Inferencia de Redes Temáticas de Colaboración en Bibliotecas Digitales

T E S I S

Maestría en Ciencias en Tecnología Informática

Instituto Tecnológico y de Estudios Superiores de Monterrey

Por

Ing. Eric Santiago Balam Sabido

Diciembre 2010

Inferencia de Redes Temáticas de Colaboración en Bibliotecas Digitales

TESIS

Maestría en Ciencias en

Tecnología Informática

Instituto Tecnológico y de Estudios Superiores de Monterrey

Por

Ing. Eric Santiago Balam Sabido

Diciembre 2010

Instituto Tecnológico y de Estudios Superiores de Monterrey

División de Graduados en Mecatrónica y Tecnologías de Información

Los miembros del comité de tesis recomendamos que la presente tesis de Eric Santiago Balam Sabido sea aceptada como requisito parcial para obtener el grado académico de Maestro en Ciencias en:

Tecnología Informática

Comité de tesis:

Dr. Juan Carlos Lavariega Jarquín
Asesor de la tesis

Dra. Lorena Gómez Martínez
Sinodal

Dr. Ramón F. Brena Pinero
Sinodal

Dr. Ramón F. Brena Pinero
Director de maestrías en Computación
División de Mecatrónica y Tecnologías
de Información

Diciembre de 2010

Inferencia de Redes Temáticas de Colaboración en Bibliotecas Digitales

Por

Ing. Eric Santiago Balam Sabido

TESIS

Presentada a la División de Mecatrónica y Tecnologías de Información
Este trabajo es requisito parcial para obtener el grado académico de Maestro en Ciencias en Tecnología Informática

Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Monterrey

Monterrey, N.L. Diciembre de 2010

A mi esposa, padres y hermano, por su inmenso cariño y apoyo.

Reconocimientos

Deseo externar un sincero agradecimiento a las personas que de alguna forma colaboraron en el desarrollo de esta tesis.

Quiero agradecer a mi esposa por su amor, apoyo y comprensión en todo momento. En especial durante los últimos dos años, que involucraron alegrías, nuevas experiencias y sin duda sacrificios como: el estar lejos de nuestras familias y la poca atención que te brinde en algunos momentos a causa de la escuela.

A mis padres y hermano por sus palabras de aliento, porque aun lejos siempre estuvieron presentes cuando los necesite, siempre impulsándome a seguir adelante. Gracias por ser una familia llena de amor.

Al Dr. Juan Carlos Lavariega, por su asesoría durante el desarrollo de esta tesis; por su constante apoyo que siempre fue un estímulo para concluir este proyecto. Muchas gracias.

A mis sinodales, Dra. Lorena Gómez y Dr. Ramón Brena, por sus valiosos consejos para mejorar la calidad de esta tesis.

A mis amigos Eric Villafaña, Mijail Espadas y Rodrigo Marquez que me animaron para estudiar la maestría y siempre me dieron ánimos en los momentos difíciles.

Eric Santiago Balam Sabido

Instituto Tecnológico y de Estudios Superiores de Monterrey
Diciembre 2010

Inferencia de Redes Temáticas de Colaboración en Bibliotecas Digitales

Eric Santiago Balam Sabido, M.C.
Instituto Tecnológico y de Estudios Superiores de Monterrey, 2010

Asesor de la tesis: Dr. Juan Carlos Lavariega Jarquín

Las bibliotecas digitales son un conjunto de recursos electrónicos y capacidades técnicas asociadas para crear, buscar y utilizar información. En este sentido el contenido de las bibliotecas digitales incluye gran variedad de datos y metadatos que describen diversos aspectos de los documentos. Las bibliotecas digitales son construidas por y para las necesidades de una comunidad de usuarios y contienen información con el fin de satisfacer las necesidades de esa comunidad.

Lo anterior las convierte en una extensión e integración de la información contenida físicamente dentro de instituciones, donde los recursos son seleccionados, recolectados, organizados, preservados y accedidos en apoyo a los usuarios.

Convergiendo así en un recurso que provee servicios e información institucional a una comunidad integrada por instituciones, investigadores, estudiantes y público en general.

Encontrar la afinidad en colecciones de documentos pertenecientes a bibliotecas digitales involucra el determinar mediante ciertas técnicas la similitud de contenido de los documentos y el poder representar dicha similitud. En este trabajo de tesis se presentan técnicas aplicadas a colecciones de documentos provenientes de diferentes bibliotecas digitales. La técnica de agrupamiento jerárquico basado en la frecuencia de de un conjunto de elementos es aplicada con la finalidad de agrupar los documentos provenientes de diferentes bibliotecas digitales en base a la similitud de su contenido. Las métricas definidas por Xiaoming Liu que originalmente han sido usadas para obtener una red de co-autoría se aplican a documentos previamente agrupados, la adaptación y adecuación de uso de dichas métricas permite obtener lo que definiremos como red temática de colaboración, red capaz de expresar el grado de afinidad entre instituciones o autores a partir de la similitud entre documentos.

Actualmente con los servicios ofrecidos por RABiD (Red Abierta de Bibliotecas Digitales) y debido a la ausencia de un servicio que permita generar una red temática de

colaboración no se pueden identificar a los investigadores o las instituciones con los que un investigador puede establecer vínculos de colaboración tomando como referencia la afinidad temática de los documentos que se encuentran en una biblioteca digital. Ante esta situación surge la necesidad de construir una red temática de colaboración a dos niveles: a nivel de instituciones y a nivel de autores, promoviendo de esta manera la creación de comunidades de contenido científico.

En el presente trabajo se utilizan para la implementación de las técnicas antes mencionadas las colecciones digitales Phronesis (biblioteca digital del Instituto Tecnológico y de Estudios Superiores de Monterrey, Monterrey), CIRIA (biblioteca digital de la Universidad de las Américas, Puebla) y Redalyc (biblioteca digital de la Universidad Autónoma del Estado de México), las cuales son repositorios de documentos científicos (tesis y artículos) ubicados en diversos servidores y forman parte de la RABiD.

La contribución de este trabajo es la aplicación del trabajo previo en agrupación jerárquica y la adaptación y adecuación del uso de métricas con el propósito de inferir redes temáticas de colaboración. Redes capaces de expresar el grado de colaboración entre instituciones y entre los autores a partir del contenido de los documentos provenientes de bibliotecas digitales. Por lo tanto, involucra el desarrollo una herramienta capaz de inferir la red temática de colaboración a partir de los documentos provenientes de las colecciones digitales de Phronesis, CIRIA y Redalyc.

Índice general

Índice de tablas

Índice de figuras

Capítulo 1

Introducción.

Las bibliotecas digitales son un conjunto de recursos electrónicos y capacidades técnicas asociadas para crear, buscar y utilizar información. En este sentido el contenido de las bibliotecas digitales incluye gran variedad de datos y metadatos que describen diversos aspectos de los documentos (por ejemplo, la representación, el creador, propietario, descripción, etc.) [4]. Las bibliotecas digitales son construidas por y para las necesidades de una comunidad de usuarios y contienen información con el fin de satisfacer las necesidades de esa comunidad, lo que permite que individuos interactúen unos con otros a través de su uso.

Lo anterior las convierte en una extensión e integración de la información contenida físicamente dentro de instituciones, donde los recursos son seleccionados, recolectados, organizados, preservados y accedidos en apoyo a los usuarios.

Convergiendo así en un recurso que provee servicios e información institucional a una comunidad integrada por instituciones, investigadores, estudiantes y público en general.

Debemos identificar el tipo de usuarios que hacen uso de la biblioteca digital, solo de esta forma podemos ofrecer un equilibrio entre el uso de la biblioteca y sus usuarios, con un soporte específico de acuerdo a las actividades de una comunidad.

En bibliotecas digitales de contenido científico la ausencia de grupos de colaboración en base a la afinidad temática de sus documentos afecta la identificación de investigadores con intereses en áreas afines y su posibilidad de interacción. La dificultad para formar estos grupos de colaboración radica en: 1) La complejidad de acceder al contenido de sus documentos, 2) El poder aplicar alguna técnica de agrupamiento en base al contenido y 3) Poder generar una representación que exprese de manera cuantitativa el grado de colaboración entre investigadores o instituciones a partir de la colección de documentos. En resumen las bibliotecas digitales no proveen de un servicio de identificación de redes de colaboración en base a la afinidad del contenido de sus documentos.

Tal servicio de búsqueda de afinidad podría permitir la formación de grupos y redes de colaboración. La principal restricción para formar estas redes radica en el acceso limitado a las colecciones digitales y la ausencia de un método que permita expresar

cuantitativamente el grado de colaboración o afinidad para potenciales colaboradores.

En México el grupo denominado RABID [1] (Red Abierta de Bibliotecas Digitales) ha promovido la colaboración entre sus miembros, tratando de explotar el contenido de sus repositorios institucionales a través de métodos limitados y tradicionales. El objetivo de RABiD es contribuir a la consolidación del desarrollo de bibliotecas digitales en México a través de una red abierta por medio de la cual puedan compartirse colecciones y servicios, a la vez que se facilita la integración de nuevas instituciones, servicios y usuarios. También se han generado servidores de metadatos bajo el protocolo estándar OAI-PMH [2], permitiendo la participación de colecciones de RABiD como proveedores de datos de la comunidad internacional de archivos abiertos y facilitando la recuperación de documentos de forma federada.

RABiD cuenta con 18 miembros, una de las recomendaciones para ser miembro de RABiD es ser al mismo tiempo ser parte de la Corporación Universitaria para el Desarrollo de Internet-2 A.C. (CUDI) [3].

Debido a la ausencia de una red temática de colaboración se limita la generación de propuestas conjuntas y existe una dificultad para detectar investigadores con aéreas de interés en común, lo que da como resultado duplicidad de esfuerzos, trabajos de investigación aislados y dificultad para acceder a fondos para la realización de proyectos.

En la Red Abierta de Bibliotecas Digitales, anteriormente se han ofrecido documentos ordenados de manera cronológica y alfabética, y que han sido visualizados mediante un método denominado Star-fish [21], sin embargo esta clasificación no permite representar afinidad temática.

Existe otro trabajo relacionado con la clasificación e identificación de afinidad temática en los contenidos de una biblioteca digital llamado OntOAIr [15], el cual permite identificar la afinidad temática de los documentos y hacer un agrupamiento jerárquico en una estructura de árbol, lo que facilita la clasificación y la búsqueda de los documentos, sin embargo esta estructura no pude ser representada como una red de colaboración.

Lee Iverson [10] menciona que merece gran atención la necesidad de las bibliotecas digitales de ofrecer servicios que satisfagan a sus usuarios y comunidades en relación a su necesidad de nuevas herramientas, para utilizar y organizar la información que se encuentra acumulada.

El agrupamiento de documentos es una operación fundamental usada para la organización sin supervisión de documentos, extracción automática de tópicos y recu-

[1] http://ict.udlap.mx/rabid/ *sitio de la Red Abierta de Bibliotecas Digitales*, consultado el 15 de octubre de 2009.

[2] http://www.openarchives.org/OAI/openarchivesprotocol.html *sitio del Protocolo de la iniciativa de archivos abiertos para la cosecha de metadatos*, consultado el 15 de octubre de 2009.

[3] http://www.cudi.edu.mx/ *sitio de la Corporación Universitaria para el Desarrollo de Internet-2 México*, consultado el 15 de octubre de 2009.

peración de información [18]. Agrupamiento implica dividir un conjunto de objetos en un número específico de grupos [11], de manera que los documentos dentro de un grupo tienen una similitud de contenido alta mientras que los documentos en grupos distintos son diferentes entre sí [7].

Trabajos anteriores han demostrado que los algoritmos jerárquicos producen grupos de alta calidad con grandes volúmenes de datos [7]. El algoritmo FIHC [9] es un algoritmo de agrupamiento jerárquico que se basa en la frecuencia de un conjunto de elementos, su hipótesis es que si un grupo de documentos hace referencia a un mismo tópico, debe compartir un conjunto de términos, los cuales son llamados conjunto de términos frecuentes.

Xiaoming Liu [13] afirma que en la última década ante la necesidad de aprender más sobre la comunidad de investigadores y sus estructuras de colaboración se han inducido patrones de colaboración aplicados al dominio de las bibliotecas digitales, ante esta situación como resultado de su trabajo se definen métricas que permiten generar una red de colaboración de co-autores.

1.1. Definición del Problema.

Al no existir un servicio para las bibliotecas digitales de contenido científico que proporcione una colección estructurada que permita obtener una red temática de colaboración a nivel de instituciones y a nivel de autores a partir del contenido de sus documentos, se pierden oportunidades de difusión de conocimiento entre investigadores e instituciones y se limita la generación de propuestas conjuntas que permitan acceder a fondos para la realización de proyectos.

Referente a los servicios que ofrece RABID no se cuenta con alguno que permita representar la afinidad temática de sus documentos y expresarla como una red de colaboración entre investigadores o instituciones.

Desde el punto de vista técnico este trabajo se enfoca en definir y proponer una estrategia de identificación de afinidades en base al contenido científico. La estrategia incluye:

- El uso de un algoritmo de agrupamiento.
- La definición de un criterio de colaboración.
- La generación de documentos para su importación/exportación a otros sistemas con fines de visualización. Documentos con la información referente a la red temática de colaboración a nivel instituciones y a nivel autores.

1.1.1. Justificación

Cuando nos enfocamos en bibliotecas digitales de contenido científico lo anterior afecta la generación de propuestas y proyectos conjuntos, teniendo como consecuencia:

- Trabajo aislado de grupos de investigación.
- Duplicidad de esfuerzos.
- Dificultad para detectar investigadores con aéreas de interés en común.
- Se desaprovecha el conocimiento de otros investigadores, impidiendo aumentar los beneficios para toda la comunidad.

Es importante considerar el hecho de que existen fondos para apoyar el desarrollo de proyectos colaborativos entre los miembros CUDI, estos fondos se otorgan a través de convocatorias por parte del CUDI y del consejo nacional de ciencia y tecnología (CONACYT [4]). De esta forma se promueve la conformación de comunidades académicas que se organicen alrededor de temáticas específicas en proyectos que contribuyan al desarrollo del país.

La figura 1.1 muestra los fondos que han sido otorgados a través de convocatorias CUDI - CONACYT entre el año 2003 y el año 2009, que suman un total de $12,800,000.00 pesos M.N..

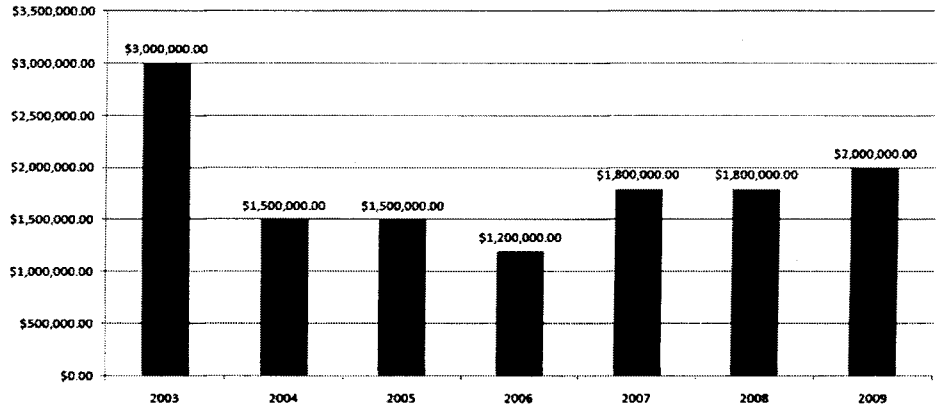


Figura 1.1: Fondos CUDI-CONACYT entre el año 2003 y el año 2009

En base a la colaboración que de alguna manera han logrado establecer investigadores de dos o más instituciones para acceder a fondos de las convocatoria CUDI - CONACYT, se han desarrollado 23 proyectos, la tabla 1.1 muestra las 40 instituciones que han accedido a estos fondos entre el año 2003 y el año 2009, su porcentaje de participación y los fondos a los que han tenido acceso debido a su participación.

[4] http://www.conacyt.mx/ *sitio del Consejo Nacional de Ciencia y Tecnología*, consultado el 15 de octubre de 2009.

Institución	% de participación	Fondos recibidos
UNAM	65 %	$ 8,347,826
UDLAP	35 %	$ 4,452,174
ITESM	30 %	$ 3,895,652
UDG	30 %	$ 3,895,652
CICESE	17 %	$ 2,226,087
UCOL	17 %	$ 2,226,087
CINVESTAV	13 %	$ 1,669,565
UAM	9 %	$ 1,113,043
BUAP	9 %	$ 1,113,043
USON	9 %	$ 1,113,043
UASLP	9 %	$ 1,113,043
UAEM	9 %	$ 1,113,043
IPN	9 %	$ 1,113,043
UACJ	9 %	$ 1,113,043
UV	9 %	$ 1,113,043
CENIDET	9 %	$ 1,113,043
LANIA	4 %	$ 556,522
IIE	4 %	$ 556,522
UABC	4 %	$ 556,522
IMP	4 %	$ 556,522
MORA	4 %	$ 556,522
UAEMEX	4 %	$ 556,522
INAOE	4 %	$ 556,522
UATX	4 %	$ 556,522
IT-VERACRUZ	4 %	$ 556,522
IMP	4 %	$ 556,522
UBJO	4 %	$ 556,522
UMNSH	4 %	$ 556,522
IPICyT	4 %	$ 556,522
UReg	4 %	$ 556,522
UM	4 %	$ 556,522
UNESCO	4 %	$ 556,522
CIDESI	4 %	$ 556,522
CIATEC	4 %	$ 556,522
CIO	4 %	$ 556,522
CIATEJ	4 %	$ 556,522
CIDETEQ	4 %	$ 556,522
UACH	4 %	$ 556,522
UAT	4 %	$ 556,522
TAMU	4 %	$ 556,522

Tabla 1.1: Porcentaje de participación a nivel institución entre el año 2003 y año 2009

Entre los requisitos para acceder a estos fondos se establece que la propuesta debe estar dirigida por un investigador líder y al menos un investigador principal de otra institución, y la participación de al menos dos instituciones miembros del CUDI en el proyecto.

1.2. Motivación.

La motivación de este trabajo es la creación de una estructura de colaboración de contenido científico para una comunidad, donde a diferencia de otras comunidades virtuales (por ejemplo una comunidad en facebook[5]) en las que los usuarios se ponen en contacto primero y después deciden de que hablar, aquí la comunidad se crea sin la necesidad de un contacto previo y en base a temas afines extraídos del contenido de documentos científicos. Lo anterior a través de la adaptación y adecuación de uso del trabajo realizado en [13].

Es importante mencionar que el resultado de esas propuestas de colaboración puede dar lugar a financiamientos otorgados por agencias nacionales como el Conacyt y CUDI.

1.3. Objetivos.

El objetivo general es desarrollar una herramienta para la Red Abierta de Bibliotecas Digitales capaz de inferir a partir de la temática de sus documentos redes de colaboración a dos niveles:

- Colaboración a nivel de instituciones.
- Colaboración a nivel de autores.

Los objetivos específicos para lograr lo anterior son:

- Realizar el agrupamiento de los documentos en base a su similitud utilizando el algoritmo FIHC.
- Realizar la adaptación y adecuación de uso de las métricas de colaboración definidas por Xiaoming Liu [13] aplicándolas sobre documentos previamente agrupados por su afinidad temática.

[5]http://www.facebook.com/ *sitio de facebook*, consultado el 28 de noviembre de 2009.

Para este trabajo se utilizarán como fuente de datos los documentos procedentes de las colecciones de Phronesis [6], CIRIA [7] y Redalyc [8], que forman parte de RABiD.

1.4. Contribución.

Las contribución de este trabajo está delimitada por las siguientes aportaciones:

- Identificar la similitud temática entre documentos pertenecientes a colecciones digitales mediante el algoritmo de agrupamiento jerárquico FIHC.
- Obtener redes temáticas de colaboración a través de la adaptación y adecuación de uso de las métricas de Xiaoming Liu [13], específicamente:
  - Red temática que expresa el nivel o grado de colaboración entre autores.
  - Red temática que expresa el nivel o grado de colaboración entre instituciones.
- El desarrollo de una herramienta capaz de generar dichas redes a partir de los documentos de las colecciones de Phronesis, CIRIA y Redalyc que forman parte de RABiD.

1.5. Organización de la tesis.

A continuación se presenta una descripción general de la organización de la presente tesis: En el capítulo dos se presenta el marco teórico relacionado con el desarrollo de este trabajo, el capítulo tres presenta el modelo de solución propuesto para la inferencia de redes temáticas de colaboración, se describe la metodología a través de la cual se infieren las redes de colaboración, en el capítulo cuatro se presentan los detalles de la implementación de las etapas descritas en la metodología, los resultados obtenidos a partir de su implementación y la validación de los resultados; por último, en el capítulo cinco se presentan las conclusiones y se incluyen sugerencias para trabajos futuros.

[6] http://copernico.mty.itesm.mx/phronesis/demo/ *sitio de la biblioteca digital Phronesis*, consultado el 15 de octubre de 2009.

[7] http://catarina.udlap.mx/u_dl_a/tales/ *sitio de la biblioteca digital CIRIA*, consultado el 15 de octubre de 2009.

[8] http://redalyc.uaemex.mx/ *sitio de la biblioteca digital Redalyc*, consultado el 15 de octubre de 2009.

Capítulo 2

Marco Teórico.

En este capítulo se presenta el marco teórico. Se describen las bibliotecas digitales, el estándar que permite la interoperatibilidad entre bibliotecas digitales; se hace mención de trabajos relacionados con la agrupación de documentos, con la generación de ontologías, y con la obtención de estructuras de colaboración en comunidades científicas.

2.1. Bibliotecas digitales.

El concepto de biblioteca digital no es simplemente el equivalente de colecciones digitalizadas con herramientas de manejo de información. Es más bien un ambiente digital para integrar colecciones, servicios y personas en apoyo a un ciclo vital de creación, diseminación, uso y preservación de datos, información y conocimiento [5]. El conjunto de características que definen a una biblioteca digital son [25]:

- Proveen acceso rápido y eficiente a través de una buena interfaz.
- Pertenecen a una organización estructurada y lógica.
- Apoyan fuertemente a la enseñanza y no sólo al acceso documental.
- Sirven a una comunidad o grupo bien definido.
- Poseen y adquieren una buena cantidad de recursos documentales.
- Trabajan en forma federada o colaborativa con otras bibliotecas.
- Invierten desarrollo en sus colecciones.
- Sus colecciones están bien definidas en cuanto a políticas de selección. Son vastas, y perduran a lo largo del tiempo.

Entre los servicios que ofrecen las bibliotecas digitales se encuentran [1]:

- **Creación de documentos digitales**. Este servicio permite la creación de un documento digital a partir de la conversión del mismo documento almacenado en otro formato. Esto se hace con la finalidad de tener disponible diferentes versiones (pdf, txt, doc) del mismo documento. Otros tipos de documentos digitales corresponden a sonido, video o imagen, un documento digital no necesariamente es textual, aunque el manejo de documentos digitales de contenido no-textual queda fuera del alcance de esta tesis.

- **Clasificación e indexamiento de contenido**. Los documentos almacenados (en sus diferentes formatos) en la biblioteca digital deben ser clasificados e indexados periódicamente, esto con la finalidad de mantener los servicios de búsqueda con las información más actualizada (cada que se agrega un nuevo documento, este debe ser indexado).

- **Búsqueda y recuperación**. Una biblioteca digital debe proporcionar servicios de búqueda y recuperación de documentos de manera fácil e intuitiva para el usuario. Los mecanismos de búsqueda pueden ser variados, pero en la mayoría de las bibliotecas digitales se permite buscar por palabras contenidas en el documento (búsqueda básica) y en los metadatos del mismo (búsqueda avanzada).

- **Distribución**. Los usuarios de la biblioteca digital deben disponer (poder acceder), de manera rápida y segura a los documentos almacenados en la biblioteca digital.

- **Administración y control de acceso**. Una biblioteca digital debe contar con un sistema de control de acceso a los documentos, así como una manera fácil de administrar y configurar usuarios y características de la biblioteca digital.

Bajo el enfoque de la organización documental los esfuerzos en bibliotecas digitales van dirigidos al análisis de: *políticas recuperación de la información, tipos de metadatos descriptivos aceptados, leguajes de marcado y clasificación e indización manual y automatizada* [25].

