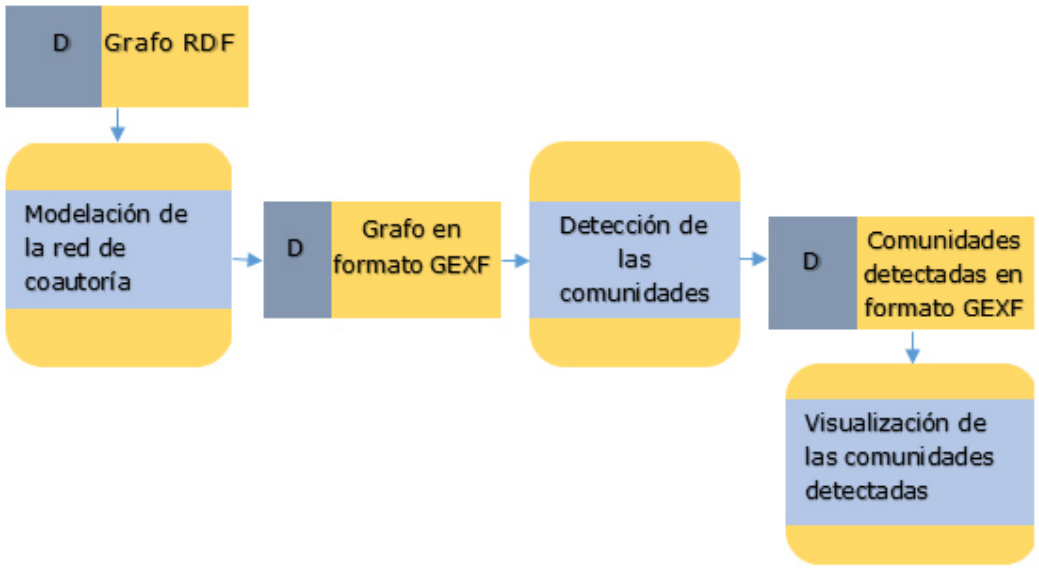
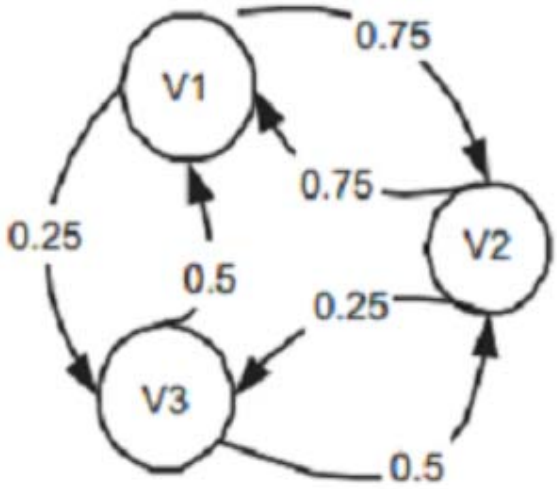


Fig. 1. Fases del método de solución propuesto.

MODELACIÓN DE LA RED DE COAUTORÍA

El modelo propuesto posee como entrada un grafo RDF que describe los metadatos pertenecientes a las publicaciones científicas de varios autores y sus coautores. Este grafo se modela como una red de coautores dirigida y ponderada. Lo anterior permite determinar una expresión de la magnitud de las relaciones entre los nodos de la red. La coautoría del grafo G es denotada como: $G= (V,E,W)$, donde V es el conjunto de nodos (autores), E es el conjunto de aristas (relaciones coautor entre

los autores) y *W* es el conjunto de los pesos W_{ij} asociados con cada arista de conexión de un par de autores. En la figura 2 se muestra una representación de la modelación de la red de coautoría ponderada y dirigida.



Tomada de: Liu X, Bollen J, Nelson ML, Van de Sompel H. Co-authorship networks in the digital library research community. Inf Process Manag. 2005;41(6):1462-80.

Fig. 2. Red de coautoría ponderada y dirigida.

Para determinar la magnitud de la relación entre dos autores en base a dos factores se propone:[10]

- *La frecuencia de coautoría:* autores que con frecuencia son coautores deben tener un mayor peso de coautoría.
- *Número total de coautores en los artículos:* si un artículo tiene muchos autores, se presupone que cada relación individual de coautoría debe tener una ponderación menor.

Para determinar el peso de las relaciones de coautoría entre los autores se denotan:[10]

- $A = \{a_1, \ldots, a_k, \ldots, a_n\}$ conjunto de artículos.
- *f* (a_k)*:* número de autores del artículo a_k.
- $V = \{v_1, \ldots, v_n\}$ conjunto de autores.

