UNIVERSIDAD CENTRAL "MARTA ABREU" DE LAS VILLAS
FACULTAD DE MATEMÁTICA, FÍSICA Y COMPUTACIÓN
DEPARTAMENTO DE CIENCIA DE LA COMPUTACIÓN

Agrupamiento basado en la intermediación diferencial y su valoración utilizando la teoría de los conjuntos aproximados

Tesis presentada en opción al grado científico de Doctor en Ciencias Técnicas

Autor: MSc. Leticia Arco García

Santa Clara, 2008

UNIVERSIDAD CENTRAL "MARTA ABREU" DE LAS VILLAS
FACULTAD DE MATEMÁTICA, FÍSICA Y COMPUTACIÓN
DEPARTAMENTO DE CIENCIA DE LA COMPUTACIÓN

Agrupamiento basado en la intermediación diferencial y su valoración utilizando la teoría de los conjuntos aproximados

Tesis presentada en opción al grado científico de Doctor en Ciencias Técnicas

Autor: MSc. Leticia Arco García

Tutores: Dr. Rafael Esteban Bello Pérez

Dr. Rudolf Kruse

Consultante: Dr. Alberto Ochoa Rodríguez

Santa Clara, 2008

A mis padres y abuelos

Agradezco a todas las personas que han contribuido al desarrollo de este trabajo:

A mis padres, abuelos y familia por todo el apoyo que me han brindado, sin su ejemplo, dedicación y paciencia para soportarme durante años hubiera sido imposible.

Al Dr. Rafael Bello por incentivarme a desarrollar este tema, por su visión y guía.

Al Dr. Alberto Ochoa por sus ideas, su exigencia, tiempo dedicado y por enseñarme tanto.

Al Prof. Jorge Marx Gómez por haber facilitado el intercambio científico con especialistas de reconocido prestigio pertenecientes a universidades alemanas.

Al Prof. Rudolf Kruse por aceptar colaborar con este tema, por sus sugerencias, libros y colecciones facilitados, así como la coordinación de importantes intercambios científicos.

Al Dr. Ramiro Pérez especialmente por la revisión exhaustiva de este trabajo.

Al Dr. Ricardo Grau por sus magníficos aportes y apoyo incondicional.

A todos mis colegas por ayudarme en cada momento que los necesité, por sus sugerencias y colaboración.

A todos los estudiantes que han trabajado conmigo en estos años con tanto entusiasmo y entrega.

A los investigadores del CENATAV por brindar valiosas sugerencias y bibliografía actualizada.

A Isis y a Yailé por estar siempre conmigo en los momentos difíciles.

A mis amigos por alegrarme la vida cada vez que este trabajo me tenía agobiada.

SÍNTESIS

En el agrupamiento sobre grafos existen métodos, basados en las relaciones de objetos, que tienen alto costo computacional porque utilizan medidas que no capturan eficientemente las propiedades topológicas. Además, las medidas de validación del agrupamiento no siempre dan criterios certeros. El objetivo de la investigación es diseñar medidas que capturen eficientemente la información topológica que codifica el problema, así como un método de agrupamiento que las utilice eficientemente, y validar el agrupamiento, utilizando una herramienta matemática que mida de manera no supervisada la calidad, precisión y consistencia de los grupos. Los resultados obtenidos son: la definición Intermediación Diferencial (DB) caracterizada por capturar eficiente y localmente la centralidad de aristas, no negociar valores de intermediación entre puentes paralelos, ser menos sensible al ruido, y comportarse como una medida de disimilitud topológica; el algoritmo para el agrupamiento basado en DB que no requiere el recálculo y tiene buen desempeño en dominios textuales; la aplicación de la Teoría de los Conjuntos Aproximados (RST) para la validación no supervisada y el etiquetamiento de grupos; el conjunto de medidas basadas en RST y el algoritmo para utilizarlas al validar agrupamientos; y los sistemas SATEX y GARLucene para manipular documentos y contribuir a la gestión de información y conocimiento.

Abstract

When clustering over graphs, one can find methods which are based on the interrelationships between the objects and exhibit a high computational cost owing to the use of metrics which do not efficiently capture the underlying topological structures. Furthermore, clustering validation measures do not always provide true criteria. The aim of this research is to design measures which are able to capture in an efficient manner the topological information that codes the problem, along with a clustering method that uses them efficiently. The study also intends to assess the clustering outcome by means of a mathematical tool capable of measuring the quality, accuracy and consistency of every cluster in an unsupervised way. The main results are: the Differential Betweenness (DB) characterized by efficient local catching of the edges centrality, not negotiation of betweenness values between parallel bridges, less susceptibility to noise, and behavior as a topological dissimilarity measure; the DB-based clustering approach, which makes no use of recalculation and achieves a good performance in textual domains; the application of Rough Set Theory (RST) for unsupervised cluster validation and labeling; the set of RST-based metrics along with the associated clustering validity algorithm; and the systems SATEX and GARLucene for document handling, thus contributing to the information and knowledge management.

Tabla de contenidos

INTRODUCCIÓN

La cantidad de información está continuamente creciendo; sin embargo, la habilidad de los humanos para procesarla y asimilarla permanece constante [1, 2]. Además, la información en sí misma tiene pocas ventajas, su sistematización, incorporación y utilización son los elementos que aportan su valor añadido: el conocimiento[1]. Es necesario crear sistemas que generen conocimiento, para asegurar el uso productivo de la información y guiar una toma de decisiones óptima, contribuyendo de esta forma a la Gestión del Conocimiento (GC) [3-6].

Aproximadamente un 80% de la información se almacena en forma textual no estructurada [7], de ésta una cantidad significativa está dada por el crecimiento de las publicaciones científicas en diferentes campos. Por ejemplo, Medline[2] ya tiene más de 10 millones de publicaciones, con un incremento semanal entre 7000 y 8000 resúmenes científicos. Gestionar el conocimiento a partir de la información encontrada es fundamental en el trabajo científico [8]. Sin embargo, la gestión de información científica se vuelve cada vez más compleja y desafiante, sobre todo porque las colecciones textuales generalmente son heterogéneas, grandes, diversas y dinámicas. Superar estos desafíos es esencial para dar a los científicos mejores condiciones de administrar el tiempo necesario para procesar la información científica, lo cual constituye la motivación principal de este trabajo.