Las bibliotecas digitales permiten el acceso en línea a colecciones de documentos de gran volumen y de carácter científico, por tal motivo las bibliotecas digitales incluyen gran variedad de datos y metadatos que describen diversos aspectos de la información que contienen [4]. Es de especial interés en la comunidad de bibliotecas digitales el desarrollo de estándares que permitan el intercambio de datos. La iniciativa de archivos abiertos (OAI), desarrolla y promueve estándares de interoperatibilidad con el objetivo de facilitar la eficiente recuperación de contenido. *Phronesis, CIRIA y Redalyc son ejemplos de bibliotecas digitales que cumplen con los estándares de interoperatibilidad de OAI.*

2.2. Iniciativa de archivos abiertos.

La iniciativa de archivos abiertos (OAI) es una organización formada por instituciones, investigadores, bibliotecarios, editores y archivistas con el objetivo de crear estándares que permitan la interoperatibilidad en bibliotecas digitales [12]. Dublin Core (DC) es el formato para metadatos recomendado por OAI usado para describir el contenido, nombre, titulo y otras características de los recursos disponibles en una biblioteca digital, DC es ampliamente usado en bibliotecas digitales federadas las cuales ofrecen colecciones descentralizadas de datos que son accedidas a través de servicios remotos.

Existen dos tipos de participantes en OAI, los proveedores de información, que publican sus metadatos y sus recursos a través de registros, y los proveedores de servicios que usan los registros ofrecidos para generar un servicio de valor agregado. Las propiedades más sobresalientes de OAI son [15]:

- **Autonomía**. Cada proveedor de información tiene sus propias políticas y administración
- **Descentralización**. Los proveedores de informacíón no tienen que reportar si sus colecciones son actualizadas o si hay cambios en sus registros.
- **Dinamismo y cantidad de información**. Nuevos miembros se incorporan frecuentemente.
- **Independencia de origen**. Las colecciones ofrecidas por los proveedores de información son construidas como respuesta a las necesidades de una comunidad en particular de manera independiente.

La figura 2.1 muestra los elementos `<dc:title>`, `<dc:description>` y `<dc:creator>` que almacenan información del contenido del documento para un registro típico de un documento digital tomada de la colección de Phronesis. El formato XML[1] es usado para codificar los registros, cada registro es asociado a un único identificador y corresponde a un documento, la estructura del registro es la siguiente:

- **Encabezado**. El cual contiene el identificador al cual está asociado el documento.
- **Metadatos**. Los cuales describen el documento.

OAI propone un mecanismo de bajo nivel llamado protocolo para cosecha de metadatos (OAI-PMH), que permite soportar la interoperatibilidad independiente de la aplicación.

[1]http://www.w3.org/XML/ *Lenguaje de marcado extensible*, consultado el 5 de abril de 2010.

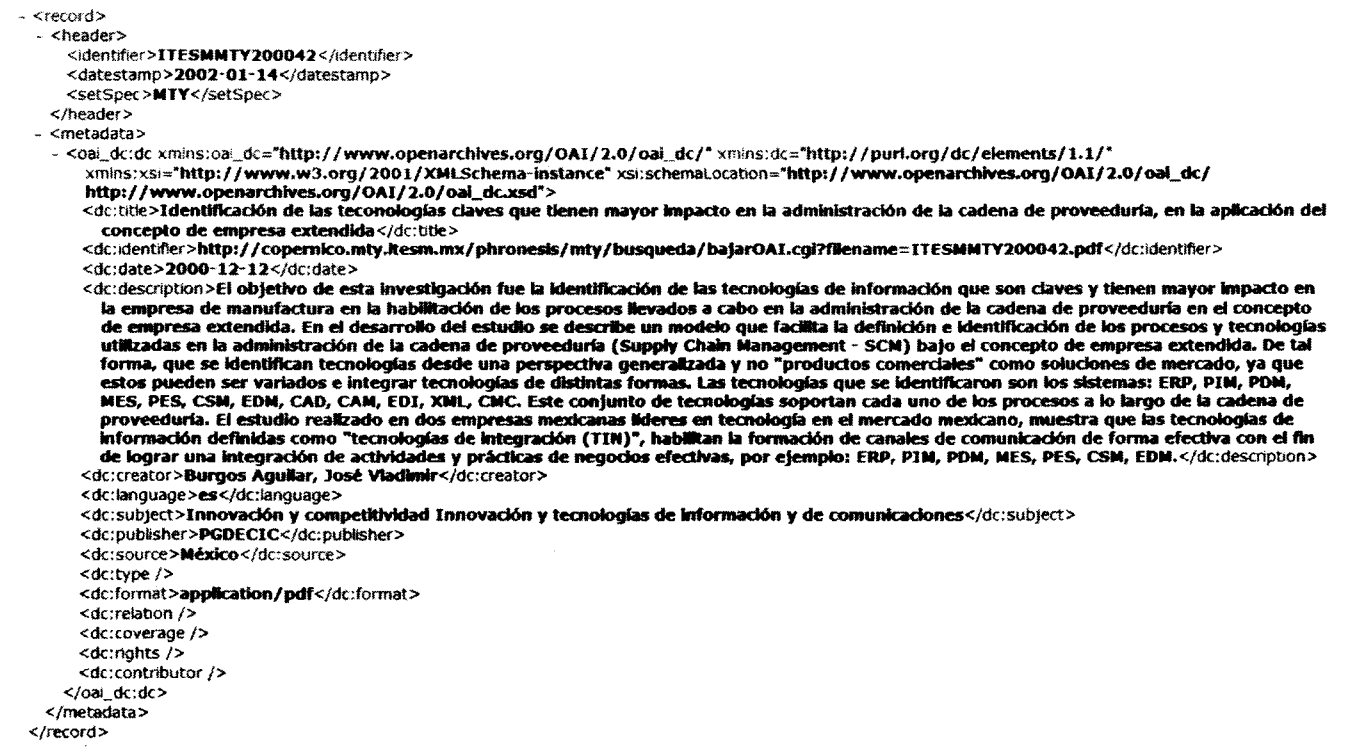


Figura 2.1: Ejemplo de un registro típico de una tesis digital de la colección de Phronesis.

2.2.1. Protocolo OAI-PMH.

El protocolo OAI-PMH proporciona acceso externo a los registros de una colección. Trabajaremos con la versión 2.0 que utiliza los metadatos DC por defecto. [22].

El protocolo OAI-PMH define las solicitudes de verbos que son utilizadas por los proveedores de servicio para cosechar registros. La tabla 2.1 muestra el contenido recuperado por cada verbo solicitado. Los proveedores de datos responden a estas solicitudes con documentos en formato XML.

Solicitud de verbo	Objetos obtenidos
GetRecord	Un único registro de metadatos
Indentify	Información acerca del proveedor de datos
ListIdentifiers	Cabeceras de los registros
ListMetadaFormat	Formatos de metadatos disponibles
ListRecords	Un listado de Registros
ListSets	Estructura establecida

Tabla 2.1: Solicitud de los verbos del protocolo OAI-PMH

Como ejemplo, la siguiente solicitud usa el verbo GetRecord para recuperar el registro de la figura 2.1:

`http://copernico.mty.itesm.mx:4000/request?verb=GetRecord`
`&identifier=ITESMMTY200042&metadaPrefix=oai_dc`

Y sus componentes son los siguientes:

El proveedor de datos es: `http://copernico.mty.itesm.mx:4000/request`
El verbo solicitado es: GetRecord
El identificador que indica que registro obtener es: `identifier=`
`ITESMMTY200042`
La especificación del formato para los metadatos es: `metadaPrefix=oai_dc`

Como se observa en la tabla 2.1, la solicitud de verbos en el protocolo OAI-PMH no implementa medidas de similitud entre registros y las solicitudes tienen como objetivo cosechar únicamente los registros indicados. En este trabajo buscamos precisamente mejorar estas limitantes.

2.3. Agrupamiento de documentos.

El agrupamiento de documentos ha sido estudiado intensivamente a causa de su amplia aplicabilidad en aéreas como la minería web, motores de búsqueda, recuperación de información y análisis topológico.

El primer reto es determinar que características de un documento son consideradas discriminatorias, la mayoría de los métodos de agrupamiento representan a cada documento como un vector. Bajo este modelo vectorial una colección de n documentos cada uno con m términos es representada en una matriz termino - documento de m x n (cada documento es un vector de m términos), comúnmente a cada vector se asocian a un peso que refleja la frecuencia de los términos en el documento multiplicado por el inverso de su frecuencia en toda la colección (Tf - Idf) [2]. La razón es que las palabras que ocurren frecuentemente en un documento pero raramente en toda la colección tiene un alto grado de poder discriminativo. Los vectores representantes de cada documento pueden tener una alta dimensión para evitar esto una serie de pasos de pre procesamiento son realizados: *Filtrado* (remover caracteres especiales), *Tokenización* (Dividir frases en palabras individuales), *Lematización* (Reducir las palabras a su forma base), *Remover Stop-words* (Palabras sin relevancia semántica) y *Poda* (Remover palabras con muy baja frecuencia a través de toda la colección, podrían formar agrupamientos muy pequeños).

El agrupamiento de documentos implica el uso de una función de similitud, la cual automáticamente agrupa documentos que tienen una alta similitud de contenido dentro de un grupo, mientras que los documentos que pertenecen a otros grupos son

diferentes en su contenido [17]. El uso de estas técnicas es un caso particular de los problemas de aprendizaje no supervisados [18].

Los algoritmos de agrupamiento se pueden clasificar en [14]:

- **Agrupamiento exclusivo**. Los elementos son agrupados de manera exclusiva, por lo tanto, cierto elemento pertenece a un grupo y no puede ser incluido a algún otro. Un ejemplo es el algoritmo K- means, explicado más adelante.

- **Agrupamiento con superposición**. Los elementos pertenecen a grupos difusos, por lo tanto, los elementos pueden pertenecer a uno o más grupos con diferentes grado de pertenencia. Un ejemplo es el algoritmo Fuzzy c-means, explicado más adelante.

- **Agrupamiento jerárquico**. Se basa en la unión de dos grupos cercanos y después de ciertas iteraciones se obtienen finalmente los grupos requeridos. El agrupamiento final corresponde a una estructura de árbol. Dependiendo de cómo un árbol de agrupaciones es construido, los algoritmos jerárquicos se dividen en aglomerativos y divisibles. Un algoritmo aglomerativo comienza con una posible agrupación y luego ajusta los grupos, mientras que uno divisible comienza con una partición muy grande y la va dividiendo en grupos más pequeños que se obtienen a partir del grupo inicial [17]. Un ejemplo es el algoritmo de agrupamiento jerárquico basado en la frecuencia de un conjunto de elementos (FIHC), explicado más a detalle en la sección 2.3.1.

- **Agrupamiento probabilístico**. Este tipo de agrupamiento asocia los elementos con grupo a través de un enfoque completamente probabilístico. Un ejemplo es al modelo de mezcla Gaussiano, en el cual los grupos se pueden considerar como distribuciones Gaussianas centradas en relación a su baricentro.

Históricamente los algoritmos jerárquicos, exclusivos y con superposición han dominado los métodos de agrupamientos [2]. A continuación se describen las etapas de los algoritmos k-means y fuzzy c means:

- **Algoritmo k-means**. El algoritmo k-means agrupa el conjunto de elementos en k grupos, cada elemento únicamente puede estar asociado con un grupo.

  El primer paso es definir k centroides, uno por cada grupo. Estos centroides deben ser elegidos cuidadosamente ya que de su elección dependen los grupos resultantes, la mejor opción es elegir como centroides elementos lejanos entre sí.

  El segundo paso es tomar cada elemento que pertenece al conjunto de elementos y asociarlo con el centroide más cercano. Cuando todos los elementos han sido asociados, el primer paso ha terminado.

El tercer paso se recalcula la posición de k nuevos centroides de acuerdo a los baricentros de cada grupo obtenido durante el segundo paso. Se realizan de nuevo el segundo y tercer paso hasta que no haya cambios en la selección de los k nuevos centroides. La finalidad es minimizar la función objetivo 2.1, que es un indicador de la distancia para los n elementos en relación a sus respectivos k centros de grupo.:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \tag{2.1}$$

Donde $\left\| x_i^{(j)} - c_j \right\|^2$ es una medida de distancia elegida entre el elemento $x_i^{(j)}$ y el centroide del grupo c_j.

- **Algoritmo fuzzy c means**. El algoritmo fuzzy c-mens permite que los elementos pertenezcan a dos o más grupos. Como resultado del fuzzy c-means cada elemento tiene un grado de pertenencia con cada grupo, representado por su centro de grupo.

  El primer paso es la inicialización, para la descripción del algoritmo utilizaremos la siguiente notación:

  - Número de grupos a encontrar: c
  - Número de elementos a agrupar: J
  - Vector de atributos del elemento j : x_j, $j = 1, \ldots, J$
  - Grado de pertenencia del objeto j a clase i : μ_{ij}, $i = 1, \ldots, c$; $j = 1, \ldots, J$

  Sea $\theta^{(0)}$ una matriz (c x J) con el elemento μ_{ij} en posición (i, j), $i = 1, \ldots, c$; $j = 1, \ldots, J$. Esta matriz se inicializa en forma aleatoria con la función de restricción 2.2:

$$\sum_{j=1}^{C} = 1, \forall j = 1, \ldots, J \tag{2.2}$$

  El segundo paso es realizar el cálculo de los centros de grupo. Dados los valores de pertenencia μ_{ij}, los centros v_i de cada grupo i están dados por la función 2.3:

$$v_i = \frac{\sum_{j=1}^{J} (\mu_{ij})^m x_j}{\sum_{j=1}^{J} (\mu_{ij})^m}, \forall j = 1, \ldots, J \tag{2.3}$$

  El parámetro m, utilizado en la función anterior, se llama difusor (fuzzifier) y determina el grado de difusión (fuzziness) para los grupos encontrados ($1 < m <$

∞). Para m "cercano a 1"se calcula una solución con grupos no-difusos (crisp); mientras mayor sea m más difusa se hace la solución.

El tercer paso es la actualización de los valores de pertenencia. Dados los centros calculados en el segundo paso, los valores de pertenencia μ_{ij} son actualizados utilizando la siguiente función 2.4:

$$\mu_{ij} = (\sum_{k=1}^{c} (\frac{d_{ij}}{d_{kj}})^{\frac{2}{m-1}}), \forall i = 1, \ldots, c, \forall j = 1, \ldots, J \tag{2.4}$$

El valor d_{ij} es la distancia entre el elemento j y el centro v_i del grupo i. En el cálculo de esta distancia se utilizan los centros de grupo v_i obtenidos en el segundo paso.

El cuarto paso es el criterio de detención, donde el segundo y tercer paso se repiten de manera iterativa hasta cumplir el siguiente criterio de detención expresado por la función 2.5:

$$\left\| \theta^{(t+1)} - \theta^{(t)} \right\| \leq \epsilon \tag{2.5}$$

Donde $\theta^{(t)}$ es la es la matriz de los valores de pertenencia en la iteración t y ϵ es un umbral a ser determinado por el usuario. En términos de resultados, el algoritmo fuzzy c-means obtiene v_i centros de grupo para los c grupos, así como los μ_{ij} valores de pertenencia de cada elemento j a agrupar en relación a cada grupo i.

Aunque técnicas de agrupamiento como K-means pueden ser aplicadas, estas usualmente no satisfacen los requerimientos para la agrupación de documentos [9]: alta dimensionalidad, aplicables a grandes volúmenes de datos, facilidad de búsqueda y etiquetado de grupos.

El algoritmo FIHC muestra gran precisión, más eficiencia y mayor escalabilidad que k-means y otros algoritmos de agrupamiento [8].

El objetivo del agrupamiento es determinar un agrupamiento intrínseco de un conjunto de datos no etiquetados. El criterio para decidir si se obtuvo un buen agrupamiento debe ser proporcionado por el usuario de manera que el resultado del agrupamiento se adapte a sus necesidades [14]. La incorporación de técnicas de agrupamiento de documentos aplicadas a repositorios de OAI, ayuda a satisfacer las necesidades de información de los usuarios de una biblioteca digital [15]. Trabajos previos sobre colecciones muestran que los algoritmos jerárquicos de agrupamiento producen grupos de alta calidad [17] [27].

2.3.1. Algoritmo FIHC.

El agrupamiento jerárquico basado en la frecuencia de un conjunto de elementos (FIHC), es un algoritmo jerárquico aglomerativo propuesto por [9], este algoritmo se basa en la hipótesis de que si un grupo de documentos se refieren al mismo tema, ellos podrían compartir un conjunto de elementos, estos elementos en común son llamados conjuntos de elementos frecuentes, documentos de diferentes grupos tienen pocos conjuntos de elementos frecuentes en común.

El algoritmo FIHC hace uso de los conceptos que se definen a continuación:

- *Conjunto global de elementos frecuentes.* Es un conjunto de elementos que aparecen juntos en más de un porcentaje mínimo en toda la colección de documentos.
- *Soporte global.* Porcentaje mínimo para determinar qué términos pertenecen al conjunto global de elementos frecuentes.
- *Elemento de frecuencia global.* Se refiere a un elemento que pertenece al conjunto global de elementos frecuentes. Un conjunto global de elementos frecuentes que contiene k elementos es llamado un conjunto de $k - elementos$ frecuentes.

Una visión general de las etapas del algoritmo FIHC se muestra en la figura 2.2, donde se menciona: el modelado vectorial de los documentos basándose en su contenido, la obtención de los elementos frecuentes a partir de los términos en común entre documentos, el uso de los elementos frecuentes para la reducción de los modelos vectoriales, la construcción de los grupos, la construcción del árbol, la etapa de reducciones (grupos similares, hijos y hermanos) y la obtención del árbol de agrupamiento.

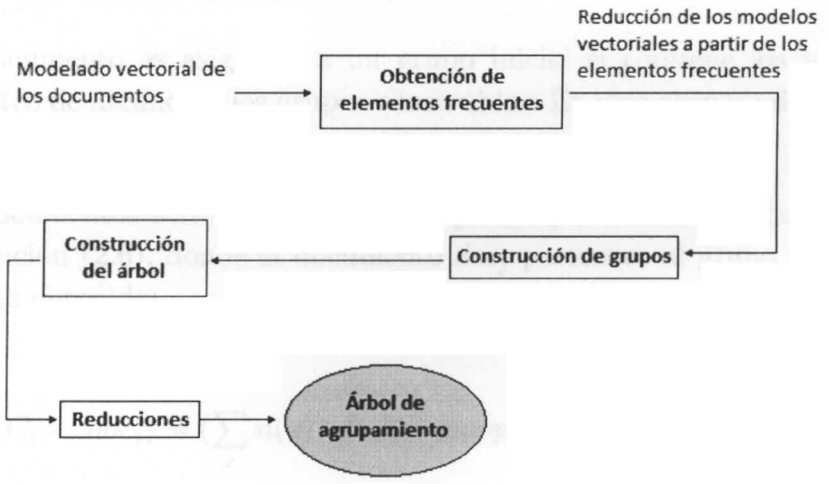


Figura 2.2: Visión general del algoritmo FIHC.

Reducción de la representación vectorial de los documentos

Similar a la mayoría de los algoritmos de agrupamiento, FIHC incluye un pre-procesamiento que involucra el remover las palabras que por su frecuencia y/o semántica no poseen valor discriminatorio entre documentos (artículos, pronombres y preposiciones), caracteres de formato y caracteres de estilo.

Posterior al pre-procesamiento cada documento es representado a través de un vector. Bajo este modelo vectorial una colección de n documentos cada uno con m términos es representada en una matriz termino - documento de m x n (cada documento es un vector de m términos).

Con el propósito de reducir la dimensionalidad del conjunto de documentos a agrupar se hace uso del *soporte global.* De manera que cada uno de los vectores representativos de los documentos únicamente conserva los *elementos de frecuencia global* y la dimensionalidad del conjunto de documentos se reduce a la dimensionalidad de los $k - elementos$ frecuentes del *conjunto global de elementos frecuentes.*

Construcción de los grupos del árbol de agrupamiento

El algoritmo FIHC sigue los pasos descritos a continuación para la construcción de los grupos del árbol de agrupamiento:

Un elemento de frecuencia global es considerado dentro de un grupo C_i, si el elemento se encuentra en un porcentaje mínimo de los documentos de C_i, este porcentaje es llamado soporte de grupo de un elemento.

1. Un grupo inicial es construido por cada conjunto global de elementos frecuentes, las etiquetas de los grupos iniciales son tomadas de los términos de cada conjunto global de elementos frecuentes.

2. Un documento es asignado a un grupo inicial si contiene los términos de las etiquetas de alguno de los grupos construidos. En este paso un documento podría pertenecer a varios grupos.

3. Los documentos son reasignados al mejor grupo usando la siguiente función de puntuación (2.6), donde el documento doc_j pertenece al grupo C_i de acuerdo al puntaje obtenido:

$$Score(C_i \leftarrow doc_j) = (\sum_{x} n(x) * cluster\_support(x)) - (\sum_{x'} n(x') * cluster\_support(x')) \tag{2.6}$$

De la función anterior tenemos que x es un elemento de frecuencia global en el documento doc_j que también es frecuente en el grupo C_i, x' es un elemento de

frecuencia global en doc_j, que no es frecuente en el grupo C_i, $n(x)$ es el peso de la frecuencia de x en el documento doc_j, $n'(x)$ es el peso de la frecuencia de x' en el documento doc_j.

4. Re calcular para cada grupo los elementos frecuentes del grupo.

5. Reducir el árbol de agrupamiento. Este proceso consiste en unir grupos similares, la similitud del grupo C_j con el grupo C_i se basa en los documentos que se agrupan en C_j y se obtiene con la siguiente función (2.7):

$$Sim(C_i \leftarrow C_j) = \frac{Score(C_i \leftarrow doc(C_j))}{\sum_x n(x) + \sum_{x'} n(x')} + 1 \tag{2.7}$$

Sean C_i y C_j dos grupos, tenemos que $doc(C_j)$ representa la combinación de todos los documentos del subárbol C_j en un solo documento, x representa un elemento de frecuencia global en $doc(C_j)$, que también es frecuente en el grupo C_i , x' representa un elemento de frecuencia global en $doc(C_j)$, que no es frecuente en el grupo C_i. $n(x)$ es el peso de la frecuencia de x en el documento doc_j, $n'(x)$ es el peso de la frecuencia de x' en el documento doc_j.

La similitud entre grupos C_i y C_j se define como la media geométrica de $sim(C_i \leftarrow C_j)$ y $sim(C_j \leftarrow C_i)$ (2.8):

$$InterSim(C_i \leftrightarrow C_j) = [Sim(C_i \leftarrow C_j) * Sim(C_j \leftarrow C_i)]^{\frac{1}{2}} \tag{2.8}$$

6. Reducción de hijos. Recorriendo el árbol de abajo hacia arriba, para cada nodo que no sea hoja y que se encuentre en el segundo nivel o un nivel superior, la similitud entre grupos es calculada para cada hijo de este nodo. El grupo del hijo es combinado con el grupo del padre si tienen mucha similitud.

7. Reducción de hermanos. Se aplica el cálculo de la similitud para todos los grupos con el objetivo de alcanzar el número de grupos especificado por el usuario.

Construcción del árbol de agrupamiento

La construcción del árbol de agrupamiento consiste en:

1. Ordenar todos los grupos en orden alfabético.

2. Para cada grupo C_i:

   a) Saltar C_i si esta vacio y no tiene hijos.

b) Coloca los $k - elementos$ en la etiqueta del grupo C_i.

c) Busca todos los grupos con $k - 1elementos$ en su etiqueta de grupo. Estos grupos son llamados padres potenciales del grupo C_i.

d) Mezcla todos los documentos del subárbol C_i en un único grupo combinado.

e) Calcula la función de puntuación para $doc(C_i)$ con respecto a cada padre potencial.

f) Selecciona como padre del grupo C_i al padre potencial que tenga la puntuación más alta.

Características del árbol de agrupamiento

Las características del árbol de agrupamiento son las siguientes:

1. Los grupos del $k - iesimo$ nivel en el árbol tienen igual número de $k - elementos$ en sus etiquetas.
2. Las etiquetas son formadas con los términos más representativos del documento de acuerdo a la función de puntuación.
3. Todos los documentos de un grupo contienen los términos de la etiqueta del grupo.
4. Existe un grupo especial cuya etiqueta es nula, los documentos bajo esta etiqueta tienen como padre la raíz del árbol.

La figura 2.3 muestra una representación de un árbol de agrupamiento, los términos en los rectángulos representan las etiquetas del grupo, y los iconos a un lado de los rectángulos representan los documentos que pertenecen al grupo. El árbol tiene 7 grupos en 4 niveles, contados a partir del nivel cero. Un grupo que se encuentra en el nivel tres, tiene una etiqueta formado por igual número de términos.