Luego se calcula como:

- Exclusividad: $G_{i,j,k} = \frac{1}{(f(a_k)-1)}$
- Frecuencia de coautoría: $C_{ij} = \sum_{k=1}^{n} G_{i,j,k}$

- Normalización del peso: $$W_{ij} = \frac{C_{ij}}{\sum_{k=1}^{n} C_{ik}}$$

Para culminar la etapa de modelación del grafo RDF como una red de coautoría dirigida y ponderada es necesario generar una red o grafo dirigido y ponderado. Para la generación del grafo se utilizó el formato de archivo para la descripción de grafos GEXF*. Este formato permite modelarestructuras de redes complejas, sus datos y las dinámicas asociadas. Es extensible y abierto, y es adecuado para aplicaciones específicas reales.

DETECCIÓN DE LAS COMUNIDADES

Teniendo como entrada el archivo GEXF generado en la fase anterior es necesario detectar las comunidades presentes en él. Sin existir una definición exacta de lo que es o debe ser una comunidad, esto ha generado multitud de inconvenientes a la hora de dividir una red en sus distintas comunidades, lo que se conoce como partición o *clustering*. Diversos métodos y algoritmos han sido desarrollados para intentar extraer la partición óptima de una red. Algunos de ellos tratan de optimizar un índice global de calidad de la partición, como son *Modularity*[11] o *Surprise*,[12] donde *Modularity* es la función de calidad más popular.

En el año 2008 se propuso el método para la detección de comunidades *Fast unfolding*.[13] Es un método simple para extraer la estructura de la comunidad a partir de redes grandes. Se basa en la optimización de la modularidad y supera otros métodos conocidos de detección de comunidades en términos de tiempo de cálculo. Además, la calidad de las comunidades detectadas es muy buena, según lo medido por *Modularity*. En esta propuesta se utiliza este algoritmo para detectar las comunidades presentes en el archivo GEXF y calcular diferentes valores y/o métricas que se encuentran en la red.

VISUALIZACIÓN DE LAS COMUNIDADES DETECTADAS

En la representación visual del conjunto de comunidades es necesario una herramienta o librería que permita explorar y modelar grafos o redes. IGraph, Java Universal Network/Graph (JUNG), Cytoscape y Gephi son algunas de las soluciones informáticas existentes con este objetivo. Se ha realizado una comparación de las herramientas anteriormente mencionadas, tomándose un grafo generado a partir de la red social Facebook con 466 nodos y 4 655 aristas, donde Gephi es la de mejor rendimiento, capacidades y extensibilidad.[14]

Gephi** se destaca por ser una herramienta de código abierto y multiplataforma. Está desarrollada en el lenguaje de programación Java y se distribuye bajo licencia GNU GPL 3. Se utiliza para desplegar gráficos representados mediante grafos, complejos gráficos de visualización de datos utilizados en análisis de redes sociales o jerarquía de datos. Además, se utiliza como herramienta de visualización en proyectos de pequeño, mediano o gran alcance.

Soporta la representación de grafos dirigidos, no dirigidos, mixtos e hipergrafos. Uno de los aspectos importantes cubiertos por Gephi es la interacción en tiempo real, que permite modificar propiedades de los nodos y aristas al mismo tiempo que se modifica la representación del grafo, y se las ofrece al usuario sin largas esperas. Asimismo, permite realizar agrupaciones, filtrado, manipulación, navegación y proveer acceso a los datos.[15]

Los desarrolladores de Gephi han desarrollado Gephi Toolkit, la cual es una librería que implementa muchas de las funcionalidades del software. Brinda una API para la visualización de grafos que contiene algoritmos de layout (visualización de grafos) como son: Kamada-Kawai,[16] Fruchterman-Reingold,[17] Force Atlas[15] y Force Atlas 2.[18]

Force Atlas 2 fue el algoritmo seleccionado para representar visualmente el conjunto de comunidades. Al ser un algoritmo de vector de fuerza, se destaca por su sencillez y por la facilidad de lectura de las redes a las que ayuda a visualizar. Presenta un modelo que optimiza la velocidad frente a la precisión, aproximación que permite una convergencia rápida en su ejecución. Esta característica permite representar redes de gran tamaño en cortos tiempos.

RESULTADOS Y DISCUSIÓN

Con el objetivo de probar la validez del método para la detección de comunidades a partir de RC en grafos RDF se desarrolla un caso de estudio utilizando un prototipo funcional que implementa el método propuesto en el entorno de desarrollo integrado IntelliJ IDEA 14.0.2 y la aplicación Gephi. Las capacidades de cómputo con las que se desarrolla el caso de estudio son: computadora personal con 4 GB de memoria RAM y un procesador Intel Core i3 380 a 2.33MHz. A continuación se detallan los resultados obtenidos en cada una de las fases del método propuesto.