Hoy día, tanto las comunidades científicas, organizaciones, así como los gobiernos, invierten en función del desarrollo de la gestión de la información y el conocimiento, a través de proyectos, congresos, postgrados y desarrollo de sistemas con este fin. Algunos ejemplos son: las políticas trazadas por la Unión Europea para incrementar la competitividad de una economía basada en el conocimiento[3], las acciones realizadas por la OPS/OMS en países en desarrollo[4] y las facilidades brindadas por la UNESCO para desarrollar software para el procesamiento de la información[5]. Universidades e institutos científicos desarrollan esta línea de investigación, como el grupo de investigación en la gestión del conocimiento y minería de

[1] Considerado un recurso estratégico para el desarrollo económico, científico y social contemporáneo.
[2] Base de datos de 11 millones de citas y resúmenes de revistas de medicina y otras fuentes. http://medline.cos.com
[3] Comunicación de la comisión de comunidades europeas al parlamento europeo "Información científica en la era digital". Bruselas. 2007. http://ec.europa.eu/research/science-society/document_library/pdf_06/communication-022007_en.pdf
[4] Newsletter BVS 056. 10 de agosto de 2006. http://newsletter.bireme.br/new/index.php?lang=pt&newsletter=20060810
[5] CDS/ISIS de UNESCO http://hrueda-isis.blogspot.com/2007_11_01_archive.html

textos de la Universidad de West Bohemia[6], el grupo de investigación en descubrimiento de conocimiento a partir de datos no estructurados de la Universidad de Texas[7], el centro nacional del Reino Unido para la minería de textos (The Nacional Centre for Text Mining; NaCTeM)[8], el centro europeo para la computación blanda (European Centre for Soft Computing)[9], el grupo de recuperación de información de la Universidad de Magdeburgo Otto-von-Guericke[10] y el Instituto Kaieteur para la gestión del conocimiento (Kaieteur Institute for Knowledge Management[11]). Cuba no está exenta del desarrollo de investigaciones que contribuyan a la gestión de la información y el conocimiento. Ejemplos de ello lo constituyen los trabajos realizados en el Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV) de la Ciudad de La Habana y en el Centro de Reconocimiento de Patrones y Minería de Datos de la Universidad de Oriente en Santiago de Cuba; así como la labor desarrollada por la Asociación Cubana de Reconocimiento de Patrones (ACRP)[12].

El conocimiento se puede gestionar de diversas formas y hacerlo requiere de la integración de varias áreas del saber: el descubrimiento de conocimiento en bases de datos, la minería de datos y de textos. Esta última integra la recuperación y extracción de información, el análisis de textos, el resumen, la categorización, la clasificación, el agrupamiento, la visualización, la tecnología de bases de datos, el aprendizaje automático y la minería de datos [1, 9]. En la actualidad los usuarios se enfrentan a grandes colecciones de información o a grandes cantidades de enlaces web recuperados por un buscador y tienen que ser muy pacientes para navegar a través de los enlaces y analizar la información que verdaderamente necesitan. La categorización, clasificación y agrupamiento se pueden utilizar para refinar resultados de la recuperación y extracción de información, y así contribuir al descubrimiento de conocimiento. Especialmente el agrupamiento y los procesos post-agrupamiento permiten organizar la información, determinar información relevante y crear nuevo conocimiento a partir de la información disponible en una colección especificada o resultante de un proceso de recuperación de información.

---

[6] http://textmining.zcu.cz
[7] http://www.cs.utexas.edu/~ml/publication/text-mining.html
[8] http://www.cs.manchester.ac.uk/research/groups/infosystems/textmining/
[9] http://www.softcomputing.es
[10] http://irgroup.cs.uni-magdeburg.de
[11] http://www.kikm.org/
[12] http://acrp.cenatav.co.cu

El agrupamiento es una tarea del aprendizaje no supervisado que tiene como objetivo descomponer el conjunto de datos, de forma tal que los objetos que pertenecen al mismo grupo sean tan similares como sea posible y los objetos que pertenecen a grupos diferentes sean tan disimilares como sea posible. El análisis de grupos es una herramienta para descubrir una estructura previamente oculta en los datos, asumiendo que existe un agrupamiento natural o cierto en ellos. Sin embargo, la asignación de los objetos a las clases y la descripción de esas clases son desconocidas [10].

El análisis de grupos forma las bases del aprendizaje y del conocimiento. Muchos investigadores consideran su primera aplicación, la realizada por John Snow, al buscar la asociación de los casos enfermos de cólera en Londres en 1854 [11]. Esta tarea puede aplicarse a disímiles campos: en la Biología para detectar genes con funcionalidades similares [12], en la informática para diseñar software donde los módulos tienen alta cohesión interna [13], en los estudios sociales para detectar comunidades de individuos [14, 15], y en la gestión de la información, para agrupar documentos afines [16] y detectar grupos en redes de enlaces web [17], entre otros.

Varios métodos de agrupamiento requieren calcular la similitud entre los objetos acorde a las características del conjunto de datos. Otros asumen un reto importante: descubrir grupos en datos que al relacionarse forman una estructura interesante para el análisis [18]. Así, varios métodos, llamados métodos de detección de comunidades en redes complejas, parten de representar los objetos y sus relaciones en un grafo y explotan su topología para descubrir los grupos. Se presupone en la actualidad que el conocimiento de la estructura de los datos es tan importante como los objetos en sí.

Un enfoque reciente de los métodos que operan sobre representaciones de los objetos en grafos consiste en calcular la centralidad de las aristas en el proceso de agrupamiento. La intermediación ha sido ampliamente utilizada como una forma de medir la centralidad [19-23]. Un autor prolífero que ha seguido este enfoque es M. J. E. Newman [24-28]. Estos métodos en su mayoría son muy costosos computacionalmente y en algunos casos no logran medir adecuadamente la centralidad de las aristas.

Los métodos basados en el cálculo de la intermediación de las aristas requieren de una fase de validación de los agrupamientos en la cual suele usarse una medida de calidad conocida como

modularidad [29]. Ante la presencia de la alta complejidad computacional del cálculo de la intermediación y de la necesidad del recálculo en el proceso del agrupamiento, en una segunda etapa del desarrollo de estos métodos, los autores prefieren optimizar directamente la modularidad y abandonan el uso de las medidas de centralidad aplicadas en la primera etapa [30].

No sólo es necesario obtener un buen agrupamiento, el post-agrupamiento es indispensable para caracterizar los grupos obtenidos, sobre todo cuando el agrupamiento ayuda a la gestión de la información. La valoración de los grupos obtenidos se puede realizar mediante su validación, etiquetamiento o alguna forma de caracterización. Los investigadores le han dedicado menor interés al post-agrupamiento, una de las causas es que aún quedan problemas sin resolver en el proceso de agrupamiento.