2.4. Construcción de ontologías.

Una ontología es una poderosa manera de representar el conocimiento para múltiples propósitos. Para fines científicos los documentos son la fuente primaria y el medio de comunicación para la difusión del conocimiento humano. En la última década, las ontologías se han convertido en uno de los métodos de modelación más populares para las taxonomías, clasificaciones y otras estructuras inteligentes. Desafortunadamente, existe una brecha sorprendentemente grande entre el conocimiento modelado a través de la ontología y el texto que documenta el mismo conocimiento. Repositorios grandes de documentos se pueden beneficiar del uso de ontologías para las tareas de búsqueda

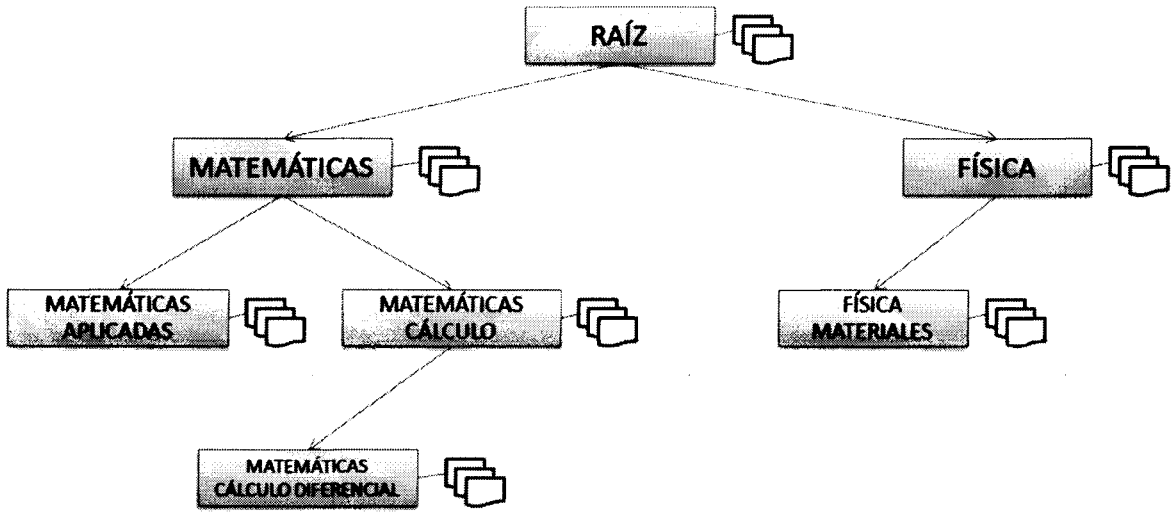


Figura 2.3: Representación de un árbol de agrupamiento construido por el algoritmo FIHC.

y recuperación de información. La semántica de documentos está encaminada a combinar documentos y ontologías, permite a los usuarios acceder a los conocimientos de múltiples maneras, su objetivo no se limita a proporcionar metadatos de documentos (palabras clave o metadatos Dublin Core) [6].

Las ontologías pueden ser creadas manualmente por expertos en un dominio o de manera semiautomática. A continuación se presenta el método llamado OntOAIr , un método para la construcción semiautomática de ontologías enfocado a proveedores de datos que cumplen los estándares de la iniciativa de archivos abiertos (OAI, por sus siglas en ingles) [15].

2.4.1. OntOAIr.

Al ser OntOAIr un método para la construcción semiautomática de ontologías enfocado a proveedores de datos que cumplen con los estándares de OAI, permite obtener ontologías de registros, las cuales son estructuras jerárquicas formadas por grupos disjuntos de registros, las ontologías de registros tienen dos usos principales [15]:

- Organizar los registros basados en sus contenidos.
- Soportar las tareas de recuperación de información para los múltiples proveedores de datos.

El proceso de construcción de la ontología de registros se realiza en cuatro etapas:

1. Cosecha.
2. Representación.
3. Agrupamiento.
4. Formalización.

1. Cosecha.

La actividad de cosecha obtiene los documentos de las colecciones digitales, usa los verbos listados en la tabla 2.1 para obtener de los proveedores de datos los registros de metadatos. Esta es una actividad repetitiva y que consume mucho tiempo debido a que los proveedores de datos típicamente listan cientos o miles de registros disponibles a través de internet.

2. Representación.

En esta actividad se construye una representación vectorial de cada documento cosechado. Una vez que los documentos han sido cosechados una representación simplificada de los mismos se produce. Para lograr lo anterior se trabaja con los metadatos Dublin Core, la información contenida en los elementos dc:title y dc:description es extraída debido a que contienen tanto el título como la descripción del recurso. La tabla 2.2 muestra la definición y un breve comentario asociado al elemento Dublin Core [15].

Etiqueta	**Definición**	**Comentario**
dc:description	El contenido del recurso	La descripción puede incluir, pero no está limitada: un resumen, una tabla de contenidos, o una representación del recurso.
dc:title	El nombre dado al recurso	Este puede ser un nombre por el cual el recurso es formalmente conocido

Tabla 2.2: Elementos Dublin Core con información de los registros

Los elementos Dublin Core contienen texto libre. Antes de procesarlos cualquier dato sensitivo es removido (artículos, preposiciones y palabras sin relevancia semántica). Posteriormente, los vectores de características son generados. Un vector de características es una representación simplificada de un registro formada por palabras claves y pesos (valores numéricos que representan la relevancia de las palabras claves)[8].

El vector de características de OntOAIr agrega dos elementos: un identificador que sirve para diferenciar cada registro y el URL del recurso proporcionado por el proveedor de datos para permitir su ubicación. Ambos elementos hacen único a un registro de OAI, la tabla 2.3 muestra un vector de características típico. Los pesos muestran que algunas palabras son más representativas para el registro que otras. El factor TF - IDF (Frecuencia de Termino en cada documentos - Inverso de la Frecuencia del término en la colección de Documentos) es usado como peso de las palabras claves [15].

Identifier	oai:thesisUDLAP:98	
URL	http://ict.udlap.mx:9090/tales	
Keyword	**TF**	**IDF**
digital	1	10.397
libraries	3	9.133
reference	1	6.089
services	1	7.783

Tabla 2.3: Ejemplos del vector de características de OntOAIr

3. Agrupamiento.

El agrupamiento consiste obtener grupos de vectores de características similares, el algoritmo utilizado en OntOAIr es una adaptación del algoritmo de agrupamiento jerárquico FIHC [8]. Las etiquetas de los grupos obtenidos permiten construir un vocabulario que describe los tópicos principales de una colección de documentos.

La adaptación del algoritmo FIHC consiste en la eliminación de dos de los pasos del algoritmo original al momento de la construcción del árbol de agrupamiento [15]:

- Reducción de hijos.
- Reducción de hermanos.

4. Formalización.

La formalización se refiere a la representación del árbol de agrupamiento en un lenguaje accesible a máquina. La representación se realiza a través de una estructura que involucra dos secciones principales [15]:

- **Jerarquía del grupo**. El concepto central es el grupo como tal, destinado a capturar la noción de cualquier cosa que involucre un conjunto de registros.
- **Descripción de registro**. En concepto central en esta sección es el registro, el cual representa a un único documento de la colección.

La representación hace uso de OWL[2] que es el lenguaje de marcado para publicar datos en internet usando ontologías. Específicamente se emplea el sub lenguaje OWL DL, llamado así debido a su correspondencia con la lógica de descripción, OWL DL esta orientado a casos donde es necesario el máximo de expresividad sin perder la integridad computacional de los sistemas de razonamiento.

Parte del conocimiento relacionado con la ontología que puede ser representado en OWL DL se describe en las siguientes sentencias:

- Una ontología de registros *es formada por* grupos (un grupo *es parte de* una ontología de registros).
- Un grupo *contiene* registros (Un registro *se encuentra* en un grupo).
- Un grupo *es descrito por* una etiqueta (Una etiqueta *describe* a un grupo).
- Un grupo *tiene un* nivel en la ontología.
- Un registro *tiene un* título, tema, descripción, identificador, URL, proveedor de datos, formato de metadatos y elementos de marca de fecha.
- Las propiedades *es formada por* y *contiene* son transitivas.

En la lista anterior, las palabras con letra itálica son nombres de propiedades y las sentencias en paréntesis representan la propiedad inversa.

La figura 2.4 muestra un extracto de la ontología de registros en el lenguaje OWL [15].

2.5. Redes sociales.

El análisis de redes sociales ha sido de gran interés en los recientes años y juega un importante papel en muchas áreas [3][16] [19]. Un ejemplo popular es el proyecto Bacon de Oracle [23], el cual determina la distancia entre cualquier actor y Kevin Bacon permitiendo examinar las relaciones entre los co-protagonistas de una película. Este ejemplo demuestra la utilidad que puede surgir de la adaptación del concepto de relación en el análisis una red social para un dominio de interés.

El análisis de una red social se basa en la premisa de que las relaciones entre los actores de una red social pueden ser descritas por un grafo. Los nodos del grafo representan a los actores y las aristas que conectan a un par de nodos representan la interacción entre estos actores [26].

Para [13] la representación de una red social de co-autoría asume la existencia de un grafo en el cual se hace uso de los términos nodo, actor o autor de manera indistinta, y también de manera indistinta hace uso de los términos arista, relación o co-autoría.

[2]http://www.w3.org/TR/owl-features/ *sitio de OWL*, consultado el 25 de junio de 2010.

```
<?xml version="1.0"?> <rdf:RDF
  xmlns:rdf=''http://www.w3.org/1999/02/22-rdf-syntax-ns#''
  xmlns:xsd=''http://www.w3.org/2001/XMLSchema#''
  xmlns:rdfs=''http://www.w3.org/2000/01/rdf-schema#''
  xmlns:owl=''http://www.w3.org/2002/07/owl#''>

<owl:Ontology rdf:about=''''>
  <rdfs:label>Ontology of records</rdfs:label>
  <owl:versionInfo>August 2007 2.1</owl:versionInfo>
</owl:Ontology>
<owl:Class rdf:ID=''OntologyOfRecords''/>

<owl:Class rdf:ID=''Cluster''>
  <rdfs:comment>Clusters form an ontology of records</rdfs:comment>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:allValuesFrom>
   <owl:Class rdf:about=''#OntologyOfRecords''/>
      </owl:allValuesFrom>
      <owl:onProperty>
   <owl:ObjectProperty rdf:ID=''isFormedBy''/>
      </owl:onProperty>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

Figura 2.4: Extracto de la ontología de registros en el lenguaje OWL

2.5.1. Red de co-aoutoría.

Una red de co-autoría es un tipo importante de red social que ha sido usada extensivamente para determinar la estructura de las colaboraciones científicas y el estatus de los autores de manera individual [13].

En [13] se presentan tres modelos para las redes de co-autoría:

- **Grafo binario no dirigido**. Se muestra en la figura 2.5 y asume que si dos autores participan en un mismo artículo una arista entre ellos es creada con un peso unitario, donde v_i representa a un autor.

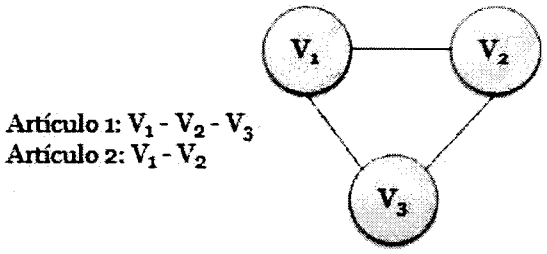


Figura 2.5: Grafo binario no dirigido.

- **Grafo binario dirigido**. Se muestra en la figura 2.6 y para convertir un grafo no dirigido en un grafo dirigido se tiene la hipótesis presentada a continuación:

  1. Cualquier red no dirigida puede ser representada como una red dirigida agregando enlaces simétricos en nodos con una conexión.
  2. Nodos simétricos representan la cooperación mutua entre autores.
  3. El peso de la arista es un valor binario que indica la ausencia o presencia de dos nodos simétricos.

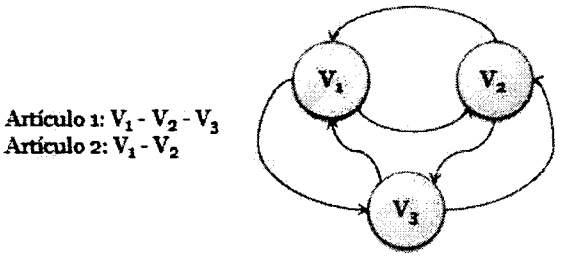


Figura 2.6: Grafo binario dirigido.

- **Grafo dirigido con pesos normalizados**. En muchos casos la noción de magnitud en una red binaria no concuerda con lo que indica el sentido común. Por

ejemplo: Si dos autores publican juntos en varios artículos, ¿su relación podría considerarse más importante que la relación entre aquellos autores que han publicado juntos con menos frecuencia? También, si un artículo tiene dos autores y otro artículo tiene cientos de autores ¿podrían considerarse los autores del primer artículo más conectados que los del segundo?

Sea G el grafo de co-autoría denotado por $G = (V, E, W)$, donde V es el conjunto de nodos (autores), E es el conjunto de aristas (relaciones entre autores) y W el conjunto de pesos W_{ij} asociados con cada arista que conecta un par de autores (v_i, v_j).

Para determinar la magnitud del enlace entre dos autores se toman en cuenta dos factores:

1. Frecuencia de co-autoría. Autores que frecuentemente publican juntos podrían tener más peso de co-autoría.
2. Número total de co-autores en artículos. Si un artículo tiene muchos autores, cada relación entre dos autores tendrá menos peso.

Ahora podemos determinar el peso de los enlaces de co-autoría. Sea $V = \{v_1, \cdots, v_n\}$ un conjunto de n autores. Sea $A = \{a_1, \cdots, a_k, \cdots, a_m\}$ un conjunto de m artículos y $f(a_k)$ el número de autores del artículo a_k, definimos los siguientes conceptos:

- *Exclusividad.* Si los autores v_i y v_j son co-autores en el artículo a_k.

$$g_{i,j,k} = \frac{1}{f(a_k) - 1} \tag{2.9}$$

  En la función 2.9 $g_{i,j,k}$ representa el grado de exclusividad en la relación de co-autoría entre v_i y v_j en un artículo en particular. Esta definición da más peso a las relaciones de co-autoría en los artículos publicados por pocos autores, y menos peso a las relaciones de co-autoría en los artículos publicados por muchos autores.

- *Frecuencia de co-autoría.* La frecuencia de co-autoría (función 2.10) consiste en la suma de todos los valores $g_{i,j,k}$ para todos los artículos publicados por v_i y v_j. Dando más peso a los autores que publiquen juntos en más ocasiones, y por lo tanto más exclusividad.

$$c_{ij} = \sum_{k=1}^{m} g_{i,j,k} \tag{2.10}$$

- *Pesos normalizados.* La normalización (función 2.11) asegura que los pesos de las relaciones de un autor sumen uno.

$$w_{ij} = \frac{c_{ij}}{\sum_{k=1}^{n} c_{i,k}} \tag{2.11}$$

Donde w_{ij} es el peso normalizado entre el autor v_i y v_j, c_{ij} es la frecuencia de co-autoría entre el autor v_i y v_j, y $\sum_{k=1}^{n} c_{i,k}$ representa la sumatoria de las frecuencias de co-autoría entre el autor v_i y cada uno de los autores v_k con los que se relaciona.

La red que se obtiene a través de un grafo dirigido con pesos normalizados se ejemplifica en la figura 2.7.

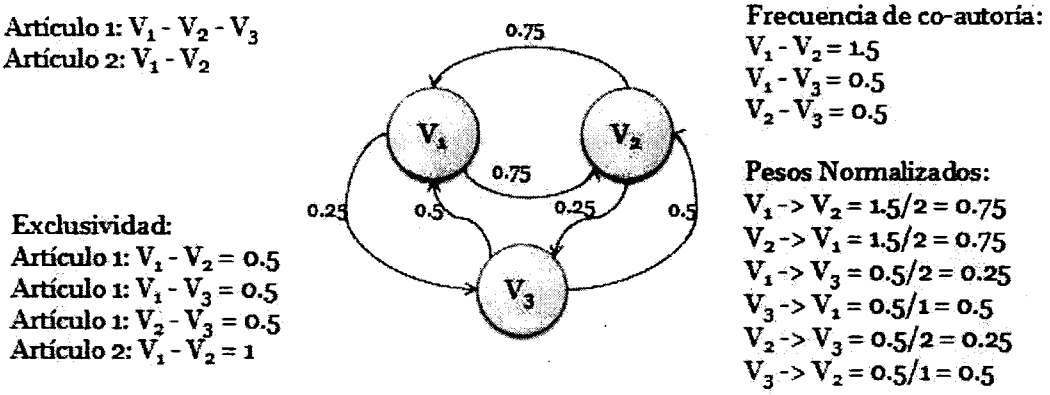


Figura 2.7: Grafo dirigido con pesos normalizados.

2.6. Resumen

Durante el desarrollo de este capítulo se describió el concepto de biblioteca digital [5], el conjunto de características que definen una biblioteca digital [25] y los servicios que ofrecen las bibliotecas digitales [1]. *Bajo el contexto de biblioteca digital este trabajo se centra en documentos digitales de contenido textual, específicamente en las colecciones de documentos provenientes de Phronesis, CIRIA y Redalyc.*

Debido a que en la comunidad de bibliotecas digitales son de especial interés los estándares que permitan el intercambio de datos, en este capítulo se hace mención de la iniciativa de archivos abiertos (OAI), que desarrolla y promueve estándares de interoperatibilidad con el objetivo de facilitar la eficiente recuperación de contenido [12]. Se detallan las características del protocolo OAI-PMH [22], el cual es un mecanismo propuesto por OAI para la cosecha de los registros de una colección. *Phronesis, CIRIA y Redalyc son ejemplos de bibliotecas digitales que cumplen con los estándares de interoperatibilidad de OAI.*

En el capítulo también se hace mención de los algoritmos de agrupamiento de documentos y su clasificación [17], [14]. Lo anterior se debe a que en este trabajo de tesis uno de los objetivos es la obtención de grupos de documentos con una alta similitud de contenido.

Como base para la obtención de las estructuras que representan las redes temáticas de colaboración podemos clasificar los trabajos relacionados en dos grupos:

- Construcción de ontologías, a partir de documentos procedentes de repositorios digitales. En este sentido se hace mención del trabajo llamado OntOAIr [15], *el cual permite identificar la afinidad temática de los documentos y obtener un agrupamiento jerárquico en una estructura de árbol, lo que facilita la clasificación y búsqueda de documentos provenientes de una biblioteca digital, sin embargo esta estructura no puede ser representada como una red de colaboración.*

- Redes sociales, específicamente estructuras de colaboración en comunidades científicas. Por lo que se hace mención del trabajo relacionado con la red de co-autoría, la cual ha sido usa para determinar la estructura de las colaboraciones científicas [13]. *En este trabajo se exploran ampliamente las relaciones de co-autoría, modelando dicha relación a través de un grafo binario no dirigido cuyos enlaces representan las relaciones de co-autoría, a diferencia del modelado de redes de colaboración propuesto en este trabajo, el cual está basado en un grafo dirigido con pesos normalizados cuyos enlaces representen las relaciones de afinidad temática entre documentos.*

En el siguiente capítulo se presenta el modelo de solución propuesto para la inferencia de redes temáticas de colaboración a partir del contenido de los documentos de las bibliotecas digitales (Phronesis, CIRIA y Redalyc). Con el objetivo de obtener las redes temáticas de colaboración, la metodología descrita en el capítulo 3 hace del uso algoritmo de agrupamiento jerárquico basado en la frecuencia de un conjunto de elementos (FIHC), del protocolo OAI-PMH y describe las métricas para inferir las redes temáticas de colaboración.

Capítulo 3

Inferencia de Redes Temáticas de Colaboración.

En este capítulo se presenta el modelo de solución propuesto para la inferencia de redes temáticas de colaboración. Para inferir las redes temáticas de colaboración a partir del contenido de los documentos de las bibliotecas digitales (Phronesis, CIRIA y Redalyc), se utiliza como fuente de datos la información contenida en los metadatos Dublin Core, los cuales se obtienen utilizando el protocolo OAI-PMH que proporciona el servicio de cosecha de los registros pertenecientes a una biblioteca digital. Lo anterior se basa en el hecho de que a través del protocolo OAI-PMH se puede tener acceso externo a los registros de una colección de documentos de una biblioteca digital que se apegue a los estándares de interoperatibilidad de la iniciativa de archivos abiertos (OAI).

3.1. Metodología.

Tomando como punto de partida el escenario antes mencionado, se propone un modelo de solución basado en tres etapas, en la figura 3.1 se describe la metodología a seguir.

En las siguientes secciones se explica a detalle cada uno de los elementos de la metodología ilustrados en la figura 3.1.

3.1.1. Recolección, pre-procesamiento e identificación de la información relevante.

La recolección, pre-procesamiento e identificación de la información relevante en los documentos de las bibliotecas digitales (Phronesis, CIRIA y Redalyc) para la inferencia de redes temáticas de colaboración corresponde a la etapa 1 de la metodología, en esta etapa se obtiene la información que sirve de entrada para el algoritmo de agrupamiento jerárquico FIHC, así como la información que es utilizada posteriormente al momento de generar las redes temáticas de colaboración. En ambos casos la información se obtiene a partir pre-procesamiento del contenido de los registros cosechados mediante el protocolo OAI-PMH, dicha información se encuentra en el formato de metadatos

Figura 3.1: Metodología para la inferencia de redes temáticas de colaboración

Dublin Core.

Esta etapa nos permite extraer, identificar y preparar la información que se encuentra en los registros cosechados (ejemplo de un registro típico en la figura 2.1) provenientes de cada una de las bibliotecas digitales para su uso en las etapas posteriores. Podemos observar lo anterior en la figura 3.2.

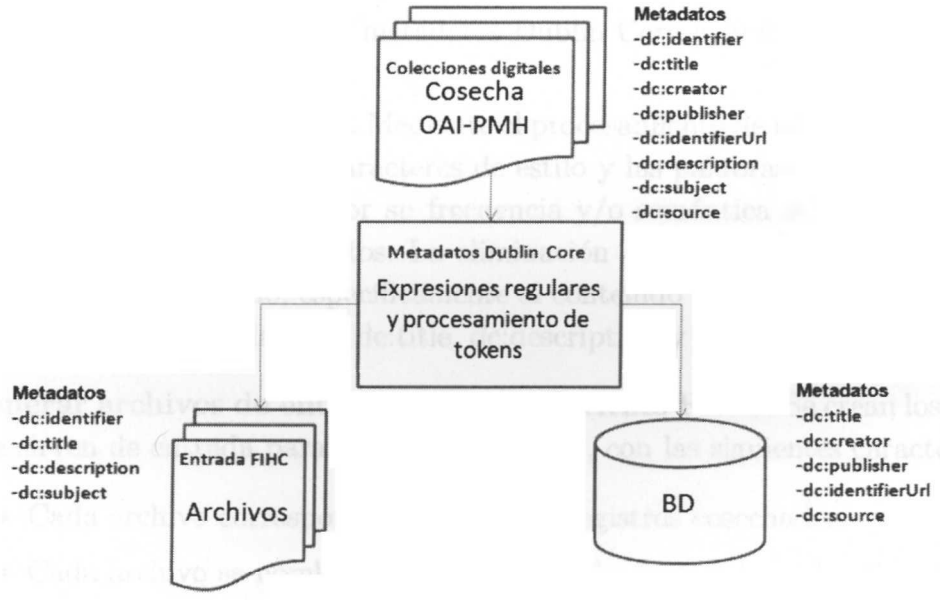


Figura 3.2: Recolección, pre-procesamiento e identificación de información relevante

La etapa 1 involucra:

- **Uso de expresiones regulares.** Mediante el uso de expresiones regulares podemos identificar en cada uno de los registros cosechados las etiquetas asociadas a aquellos metadatos Dublin Core que contienen información relevante para realizar el agrupamiento de dichos registros mediante el algoritmo FIHC, así como a aquellos metadatos con información relevante para crear las redes temáticas de colaboración. En la tabla 3.1 se muestran las etiquetas asociadas a los metadatos Dublin Core que contienen información relevante, una descripción de su contenido y el ámbito de su aplicación.

Etiqueta	**Contenido**	**Entrada FIHC**	**Base de datos**
dc:title	Titulo del documento	X	X
dc:creator	Autor del documento	-	X
dc:publisher	Institución que publica el documento	-	X
dc:identifier	Url asociada al documento	-	X
dc:description	Resumen del contenido del documento	X	-
dc:subject	Palabras claves asociadas al documento	X	-
dc:source	País de origen de la publicación	-	X
*Se marcan con una X los casos donde aplica el uso del metadato Dublin Core.			

Tabla 3.1: Etiquetas asociadas a metadatos Dublin Core y su ámbito de aplicación

- **Procesamiento de tokens.** Mediante el procesamiento de tokens se eliminan los caracteres de formato, los caracteres de estilo y las palabras (artículos, pronombres y preposiciones) que por su frecuencia y/o semántica no poseen valor discriminatorio entre documentos. La eliminación de dichos caracteres se aplica a cada registro cosechado, específicamente al contenido de los siguientes metadatos Dublin Core (ver tabla 3.1): dc:title, dc:description y dc:subject.

- **Generar archivos de entrada para el algoritmo FIHC.** Se crean los archivos que sirven de entrada para el algoritmo FIHC, con las siguientes características:

  - Cada archivo corresponde a uno de los registros cosechados.
  - Cada archivo se nombra con la información contenida en la etiqueta *identifier*, de esta manera podemos relacionar a cada uno de los archivos que se crean con uno de los registros cosechados. (ver figura 3.3).
  - La información que se almacena en cada archivo proviene de los metadatos

(dc:title, dc:description y dc:subject) a los que se les ha aplicado el procesamiento de tokens descrito anteriormente.

Como resultado se tiene un archivo por cada documento proveniente de alguna de las bibliotecas digitales cuyo contenido cumple con las características para ser agrupado mediante el algoritmo FIHC. La figura 3.4 representa la estructura de cada uno de los archivos que se generan.

