UNIVERSIDAD DE LAS CIENCIAS INFORMÁTICAS

FACULTAD 3

Grupo de Investigación de Web Semántica

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

Trabajo de Diploma para optar por el título de
Ingeniero en Ciencias Informáticas

Autor:
Ernesto Ortiz Muñoz

Tutor:
Ing. Yusniel Hidalgo Delgado

La Habana, junio de 2015
"Año 57 de la Revolución"

DECLARACIÓN DE AUTORÍA

Declaro ser el autor de la presente tesis y reconozco a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmo la presente a los ____ días del mes de _____ del año _____.

_____	_____
Ernesto Ortiz Muñoz	Ing. Yusniel Hidalgo Delgado
Autor	Tutor

DATOS DE CONTACTO

Síntesis del Tutor

El Ingeniero Yusniel Hidalgo Delgado se graduó con Título de Oro en la Universidad de Ciencias Informáticas en el año 2010. En su primer año de adiestramiento desempeñó diversos roles dentro del proyecto de desarrollo del ERP Cubano. Actualmente se desempeña como profesor asistente del departamento docente de técnicas de programación de la Facultad 3. Es coordinador del grupo de investigación de Web Semántica de la UCI. Es miembro de la Asociación Cubana de Reconocimiento de Patrones, de la Sociedad Cubana de Matemática y Computación y de la *International Association for Pattern Recognition*.

DEDICATORIA

A Lily y Ortiz, mis padres y mejores maestros.
A Pipa, el abuelo más jovial del mundo.
A Yake, mi hermana y amiga por siempre.
A toda mi familia y amigos.
Al grupo de Web Semántica.
A la Revolución Cubana.

AGRADECIMIENTOS

Gracias a Lily y Ortiz, por la formación, el amor y consejos que siempre me han dado en todos estos años.

Gracias a Pipa, por haber sido el abuelo de las travesuras y de los chistes inolvidables.

Gracias a Yake, por siempre aconsejarme y acompañarme en todos estos años.

Gracias a Katy, por su amor, por siempre estar ahí, en las buenas y en las malas.

Gracias a Lidia, mi segunda madre, por aceptarme como su hijo varón.

Gracias a mi familia toda, por todo el apoyo y el cariño.

Gracias a mis amigos camagüeyanos, por enseñarme el verdadero valor de la amistad, en especial a Alexei, Anier, Kiki, Molina, Yoan y Yuliano.

Gracias a mis amigos de la Universidad y a mis compañeros de aula, por sus lecciones y comprensión, en especial a, Anchel, Aniel, Baby, Isis, Lilian y Michel.

Gracias a mi tutor, por sus conocimientos y por adentrarme en el camino de la investigación.

Gracias a los profesores que contribuyeron a mi formación a lo largo de estos cinco años, especialmente a los profes: Ana Maris, Eliober, Eylis, Leandro, Yalice, Yoan.

Gracias a la FEU y a todos los que implica, por formarme integralmente.

Gracias al Grupo de Web Semántica, por sus recomendaciones.

A todos, muchas gracias.

RESUMEN

La detección de comunidades se refiere al problema de identificar comunidades o particiones de vértices en una red que comparten propiedades comunes. Las redes de coautoría en metadatos bibliográficos se consideran redes complejas, donde los vértices de la red son los autores y los enlaces entre los vértices establecen la relación de coautoría en una o varias publicaciones. En los últimos años se han desarrollado proyectos de investigación 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. Sin embargo, aún resultan insuficientes los resultados obtenidos en la detección de comunidades en grafos RDF a partir de las redes de coautoría. En este trabajo de diploma 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 bibliométricos en el área.

Palabras claves: colaboración científica, detección de comunidades, RDF, redes de coautoría

ABSTRACT

COMMUNITIES DETECTION USING COAUTHORSHIP NETWORKS FROM RDF GRAPHS

The community's detection refers to the problem to identify communities or partitions of nodes in a network that shares common properties. The Co-authorship networks in bibliographic metadata are considered complex networks, where the nodes of the network are authors and the relation between nodes establish the co-author relationship in one or more publications. In recent years there have been developed research projects with aim of publishing bibliographic metadata following the Linked Data principles. These research projects produce several RDF graphs that contain the authors and co-authorship relations established between them. However, still insufficient results obtained in communities detection from RDF graphs using co-authorship networks. In this diploma work we propose a method for detecting and visualizing communities from RDF graphs considering co-authorship relations as an indicator to measure scientific collaboration. With the implementation of the method proposed we provide an analysis tool for specialists in information science that helps in the process of decision making and implementation of bibliometrics studies in the area.

Keywords: *co-authorship networks, communities detection, RDF, scientific collaboration*

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

En los últimos años ha cobrado importancia la publicación de metadatos bibliográficos siguiendo los principios de los datos enlazados. Estos metadatos son publicados utilizando alguna de las serializaciones del modelo de datos *Resource Description Framework* (RDF), un estándar del *World Wide Web Consortium* (W3C[1]) para la descripción de recursos en la Web. Instituciones como la IEEE utilizan el modelo RDF para publicar los metadatos bibliográficos de las publicaciones que indexan. Los grafos RDF de estas instituciones constituyen fuentes de datos de gran valor para realizar análisis y detección de comunidades científicas, cuantificar las relaciones de colaboración y realizar análisis métricos de las mismas.

En estudios realizados en (Garvey y Griffith 1964), (Frame y Carpenter 1979) y (Miquel et al. 1989) se han ido presentando diferentes formas de medir la colaboración científica, los cuales han evolucionado a lo largo de los años. Prevaleciendo según la investigación de (Stokes y Hartley 1989) la coautoría como un indicador disponible para analizar la colaboración científica. A partir del cual los resultados que con ella se obtienen, surgen instrumentos estratégicos tanto para los investigadores que participan en esos trabajos como para las instituciones involucradas en su elaboración.

Por otro lado, se han propuesto algoritmos, técnicas y herramientas para la detección de comunidades en redes complejas, influenciados principalmente por las investigaciones de Newman en la década de los 90. En el marco de la detección de comunidades se emplean diferentes algoritmos basados en modularidad, información, teoría de códigos y algoritmos basados en camarilla (Saha et al. 2015). Dichos algoritmos detectan las comunidades a partir de datos presentes en redes o grafos, no existiendo evidencia científica en los últimos años de su aplicación a partir de las redes presentes en grafos RDF.

En el grupo de investigación de Web Semántica de la Universidad de las Ciencias Informáticas se desarrolla el proyecto Biblioteca Digital Semántica (BDS). El objetivo del proyecto es publicar y consumir metadatos bibliográficos siguiendo los principios de los datos enlazados. Una etapa del proyecto es la fase de análisis de los datos enlazados que se han sido publicados. En esta fase, se hace necesario contar con una herramienta informática que permita detectar y visualizar comunidades entre los autores presentes en las redes de coautoría (RC) que se pueden obtener a partir de grafos RDF. Las RC constituyen una herramienta de análisis de impacto en estudios bibliométricos y para los propios investigadores. Las RC permiten además identificar y cuantificar las

[1] Comunidad internacional donde las organizaciones miembros, personal a tiempo completo y el público en general trabajan conjuntamente para desarrollar estándares Web. Fuente: http://www.w3c.org

relaciones de colaboración existentes entre los autores de diversas instituciones y áreas del conocimiento. Actualmente, no es posible realizar este tipo de análisis en los grafos RDF generados en el proyecto de investigación, lo cual influye negativamente en el estudio y en el análisis de los resultados del mismo.

De acuerdo con la situación descrita anteriormente, se plantea el siguiente **problema a resolver:** ¿Cómo ***detectar comunidades a partir de redes de coautoría en grafos RDF*** de manera que se logre ***identificar y cuantificar las relaciones de colaboración*** que se establecen entre los autores presentes en los metadatos bibliográficos?

El **objeto de estudio** donde se enmarca la investigación está constituido por los datos enlazados y como **campo de acción** las redes de coautoría en grafos RDF.

Para resolver el problema se identifica el siguiente **objetivo general:**

Detectar comunidades en las redes de coautoría existentes en los grafos RDF aplicando teoría de redes complejas que permitan identificar y cuantificar las relaciones de colaboración que se establecen entre los autores presentes en los metadatos bibliográficos.

A partir de lo planteado anteriormente se desglosan los siguientes **objetivos específicos:**

1. Elaborar el marco teórico y el estado del arte del objeto de estudio de la investigación mediante el análisis bibliográfico documental para identificar tendencias y adoptar posiciones al respecto.
2. Diseñar un método para la detección de comunidades a partir de las redes de coautoría en grafos RDF aplicando teoría de redes complejas.
3. Implementar el método para la detección de comunidades a partir de las redes de coautoría en grafos RDF.
4. Validar los resultados del método implementado mediante la utilización de un caso de estudio.

Se obtienen como **posibles resultados** en artefactos a entregar: método para la detección de comunidades a partir de redes de coautoría en grafos RDF y prototipo funcional que implemente el método propuesto.

Se definen como **tareas a cumplir**:

1. Estudio de las principales aproximaciones existentes para la detección de comunidades a partir de redes de coautoría en grafos RDF.

2. Análisis de las fuentes de información que servirán de entrada al método.
3. Diseño del método para la detección de comunidades a partir de redes de coautoría en grafos RDF.
4. Implementación del método para la detección de comunidades a partir de redes de coautoría en grafos RDF.
5. Análisis de los resultados obtenidos por el método propuesto.
6. Validación de los resultados obtenidos por el método propuesto.

Se plantea como **idea a defender**:

Con la obtención de un método para la detección de comunidades a partir de redes de coautoría en grafos RDF, se identificarán y cuantificarán las relaciones de colaboración que se establecen entre los autores presentes en los metadatos bibliográficos. Para el desarrollo de la investigación se tiene como **población: datos bibliográficos de revistas científicas en forma de grafos RDF.** Por otra parte, la **muestra** seleccionada es: **datos bibliográficos de ocho revistas científicas cubanas en forma de grafos RDF**.

Durante la investigación se han empleado un conjunto de **métodos científicos** como procedimientos lógicos, que se han seguido para la obtención y el procesamiento de la información.

Métodos teóricos

El método **Histórico-Lógico** ha permitido analizar la evolución, de forma cronológica, de los elementos relacionados con las RC, así como la evolución de tecnologías usadas en la detección de dichas redes en grafos RDF, teniendo en cuenta el desarrollo de las aproximaciones realizadas.

El método **Analítico-Sintético** ha permitido realizar un análisis sobre la teoría y las tendencias de los componentes relacionados con la colaboración científica y las RC, incluyendo también su análisis en grafos RDF, de manera que se hayan podido estudiar a profundidad cada uno de ellos por separado, así como las técnicas o tecnologías involucradas en el proceso de detección de dichas RC. Ha permitido, además, resumir y centrar la atención en los conceptos e ideas principales de cada uno de los componentes analizados previamente.

El método **Inductivo-Deductivo** ha permitido realizar una generalización de los procesos involucrados en la detección de RC teniendo como base los elementos comunes de las aproximaciones previamente analizadas. A partir de dichos conocimientos, se ha podido generar una propuesta de solución con elementos más generales a través del razonamiento lógico.

El método de **Modelación** ha permitido la descripción de la propuesta de solución, basándose en la aplicación de restricciones y supuestos que posibilitaron la conceptualización y modelado del método para la detección de comunidades a partir de RC en grafos RDF.

Métodos empíricos

El método de **caso de estudio** se empleó con el objetivo de probar la validez de la propuesta de solución permitiendo obtener resultados medibles. Es decir, comprobar que se cumplieran cada una de las restricciones del método propuesto a partir de su implementación y ejecución en un prototipo funcional.

El método de la **medición** ha permitido la aplicación de métricas para comparar los resultados obtenidos con la propuesta de solución.

La investigación está estructurada en tres capítulos.

Capítulo I: Se definen los conceptos necesarios para el desarrollo de la propuesta de solución, que pertenecen al ámbito de la Web Semántica. Se analiza un conjunto de investigaciones relacionadas con la detección de comunidades a partir de RC en grafos RDF. Se realiza un estudio de la literatura para identificar los elementos que formarán parte de la propuesta de solución y su impacto social.

Capítulo II: Se define el método para la detección de comunidades a partir de RC en grafos RDF. Se exponen las características relevantes en cuanto a su diseño e implementación en un prototipo funcional.

Capítulo III: Se describe el proceso de validación de la propuesta de solución. Se expone un caso de estudio para validar el método para la detección de comunidades a partir de RC en grafos RDF, se hace un análisis de la ejecución de las fases del método propuesto. Por último se ilustran ejemplos y resultados finales.

CAPÍTULO 1. FUNDAMENTACIÓN TEÓRICA

1.1. Introducción

En este capítulo se definen los principales conceptos, relacionados con la Web Semántica, la colaboración científica, las comunidades y las redes. Se analiza un conjunto de investigaciones afines con las RC y la detección de comunidades. Se realiza un estudio de la literatura para identificar los elementos que formarán parte de la propuesta de solución y su impacto social.

1.2. Análisis bibliométrico y documental

En esta sección se realiza un análisis bibliométrico con el objetivo de mostrar la novedad de la revisión bibliográfica realizada, basándose en las fechas de las publicaciones consultadas. Las bases de datos utilizadas son: *IEEE, Google Scholar y Scielo*. Además, se relacionan los tipos de fuentes bibliográficas más citadas. Las fuentes bibliográficas son: artículos de revistas, libros, tesis (específicamente de maestrías o doctorados), artículos en congresos y sitios web. El análisis realizado se muestra en la siguiente tabla.

Tabla I: *Análisis bibliométrico y documental*

Tipo de fuente bibliográfica	Cantidad consultada	Cantidad publicada en los últimos cinco años (2010-2015)
Artículos de revista	41	25
Libro	8	4
Tesis (maestrías o doctorados)	4	3
Artículos en congreso	10	6
Sitio web	6	6
Total:	**69**	**44**

De la tabla anterior se obtiene el resultado mostrado en la **Figura I**, donde se muestra que el 64% de la literatura consultada pertenece a los últimos cinco años, por lo que se evidencia la actualidad de la bibliografía consultada.

Figura I: *Actualidad de la bibliografía consultada*

1.3. Marco teórico. Conceptos y definiciones

En la transición de la Web actual a la Web Semántica se debe cumplir con los **principios de los datos enlazados**, término definido en el 2006 por Timothy Berners-Lee. Los **datos enlazados** se refieren a un conjunto de principios o buenas prácticas para la publicación y enlazado de datos estructurados a través de *Uniform Resource Identifier* (URIs) dereferenciables en la Web (Berners-Lee 2006). Dichos principios se enuncian a continuación:

1. Utilizar el identificador de recurso uniforme, del acrónimo en inglés *Uniform Resource Identifier* (URI), para identificar cada recurso[2] publicado en la Web. (Identificar el recurso).

2. Publicar los datos en una URI basada en *Hypertext Transfer Protocol* (HTTP) con el que puedan ser fácilmente localizados y consultados. (Publicar el recurso)

3. Proporcionar información útil, detallada o extra acerca del recurso cuando se acceda a esta URI basada en HTTP. (Describir el recurso).

4. Incluir enlaces a otras URIs relacionadas con los datos contenidos en el recurso, de manera que se potencie el descubrimiento de información en la Web. (Enlazar el recurso con otros recursos similares).

La aplicación de los principios de los datos enlazados se ha extendido por todo el mundo a través de la utilización en sistemas de todo tipo. Sin embargo, su empleo requiere compatibilidad y estandarización para la representación de los datos. Es por ello que se utilizan estándares establecidos por el W3C como el Marco de Descripción del Recurso (RDF, por sus siglas en inglés). **RDF** es un modelo de datos basado en grafos dirigidos para la publicación y enlazado de datos

[2] El término de la arquitectura web "recurso" se refiere al contenido de interés en la Web que se publica mediante URIs HTTP

estructurados en la Web. Con RDF se hace uso de las ontologías para su descripción formal**.** Su sintaxis está basada en tripletas del tipo sujeto-predicado-objeto.

En 2007 Studer refiere que: "una ontología es una especificación formal y explícita de una conceptualización compartida" (Studer y Benjamins 2007). Esta última definición ha sido a menudo citada en la literatura y la usada por la comunidad de desarrolladores de ontologías. La evolución del término ha estado marcada por aportes al concepto dado por Studer según el área evidente en cada investigación. En este trabajo se adopta la definición dada por Studer, que se refiere a: "formal" como la necesidad de que sea comprensible por las computadoras, "explícita" a su descripción en un determinado lenguaje, "conceptualización" a la forma de entender y describir un dominio y "compartida" a ser consensuada por un grupo, o compartida por varias partes.

Las ontologías se pueden clasificar atendiendo a sus diferencias en cuanto a dependencias y relación con tareas específicas. Pueden ser de **Nivel Superior**: se describen conceptos de forma general; de **dominio:** se utilizan en un dominio específico, definiendo además, sus relaciones; de t**areas**: se propone un vocabulario sistemático de términos utilizados para resolver problemas asociados a tareas que pueden ser parte o no de un mismo dominio y de **aplicación**: dependen de la aplicación en particular, pueden extenderse en el vocabulario del dominio y de las ontologías de tareas. En esta investigación se emplean las ontologías de dominio enmarcadas en el ámbito de los metadatos bibliográficos.

Resulta necesario modelar RC en grafos RDF. Las **RC** son un grafo de colaboración donde los vértices son científicos. Se considera que dos científicos están conectados si han aparecido como autores en la misma publicación (Miquel et al. 1989).

Un grafo, o grafo no dirigido, $G = (V, E)$ se define como un conjunto V finito y no vacío de vértices y un multiconjunto E de aristas, donde cada arista $(v_i, v_j) \in E;\ v_i, v_j \in V$ es un par no ordenado de vértices. Opcionalmente una arista puede tener un valor que la identifique y una lista de atributos (Tucker 2004).

Por otra parte, las **redes complejas** son conjuntos de muchos vértices conectados que interactúan de alguna forma. A los nodos de una red también se les llama *vértices* o *elementos* y se representan por los símbolos $v_1, v_2, \ldots, v_n$, donde n es el número total de nodos en la red. Si un nodo v_i está conectado con otro nodo v_j, esta conexión se representa por una pareja ordenada (v_i, v_j). La definición de una red es también llamada *grafo* por los matemáticos (Newman 2012).

En las RC es posible identificar comunidades. Una **comunidad** es un grupo o conjunto de individuos, seres humanos, o de animales (o de cualquier otro tipo de vida) que comparten elementos en común, tales como un idioma, costumbres, valores, tareas, visión del mundo, edad, ubicación geográfica (un barrio por ejemplo), estatus social y roles. Por lo general, en una comunidad se crea una identidad común, mediante la diferenciación de otros grupos o comunidades (generalmente por signos o acciones), que es compartida y elaborada entre sus integrantes y socializada. Generalmente, una comunidad se une bajo la necesidad o meta de un objetivo en común; si bien esto no es imprescindible, basta una identidad común para conformar una comunidad sin la necesidad de un objetivo específico. En esta investigación se realiza un estudio de las comunidades científicas, compuestas por investigadores.

1.4. Estado del arte

1.4.1. Colaboración científica

En la etapa del Renacimiento la ciencia se consideraba como actividad solitaria y no fue hasta el desarrollo de la Revolución Industrial que comenzó a evidenciarse como una actividad de grupo. 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 (Beaver y Rosen 1978). La colaboración es considerada un desarrollo complejo, una forma de intercambiar información, para trabajar juntos, para utilizar los recursos de forma racional y perpetuar comunidades de científicos y tecnólogos. Lo anterior evidencia que la colaboración más que una necesidad se ha convertido en una elección.

Beaver y Rosen definen que la colaboración e interacción personal está asociada al intercambio de conocimientos a través de actividades que responden a las habilidades como hablar, escuchar o redactar conjuntamente. Se trata de una combinación de intercambios de conocimientos tácitos y explícitos. Las colaboraciones personales entre científicos pueden dividirse en cuatro tipos, reflejo de los diferentes perfiles, conocimientos previos y funciones de cada uno de los colaboradores:

Empleador/empleado: es la forma más débil de colaboración, en la que el científico delega en el empleado tareas rutinarias que este último conoce. Ejemplo de este caso es la realización de experimentos, la elaboración de programas o subrutinas de programación, la construcción de circuitos, entre otros. Los técnicos y los ayudantes de laboratorio no suelen ser considerados colaboradores, pero algunos de ellos pueden pasar a formar parte de la siguiente categoría.