Se utiliza frecuentemente la clasificación de las medidas de validación del agrupamiento en internas, externas y relativas [31, 32]. En situaciones del mundo real donde no se tiene conocimiento del dominio de aplicación, como es el caso de algunas aplicaciones textuales, es necesario utilizar medidas internas [29, 33-40]. Cada medida interna existente logra captar algunas de las propiedades deseadas de un agrupamiento, aplicar varias medidas permite llegar a mejores conclusiones del agrupamiento que se desea validar. El etiquetamiento de los grupos resultantes se investiga en la actualidad como otra acción post-agrupamiento [41-47]. Las variantes para etiquetar generalmente se basan en heurísticas, requieren definición de umbrales y están influenciadas por elementos subjetivos.

La evaluación y otras tareas post-agrupamiento poseen incertidumbre y vaguedad, sobre todo en dominios textuales. Varios autores han manifestado que la Teoría de los Conjuntos Aproximados (Rough Set Theory; RST) es una buena herramienta para modelar la incertidumbre cuando ésta se manifiesta en forma de inconsistencia [48-51]. Dos ventajas de RST se pueden utilizar para realizar una validación no supervisada del agrupamiento: no requiere información adicional o preliminar sobre el conjunto de datos, ni suposición sobre éstos, y se usa en circunstancias caracterizadas por vaguedad e incertidumbre.

El análisis histórico-lógico en torno al agrupamiento y el post-agrupamiento conduce a inferir que aún persisten insuficiencias al agrupar automáticamente objetos y brindar una valoración útil de los resultados. Todo ello constituye una problemática a la cual aún la ciencia no ha

dado respuestas definitivas, lo cual justifica el planteamiento del **problema de investigación** siguiente:

Por un lado, existen métodos que descubren grupos de objetos, basados en las relaciones de éstos representados en un grafo, que tienen alto costo computacional porque utilizan mediciones de la centralidad que no capturan eficientemente las propiedades topológicas que codifican la estructura del problema. Por otro, las medidas existentes para validar el agrupamiento no siempre dan criterios certeros, principalmente en dominios textuales donde frecuentemente se carece de la clasificación de referencia.

El **objetivo general** de la investigación consiste en diseñar medidas que capturen eficientemente la información topológica que codifica la estructura del problema, así como un nuevo método que las utilice eficientemente para el agrupamiento, en particular en dominios textuales, y validar los grupos obtenidos, mediante el uso de una herramienta matemática que permita medir la calidad, precisión, posibles inconsistencias y manejar incertidumbre, sin requerir información adicional de los datos.

Este se desglosa en los siguientes **objetivos específicos**:

1. Diseñar medidas que capturen eficientemente la información topológica que codifica la estructura del problema y un método que las utilice eficientemente en el agrupamiento de objetos.
2. Validar los grupos a través del uso de RST permitiendo medir la precisión, calidad y consistencia del agrupamiento sin requerir conocimiento del dominio y caracterizar los grupos mediante sus objetos más representativos y relacionados.
3. Aplicar los métodos diseñados para el agrupamiento y la valoración[13] de los grupos en la manipulación de documentos contribuyendo a gestionar información y conocimiento.

Las **preguntas de investigación** planteadas son:

¿Cómo medir adecuadamente la intermediación de las aristas de forma que se capture

[13] A pesar de que existen autores que diferencian los términos valoración, evaluación y validación, en esta tesis se utilizan indistintamente los términos validar y evaluar. El término valorar es aquí utilizado para referirse a procesos post-agrupamiento, tales como validar, etiquetar y fusionar.

eficientemente la información topológica de los grafos que representan el problema y cómo utilizarla en un algoritmo de agrupamiento?

¿Cómo aplicar y qué extensiones realizar a RST, para utilizarla en la validación del agrupamiento?

¿Qué consideraciones se requieren para aplicar el algoritmo de agrupamiento y la valoración de los grupos en la manipulación de documentos?

Después de haber realizado el marco teórico se formularon las siguientes **hipótesis de investigación** como presuntas respuestas a las preguntas de investigación:

H1: Una medición de la centralidad de las aristas que capture eficientemente la información topológica que describe el problema contribuye al desarrollo de métodos de agrupamiento que no requieran la optimización directa de las medidas de calidad ni el recálculo.

H2: La Teoría de los Conjuntos Aproximados es útil para realizar la validación no supervisada mediante el cálculo local y global de la calidad, precisión y consistencia de los resultados de agrupamientos, así como para caracterizar los grupos como parte del proceso post-agrupamiento. A ello puede ayudar la definición de nuevos algoritmos y medidas que enriquecen las posibilidades de la teoría para estos análisis.

H3. Esos procedimientos, formulados en general, tienen posibilidades efectivas de aplicación en los agrupamientos y la manipulación de documentos textuales.

Para lograr los objetivos trazados y demostrar las hipótesis establecidas se acometieron las **tareas de investigación** siguientes:

- Análisis de los métodos de agrupamiento, medidas de validación de sus resultados y principales estrategias para realizar post-agrupamiento, así como la implementación de algunos métodos encontrados en la revisión bibliográfica.
- Definición de nuevas expresiones que permitan medir la centralidad de los enlaces entre objetos mediante un uso adecuado de las propiedades topológicas obtenidas por las interrelaciones entre éstos.
- Diseño de un método de agrupamiento que permita descubrir grupos mediante el uso eficiente de las expresiones que miden la intermediación de aristas.
- Comparación del método de agrupamiento propuesto y la medición de la intermediación

de las aristas con métodos que lo antecedieron y con métodos ampliamente utilizados en el agrupamiento de documentos.

- Definición mediante el uso de RST de las medidas de calidad y precisión de los agrupamientos, y pertenencia aproximada de los objetos a grupos.
- Diseño del algoritmo que permita aplicar RST a la validación no supervisada de los resultados de agrupamientos y utilización de las aproximaciones inferiores y superiores obtenidas de éste para caracterizar los grupos.
- Evaluación de los conjuntos aproximados como instrumentos para la validación de los resultados de agrupamientos, particularizando en dominios textuales.
- Diseño de la aplicación de los resultados alcanzados en el agrupamiento y post-agrupamiento en la manipulación de documentos y su implementación mediante los sistemas SATEX y GARLucene.

Entre los **métodos de trabajo científico** utilizados se destacan los siguientes:

Métodos generales. El método hipotético-deductivo para elaborar las hipótesis de investigación y proponer líneas de trabajo a partir de resultados parciales; el método sistémico para el desarrollo de los sistemas computacionales y lograr que los elementos que formen parte de la aplicación real sean un todo que funcione de manera armónica; el método histórico-lógico y el dialéctico para el estudio crítico de los trabajos anteriores, y para utilizar éstos como punto de referencia y comparación de los resultados alcanzados.

Métodos lógicos. El método analítico-sintético al descomponer el problema de investigación en elementos por separado y profundizar en el estudio de cada uno de ellos, para luego sintetizarlos en la solución de la propuesta; el método inducción-deducción como vía de la constatación teórica durante el desarrollo de la tesis; el método de modelación para el desarrollo de los algoritmos.