```
- <record>
  - <header>
      <identifier>ITESMMTY200042</identifier>
```

Figura 3.3: Información contenida en la etiqueta identifier de uno de los registros cosechados de Phronesis.

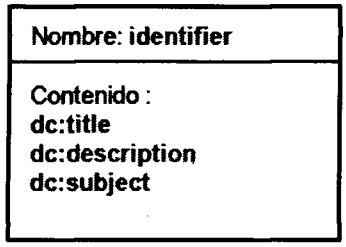


Figura 3.4: Estructura de archivo de entrada para el algoritmo FIHC

- **Almacenar la información asociada a las redes temáticas de colaboración.** Consiste en guardar en base de datos los metadatos Dublin Core que son utilizados al momento de generar las redes temáticas de colaboración, específicamente (ver tabla 3.1): dc:title, dc:creator, dc:publisher, dc:identifier y dc:source. El almacenamiento de estos metadatos se hace por cada uno de los registros cosechados provenientes de las bibliotecas digitales.

3.1.2. Agrupamiento jerárquico de registros cosechados.

El agrupamiento jerárquico de registros cosechados a través del algoritmo FIHC corresponde a la etapa 2, esta etapa tiene como objetivo agrupar los registros cosechados provenientes de las bibliotecas digitales de acuerdo a su similitud de contenido, para lograr su objetivo esta etapa involucra (figura 3.5):

- **Datos de entrada**. Como datos de entrada para el algoritmo FIHC se tiene a cada uno de los archivos obtenidos durante la etapa 1 y cuyo contenido cumple

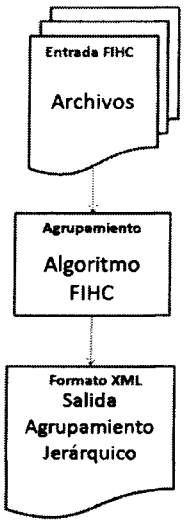


Figura 3.5: Agrupamiento jerárquico de registros cosechados

con un pre-procesamiento necesario para la realizar el agrupamiento de las colecciones digitales. Pre-procesamiento que implica el remover las palabras que por su semántica no poseen valor discriminatorio entre documentos, remover caracteres de estilo y de formato.

- **Agrupamiento jerárquico.** El agrupamiento jerárquico se realiza a través del algoritmo FIHC.

- **Datos de salida.** Archivo en formato XML que representa el agrupamiento jerárquico de los documentos. En la figura 3.6 podemos observar un ejemplo de la salida en formato XML que corresponde a la representación del árbol de agrupamiento mostrado en la figura 2.3

3.1.3. Generación de redes temáticas de colaboración.

La generación de redes temáticas de colaboración corresponde a la etapa 3, en esta etapa se obtienen las redes que expresan la posibilidad o grado de colaboración entre autores o entre instituciones, y que se infieren mediante la adecuación de uso y adaptación de las métricas definidas por Xiaoming Liu [13] que permiten obtener un grafo dirigido con pesos normalizados.

En esta etapa se usa como fuente de datos la siguiente información:

- El agrupamiento jerárquico de documentos (resultado de la etapa 2) se usa como fuente de datos ya que permite tratar a cada uno de los grupos como un único espacio en el cuál se encuentran agrupados aquellos documentos que comparten una afinidad temática. Es importante resaltar que posterior al agrupamiento

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
- <root num_clusters="6" label="raiz" num_children="2" num_docs="15">
  - <documents num_docs="2">
      <document>ITESMMTY19991</document>
      <document>ITESMMTY199911</document>
    </documents>
  - <cluster label="matematicas" num_children="2" num_docs="8">
    - <documents num_docs="2">
        <document>ITESMMTY19995</document>
        <document>ITESMMTY19997</document>
      </documents>
    - <cluster label="matematicas aplicadas" num_children="0" num_docs="2">
      - <documents num_docs="2">
          <document>ITESMMTY19993</document>
          <document>ITESMMTY199912</document>
        </documents>
      </cluster>
    - <cluster label="matematicas calculo" num_children="1" num_docs="4">
      - <documents num_docs="2">
          <document>ITESMMTY199935</document>
          <document>ITESMMTY199921</document>
        </documents>
      - <cluster label="matematicas calculo diferencial" num_children="0" num_docs="2">
        - <documents num_docs="2">
            <document>ITESMMTY19992</document>
            <document>ITESMMTY199956</document>
          </documents>
        </cluster>
      </cluster>
    </cluster>
  - <cluster label="fisica" num_children="1" num_docs="5">
    - <documents num_docs="3">
        <document>ITESMMTY19996</document>
        <document>ITESMMTY19994</document>
        <document>ITESMMTY19998</document>
      </documents>
    - <cluster label="fisica materiales" num_children="0" num_docs="2">
      - <documents num_docs="2">
          <document>ITESMMTY199915</document>
          <document>ITESMMTY199972</document>
        </documents>
      </cluster>
    </cluster>
</root>
```

Figura 3.6:

jerárquico la única información que se posee de los documentos es el identificador del documento y la etiqueta del grupo al que pertenece (ver figura 3.6).

- Los metadatos almacenados en base de datos (resultado de la etapa 1) proporcionan información adicional asociada a cada documento agrupado, posterior al agrupamiento jerárquico la recuperación de dicha información en la base de datos es posible a través del identificador del documento. Los metadatos asociados a la institución que publica el documento o al autor del documento (ver tabla 3.1) permiten generar las redes de colaboración para autores o instituciones respectivamente, por esta razón estos metadatos son nombrados entidades de colaboración.

Se menciona la adecuación de uso de las métricas definidas por Xiaoming Liu debido a que originalmente evalúan la relación entre los autores de un artículo para obtener una red de co-autoría, pero en el caso de la inferencia de redes temáticas de colaboración las métricas se usan para evaluar la relación entre los autores o instituciones (entidades de colaboración) de un grupo de documentos con temática afín, lo anterior podemos observarlo en la figura 3.7.

Por otra parte, hablamos de adaptación debido a que es necesaria una modificación al momento de calcular la exclusividad de relación entre las entidades de colaboración (autores o instituciones). La modificación permite considerar el hecho de que una misma entidad de colaboración esté presente en más de una ocasión dentro de un mismo grupo (es el caso de la entidad v_3 dentro del grupo c_1 en la figura 3.7), situación que no se presenta en un artículo (ningún autor se repite más de una vez en un artículo).

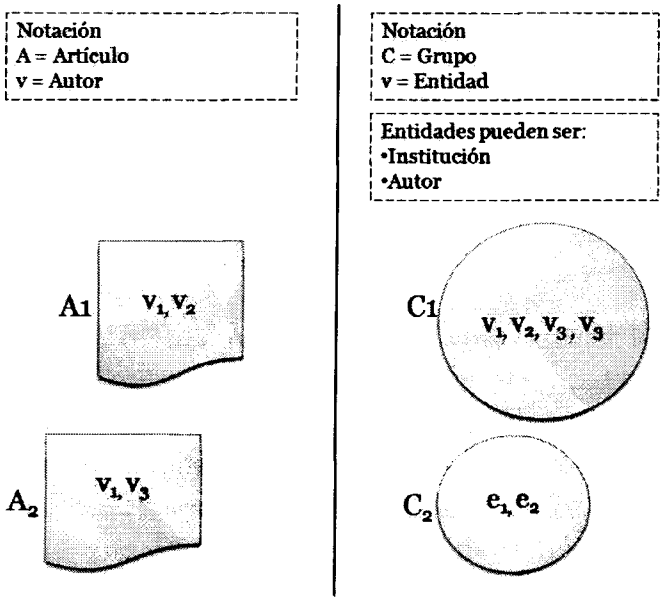


Figura 3.7: Adecuación de uso de las métricas de Xiaoming Liu

Los cálculos realizados en las funciones de *frecuencia de co-autoría* y *pesos normalizados* permanecen sin cambios. Sin embargo, la terminología cambia para todas las funciones, ahora nos referimos a grupos en vez de a artículos, usamos el término entidad (autor o institución) en lugar de autor y empleamos el concepto de afinidad temática para reemplazar el concepto de co-autoría.

Modelo de inferencia de afinidad temática.

A través del modelo descrito a continuación se puede determinar la estructura de las redes temáticas de colaboración.

Sea G el grafo de afinidad temática denotado por $G = (V, E, W)$, donde V es el conjunto de nodos (entidades), E es el conjunto de aristas (relaciones entre entidades) y W el conjunto de pesos W_{ij} asociados con cada arista que conecta un par de entidades (v_i, v_j).

Para determinar la magnitud del enlace entre dos entidades se toman en cuenta dos factores:

1. Frecuencia de afinidad temática. Entidades que frecuentemente se encuentran juntos en algún grupo podrían tener más peso de afinidad temática.

2. Número total de entidades en grupos. Si un grupo tiene muchas entidades, cada relación entre dos entidades tendrá menos peso.

Ahora podemos determinar el peso de los enlaces de la afinidad temática. Sea $V = \{v_1, \cdots, v_n\}$ un conjunto de n entidades (autores o instituciones). Sea $C = \{c_1, \cdots, c_k, \cdots, c_m\}$ un conjunto de m grupos de documentos y $f(c_k)$ el número de entidades en el grupo c_k, definimos los siguientes conceptos:

- **Exclusividad**. Si las entidades v_i y v_j están presentes en el grupo c_k.

$$g_{i,j,k} = \frac{1}{f(c_k) - 1} * f(j_k) \tag{3.1}$$

  En la función 3.1 $g_{i,j,k}$ representa el grado de exclusividad en la relación de afinidad temática entre v_i y v_j en un grupo en particular, $f(j_k)$ (adaptación) representa el número de veces que la entidad v_j está presente en el grupo (dando mayor peso a la relación con una entidad con mayor presencia en el grupo). Esta definición da más peso a las relaciones de afinidad temática en los grupos conformados por pocas entidades, y menos peso a las relaciones de afinidad temática en los grupos conformados por muchas entidades.

- **Frecuencia de afinidad temática**. La frecuencia de afinidad temática (función 3.2) consiste en la suma de todos los valores $g_{i,j,k}$ para todos los grupos donde

están presentes v_i y v_j. Dando más peso a las entidades que se encuentren juntas en más ocasiones, y por lo tanto más exclusividad.

$$s_{ij} = \sum_{k=1}^{m} g_{i,j,k} \tag{3.2}$$

- **Pesos normalizados**. La normalización (función 3.3) asegura que los pesos de las relaciones de una entidad sumen uno.

$$w_{ij} = \frac{s_{ij}}{\sum_{k=1}^{n} s_{i,k}} \tag{3.3}$$

Donde w_{ij} es el peso normalizado entre la entidad v_i y v_j, s_{ij} es la frecuencia de afinidad temática entre la entidad v_i y v_j, y $\sum_{k=1}^{n} s_{i,k}$ representa la sumatoria de las frecuencias de afinidad temática entre la entidad v_i y cada uno de las entidades v_k con las que se relaciona.

Una representación del grafo dirigido con pesos normalizados que corresponde a la estructura de una red temática de colaboración la podemos observar en figura 3.8 y se obtiene al aplicar las funciones antes mencionadas .

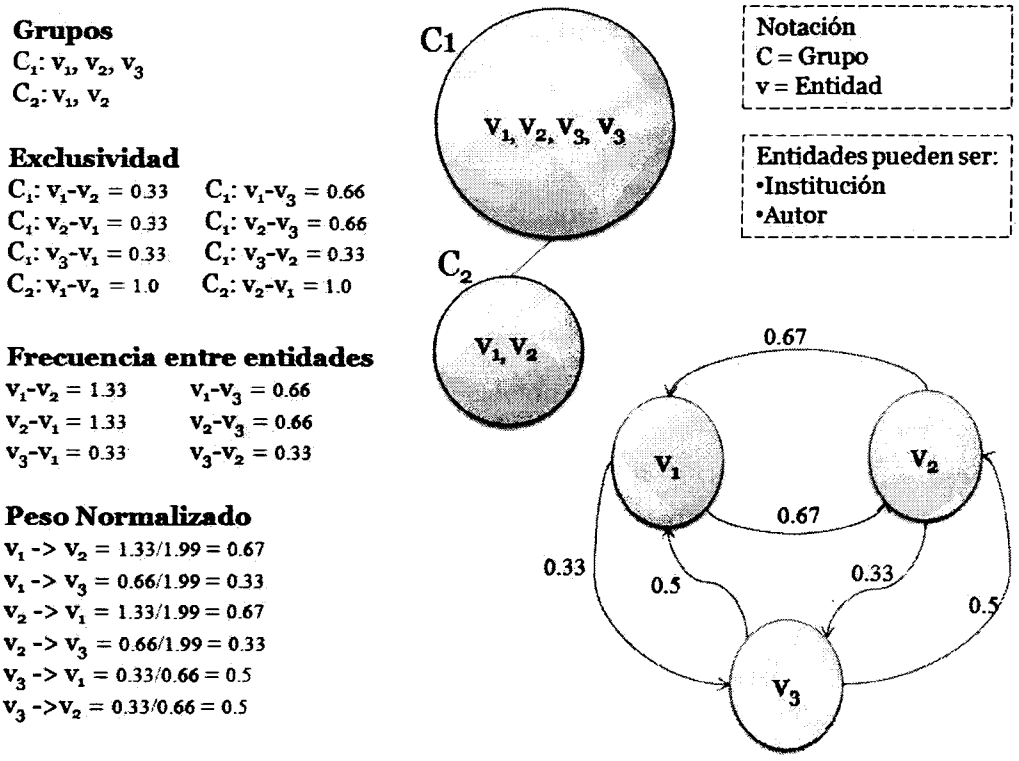


Figura 3.8: Red temática de colaboración

Representación de las redes temáticas de colaboración.

A continuación se describe la estructura de los archivos de salida en formato XML utilizados para representar las redes de colaboración. En la figura 3.9 se muestra una representación de una red de colaboración a nivel de instituciones. El alcance e información contenida en sus etiquetas es:

- <**institutionQuery**>. La etiqueta global.
- <**dataSet**>. Dentro de esta etiqueta están presentes todos y cada uno de los elementos de la red.
- <**institution**>. Etiqueta que hace referencia a una institución. Atributos:
  - id = Identificador.
  - name = Nombre de la institución.
  - country = País al que pertenece la institución.
- <**author**>. Etiqueta que hace referencia a un autor. Atributos:
  - Name = Nombre del autor.
- <**subject**>. Tema de interés del autor (proviene de la etiqueta de agrupamiento a través del uso del algoritmo FIHC).
- <**relationSet**>. Conjunto de relaciones entre los elementos del dataset.
- <**relation**>. Conexión entre dos entidades institución. Atributos:
  - from = Id de la institución de donde proviene la conexión (origen).
  - to = Id de la institución a donde llega la conexión (destino).
  - strength = Fortaleza de la relación (pesos normalizados).

La figura 3.10 muestra la estructura y sintaxis correspondiente al archivo XML de la red de colaboración a nivel de instituciones a través de una definición del tipo de documento o DTD (por sus siglas en ingles).

La figura 3.11 muestra una representación de una red de colaboración a nivel de autores. El alcance e información contenida en sus etiquetas es:

- <**authorQuery**>. La etiqueta global.
- <**dataSet**>. Dentro de esta etiqueta están presentes todos y cada uno de los elementos de la red.

```
< ?xml version="1.0" encoding="utf-8" ? >
<intitutionQuery>
    <dataSet>
        <institution id="AnyID 1" name="Institution Name" country="Institution Country">
            <author name="Author Name">
                <subject>subject 1</subject>
                <subject>subject 2</subject>
            </author>
            <author name="Author Name">
                <subject>subject 1</subject>
            </author>
            . . .
            <author name="Author Name">
                <subject>subject 1</subject>
                . . .
                <subject>subject N</subject>
            </author>
        </institution>
        <institution id="AnyID 2" name="Institution Name" country="Institution Country">
            Authors & Subjects
        </institution>
        . . .
        <institution id="AnyID N" name="Institution Name" country="Institution Country">
            Authors & Subjects
        </institution>
    </dataSet>
    <relationSet>
        <relation from="Any ID 1" to="Any ID 2" strength="Nivel de Relacion(float)" />
        <relation from="Any ID 1" to="Any ID 3" strength="Nivel de Relacion(float)" />
        <relation from="Any ID 2" to="Any ID 1" strength="Nivel de Relacion(float)" />
        . . .
        <relation from="Any ID 3" to="Any ID 1" strength="Nivel de Relacion(float)" />
        <relation from="Any ID 3" to="Any ID 2" strength="Nivel de Relacion(float)" />
    </relationSet>
</intitutionQuery>
```

Figura 3.9: Representación de la red de colaboración a nivel de instituciones

```
<!ELEMENT institutionQuery (dataSet, relationSet)>
<!ELEMENT dataSet (institution* )>
<!ELEMENT institution (author* )>
<!ATTLIST institution
    id      CDATA   #REQUIRED
    name    CDATA   #REQUIRED
    country CDATA   #REQUIRED>
<!ELEMENT author (subject* )>
<!ATTLIST author
    name    CDATA   #REQUIRED>
<!ELEMENT subject (#PCDATA)>
<!ELEMENT relationSet (relation* )>
<!ELEMENT relation EMPTY>
<!ATTLIST relation
    from    CDATA   #REQUIRED
    to      CDATA   #REQUIRED
    stregth CDATA   #REQUIRED>
```

Figura 3.10: DTD de la red de colaboración a nivel de instituciones

- **<author>**. Etiqueta que hace referencia a un autor. Atributos:
  - id = Identificador.
  - name = Nombre del autor.
  - institution = Institución a la que pertenece el autor.
- **<document>**. Hace referencia a un documento. Atributos:
  - title = Título del documento.
  - url = Dirección virtual del documento.
  - date = Fecha de creación del documento.
- **<subject>**. Tema del documento (proviene de la etiqueta de agrupamiento a través del uso del algoritmo FIHC).
- **<relationSet>**. Conjunto de relaciones entre los elementos del dataset.
- **<relation>**. Conexión entre dos entidades autor. Atributos:
  - from = Id del autor de donde proviene la conexión (origen).
  - to = Id del autor a donde llega la conexión (destino).
  - strength = Fortaleza de la relación (pesos normalizados).

```
< ?xml version="1.0" encoding="utf-8" ? >
<authorQuery>
    <dataSet>
        <author id="AnyID 1" name="Author Name" Institution="Institution Name">
            <document title="Document Name" url="Document URL" >
                <subject>subject 1</subject>
            </document>
            <document title="Document Name" url="Document URL" >
                <subject>subject 1</subject>
            </document>
            . . .
            <document title="Document Name" url="Document URL" >
                <subject>subject 1</subject>>
            </document>
        </author>
        <author id="AnyID 2" name="Author Name" Institution="Institution Name">
            Documents & Subjects
        </author>
        . . .
        <author id="AnyID N" name="Author Name" Institution="Institution Name">
            Documents & Subjects
        </author>
    </dataset>
    <relationSet>
        <relation from="Any ID 1" to="Any ID 2" strength="Nivel de Relacion(float)" />
        <relation from="Any ID 1" to="Any ID 3" strength="Nivel de Relacion(float)" />
        <relation from="Any ID 2" to="Any ID 1" strength="Nivel de Relacion(float)" />
        . . .
        <relation from="Any ID 3" to="Any ID 1" strength="Nivel de Relacion(float)" />
        <relation from="Any ID 3" to="Any ID 2" strength="Nivel de Relacion(float)" />
    </relationSet>
</authorQuery>
```

Figura 3.11: Representación de la red de colaboración a nivel de autores

La figura 3.12 muestra la estructura y sintaxis correspondiente al archivo XML de la red de colaboración a nivel de autores a través de una definición del tipo de documento o DTD (por sus siglas en ingles).

```
<!ELEMENT authorQuery (dataSet, relationSet)>
<!ELEMENT dataSet (author* )>
<!ELEMENT author (document* )>
<!ATTLIST author
    id          CDATA   #REQUIRED
    name        CDATA   #REQUIRED
    institution CDATA   #REQUIRED>
<!ELEMENT document (subject* )>
<!ATTLIST document
    title   CDATA   #REQUIRED
    url     CDATA   #REQUIRED>
<!ELEMENT subject (#PCDATA)>
<!ELEMENT relationSet (relation* )>
<!ELEMENT relation EMPTY>
<!ATTLIST relation
    from    CDATA   #REQUIRED
    to      CDATA   #REQUIRED
    stregth CDATA   #REQUIRED>
```

Figura 3.12: DTD de la red de colaboración a nivel de autores

3.2. Resumen

Durante el desarrollo de este capítulo se describió de forma detallada el modelo de solución propuesto. Primero se describió el proceso general y posteriormente se definieron cada una de las etapas que forman el modelo de solución global y cada uno de los conceptos involucrados. En el próximo capítulo se mencionan las características, situaciones presentes durante la implementación y desarrollo de la herramienta (herramienta realizada de acuerdo al modelo descrito en este capítulo) para la inferencia de redes temáticas de colaboración, así como resultados y experimentos de validación.

Capítulo 4

Implementación y Resultados obtenidos.

A continuación se mencionan características, situaciones y resultados presentes durante la implementación y desarrollo de la herramienta para la inferencia de redes temáticas de colaboración. Herramienta desarrollada de acuerdo a lo descrito en la sección 3.1 con el propósito de inferir las redes temáticas de colaboración a partir de las colecciones digitales provenientes de Phronesis, CIRIA y Redalyc. En la sección de resultados se incluyen los experimentos de validación realizados.

4.1. Implementación.

Con el objetivo de presentar en forma organizada las características y situaciones que se presentaron durante la implementación del modelo propuesto, los detalles de la implementación se abordan en relación a cada una de las etapas de la metodología:

- Recolección, pre-procesamiento e identificación de la información relevante.
- Agrupamiento jerárquico de registros cosechados.
- Generación de redes temáticas de colaboración.

4.1.1. Características y situaciones presentes durante la implementación de la etapa 1.

La etapa 1 de la metodología involucra la recolección, pre-procesamiento e identificación de la información relevante, a través de la etapa 1 se obtiene la información que sirve de entrada para el algoritmo de agrupamiento jerárquico FIHC, así como la información que es utilizada posteriormente al momento de generar las redes temáticas de colaboración.. Los detalles de la etapa 1 se describen en la sección 3.1.1.

Podemos identificar tres procesos dentro de la etapa 1 de los cuales se describen los aspectos relacionados con su implementación:

- Recolección de registros.

- Uso de expresiones y procesamiento de tokens.
- Almacenamiento en base de datos de información relevante.

Recolección de registros.

La recolección o cosecha de registros procedentes de las colecciones de Phronesis, Redalyc y CIRIA se representa en la figura 4.1.