Profesor/ayudante: este tipo de colaboración es similar a la anterior, se muestra una asimetría entre el conocimiento y el nivel académico, pero sus objetivos son diferentes. Los ayudantes adquieren

habilidades técnicas que les permiten trabajar por sí mismos. El diseño de experimentos y la interpretación de resultados solo se aprenden a través del trabajo en proyectos con investigadores especializados.

Pares similares: es frecuente que científicos con conocimientos, intereses y estatus similares, encuentren valioso el trabajo en equipo. Sin embargo, "similar" no quiere decir idéntico. Aunque pertenezcan a la misma disciplina científica, poseerán habilidades y conocimientos distintos.

Pares diferentes: la investigación interdisciplinar es la que se produce entre investigadores con objetivos similares, pero con conocimientos y habilidades diferentes. En las ciencias cognitivas, esta colaboración es la que se da entre un psicólogo y un informático. El primero tiene formación teórica y experimental, incluyendo la forma de realizar experimentos, mientras que el segundo sabe desarrollar programas que simulen ciertos aspectos del razonamiento humano.

Aunque en el nivel más bajo se encuentran los individuos, la colaboración científica no se produce solo en el nivel personal y entre científicos. La asociación directa entre dos o más científicos es la unidad mínima fundamental de colaboración. También se puede hablar de colaboración en otros niveles: entre grupos de investigación de un mismo departamento, entre departamentos, entre instituciones, entre sectores y entre regiones o países.

Otros autores, en cambio, mantienen que la literatura es solo una parte de un sistema que posee muchas formas de intercambio de información o colaboración. Garvey y Griffith, en su estudio de intercambio de información entre psicólogos encontraron, que los principales productores y consumidores de este tipo de información no esperan a verla publicada en revistas, sino que hacen uso de su red de contactos, la mayor parte de ellos informales, pero muy eficientes, para su difusión y utilización (Garvey y Griffith 1964).

Los trabajos de Frame y Carpenter son pioneros en el uso de la publicación signada por dos o más autores como indicador de colaboración. A este respecto señalan que la colaboración científica puede tener múltiples formas: participación de fuentes de datos, intercambio de ideas, estancia en centros en el extranjero o intercambio de artículos, son algunos ejemplos de actividad colaboradora. Sin embargo, la forma más obvia de colaboración y la más fácilmente mensurable, es la colaboración en forma de artículos científicos (Frame y Carpenter 1979).

Miquel y sus colaboradores 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 o ministerios) involucradas en su elaboración (Miquel et al. 1989).

1.4.2. Redes de colaboración y coautoría

En el siglo XXI con el desarrollo de las Tecnologías de la Información y las Comunicaciones se han producido avances en la perspectiva científica de los estudios estructurales, se ha convertido a la teoría de redes en un área de investigación interdisciplinar emergente. Las investigaciones están orientadas al desarrollo de teorías y técnicas para aumentar el conocimiento existente tanto de las redes biológicas o tecnológicas, como de las académicas y de los sistemas complejos (He, Ding y Ni 2011).

Estos estudios tienen su origen en disciplinas como las matemáticas discretas y la teoría de grafos, la sociología matemática, la psicología de grupos o la biología, pero también en la bibliometría, la infometría y la cibermetría, y más recientemente en la física. Los estudios de redes comprenden múltiples enfoques, que se centran en la descripción y el análisis de sus propiedades, la investigación de su modelado y sus dinámicas y el establecimiento de nuevas técnicas para su visualización. Su ámbito de análisis engloba problemas derivados de la existencia de redes sociales, biológicas, de información o tecnológicas, como los sistemas de comunicación (Internet, redes telefónicas), las infraestructuras de transporte (carreteras, líneas ferroviarias o aéreas), los sistemas biológicos (interacción de proteínas, ácido desoxirribonucleico (ADN), evolución de epidemias) y un amplio abanico de estructuras de interacción social, entre las que se encuentran las redes de colaboración entre científicos (Donetti y Muñoz 2004).

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 (Melin y Persson 1996). La misma puede manifestarse en proyectos científicos, publicaciones, contratos de investigación, contactos informales, patentes, formación a través de cursos, seminarios, tesis de pregrado, maestrías, doctorados, intercambios de investigadores y la participación en comités científicos (Vidal y Villarroel 1995).

Es posible capturar la estructura y evolución del desarrollo científico desde dos modelos distintos: el modelo descriptivo, destinado a definir los principales rasgos de un conjunto de datos casi siempre estáticos, y el modelo de proceso, que trata de determinar los mecanismos y dinámicas temporales mediante el empleo de redes del mundo real, como las redes de colaboración basadas en coautorías (Börner, Maru y Goldstone 2004).

Desde esta perspectiva, las redes de filiación de científicos, en las que los enlaces se obtienen a partir de la coautoría en una o más publicaciones, son más reales desde el punto de vista de red social que muchas redes de filiación institucional, puesto que este tipo de redes refleja de forma genuina la interacción profesional entre científicos y se postula como las redes sociales más grandes hasta ahora analizadas (Martin et al. 2013). Aunque similares a las redes de citación y co-citación, las redes de colaboración implican lazos sociales más fuertes que los existentes en las primeras. Las citas ocurren sin que los autores se conozcan entre sí y pueden prolongarse a lo largo del tiempo, las RC, en cambio, implican una relación colegiada y temporal que tiene lugar en el ámbito de los estudios de análisis de redes sociales (Liu et al. 2005).

La aplicación de la teoría de redes al estudio de la coautoría parte de la idea de entender a la ciencia como un sistema auto-organizado en el que la selección de colaboradores y la localización de la investigación se da a partir de las elecciones de los propios investigadores, más que a través de incentivos institucionales, nacionales o de cualquier otra índole, y sin que exista autoridad alguna en el seno de la organización científica mundial encargada de arbitrar o resolver posibles disputas que puedan surgir durante los intercambios que en ella se realizan.

El estudio de la colaboración científica ayuda a establecer grupos y redes de trabajo, pudiendo ser analizada y evaluada mediante el empleo de técnicas bibliométricas y representada mediante lo que algunos autores denominan RC o mapas bibliométricos. A lo largo de la investigación se utiliza el término “coautoría” para hacer referencia a la firma conjunta de un trabajo científico por dos o más autores.

Las RC hacen referencia a una clase importante de las redes sociales y se han utilizado ampliamente para determinar la estructura de colaboraciones científicas y la situación de los investigadores individuales. Ejemplos de RC se muestran en el Número del Proyecto Erdos, el menor número de enlaces de coautoría entre cualquier individuo matemático y el matemático húngaro Erdos (De Castro y Grossman 1999).

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 (Stokes y Hartley 1989). Newman abunda en el concepto de RC indicando que se está en presencia de redes que además de describir la sociedad académica, muestran la estructura del conocimiento (Newman y Girvan 2004).

Medir la colaboración científica implica incluir varios factores. El estudio de los intercambios informales entre científicos requiere observación directa. Este contacto puede cristalizar en forma de artículos científicos que son en los que los especialistas de la colaboración ponen su atención. Sin embargo, además de la coautoría, también se podrían tener en cuenta aspectos como la movilidad de los investigadores, el desarrollo de proyectos científicos o la creación de centros de investigación compartidos (Maltras, Vega y Quintanilla 1995).

El análisis de coautoría, también ha sido aplicado a diversas conferencias de ACM (acrónimo de *Association for Computing Machinery)*, Recuperación de Información (SIGIR) y Gestión de Datos (SIGMOD) (Nascimento, Sander y Pound 2003), así como las matemáticas y la neurociencia y el campo de análisis de redes sociales (Ball y Newman 2013), (Azizifard 2014). Las RC internacionales se han estudiado en la JASIST (acrónimo de *Journal of American Society for Information Science & Technology*) y en *Science Citation Index* (Wagner, Leydesdorff y Bornmann 2014).

Los elementos anteriores permiten aseverar que el trabajo con redes de colaboración y coautoría es un área de investigación en desarrollo. Posibilita el análisis de la colaboración e intercambio científico de los investigadores. Definiéndose así que la coautoría constituye un indicador de impacto a tener en cuenta para la modelación y análisis de redes de colaboración científica.

1.4.3. Modelado de RC

1.4.3.1. Binaria y no dirigida

Modelo sencillo y ampliamente utilizado de las RC. Se basa en un grafo binario no dirigido, donde cada vértice representa a un autor.

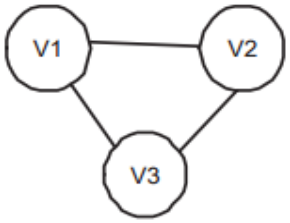

Figura II: *Grafo G. Red de coautoría binaria no dirigida. Fuente: Liu, Xiaoming (2005). Co-Authorship Networks in the Digital Library Research Community.*

En la Tabla II se representan dos artículos y los respectivos autores.

Tabla II: *Ejemplo de relación entre artículos y autores*

Artículos	Autores
Artículo 1	V1, V2, V3
Artículo 2	V1, V2

Por cada coautor (vértices) se crea una arista que los une si pertenecen al mismo artículo. El grafo resultante se denota como una unidad ponderada no dirigida $G = grafo\ (V, E)$, donde el conjunto de n autores se denota $V = \{v_1, \ldots, v_n\}$ y E representa las aristas entre autores (Liu et al. 2005).

1.4.3.2. Binaria y dirigida

Con el fin de medir el prestigio de un autor, hay que distinguir el "aval" otorgado de respaldo recibido por los autores. En el análisis de redes sociales, el concepto de prestigio se define por las relaciones direccionales. Con el fin de convertir una red de coautoría a un grafo dirigido, se realizan las siguientes suposiciones:

- Cualquier red no dirigida se puede representar como una red dirigida con vinculación simétrica, es decir, todas las aristas en la red dirigida G se sustituyen por dos aristas dirigidas simétricas.
- En la dirección resultante, las aristas simétricas representan el respaldo mutuo de los autores.
- El peso de cada arista es un valor binario, lo que indica la presencia o ausencia de dos aristas simétricas.

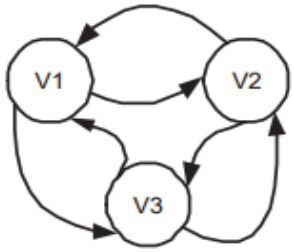

Figura III: *Grafo G. Red de coautoría binaria dirigida. Fuente: Liu, Xiaoming (2005). Co-Authorship Networks in the Digital Library Research Community.*

1.4.3.3. Ponderada y dirigida

Para permitir que una expresión de la magnitud de relación se representan las RC como un grafo dirigido ponderado. La coautoría del grafo G es denotada como $G = (V, E, W)$. Donde V es el conjunto de vértices (autores), E es el conjunto de aristas (relaciones coautor entre los autores) y W es el conjunto de los pesos $w_{i,j}$ asociados con cada arista de conexión de un par de autores.

Se propone, para determinar la magnitud de la relación entre dos autores, la base de dos factores:

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

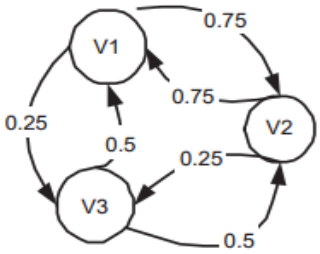

Figura IV: *Grafo G. Red de coautoría ponderada dirigida. Fuente: Liu, Xiaoming (2005). Co-Authorship Networks in the Digital Library Research Community.*

1.4.4. Estudio de comunidades científicas

En 1945 se adoptó una forma de trabajo denominada colaboratorios, del inglés *collaboratory*[3], o centro de investigación sin muros, un paradigma en la práctica científica en el que los investigadores pueden desarrollar su trabajo sin tener en cuenta la localización física (interaccionar con colegas, utilizar instrumentos, compartir datos y acceder a la información a través de bibliotecas digitales). El colaboratorio proporciona un excelente mecanismo para el aprovechamiento de las tecnologías de la información y las comunicaciones y el trabajo organizado en forma de redes (Ji Liu 2011a), (Chang y Huang 2014).

La comunicación es incuestionablemente necesaria para garantizar e incluso alentar la colaboración científica. Pero el trabajo en equipo cada vez más distribuido geográficamente también requiere, como predice Zare, el acceso a equipamiento especializado y conjuntos de datos, por lo que es preciso abordar la necesidad de elaborar una red de aplicaciones que permita compartir datos, visualizaciones y programas que faciliten el uso remoto de instrumentos (Zare 1997).

El colaboratorio se refiere a una entidad organizativa que traspasa las fronteras espaciales, alienta la interacción humana hacia el logro de resultados comunes, lo que fomenta el contacto entre investigadores que pueden o no conocerse, ofrece acceso a datos, instrumentos y herramientas comunes para el logro de las tareas de investigación. Los grupos de investigación distribuidos geográficamente pueden clasificarse en siete categorías atendiendo a sus objetivos o a sus métodos de trabajo (Bos et al. 2007):

[3] *Collaboratory* es una palabra compuesta a partir de *collaboration* y *laboratory*. Entre los visionarios de la importancia de las nuevas tecnologías en la ciencia se encuentra Vannebar Bush, que anticipó algunas de las funcionalidades del colaboratorio, imaginando una máquina, Memex, que permitiría a los científicos acceder y recuperar datos y resultados de una gran matriz de publicaciones científicas (Bush y Think 1945).

- ✓ **Instrumentos compartidos:** este tipo de colaboratorio tiene como función principal el acceso a instrumental científico. Compartir instrumentos a menudo supone el manejo de aparatos de alto costo como telescopios o supercomputadoras. Si ese acceso, además, es remoto, requerirá el suplemento de equipos de videoconferencia, ordenadores portátiles, dispositivos de almacenamiento masivo de datos y otras herramientas de comunicación.
- ✓ **Sistemas comunes de datos:** este tipo de colaboración se basa en la creación, mantenimiento y explotación de recursos de información comunes para una comunidad geográficamente dispersa. No se considerarán como tales los grupos pequeños que comparten documentos de trabajo en línea. Los sistemas de datos comunitarios se encuentran en la vanguardia de los esfuerzos de normalización de datos. Las grandes bases de datos compartidas no pueden construirse ni utilizarse hasta que sus usuarios establezcan los criterios de almacenamiento, búsqueda y explotación de datos.
- ✓ **Comunidad abierta:** es un proyecto abierto que aglutina esfuerzos de diversos individuos geográficamente dispersos involucrados en la resolución de un problema común. Se distingue de los sistemas comunes de datos en que las contribuciones son en forma de trabajo y no de datos. Por otro lado, se diferencia de los centros de investigación distribuidos porque sus requisitos de participación son más abiertos. Uno de los retos tecnológicos que plantea este tipo de colaboratorio, es la creación de un sistema universal capaz de trabajar sobre cualquier plataforma y que sea fácil de usar, para que los participantes puedan obtener resultados de forma rápida sin necesidad de perder demasiado tiempo en su aprendizaje. En este caso, los problemas organizativos deben dirigirse a la consecución de controles de calidad efectivos que eviten los problemas de obtención masiva de trabajos de un conjunto heterogéneo de contribuyentes.
- ✓ **Comunidades virtuales de prácticas:** son redes de individuos que comparten un área de investigación y que se comunican a través de las nuevas tecnologías. Estas comunidades virtuales pueden compartir noticias profesionales de interés, técnicas, recursos en línea sobre la materia, entre otras. Se distinguen de los centros de investigación distribuidos porque no realizan trabajos o proyectos en común.
- ✓ **Comunidades virtuales de aprendizaje:** su principal objetivo es el aumento de las capacidades de los participantes, no necesariamente dirigidos a investigación original. Suelen ser comunidades formales de educación, pero también pueden darse en forma de servicios de desarrollo profesional.
- ✓ **Centros de investigación distribuidos:** funcionan como cualquier centro de investigación tradicional pero sus miembros no comparten un espacio físico, sino que trabajan a distancia. Es un intento de aglutinar talento, esfuerzo y recursos más allá de investigadores individuales.

Se configuran en un área temática de interés común y desarrollan proyectos en la misma. Este tipo de colaboratorio se enfrenta con problemas tecnológicos, incluyendo las dificultades de normalización de datos y los derivados del necesario soporte técnico a distancia. Al tratarse de uno de los proyectos organizativos más ambiciosos, requiere el mantenimiento de la participación de los diversos actores, normalizar los protocolos de comunicación, facilitar la toma de decisiones distribuidas y garantizar el soporte administrativo.

- ✓ **Infraestructura comunitaria:** este tipo de proyecto trata de desarrollar infraestructuras que permitan la elaboración de trabajos que profundicen en dominios científicos concretos. Por infraestructuras se entienden recursos comunes que facilitan el trabajo investigador, como programas informáticos, protocolos de normalización, nuevos instrumentos científicos o métodos educativos. Este tipo de proyecto es, a menudo, interdisciplinar. Por tanto, la colaboración científica actual no puede entenderse sin las nuevas tecnologías. Resulta fundamental la presencia y también la necesidad, de infraestructura tecnológica suficiente para desarrollar el trabajo investigador compartido con éxito pero, por otro lado, también exige la disponibilidad local de las habilidades necesarias para su explotación. En resumen, la aparición del trabajo investigador en forma de grupos y su nueva organización en forma de colaboratorio supone la convergencia de la tecnología informática y la práctica científica.

Puede decirse que la comunidad científica internacional es una gran colaboración y que la investigación básica es una actividad global donde todos los investigadores trabajan juntos para hacer avanzar el conocimiento científico. Se intercambian ideas sobre cómo hacer experimentos, qué hipótesis probar, qué nuevos instrumentos construir, cómo relacionar los resultados experimentales con los modelos teóricos, por solo mencionar algunas. En estas y otras tareas, los miembros de un grupo de investigación no solo hablan entre ellos, sino que reciben ayuda y consejo de otros.

Otros trabajos muestran nuevas ventajas para la autoría múltiple. Nudelman y Landers sugieren que el crédito de la comunidad científica es mayor para los autores de trabajos en colaboración que para aquellos que firman en solitario. El número de coautores también parece correlacionarse con el impacto de los documentos. Lawani muestra que cuando el número de autores por documento aumenta, la proporción de documentos de alto impacto también aumenta. Diversos autores indican que la investigación realizada por grandes grupos tiende a ser más influyente (Rodríguez y de Moya Anegón 2007).

Por otra parte, Rousseau señala que no es cierto que los artículos con autoría múltiple reciban más citas que los firmados en solitario, ni siempre es correcta la suposición de que mientras más autores tenga un trabajo, más citado será. La mayor frecuencia de citas se da, a menudo, más entre artículos

de autoría múltiple que en artículos firmados en solitario, pero esta relación no es tan fuerte como para que se mantenga bajo cualquier circunstancia y para todos los dominios científicos (Rousseau 2001).

Para Zare el progreso de la investigación demanda mecanismos que soporten la mega-colaboración de grupos de investigación, que será la forma de trabajo necesaria para lograr la resolución de problemas globales a los que se enfrentarán los científicos. La actividad investigadora, por tanto, se organizará en forma de inteligencia distribuida, en la que la experiencia y el conocimiento se sustentarán en una única ubicación física que podrá ser accesible y compartida por el resto de la comunidad con independencia del lugar en el que se encuentren. De esta manera, la información estará disponible para cualquiera, en cualquier lugar y en cualquier momento, trasladando la información, el control y el poder de los sistemas centralizados a sistemas distribuidos centrados en los individuos. Este modelo supone una profunda revisión de la organización de la ciencia, que se aleja cada vez más de los "colegios invisibles", donde el centro del conocimiento se basaba en la intervención de un pequeño número de investigadores de élite que trabajaban en conjunto. La noción de inteligencia distribuida, en cambio, supone la movilización del esfuerzo científico colectivo y el consecuente aumento de los resultados y de las capacidades, dado que existirá un mayor número de participantes en la creación del nuevo conocimiento (Zare 1997).

La comunidad científica estructura sus relaciones conforme a modelos de redes, donde los vértices representan a los individuos, disciplinas o instituciones; y las aristas, a los flujos de información que intercambian esos vértices. La producción de trabajos de investigación en el campo de redes complejas en los últimos años ha sido considerable. Se han descrito y analizado los rasgos topológicos y dinámicos de redes de tipo social, biológico o tecnológico. Uno de los aspectos a los que se ha prestado mayor atención ha sido a la existencia de subconjuntos de vértices fuertemente entrelazados, unas veces conectados y otras veces desconectados de la red.