Métodos empíricos. El método coloquial para la presentación y discusión de los resultados en sesiones científicas; el método experimental para comprobar la utilidad de los resultados obtenidos y la comparación con otros métodos reportados.

Métodos matemáticos. Los métodos de experto para la validación del sistema de gestión de la información; métodos estadísticos para evaluar los conjuntos aproximados como instrumento

de medición.

La **novedad científica** de la investigación radica en:

- La nueva propuesta de medición de la intermediación diferencial de las aristas en su localidad, tanto para grafos no ponderados como ponderados, y el nuevo método que la utiliza eficientemente para el agrupamiento sobre grafos.
- La valoración de los grupos resultantes de un proceso de agrupamiento mediante el uso de RST: algoritmo para validar agrupamientos a partir de la aplicación y definición de medidas de precisión y calidad; y la determinación de los objetos más representativos y relacionados por grupos usando las aproximaciones inferiores y superiores.
- La aplicación de los resultados teóricos de alcance general del agrupamiento y valoración de los grupos en dominios textuales a través de nuevos sistemas elaborados por la autora para gestionar documentos, denominados SATEX y GARLucene.

El **valor práctico** del trabajo está dado por:

- Disponer de un algoritmo de agrupamiento, medidas de validación y el algoritmo para aplicarlas, así como la forma para etiquetar los grupos, que integralmente permiten el agrupamiento y post-agrupamiento de colecciones de objetos de diversos dominios de aplicación con mayor calidad.
- Utilizar los resultados teóricos del agrupamiento y el post-agrupamiento de forma automatizada hace su aplicación más fácil y cómoda para los investigadores de esta área. Se incorporaron al módulo de validación del agrupamiento de Weka, las medidas clásicas de validación interna y externa y la propuesta usando RST; permitiendo estudios comparativos en investigaciones de la Ciencia de la Computación por su grado de generalidad. Además, el desarrollo de un paquete de funciones en Matlab que permite la experimentación con diversas formas del cálculo de la centralidad de las aristas y vértices en grafos ponderados y no ponderados usando información local y global.
- Aplicar los resultados teóricos en dominios textuales mediante el sistema CorpusMiner, elaborado anteriormente por la autora, que permite la representación, agrupamiento y valoración de colecciones textuales. Además se ilustra como preprocesar conjuntos de documentos con TextLynx, otro sistema realizado en la génesis de esta investigación.

- Manipular documentos mediante los nuevos sistemas SATEX y GARLucene que disponen de los métodos y las medidas que permiten el agrupamiento y el post-agrupamiento. Esta forma de gestión permite mostrar organizadamente y caracterizar colecciones textuales previamente establecidas o resultantes de procesos de recuperación de información.

Estos sistemas facilitan a los investigadores y docentes comenzar una revisión del estado del arte, organizar materiales por equipos de estudiantes para la docencia, organizar por temáticas los artículos que le han llegado al comité científico del programa de un evento, así como tener una idea de las asociaciones que existen entre los documentos recuperados.

La **tesis** está **estructurada** en cuatro capítulos. En el Capítulo 1 se trata el agrupamiento y las principales técnicas para validarlo. Se enfatiza en el análisis de los métodos basados en la centralidad de las aristas y se mencionan las posibilidades y limitantes de varias medidas de validación internas y externas. En los capítulos 2 y 3 se presentan los resultados teóricos alcanzados en esta investigación y su evaluación. En el capítulo 2 se define la medida de intermediación diferencial de las aristas calculada localmente y el método de agrupamiento que la utiliza eficientemente; mientras que en el capítulo 3 se presenta el uso de RST para la valoración (validación y etiquetamiento) de los grupos resultantes de agrupamientos. En el Capítulo 4 se aplican los resultados teóricos de la investigación en dominios textuales. Se presentan los sistemas SATEX y GARLucene para la manipulación de documentos, contribuyendo a la gestión de la información mediante el agrupamiento y post-agrupamiento. Este documento culmina con las conclusiones, recomendaciones, referencias bibliográficas, producción científica de la autora sobre el tema de la tesis y los anexos.

1 Acerca de métodos de agrupamiento y medidas de validación

El volumen de datos es cada día mayor; una de sus causas es el crecimiento exponencial de las colecciones de datos no estructurados; por ejemplo, textuales. Por tanto, es fundamental el desarrollo de técnicas que permitan el análisis exploratorio de los datos. El análisis de grupos[14] permite descubrir la estructura interna de éstos e identificar distribuciones interesantes y patrones subyacentes en ellos, considerando muy poca o ninguna información a priori [52]. Dos tareas estrechamente relacionadas con los algoritmos de agrupamiento son la validación y el etiquetamiento de los grupos encontrados. Éstas permiten conocer con qué grado de certeza los grupos fueron obtenidos y cómo es posible caracterizarlos. A continuación se citarán las principales técnicas de agrupamiento, se particularizará en los algoritmos que parten de una representación en grafos de los objetos y sus relaciones, específicamente en aquellos métodos para la detección de comunidades en redes complejas basados en el cálculo de la centralidad de las aristas. Finalmente, se mencionarán algunas clasificaciones de las medidas de validación y se discutirán las ventajas y desventajas de medidas de validación internas y externas.

1.1 Agrupamiento

El análisis de grupos es descrito como una herramienta para el descubrimiento porque tiene la potencialidad de revelar relaciones basadas en datos complejos no detectadas previamente. Los algoritmos de agrupamiento son usados para encontrar una estructura de grupos que se ajuste al conjunto de datos, logrando homogeneidad dentro de los grupos y heterogeneidad entre ellos [53].

Debe existir un alto grado de asociación entre los objetos de un mismo grupo y un bajo grado entre los miembros de grupos diferentes [53]. Cuando el agrupamiento se basa en la similitud de los objetos, se desea que los objetos que pertenecen al mismo grupo sean tan similares como se pueda y los objetos que pertenecen a grupos diferentes sean tan diferentes como sea

[14] En esta tesis se utilizarán indistintamente los términos: grupos, conglomerados, comunidades, subconjuntos y clases

posible [10, 54]. En otras palabras, seguir el principio de maximizar la similitud dentro del grupo y minimizar la similitud entre los grupos.

El concepto de "similitud" tiene que ser especificado acorde a los datos. En la mayoría de los casos los datos son vectores de valores reales, entonces se requieren algunas medidas (distancias, similitudes, o disimilitudes) para cuantificar el grado de asociación entre ellos. Las medidas más usadas para comparar objetos se muestran en el Anexo 2.