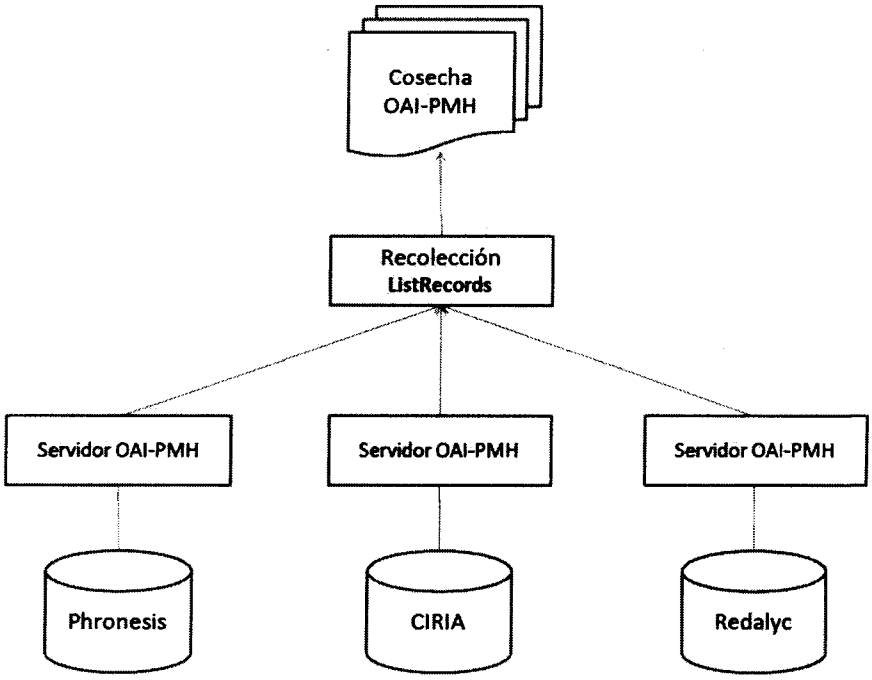


Figura 4.1: Recolección o cosecha de registros en las colecciones: Phronesis, CIRIA y Redalyc

Es importante mencionar que los tres repositorios cumplen con el estándar OAI, lo cual es una condición indispensable para la aplicación de la metodología descrita en la sección 3.1.

En relación a la implementación del proceso de cosecha vale la pena resaltar los siguientes aspectos:

- Para realizar la cosecha se hace uso del estándar OAI-PMH.
- La cosecha se realiza con el objetivo de recuperar el total de registros contenidos en los repositorios.

En el caso de la colecciones de Phronesis y Redalyc el proceso de recolección fue realizado de manera manual mediante el uso de la solicitud ListRecords del protocolo

OAI-PMH descrito en la sección 2.1.1. En el caso de Redalyc la cosecha de los registros pertenecientes a su colección está disponible en el sitio ftp:

`http://sis.redalyc.uaemex.mx/ftp/`

Este proceso de recolección se abordó como se comenta en el párrafo anterior debido a que este trabajo se centra en la inferencia de la red temática de colaboración y no en el proceso de recolección, y a pesar de que la recolección es una actividad repetitiva es posible realizarla manualmente (el caso de Phronesis y de CIRIA).

El proceso de recolección hecho manualmente se describe en la figura 4.2. Para una mayor claridad de la forma en la que se realiza el proceso de cosecha de registros se proporciona la solicitud ListRecords con la que se inicia la cosecha de la colección de CIRIA:

`http://catarina.udlap.mx/u_dl_a/tales/oai/requestETD.jsp?`
`verb=ListRecords&metadataPrefix=oai_dc`

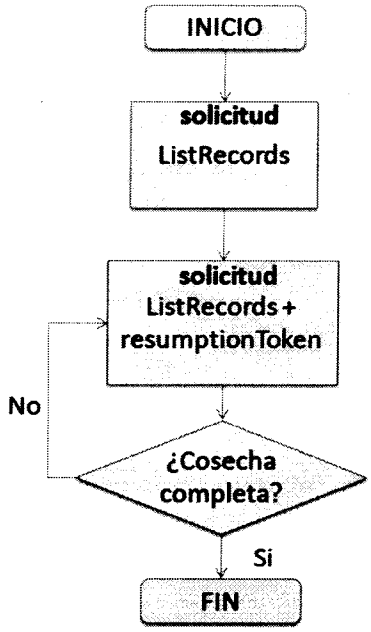


Figura 4.2: Proceso de cosecha manual

La figura 4.3 muestra parte del resultado tras la ejecución de la solicitud ListRecords con la que se inició la cosecha de registros. Para continuar con la cosecha de registros es necesario agregar el atributo resumptionToken (ver figura 4.4) en la solicitud ListRecords:

`http://catarina.udlap.mx/u_dl_a/tales/oai/requestETD.jsp?`
`verb=ListRecords&metadataPrefix=oai_dc&resumptionToken=mmumcmuumiirr`

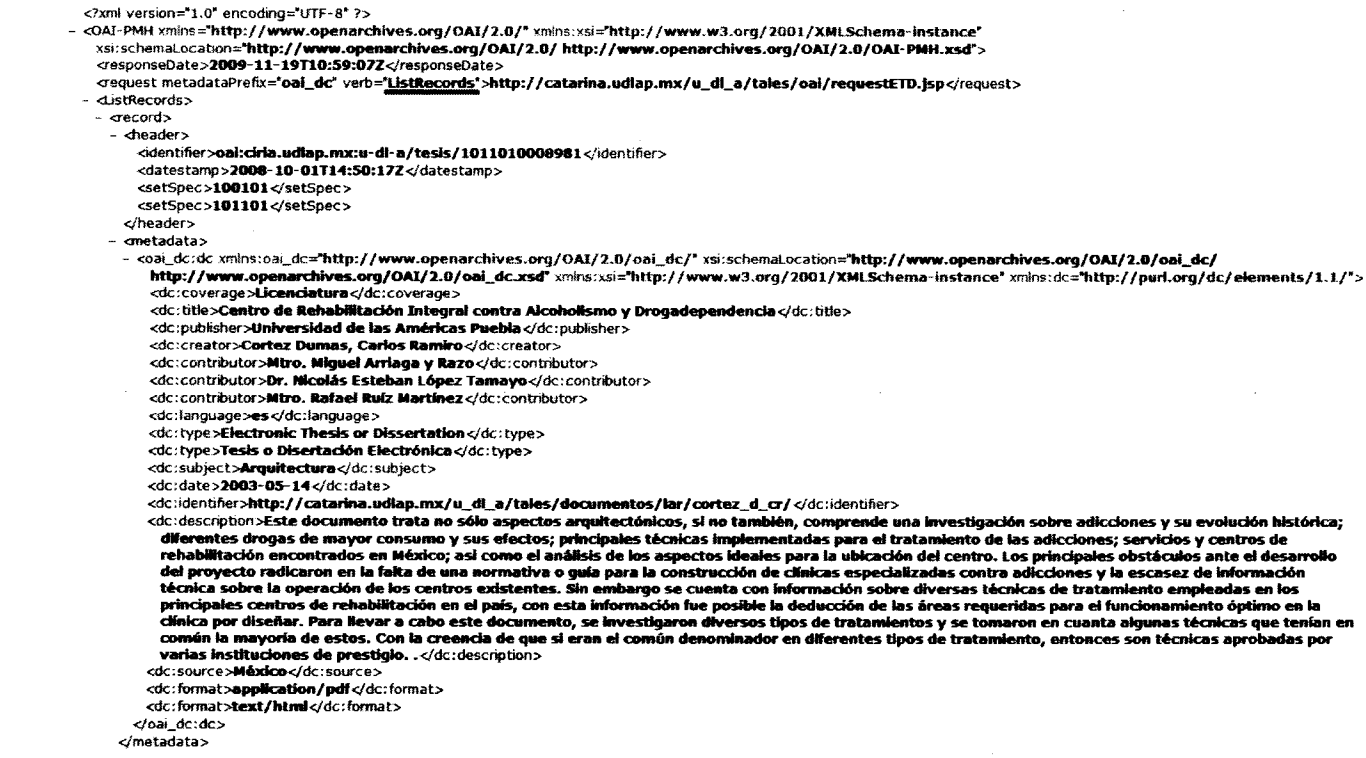


Figura 4.3: Resultado de la solicitud ListRecords

Mientras existan registros pendientes por ser cosechados se obtiene un valor en el atributo resumptionToken como resultado de cada cosecha. La cosecha finaliza cuando el atributo resumptionToken no tenga ningún valor.

```
      </metadata>
    </record>
    <resumptionToken cursor="0" completeListSize="1628">maruuuoumeoma</resumptionToken>
  </ListRecords>
</OAI-PMH>
```

Figura 4.4: Valor del atributo resumptionToken tras la ejecución de la solicitud ListRecords

Uso de expresiones regulares y procesamiento de tokens.

El uso de expresiones regulares permite identificar para cada uno de los registros cosechados los metadatos Dublin Core (ver tabla 3.1) que contienen información relevante para la inferencia de las redes temáticas de colaboración.

Las figuras 4.5, 4.6 y 4.7 corresponden a un registro proveniente de Phronesis, CIRIA y Redalyc respectivamente. En ellas se puede observar la presencia de los

metadatos enlistados en la tabla 3.1.

```
- <record>
  - <header>
      <identifier>ITESMMTY199913</identifier>
      <datestamp>2000-12-15</datestamp>
      <setSpec>MTY</setSpec>
    </header>
  - <metadata>
    - <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:dc="http://purl.org/dc/elements/1.1/"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
        http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
        <dc:title>Planeación de Trayectorias en Espacio-C Mediante B-Splines y Recocido Simulado</dc:title>
        <dc:identifier>http://copernico.mty.itesm.mx/phronesis/mty/busqueda/bajarOAI.cgi?filename=ITESMMTY199913.ps</dc:identifier>
        <dc:date>1996-11-11</dc:date>
        <dc:description>En este artículo se describe una técnica para planear trayectorias para un objetopoligonal entre obstáculos, que sean libres de
          colisiones y que además sean de mínima distancia. Para la representación del espacio de trabajo se utiliza EspacioC (CSpace), lo cual reduce el
          problema a la planeación de la trayectoria de un sólo punto entre obstáculos expandidos. Para la representación de las trayectorias, se utilizan
          curvas BSplines, y para generarlas, se usa una técnica de optimización conocida como Recocido Simulado. Se presentan algunos ejemplos para
          demostrar la validez del método.</dc:description>
        <dc:creator>Martínez Alfaro, Horacio</dc:creator>
        <dc:contributor>Ulloa Pérez, Antonio</dc:contributor>
        <dc:language>es</dc:language>
        <dc:subject>Inteligencia Artificial, recocido simulado, planeación de trayectorias, b-spline, espacio-C</dc:subject>
        <dc:publisher>International Symposium on Artificial Intelligence</dc:publisher>
        <dc:source>México</dc:source>
        <dc:type />
        <dc:format>application/postscript</dc:format>
        <dc:relation />
        <dc:coverage />
        <dc:rights />
        <dc:contributor />
      </oai_dc:dc>
    </metadata>
  </record>
```

Figura 4.5: Registro procedente de Phronesis

```
- <record>
  - <header>
      <identifier>oai:ciria.udlap.mx:u-dl-a/tesis/1021012742981</identifier>
      <datestamp>2008-10-01T14:50:17Z</datestamp>
      <setSpec>100102</setSpec>
      <setSpec>102101</setSpec>
    </header>
  - <metadata>
    - <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
        xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:dc="http://purl.org/dc/elements/1.1/">
        <dc:coverage>Licenciatura</dc:coverage>
        <dc:title>En Transformación</dc:title>
        <dc:publisher>Universidad de las Américas Puebla</dc:publisher>
        <dc:creator>Vargas Lugo Salinas, Miriam</dc:creator>
        <dc:contributor>Mtra. Adriana Martínez Mendoza</dc:contributor>
        <dc:contributor>Mtro. Joaquín Conde García</dc:contributor>
        <dc:contributor>Mtro. Sergio González Angulo Aguirre</dc:contributor>
        <dc:contributor>Mtro. Antonio Miguel Audirac Camarena</dc:contributor>
        <dc:language>es</dc:language>
        <dc:type>Electronic Thesis or Dissertation</dc:type>
        <dc:type>Tesis o Disertación Electrónica</dc:type>
        <dc:subject>Artes Plásticas</dc:subject>
        <dc:date>2003-05-15</dc:date>
        <dc:identifier>http://catarina.udlap.mx/u_dl_a/tales/documentos/lap/vargas_l_m/</dc:identifier>
        <dc:description>La creación de un equilibrio formal y plástico por medio de la transformación, variación y modulación de la
          materia, donde el juego y la experimentación fueron condiciones del proceso creativo, fue la propuesta a trabajarse en este
          proyecto de tesis. El objetivo de este fue el crear estructuras y formas que se pudieran relacionar entre sí, para así
          conformar formas tridimensionales, las cuales puedan ser modificadas en diversas posibilidades conforme a su relación con el
          espacio para crear nuevas composiciones. Estas modificaciones fueron sujetas a dos transformaciones expansión y
          compresión, tomando en cuenta que en ellas existió la experimentación mediante una situación controlada. Los conceptos de
          equilibrio, contraposición, modulación, variación y espacio, fueron los que se trabajaron a lo largo de esta tesis para mostrar
          que la escultura, a través de la representación de sus propios materiales y de el proceso de construcción o concepción, nos
          muestra su propia autonomía.</dc:description>
        <dc:source>México</dc:source>
        <dc:format>application/pdf</dc:format>
        <dc:format>text/html</dc:format>
      </oai_dc:dc>
    </metadata>
  </record>
```

Figura 4.6: Registro procedente de CIRIA

```
- <record>
  - <header>
      <identifier>oai:redalyc.uaemex.mx:10202904</identifier>
      <datestamp>2006-05-05</datestamp>
      <setSpec>1870-3925</setSpec>
    </header>
  - <metadata>
    - <oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/" xmlns:dc="http://purl.org/dc/elements/1.1/"
        xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
        xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
        <dc:title>Evaluación de la sustentabilidad del desarrollo regional. El marco de la agricultura</dc:title>
        <dc:creator>Pablo Torres Lima</dc:creator>
        <dc:creator>Luis Rodríguez Sánchez</dc:creator>
        <dc:creator>Óscar Sánchez Jerónimo</dc:creator>
        <dc:subject>Estudios Territoriales</dc:subject>
        <dc:description>sustentabilidad, desarrollo regional, sistemas agrícolas.</dc:description>
        <dc:publisher>El Colegio de Sonora</dc:publisher>
        <dc:source>México</dc:source>
        <dc:date>05-05-2004</dc:date>
        <dc:type>Artículo científico</dc:type>
        <dc:format>application/pdf</dc:format>
        <dc:identifier>http://redalyc.uaemex.mx/redalyc/src/inicio/HomRevRed.jsp?iCveEntRev=102</dc:identifier>
        <dc:relation>1</dc:relation>
        <dc:rights>Región y Sociedad</dc:rights>
      </oai_dc:dc>
    </metadata>
</record>
```

Figura 4.7: Registro procedente de Redalyc

Durante el proceso de identificación de los metadatos Dublin Core se hicieron evidentes algunas particularidades en los registros cosechados. Estas particularidades son comunes entre los registros procedentes de un mismo repositorio (Phronesis, CIRIA o Redalyc) y se muestran en la tabla 4.1.

	Particularidades		
Etiqueta	**Phronesis**	**CIRIA**	**Redalyc**
dc:creator	Una etiqueta dc:creator	Una etiqueta dc:creator	*Más de una etiqueta dc:creator (ver figura 4.9)*
dc:description	Una etiqueta dc:description	*Más de una etiqueta dc:description (ver figura 4.8)*	Una etiqueta dc:description

Tabla 4.1: Particularidades asociadas a los metadatos Dublin Core comunes entre los registros procedentes de un mismo repositorio

Con la finalidad de mantener consistente el proceso de identificación de la información mediante el uso de expresiones regulares fue necesario considerar las particularidades mostradas en la tabla 4.1, las particularidades se trataron la siguiente manera:

- *Metadato dc:creator.* En el caso de los registros procedentes del repositorio de Redalyc la información que se toma en cuenta para cada registro es la que corresponde al primer metadato dc:creator encontrado.

<dc:description>**En ésta tesis deseo investigar el retrato como un medio de representación y como género pictórico y estudiar sus diferentes posibilidades plásticas. El retrato es la representación de una persona determinada donde se muestran sus características principales. Rasgos físicos, carácter, gestos, entorno, estado de ánimo, personalidad, lenguaje corporal y vestimenta, son algunas de estas características. Mi proyecto consiste en realizar una serie de obras plásticas tomando como modelo a una sola persona. Estas obras fueron realizadas en diferentes técnicas: dibujos, acuarelas, escultura y pinturas. Conocí a José Alejandro Rodríguez y Soto cuando tuve mi primera clase de escultura y él trabajaba como modelo. Esta situación se repitió a lo largo de mi carrera en clases de dibujo y pintura. Al plantear mi trabajo de tesis decidí elegirlo ya que al hablar con él se mostró interesado y aceptó participar. Para empezar mi proyecto fue necesario adentrarme en su vida personal y conocer más sobre él ya que quería que en mis retratos se reflejara su vida diaria, más allá de su apariencia física como modelo.**</dc:description>
<dc:description>**Tuvimos varias pláticas en las que me habló de su trabajo, familia, viajes y gustos. Fui a su lugar de trabajo y a su casa en donde conocí a su esposa e hija y pude conocer un poco más de su vida privada. Ahí pude tomar fotografías y observar detalles que posteriormente me ayudaron a proyectar los cuadros. A través de esta serie de retratos pretendo acercarme a la esencia de la persona retratada, reflejar parte de su personalidad, características que lo identifican y su entorno. Al mismo tiempo plasmar mi naturaleza, evidente en mi forma de pintar, los trazos, manchas y pinceladas que me definen. Palabras Claves: Retrato, Percepción, Representación, Esencia .**</dc:description>

Figura 4.8: Parte de un registro procedente del repositorio de CIRIA con más de una etiqueta dc:description

<dc:creator>**Pablo Torres Lima**</dc:creator>
<dc:creator>**Luis Rodríguez Sánchez**</dc:creator>
<dc:creator>**Óscar Sánchez Jerónimo**</dc:creator>

Figura 4.9: Parte de un registro procedente del repositorio de Redalyc con más de una etiqueta dc:cretator

- *Metadato dc:description.* En el caso de los registros procedentes del repositorio de CIRIA la información que se toma en cuenta para cada registro es la que corresponde al total de metadatos dc:description encontrados.

Por otra parte, si bien el uso de los metadatos Dublin Core es un estándar, la información contenida en ellos no lo es. Esta situación se muestra en la tabla 4.2 donde se mencionan las diferencias encontradas entre los datos procedentes de repositorios distintos pero etiquetados con un mismo metadato Dublin Core.

	Diferencias en relación al contenido		
Etiqueta	**Phronesis**	**CIRIA**	**Redalyc**
dc:creator	*Apellido(s), Nombre*	*Apellido(s), Nombre*	Nombre Apellido(s)
dc:publisher	*Revista donde se publico el documento*	Institución que publica el documento	Institución que publica el documento
dc:description	Resumen del documento	Resumen del documento	Palabras clave del documento

Tabla 4.2: Diferencias entre datos procedentes de diferentes repositorios (Phronesis, CIRIA o Redalyc) etiquetados con un mismo metadato Dublin Core.

La estandarización del contenido en el caso de los metadatos dc:creator y dc:publisher (ver diferencias en la tabla 4.2) se llevo a cabo mediante el procesamiento de tokens, la estandarización consistió en:

- *Metadato dc:creator.* En el caso de los registros procedentes de los repositorios de

Phronesis y CIRIA su contenido se ajusto al formato de los registros procedentes de Redalyc (Nombre Apellido(s)).

- *Metadato dc:publisher.* En el caso de los registros procedentes del repositorio de Phronesis su contenido se cambio por el nombre de la institución de donde proceden (Instituto Tecnológico y de Estudios Superiores de Monterrey).

El procesamiento de tokens es utilizado durante la etapa 1 con el propósito de generar los archivos de entrada (ver figura 3.4). El objetivo es obtener un archivo de entrada por cada uno de los registros cosechados y cada uno de estos archivos debe cumplir con las características definidas en la sección 3.1.1 para ser utilizados como datos de entrada del agrupamiento FIHC.

El contenido de cada archivo de entrada se obtiene a partir de una copia de la información que se encuentra en los metadatos *dc:tilte, dc:subject y dc:description.* Mediante el procesamiento de tokens se eliminan los caracteres de formato (mayúsculas, espaciado, signos de puntuación). También son eliminadas las palabras (artículos, preposiciones y pronombres) que por su frecuencia y/o semántica no poseen valor discriminatorio entre documentos, son eliminadas mediante un listado que contiene dichas palabras y a través del procesamiento de tokens son identificadas.

Almacenamiento en base de datos de información relevante.

La información que se almacena en base de datos proviene de cada registro que es procesado (mediante el uso de expresiones regulares) en busca de la información que es relevante para la inferencia de las redes temáticas de colaboración, en este caso, la información que por su ámbito de aplicación debe almacenarse en base de datos proviene de los metadatos especificados en la tabla 3.1.

En lo referente al almacenamiento de la información fue necesario crear reglas que permitan determinar la unicidad para los autores, instituciones y los documentos procedentes de los registros cosechados. En cada uno de los registros cosechados los metadatos dc:creator, dc:publisher y dc:title contienen el nombre del autor, nombre de la institución y titulo del documento respectivamente.

Las reglas de unicidad se crearon en base a la información disponible en los registros cosechados y son:

- *Unicidad de autor.* Ante una coincidencia entre el nombre de un autor que intenta almacenarse en base de datos y el nombre de un autor ya existente en base datos, se utiliza como criterio de unicidad la institución donde publica el autor ya existente en base de datos. Si la institución en donde publican ambos es la misma, se trata del mismo autor y no se crea una nueva entrada en la base de datos. En caso de que los autores en "conflicto"publiquen en instituciones diferentes, se

trata de autores diferentes y se crea una nueva entrada en la base de datos. El flujo de datos involucrado al momento de verificar la unicidad de un autor se muestra en la figura 4.10.

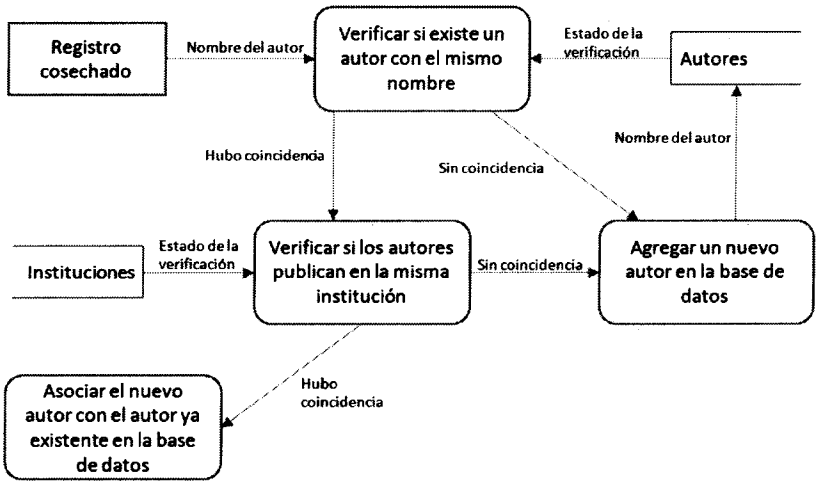


Figura 4.10: Verificación de la unicidad de un autor

- *Unicidad de institución.* Ante una coincidencia entre el nombre de una institución que intenta almacenarse en base de datos y el nombre de una institución ya existente en base datos, se utiliza como criterio de unicidad el país de origen de la institución ya existente en base de datos. Si el país donde se encuentran ambas es el mismo, se trata de la misma institución y no se crea una nueva entrada en la base de datos. En caso de que las instituciones en "conflicto"se encuentren en países diferentes, se trata de instituciones diferentes y se crea una nueva entrada en la base de datos. El flujo de datos involucrado al momento de verificar la unicidad de una institución se muestra en la figura 4.11.

- *Unicidad de documento.* Ante una coincidencia entre el titulo de un documento que intenta almacenarse en base de datos y el titulo de un documento ya existente en base datos, se utilizan como criterio de unicidad el autor y la institución asociados con el documento ya existente en base de datos. Si el autor y la institución son los mismos en ambos, se trata del mismo documento y no se crea una nueva entrada en la base de datos. En caso de que los documentos en "conflicto"sean de diferente autor y de diferente institución, se trata de documentos diferentes y se crea una nueva entrada en la base de datos. El flujo de datos involucrado al momento de verificar la unicidad de un documento se muestra en la figura 4.12.