Las comunidades juegan un papel fundamental en las propiedades de las estructuras complejas, por lo que identificar y analizar su naturaleza es una importante tarea que se está llevando a cabo desde campos tan diversos como la informática o la sociología, pasando por la bibliometría o la bioquímica, con el objeto de revelar la organización informal y la naturaleza de los flujos de información dentro de esos sistemas complejos. Sin embargo, el concepto de comunidad, entendida como patrón estable de transacciones entre individuos o grupos de individuos, es variable (Montfort 2004). Dependiendo de la red motivo de análisis, estas agrupaciones pueden surgir de forma natural como herramienta valiosa para el análisis de su estructura en unos casos, mientras que en otros su extracción debe realizarse de forma artificial (Ichise, Takeda y Ueyama 2006).

Las comunidades se componen por distintos subconjuntos de vértices en un grafo que están relacionados. La extracción de esas comunidades facilita la comprensión de las redes, para lo que se hace necesario:

- **El descubrimiento de la comunidad:** consiste en la extracción de las distintas comunidades de una red determinada.
- **La identificación de la comunidad:** se centra en la caracterización de cada uno de los subconjuntos de vértices extraídos de la red original.

Para la detección de comunidades en la literatura se hace referencia a técnicas y herramientas para la localización de subconjuntos de vértices fuertemente entrelazados, siendo denominadas comunidades. Estos procesos tienen por objeto el análisis de las propiedades de esas estructuras de manera que permitan revelar y conocer, con el mayor grado de detalle posible, sus formas de organización. Además analizan la naturaleza de sus flujos de información y sus funciones internas dentro del sistema complejo del que forman parte.

Los métodos existentes para hallar comunidades en grandes redes son útiles si la estructura de comunidad ha de interpretarse en términos de conjuntos de comunidades separadas. Sin embargo, la mayor parte de las redes del mundo real se caracterizan por contener comunidades muy entrelazadas y solapadas. Además, los miembros de esas comunidades tienen a su vez las suyas propias, configurando una intrincada red entre todos ellos (Palla et al. 2005). Estos métodos y algoritmos, por tanto, tratan de detectar o extraer esas comunidades para que, posteriormente, puedan ser identificadas y caracterizadas. La mayor parte de ellos proceden de campos como la física o las matemáticas y realizan la extracción mediante técnicas automáticas que se han mostrado poco eficaces hasta la fecha. Esto se debe a que en unos casos, requieren el conocimiento de determinados valores a priori, mientras que en otros, es preciso indicar cuándo debe detenerse la extracción. Varios de ellos son invasivos, transformando la apariencia de la red original, optando por convertirla en un grafo distinto o por la eliminación sistemática de aristas. Además, suelen forzar la pertenencia de las variables a un único factor.

Aunque los avances para la resolución de estos inconvenientes son constantes y su utilización futura augura buenos resultados, existe un análisis procedente de las ciencias sociales: el análisis factorial. Su uso para la identificación de comunidades se basa en la posibilidad de definir los subgrupos en función de la estructura de elección de sus enlaces, basándose en la premisa de que los miembros de cada uno de ellos tienden a elegir a los mismos colaboradores y a ser elegidos por los mismos autores. Por tanto, la pertenencia a un subgrupo se establecerá en función de las similitudes de

elección dadas y recibidas por cada autor. Estas condiciones hacen que las elecciones muestren tendencia a la reciprocidad y que los factores obtenidos y rotados permitan obtener una estructura simple (Perianes-Rodríguez 2007).

En resumen, el análisis factorial permite la localización de las distintas comunidades de la red a través del estudio de los componentes de la matriz original en redes no dirigidas y sin aristas ponderadas. Posteriormente, será posible identificar y agregar los subgrupos en función de características comunes. En este sentido, es importante enfatizar el hecho de que si una determinada característica (el trabajo en una misma especialidad, por ejemplo), es relevante en la elección de autorías, no hay ningún impedimento para que existan dos o más subgrupos en torno a esa característica común, pero en factores distintos.

1.4.5. Detección de comunidades en redes

El objetivo que persigue la detección de comunidades consiste en identificar, a partir de utilizar la tipología de un grafo, grupos de vértices conectados entre sí y que compartan características comunes o tengan un rol similar dentro del conjunto de investigaciones. Se hace referencia a una publicación de Rice en 1927 donde se aplicaban técnicas de detección de comunidades para identificar grupos políticos de personas de acuerdo con la similitud de sus patrones de voto. Sin embargo, una investigación más popular la han desarrollado los investigadores Girvan y Newman en el 2002 (Perianes-Rodríguez 2007).

La detección de comunidades presenta un problema asociado a la no existencia de una definición de comunidad universalmente aceptada, más allá de la noción de que debe haber más aristas entre los vértices de una comunidad que con los vértices de otras comunidades (Fortunato 2010). Algoritmos desarrollados en la última década poseen su propia definición, la cual depende del estudio que se realice. Los algoritmos pueden clasificarse, además de por la técnica utilizada para llevar a cabo la identificación de comunidades, de acuerdo con tres criterios comunes a ellos. El primero hace referencia al determinismo del algoritmo, puede serlo o no. El segundo considera los tipos de grafos para los que el algoritmo es capaz de detectar comunidades, atendiendo a la dirección de sus aristas, puede manipular grafos dirigidos, no dirigidos o ambos; y a su peso, manipulando grafos ponderados, no ponderados o ambos. El tercer criterio contempla los tipos de comunidades que genera el algoritmo, detectando comunidades solapadas, no solapadas o ambas (Rodríguez, Gómez y de Moya Anegón 2010).

Las RC pueden desenvolverse como comunidades científicas en el ámbito investigativo, por lo que resulta imprescindible el estudio de algoritmos existentes para la detección de comunidades en redes.

1.4.6. Algoritmos para la detección de comunidades en redes

Una gran variedad de métodos y algoritmos, cada uno de ellos con su propia definición intrínseca de comunidad, han sido desarrollados para intentar extraer la partición óptima de una red. Muchos de los algoritmos o métodos propuestos para la detección de comunidades en redes basados en agrupación o comunidades son versiones modificadas o inspiradas en los conceptos de algoritmos centrados en el mínimo corte, algoritmos basados en jerarquía, el algoritmo original de Girvan-Newman, conceptos de maximización de modularidad, algoritmos que utilizan métricas de la información y la teoría de códigos y algoritmos basados en camarilla (Saha et al. 2015).

1.4.6.1. Algoritmos divisivos basados en la intermediación

Los algoritmos divisivos se basan en eliminar las aristas que conectan vértices pertenecientes a diferentes comunidades, de manera que estas queden aisladas unas de otras. La dificultad de este enfoque reside en identificar tales aristas. Uno de los algoritmos divisivos más conocidos es el de los investigadores Newman y Girvan, el cual determina las aristas que conectan vértices pertenecientes a diferentes comunidades a partir de una extensión de la intermediación (Newman y Girvan 2004). La determinación se refiere a una medida de centralidad que indica la influencia de un vértice de un grafo en base al número de caminos mínimos, entre todos los pares de vértices, que pasen por dicho vértice (Freeman 1977).

1.4.6.2. Algoritmos basados en dinámica social y redes

La aparición de servicios como Facebook o Twitter ha brindado la oportunidad de observar el proceso natural de creación de las comunidades. En el 2010 fue propuesto un algoritmo basado en dos principios: naturaleza intrínseca de las comunidades y la detección longitudinal. El primer principio hace referencia a la posibilidad de encontrar comunidades de diversos tamaños en el mismo grafo. El segundo principio tiene por objetivo recoger la dinámica del grafo, esto es el momento en que se crea un vértice, una arista o una comunidad (Cazabet, Amblard y Hanachi 2010).

Los grafos clásicos se han convertido en un instrumento importante en la evaluación de los algoritmos de detección de comunidades debido a su similitud en estructura y tamaño a las redes sociales bajo estudio. El creciente avance de las redes sociales hizo necesarios la investigación y desarrollo de algoritmos que permitieran generar redes con características semejantes a las nuevas redes sociales.

Los generadores de grafos o redes comparten con los algoritmos de detección de comunidades dos criterios de acuerdo con los cuales pueden clasificarse. El tercer criterio de los algoritmos de detección de comunidades, el determinismo, no se tiene en cuenta, puesto que los algoritmos

generadores de grafos por definición son no deterministas (Rodríguez, Gómez y de Moya Anegón 2010).

1.4.6.3. Algoritmos basados en jerarquía

Los algoritmos basados en jerarquía es uno de los principales tipos de algoritmos de detección de comunidades en redes complejas. En (Silva y Zhao 2007) utilizaron el concepto de órdenes tipológicos entre los datos de entrada representados como grafos y desarrollaron un algoritmo para obtener las agrupaciones en diferentes escalas. El algoritmo inicialmente consistía en la construcción de la red de datos de entrada y en segundo lugar, la partición jerárquica de la red formada. El algoritmo, aunque era libre completamente del cálculo de las distancias físicas entre los datos de entrada, constantemente produce un grafo conexo con vértices fuertemente vinculadas dentro de una comunidad y esporádicamente vinculaba vértices entre las diferentes comunidades. Los autores aplicaron su algoritmo para el problema de la agrupación de píxeles en imágenes.

Guang Xu propuso un nuevo algoritmo, llamado Descubrimiento Latente de Comunidad, para la detección de la comunidad en las redes sociales complejas (Xun et al. 2012). Específicamente, su algoritmo divide los actores principales, basados en un modelo probabilístico jerárquico y un modelo estadístico por temas, que se normaliza por el ordenamiento de los datos de la red. Su algoritmo se inspira en el Principio de Pareto, que da cuenta de la existencia irregular de dos tipos diferentes de actores en la red, centrándose en los actores principales que suelen compartir una pequeña parte de los vértices, pero que tienen una gran influencia en la red compleja. Probaron su algoritmo en tres grandes sociales redes y encontraron su desempeño competitivo de los populares algoritmos existentes para esta categoría de problemas.

Oliveira y colaboradores proponen un algoritmo de agrupamiento basado en las representaciones teóricas de grafos y el descubrimiento de la comunidad en redes complejas. Inicialmente, representan los datos de entrada como una red y luego dividen la red en subredes para crear grupos de datos (de Oliveira et al. 2008). En la primera etapa, cada uno de los vértices tiene un ángulo inicial asignado al azar. Este ángulo inicial es gradualmente modificado de acuerdo con el ángulo de los vecinos. En última instancia, la red alcanza un estado de equilibrio. En este estado los vértices en el mismo grupo tienen ángulos comparables. Este proceso se repite y los resultados se obtienen luego de un agrupamiento jerárquico. Simulaciones por los autores demuestran que este algoritmo tiene el potencial para encontrar grupos en diferentes formas, compactibilidad y proporciones. El algoritmo también tiene la capacidad de generar grupos con diversos grados de refinamiento. Además, el algoritmo propuesto también es robusto y eficiente.

1.4.6.4. Algoritmos basados en teoría de la información

Los algoritmos de teoría de la información son otro grupo importante de algoritmos de agrupamientos para la detección de la comunidad en redes complejas. En el 2012 se emplea el solapamiento de las comunidades de acuerdo con la vinculación por etiquetas para mejorar la agrupación de textos (Cravino, Devezas y Figueira 2012). Basado en un pequeño conjunto de noticias y resúmenes, los autores construyen una red de co-ocurrencia de etiquetas definidas por el usuario a partir de los campos de los metadatos extraídos de las noticias. Describen una medida de cercanía ponderada según la similitud del coseno, que tiene en cuenta tanto el extracto de los vectores y las etiquetas de los mismos. A partir de entonces, se calcula el peso de las etiquetas con las etiquetas correlacionadas que existen en la comunidad descubierta y usan una novedosa métrica de distancia, para identificar grupos de documentos socialmente sesgados.

Aunque la detección de la comunidad en las redes sociales se basa generalmente en la agrupación del grafo empleando estructuras de información como la estructura de las aristas o topología de vértices, para la identificación de grupos Huang y Yang utilizaron la información semántica presente en los puestos de medios de comunicación social para encontrar comunidades ocultas en estos medios. Incorporando el supuesto de que el contenido publicado por los usuarios podrá expresar las relaciones entre los usuarios o entidades (Huang y Yang 2012). Este método es adecuado para detectar las comunidades en las redes que evolucionan continuamente, por ejemplo las redes sociales.

En el 2011 se propone un nuevo algoritmo para la detección de estructuras de la comunidad en redes complejas ponderadas. El método construye una red compleja ponderada con respecto a la similitud entre los pares de documentos calculados por la función coseno, y entonces el algoritmo realiza búsquedas de los conjuntos densos (Xie y Szymanski 2011). Siendo útil en agrupamiento de documentos de texto representados por modelo de espacio vectorial.

Zhang y colaboradores investigaron el tema de la confianza en el comercio electrónico basado en redes sociales utilizando la métrica de confianza grado de información sobre la base de información mutua entre los sujetos. Incorporaron métricas desarrolladas con anterioridad basadas en el grado de confianza directa de información y grado de confianza de información global para construir relaciones de confidencia entre los sujetos (Zhang et al. 2013). El coeficiente de agrupamiento y grado de confianza de la información mundial se adoptaron para construir comunidades de confianza.

1.4.6.5. Algoritmos basados en modularidad

La modularidad es una función propuesta por Girvan y Newman en el 2002 como criterio de parada de su algoritmo. Indica la calidad de las comunidades detectadas; cuanto mayor es la modularidad, mejor es el resultado obtenido por el algoritmo. Se distinguen tres tipos de técnicas basadas en modularidad: algoritmos voraces con un grado bajo de optimización, algoritmos de complejidad elevada con una precisión mayor y algoritmos que equilibran la complejidad computacional y el grado de optimización (Rodríguez, Gómez y de Moya Anegón 2010).

En el 2011 se desarrolló una novedosa métrica de representación, la matriz modularidad de co-vecino, para evaluar la calidad de la comunidad, por lo que el problema de la detección de la comunidad es transformado a la de un problema de la agrupación de los vectores propios en el espacio euclidiano. A partir de entonces, la arquitectura de la red de la comunidad se identifica con el algoritmo de agrupamiento espectral (Ji Liu 2011b). Una ventaja principal de este algoritmo es que es libre del ruido generado por los puntos iniciales medios de la agrupación, por ejemplo, en *k-means*.

Scibetta abordó la estrategia de división de la red en grupos o distribuciones por áreas medidas para la detección de pérdidas de agua en las redes de distribución, dado que la medición de los flujos entrantes y salientes para cada grupo o distrito (zona) medida, permite una cuantificación de las pérdidas de agua. Los autores utilizan el método de detección de la comunidad desarrollado en la teoría de redes complejas para identificar grupos o áreas del distrito que pertenecen al sistema de distribución de agua (Scibetta et al. 2013). El método tiene como objetivo encontrar soluciones que satisfagan las restricciones de maximización de la modularidad y la reducción del número de comunidades. Los autores afirman que el método es suficientemente escalable.

Investigaciones hacen referencia a una propuesta de diseño de un sistema de recomendación visual para recomendar recursos de aprendizaje a los alumnos cibernéticos dentro de la misma comunidad, mediante el uso de un algoritmo de detección comunidad en las redes complejas a gran escala de los estudiantes cibernéticos, basado en el uso de los datos y recursos de la Web de dichos estudiantes (Zhuhadar, Yang y Nasraoui 2012). Su algoritmo utiliza una heurística que logra inicialmente una agrupación por un algoritmo de visualización basado en la fuerza. Posteriormente, el algoritmo utiliza la información en la modularidad de red para elegir buenas descomposiciones de las que se encuentran utilizando el algoritmo de visualización.

1.4.7. Visualización de la información

La visualización de la información es el proceso de pasar de representaciones gráficas a representaciones perceptivas, se eligen las técnicas de codificación que maximicen la comprensión

humana y la comunicación. El enfoque de la exploración de datos a través de la visualización busca combinar la flexibilidad, creatividad y conocimiento presentes en grandes volúmenes de datos. Actualmente una de las técnicas de visualización con mayor aceptación son los grafos. Para la cual se han desarrollado diferentes herramientas especializadas en la visualización y comprensión de grafos.

Dichas herramientas usan diferentes técnicas, algoritmos y marcos de trabajos que agilizan el proceso de visualización. Se señala que tienen un mayor impacto aquellas que combinan diferentes patrones de representación para un gran número de datos. Estas son fundamentales para el análisis y evaluación de grafos de gran tamaño. Herramientas como IGraph, JUNG y Gephi se han ido desarrollando en este campo; en las cuales se centra el análisis y posterior comparación de las mismas atendiendo a sus características.

1.4.8. Visualización de redes

1.4.8.1. IGraph

IGraph se define como una biblioteca de código abierto, distribuida bajo licencia GPL (del inglés *General Public License*), para el estudio y análisis de redes/grafos. Los principales objetivos de esta biblioteca se centran en proveer un conjunto de tipos de datos y funciones para una fácil implementación de algoritmos de grafos, un manejo rápido de grandes grafos con millones de vértices y aristas y permitir un prototipado rápido por medio de un lenguaje de alto nivel como R[4]. IGraph permite manipular tanto grafos dirigidos como no dirigidos, no admite la implementación de hipergrafos. Por otro lado, cuenta con implementaciones de problemas típicos de teoría de grafos como árboles de expansión mínima y flujo de red, también implementa algoritmos para algunos métodos de análisis estructural dentro de una red (Web IGraph, 2015).

IGraph puede ser instalado como una biblioteca del lenguaje C, como un paquete de R, como un módulo de extensión de Python[5] o como una extensión de Ruby[6].

1.4.8.2. JUNG

Java Universal Network/Graph (JUNG) que es un marco de trabajo de código abierto que provee un lenguaje común y extensible para la manipulación, análisis y visualización de datos que pueden ser representados como un grafo o una red. Está desarrollado en el lenguaje de programación Java, el cual permite el desarrollo de aplicaciones basadas en JUNG por medio de la interfaz de programación

[4] Lenguaje de programación R. http://www.r-project.org/.
[5] Lenguaje de programación Python. http://www.python.org/.
[6] Lenguaje de programación Ruby. http://www.ruby-lang.org/es/.

de aplicaciones (IPA) o API (del inglés *Application Programming Interface*) disponible para su uso, así mismo permite la integración de bibliotecas de terceros.

Se considera que JUNG no es un producto final y no intenta serlo, es una biblioteca que permite el trabajo con grafos pero que necesita conocimientos de programación ya que debe estar implementada en una aplicación Java. Puede ser usada para el desarrollo de herramientas flexibles orientadas al análisis de redes/grafos (O'Madadhain et al. 2005). Posee varias características como:

- Soporte a una variedad de representaciones de entidades y sus relaciones, incluyendo grafos dirigidos y no dirigidos, grafos multi-modales (grafos que contienen más de un tipo de vértices o aristas), multigrafos e hipergrafos.
- Mecanismo para etiquetar grafos, entidades y relaciones con metadatos. Esta capacidad facilita la creación de herramientas analíticas para conjuntos complejos de datos que necesitan analizar las relaciones entre las entidades, así como los metadatos asociados a cada entidad y relación.
- Implementación de algoritmos de teoría de grafos, análisis exploratorio de datos, análisis de redes sociales y aprendizaje automático. Esto incluye rutinas para agrupamiento, descomposición, optimización, generación aleatoria de grafos, análisis estadístico y cálculo de distancia de una red, flujo y medidas de clasificación (centralidad, *PageRank*, *HITS*, entre otras).
- Marco de trabajo de visualización que permite la construcción de herramientas para la exploración interactiva de redes de datos. Los usuarios pueden elegir de una cantidad de diseños y algoritmos de dibujo, o pueden usar el marco de trabajo para crear sus propios algoritmos.
- Mecanismo de filtrado los cuales extraen subconjuntos de una red; permitiendo a los usuarios centrar su atención o sus algoritmos en una porción específica de la red.

1.4.8.3. Gephi

Gephi es una herramienta para la exploración, navegación y análisis de grafos. Permite a los usuarios interactuar con las distintas representaciones, manipular las estructuras, formas y colores que revelan propiedades ocultas. Utiliza un motor de renderizado en tres dimensiones (3D) para mostrar las redes en tiempo real y acelerar la exploración. Su objetivo es ayudar a los analistas de datos a hacer hipótesis, descubrir patrones, aislar singularidades en las estructuras o encontrar fallas en los datos (Web Gephi 2015).

Gephi se destaca por ser una herramienta libre 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 (Web Gephi 2015).