Por otra parte, un reto para la minería de datos es descubrir grupos en datos que al relacionarse forman una estructura interesante para el análisis. Este tipo de datos tiene una mejor descripción cuando se representa como una colección de objetos interrelacionados y enlazados [18]. El enlace entre objetos es un conocimiento que puede explotarse en el agrupamiento, ya que rasgos de objetos enlazados están correlacionados, y es probable la existencia de enlaces entre objetos que tienen elementos comunes. Así, varios métodos parten de representar los objetos y sus relaciones en un grafo y explotan su topología para descubrir los grupos. Estas propuestas ven el conjunto de datos desde la perspectiva de las conexiones entre los objetos más que los objetos en sí mismos [55]. Los conjuntos de datos pueden intrínsicamente formar un grafo o se pueden obtener grafos de similitud a partir de la matriz de similitud entre los objetos. En el Anexo 2 se muestran las definiciones y notaciones que se asumen en esta tesis respecto a la Teoría de Grafos.

En la actualidad se presupone que el conocimiento de la estructura de los datos es tan importante como los objetos en sí. Ese conocimiento puede ayudar a descubrir grupos que se ocultan en las comunicaciones entre los objetos [18, 56, 57]. Propiedades de los grafos, sobre todo cuando éstos representan redes complejas, pueden ser indicadores importantes para el agrupamiento. En el Anexo 4 se muestran algunas de estas propiedades. Ellas, en su mayoría, son computacionalmente difíciles de verificar, de ahí que muchos algoritmos sobre grafos tengan una complejidad temporal exponencial [19].

1.1.1 Clasificación de las técnicas de agrupamiento y sus principales algoritmos

Los métodos de agrupamiento se clasifican siguiendo varios criterios: tipo de los datos de entrada del algoritmo, criterios para definir la similitud entre los objetos, conceptos en los cuales se basa el análisis y forma de representación de los datos, entre otros [58]. Una

clasificación general distingue dos tipos: aquellos basados en una función objetivo y los jerárquicos [59, 60].

Esta primera categoría más general se refiere a la construcción de particiones (grupos) del conjunto de datos sobre la base del perfeccionamiento de algún índice, conocido también como función objetivo. En esencia, dividir n objetos en un número positivo k de grupos, generalmente especificado a priori. El objetivo de estos métodos es encontrar la mejor división de los datos en k grupos basada en una medida de similitud dada y conservar el espacio de particiones posibles en k subconjuntos solamente. La mayoría de los algoritmos que siguen esta técnica son esencialmente basados en prototipos, comienzan con una partición inicial, usualmente aleatoria, y proceden con su refinamiento [10].

Uno de los algoritmos que sigue esta primera categoría y que ha sido ampliamente utilizado es el *k*-medias (*k*-means) [32, 61-63]. Este algoritmo funciona mejor con grupos que tienen forma convexa y requiere que el número de grupos a obtener se especifique a priori, por tanto requiere un cierto conocimiento del dominio, ya que es sensible a cómo se hizo inicialmente la partición. El algoritmo *k*-medias tiene una complejidad temporal $O(Ikn)$, donde I se utiliza para indicar número de iteraciones, n el número de objetos y k el número de grupos. Esta notación será válida en todo el epígrafe. A partir de él se han derivado varios como el *x*-medias (*x*-means) que estima eficientemente el número de grupos, el llamado conjunto *k*-medias (batch *k*-means) [64], PAM [32] y sus mejoradas CLARA y CLARANS [65], y la modificación propuesta en [66]. Todos intentan resolver las desventajas del *k*-medias, pero son costosos computacionalmente en su mayoría. Estos algoritmos funcionan bien cuando los datos tienen baja dimensionalidad y los grupos están bien separados y tienen alta densidad. Otro que mejora el *k*-medias es el denominado primero el más lejano (Farthest First) [67], que realiza selección aleatoria de los centros de grupos y cada instancia que quede más lejos del centro más cercano forma un nuevo centro, hasta que el número de grupos supere un umbral; en [68] lo utilizaron en dominios textuales. Propuestas que usan análisis discriminatorio de grupos [69] y sus combinaciones con SVM [70, 71] también superan los resultados del *k*-medias.

Los algoritmos jerárquicos, por su parte, hacen una descomposición jerárquica de los objetos. Dentro de ellos, los aglomerativos (bottom-up), consideran que cada objeto constituye un

grupo, por tanto, inicialmente existen tantos grupos como objetos tiene la colección, y sucesivamente los unen, hasta que todos los objetos formen un único grupo, generalmente considerando alguna de las medidas de distancia entre grupos mostradas en el Anexo 5; dentro de ellas, el enlace simple [72, 73] y el enlace completo [74] son ampliamente utilizadas. Mientras que los divisivos (top-down) consideran inicialmente que existe un único grupo al cual pertenecen todos los objetos y sucesivamente dividen los grupos en grupos más pequeños, hasta que cada grupo contenga un único objeto. La construcción de la jerarquía se puede detener por criterios automáticos o del usuario. Muchas veces combinar o dividir grupos es comprometido, y no puede ser deshecho o refinado. En [75, 76] se presenta una nueva metodología que combina la estrategia divisiva y la aglomerativa.

Las técnicas jerárquicas han sido utilizadas en problemas de minería de datos [77], a pesar de tener alta complejidad temporal, generalmente cuadrática. BIRCH [78] es una variante con complejidad lineal, pero no descubre grupos con calidad, requiere de parámetros de entrada que pueden forzar el tamaño de los grupos, es sensible al orden de los datos de entrada y es cuestionable su uso en datos con alta dimensionalidad. CURE es capaz de captar grupos de varias formas y tamaños, tiene una alta complejidad, $O(n^2 \log n)$, y es sensible a varios parámetros de entrada [79]. BIRCH y CURE manejan bien los puntos fuera de rango. Algoritmos jerárquicos divisivos muy referenciados son PDDP [80] y su mejora sPDDP [64].

Los métodos jerárquicos organizan los datos en una secuencia anidada de grupos, que puede visualizarse en forma de dendrograma o árbol. Basándose en el dendrograma se decide cuál es el número de grupos para el cual los datos están mejor representados según un propósito dado. Usualmente el número "verdadero" de grupos para un conjunto de datos dado se desconoce a priori. Sin embargo, al usar métodos que crean particiones usualmente se requiere especificar el número de grupos como un parámetro de entrada [10]. Estimar este número es de gran interés, dos técnicas básicas son optimizar una función de validación de toda la partición, o evaluar cada grupo individualmente y refinar el agrupamiento [54]. Muchas veces métodos que forman jerarquías son refinados mediante la relocalización iterativa de puntos en la conformación de particiones; por ejemplo, las salidas de PDDP y sPDDP se refinan con el algoritmo denominado medias [64].