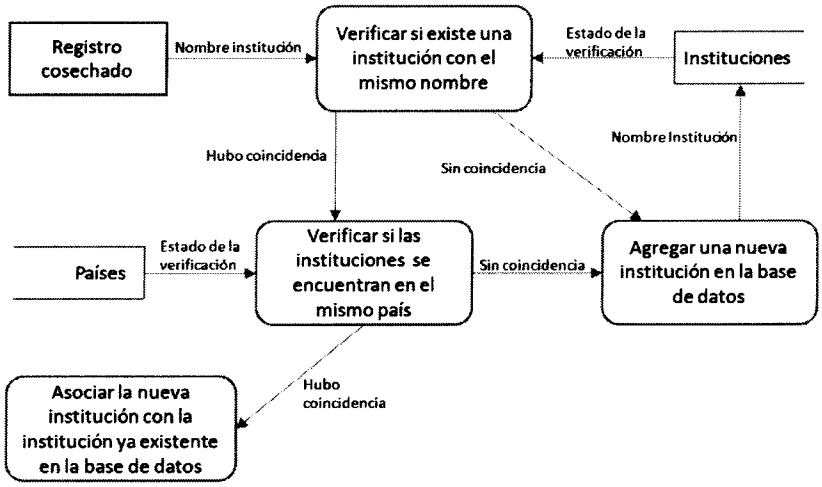


Figura 4.11: Verificación de la unicidad de una institución

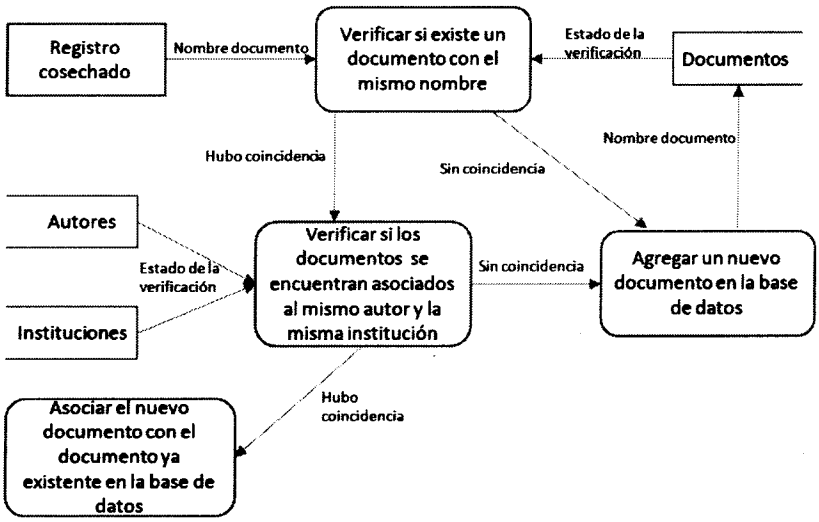


Figura 4.12: Verificación de la unicidad de un documento

4.1.2. Características y situaciones presentes durante la implementación de la etapa 2.

La etapa 2 de la metodología corresponde a lo descrito en la sección 3.1.2. Para realizar el agrupamiento mediante el algoritmo FIHC fue utilizada la versión FIHC 1.0 disponible en [24] y desarrollada por [9], la cual es gratuita y su modificación está permitida para fines académicos y de investigación.

Esta versión provee un mecanismo de pre-procesamiento para los documentos que se desean agrupar, sin embargo dicho mecanismo fue eliminado por las siguientes razones:

- Los archivos de entrada que se generan como resultado de lo descrito en la etapa 1 (ver sección 3.1.1) ya cumplen con el pre-procesamiento requerido para ser agrupados mediante el algoritmo FIHC.
- El pre-procesamiento de esta aplicación fue desarrollada para agrupar documentos en ingles.

En cuanto a las características de agrupamiento que valen la pena resaltar para hacer uso del algoritmo FIHC se encuentran [8]:

- *Escalabilidad.* Este algoritmo ha demostrado ser eficiente en colecciones de gran tamaño. Algunos algoritmos trabajan bien sobre colecciones pequeñas, pero en el mundo real las colecciones son de cientos o miles de documentos.

  El algoritmo FIHC es eficiente y escalable ya que los vectores representativos de sus documentos contienen únicamente términos de frecuencia global con lo que reduce significativamente la dimensionalidad del conjunto de documentos a agrupar.
- *Calidad.* Este algoritmo ha demostrado tener una alta similitud interna en sus grupos grupo y baja similitud entre grupos.
- *Facilidad de búsqueda.* Resultado de la estructura jerárquica del agrupamiento resultante y la descripción de cada unos de los grupos.
- *No es necesario un experto del dominio.* Algunos algoritmos como k-means requieren de un experto del tipo de documentos a agrupar que especifique en base a su "experiencia.[el] número de grupos que se esperan obtener. De lo contrario la calidad del agrupamiento se degrada drásticamente.

Los parámetros de entrada necesarios para la ejecución de la aplicación FIHC 1.0 son los siguientes:

- *Soporte global.* Porcentaje mínimo para determinar qué términos pertenecen al conjunto global de elementos frecuentes.
- *Soporte de grupo.* Porcentaje mínimo para determinar si un elemento de frecuencia global es considerado dentro de un grupo C_i, dado que el elemento se encuentra en un porcentaje mínimo de los documentos de C_i.
- *Bandera booleana.* Indica que algoritmo se utilizara para realizar el agrupamiento (k-means/FIHC).
- *Número de grupos deseados.* Indica el número deseado de grupos (aplica para k-means o FIHC).
- *Directorio fuente.* Indica la ruta donde se encuentran los archivos de entrada para realizar el agrupamiento.

Los valores asignados a los parámetros de entrada son los siguientes:

- *Soporte global.* El valor asignado es 0.05, se usa este porcentaje en base a la recomendación dada en [9], resultado de numerosos experimentos y debido a que el número de documentos a agrupar es de alrededor de los 100,000.
- *Soporte de grupo.* El valor asignado es 0.25, se usa este porcentaje en base a [9], donde se indica que con este porcentaje se siempre se producen de manera consistente grupos de alta calidad.
- *Bandera booleana.* El valor asignado es FALSE para seleccionar el algoritmo de agrupamiento FIHC.
- *Número de grupos deseados.* El valor asignado es 0 debido a que en [8] se menciona que una calidad cercana al óptimo se obtiene si el usuario no especifica este parámetro.
- *Directorio fuente.* Se le asigna la ruta relativa donde se encuentra el directorio que contiene los archivos de entrada.

4.1.3. Características y situaciones presentes durante la implementación de la etapa 3.

La etapa 3 de la metodología corresponde a lo descrito en la sección 3.1.3, mediante la cual se obtienen las redes que expresan la posibilidad o grado de colaboración entre autores o entre instituciones.

Durante la implementación de la etapa 3 se pueden identificar dos procesos para describir los detalles de su implementación:

- Uso de las métricas para inferir las redes temáticas de colaboración.
- Representación de las redes temáticas de colaboración.

Uso de las métricas para inferir las redes temáticas de colaboración.

En relación a las métricas para inferir las redes temáticas de colaboración, estas nos permiten determinar la relación entre dos entidades de colaboración (autores o instituciones) en base a su *frecuencia de afinidad temática* y al *número total de entidades en grupos*. Basados en lo anterior la implementación de la etapa consistió en recuperar la información y definir las estructuras de datos involucrados en el cálculo de los siguientes conceptos:

- Exclusividad.
- Frecuencia de afinidad temática.
- Pesos normalizados.

La información recuperada proviene de dos fuentes:

- *El agrupamiento jerárquico.* El agrupamiento jerárquico posee los identificadores de los documentos y el grupo al que pertenece cada documento.
- *La información almacenada en la base de datos.* En la base de datos se encuentra la información asociada a cada uno de los identificadores de los documentos agrupados.

Mediante un cruce de información entre el agrupamiento jerárquico y la base de datos podemos identificar plenamente que entidades de colaboración (autores o instituciones) están asociadas a los documentos de un grupo en particular, y obtener una representación similar a la que se muestra en la figura 4.13 para cada uno de los grupos.

De esta forma cada grupo está asociado a sus entidades de colaboración (autores e instituciones). Con el nuevo conocimiento acerca de los grupos obtenidos podemos aplicar las métricas para determinar la red temática de colaboración de autores o la red temática de colaboración de instituciones, considerando como entidad de colaboración a los autores o a las instituciones, según sea el caso, al momento de aplicar las métricas.

En relación a la estructura de datos necesaria para el cálculo de las métricas de *exclusividad, frecuencia de afinidad temática y pesos normalizados* es necesaria una matriz tamaño nxn, donde n es el número total de entidades a relacionar. De esta manera el pseudocódigo del algoritmo mediante el cual se hace el cálculo de las métricas

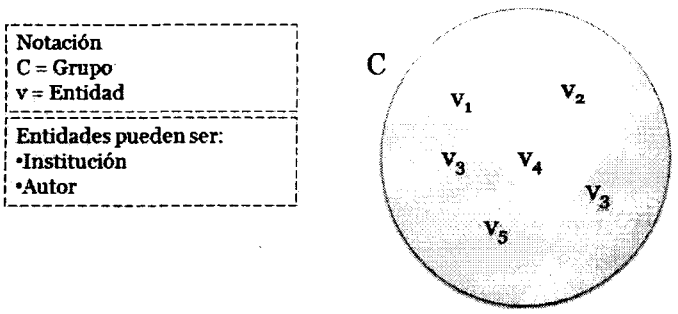


Figura 4.13: Representación de un grupo en base a entidades de colaboración

Pseudocódigo : Cálculo de las métricas: exclusividad, frecuencia de afinidad temática y pesos normalizados.

for all $grupo \in Agrupamiento$ **do**
  $fck \leftarrow numeroEntidadesGrupo(grupo)$
  for all $entidad_i \in grupo$ **do**
    for all $entidad_j \in grupo$ **do**
      if $entidad_i = entidad_j$ **then**
        $g[entidad_i][entidad_j] \leftarrow 0$
        $s[entidad_i][entidad_j] \leftarrow 0$
      else
        $jck \leftarrow presenciaEntidad(entidad_j)$
        $g[entidad_i][entidad_j] \leftarrow (1 \div (1 - fck)) * jck$
        $s[entidad_i][entidad_j] \leftarrow s[entidad_i][entidad_j] + g[entidad_i][entidad_j]$
      end if
      $sum[entidad_i] \leftarrow sum[entidad_i] + s[entidad_i][entidad_j]$
    end for
  end for
end for
for all $entidad_i \in entidadesAgrupamiento$ **do**
  for all $entidad_j \in entidadesAgrupamiento$ **do**
    if $entidad_i = entidad_j$ **then**
      $W[entidad_i][entidad_j] \leftarrow 0$
    else
      $W[entidad_i][entidad_j] \leftarrow s[entidad_i][entidad_j] \div sum[entidad_i]$
    end if
  end for
end for

Figura 4.14: Cálculo de las métricas: exclusividad, frecuencia de afinidad temática y pesos normalizados.

se muestra en la figura 4.14. La complejidad del algoritmo es $O(n^2)$ en relación al número de *entidades de colaboración* (n) a relacionar.

En el caso de los autores se requiere una matriz de tamaño 84673x84673 para almacenar los resultados de las métricas (84673 son los autores encontrados en las colecciones digitales phronesis, CIRIA y Redalyc). Para mantener la precisión de dichos resultados es necesario utilizar el tipo de dato float que ocupa 4 bytes. En consecuencia, una matriz de tamaño 84673x84673 consume 26,7 GB de espacio y se exceden las capacidades de almacenamiento en memoria RAM.

La solución fue implementar un mapeo para almacenar cada uno de los elementos de la matriz en una posición específica dentro de un archivo, de manera que cada elemento de la matriz ocupa 4 bytes en el archivo, el mapeo se representa en la figura 4.15.

Para establecer la correspondencia entre la posición de los elementos de la matriz y la posición de los elementos en el archivo se utiliza la función 4.1, la cual permite calcular en qué byte se coloca el puntero del archivo para la lectura o escritura de cada uno de los elementos de la matriz.

$$m = (indice(e1) * n) + indice(e2)) * 4 \tag{4.1}$$

Donde m es el byte donde se coloca el puntero del archivo, $indice(e1)$ e $indice(e2)$ corresponden a la posición de un elemento de la matriz nxn, n es el número de entidades de colaboración y 4 corresponde al tamaño en bytes del tipo de dato float. El tamaño del archivo en bytes para realizar el mapeo es igual a $n^2 * 4$.

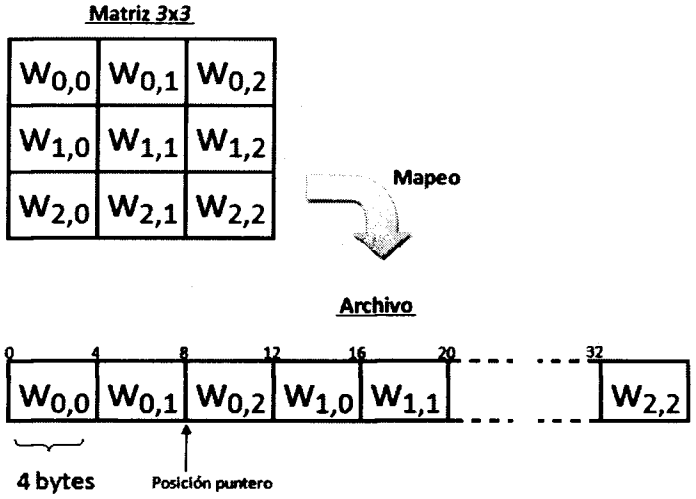


Figura 4.15: Representación del mapeo de los elementos de una matriz a un archivo

La matriz de tamaño 3x3 mostrada en la figura 4.15 almacena el resultado de las métricas para 3 entidades, el tamaño del archivo requerido para almacenar una matriz

de estas características es de 36 bytes ($3^2 * 4$). El valor almacenado en el elemento (0,2) de la matriz es mapeado al archivo colocando el puntero del archivo en el byte 8 $(((0 * 3) + 2) * 4)$.

Representación de las redes temáticas de colaboración.

Los datos a partir de los cuales se generan tanto la red temática de colaboración a nivel de instituciones (ver figura 3.9) como la red temática de colaboración a nivel de autores (ver figura 3.11) provienen de distintas fuentes. Para crear los archivos XML correspondientes a cada una de las estructuras de colaboración es necesario identificar el origen de dichos datos. La tabla 4.3 muestra para ambas estructuras el origen de los datos requeridos en las etiquetas dataSet y relationSet, respectivamente.

Etiqueta	**Red temática de colaboración a nivel de instituciones**	**Red temática de colaboración a nivel de autores**
<dataSet>	Los atributos requeridos en la etiqueta <institution> provienen de la base de datos. Los atributos requeridos en la etiqueta <author> provienen de la base de datos. El contenido requerido en la etiqueta <subject> proviene del agrupamiento.	Los atributos requeridos en la etiqueta <author> provienen de la base de datos. Los atributos requeridos en la etiqueta <document> provienen de la base de datos. El contenido requerido en la etiqueta <subject> proviene del agrupamiento.
<relationSet>	Los atributos requeridos en la etiqueta <relation> provienen del uso de las métricas para inferir las redes temáticas de colaboración tomando como entidad de colaboración a las instituciones.	Los atributos requeridos en la etiqueta <relation> provienen del uso de las métricas para inferir las redes temáticas de colaboración tomando como entidad de colaboración a los autores.

Tabla 4.3: Origen de los datos requeridos para generar los archivos XML de las estructuras de colaboración

En base a la información de la tabla 4.3 la implementación consistió en recuperar dicha información y crear los archivos XML correspondientes a la red de colaboración a nivel de instituciones y la red de colaboración a nivel de autores.

4.2. Resultados.

Los resultados análogamente a los detalles de la implementación se muestran desde una perspectiva en relación a cada una de las etapas de la metodología, lo anterior con el propósito de dar mayor claridad al análisis de resultados.

4.2.1. Resultados correspondientes a la etapa 1.

Los resultados de la etapa 1 se presentan en relación a cada uno de los procesos cuyos detalles de implementación son mencionados en la sección 4.1, los procesos son: Recolección de registros. Uso de expresiones regulares y procesamiento de tokens Almacenamiento en base de datos de información relevante

Recolección de Registros

Los resultados de la recolección de registros representada en la figura 4.1 se muestran en la tabla 4.4. Por lo tanto, el número total de documentos procedentes de las bibliotecas digitales Phronesis, CIRIA y Redalyc es de 114,912.

Biblioteca Digital	Registros cosechados
Phronesis	404
CIRIA	1,628
Redalyc	112,880

Tabla 4.4: Cantidad de registros provenientes de las bibliotecas digitales Phronesis, CIRIA y Redalyc

Uso de expresiones regulares y procesamiento de tokens

Como resultado del uso de expresiones regulares se logra procesar la información contenida en los 114,912 registros procedentes de las colecciones de Phronesis, CIRIA y Redalyc. Mediante el procesamiento de tokens se obtienen el mismo número de archivos de entrada para el agrupamiento FIHC (uno por cada registro cosechado).

Sin embargo, aplicando la regla de unicidad de documentos son descartados 860 registros y únicamente se conservan 114,052 registros (descartando los documentos repetidos) como archivos de entrada para el agrupamiento FIHC.

La figura 4.16 representa al archivo de entrada que se obtiene a partir del contenido del registro mostrado en la figura 4.6.

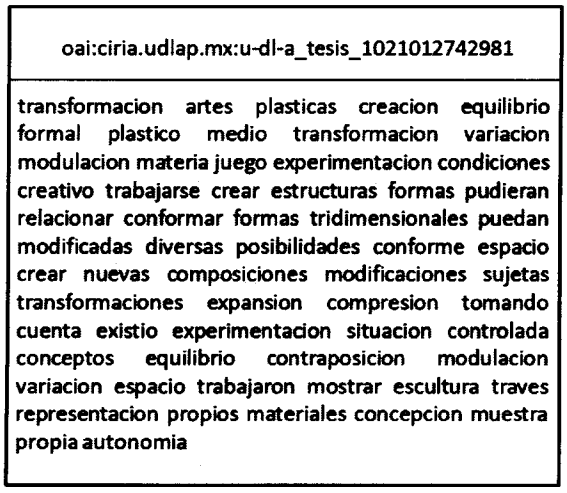


Figura 4.16: Archivo de entrada para agrupamiento FIHC correspondiente al registro de la figura 4.6

Almacenamiento en base de datos de información relevante

Como resultado del procesamiento de los registros cosechados es posible almacenar en base de datos la información relevante para la inferencia de las redes temáticas de colaboración. La tabla 4.5 muestra el número de documentos, autores e instituciones almacenados en base de datos provenientes de los repositorios Phronesis, CIRIA y Redalyc.

	Phronesis	**CIRIA**	**Redalyc**	**TOTAL**
Documentos	404	1,628	112,020	114,052
Autores	383	1,620	82,670	84,673
Instituciones	1	1	311	313

Tabla 4.5: Número de documentos, autores e instituciones almacenados en base de datos provenientes de los repositorios Phronesis, CIRIA y Redalyc.

Al comparar el número total de documentos (114,912) procedentes de la bibliotecas digitales Phronesis, CIRIA y Redalyc con el número total de documentos almacenados en base de datos (114,052) existe una diferencia de 860 documentos. Esta diferencia se debe a la aplicación de la regla de unicidad de documento, e implica que entre las colecciones existen documentos en común.

En el caso de los documentos procedentes de los repositorios Phronesis y CIRIA, todos son publicados por el Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM) y la Universidad de las Américas Puebla (UDLAP) respectivamente. Las tablas 4.6 y 4.7 muestran las diferencias entre la información de la base de datos y la información de los registros cosechados de Phronesis y CIRIA respectivamente. En

ambos casos las diferencias observadas se deben a la presencia de documentos publicados por el ITESM y la UDLAP que no se encuentran en los repositorios Phronesis y CIRIA, pero si en el repositorio de Redalyc.

		Phronesis	**Base de datos**
ITESM	Documentos	404	823
	Autores	383	728

Tabla 4.6: Diferencia entre la información de la base de datos y los registros cosechados de Phronesis.

		CIRIA	**Base de datos**
UDLAP	Documentos	1,628	1,857
	Autores	1,620	1,772

Tabla 4.7: Diferencia entre la información de la base de datos y los registros cosechados de CIRIA.

Las situaciones observadas en las tablas 4.5, 4.6 y 4.7 hacen evidentes la necesidad de las reglas de unicidad de autor, institución y documento.

4.2.2. Resultados correspondientes a la etapa 2.

El resultado de la etapa 2 es un árbol de agrupamiento construido mediante el algoritmo FIHC. El árbol de agrupamiento se obtiene a partir de un total de 114,052 documentos (archivos de entrada). Parte de la estructura del árbol de agrupamiento obtenido se muestra en la figura 4.17.