Se plantea además que soporta la representación de grafos dirigidos, no dirigidos y mixtos e hipergrafos. Uno de los aspectos importantes cubiertos por Gephi es la interacción en tiempo real, la que permite modificar propiedades de los vértices y aristas al mismo tiempo que se modifica la representación del grafo, ofreciéndoselas al usuario sin largas esperas. Así mismo permite realizar agrupaciones, filtrado, manipulación, navegación y proveer acceso a los datos (Bastian, Heymann y Jacomy 2009).

Gephi dispone del código fuente para su utilización y de una API denominada Gephi Toolkit[7] para desarrollar aplicaciones propias basadas en dicha herramienta. Actualmente está disponible la versión de Gephi 0.8.2.

En (Medrano, Berrocal y Figuerola 2011) se realiza una comparación de las herramientas anteriormente mencionadas. Se toma un grafo generado a partir de la red social Facebook[8] con 466 vértices y 4655 aristas. Los aspectos que se evaluaron de cada herramienta son:

- Los distintos algoritmos de representación o *layout* que dispone.
- Los distintos algoritmos para la detección/identificación de comunidades o agrupaciones.
- El grado de interacción otorgado por la herramienta.
- El desempeño y rendimiento de la herramienta.

Se muestran diferentes resultados con la ejecución de métodos y algoritmos de *layout* (visualización de grafos) como son Kamada-Kawai, Fruchterman-Reingold y *ForceAtlas* (algoritmo implementado por Gephi) al ejecutar algoritmos de agrupamiento para la detección de comunidades a partir de un análisis en la red social Facebook.

La ejecución de *ForceAtlas* demostró ser la opción más acertada de todas las estudiadas, para este caso y para este conjunto de datos, principalmente por la velocidad de ejecución y por la disposición correcta de las comunidades encontradas (cada comunidad identificada con un color específico se

[7] Toolkit Gephi. http://gephi.org/toolkit/.
[8] Facebook. http://www.facebook.com/.

encuentra geográficamente separada del resto de las comunidades y los elementos pertenecientes a dichas comunidades se encuentran muy cercanos unos de otros).

1.5. Conclusiones parciales

Las variantes para modelar una red de coautoría a partir de un grafo se dividen teniendo en cuenta la tipología y direccionalidad de las aristas en la red. Cada una aporta o no diferentes características y datos a la red de coautoría, destacando la modelación dirigida y ponderada como la opción factible para analizar la frecuencia y exclusividad de publicación entre los autores presentes en la red.

Durante la revisión de la literatura no se ha encontrado evidencia científica de la existencia de métodos o algoritmos para la detección de comunidades a partir de redes modeladas en grafos RDF. Sin embargo, existen testimonios de algoritmos para la detección de comunidades en redes.

Los algoritmos para la detección de comunidades en redes son heterogéneos. Varían en cada uno el determinismo de los algoritmos, los tipos de redes/grafos que pueden manejar y los tipos de comunidades que genera cada uno. Se considera que los algoritmos basados en modularidad son los más eficientes referentes al problema de detectar comunidades en redes.

En la visualización de redes las herramientas analizadas presentan diferentes formatos de entrada, algoritmos de visualización y parámetros de personalización. Se destaca la herramienta Gephi por brindar una mejor representación gráfica y varias opciones a la hora de interactuar con el resultado, por ello es utilizada como parte de la propuesta de solución.

CAPÍTULO 2. DESCRIPCIÓN Y REPRESENTACIÓN DEL MÉTODO

2.1. Introducción

El presente capítulo está compuesto por dos secciones. En la primera sección se describe la propuesta de solución desde un enfoque teórico, basada en un método compuesto por tres fases para la detección de comunidades en RC a partir de un grafo RDF. En la segunda sección se describe la arquitectura, componentes, tecnologías y algoritmos aplicados en el desarrollo de un prototipo funcional que implementa el método propuesto.

2.2. Método propuesto

Al modo estructurado y ordenado de obtener un resultado, descubrir la verdad y sistematizar los conocimientos se le conoce como método. En las matemáticas, un método de cómputo se utiliza para encontrar una respuesta con respecto a un problema determinado. La mayoría de las funciones matemáticas básicas como suma, resta, multiplicación y división son los métodos de cálculo más comunes.

En el trabajo se asume que se tiene un grafo RDF que contiene los metadatos de los artículos científicos y sus autores. A partir de este se modela una red de coautoría, se detectan las comunidades presentes y se visualizan las comunidades detectadas.

El método de solución para la detección de comunidades a partir de una red de coautoría en grafos RDF está compuesto por tres fases. La salida de una fase constituye la entrada de la próxima. Las fases propuestas son:

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

Para el desarrollo del método propuesto y teniendo como base la revisión realizada con anterioridad se muestra en la Figura V la representación del mismo. Se describen además en orden lógico los diferentes conceptos, artefactos, algoritmos y herramientas involucrados en cada una de las fases.

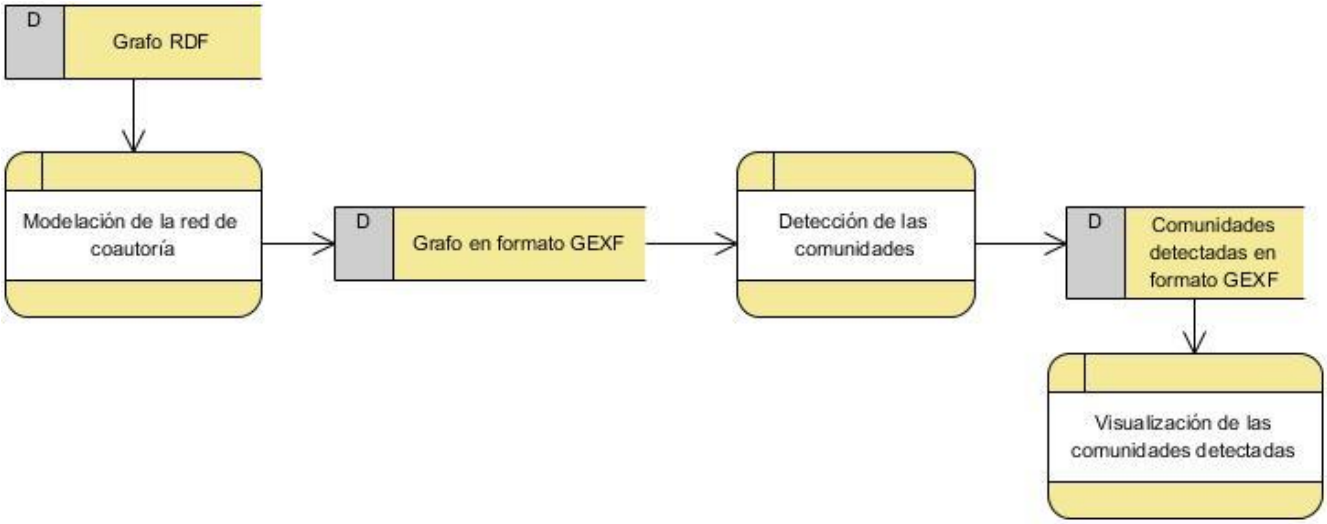

Figura V: *Método propuesto*

El método representado en la Figura V se puede describir aplicando teoría de conjuntos con un grupo de restricciones.

- Sea (G) un grafo RDF definido de la forma $G = \{J^G, R^G, A^G\}$ donde:
  - $J^G = \{J_1^G, \dots, J_n^G\}$: Conjunto de tripletas $(J^G, subconjunto(G))$ que describen las revistas
  - $R^G = \{R_1^G, \dots, R_n^G\}$: Conjunto de tripletas $(R^G, subconjunto(G))$ que describen los artículos en una revista
  - $A^G = \{A_1^G, \dots, A_n^G\}$: Conjunto de tripletas $\left(A^G, subconjunto(G)\right)$ que describen los autores de los artículos
- Se modela a partir del grafo (G) una red de coautoría (RC) generándose un grafo de la forma $RC = \{N^{RC}, E^{RC}\}$ donde:
  - $N^{RC} = \{N_1^{RC}, \dots, N_n^{RC}\}$: Conjunto de vértices del grafo (RC)
  - $E^{RC} = \{E_1^{RC}, \dots, E_n^{RC}\}$: Conjunto de aristas del grafo (RC)
- Se detectan las comunidades presentes en (RC) generándose el grafo $VC = \{N^{VC}, E^{VC}, A^{VC}\}$, donde:
  - $N^{VC} = \{N_1^{VC}, \dots, N_n^{VC}\}$: Conjunto de vértices del grafo (VC)
  - $E^{VC} = \{E_1^{VC}, \dots, E_n^{VC}\}$: Conjunto de aristas del grafo (VC)
  - $A^{VC} = \{A_1^{VC}, \dots, A_n^{VC}\}$: Conjunto de atributos que describen las comunidades presentes en el grafo (VC)
- Se visualizan las comunidades detectadas mediante un algoritmo de *layout* que pueden variar de acuerdo al grafo (VC):

- Si $N^{VC} < 100\,000$ se recomienda un algoritmo de atracción – repulsión que priorice la rapidez ante la precisión con un criterio de parada.
- Si $N^{VC} \geq 100\,000$ se recomienda un algoritmo de atracción – repulsión que priorice la calidad evitando el solapamiento entre los vértices.

2.2.1. Modelación de la red de coautoría

A partir de un grafo RDF, que describe los metadatos pertenecientes a las publicaciones de diferentes coautores, el método propuesto debe procesar y transformar dicho grafo en una red de coautoría dirigida y ponderada. Esto permite obtener una expresión de la magnitud de las relaciones entre los vértices de la red.

Se hace necesario definir elementos para la transformación. Las relaciones de coautoría en el grafo (G) son denotadas como $G = (V, E, W)$ donde:

- V es el conjunto de vértices (autores)
- E es el conjunto de aristas (relaciones de coautoría entre autores)
- W es el conjunto de los pesos w_{ij} asociados con cada arista de conexión de un par de autores.

Se propone para determinar la magnitud de la relación entre dos autores en base a tres factores:

- La frecuencia de coautoría: autores que con frecuencia son coautores deben tener un mayor peso de coautoría.
- Exclusividad: autores que publican exclusivamente entre ellos.
- Número total de coautores en los artículos: si un artículo tiene varios autores, cada relación individual del coautor debe tener una ponderación menor (Liu et al. 2005).

Para determinar el peso de las relaciones de coautoría entre los autores se definen:

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

Y se calcula:

- 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. La red definida como $RC = \{N, E\}$ contiene toda la información de los autores y relaciones de coautoría para proceder a la próxima fase del método.

2.2.2. Detección de las comunidades

A partir de tener como entrada la red definida como $RC = \{N, E\}$ generada en la fase anterior, es necesario proceder a detectar las comunidades presentes en la misma. Al no 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*. Existen varios algoritmos para realizar este proceso (ver sección **1.4.6. Algoritmos para la detección de comunidades en redes**).

La aplicación de algún tipo de restricción en cuanto a la forma de detectar las comunidades presentes en la red (RC) limitaría el resultado basándose en que el concepto de comunidad está pobremente identificado y definido. Es necesario definir entonces, una función global de calidad que permita comparar los resultados en cada caso. Según la revisión bibliográfica desarrollada y un análisis realizado en (Aldecoa y Marín 2011) y (Aldecoa García 2013) la función de calidad más popular en la actualidad es *Modularity*, propuesta en 2004 por Newman y Girvan.

Para calcular la *Modularity* (Q) de una partición determinada se define que:

- $Q = \frac{1}{2m} \Sigma \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(C_i, C_j)$

Donde:

- $A_{i,j}$: es la matriz de adyacencia de (RC)
- k_i : es el grado del vértice i
- m : número de aristas
- $\frac{k_i k_j}{2m}$: número esperado de aristas entre los vértices i y j
- C_i, C_j : comunidades a las que pertenecen los vértices i y j
- $\delta(C_i, C_j) = \begin{cases} 1 \; si \; C_i = C_j \\ 0 \; e.o.c \end{cases}$

Una vez aplicada esta medida de calidad mediante un algoritmo de detección de comunidades se genera el grafo $VC = \{N, E, A\}$ encontrándose la información de los vértices, aristas y atributos de las comunidades detectadas necesarias para la próxima fase.

2.2.3. Visualización de las comunidades detectadas

En la visualización de las comunidades detectadas se separan y clasifican las mismas de acuerdo a diferentes métricas y algoritmos de *layout*. Para identificar y cuantificar las relaciones de colaboración que se establecen entre los autores presentes en el grafo (VC) se definen los siguientes criterios de clasificación:

- Grado de los vértices: permite cuantificar la cantidad de coautores dado un autor.
- Grado de entrada/salida de los vértices: permite cuantificar la cantidad de coautores con los que se publica atendiendo al nivel de coautoría en las publicaciones.
- Intermediación: representa la importancia de un autor entre las relaciones de coautoría de una comunidad.
- Cercanía: representa la cercanía que mantienen los autores dentro de una misma comunidad.
- Autoridad: identifica cuán representativo es un autor dentro de una comunidad atendiendo a las relaciones de coautoría en la red.

Los algoritmos de *layout* resuelven el problema de la distribución visual de los vértices y aristas presentes en el grafo (VC). Un caso particular de los algoritmos de *layout* son los basados en fuerza directa. Se basan en la asignación de fuerzas atractivas y repulsivas entre los vértices para encontrar una distribución óptima reduciendo al mínimo la energía del sistema. Se caracterizan porque intentan representar el grafo de forma que se crucen el menor número de aristas posibles. Para conseguirlo se basan en una idea similar a los movimientos de las partículas debido a las cargas magnéticas. Todos los vértices se repelen entre sí, pero los que tienen aristas en común se atraen. Esto provoca que en el resultado final los vértices aparezcan agrupados en comunidades en la mayoría de lo posible. Este tipo de algoritmo de *layout* no produce un único resultado de salida, no son deterministas. El resultado final depende del orden de exploración de los vértices. Pueden tener en cuenta o no el valor de las aristas para hacer los cálculos. Para la elección del mismo debe tenerse en cuenta que:

- La cantidad de vértices del grafo (VC) influye en el tiempo de ejecución del mismo.
- El modelo de energía debe irse adaptando y converger en un tiempo finito.

2.3. Implementación del método propuesto

Una vez descrito el método propuesto en la sección anterior, se hace necesario llevar a un escenario real la materialización del mismo. Se aplica una aproximación al modelo arquitectónico de tuberías y filtros y se desarrolla una implementación del método en un prototipo funcional. Se irá describiendo el modelo ontológico, la arquitectura, estándares, tecnologías y algoritmos que se aplican indistintamente en las diferentes fases definidas.

2.3.1. Modelo ontológico

Un modelo ontológico proporciona definiciones precisas de los conceptos fundamentales de un dominio específico. La aplicación de un modelo ontológico permitirá la creación o reutilización de ontologías para la modelación de los datos en un grafo RDF. El dominio de los metadatos bibliográficos ha sido ampliamente estudiado por ingenieros de ontologías y expertos en esta área del conocimiento. Para esta investigación se utilizó un grafo RDF que utiliza las ontologías de la Tabla III.

Tabla III: *Descripción de las ontologías presentes en el grafo RDF*

Nombre	Fuente de la ontología	Descripción
rdfs	http://www.w3.org/2000/01/rdf-schema#	Esquema de vocabulario RDF.
fabio	http://purl.org/spar/fabio/	Es una ontología para el registro y publicación de registros bibliográficos en la Web Semántica de trabajos académicos.
foaf	http://xmlns.com/foaf/0.1/	Ontología que describe a las personas, sus actividades y sus relaciones con otras personas y objetos.
bibo	http://purl.org/ontology/bibo/	Ontología bibliográfica. Puede ser usada como una ontología de clasificación de documentos o simplemente como una forma de describir cualquier tipo de documento en RDF.
dc	http://purl.org/dc/elements/1.1/	La Iniciativa de Metadatos Dublin Core es una organización para apoyar la innovación en el diseño de los metadatos y las mejores prácticas en toda la

		tecnología de metadatos.
swrc	http://swrc.ontoware.org/ontology#	SWRC (Web Semántica para Comunidades de Investigación). Es una ontología para el modelado de entidades tales como: personas, organizaciones, publicaciones (metadatos bibliográficos) y sus relaciones.

Estas ontologías permiten modelar los datos en función del tipo de información que se desea manejar. Este vocabulario se usa de manera combinada para crear modelos de datos más ricos que permiten representar la información de una manera más completa. En la Figura VI se describe el modelo ontológico utilizado para modelar el grafo RDF.

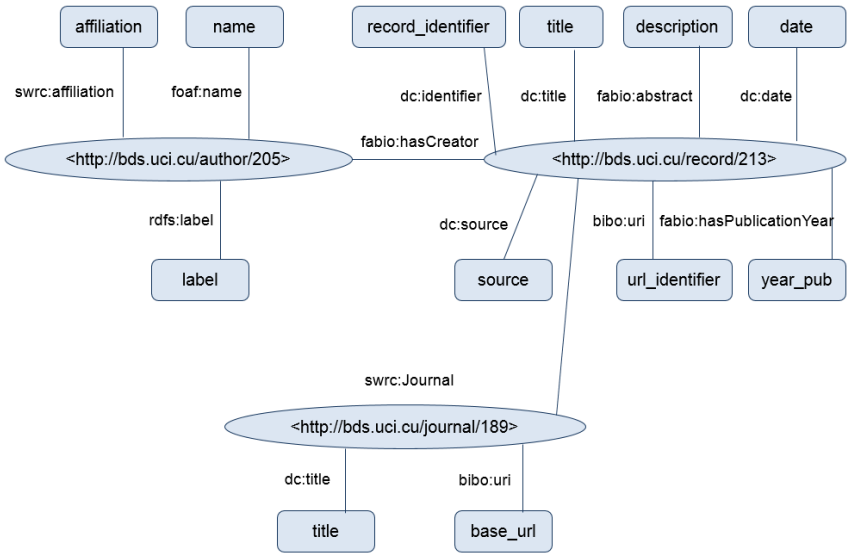

Figura VI: *Modelo ontológico del grafo RDF*

2.3.2. Arquitectura

La propuesta de solución sigue una arquitectura de flujo de datos. Es aplicada cuando los datos de entrada se transforman en datos de salida mediante una serie de componentes aplicando el cálculo o la manipulación. La estructura utilizada es de tuberías y filtros. Posee un conjunto de componentes, denominados filtros, conectados por tuberías que transmiten datos de un componente al siguiente. Cada filtro se diseña para esperar la entrada de datos con cierta forma y producir su salida (al siguiente filtro) de una forma específica (Pressman 2002).

Para lograr un mejor entendimiento de la arquitectura propuesta, en la Figura VII se describe cómo interactúan cada uno de los componentes y paquetes según lo definido en el método propuesto. Además, se describen las principales características y funcionalidades presentes en los diferentes paquetes de la solución desarrollada.

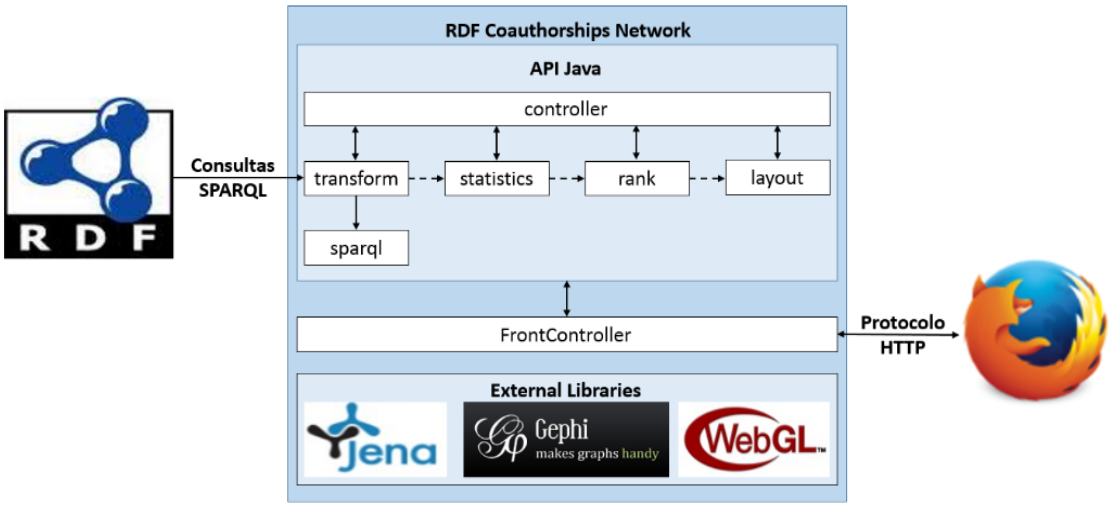

Figura VII: *Arquitectura de la propuesta de solución*

A partir de un grafo RDF dado, los paquetes de la aplicación van procesando y enviando la información de un paquete a otro a través del ***controller*** para posteriormente visualizar la red de coautoría en un navegador web. El paquete ***transform*** extrae la información de los autores y sus publicaciones usando las consultas SPARQL del paquete ***sparql***. Transforma la información necesaria para medir la colaboración científica a partir de la coautoría y genera un grafo ponderado y dirigido. Al aplicar la estadística correspondiente en el paquete ***statistics*** se clasifican los elementos de la red según lo definido en ***rank*** para que los algoritmos de distribución visual de los vértices y aristas presentes en ***layout*** ordenen los vértices y aristas de la red. A continuación el navegador web interpreta los datos y muestra las comunidades detectadas en la red de coautoría.