Se aplicó el método propuesto utilizando de entrada un grafo RDF con 51329 tripletas. Este grafo modela los metadatos bibliográficos procedentes de revistas cubanas de las áreas de ciencias médicas y ciencias informáticas. Se procesan y transforman los metadatos en una red de coautoría ponderada y dirigida con un total de 5 203 vértices y 24 498 aristas. Se detectaron 826 comunidades, que se identificaron con una clase de modularidad y un color determinado. Agrupó visualmente los elementos de la red en comunidades, lo que evitó el solapamiento de los vértices y mejoró en gran medida la identificación visual de las comunidades. Los algoritmos utilizados caracterizan las relaciones de coautoría y en muchos casos se benefician de la tipología y ponderación de la red. Los criterios de medidas que se plantean en el método permiten identificar y cuantificar a las comunidades y autores dentro de la red y realizar análisis y comparaciones de estos.

El prototipo funcional desarrollado implementa cada una de las etapas del método propuesto. La ejecución del método siguiendo las configuraciones establecidas en el caso de estudio tiene una duración total de 179 407 milisegundos, donde la fase 3 es la que ocupa un 85 % del tiempo total (Fig. 3). Dicho tiempo de ejecución puede variar según las características de la red y las configuraciones establecidas, pero siempre va a ser la fase que ocupa mayor tiempo de ejecución. En el prototipo funcional solo se ejecutan las tres fases una única vez por cada grafo RDF. Si este no es modificado solo se accede a los archivos GEXF que fueron generados con anterioridad. En la figura 4 se muestra una vista del prototipo funcional implementado.

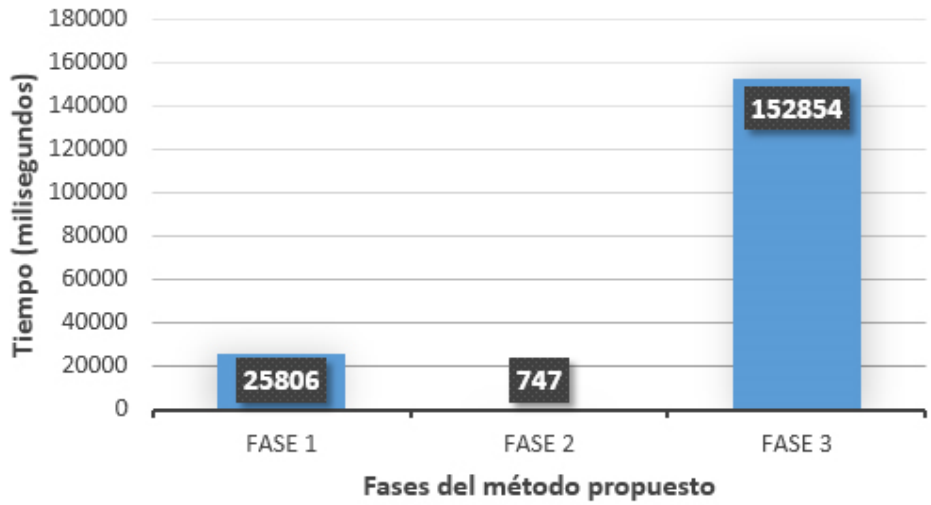


Fig. 3. Tiempo de ejecución de las fases del método propuesto.

Fig. 4. Vista del prototipo funcional implementado.

CONCLUSIONES Y TRABAJOS FUTUROS

En este artículo se presenta un método para la detección y visualización de comunidades existentes en grafos RDF a partir de redes de coautoría. La revisión de las principales contribuciones encontradas en la literatura evidencia la necesidad de contar con métodos para la detección de comunidades en este tipo de grafos,

teniendo en cuenta el avance de las tecnologías de la web semántica y lo datos enlazados en la publicación de metadatos bibliográficos. Con la implementación del método propuesto en un caso de estudio se ha demostrado su viabilidad para detectar y visualizar las comunidades a partir de redes de coautoría existentes en grafos RDF.

Como trabajo futuro se propone refinar el método propuesto variando sus parámetros de configuración y calculando métricas que evidencien la eficiencia y eficacia de este en la detección de las comunidades. Adicionalmente, se recomienda diseñar un algoritmo para etiquetar las comunidades detectadas atendiendo a las palabras clave presentes en los artículos científicos utilizados como muestras.