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
- <root num_clusters="19" label="root" num_children="19" num_docs="140052">
  + <documents num_docs="1349">
  - <cluster label="administracion" num_children="0" num_docs="4386">
    + <documents num_docs="4386">
    </cluster>
  - <cluster label="agrociencia" num_children="0" num_docs="7880">
    + <documents num_docs="7880">
    </cluster>
  - <cluster label="antropologia" num_children="0" num_docs="4506">
    + <documents num_docs="4506">
    </cluster>
  - <cluster label="biologia" num_children="0" num_docs="7707">
    + <documents num_docs="7707">
    </cluster>
  - <cluster label="comunicacion" num_children="0" num_docs="4816">
    + <documents num_docs="4816">
    </cluster>
  - <cluster label="cultura" num_children="0" num_docs="3534">
    + <documents num_docs="3534">
    </cluster>
  - <cluster label="derecho" num_children="0" num_docs="4872">
    + <documents num_docs="4872">
    </cluster>
  - <cluster label="economia" num_children="0" num_docs="6131">
    + <documents num_docs="6131">
    </cluster>
  - <cluster label="educacion" num_children="0" num_docs="9070">
    + <documents num_docs="9070">
    </cluster>
  - <cluster label="historia" num_children="0" num_docs="6830">
    + <documents num_docs="6830">
    </cluster>
```

Figura 4.17: Parte del árbol de agrupamiento obtenido mediante el algoritmo FIHC.

El árbol de agrupamiento cuenta un total de 19 grupos cuyas etiquetas de agrupamiento y detalles se muestran en la tabla 4.8. En relación al árbol de agrupamiento podemos mencionar que ningún grupo es de nivel dos en el árbol de agrupamiento (todos los grupos son hijos de la raíz del árbol). De igual forma podemos comentar que existen documentos (con términos sin la frecuencia necesaria para pertenecer a algún grupo) agrupados en la raíz del árbol sin asociarse a alguna etiqueta de agrupamiento y que se pueden encontrar grupos con grandes diferencias en cuanto a tamaño (número de elementos agrupados).

Etiqueta de agrupamiento	Padre del grupo	Número de documentos agrupados
Root	-	1,349
Administración	Root	4,386
Agrociencia	Root	7,880
Antropología	Root	4,506
Biología	Root	7,707
Comunicación	Root	4,816
Cultura	Root	3,534
Derecho	Root	4,872
Economía	Root	6,131
Educación	Root	9,070
Historia	Root	6,830
Humanidad	Root	1,267
Ingeniería	Root	5,729
Literatura	Root	4,044
Medicina	Root	5,766
Política	Root	6,001
Psicología	Root	10,503
Química	Root	3,932
Salud	Root	8,290
Sociología	Root	7,439

Tabla 4.8: Características del árbol de agrupamiento obtenido mediante el algoritmo FIHC.

4.2.3. Resultados correspondientes a la etapa 3.

Como resultado de la etapa 3 se obtienen las estructuras de colaboración que se encuentran representadas en la figuras 3.9 y 3.11, las cuales corresponden a la red de colaboración a nivel de instituciones y a la red de colaboración a nivel de autores respectivamente.

Debido a la cantidad de elementos que conforman las estructuras de colaboración obtenidas, en las figuras 4.18 y 4.19 se muestra parte de la red de colaboración a nivel de instituciones. Y en las figuras 4.20 y 4.21 se muestra parte de la red de colaboración a nivel de autores.

La tabla 4.9 contiene el detalle de la red temática de colaboración a nivel instituciones, y la tabla 4.10 contiene el detalle de la red temática de colaboración a nivel de autores. En ambos casos el detalle es en relación al contenido de las etiquetas dataSet y relationSet de los archivos XML obtenidos.

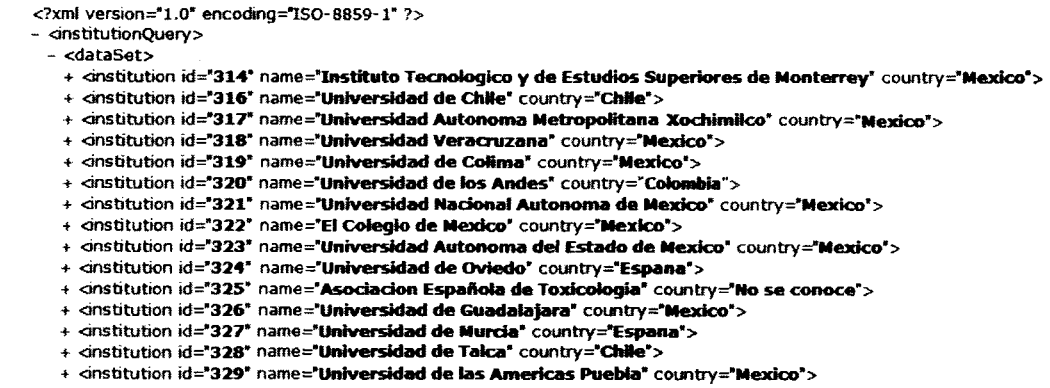


Figura 4.18: Parte del dataSet de la red de colaboración a nivel de instituciones

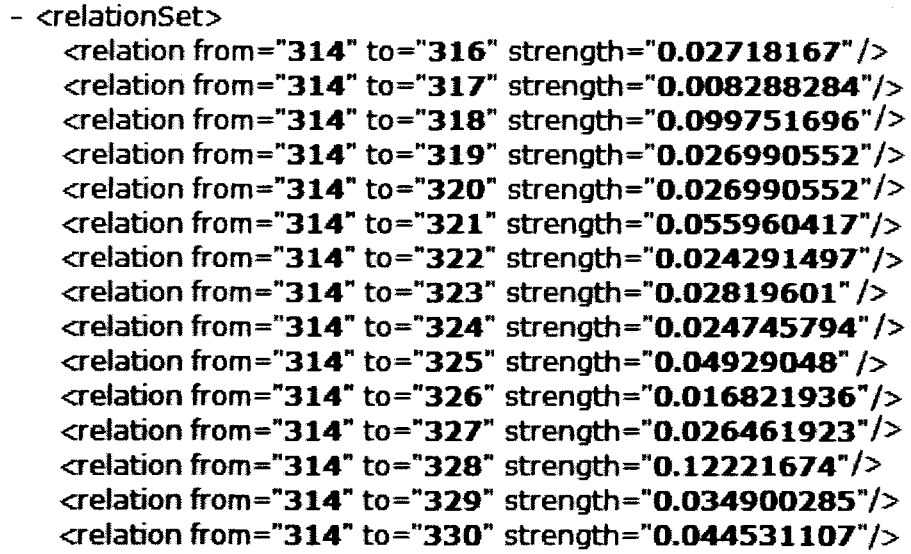


Figura 4.19: Parte del relationSet de la red de colaboración a nivel de instituciones.

<?xml version="1.0" encoding="ISO-8859-1" ?>
- <authorQuery>
  - <dataSet>
    + <author id="85569" name="David A. Garza Salazar" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">
    + <author id="85570" name="Horacio Martinez Alfaro" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">
    + <author id="85572" name="Carlos Gaitan Chavez" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">
    + <author id="85573" name="Jose Vladimir Burgos Aguilar" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">
    + <author id="85574" name="Roberto Jose Garcia Flores" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">
    + <author id="85575" name="Alejandro Esteban Marcus Martinez" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">
    + <author id="85576" name="Guillermo Ernesto Ponce Campos" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">
    + <author id="85578" name="Blanca Nallely Villarreal Dominguez" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">
    + <author id="85579" name="Beatriz Adriana Flores Clemente" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">
    + <author id="85580" name="Ruth Josefina Sanchez Zambrano" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">
    + <author id="85581" name="Maria Enriqueta Lopez Galvan" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">
    + <author id="85582" name="Arturo Medrano Leal" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">
    + <author id="85583" name="Deyanira Meza Martell" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">
    + <author id="85584" name="Alejandro Melchor Leon" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">
    + <author id="85585" name="Javier Ivan Limon Pavia" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">
    + <author id="85586" name="Angel Omar Hernandez Aguilera" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">
    + <author id="85587" name="Pedro Hernandez" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">
    + <author id="85588" name="Luis Daniel Abella Reyes" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">
    + <author id="85589" name="Delia Magaly Rotuno Espino" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">
    + <author id="85590" name="Ileana Melisa Pimentel Sopalda" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">
    + <author id="85592" name="Jose Jorge Furber Cano" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">
    + <author id="85593" name="Ismaylia Saucedo Ugalde" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">
    - <author id="85594" name="Ricardo Jimenez Torres" Institution="Instituto Tecnologico y de Estudios Superiores de Monterrey">

Figura 4.20: Parte del dataSet de la red de colaboración a nivel de autores.

- <relationSet>
  <relation from="85569" to="85570" strength="0.011235955" />
  <relation from="85569" to="85593" strength="0.011235955" />
  <relation from="85569" to="85605" strength="0.011235955" />
  <relation from="85569" to="85612" strength="0.011235955" />
  <relation from="85569" to="85624" strength="0.011235955" />
  <relation from="85569" to="85665" strength="0.011235955" />
  <relation from="85569" to="85670" strength="0.011235955" />
  <relation from="85569" to="85672" strength="0.011235955" />
  <relation from="85569" to="85708" strength="0.011235955" />
  <relation from="85569" to="85718" strength="0.011235955" />
  <relation from="85569" to="85755" strength="0.011235955" />
  <relation from="85569" to="85765" strength="0.011235955" />
  <relation from="85569" to="85777" strength="0.011235955" />
  <relation from="85569" to="85778" strength="0.011235955" />
  <relation from="85569" to="85779" strength="0.011235955" />
  <relation from="85569" to="85853" strength="0.011235955" />
  <relation from="85569" to="85858" strength="0.011235955" />
  <relation from="85569" to="85878" strength="0.011235955" />
  <relation from="85569" to="85885" strength="0.011235955" />
  <relation from="85569" to="85947" strength="0.011235955" />
  <relation from="85569" to="86089" strength="0.011235955" />
  <relation from="85569" to="86145" strength="0.011235955" />
  <relation from="85569" to="86146" strength="0.02247191" />

Figura 4.21: Parte del relationSet de la red de colaboración a nivel de autores.

Etiqueta	Contenido
<dataSet>	Contiene un total de 313 instituciones, y los 84,673 autores asociados a sus respectivas instituciones, así como las asociaciones entre cada uno de los autores y sus respectivas temáticas, de un total de 19 temáticas.
<relationSet>	Contiene un máximo de 97,656 relaciones entre pares de instituciones (descartando la relación entre una institución consigo misma), de manera, que para cada institución se tiene el grado de colaboración con un máximo de 312 instituciones (descartando las relaciones con peso igual a cero), ya que como se muestra en la figura 3.8 no necesariamente es reciproca la colaboración entre dos instituciones.

Tabla 4.9: Detalle del contenido de la red temática de colaboración a nivel de instituciones.

Etiqueta	Contenido
<dataSet>	Contiene un total de 84,673 autores, y los 114,052 documentos asociados a sus respectivos autores, así como las asociaciones entre cada uno de los documentos y su temática, de un total de 19 temáticas.
<relationSet>	Contiene un máximo de 7,169,432,256 relaciones entre pares de autores (descartando la relación entre un autor consigo mismo), de manera, que para cada autor se tiene el grado de colaboración con un máximo de 84,672 autores (descartando las relaciones con peso igual a cero), ya que como se muestra en la figura 3.8 no necesariamente es reciproca la colaboración entre dos autores.

Tabla 4.10: Detalle del contenido de la red temática de colaboración a nivel de autores.

Alcance de la inferencia de redes temáticas de colaboración en bibliotecas digitales

A continuación se analiza el alcance de la inferencia de redes temáticas de colaboración en base a una muestra de documentos provenientes del repositorio de CIRIA, la tabla 4.11 contiene la información de la muestra.

Debido a que todos los documentos provienen de la misma institución (UDLAP) la inferencia de la red temática de colaboración se realizó a nivel de autores. En base a la clasificación (ver tabla 4.11) de los documentos que se proporciona en el sitio de web de CIRIA se puede realizar un agrupamiento lógico a través del cual se puede considerar la posibilidad de colaboración entre los autores asociados a una misma clasificación o temática, sin embargo no puede cuantificarse la fortaleza de dicha colaboración. La tabla 4.12 muestra el agrupamiento lógico basado en la clasificación del documento.

Aplicando la metodología descrita en la sección 3.1 a los documentos de la muestra, se obtiene lo siguiente:

- Como resultado de la recolección, pre-procesamiento e identificación relevante (etapa 1), se obtienen 10 archivos de entrada para realizar el agrupamiento jerárquico y se almacena en la base de datos la información que es utilizada al momento de generar las redes temáticas de colaboración de acuerdo a lo especificado en la tabla 3.1.

- El resultado del agrupamiento jerárquico (etapa 2) se muestra en la figura 4.22.

- El resultado de la generación de la red temática de colaboración se muestra en las figuras 4.23 y 4.24, secciones dataSet y relationSet, respectivamente.

La colaboración mostrada en la tabla 4.13 corresponde a la red temática de colaboración de las figura 4.23 y 4.24. Comparando la tabla 4.13 con respecto a la colaboración mostrada en la tabla 4.12 puede observarse que:

- Las relaciones de colaboración han cambiado (inclusive mezclando elementos de los grupos lógicos).

- La fortaleza de colaboración entre autores es cuantificable.

- La exploración de la de colaboración entre los autores es más detallada.

Es importante enfatizar que debido al tamaño de la muestra y que todos los documentos provienen de un único repositorio (CIRIA) es posible identificar grupos lógicos de colaboración fácilmente y de una manera manual. Sin embargo, ampliando la situación a colecciones provenientes de diferentes repositorios y con un número mucho mayor de documentos, la identificación manual de grupos lógicos (en base a su clasificación) sería muy complicada.

Clasificación	Autor	Título del documento
Ingeniería de procesos	Estrella Urania Ávalos Santos	Desarrollo de la batería de separación modular del campo Puerto Ceiba.
Ingeniería de procesos	Roberto Carreón Sierra	Diseño conceptual de la Infraestructura de Explotación del Campo Costero.
Ingeniería de procesos	Juan Ulises Martínez López	Opciones para la optimización en el manejo de la producción en la Batería de Separación y Estación de Compresión Tecominoacán, Ubicada en el municipio de Huamanguillo, en el Estado de Tabasco.
Ingeniería de procesos	Osiris Priego Feria	Propuesta e implantación de mejora del desalado de Petróleo Crudo del Campo Samario III. Efecto del uso de dos etapas y modificación del mezclado del agua de dilución.
Ingeniería Química	Carmen Ivette Arzate Echeverría	Estudio de la factibilidad técnica y económica de la instalación de un taller para elaborar cerámica usando como materia prima residuos industriales.
Ingeniería Química	María Isabel Gómez Suárez	Segunda etapa de hidrotratamiento catalítico de gasóleos del petróleo maya empleando catalizadores CoMo y NiMo soportados.
Ingeniería Química	Arlen de María Méndez Martínez	Obtención de capsicinoides a partir de la merma en el enlatado de chiles de conserva.
Ingeniería Química	Laura Teresa Morales Gámez	Desarrollo de una propuesta para la construcción de una planta piloto para el reciclaje de los componentes de las baterías níquel-cadmio.
Ingeniería Química Industrial	Miguel Angel Gutiérrez Fernández	Arenas sílicas
Ingeniería Química Industrial	María del Carmen Meneses Herrera	Hidrodesulfurización de Gasóleos a partir de Crudo Maya II

Tabla 4.11: Detalles de la muestra de documentos provenientes del repositorio CIRIA.

Agrupamiento lógico	Autores
Ingeniería de procesos	Estrella Urania Ávalos Santos
	Roberto Carreón Sierra
	Juan Ulises Martínez López
	Osiris Priego Feria
Ingeniería Química	Carmen Ivette Arzate Echeverría
	María Isabel Gómez Suárez
	Arlen de María Méndez Martínez
	Laura Teresa Morales Gámez
Ingeniería Química Industrial	Miguel Angel Gutiérrez Fernández
	María del Carmen Meneses Herrera

Tabla 4.12: Agrupamiento lógico de la muestra en base a la clasificación del documento.

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
- <root num_clusters="4" label="null" num_children="4" num_docs="10">
    <documents num_docs="0" />
  - <cluster label="medio ambiente" num_children="0" num_docs="5">
    - <documents num_docs="5">
        <document>OAI_CIRIA.UDLAP.MX_U_DL_A_TESIS_4062015154971</document>
        <document>OAI_CIRIA.UDLAP.MX_U_DL_A_TESIS_5021015700981</document>
        <document>OAI_CIRIA.UDLAP.MX_U_DL_A_TESIS_4062012675981</document>
        <document>OAI_CIRIA.UDLAP.MX_U_DL_A_TESIS_5021015676981</document>
        <document>OAI_CIRIA.UDLAP.MX_U_DL_A_TESIS_4062063310981</document>
      </documents>
    </cluster>
  - <cluster label="extraccion hidrocarburos" num_children="0" num_docs="3">
    - <documents num_docs="3">
        <document>OAI_CIRIA.UDLAP.MX_U_DL_A_TESIS_5021015433981</document>
        <document>OAI_CIRIA.UDLAP.MX_U_DL_A_TESIS_5021015442981</document>
        <document>OAI_CIRIA.UDLAP.MX_U_DL_A_TESIS_5021015542982</document>
      </documents>
    </cluster>
  - <cluster label="refinamiento" num_children="0" num_docs="1">
    - <documents num_docs="1">
        <document>OAI_CIRIA.UDLAP.MX_U_DL_A_TESIS_5021015543981</document>
      </documents>
    </cluster>
  - <cluster label="energia" num_children="0" num_docs="1">
    - <documents num_docs="1">
        <document>OAI_CIRIA.UDLAP.MX_U_DL_A_TESIS_4062012604971</document>
      </documents>
    </cluster>
</root>
```

Figura 4.22: Agrupamiento jerárquico resultante.

<?xml version="1.0" encoding="ISO-8859-1" ?>
- <authorQuery>
  - <dataSet>
    - <author id="1" name="Carmen IvetteArzate Echeverría" Institution="Universidad de las Américas Puebla">
      - <document title="Estudio de la factibilidad técnica y económica de la instalación de un taller para elaborar cerámica usando como materia prima residuos industriales" url="http://catarina.udlap.mx/u_dl_a/tales/documentos/mqi/arzate_e_ci/" date="2005-08-16">
        <subject>medio ambiente</subject>
      </document>
    </author>
    - <author id="2" name="Arlen de MariaMéndez Martínez" Institution="Universidad de las Américas Puebla">
      - <document title="Obtención de capsicinoides a partir de la merma en el enlatado de chiles de conserva" url="http://catarina.udlap.mx/u_dl_a/tales/documentos/mqi/mendez_m_ad/" date="2005-09-30">
        <subject>medio ambiente</subject>
      </document>
    </author>
    - <author id="3" name="Laura TeresaMorales Gámez" Institution="Universidad de las Américas Puebla">
      - <document title="Desarrollo de una propuesta para la construcción de una planta piloto para el reciclaje de los componentes de las baterías níquel-cadmio" url="http://catarina.udlap.mx/u_dl_a/tales/documentos/mqi/morales_g_lt/" date="2005-07-01">
        <subject>energía</subject>
      </document>
    </author>
    - <author id="4" name="María IsabelGómez Suárez" Institution="Universidad de las Américas Puebla">
      - <document title="Segunda etapa de hidrotratamiento catalítico de gasóleos del petróleo maya empleando catalizadores CoMo y NiMo soportados" url="http://catarina.udlap.mx/u_dl_a/tales/documentos/mqi/gomez_s_mi/" date="2005-12-15">
        <subject>medio ambiente</subject>
      </document>
    </author>
    - <author id="5" name="Estrella UraniaÁvalos Santos" Institution="Universidad de las Américas Puebla">
      - <document title="Desarrollo de la batería de separación modular del campo Puerto Ceiba" url="http://catarina.udlap.mx/u_dl_a/tales/documentos/mip/avalos_s_eu/" date="2002-05-19">
        <subject>extraccion hidrocarburos</subject>
      </document>
    </author>
    - <author id="6" name="RobertoCarreón Sierra" Institution="Universidad de las Américas Puebla">
      - <document title="Diseño conceptual de la Infraestructura de Explotación del Campo Costero" url="http://catarina.udlap.mx/u_dl_a/tales/documentos/mip/carreon_s_r/" date="2002-05-19">
        <subject>extraccion hidrocarburos</subject>
      </document>
    </author>
    - <author id="7" name="Juan UlisesMartínez López" Institution="Universidad de las Américas Puebla">
      - <document title="Opciones para la optimización en el manejo de la producción en la Batería de Separación y Estación de Compresión Tecominoacán, Ubicada en el municipio de Huamanguillo, en el Estado de Tabasco" url="http://catarina.udlap.mx/u_dl_a/tales/documentos/mip/martinez_l_ju/" date="2002-05-19">
        <subject>extraccion hidrocarburos</subject>
      </document>
    </author>
    - <author id="8" name="OsirisPriego Feria" Institution="Universidad de las Américas Puebla">
      - <document title="Propuesta e implantación de mejora del desalado de Petróleo Crudo del Campo Samario III. Efecto del uso de dos etapas y modificación del mezclado del agua de dilución" url="http://catarina.udlap.mx/u_dl_a/tales/documentos/mip/priego_f_o/" date="2002-05-19">
        <subject>refinamiento</subject>
      </document>
    </author>
    - <author id="9" name="Miguel AngelGutiérrez Fernández" Institution="Universidad de las Américas Puebla">
      - <document title="Arenas silicas" url="http://catarina.udlap.mx/u_dl_a/tales/documentos/liq/gutierrez_f_ma/" date="2003-05-09">
        <subject>medio ambiente</subject>
      </document>
    </author>
    - <author id="10" name="María del CarmenMeneses Herrera" Institution="Universidad de las Américas Puebla">
      - <document title="Hidrodesulfurización de Gasóleos a partir de Crudo Maya II" url="http://catarina.udlap.mx/u_dl_a/tales/documentos/liq/meneses_h_md/" date="2004-12-10">
        <subject>medio ambiente</subject>
      </document>
    </author>
  </dataSet>

Figura 4.23: Red temática de colaboración a nivel de autores (dataSet).

```
- <relationSet>
    <relation from="1" to="2" strength="0.25" />
    <relation from="1" to="4" strength="0.25" />
    <relation from="1" to="9" strength="0.25" />
    <relation from="1" to="10" strength="0.25" />
    <relation from="2" to="1" strength="0.25" />
    <relation from="2" to="4" strength="0.25" />
    <relation from="2" to="9" strength="0.25" />
    <relation from="2" to="10" strength="0.25" />
    <relation from="4" to="1" strength="0.25" />
    <relation from="4" to="2" strength="0.25" />
    <relation from="4" to="9" strength="0.25" />
    <relation from="4" to="10" strength="0.25" />
    <relation from="5" to="6" strength="0.5" />
    <relation from="5" to="7" strength="0.5" />
    <relation from="6" to="5" strength="0.5" />
    <relation from="6" to="7" strength="0.5" />
    <relation from="7" to="5" strength="0.5" />
    <relation from="7" to="6" strength="0.5" />
    <relation from="9" to="1" strength="0.25" />
    <relation from="9" to="2" strength="0.25" />
    <relation from="9" to="4" strength="0.25" />
    <relation from="9" to="10" strength="0.25" />
    <relation from="10" to="1" strength="0.25" />
    <relation from="10" to="2" strength="0.25" />
    <relation from="10" to="4" strength="0.25" />
    <relation from="10" to="9" strength="0.25" />
  </relationSet>
</authorQuery>
```

Figura 4.24: Red temática de colaboración a nivel de autores (relationSet).

Temática	Del Autor	Al Autor	Fuerza de colaboración
medio ambiente	Carmen Ivette Arzate Echeverría	Arlen de María Méndez Martínez	0.25
	Carmen Ivette Arzate Echeverría	María Isabel Gómez Suárez	0.25
	Carmen Ivette Arzate Echeverría	Miguel Angel Gutiérrez Fernández	0.25
	Carmen Ivette Arzate Echeverría	María del Carmen Meneses Herrera	0.25
medio ambiente	Arlen de María Méndez Martínez	Carmen Ivette Arzate Echeverría	0.25
	Arlen de María Méndez Martínez	María Isabel Gómez Suárez	0.25
	Arlen de María Méndez Martínez	Miguel Angel Gutiérrez Fernández	0.25
	Arlen de María Méndez Martínez	María del Carmen Meneses Herrera	0.25
medio ambiente	María Isabel Gómez Suárez	Carmen Ivette Arzate Echeverría	0.25
	María Isabel Gómez Suárez	Arlen de María Méndez Martínez	0.25
	María Isabel Gómez Suárez	Miguel Angel Gutiérrez Fernández	0.25
	María Isabel Gómez Suárez	María del Carmen Meneses Herrera	0.25
extraccion hidrocarburos	Estrella Urania Ávalos Santos	Roberto Carreón Sierra	0.5
	Estrella Urania Ávalos Santos	Juan Ulises Martínez López	0.5
extraccion hidrocarburos	Roberto Carreón Sierra	Estrella Urania Ávalos Santos	0.5
	Roberto Carreón Sierra	Juan Ulises Martínez López	0.5
extraccion hidrocarburos	Juan Ulises Martínez López	Estrella Urania Ávalos Santos	0.5
	Juan Ulises Martínez López	Roberto Carreón Sierra	0.5
refinamiento	Miguel Angel Gutiérrez Fernández	Carmen Ivette Arzate Echeverría	0.25
	Miguel Angel Gutiérrez Fernández	Arlen de María Méndez Martínez	0.25
	Miguel Angel Gutiérrez Fernández	María Isabel Gómez Suárez	0.25
	Miguel Angel Gutiérrez Fernández	María del Carmen Meneses Herrera	0.25
energia	María del Carmen Meneses Herrera	Carmen Ivette Arzate Echeverría	0.25
	María del Carmen Meneses Herrera	Arlen de María Méndez Martínez	0.25
	María del Carmen Meneses Herrera	María Isabel Gómez Suárez	0.25
	María del Carmen Meneses Herrera	Miguel Angel Gutiérrez Fernández	0.25

Tabla 4.13: Relación entre los autores de la red temática de colaboración.

Lo anterior da una perspectiva mucho más clara del alcance que se tiene mediante la inferencia de las redes temáticas de colaboración como medio para encontrar grupos de colaboración en la comunidad científica que de otra forma mediante un análisis humano pasaría desapercibido.

4.2.4. Validación de resultados

Para validar el hecho de que las relaciones obtenidas corresponden a relaciones con una posibilidad real de colaboración en relación a su afinidad temática se llevaron a cabo tres experimentos:

- Validación de relaciones de colaboración.
- Validación de la fortaleza de afinidad.
- Validación de la fortaleza de colaboración.

Los experimentos están basados en una muestra de documentos provenientes de la colección de Redalyc que se detalla en la tabla 4.14.

Número de documentos	**Autores identificados**	**Relaciones de co-autoría**	**Autores con co-autoría**
1,070	910	127	106

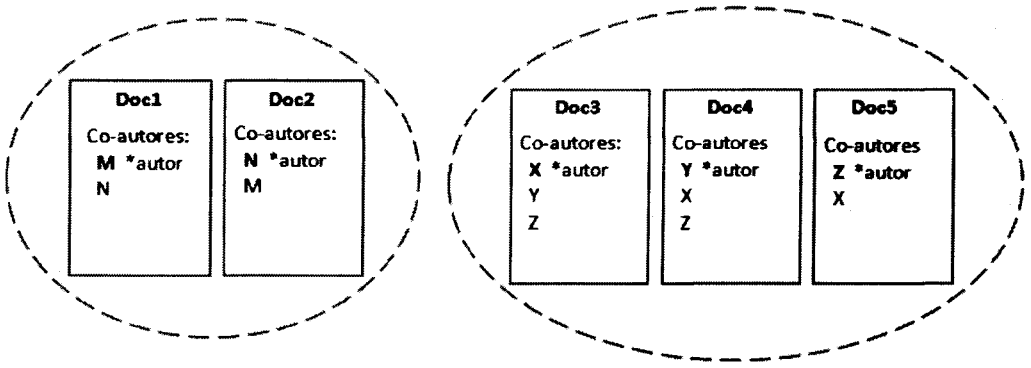
Tabla 4.14: Características de la muestra para validación de resultados.

Características de las relaciones de co-autoría:
De la muestra fueron tomadas 127 relaciones de co-autoría, los documentos asociados a las relaciones de co-autoría tienen en común dos o más autores y ninguno de estos documentos tienen al mismo autor como autor principal del documento (ver figura 4.25).

Validación de relaciones de colaboración

La validación de las relaciones de colaboración consistió en verificar que están presentes en la red temática de colaboración las relaciones entre aquellos autores que se sabe son co-autores a partir de la información de la muestra.