Como se puede evidenciar la solución fue dividida en diferentes paquetes con el fin de agrupar las funcionalidades y proveer de un orden coherente con el estilo arquitectónico aplicado. Los paquetes son:

controller: paquete principal que controla todo el flujo de información que se va generando según las diferentes fases del método. Es el encargado de desencadenar la ejecución del mismo y garantizar el correcto funcionamiento de la solución.

layout: paquete que gestiona los algoritmos de distribución visual de los vértices y aristas presentes en la red de coautoría. Según las restricciones detalladas en el método propuesto los algoritmos de distribución que se emplean convergen en un tiempo finito y se aplican uno a continuación del otro para obtener una mejor distribución de la red sin afectar, en gran medida, el tiempo de respuesta. Los algoritmos pueden ser configurados adaptándose a las características de la red.

rank: paquete de utilidades que gestiona gran parte de los elementos de la fase de visualización de las comunidades. Una vez calculada una estadística en la red de coautoría este paquete se encarga de clasificar los elementos de la red de acuerdo a la misma. Entre las principales funcionalidades que maneja están: la configuración del tamaño de los vértices y aristas en la red y los algoritmos para clasificar la red. Para poder clasificar los elementos de la red es necesario haber aplicado la estadística correspondiente en el paquete ***statistics.***

sparql: paquete auxiliar que maneja todo lo referente a las consultas SPARQL que son utilizadas en la fase inicial del método. Utiliza funcionalidades implementadas por la biblioteca Apache Jena.

statistics: paquete que maneja las estadísticas definidas en las fases de detección de las comunidades y en la visualización de las comunidades detectadas. Utiliza diferentes criterios, métricas y algoritmos, gestiona criterios de clasificación definidos con anterioridad: agrupamiento en comunidades, grado de los vértices, grado de entrada/salida de los vértices, intermediación, cercanía y autoridad. Permite la configuración de los algoritmos según características y criterios de los mismos, variando los resultados que se muestran con la variación de cada uno.

transform: paquete encargado de manejar las transformaciones de la fase inicial del método. Tiene todas las clases y métodos necesarios para transformar, dado un grafo RDF o SPARQL *endpoint*, los metadatos de los autores y sus publicaciones en un grafo dirigido y ponderado. Implementa las funcionalidades necesarias para calcular la frecuencia de publicación de los autores, exclusividad y la normalización de dichos indicadores. A partir de la modelación del grafo RDF se debe utilizar el método de transformación indicado para este.

Los patrones arquitectónicos y de diseño se consideran una serie de buenas prácticas basadas en la experiencia y se ha demostrado que funcionan en el desarrollo del software. Son soluciones a problemas específicos y comunes del diseño orientado a objeto. En el desarrollo de la solución propuesta fueron utilizados, entre otros, los patrones que se detallan a continuación:

Controlador (en inglés ***Controller*****):** es un patrón de tipo *General Responsibility Assignment Software Patterns* (GRASP). Consiste en asignar la responsabilidad de controlar el flujo de eventos del sistema a clases específicas. La solución posee dos clases controladoras, una de ellas controla el flujo normal de eventos y de la interacción entre los paquetes y la otra clase controla el flujo de las peticiones HTTP entre el navegador web y el prototipo funcional desarrollado.

Fachada (en inglés ***Facade*****):** conoce qué clases del subsistema son responsables de una determinada petición y delega esas peticiones de los clientes a los objetos apropiados del subsistema. Se evidencia en la interacción de los componentes de la aplicación al acceder a la API desarrollada en Java, donde se proporciona una interfaz simple que delega las acciones evitando la interacción directa con los paquetes internos de la API.

Instancia única (en inglés ***Singleton*****):** está diseñado para restringir la creación de objetos pertenecientes a una clase o el valor de un tipo a un único objeto. Su intención consiste en garantizar que una clase solo tenga una instancia y proporcionar un punto de acceso global a ella. El patrón se implementa instituyendo en nuestra clase un método que crea una instancia del objeto solo si no existe alguna. Para asegurar que la clase no puede ser instanciada nuevamente se regula el alcance del constructor (con atributos como protegido o privado) en muchas de las clases de la solución como por ejemplo en la clase principal de la API.

Inyección de Dependencias (en inglés ***Dependency Injection*****):** es un patrón de diseño orientado a objetos, en el que se suministran objetos a una clase en lugar de ser la propia clase quien cree el objeto. Se implementa por defecto en las aplicaciones desarrolladas en el marco de trabajo Groovy and Grails. Su uso se pone en práctica en la llamada a los objetos o componentes de la solución desarrollada, ejemplo, cada vez que se utiliza el objeto *render* para mostrar las plantillas GSP de Groovy and Grails.

Fábrica Abstracta (en inglés ***Abstract Factory*****):** es un patrón de diseño que consiste en utilizar una clase constructora abstracta con unos cuantos métodos definidos y otro(s) abstracto(s); el dedicado a la construcción de objetos de un subtipo de un tipo determinado. La clase abstracta tiene métodos concretos que usan algunos de los abstractos; según usemos una u otra hija de esta clase abstracta, tendremos uno u otro comportamiento. Este patrón se evidencia en la creación de espacios de trabajo, contenedores del grafo y modificadores visuales que utiliza la biblioteca Gephi Toolkit.

2.3.3. Bibliotecas

***Apache Jena*:** marco de trabajo para Java que permite la construcción de aplicaciones para la Web Semántica, ofreciendo una colección de herramientas y bibliotecas Java. Incluye la utilización del lenguaje de consultas para grafos RDF *SPARQL 1.1 Update*, elemento que lo destaca entre otros marcos de trabajo con funcionalidades afines. Entre sus principales características se encuentran:

- Posee una interfaz de programación de aplicaciones o API para RDF, la misma que soporta la creación, manipulación y consulta de grafos RDF.
- Posee un API para el manejo de ontologías.
- Permite realizar lectura y escritura de documentos en formato RDF/XML, N3 y NTriples.
- Posee un almacenamiento persistente para poder recopilar gran cantidad de tripletas RDF.

***Gephi Toolkit*:** paquete o biblioteca Java con algunos de los módulos esenciales del proyecto Gephi. Entre las principales funcionalidades que brinda están: una API para el manejo de grafos, ficheros de entrada/salida y configuraciones y algunos filtros, métricas y algoritmos de visualización para grafos. Es una biblioteca que no tiene dependencias de ninguna otra y puede ser incorporada para el desarrollo de aplicaciones reales. Sus funcionalidades pueden ser redefinidas o extendidas según las necesidades de los desarrolladores.

***SigmaJS*:** biblioteca JavaScript para la visualización de grafos. Está diseñada usando un motor que permite la personalización y un desarrollo altamente interactivo en el desarrollo de herramientas de visualización de grafos en la web. Se destaca por las siguientes características:

- Renderizado personalizable: se puede usar *canvas* o WebGL en la construcción del *render* o se puede desarrollar uno propio. Provee diferentes vías de personalización según las necesidades de los desarrolladores.
- Orientado a la interactividad: posibilita la gestión de eventos con el puntero, acercar/alejar la vista, mover los elementos visuales dentro del grafo, posicionar la pantalla en determinados puntos, entre muchos otros.
- Modelo extensible para grafos: brinda la posibilidad de desarrollar algoritmos para el trabajo con los grafos. Haciendo al modelo personalizable y extensible mediante *plugins* para el manejo de archivos de descripción de grafos, exportar en diferentes formatos, entre muchos otros.

- Compatibilidad: soportado por los navegadores Mozilla Firefox, Google Chrome, Internet Explorer y Opera que brindan compatibilidad con *canvas* y *WebGL* para una mejor velocidad y mayor rendimiento.

2.3.4. Estándares y Tecnologías

***RDF (Resource Description Framework)*:** modelo de datos que utiliza XML, Ntriples, Turtle, entre otros, como lenguaje para serializar los datos y metadatos de los recursos de la web. Permite la interoperabilidad entre aplicaciones que intercambian información comprensible por la página Web, para proporcionar una infraestructura que soporte actividades de metadatos. Dicho modelo de datos constituye la fuente principal de información que almacena los metadatos bibliográficos necesarios para el desarrollo del prototipo funcional. Al ser un estándar a nivel mundial, posibilita la utilización de cualquier grafo RDF que almacene metadatos bibliográficos como fuente de datos del prototipo funcional.

***SPARQL*:** lenguaje de consulta para RDF. Se puede utilizar para expresar consultas que permiten interrogar diversas fuentes de datos, si los datos se almacenan de forma nativa como RDF o son definidos mediante vistas RDF a través de algún sistema *middleware*. SPARQL también soporta la ampliación o restricciones del ámbito de las consultas indicando los grafos sobre los que se opera. Es aplicado como el lenguaje para hacer obtener información de la fuente de datos principal del prototipo funcional. A partir del uso de la biblioteca Apache Jena este lenguaje de consulta constituye la herramienta esencial para la interoperabilidad con las fuentes de datos que puede usar el prototipo funcional. Permite acceder a un grafo RDF a partir de un archivo físico o a través de un *SPARQL endpoint.*

***WebGL*:** es una especificación estándar en desarrollo, para mostrar gráficos en 3D en navegadores web. El WebGL permite mostrar gráficos en 3D acelerados por hardware (GPU) en páginas Web, sin la necesidad de *plugins* en cualquier plataforma que soporte OpenGL 2.0 u OpenGL ES 2.0. Técnicamente, es un API para JavaScript que permite usar la implementación nativa de OpenGL ES 2.0 que será incorporada en los navegadores. El uso de WebGL en el prototipo funcional provee un mejor rendimiento que el uso del elemento *canvas* de HTML5. El mismo posibilita el renderizado de las comunidades en tiempo real, mediante una correcta gestión y optimización del código a partir del lenguaje JavaScript.

***GEXF (Graph Exchange XML Format)*:** lenguaje para describir estructura de redes complejas, sus datos asociados y su dinamismo. Iniciado en 2007 en el desarrollo del proyecto *Gephi*,

profundamente involucrado en la descripción e intercambio de las redes o grafos. Las especificaciones de *GEXF* han ido madurando desde entonces convirtiéndose en un lenguaje extensible y abierto para el uso en diversas aplicaciones de dominio específico. Las características que soporta son comparadas con otros lenguajes de descripción de redes en la Figura VIII. *GEXF* facilita el manejo de los diferentes atributos y características que son modelados en el método propuesto. Su utilización en el prototipo funcional junto a la biblioteca Gephi Toolkit permiten explotar al máximo características correspondientes a la personalización de la red, manipulación de los elementos y características presentes en el grafo para lograr una mejor visualización. Además *GEXF* puede ser utilizado como fuente de datos para la renderización en los navegadores web.

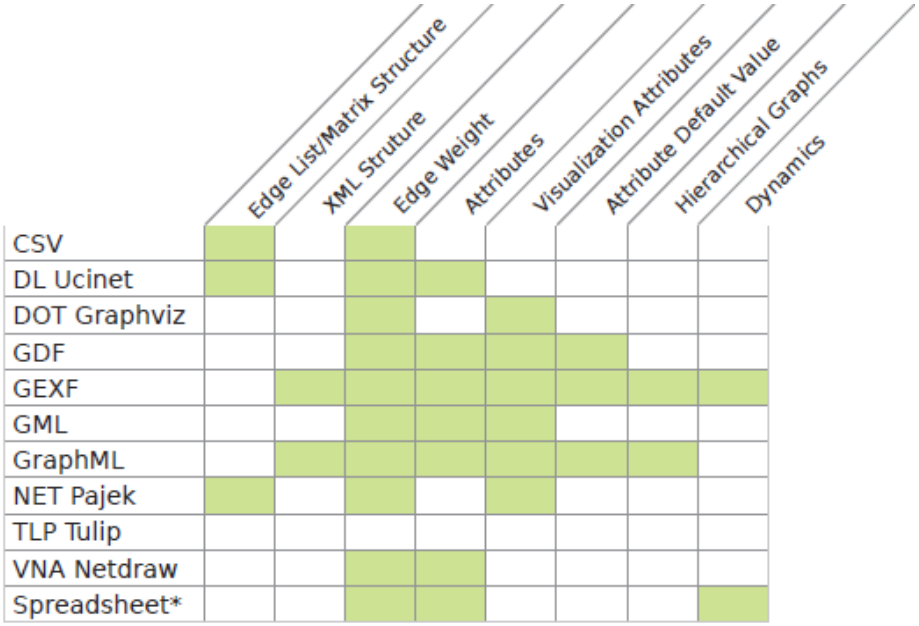

Figura VIII: *Comparación de lenguajes para describir estructuras de grafos. Fuente: http://gephi.github.io/features/*

***Groovy and Grails*:** marco de trabajo web para la plataforma Java que se basa en el lenguaje dinámico *Groovy*. *Groovy and Grails* es la base para el desarrollo del prototipo funcional. Facilita la interacción con la API desarrollada en Java que maneja las principales funcionalidades de la solución. A través de este lenguaje de programación se interactúa de una forma ágil y sencilla con las propiedades y métodos dinámicos en los objetos de la aplicación. Las bibliotecas de *Spring* que implementa *Groovy and Grails* para los flujos de trabajo e inyección de dependencias son una de las características que influyó en un menor tiempo de desarrollo de la solución. La adaptabilidad y rendimiento que brindan las plantillas GSP en la composición de las vistas son factores que favorecieron el rendimiento del prototipo funcional y su adaptabilidad a los diferentes navegadores web.

2.3.5. Algoritmos

En el desarrollo de las fases del método y atendiendo a las restricciones presentes en cada una se emplearon diferentes algoritmos. Algunos de estos fueron desarrollados con anterioridad y otros modificados o adaptados para obtener mejores resultados de acuerdo con el método definido. A continuación se describen los algoritmos utilizados que intervienen en cada una de las fases del método propuesto.

Algoritmos de la fase: Modelación de la red de coautoría

TransformarRDFaRC: algoritmo desarrollado modelar una red de coautoría ponderada y dirigida a partir de un grafo RDF. La implementación del método sigue las restricciones mencionadas en el método propuesto (ver sección **2.2 Método propuesto**). El algoritmo tiene como entrada un grafo RDF de donde extrae la información de todos los autores, por cada publicación del autor se calcula la exclusividad (Algoritmo 2) y la frecuencia de publicación con sus coautores (Algoritmo 3) y se normaliza dicho valor entre 0 y 1 (Algoritmo 4). Genera las relaciones correspondientes entre los autores y devuelve una red de coautoría ponderada y dirigida.

Algoritmo 1 TransformarRDFaRC

Entrada: GrafoRDF $G = \{J^G, R^G, A^G\}$
Salida: Red de coautoría $RC = \{N^{RC}, E^{RC}\}$

```
1:  G <- GrafoDirigido()
2:  N^RC <- Lista()
3:  E^RC <- Lista()
4:  Para todo A_i^G ∈ G hacer
5:      N_i^RC <- ExtraerDatos(A_i^G)
6:      Adicionar(N_i^RC, N^RC)
7:  Fin Para
8:  EP <- Lista()
9:  Para todo R_i^G ∈ G hacer
10:     P_i <- ExtraerDatos(R_i^G)
11:     EP_i <- CalcularExclusividad(P_i)
12:     Adicionar(EP_i , EP )
13: Fin Para
14: Para todo N_i^RC ∈ RC hacer
15:     P^Ni <- ExtraerPublicaciones(N_i^RC, R^G)
16:     FP^Ni <- Mapa()
17:     Para todo P_j^Ni ∈ P^Ni hacer
18:         CalcularFrecuencia(P_j^Ni, EP, N_i^RC, FP^Ni)
19:     Fin Para
20:     FT <- SumarFrecuencias(FP^Ni)
```

```
21:      Para todo FP_j^{N_i} ∈ FP^{N_i} hacer
22:          WP <- NormalizarValores(FP_j^{N_i}, FT)
23:          E_i^{RC} <- CrearArista(N_i^{RC}, WP)
24:          Adicionar(E_i^{RC}, E^{RC})
25:      Fin Para
26:  Fin Para
27:  AdicionarAristas(E^{RC}, RC)
28:  Retornar RC
```

Algoritmo 2 CalcularExclusividad

Entrada: Publicación P
Salida: Valor de la exclusividad E en la publicación P

```
1:  n <- CantidadAutores(P)
2:  E <- 1 / (n − 1)
3:  Retornar E
```

Algoritmo 3 CalcularFrecuencia

Entrada: Publicación P, Listado de exclusividad EP, Datos autor N, Listado de frecuencias FP

```
1:  CA <- ListadoCoautores(P_j, N_i)
2:  Para todo CA_i ∈ CA hacer
3:      F <- ObtenerFrecuencia(CA_i, FP)
4:      Si F = 0 hacer
5:          F <- ObtenerExclusividad(P, EP)
6:          CrearFrecuencia(F, CA_i, FP)
7:      Sino hacer
8:          F <- F + ObtenerExclusividad(P, EP)
9:          ActualizarFrecuencia(F, CA_i, FP)
10:     Fin Si
11: Fin Para
```

Algoritmo 4 NormalizarValores

Entrada: Frecuencia de publicación con el autor FP, Frecuencia total de publicaciones FT
Salida: Frecuencia normalizada con el autor WP

```
1:  fp <- ObtenerValor(FP)
2:  N <- OptenerNodo(FP)
3:  wp <- fp / FT
4:  WP <- Mapa(N, wp)
5:  Retornar WP
```

Algoritmos de la fase: Detección de las comunidades

Fast Unfolding (Blondel et al. 2008): este algoritmo se basa en el concepto de modularidad. Tiene dos fases, en la primera todos los vértices empiezan perteneciendo a comunidades distintas. Cada vértice i se prueba con cada vértice adyacente j introduciéndolo en su comunidad. Si la fórmula para el cálculo de la modularidad toma valor positivo, se asigna a la comunidad que alcanzó un valor mayor. Después de varias iteraciones los valores convergen. La segunda fase consiste en crear un nuevo grafo en el que cada vértice es una comunidad encontrada anteriormente, y aplicar de nuevo el paso anterior para intentar unir comunidades hasta que los valore converjan tras varias iteraciones. El algoritmo se utiliza para la detección de las comunidades de coautores a partir de la red de coautoría que se modelo en la primera fase del método.

Algoritmos de la fase: Visualización de las comunidades detectadas

Betweenness centrality (Brandes 2001)***:*** algoritmo desarrollado para obtener el camino más corto de un grafo. Establece que los vértices conectados tienen distancia 1. El algoritmo se basa en que el diámetro entre dos vértices cualquiera de la red es la distancia de grafo más larga entre ellos (es decir, cuán distantes están los dos vértices más alejados). Calcula criterios como la intermediación, cercanía y centralidad de los vértices en la red, brindando el indicador de centralidad de los autores en la red de coautoría a partir del criterio de distancia de grafo.

HITS (Kleinberg 1999): del acrónimo del inglés *Hypertext Induced Topic Selection* diseñado en un principio para valorar y de paso clasificar la importancia de una página Web. Ha sido adaptado para su uso como criterio de medida en redes complejas. Computa dos valores separados para cada vértice. El primer valor llamado *authority* mide cuán valiosa es la información almacenada en ese vértice. El segundo valor llamado *hub* mide la calidad de las aristas de ese vértice. Cuanto menor sea el valor del criterio de parada, más tiempo tomará la convergencia del algoritmo. Se utiliza para medir el indicador de autoridad en la red atendiendo al criterio definido por *HITS*.