Otros tipos de métodos han emergido para el análisis de grupos, principalmente motivados en problemas específicos de minería de datos [59]. El agrupamiento basado en densidad (density-based clustering) agrupa objetos vecinos de un conjunto de datos basándose en condiciones de densidad. Éstos difieren de los algoritmos que obtienen particiones mediante la relocalización iterativa de puntos a partir del número de grupos. Los algoritmos DBSCAN [81] y DENCLUE [82] son ejemplos de algoritmos basados en densidad. Ambos tienen una complejidad $O(n\log n)$, no funcionan correctamente con datos de alta dimensionalidad y dependen altamente de los parámetros iniciales. OPTICS [83] y el algoritmo propuesto en [84], son variantes mejoradas del DBSCAN. Otros algoritmos basados en densidad se reportan en [85-87].

El agrupamiento basado en celdas (grid-based clustering) es esencialmente propuesto para la minería de datos espaciales [37]. Algunos algoritmos basados en celdas son STING [88], WaveCluster [89, 90] y CLIQUE [91]. Estos algoritmos son escalables, tienen complejidad $O(n)$, pero no son buenos para datos con alta dimensionalidad, porque se focalizan en la modelación de la estructura geométrica de objetos en el espacio y no dependen de una medida de distancia. El agrupamiento basado en densidad permite descubrir grupos de varias formas, mientras que el agrupamiento basado en celdas se conoce por su alta velocidad. Los algoritmos AGRID, CLONE [92] y GARDEN [93] combinan ambos enfoques.

Existe una clasificación que divide el agrupamiento en conceptual y estadístico [37]. Por otra parte, aquellos algoritmos que utilizan métodos geométricos y técnicas de proyección se clasifican en agrupamiento incompleto o heurístico [54]. Otras propuestas agrupan utilizando redes neuronales artificiales, por ejemplo, mapas auto-organizativos (Self Organizing Maps; SOM) [37]. Algunos agrupamientos se basan en modelos (model-based clustering) encontrando buenas aproximaciones de los parámetros del modelo que mejor ajusten a los datos.

Otra clasificación, no mutuamente excluyente a las ya presentadas, considera la forma de manipular la incertidumbre en términos del solapamiento de los grupos: agrupamiento duro y borroso [54]. Las técnicas duras pueden ser deterministas o con solapamiento. Las deterministas crean una partición, donde los grupos son mutuamente excluyentes y exhaustivos del universo de objetos. Los algoritmos con solapamiento crean un cubrimiento,

donde un objeto puede pertenecer a más de un grupo. Las técnicas borrosas se subdividen en probabilísticas y posibilísticas [10]. El algoritmo EM [94], base del agrupamiento basado en modelos probabilísticos, y su mejora FREM [95], son variantes del algoritmo *k*-medias y asignan a los objetos una distribución de probabilidad de pertenencia a cada grupo. Éstos manipulan datos de alta dimensionalidad, pero realizan un refinamiento muy costoso.

Otra clasificación divide los algoritmos en estáticos e incrementales. Estos últimos tienen la habilidad de procesar nuevos datos que son adicionados a la colección; por ejemplo, el flujo de noticias [96]. El algoritmo *k*-medias incremental (incremental *k*-means) es una muestra de este tipo de algoritmos [64].

El análisis mixto de grupos (joint cluster analysis) es otra técnica de agrupamiento que integra aquellos métodos que trabajan tanto con las propiedades endógenas de los objetos así como con las relaciones que existen entre ellos [56, 97]. El agrupamiento basado en restricciones (constraint-based clustering) reúne a aquellos algoritmos que consideran aspectos más significativos acorde a los requerimientos de la aplicación [58].

1.1.2 Métodos que parten de la representación en un grafo de los objetos a agrupar

La teoría de grafos constituye una herramienta valiosa para desarrollar modelos de abstracción para el agrupamiento, proporcionando el formalismo matemático requerido. Sus conceptos básicos han sido utilizados para el desarrollo de algoritmos de agrupamiento y de índices de validación. Los grafos proveen modelos estructurales para el análisis de grupos. Generalmente este tipo de algoritmos no requieren que se especifique el número de grupos a obtener, descubren grupos de formas arbitrarias, no uniformes y de densidades diversas, y son menos sensibles a la presencia de ruido.

En problemas de la minería de textos existen conjuntos de datos que pueden intrínsicamente formar un grafo para aplicar posteriormente métodos de agrupamiento. Por ejemplo, los términos que describen una colección de documentos, donde los nodos representan los términos y las aristas entre ellos significan que los dos términos enlazados co-ocurren en una misma unidad textual [98]. Los enlaces de páginas Web constituyen otro caso donde cada nodo puede representar una página y las aristas los enlaces entre ellas [29]. Por otra parte, al representar la presencia de términos en documentos se puede utilizar un grafo bipartito con un conjunto de nodos documentos y otro de palabras, donde las aristas indican la presencia de las

palabras en los documentos [99]. También se puede representar la co-ocurrencia de conceptos en documentos en un grafo bipartito [100]. Varios algoritmos se han desarrollado recientemente para agrupar sobre este tipo de grafos [101, 102]. Por otra parte, también se pueden obtener grafos de similitud a partir de la matriz de similitud entre los objetos. Por ejemplo, se puede obtener la matriz de similitud coseno entre documentos de una colección y a partir de ella obtener un grafo donde cada nodo representa un documento y las aristas entre ellos son ponderadas con la similitud coseno[15], en [103, 104] hacen uso de esta representación. Igualmente es posible obtener un grafo donde los nodos representen documentos y las aristas la información mutua entre ellos.

Algunos ejemplos de algoritmos que parten de la representación en un grafo de similitud de los objetos a agrupar son el algoritmo Estrella (Star) [105] y sus extensiones [103, 104, 106]. Éstos generan cubrimientos sobre los datos y no requieren especificar el número de grupos a obtener, pero son sensibles al umbral de similitud fijado inicialmente. Estrella Extendido (Extended Star; ES) es independiente del orden de los datos y obtiene menor número de grupos respecto al algoritmo Estrella [105]. Los algoritmos Estrella Generalizada (Generalized Star; GStar) [104] y Estrella Condensada (Condensed Star; ACONS) [103] introducen nuevos conceptos de estrella y obtienen un menor número de grupos. En [107] se muestra un algoritmo concatenado que permite refinar la salida del algoritmo estrella extendido [106] mediante la aplicación de los algoritmos SKWIC duro y borroso [64]. Así, se obtiene un agrupamiento de mejor calidad y no es necesario especificarle a SKWIC el número de grupos ni sus centros.