Tomando en cuenta que durante el modelado de la red de colaboración no se consideran las relaciones de co-autoría como una situación condicionante para establecer una relación entre autores (recordar el hecho de que únicamente se toma al autor principal de cada documento para generar la red de colaboración), es relevante verificar el hecho de que autores para los cuales se tiene la certeza de un área de interés afín

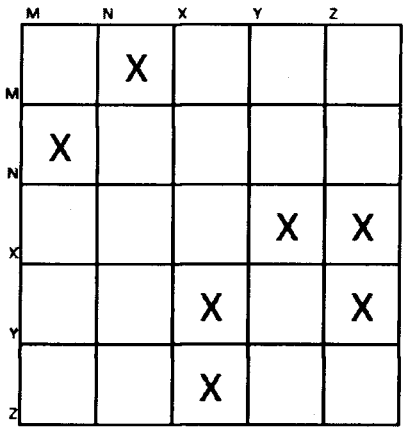


*autor: Es considerado el autor asociado a un documento durante el proceso de inferencia de la red temática de colaboración.

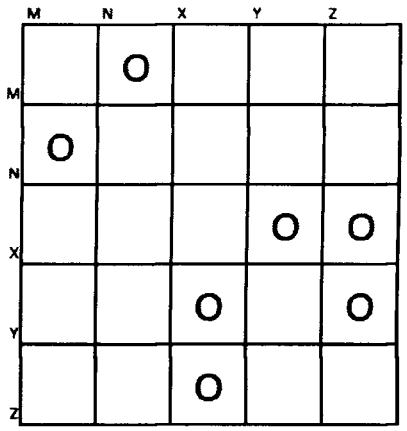
Figura 4.25: Representación de las relaciones de co-autoría identificadas en la muestra para validación de resultados.

(debido a su relación de co-autoría identificada en la muestra) estén presentes en la red temática de colaboración y tengan una relación de colaboración dentro de la red.

La figura 4.26 representa la comparación realizada entre las relaciones previamente conocidas de co-autoría y las relaciones obtenidas durante el proceso de inferencia de la red temática de colaboración. Donde el objetivo es encontrar una coincidencia entre las relaciones de la muestra y las relaciones que se obtienen a través de la red temática de colaboración.



Relaciones identificadas en la muestra.



Relaciones identificadas en la red temática de colaboración

Figura 4.26: Comparación entre las relaciones de co-autoría de la muestra y las relaciones de la red temática de colaboración.

Como resultado del experimento se obtuvieron 119 coincidencias entre las relaciones de co-autoría identificadas en la muestra y las relaciones identificadas en la red

temática de colaboración. En 8 casos no se obtuvo una coincidencia debido a que los documentos asociados a dichos autores quedaron en la raíz del árbol de agrupamiento jerárquico y dichos autores son descartados de la red temática de colaboración debido a que sus documentos no tuvieron las características para ser agrupados en ninguno de los 5 grupos obtenidos a partir del total de documentos de la muestra (la figura 4.27corresponde al árbol de agrupamiento que se obtuvo).

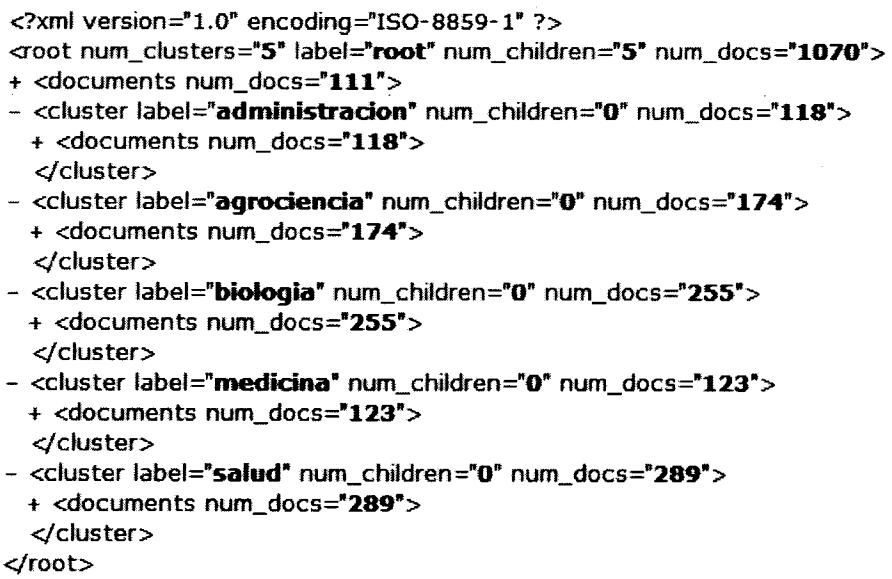


Figura 4.27: Árbol de agrupamiento jerárquico obtenido a partir de la muestra de 1070 documentos.

Para todos los casos donde previamente existía una relación de co-autoría y dado que los documentos de dichos autores fueran agrupados en una de las temáticas del árbol de agrupamiento, puede concluirse que se logra validar la existencia de relaciones de colaboración en la red temática.

Validación de la fortaleza de afinidad

La validación de la fortaleza de afinidad consistió en verificar que existe una dependencia directa entre la fortaleza de afinidad y aquellos autores que se sabe son co-autores a partir de la información de la muestra, de manera que los autores involucrados en una relación de co-autoría tengan una mayor fortaleza de afinidad.

A partir de los 106 autores involucrados en relaciones de co-autoría identificados en la muestra se obtuvo una tabla con las características mostradas en figura 4.28, la primera columna corresponde a todas las posibilidades de relación (combinaciones) entre los 106 autores, la segunda columna corresponde a la fortaleza de afinidad (%) y

la tercera columna indica si tienen (asignando un valor de 1) o no (asignando un valor de 0) relación de co-autoría previamente identificada a partir de la muestra.

En base a las posibles combinaciones entre los 106 autores involucrados en relaciones de co-autoría se obtuvo la grafica mostrada en la figura 4.29 en donde

- El eje de las abscisas corresponde al valor de la afinidad (x = % de fortaleza de afinidad).
- El eje de las ordenadas corresponde a la existencia de una relación de co-autoría ($y = 1$) o su ausencia ($y = 0$).

Relaciones	Fortaleza	Co-autoría
A1 , A2	%	1
A1 , A3	%	0
A1 , A4	%	1
.....	...	...
A1 , A106	%	0
A2 , A1	%	1
A2 , A3	%	0
.....	...	...
A106, A104	%	0
A106, A105	%	1

Figura 4.28: Representación de la tabla de fortaleza de afinidad entre los autores con co-autoría.

En relación a los resultados observados en la grafica se obtuvo la siguiente información:

- El promedio de la fortaleza de afinidad de autores con co-autoría (promedio de X con $Y = 1$) fue de 0,58.
- El promedio de la fortaleza de afinidad de autores de un mismo grupo y sin relación de coautoría ($Y = 0$) fue de 0,34.
- El promedio de fortaleza de afinidad de autores de distintos grupos y sin relación de co-autoría ($Y = 0$) fue de 0,19.

Cabe aclarar que lo anterior no contradice el porcentaje de soporte grupo, uno de los parámetros utilizados en la aplicación FIHC 1.0 al que se le asigna un valor del 0,25. El soporte de grupo es el porcentaje mínimo de afinidad que debe tener un documento

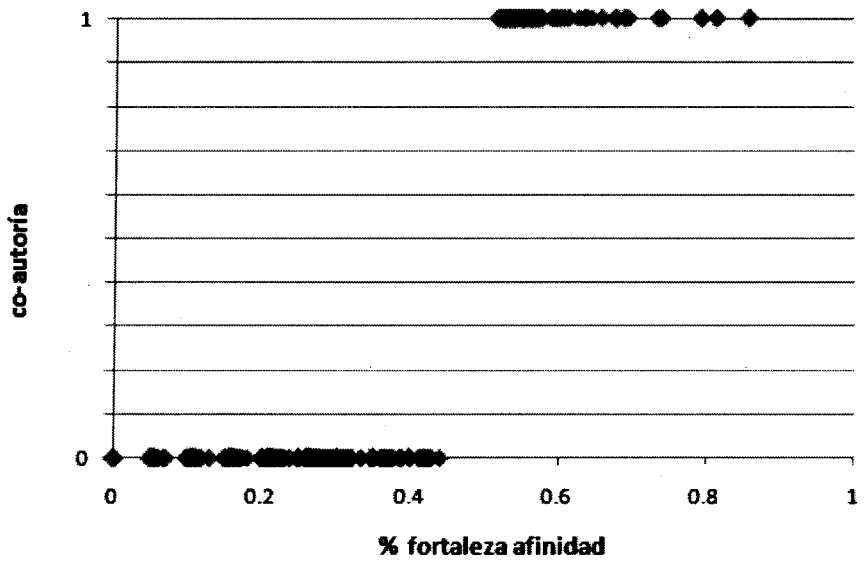


Figura 4.29: Grafica de dispersión de la fortaleza de afinidad correspondiente a los autores identificados en la muestra.

en relación con los otros documentos de un grupo en particular para ser asignado a dicho grupo.

Para validar que el porcentaje de fortaleza de afinidad se ve afectado directamente por la relación de co-autoría (conocida a partir de la muestra) se aplico un análisis de correlación obteniendo un porcentaje de 0,31. Por lo tanto se comprueba la existencia de una dependencia directa entre la fortaleza de afinidad y los autores que a partir de la información de la muestra se sabe que son co-autores, dependencia que permite obtener un mayor porcentaje de afinidad en la relación entre co-autores.

Validación de la fortaleza de Colaboración

La validación de la fortaleza de colaboración consistió en verificar que existe una dependencia directa entre la fortaleza de colaboración y aquellos autores con mayor presencia en el agrupamiento (validando la adaptación de la métrica de exclusividad que aumenta la probabilidad de colaboración si la relación es con algún autor que tenga más de un documento en el grupo), de tal forma que las relaciones con autores con una mayor presencia en el grupo tengan una mayor fortaleza de colaboración.

A partir de los 106 autores involucrados en relaciones de co-autoría identificados en la muestra se obtuvo una tabla con las características mostradas en figura 4.30, la primera columna corresponde a todas las posibilidades de relación (combinaciones) entre los 106 autores, la segunda columna corresponde a la fortaleza de colaboración (%) y la tercera columna indica si la relación es con un autor asociado con más de un

documento en el grupo (asignando un valor de 1) o no (asignando un valor de 0).

Relaciones	Fortaleza	Presencia >1
A1 , A2	%	1
A1 , A3	%	0
A1 , A4	%	1
.....	...	...
A1 , A106	%	0
A2 , A1	%	1
A2 , A3	%	0
.....	...	...
A106, A104	%	0
A106, A105	%	1

Figura 4.30: Representación de la tabla de fortaleza de colaboración entre los autores de la muestra.

En base a las posibles combinaciones entre los 106 autores y su presencia en el agrupamiento se obtuvo la grafica mostrada en la figura 4.31 en donde:

- El eje de las abscisas corresponde al valor de la colaboración (x = % fortaleza de colaboración) entre los autores identificados en la muestra.
- El eje de las ordenadas indica la relación con un autor asociado a más de un documento en el grupo ($y = 1$) o no ($y = 0$).

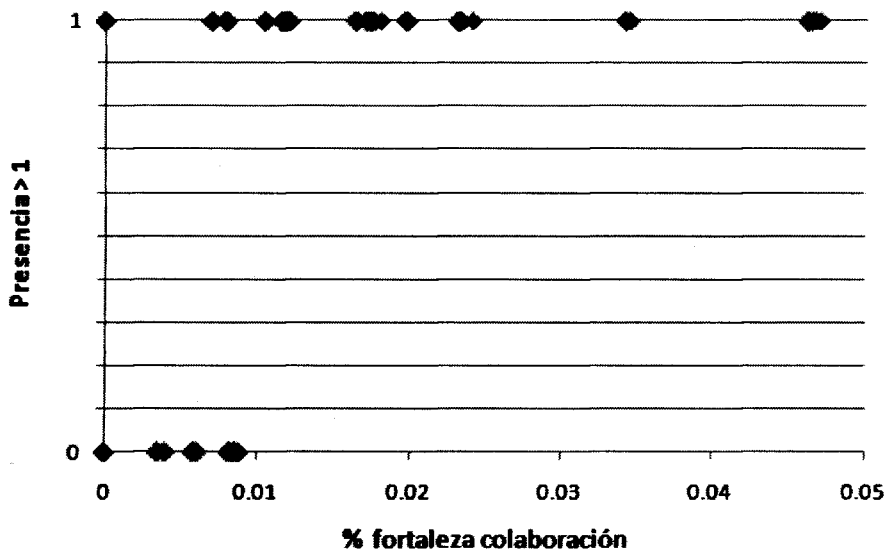


Figura 4.31: Grafica de dispersión de la fortaleza de colaboración correspondiente a los autores identificados en la muestra.

Para validar que el porcentaje de fortaleza de colaboración se ve afectada directamente por la relación con un autor con mayor presencia (asociado a más de un documento) en el grupo se aplico un análisis de correlación obteniendo un porcentaje de 0,19. Por lo tanto se comprueba la existencia de una dependencia directa entre la fortaleza de colaboración y los autores con mayor presencia en un grupo, dependencia que permite obtener un mayor porcentaje de colaboración en las relaciones con autores con más presencia en el grupo.

4.3. Resumen

En este capítulo se describe con respecto a la implementación y resultados obtenidos a través de la herramienta para inferencia de redes temáticas de colaboración.

En relación a la implementación, se hicieron evidentes características y situaciones las cuales son documentadas en relación a la etapa de la metodología (ver sección 3.1.) en las que se presentaron:

- Durante la recolección, pre-procesamiento e identificación de la información relevante (etapa 1) una situación relevante es la estandarización realizada sobre el contenido de los metadatos dc:creator y dc:publisher, debido a las diferencias encontradas entre los registros procedentes de repositorios distintos (Phronesis, CIRIA y Redalyc), ya que si bien el uso de los metadatos Dublin Core es un estándar, la información contenida en ellos no lo es. De igual forma, fueron necesarias crear reglas que permitan determinar la unicidad para los atores, instituciones y documentos procedentes de los registros cosechados.

- Durante el agrupamiento jerárquico de registros cosechados (etapa 2) fue utilizada la versión FIHC 1.0 del algoritmo de agrupamiento jerárquico basado en la frecuencia de un conjunto de elementos disponible en [24]. FIHC 1.0 provee de un mecanismo de pre-procesamiento para los documentos que se deseen agrupar, el cual fue eliminado debido a que los archivos de entrada procedentes de la etapa 1 ya cumplen con las características de pre-procesamiento (ver sección 3.1.1) requeridas para agrupar los documentos mediante el algoritmo FIHC, y también porque el pre-procesamiento del algoritmo FIHC 1.0 fue desarrollado para documentos en ingles. También se hace mención de los parámetros de entrada y los valores asignados a dichos parámetros para la ejecución de la aplicación FIHC 1.0.

- Durante la generación de las redes temáticas de colaboración (etapa 3) se definen las estructuras de datos involucradas en el cálculo de los conceptos: Exclusividad, Frecuencia de afinidad temática y Pesos normalizados. Se describe también el proceso de identificación de las entidades de colaboración en base a la información

procedente de la base de datos y del agrupamiento jerárquico, correspondiente a las etapas 1 y 2 respectivamente. Se especifica el origen de los datos involucrados en la creación de los archivos XML que permiten representar las redes temáticas de colaboración.

En relación a los resultados, de manera análoga a los detalles de implementación se muestran en relación a la etapa de la metodología en la que se obtienen:

- Durante la recolección, pre-procesamiento e identificación de la información relevante (etapa 1) se identificaron 114,912 registros provenientes de las bibliotecas digitales de Phronesis, CIRIA y Redalyc. Al momento de realizar el almacenamiento y como resultado de la aplicación de la regla de unicidad de documentos son descartados 860 registros y únicamente se conservan 114,052. La necesidad de aplicación la regla de unicidad institución se hacen evidentes al encontrar en la base de datos 823 publicados por el Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM), siendo que en el directorio de Phronesis únicamente existen 404 documentos (todos publicados por el ITESM). La necesidad de la regla de unicidad de autor se hacen evidentes al encontrar en la base de datos 728 autores por el Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM), siendo que en el directorio de Phronesis únicamente existen 383 autores. En relación al número de archivos de entrada para el agrupamiento jerárquico (etapa 2) , se obtienen un total de 114,912 archivos.

- Durante el agrupamiento jerárquico de registros cosechados (etapa 2) el resultado es un árbol de agrupamiento con un total de 19 grupos cuyas etiquetas de agrupamiento y detalles se muestran en la tabla 4.8.

- Durante la generación de las redes temáticas de colaboración (etapa 3) se obtienen los archivos XML correspondientes a la red temática de colaboración a nivel de instituciones y a la red temática de colaboración a nivel de autores. Los detalles de la información contenida a nivel de instituciones y a nivel de autores se muestran en las tablas 4.9 y 4.10, respectivamente.

En el siguiente capítulo se presentan las conclusiones derivadas de este trabajo, así como también se presentan algunas sugerencias para extender el trabajo presentado en este documento.

Capítulo 5

Conclusiones y Trabajo futuro.

En este capítulo se presentan las conclusiones derivadas de este trabajo, así como también se presentan algunas sugerencias para extender el trabajo presentado en este documento.

5.1. Conclusiones

La investigación realizada en esta tesis se enfocó en demostrar la factibilidad de crear una estructura de colaboración de contenido científico para una comunidad sin la necesidad de un contacto previo entre sus miembros. Determinando dicha estructura de colaboración en base a temas afines entre los miembros de la comunidad, específicamente, en base a la afinidad de contenido en los documentos digitales de la comunidad.

Tomando como base la comunidad formada por las instituciones y autores procedentes de las bibliotecas digitales de Phronesis, CIRIA y Redalyc se logró implementar el modelo de solución propuesto en la sección 3.1, dando como resultado una herramienta capaz de inferir a partir de documentos digitales, las redes temáticas de colaboración a nivel de instituciones y a nivel de autores, respectivamente.

Otras aportaciones obtenidas a través del modelo de solución propuesto son:

- El identificar la afinidad temática entre documentos pertenecientes a colecciones digitales a partir de la similitud de su contenido, mediante el uso del algoritmo de agrupamiento jerárquico basado en la frecuencia de un conjunto de elementos (FIHC).

- Se logró la adaptación y adecuación de uso de las métricas originalmente usadas en las redes de co-autoría, a través de las cuales se obtiene un grafo dirigido con pesos normalizados cuyos enlaces representen las relaciones de afinidad temática entre los miembros de la comunidad.

Con base a los resultados obtenidos es posible concluir que el modelo de solución propuesto en la sección 3.1 es una alternativa viable para la inferencia de redes

temáticas de colaboración en bibliotecas digitales. Sin embargo, debido a la complejidad polinómica ($O(n^2)$) del algoritmo mostrado en la figura 4.14 el tiempo de ejecución puede volverse intratable (resolverse pero no lo suficientemente rápido para que la solución sea útil [20]) dependiendo del número de *entidades de colaboración* (n) a relacionar.

5.2. Trabajo futuro

A continuación se presentan algunas ideas sobre las cuales se podría extender el trabajo realizado en esta tesis:

- El trabajo presentado en esta tesis puede ser aplicado en otras áreas o comunidades. Por ejemplo, en el área de la salud podría servir para encontrar relaciones entre instituciones o países con problemas sanitarios en común, y a partir del grado (fuerza) de afinidad de su problemática puedan colaborar en propuestas o estrategias para abatir su problema. Habría que comenzar considerando aspectos como: contar con los repositorios con la información necesaria, la estandarización de la información de dichos repositorios y cuestiones de acceso a la información.

- Otra de las áreas donde es posible extender el presente trabajo en relación a la complejidad del algoritmo (ver figura 4.14) usado para realizar el cálculo de las métricas de colaboración. La complejidad actual del algoritmo es $O(n^2)$, sin embargo, contando con la infraestructura necesaria (un cluster de computadoras) y mediante el uso de técnicas de computo paralelo los esfuerzos se pueden dirigir en implementar un algoritmo con mayor escalabilidad con respecto al número de entidades de colaboración.

Bibliografía

[1] N. Adam, R. Holowczak, M. Halem, N. Lal, and Y. Yesha. Digital library task force. *Computer*, 29(8):89–91, 1996.

[2] N. Andrews and E. Fox. Recent developments in document clustering. Technical report, Citeseer, 2007.

[3] A. Barabasi and R. Crandall. Linked: The new science of networks. *American journal of Physics*, 71:409, 2003.

[4] C. L. Borgman. Challenges in building digital libraries for the 21(st) century. In E. P. Lim, S. Foo, C. Khoo, S. Urs, T. Costantino, E. Fox, and H. Chen, editors, *5th International Conference on Asian Digital Libraries (ICADL 2002)*, volume 2555, pages 1–13, Singapore, Singapore, 2002. Springer-Verlag Berlin. 29 Berlin BW29S.

[5] P. Duguid and D. Atkins. Report of the Santa Fe planning workshop on distributed knowledge work environments: Digital libraries. *URL: http://www. si. umich. edu/SantaFe*, 1997.

[6] H. Eriksson. The semantic-document approach to combining documents and ontologies. *International Journal of Human-Computer Studies*, 65(7):624 – 639, 2007. Knowledge representation with ontologies: Present challenges - Future possibilities.

[7] W. Frakes and R. Baeza-Yates. *Information retrieval: data structures and algorithms*. Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1992.

[8] B. Fung, K. Wang, and M. Ester. Hierarchical document clustering. *The Encyclopedia of Data Warehousing and Mining, John Wang (ed.), Idea Group*, 2005.

[9] B. C. M. Fung, K. Wang, and M. Ester. Hierarchical document clustering using frequent itemsets. In D. Barbara and C. Kamath, editors, *3rd SIAM International Conference on Data Mining*, pages 59–70, San Francisco, Ca, 2003. Siam. 19 Philadelphia BX26M.

[10] L. Iverson. Collaboration in digital libraries: A conceptual framework. In H. Chen, M. Christel, and E. P. Lim, editors, *4th Joint Conference on Digital Libraries*, pages 380–380, Tucson, AZ, 2004. Assoc Computing Machinery. 2 New York BAM92.

[11] A. K. Jain, M.Ñ. Murty, and P. J. Flynn. Data clustering: A review. *Acm Computing Surveys*, 31(3):264–323, 1999. 204 Assoc Computing Machinery New York 302KZ.

[12] C. Lagoze and H. Van de Sompel. The Open Archives Initiative: Building a low-barrier interoperability framework. In *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, pages 54–62. ACM New York, NY, USA, 2001.

[13] X. M. Liu, J. Bollen, M. L. Nelson, and H. Van de Sompel. Co-authorship networks in the digital library research community. *Information Processing and Management*, 41(6):1462–1480, 2005. 36 Pergamon-Elsevier Science Ltd Oxford 956XE.

[14] M. Matteucci. A tutorial on clustering algorithms. *http://home.dei.polimi.it/matteucc/Clustering/tutorialhtml/index.html.*, Retrieved on May 10, 2010.

[15] M. A. Medina and J. A. Sanchez. Ontoair: a method to construct lightweight ontologies from document collections. *Ninth Mexican International Conference on Computer Science, Proceedings*, pages 115–125, 2008. Gelbukh, A Adiba, M 9th Mexican International Conference on Computer Science OCT 06-10, 2008 Mexicali, México.

[16] E. Otte and R. Rousseau. Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science*, 28(6):441, 2002.

[17] E. Rasmussen. Information Retrieval: Data Structures and Algorithms, chapter Clustering algorithms, 1991.

[18] B. Ricardo, R. Berthier, et al. Modern information retrieval. *England: Pearson Education Limited*, 1999.

[19] J. Scott. Social network analysis: A handbook . Thousands Oaks. *Cal.: SAGE Publications*, 2000.

[20] S. Smale. Complexity theory and numerical analysis. *Acta Numerica*, 6:523–551, 1997.

[21] J. A. Sánchez, M. G. Quintana, and A. Razo. Star-fish: Starfields+fisheye visualization and its application to federated digital libraries. *CLIHC 2007*, pages 1–11,

2007. 3rd Latin American Conference on Human-Computer Interaction, 2007, Rio de Janeiro, Brazil.

[22] H. Sompel and C. Lagoze. Notes from the interoperability front: a progress report on the open archives initiative. In *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, page 157. Springer-Verlag, 2002.

[23] B. Tjaden, P. Reynolds, et al. The oracle of bacon. Technical report, Technical report, 2003.

[24] S. F. University. FIHC 1.0: Frequent Itemset-based Hierarchical Clustering. *URL: http://ddm.cs.sfu.ca/software.html*, 2007.

[25] J. Voutssás. *Bibliotecas y publicaciones digitales.* Unam, 2006.

[26] S. Wasserman and K. Faust. *Social network analysis: Methods and applications.* Cambridge Univ Pr, 1994.

[27] Y. Zhao and G. Karypis. Evaluation of hierarchical clustering algorithms for document datasets. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 515–524. ACM New York, NY, USA, 2002.