PageRank (Brin y Page 1998): Clasifica los vértices de una red atendiendo a la frecuencia con la que un usuario siguiendo las aristas llega al vértice de forma no aleatoria. Permite establecer una probabilidad para simular aleatoriamente el reinicio de la navegación por las aristas. Puede utilizar el peso de las aristas en su recorrido. Cuanto menor sea el valor del criterio de parada, más tiempo tomará la convergencia del algoritmo. Se utiliza para medir el indicador de autoridad en la red atendiendo al criterio definido por *PageRank*.

***AuthorRank*:** modificación del algoritmo *PageRank* que tiene en cuenta el peso de las aristas en un grafo. El peso de las aristas expresa cómo de fuerte son las relaciones entre los autores que conecta. Este peso es usado para determinar el *AuthorRank*, determinado por el respaldo que transfieren los pesos de las aristas que conectan a los autores en la red. Se utiliza para medir el indicador de autoridad de los autores en la red atendiendo al criterio de la frecuencia y exclusividad modelado y reflejado en el peso de las aristas. Para el cálculo del mismo se define que:

- $AR(i) = (1 - d) + \ d \sum_{j=0}^{n} AR(j) \ * \ w_{i,j}$

Donde:

- $AR(i)$ es el *AuthorRank* del autor i.
- d es un factor de amortiguación que tiene un valor entre 0 y 1.
- $AR(j)$ son los valores de *AuthorRank* que tienen cada uno de los autores que se relacionan con i.
- $W_{i,j}$ es el peso de coautoría entre los autores.

Algoritmo 5 *AuthorRank*

Entrada: Red de coautoría $RC = \{N, E\}$
Salida: Lista del *AuthorRank* (AR)

```
 1:  d <- FactorAmortiguacion();
 2:  AR <- Lista()
 3:  TAR <- Lista()
 4:  W <- Lista()
 5: Para todo N_i ∈ RC hacer
 6:      AR_i <- 0.1
 7:      Adicionar(AR_i, AR)
 8:  Fin Para
 9:  Hacer
10:      parada <- verdadero
11:      Iterador <- IteradorNodos(RC)
12:      Mientras Iterador > 0 hacer
13:          N_i <- ObtenerSiguiente(Iterador)
14:          TAR_i <- 0
15:          IteradorE ->IteradorEnlaces(N_i)
16:          Mientras IteradorE > 0 hacer
17:              E_k <- ObtenerSiguiente(IteradorE)
18:              N_j <- ObtenerNodo(E_k, N_i)
19:              W_i,j <- ExtraerPeso(N_i, N_j)
20:              TAR_i <- TAR_i + (1 – d) + d * AR_j * W_i,j
21:              IteradorE <- IteradorE – 1
22:          Fin Mientras
```

```
23:          Si (TAR_i - AR_i) / AR_i ≥ CriterioParada()
24:              parada <- falso
25:          Fin Si
26:          Iterador <- Iterador – 1
27:      Fin Mientras
28:      AR <- TAR
29:      TAR <- Lista()
30:  Mientras parada ≠ verdadero Fin Hacer
31:  Retornar AR
```

YifanHu Multilevel (Hu 2011)**:** algoritmo de distribución visual en grafos de gran eficiencia y calidad. Combina una aproximación multinivel que aprovecha efectivamente los mínimos locales calculados con Barnes y Hut aproximando mediante fuerzas de corto y largo alcance los vértices en la red. Usa un esquema de enfriamiento adaptativo y un modelo general de repulsión a partir de los algoritmos de fuerza directa. Se aplica como primera aproximación visual al resultado requerido en la representación de las comunidades en la red.

Noverlap (Li et al. 2012): algoritmo que evita que los vértices circulares se superpongan. Está optimizado para grafos grandes. Elije la velocidad frente a la precisión y establece más espacio alrededor de grandes vértices. En el prototipo funcional dicho algoritmo se ejecuta posteriormente a la aplicación del *YifanHu Multilevel* para desagregar aquellos vértices que quedaron solapados.

ForceAtlas2 (Jacomy et al. 2014): es un algoritmo de vector de fuerza propuesto en el desarrollo de la aplicación Gephi. Se destaca por su simplicidad y legibilidad de las redes que ayudan en la visualización. El modelo de energía y la forma de optimizar la "velocidad versus la precisión" lo hace una aproximación única de rápida convergencia pero sin llegar a ser determinista sobre la misma red. Aplica una técnica de repulsión entre los vértices de la red según el grado de ellos que evita el solapamiento de los grandes subgrupos sobre los pequeños subgrupos en la red. *ForceAtlas2* se implementa en el prototipo funcional para que el usuario pueda visualizar y distribuir en tiempo real la visualización que se logra con los algoritmos anteriormente mencionados.

2.4. Conclusiones parciales

El método de solución para la detección de comunidades a partir de una red de coautoría en grafos RDF se basa en tres fases, siguiendo un enfoque basado en filtros y tuberías. Se detectan comunidades a partir de una red de coautoría en grafos RDF, solucionándose los problemas de identificación de las comunidades y cuantificación de las relaciones de colaboración en las RC con diferentes criterios de medida.

La arquitectura seleccionada a partir del método propuesto define el desarrollo del prototipo funcional. Agrupa las funcionalidades en paquetes y evidencia el flujo de información planteado en las restricciones del método. Los estándares y tecnologías constituyeron los pilares fundamentales en el modelado de la solución e implementación del prototipo funcional. Se aprovechan las principales características y beneficios de ellos, lográndose transformar las restricciones definidas en el método propuesto a las correspondientes funcionalidades del prototipo funcional.

CAPÍTULO 3. VALIDACIÓN DE LA SOLUCIÓN PROPUESTA

3.1. Introducción

En este capítulo se realizará la validación del método para la detección de comunidades a partir de RC en grafos RDF, aplicando un caso de estudio con el objetivo de identificar y cuantificar las relaciones de colaboración existentes. El capítulo posee una única sección donde se describe paso a paso el desarrollo de las fases de la propuesta de solución para posteriormente realizar un análisis de los resultados obtenidos.

3.2. Caso de estudio

Un caso de estudio ofrece importantes resultados e información que no puede ser encontrada por medio de los métodos cuantitativos y cualitativos tradicionales. Su aplicación posibilita ir observando y comparando los resultados que se esperan alcanzar en el desarrollo de una investigación de una forma real y con resultados medibles (Monge 2010).

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. Dicho caso será desarrollado utilizando el 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 en las que se desarrolla el caso de estudio son de 4 GB de memoria RAM y un procesador Intel Core i3 380 a 2.33GHz. A continuación se detallan los resultados obtenidos en cada una de las fases del método propuesto. Se realizarán observaciones en diferentes momentos que permitan comparar resultados y arribar a conclusiones.

3.2.1. Fase 1: Modelación de la red de coautoría

En la primera fase es necesario tener un grafo RDF con los metadatos de las publicaciones de las revistas científicas. Para el desarrollo del caso de estudio la muestra es un grafo RDF que contiene los metadatos de las publicaciones en las revistas científicas cubanas como se puede apreciar en la Tabla IV.

Tabla IV: *Revistas Científicas Cubanas usadas en el caso de estudio*

Número	Nombre
1	Revista Cubana de Información en Ciencias de la Salud
2	Revista Cubana de Ingeniería
3	Revista de Ingeniería Industrial
4	Revista Cubana de Ciencias Informáticas
5	Ingeniería Electrónica, Automática y Comunicaciones
6	Serie Científica de la UCI
7	Avanzada Científica
8	Ingeniería Energética

Observación 1: El grafo RDF se muestra en la Figura IX con un total de 51329 tripletas y estructurado mediante sujeto – predicado – objeto tiene un espacio de almacenamiento de 8 MB. Almacena todos los metadatos de las publicaciones existentes en dichas revistas hasta el año 2014. Modela las relaciones entre las tripletas a partir de URIs únicas para los diferentes recursos, pero no refleja una estructura identificable de las relaciones de colaboración basadas en la coautoría ni la ponderación de las mismas. Contiene además metadatos adicionales de las revistas y de los artículos que no aportan información significativa para el problema de la detección de las comunidades.

```
<rdf:Description rdf:about="author/242">
  <fabio:has_creator rdf:resource="article/126"/>
  <foaf:name>Rolando Bonal Cáceres</foaf:name>
  <rdfs:label>Rolando Bonal Cáceres</rdfs:label>
  <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
</rdf:Description>
<rdf:Description rdf:about="author/688">
  <fabio:has_creator rdf:resource="article/440"/>
  <swrc:affiliation>Universidad de las Ciencias Informáticas</swrc:affiliation>
  <foaf:name>Ernesto Ortiz Muñoz</foaf:name>
  <rdfs:label>Ernesto Ortiz Muñoz</rdfs:label>
  <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
</rdf:Description>
<rdf:Description rdf:about="author/6975">
  <fabio:has_creator rdf:resource="article/3369"/>
  <swrc:affiliation>Especialista segundo Grado en Anatomía Patológica. Profesor Auxiliar. Hospital General Docente "Abel
  Santamaría Cuadrado", Pinar del Río.</swrc:affiliation>
  <foaf:name>Miguel Angel Pérez Herrera</foaf:name>
  <rdfs:label>Miguel Angel Pérez Herrera</rdfs:label>
  <rdf:type rdf:resource="http://xmlns.com/foaf/0.1/Person"/>
</rdf:Description>
```

Figura IX: *Ejemplo de metadatos en el grafo RDF*

Se configura el prototipo funcional para que utilice como fuente de datos el grafo RDF anteriormente mencionado. Se ejecuta el algoritmo **TransformarRDFaRC** descrito en la sección **2.3.5. Algoritmos**, se genera como salida un archivo en formato GEXF con un tiempo total de ejecución de la fase 1 de 25806 milisegundos como se evidencia en la captura de la consola de ejecución del IntelliJ IDEA (ver Figura X).

Figura X: *Ejecución de la Fase 1: Modelación de la red de coautoría*

Observación 2: El archivo GEXF generado con 5203 vértices y 24498 aristas y un espacio de almacenamiento de 5.3 MB describe una red de coautoría ponderada y dirigida como se evidencia en la Figura XI. Los vértices están etiquetados con el nombre y apellidos de los autores y la cantidad de publicaciones que poseen. Del mismo modo los vértices tienen una posición predefinida en el plano de las coordenadas y las ordenadas, un tamaño por defecto, un color estándar y no presentan atributos adicionales que los caractericen dentro de la red. Las aristas ya cuentan con la ponderación deseada para cuantificar las relaciones de coautoría y unen a los vértices a través de identificadores únicos, los cuales corresponden a la descripción del recurso del grafo RDF original.

```
<graph defaultedgetype="directed" mode="static">
    <nodes>
        <node id="http://localhost:2020/author/429" label="Yusniel Hidalgo Delgado (3 pub)">
            <attvalues></attvalues>
            <viz:size value="1.0"></viz:size>
            <viz:position x="241.45618" y="270.40875" z="0.0"></viz:position>
            <viz:color r="153" g="153" b="153"></viz:color>
        </node>
        <node id="http://localhost:2020/author/688" label="Ernesto Ortiz Muñoz (1 pub)">
            <attvalues></attvalues>
            <viz:size value="1.0"></viz:size>
            <viz:position x="-74.157135" y="-360.18652" z="0.0"></viz:position>
            <viz:color r="153" g="153" b="153"></viz:color>
        </node>
    </nodes>
    <edges>
        <edge source="http://localhost:2020/author/429" target="http://localhost:2020/author/688" weight="0.25">
            <attvalues></attvalues>
        </edge>
        <edge source="http://localhost:2020/author/688" target="http://localhost:2020/author/429" weight="0.5">
            <attvalues></attvalues>
        </edge>
    </edges>
</graph>
```

Figura XI: *Ejemplo de archivo en formato GEXF generado en Fase 1: Modelación de la red de coautoría*

3.2.2. Fase 2: Detección de las comunidades

En esta fase corresponde comprobar la ejecución del algoritmo ***Fast Unfolding*** propuesto en la sección **2.3.5. Algoritmos** teniendo en cuenta la función de calidad *Modularity* que utiliza. Se ejecutará el mismo con diferentes configuraciones para medir si la direccionalidad y ponderación de las relaciones de coautoría tiene o no influencia en el proceso de detección de las comunidades presentes en la red. Se utiliza como valor de control el resultado de la ejecución de la fase 2 en el prototipo funcional configurado con los valores de la Tabla V. Luego se compara con los resultados de la herramienta Gephi utilizando otras configuraciones a partir del mismo archivo GEXF generado anteriormente (ver Tabla VI).

Tabla V: *Configuración del algoritmo Fast Unfolding*

Parámetro	Valor	Descripción
Aleatorio	Verdadero	Produce una mejor descomposición de las comunidades pero aumenta el tiempo de cómputo.
Utilizar Pesos	Verdadero	Tiene en cuenta el peso de las aristas para la detección de las comunidades.

Resolución	1.0	Valores pequeños agrupa en comunidades más pequeñas y valores mayores agrupa en comunidades más grandes. El valor por defecto es 1.

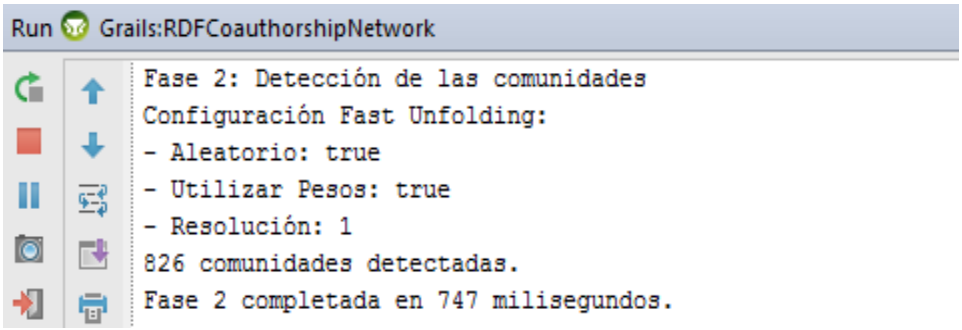

Figura XII: *Ejecución de la Fase 2: Detección de las comunidades*

Terminada la ejecución del algoritmo como se muestra en la Figura XII, fueron detectadas 826 comunidades. Cada comunidad fue agrupada atendiendo al criterio de modularidad, se les asignó de manera consecutiva un número o clase de modularidad según propone el algoritmo. A los vértices se les asignó un color aleatorio, siendo igual para todos los que pertenecen a una misma clase de modularidad. La fase se completó en un tiempo total de 747 milisegundos.

Tabla VI: *Ejecución del algoritmo Fast Unfolding cambiando su configuración*

Parámetro	**Configuraciones**					
	Defecto	**1**	**2**	**3**	**4**	**5**
Aleatorio	Verdadero	Falso	Verdadero	Falso	Verdadero	Verdadero
Utilizar Pesos	Verdadero	Verdadero	Falso	Falso	Verdadero	Verdadero
Resolución	1.0	1.0	1.0	1.0	0.1	10.0
Comunidades detectadas	826	825	817	815	922	795

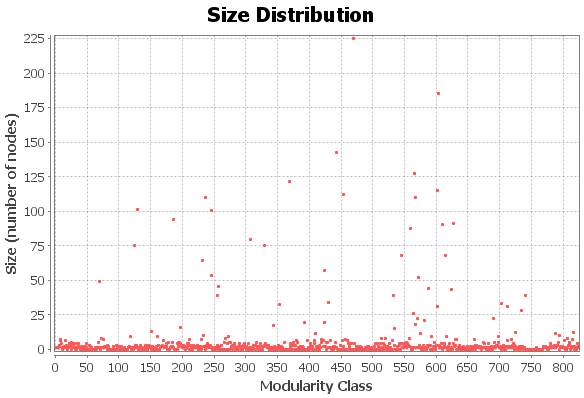

Figura XIII: *Fast Unfolding C. Defecto*

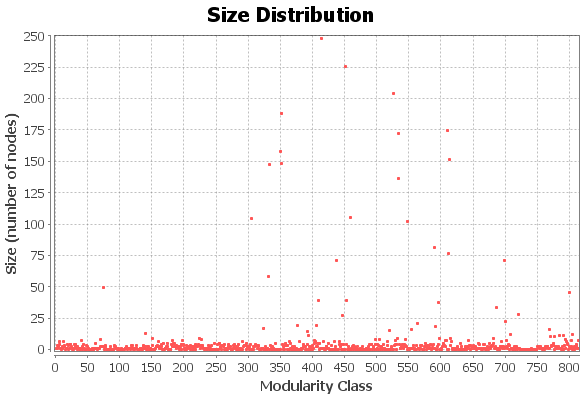

Figura XIV: *Fast Unfolding C. 3*

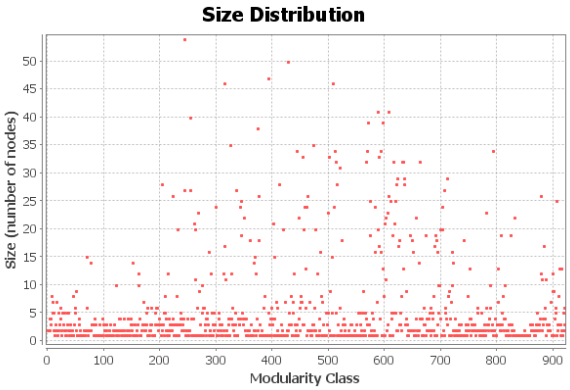

Figura XV: *Fast Unfolding C. 4*

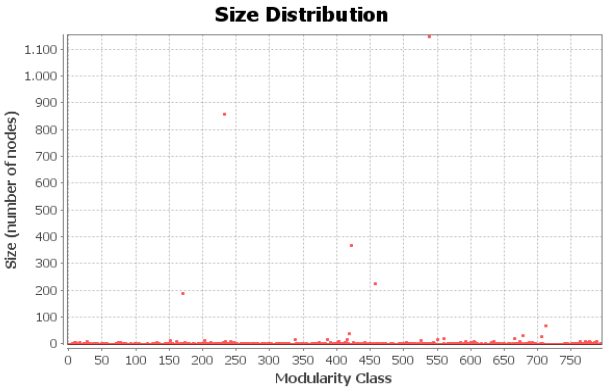

Figura XVI: *Fast Unfolding C. 5*

Observación 3: La ejecución del algoritmo ***Fast Unfolding*** con las configuraciones establecidas en la Tabla VI y los reportes mostrados en las Figuras XIII hasta la XVI (distribución de la cantidad de vértices agrupados en las diferentes comunidades), permite corroborar el correcto funcionamiento del algoritmo y arribar a las siguientes conclusiones:

- El parámetro aleatorio no tiene gran influencia en la detección de comunidades, el tiempo de cómputo que aumenta o disminuye según su activación o desactivación no supera los 100 milisegundos para la cantidad de datos de entrada.
- El uso del peso de las aristas influye en la detección de comunidades pero no de manera significativa ya que solo varía la cantidad en aproximadamente 10 comunidades para un total del 98%.
- La variación de la resolución tiene un mayor peso en los resultados del algoritmo llegando a variar según las diferentes configuraciones entre 0.1, 1.0 y 10.0 un total de 127 comunidades. Existe gran variación entre el número de vértices presentes por comunidades

que oscilan entre 1 y 25 normalmente, en valores menores de resolución llegan hasta 60 por comunidad y superan los 1100 vértices para la resolución 10.0.

3.2.3. Fase 3: Visualización de las comunidades detectadas

Una vez detectadas e identificadas las comunidades y siguiendo las fase del método corresponde visualizar las mismas. Para visualizar las comunidades, a partir del prototipo funcional en esta fase, se aplican dos algoritmos de *layout* ***YifanHu Multilevel*** y ***Noverlap*** (ver sección **2.3.5. Algoritmos**). El primero para lograr una mayor distribución visual de los vértices y aristas en la red y el segundo para evitar el solapamiento de los vértices una vez que son aplicados los criterios de clasificación a la red (ver sección **2.2.3. Visualización de las comunidades detectadas**). Los parámetros de configuración de los algoritmos se muestran en la Tabla VII y VIII respectivamente.