Los métodos jerárquicos a partir de una representación de los objetos en grafos han sido muy trabajados. Varios siguen estrategias divisivas, entre ellos uno de los más conocidos es el método basado en la construcción de un árbol de expansión mínimo [108]. Otras propuestas que construyen jerarquías a partir de grafos son: STIRR [109] que aplica métodos espectrales; el algoritmo que divide y combina los grupos por su conectividad interna y cercanía [110]; los algoritmos jerárquicos utilizados por el sistema SUBDUE para agrupar datos estructurados o

[15] En el Anexo 2 se muestra la expresión de la similitud coseno.

no [77] y el algoritmo jerárquico propuesto en [111] que descubre grupos de formas irregulares en datos de alta dimensionalidad, pero tiene complejidad cuadrática.

En [112] se presenta un marco de trabajo para algoritmos jerárquicos aglomerativos basados en grafos. Sin embargo, en [29] afirman que los métodos jerárquicos aglomerativos tienen problemas en el análisis de grupos de objetos representados en un grafo. Un problema es que fallan con cierta frecuencia al encontrar grupos en grafos donde se conoce la estructura, lo cual hace difícil tener credibilidad cuando funcionan correctamente. Otro problema es su tendencia a encontrar solamente los centros de los grupos y no incluir la periferia. Los nodos centrales de un grupo usualmente tienen una similitud alta, y son conectados tempranamente en un proceso aglomerativo, pero los nodos de la periferia, que no tienen una similitud tan fuerte con los otros y generalmente tienen un único enlace al grupo al que pertenecen, tienden a quedar abandonados.

Varios métodos no sólo parten de una representación de los objetos en un grafo, sino que explotan las propiedades estructurales de las conexiones entre los objetos a agrupar; por ejemplo, los algoritmos basados en la densidad local de los nodos [113, 114]. Éstos utilizan las características de los grafos scale-free [115-117]. Otra propuesta, basada en el coeficiente de agrupamiento local, se publicó en [20]. Es exitosa para grafos densos, donde se definen grupos fuertes y débiles considerando el grado de las conexiones internas en los grupos y las externas hacia otros grupos. El algoritmo SCAN [118], con complejidad $O(m)$ donde m es el número de aristas, utiliza la vecindad de los nodos como un criterio de agrupamiento, así los nodos se agrupan considerando los vecinos que comparten. El algoritmo SMTIN permite minar datos espaciales pero requiere la especificación de un umbral de distancia y con un único umbral no es posible obtener grupos con múltiples resoluciones [119]. Esta desventaja se supera con la extensión propuesta en [120]. Otras propuestas basadas en las propiedades estructurales se presentan en [121-123].

Las técnicas de agrupamiento tienen ventajas y desventajas en dependencia del tipo de problema al cual son aplicadas. El conocimiento del dominio puede, en muchos casos, ayudar a determinar qué tipo de grupo se va a formar y qué tipo de agrupamiento se va utilizar con el objetivo de obtener los mejores resultados. Se han desarrollado muchos algoritmos para el agrupamiento de documentos, pero muy pocas propuestas toman en consideración las ventajas

de las estructuras inherentes en el lenguaje. Varias investigaciones muestran que el lenguaje existe en una red small-world [124-126]; sin embargo, raras veces ha sido utilizado en el agrupamiento en dominios textuales [98]. Por otra parte, cuando se representa una colección textual como un grafo de similitud coseno entre los documentos, sucede que, como se muestra en la Figura 1.1, algunos documentos tienen alta pertenencia a los grupos, otros pertenecen a puentes entre grupos y algunos tienen una pertenencia débil a los grupos; por tanto, es fundamental explotar las propiedades topológicas del grafo y el flujo de información entre los nodos para detectar las estructuras ocultas en él [118]. En algunos casos dos documentos pueden tratar temas diferentes y tener alta similitud coseno. Estos documentos no necesariamente son tan similares a sus vecinos respectivos, y sólo el uso de la similitud no logra identificarlos en grupos diferentes, se requiere analizar los enlaces entre ellos.

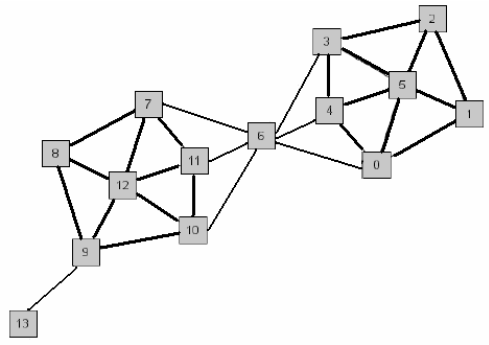


Figura 1.1 Grafo donde se han representado dos grupos, el nodo 6 es puente entre los grupos y el nodo 13 tiene muy baja pertenencia a los grupos. Fuente: [118].

A partir de la motivación principal de este trabajo y considerando de gran utilidad las propiedades topológicas de los lenguajes y corpus textuales, se pretende evaluar las posibilidades de los métodos de análisis de redes complejas existentes para abordar problemas de la minería de textos. Por tanto, se focalizará en el estudio de las técnicas que trabajan sobre una representación de los objetos en un grafo y que explotan la estructura de las interrelaciones de los objetos en el proceso de agrupamiento.

1.1.3 Métodos basados en la intermediación de las aristas

Los métodos jerárquicos divisivos sobre grafos generalmente averiguan cuáles son los pares de nodos que están más débilmente conectados. Otro enfoque posible, cuando estos métodos operan sobre representaciones gráficas, consiste en averiguar cuáles son las aristas en el grafo

que están más "entre" otros nodos, significando que la arista es, en algún sentido, la responsable de conectar muchos otros pares de nodos, aunque tales aristas no necesitan ser débiles en el sentido de similitud. La esencia es detectar cuáles son las aristas con mayor centralidad e ir eliminándolas en un proceso divisivo y así ir construyendo la jerarquía de grupos. En el Anexo 4 se describe la propiedad estructural: centralidad.

En la actualidad varias investigaciones reportan resultados favorables siguiendo este enfoque [19-23, 55, 127-131], siendo M. J. E. Newman un autor prolífero en este tópico [24-27, 30, 132-134]. La mayoría de estos trabajos miden la centralidad de la arista contando el número de caminos más cortos que pasan a través de ella, utilizando las llamadas medidas de la intermediación de los caminos más cortos (shortest-path betweenness) [55]. Estas medidas indican el potencial que tiene una arista para controlar el flujo de información en el grafo; así, favorecen a las aristas que se encuentran entre grupos y desfavorecen a aquellas incidentes a nodos de un mismo grupo. Intuitivamente, si una arista actúa en la interacción de muchos nodos su nivel de intermediación debe ser alto[16].