REFERENCIAS BIBLIOGRÁFICAS

1. Beaver D, Rosen R. Studies in scientific collaboration. Scientometrics [Internet]. 1978 [citado 30 de marzo de 2015];1(1):65-84. Disponible en: http://link.springer.com/article/10.1007/BF02016840

2. Maltras B, Vega J, Quintanilla MA. Measuring Multinational Cooperation in Science & Technology : Different Methods Applied to the European Framework Programs. EE.UU.: Proceedings of the Fifth Biennial International Conference of the International Society for Scientometrics and Infometrics [Internet]. 1995 [citado 30 de marzo de 2015]. p. 303-12. Disponible en: http://cat.inist.fr/?aModele=afficheN&cpsidt=3146004

3. Melin G, Persson O. Studying research collaboration using co-authorships. Scientometrics [Internet]. 1996 [citado 30 de marzo de 2015];36(3):363-77. Disponible en: http://link.springer.com/article/10.1007/BF02129600

4. Miquel JF, Okubo Y, Narvaez N, Frigoletto L. Les scientifiques sont-ils ouverts à la coopération internationale. La Recherche. 1989;20(206):116-8.

5. Stokes TD, Hartley JA. Coauthorship, social structure and influence within specialties. Soc Stud Sci. 1989;19(1):101-25.

6. Martin T, Ball B, Karrer B, Newman MEJ. Coauthorship and citation in scientific publishing. Phys Rev E [Internet]. 2013 [citado 30 de marzo de 2015];88(1). Disponible en: http://arxiv.org/abs/1304.0473

7. Newman MEJ. Community detection and graph partitioning. EPL Europhys Lett [Internet]. 2013 [citado 26 de marzo de 2015];103(2):28003. Disponible en: http://arxiv.org/abs/1305.4974

8. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D-U. Complex networks: Structure and Dynamics. Phys Rep [Internet]. 2006 [citado 20 de julio de 2015];424(4):175-308. Disponible en: http://www.sciencedirect.com/science/article/pii/S037015730500462X

9. Fortunato S. Community detection in graphs. Phys Rep. 2010;486(3):75-174.

10. Liu X, Bollen J, Nelson ML, Van de Sompel H. Co-authorship networks in the digital library research community. Inf Process Manag. 2005;41(6):1462-80.

11. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Phys Rev E [Internet]. 2004 [citado 30 de marzo de 2015];69(2). Disponible en: http://www.bibsonomy.org/bibtex/1b9145040e35ccb4d2a0ce18105e64ff4/kibanov

12. Aldecoa R, Marín I. Deciphering Network Community Structure by Surprise. PLoS ONE [Internet]. 2011 [citado 30 de marzo de 2015];6(9):e24195. Disponible en: http://dx.doi.org/10.1371/journal.pone.0024195

13. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech Theory Exp [Internet]. 2008 [citado 30 de marzo de 2015];(10):P10008. Disponible en: http://arxiv.org/abs/0803.0476

14. Medrano JF, Berrocal JLA, Figuerola CG. Visualización de Grafos Web. 2011 [citado 30 de marzo de 2015]. Disponible en: http://www.researchgate.net/profile/Jose_Luis_Berrocal/publication/247936803_Visualizacion_de_Grafos_Web_-_Web_Graph_Visualization/links/5474a3cb0cf245eb436deaeb.pdf

15. Bastian M, Heymann S, Jacomy M. Gephi: An Open Source Software for Exploring and Manipulating Networks. 2009 [citado 25 de marzo de 2015]. Disponible en: https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154

16. Kamada T, Kawai S. An algorithm for drawing general undirected graphs. Inf Process Lett [Internet]. 1989 [citado 30 de marzo de 2015];31(1):7-15. Disponible en: http://www.sciencedirect.com/science/article/pii/0020019089901026

17. Fruchterman TMJ, Reingold EM. Graph Drawing by Force-directed Placement. Softw Pr Exper [Internet]. 1991 [citado 30 de marzo de 2015];21(11):1129-64. Disponible en: http://dx.doi.org/10.1002/spe.4380211102

18. Jacomy M, Venturini T, Heymann S, Bastian M. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. PLoS ONE [Internet]. 2014 [citado 30 de marzo de 2015];9(6):e98679. Disponible en: http://dx.doi.org/10.1371/journal.pone.0098679

Recibido: 4 de julio de 2015.
Aprobado: 2 de septiembre de 2015.

Ernesto Ortiz Muñoz. Universidad de las Ciencias Informáticas. La Habana, Cuba. Correo electrónico: ernesto@uci.cu

* http://gexf.net/format/
** http://gephi.github.io/