Tabla VII: *Configuración del algoritmo YifanHu Multilevel*

Parámetro	Valor	Descripción
Nivel máximo de Quadtree	10	Define un máximo de descomposición recursiva de los vértices dentro del espacio de la red. Valores mayores mejoran la precisión del algoritmo.
Theta	1.2	Criterio de parada para el cálculo de los mínimos locales aplicando Barnes-Hut. Valores menores mejoran la precisión del algoritmo.
Tamaño mínimo del nivel	5	Cantidad mínima de vértices que cada nivel debe tener. Valores mayores implican menos niveles.
Radio mínimo de aspereza	0.75	Tamaño mínimo relativo que debe existir entre dos niveles. Valores menores implican menos niveles.
Radio del paso	0.97	Valor utilizado para actualizar el tamaño de paso entre las iteraciones del algoritmo.
Distancia óptima	200.0	Establece la distancia elástica natural de los elementos en la red. Valores mayores hacen que los vértices queden más alejados.

Tabla VIII: *Configuración del algoritmo Noverlap*

Parámetro	Valor	Descripción
Velocidad	3.0	Establece la velocidad de ejecución del algoritmo.
Radio	1.2	Define el radio mínimo de los vértices de la red.
Margen	10.0	Establece el espacio que separa los elementos dentro de la red.

Al terminar de ejecutarse la fase 3 del método con un tiempo de 152854 milisegundos (ver Figura XVII) se tienen todas las condiciones necesarias para aplicar los criterios de clasificación a la red. La salida de la fase generó el archivo en formato GEXF con un espacio de almacenamiento de 6.5 MB (ver Figura XVIII). En el archivo los vértices poseen atributos tales como la clase de modularidad a la que pertenecen, los grados de entrada/salida y grado total correspondiente a la cantidad de vértices con los que se relacionan, el tamaño según el criterio de medida por grados y un color de acuerdo a la clase de modularidad a la que están asociados. Las aristas se mantienen con el mismo nivel de información que tenían al finalizar la fase 1.

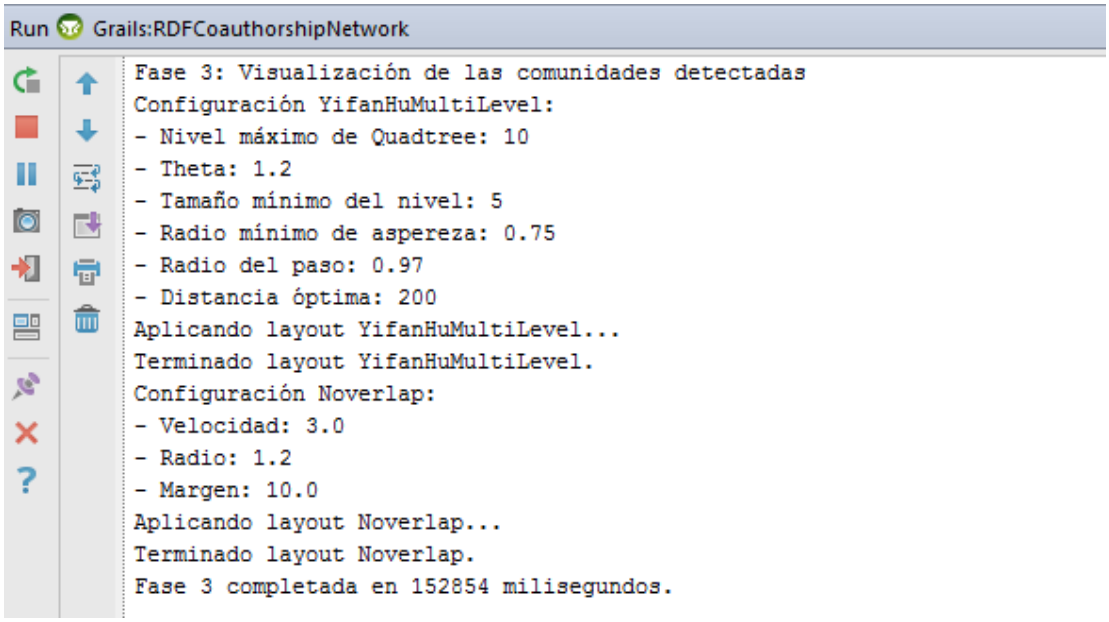

Figura XVII: *Ejecución de la Fase 3: Visualización de las comunidades detectadas*

```
<node id="http://localhost:2020/author/688" label="Ernesto Ortiz Muñoz (1 pub)">
    <attvalues>
        <attvalue for="modularity_class" value="679"></attvalue>
        <attvalue for="indegree" value="2"></attvalue>
        <attvalue for="outdegree" value="2"></attvalue>
        <attvalue for="degree" value="4"></attvalue>
    </attvalues>
    <viz:size value="6.40625"></viz:size>
    <viz:position x="-937.23456" y="1443.6401" z="0.0"></viz:position>
    <viz:color r="23" g="128" b="154"></viz:color>
</node>
<node id="http://localhost:2020/author/429" label="Yusniel Hidalgo Delgado (3 pub)">
    <attvalues>
        <attvalue for="modularity_class" value="679"></attvalue>
        <attvalue for="indegree" value="3"></attvalue>
        <attvalue for="outdegree" value="3"></attvalue>
        <attvalue for="degree" value="6"></attvalue>
    </attvalues>
    <viz:size value="7.109375"></viz:size>
    <viz:position x="-966.36615" y="1515.4722" z="0.0"></viz:position>
    <viz:color r="23" g="128" b="154"></viz:color>
</node>
```

Figura XVIII: *Ejemplo de archivo en formato GEXF generado en Fase 3: Visualización de las comunidades detectadas*

Observación 4: Los resultados de la ejecución del prototipo funcional se pueden evidenciar en un antes y un después del inicio y culminación de la fase 3. Al inicio (ver Figura XIX) los elementos de la red se encontraban distribuidos de manera aleatoria, no se identificaba la agrupación por comunidades, salvo por los colores que tienen asignados de acuerdo a las clases de modularidad a que pertenecen. Los vértices son de igual tamaño y no evidencian diferencia alguna entre ellos. Posterior a la ejecución de la fase 3 (ver Figura XX) ya se puede observar una mejor distribución visual de las comunidades de coautores presentes en la red. Los vértices están organizados por comunidades, las comunidades más pequeñas al exterior y las más densas en el centro. También los vértices no se solapan unos con otros y varían su tamaño de acuerdo al grado que ostentan. Las aristas toman el color de acuerdo a las comunidades a las que relacionan.

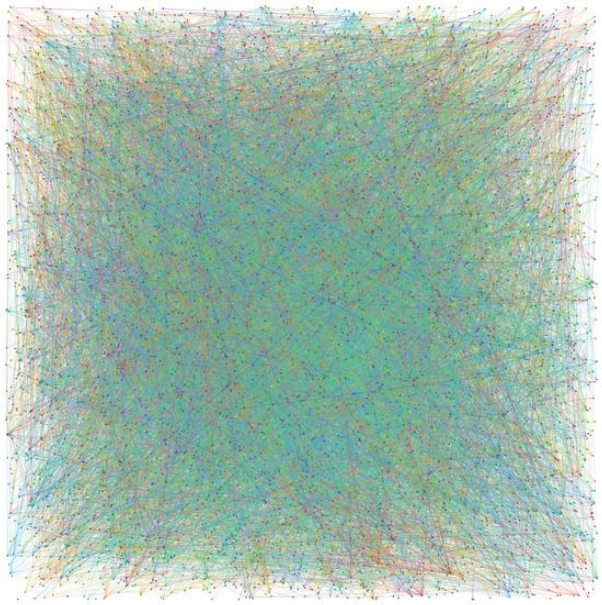

Figura XIX: *Vista de la red al inicio de la Fase 3*

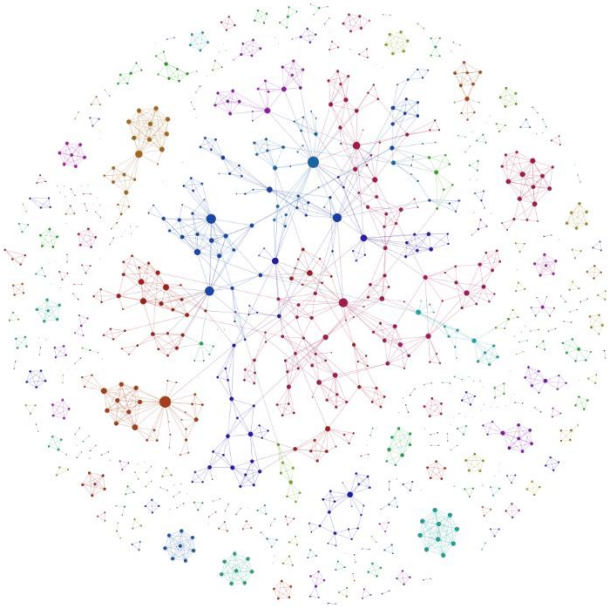

Figura XX: *Vista de la red al terminar de la Fase 3*

Observación 5: La ejecución de la fase 3 es la que mayor impacto tiene atendiendo al tiempo de ejecución del método en el prototipo funcional. El mismo está determinado de acuerdo a la cantidad de vértices y aristas presentes en la red, las configuraciones establecidas para los algoritmos de *layout* y las capacidades de cómputo del entorno en que es ejecutado. ***YifanHu Multilevel*** tiene mayor impacto en el proceso de distribuir visualmente los elementos de la red, por lo que su correcta configuración según la cantidad de datos de entrada afecta directamente el resultado final de la fase. ***Noverlap,*** por su parte, suplementa el proceso de visualización, evita el solapamiento de los vértices que ***YifanHu Multilevel*** no separa de forma óptima. En el prototipo funcional se brinda la posibilidad de mejorar la distribución visual ejecutando el algoritmo ***ForceAtlas2*** con sus respectivas configuraciones. ***ForceAtlas2*** permite al usuario que interactúe con una distribución visual inicial y detener el algoritmo según va organizando los elementos por comunidades dentro de la red (ver Figura XXI).

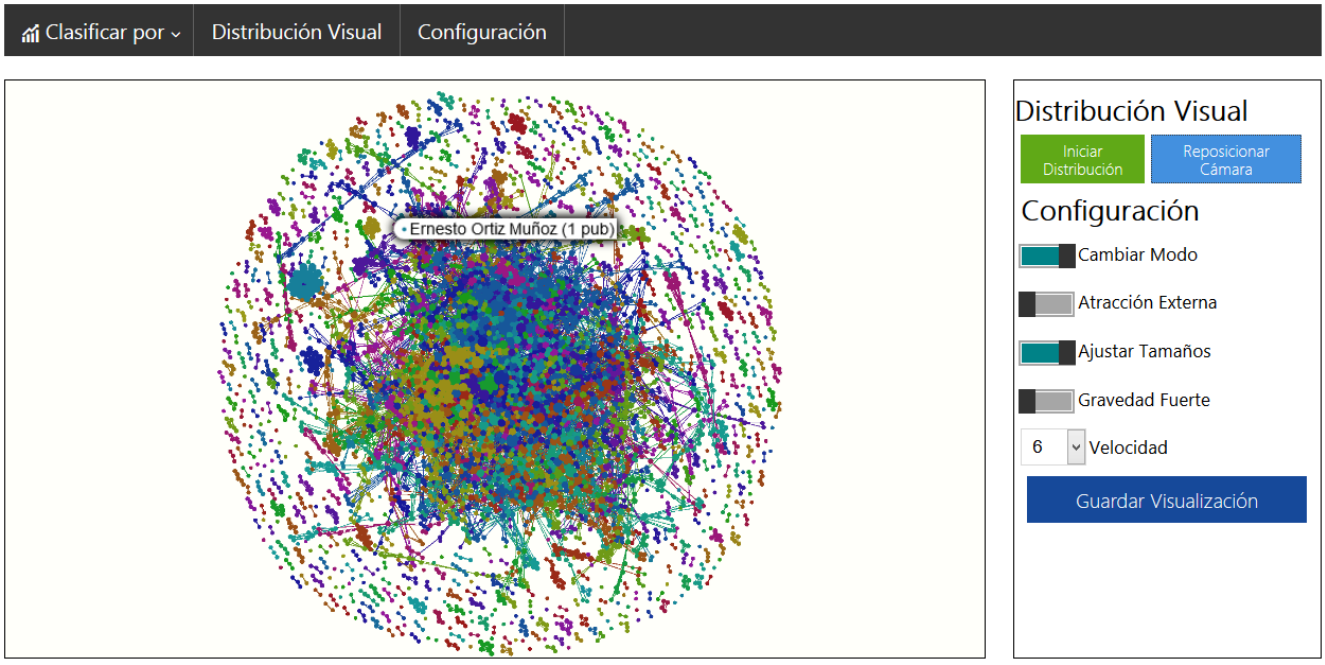

Figura XXI: *Distribución visual utilizando ForceAtlas2*

A continuación se muestran los resultados de la aplicación de algunos de los criterios de clasificación definidos en el método en la sección **2.2.3. Visualización de las comunidades detectadas**. Se comparan los resultados de la ejecución de los algoritmos correspondientes a cada clasificación. Los resultados se ordenan en las siguientes tablas cargando los archivos generados por el prototipo en el Laboratorio de Datos de la herramienta Gephi.

Tabla IX: *Criterio de clasificación: Intermediación*

Nombre y apellidos	Comunidad	Coautores	Intermediación
Julio López Arguelles (7 pub)	809	14	0.13646157026961536
Ariel E. Delgado Rodríguez (7 pub)	815	18	0.13630459133693118
Heydi Flores Podadera (2 pub)	815	8	0.13601027986736353
Osbert Rodríguez Miranda (2 pub)	815	4	0.1358568181184806
Erélido Hernández Valero (2 pub)	815	2	0.13580536851480887
Ángel Julio Romero Cabrera (10 pub)	809	18	0.11845569194476427
Raúl López Fernández (16 pub)	807	48	0.06693892039140856
Mikhail Benet Rodríguez (21 pub)	807	64	0.06676278808523266
Odalys Orraca Castillo (8 pub)	780	19	0.06373075760770841

Teddy Osmín Tamargo Barbeito (2 pub)	780	9	0.060624333930126846

Tabla X: *Criterio de clasificación: Centralidad*

Nombre y apellidos	Comunidad	Coautores	Centralidad
Maylen Cepero Madruga (1 pub)	291	1	1.0
Antonio Gutiérrez Laborit (1 pub)	291	1	1.0
Yusel Arias Guerra (2 pub)	292	1	1.0
Yulaine Arias Guerra (2 pub)	292	1	1.0
Malena García Izquierdo (1 pub)	293	1	1.0
Manuel Macías Martínez (1 pub)	293	1	1.0
Annies Cedeño López (1 pub)	294	1	1.0
Yuneisy Barrios Pérez (1 pub)	294	1	1.0
Yudit Ponce Toste (1 pub)	295	1	1.0
Félix Oscar Fernández Peña (1 pub)	295	1	1.0

Observación 6: El criterio de clasificación por intermediación y centralidad al aplicar el algoritmo ***Betweenness centrality*** pondera los elementos de la red utilizando la direccionalidad de las aristas. La intermediación no se ve afectada por la cantidad de relaciones de coautoría de los autores ni por el tamaño de las comunidades. La cercanía si tiene en cuenta la cantidad de coautores que publican juntos, pondera con mayor peso a las comunidades de menor tamaño y con menor peso a las de gran tamaño. Ambos criterios no se ven afectados por la ponderación de las relaciones de coautoría existente en la red.

El criterio de autoridad en el escenario de una red de coautoría mide cuán representativo es un autor atendiendo a sus relaciones de coautoría. Dicho criterio es medido aplicando tres algoritmos: ***HITS***, ***PageRank*** y ***AuthorRank***. Los dos primeros son algoritmos ya probados en diferentes situaciones reales y constituyen los valores o resultados de control para comparar el ***AuthorRank***, desarrollado como parte del método propuesto (ver sección **2.3.5. Algoritmos** para más detalles). Los resultados de los algoritmos se evidencian en las tablas XI, XII y XIII respectivamente.

Tabla XI: *Criterio de clasificación: Autoridad (HITS)*

Nombre y apellidos	Comunidad	Coautores	Autoridad (*HITS*)
Mikhail Benet Rodríguez (21 pub)	807	64	0.0022100576
Luis Roberto Llerena Rojas (24 pub)	788	55	0.0019040495

José G. Sanabria Negrín (18 pub)	791	52	0.0018020469
Vladimir Mendoza Rodríguez (20 pub)	788	50	0.0017340451
Raúl López Fernández (16 pub)	807	48	0.0016660433
Juan Raúl Hernández Silva (15 pub)	779	44	0.0015300398
Raymid García Fernández (10 pub)	788	42	0.001462038
Joaquín Pérez Labrador (25 pub)	793	40	0.0013940362
Carmen María Padilla González (10 pub)	779	37	0.0012920336
Juan Valiente Mustelier (13 pub)	788	36	0.0012580327

Tabla XII: *Criterio de clasificación: Autoridad (PageRank)*

Nombre y apellidos	Comunidad	Coautores	Autoridad (*PageRank*)
Mikhail Benet Rodríguez (21 pub)	807	64	0.0018502261540461562
Luis Roberto Llerena Rojas (24 pub)	788	55	0.0016894469199980808
Lidia Torres Aja (26 pub)	803	28	0.0015154203313968019
Vladimir Mendoza Rodríguez (20 pub)	788	50	0.0014723487064213848
Raúl López Fernández (16 pub)	807	48	0.001434643665735225
Jesús Juan Rodríguez (17 pub)	766	35	0.0013784020331724187
José G. Sanabria Negrín (18 pub)	791	52	0.001339503651922501
Fidel Castro Pérez (13 pub)	775	32	0.0012178433354775326
Rubén Cañedo Andalia (14 pub)	676	26	0.001187498085298377
Alfredo Darío Espinosa Brito (26 pub)	817	33	0.001157068764809578

Tabla XIII: *Criterio de clasificación: Autoridad (AuthorRank)*

Nombre y apellidos	Comunidad	Coautores	Autoridad (*AuthorRank*)
Luis Roberto Llerena Rojas (24 pub)	788	55	80.556
Vladimir Mendoza Rodríguez (20 pub)	788	50	72.337
Mikhail Benet Rodríguez (21 pub)	807	64	70.013
Juan Raúl Hernández Silva (15 pub)	779	44	54.503
Raúl López Fernández (16 pub)	807	48	54.180
José G. Sanabria Negrín (18 pub)	791	52	52.625
Juan Valiente Mustelier (13 pub)	788	36	49.255
Jesús Juan Rodríguez (17 pub)	766	35	48.343

Alfredo Darío Espinosa Brito (26 pub)	817	33	44.115
Lidia Torres Aja (26 pub)	803	28	44.001

Observación 7: El algoritmo ***HITS*** no tiene en cuenta la ponderación de las aristas para medir las autoridades en una red. En la Tabla XI se refleja que los valores que asigna a cada uno de los autores están en correspondencia con la cantidad de coautores que han publicado. ***PageRank,*** sin embargo, tiene en cuenta la ponderación de las aristas en la ejecución del paseo aleatorio que implementa y muestra resultados similares a los calculados por ***HITS***. En el cálculo del ***AuthorRank*** se evidencia que los autores con mayores ponderaciones se van clasificando según el peso de sus relaciones de coautoría, influenciado además por la cantidad de coautores en la misma comunidad y que con mayor frecuencia publican exclusivamente entre ellos. Los valores de ***AuthorRank***, en comparación con los otros dos algoritmos, ratifican su correcto funcionamiento ya que pondera a los autores en la red aprovechando todas las características de la misma; siendo los resultados similares a los alcanzados por ***HITS*** y ***PageRank***.

3.2.4. Análisis de los resultados

El método para la detección de comunidades en RC a partir de un grafo RDF de 51329 tripletas procesa y transforma los metadatos en una red de coautoría ponderada y dirigida con un total de 5203 vértices y 24498 aristas. Detecta 826 comunidades y las identifica con una clase de modularidad y un color determinado. Agrupa visualmente los elementos de la red en comunidades evitando el solapamiento de los vértices y mejorando 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 posibilitando realizar análisis y comparaciones de los mismos.

El prototipo funcional desarrollado implementa cada una de las restricciones por etapas del método propuesto. La ejecución del método siguiendo las configuraciones establecidas en el caso de estudio tiene una duración total de 179407 milisegundos, siendo la fase 3 la que ocupa un 85% del tiempo total (ver Figura XXII). 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. Es válido aclarar que en el prototipo funcional solo se ejecutan las tres fases una única vez por grafo RDF, si este no es modificado solo se accede a los archivos GEXF que fueron generados con anterioridad.

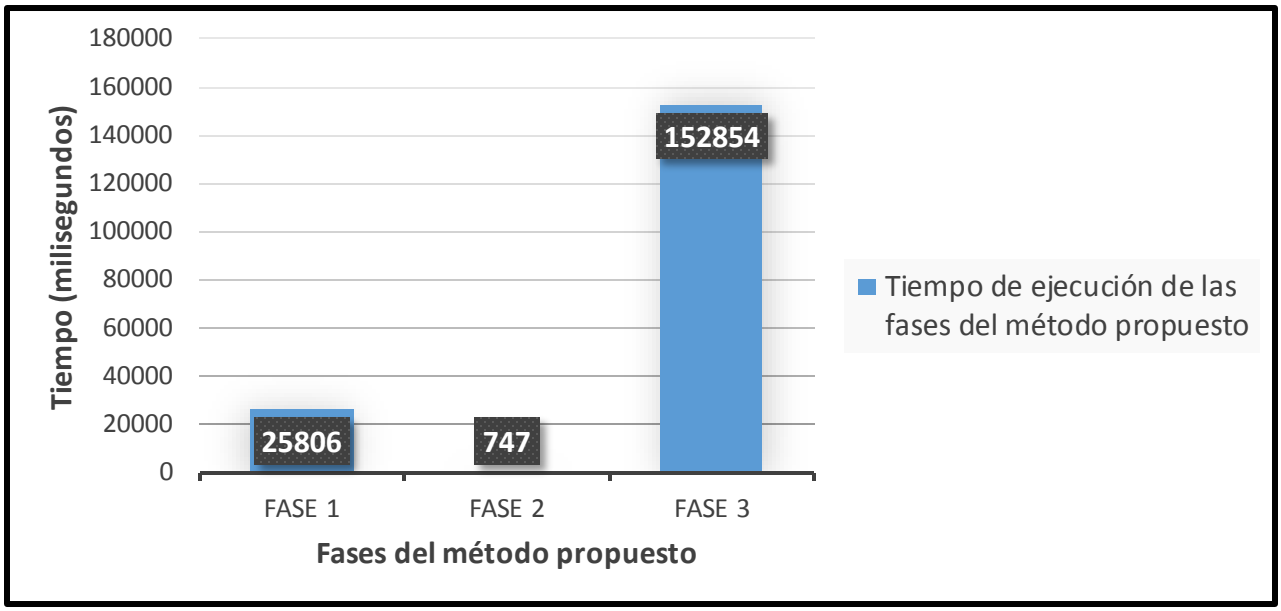

Figura XXII: *Tiempo de ejecución de las fases del método propuesto*

3.3. Conclusiones parciales

El caso de estudio permitió evaluar el método propuesto a partir de los resultados obtenidos mediante su ejecución en un prototipo funcional implementado. Las observaciones realizadas en el desarrollo de cada fase del método constataron la utilidad del mismo en la solución del problema de la detección y visualización de comunidades a partir de grafos RDF. Con los criterios de medidas aplicados se logró caracterizar las relaciones de colaboración que se establecen entre los autores a partir de redes de coautoría extraídas de grafos RDF, evidenciándose la viabilidad del método propuesto.

CONCLUSIONES GENERALES

La revisión bibliográfica evidenció que la detección de comunidades en redes es un área de investigación en constante desarrollo, principalmente en el campo de las redes sociales y redes de colaboración científica. Sin embargo, no existen evidencias de mecanismos o métodos que demuestren su aplicación en RC a partir de grafos RDF que almacenan los metadatos de las revistas y artículos científicos.

Las RC pueden ser modeladas teniendo en cuenta la tipología y direccionalidad de las aristas en la red. Los algoritmos para la detección de comunidades se basan en intermediación, dinámica social y redes, jerarquía, teoría de la información o modularidad. En la visualización de redes existen soluciones que brindan bibliotecas o extensiones para el desarrollo de aplicaciones reales como son IGraph, JUNG y Gephi. Lo anterior permitió determinar que el método a proponer debía modelar una red de coautoría ponderada y dirigida, utilizar un algoritmo basado en modularidad para la detección de las comunidades y la biblioteca de Gephi Toolkit para la visualización de la red.

Se propone un método basado en tres fases para la detección de comunidades en RC a partir de grafos RDF. Este método se implementa en un prototipo funcional siguiendo las restricciones y características definidas haciendo uso de una arquitectura basada en tuberías y filtros. La utilización de estándares, tecnologías, bibliotecas y algoritmos, unido a la aplicación de buenas prácticas y patrones de diseño en el desarrollo del prototipo potencian su correcto funcionamiento.

El desarrollo de un caso de estudio posibilitó demostrar la viabilidad del método para la detección de comunidades a partir de las redes de coautoría existentes en los grafos RDF, permitiendo identificar y cuantificar las relaciones de colaboración que se establecen entre los autores presentes en los metadatos bibliográficos.

RECOMENDACIONES

Continuar con el desarrollo del prototipo funcional para su incorporación a la plataforma de la Biblioteca Digital Semántica, desarrollada por el grupo de investigación de Web Semántica. Así como ir incorporando nuevos algoritmos y técnicas de visualización para aumentar el grado de personalización y configuración de la solución.

Fundamentar en el estudio de la detección de comunidades con el objetivo de poder identificar o desarrollar algoritmos que propongan mejores soluciones a las ya existentes. Además, se recomienda investigar posibles soluciones para el etiquetado de las comunidades de coautores a partir de las áreas temáticas a las que pertenecen.

REFERENCIAS BIBLIOGRÁFICAS

ALDECOA GARCÍA, R. 2013. Detección de comunidades en redes complejas. *Riunet* [en línea], [Consulta: 31 marzo 2015]. Disponible en: http://riunet.upv.es/handle/10251/31638.

ALDECOA, R. y MARÍN, I. 2011. Deciphering Network Community Structure by Surprise. *PLoS ONE*, vol. 6, no. 9, pp. e24195. DOI 10.1371/journal.pone.0024195.

AZIZIFARD, N. 2014. Social Network Clustering. *International Journal of Information Technology and Computer Science*, vol. 6, no. 1, pp. 76-81. ISSN 20749007, 20749015. DOI 10.5815/ijitcs.2014.01.09.

BALL, B. y NEWMAN, M.E.J. 2013. Friendship networks and social status. En: arXiv: 1205.6822, *Network Science*, vol. 1, no. 01, pp. 16-30. ISSN 2050-1242, 2050-1250. DOI 10.1017/nws.2012.4.

BASTIAN, M., HEYMANN, S. y JACOMY, M. 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks. *Third International AAAI Conference on Weblogs and Social Media* [en línea]. S.l.: s.n., [Consulta: 25 marzo 2015]. Disponible en: https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

BEAVER, D. deB y ROSEN, R. 1978. Studies in scientific collaboration. *Scientometrics*, vol. 1, no. 1, pp. 65-84. ISSN 0138-9130, 1588-2861. DOI 10.1007/BF02016840.

BERNERS-LEE, T. 2006. Linked data-design issues. [en línea], [Consulta: 2 abril 2015]. Disponible en: http://www.w3.org/DesignIssues/LinkedData.html.

BLONDEL, V.D., GUILLAUME, J.-L., LAMBIOTTE, R. y LEFEBVRE, E. 2008. Fast unfolding of communities in large networks. En: arXiv: 0803.0476, *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, pp. P10008. ISSN 1742-5468. DOI 10.1088/1742-5468/2008/10/P10008.

BÖRNER, K., MARU, J.T. y GOLDSTONE, R.L. 2004. The simultaneous evolution of author and paper networks. En: PMID: 14976254, *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5266-5273. ISSN 0027-8424, 1091-6490. DOI 10.1073/pnas.0307625100.

BOS, N., ZIMMERMAN, A., OLSON, J., YEW, J., YERKIE, J., DAHL, E. y OLSON, G. 2007. From shared databases to communities of practice: A taxonomy of collaboratories. *Journal of Computer-Mediated Communication*, vol. 12, no. 2, pp. 652-672.

BRANDES, U. 2001. A faster algorithm for betweenness centrality*. *Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163–177.

BRIN, S. y PAGE, L. 1998. The Anatomy of a Large-scale Hypertextual Web Search Engine. [en línea]. Amsterdam, The Netherlands, The Netherlands: Elsevier Science Publishers B. V., pp. 107–117. [Consulta: 2 abril 2015]. Disponible en: http://dl.acm.org/citation.cfm?id=297805.297827.

BUSH, V. y THINK, A.W.M. 1945. The atlantic monthly. *As we may think*, vol. 176, no. 1, pp. 101-108.

CAZABET, R., AMBLARD, F. y HANACHI, C. 2010. Detection of overlapping communities in dynamical social networks. . S.l.: IEEE, pp. 309-314.

CHANG, H.-W. y HUANG, M.-H. 2014. Cohesive subgroups in the international collaboration network in astronomy and astrophysics. *Scientometrics*, vol. 101, no. 3, pp. 1587-1607. ISSN 0138-9130, 1588-2861. DOI 10.1007/s11192-014-1312-9.

CRAVINO, N., DEVEZAS, J. y FIGUEIRA, Á. 2012. Using the Overlapping Community Structure of a Network of Tags to Improve Text Clustering. [en línea]. New York, NY, USA: ACM, pp. 239–244. [Consulta: 27 marzo 2015]. ISBN 978-1-4503-1335-3. Disponible en: http://doi.acm.org/10.1145/2309996.2310036.

DE CASTRO, R. y GROSSMAN, J.W. 1999. Famous trails to Paul Erdős. *The Mathematical Intelligencer*, vol. 21, no. 3, pp. 51-53.

DE OLIVEIRA, T.B.S., ZHAO, L., FACELI, K. y DE CARVALHO, A.C.P.L.F. 2008. Data clustering based on complex network community detection. *IEEE Congress on Evolutionary Computation, 2008. CEC 2008. (IEEE World Congress on Computational Intelligence)*. S.l.: s.n., pp. 2121-2126.

DONETTI, L. y MUNOZ, M.A. 2004. Detecting network communities: a new systematic and efficient algorithm. *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2004, no. 10, pp. P10012.

FORTUNATO, S. 2010. Community detection in graphs. *Physics Reports*, vol. 486, no. 3, pp. 75-174.

FRAME, J.D. y CARPENTER, M.P. 1979. International research collaboration. *Social Studies of Science*, vol. 9, no. 4, pp. 481–497.

FREEMAN, L.C. 1977. A set of measures of centrality based on betweenness. *Sociometry*, pp. 35-41.

GARVEY, W.D. y GRIFFITH, B.C. 1964. Scientific information exchange in psychology. *Science (New York, NY)*, vol. 146, no. 3652, pp. 1655–1659.

HE, B., DING, Y. y NI, C. 2011. Mining enriched contextual information of scientific collaboration: A meso perspective. *Journal of the American Society for Information Science and Technology*, vol. 62, no. 5, pp. 831-845. ISSN 1532-2890. DOI 10.1002/asi.21510.

HUANG, H.-H. y YANG, H.-C. 2012. Semantic Clustering-Based Community Detection in an Evolving Social Network. *2012 Sixth International Conference on Genetic and Evolutionary Computing (ICGEC)*. S.l.: s.n., pp. 91-94.

HU, Y. 2011. Algorithms for visualizing large networks. *Combinatorial Scientific Computing*, vol. 5, no. 3, pp. 180–186.

ICHISE, R., TAKEDA, H. y UEYAMA, K. 2006. Exploration of researchers' social network for discovering communities. *New Frontiers in Artificial Intelligence*. S.l.: Springer, pp. 458-469. ISBN 3540354700.

JACOMY, M., VENTURINI, T., HEYMANN, S. y BASTIAN, M. 2014. ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software. *PLoS ONE*, vol. 9, no. 6, pp. e98679. DOI 10.1371/journal.pone.0098679.

JI LIU, L.L. 2011a. Network Community Detection Based on Co-Neighbor Modularity Matrix with Spectral Clustering. *Applied Mechanics and Materials*, ISSN 1662-7482. DOI 10.4028/www.scientific.net/AMM.55-57.1237.

JI LIU, L.L. 2011b. Network Community Detection Based on Co-Neighbor Modularity Matrix with Spectral Clustering. *Applied Mechanics and Materials*, ISSN 1662-7482. DOI 10.4028/www.scientific.net/AMM.55-57.1237.

KLEINBERG, J.M. 1999. Authoritative Sources in a Hyperlinked Environment. *J. ACM*, vol. 46, no. 5, pp. 604–632. ISSN 0004-5411. DOI 10.1145/324133.324140.

LIU, X., BOLLEN, J., NELSON, M.L. y VAN DE SOMPEL, H. 2005. Co-authorship Networks in the Digital Library Research Community. *Inf. Process. Manage.*, vol. 41, no. 6, pp. 1462–1480. ISSN 0306-4573. DOI 10.1016/j.ipm.2005.03.012.

LI, Z., CHEN, Y., MU, D., YUAN, J., SHI, Y., ZHANG, H., GAN, J., LI, N., HU, X., LIU, B. y OTHERS 2012. Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in functional genomics*, vol. 11, no. 1, pp. 25–37.

MALTRAS, B., VEGA, J. y QUINTANILLA, M.A. 1995. Measuring multinational cooperation in science & technology : different methods applied to the European Framework Programs. *International Society for Scientometrics and Informetrics. International conference* [en línea]. S.l.: s.n., pp. 303-312. [Consulta: 30 marzo 2015]. ISBN 1-57387-010-2. Disponible en: http://cat.inist.fr/?aModele=afficheN&cpsidt=3146004.

MARTIN, T., BALL, B., KARRER, B. y NEWMAN, M.E.J. 2013. Coauthorship and citation in scientific publishing. En: arXiv: 1304.0473, *Physical Review E* [en línea], vol. 88, no. 1. [Consulta: 30 marzo 2015]. ISSN 1539-3755, 1550-2376. DOI 10.1103/PhysRevE.88.012814. Disponible en: http://arxiv.org/abs/1304.0473.

MEDRANO, J.F., BERROCAL, J.L.A. y FIGUEROLA, C.G. 2011. *Visualización de Grafos Web*. S.l.: s.n.

MELIN, G. y PERSSON, O. 1996. Studying research collaboration using co-authorships. *Scientometrics*, vol. 36, no. 3, pp. 363-377. ISSN 0138-9130, 1588-2861. DOI 10.1007/BF02129600.

MIQUEL, J.F., OKUBO, Y., NARVAEZ, N. y FRIGOLETTO, L. 1989. Les scientifiques sont-ils ouverts à la coopération internationale. *La Recherche*, vol. 20, no. 206, pp. 116–118.

MONGE, E.C. 2010. El estudio de casos como metodología de investigación y su importancia en la dirección y administración de empresas. *Revista Nacional de Administración*, vol. 1, no. 2, pp. 31–54.

MONTFORT, N. 2004. Discovering communities through information structure and dynamics: A review of recent research. *Pennsylvania State University (Technical Report, no MS-CIS-04–18)*,

NASCIMENTO, M.A., SANDER, J. y POUND, J. 2003. Analysis of SIGMOD's co-authorship graph. *ACM Sigmod record*, vol. 32, no. 3, pp. 8-10.

NEWMAN, M.E.J. 2012. Communities, modules and large-scale structure in networks. *Nature Physics*, vol. 8, no. 1, pp. 25–31.

NEWMAN, M.E.J. y GIRVAN, M. 2004. Finding and evaluating community structure in networks. *Physical Review E* [en línea], vol. 69, no. 2. [Consulta: 30 marzo 2015]. ISSN 1539-3755, 1550-2376. DOI 10.1103/PhysRevE.69.026113. Disponible en: http://www.bibsonomy.org/bibtex/1b9145040e35ccb4d2a0ce18105e64ff4/kibanov.

O'MADADHAIN, J., FISHER, D., SMYTH, P., WHITE, S. y BOEY, Y.-B. 2005. Analysis and visualization of network data using JUNG. *Journal of Statistical Software*, vol. 10, no. 2, pp. 1-35.

PALLA, G., DERÉNYI, I., FARKAS, I. y VICSEK, T. 2005. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, vol. 435, no. 7043, pp. 814-818.

PERIANES-RODRÍGUEZ, A. 2007. Análisis y visualización de redes de colaboración científica: grupos de investigación en la Universidad Carlos III de Madrid (ISI, Web of Science, 1990-2004). ,

PRESSMAN, R.S. 2002. *Ingeniería del software: un enfoque práctico*. 5ta. S.l.: Mikel Angoar.

RODRÍGUEZ, A.P., GÓMEZ, C.O. y DE MOYA ANEGÓN, F. 2010. *Redes de colaboración científica: análisis y visualización de patrones de coautoría*. S.l.: Tirant lo Blanch.

ROUSSEAU, R. 2001. Are multi-authored articles cited more than single-authored ones? Are collaborations with authors from other countries more cited than collaborations within the country? A case study. ,

SAHA, B., MANDAL, A., TRIPATHY, S.B. y MUKHERJEE, D. 2015. Complex Networks, Communities and Clustering: A survey. En: arXiv: 1503.06277, *arXiv:1503.06277 [cs]* [en línea], [Consulta: 25 marzo 2015]. Disponible en: http://arxiv.org/abs/1503.06277.

SCIBETTA, M., BOANO, F., REVELLI, R. y RIDOLFI, L. 2013. Community detection as a tool for complex pipe network clustering. *EPL (Europhysics Letters)*, vol. 103, no. 4, pp. 48001. ISSN 0295-5075. DOI 10.1209/0295-5075/103/48001.

SILVA, T.C. y ZHAO, L. 2007. Pixel Clustering by Using Complex Network Community Detection Technique. [en línea]. Washington, DC, USA: IEEE Computer Society, pp. 925–932. [Consulta: 27 marzo 2015]. ISBN 0-7695-2976-3. Disponible en: http://dl.acm.org/citation.cfm?id=1317534.1318246.

STOKES, T.D. y HARTLEY, J.A. 1989. Coauthorship, social structure and influence within specialties. *Social Studies of Science*, vol. 19, no. 1, pp. 101-125.

STUDER, R. y BENJAMINS, V.R. 2007. Dieter Fense «Knowledge Engineering Principles and Methods». . S.l.:

TUCKER, A.B. 2004. *Computer science handbook*. S.l.: CRC press. ISBN 0203494458.

VIDAL, J. y VILLARROEL, R. 1995. The dynamics of research groups : representation and interpretation problems in collaboration analysis. *International Society for Scientometrics and Informetrics. International conference* [en línea]. S.l.: s.n., pp. 607-616. [Consulta: 30 marzo 2015]. ISBN 1-57387-010-2. Disponible en: http://cat.inist.fr/?aModele=afficheN&cpsidt=3146158.

WAGNER, C., LEYDESDORFF, L. y BORNMANN, L. 2014. Recent Developments in China-US Science Cooperation. *arXiv preprint arXiv:1404.6545* [en línea], [Consulta: 13 abril 2015]. Disponible en: http://arxiv.org/abs/1404.6545.

WEB GEPHI 2015. Gephi - Makes graphs handy. [en línea]. Disponible en: http://gephi.github.io/.

XIE, J. y SZYMANSKI, B.K. 2011. Community Detection Using A Neighborhood Strength Driven Label Propagation Algorithm. *arXiv:1105.3264 [physics]* [en línea], [Consulta: 27 marzo 2015]. Disponible en: http://arxiv.org/abs/1105.3264.

XUN, G., YANG, Y., WANG, L. y LIU, W. 2012. Latent Community Discovery with Network Regularization for Core Actors Clustering. *Proceedings of COLING 2012: Posters*, pp. 1351-1360.

ZARE, R.N. 1997. Knowledge and distributed intelligence. *Science*, vol. 275, no. 5303, pp. 1047-1047.

ZHANG, S.-Z., FANG, Z.-X., CHEN, J.-G. y SHI, J. 2013. Community clustering model for E-commerce trust based on social network. *Journal of Zhejiang University. Engineering Science*, vol. 47, no. 4, pp. 656–661.

ZHUHADAR, L., YANG, R. y NASRAOUI, O. 2012. Toward the Design of a Recommender System: Visual Clustering and Detecting Community Structure in a Web Usage Network. [en línea]. Washington, DC, USA: IEEE Computer Society, pp. 354–361. [Consulta: 27 marzo 2015]. ISBN 978-0-7695-4880-7. Disponible en: http://dl.acm.org/citation.cfm?id=2457524.2457614.