Uno de los primeros algoritmos jerárquicos divisivos que usa el cálculo de la intermediación de las aristas a partir de la determinación de los geodésicos[17] en un grafo se propuso bajo el supuesto: los caminos más cortos entre grupos viajan por el pequeño número de aristas que los comunican, produciendo en éstas alta intermediación [55]. Este algoritmo, denominado GN debido a Girvan y Newman, sigue un proceso divisivo mediante la eliminación de las aristas con mayor intermediación. En el Anexo 6 se muestran los cuatro pasos que conforman este algoritmo. Aunque GN es poco conocido en el área del agrupamiento de documentos, pues se desarrolló para analizar las redes complejas, se ha incluido en este estudio porque constituye el antecesor directo del método de agrupamiento propuesto. En [127] se presenta una variante de GN que selecciona aleatoriamente una de las aristas con máxima intermediación en el proceso de eliminación.

Las aristas con altos valores de intermediación producen un incremento de la distancia geodésica[18] entre un gran número de pares de nodos cuando éstas son eliminadas del grafo

[16] Intuición original para la intermediación de nodos. 135. Bavelas, A., *A mathematical model for group structures.* Human Organization, 1948. **7**: p. 16-30.
[17] Geodésico o camino más corto que enlaza un par de nodos dado en un grafo.
[18] Costo del camino mínimo entre un par de nodos en un grafo.

[55, 127, 136]. Así, el recálculo de la intermediación después de cada eliminación es distintivo de estas propuestas. Esto significa que no hay una función que pueda ser definida para cada arista en el grafo inicial tal que el dendrograma resultante sea la representación del agrupamiento jerárquico llevado a cabo usando dicha función [127].

Una variante dura y con solapamiento, a diferencia de las dos anteriores deterministas, fue propuesta en [23]. Este algoritmo elimina aquellas aristas con alto valor de intermediación y que sean incidentes a nodos con valores de intermediación[19] similares. Un algoritmo de agrupamiento escalable que calcula los geodésicos se presenta en [19]. Éste funciona correctamente con grafos grandes y tiene una complejidad temporal $O(n \log n)$. En este epígrafe se utiliza n para indicar el número de nodos y m el número de aristas. Otros algoritmos, basados en la intermediación a partir de los geodésicos, intentan mejorar la calidad de los grupos o realizar este proceso de una forma menos costosa [21, 22, 27].

Los algoritmos jerárquicos divisivos que utilizan la intermediación según los caminos geodésicos tienen desventajas debido a la complejidad del cálculo y a problemas presentados por esta forma de medición de la centralidad de las aristas.

Alta complejidad. El cálculo de la intermediación según los caminos más cortos es costoso, ya que encontrar caminos de costo mínimo entre un par de nodos lo es, y adicionalmente es necesario calcularlos entre todos los pares de nodos del grafo. A pesar de que se han propuesto variantes que intentan reducir la complejidad de este cálculo [133, 137, 138], éste continúa siendo costoso. En [133, 138] se proponen modificaciones del cálculo de la intermediación en un tiempo $O(mn)$ y $O(n^2)$ para grafos dispersos. Modificaciones de la búsqueda primero a lo ancho para considerar todos los caminos mínimos en grafos no ponderados hacen más costoso el cálculo de la intermediación [137]. El cálculo de esta medida en grafos ponderados es aún más costoso, sólo calcular las distancias mínimas entre todos los pares de nodos tiene una complejidad $O(mn \log n)$ [133, 137] cuando se utiliza una variante optimizada del algoritmo Dijkstra [139]. Otra propuesta para grafos ponderados consume $O(mn+n^2 \log n)$ [138].

Medición deficiente de la centralidad. La forma en que se calcula la intermediación de las aristas no garantiza que todas las aristas que conectan grupos tengan intermediación alta. Esto

[19] En el Anexo 4 se muestra que es posible calcular la intermediación de los nodos; así se originó este concepto.

ocurre cuando existen puentes paralelos[20] entre los grupos a descubrir. Newman en [127] lo expresó como sigue:

> "Cuando hay más de una conexión entre dos grupos de vértices, no existe garantía de que todas las conexiones existentes reciban valores altos de intermediación; en algunos casos la mayoría de los caminos geodésicos fluyen a través de una arista y solamente esa arista recibirá un valor alto de intermediación."

Por esta razón, los algoritmos de agrupamiento que utilizan la intermediación según los geodésicos necesitan el recálculo, produciendo un incremento de la complejidad temporal de estas propuestas. Agrupar según [55, 127] consume $O(m^2n)$ en el peor de los casos, o un $O(n^3)$ para grafos dispersos.

Estos métodos jerárquicos divisivos basados en el cálculo de la intermediación de las aristas requieren de una fase de validación de los agrupamientos en la cual suele usarse una medida de calidad conocida como modularidad [29]. Esta medida será comentada en el epígrafe 1.2.3, cuando se aborden las medidas de validación internas. Ante la presencia de la alta complejidad computacional del cálculo de la intermediación y de la necesidad del recálculo en el proceso del agrupamiento, en una segunda etapa del desarrollo de estos métodos, los autores prefieren optimizar directamente la modularidad y abandonan el uso de las medidas de centralidad aplicadas en la primera etapa [30, 140-143].

No obstante a este giro en el desarrollo de estas técnicas, existen otras expresiones que fueron creadas en el afán de captar, adecuadamente y con menor costo, la centralidad de las aristas [21-23, 26, 29, 144-150]. Estas nuevas propuestas, en su mayoría, son fieles a la idea de la intermediación, algunas se alejan de esta técnica siguiendo otras estrategias para medir la centralidad.

En muchas situaciones reales la comunicación no viaja a través de los caminos geodésicos solamente. Así, en [146] se propone el flujo de la intermediación (flow betweenness) basado en todos los caminos posibles entre un par de nodos y se consideran las aristas como canales de comunicación. Su cálculo tiene una complejidad temporal $O(m^2n)$ y es necesario conocer la ruta ideal de cada fuente al destino en casos donde el flujo no tome un camino ideal, desventajas que mantiene de la intermediación a partir de los geodésicos. La intermediación a

[20] Se dice puente a una interconexión –arista o camino– entre dos grupos de vértices. Dos o más puentes se dicen paralelos si conectan el mismo par de grupos.

partir de caminos aleatorios (random-walk betweenness) es una propuesta que supera las desventajas anteriores [26, 29]. Su cálculo tiene una complejidad de $O(mn^2)$ y el algoritmo que la utiliza, incluyendo el recálculo, consume $O((n+m)mn^2)$, o $O(n^4)$ para grafos dispersos [29]. Otras medidas de centralidad basadas en caminos aleatorios son la centralidad poderosa de Bonacich (Bonacich's power centrality) [144] y la centralidad de paso aleatorio (random-walk centrality) [147]. En [151-155] se proponen expresiones de intermediación que no sólo involucran los caminos geodésicos en su cálculo, sino que incluyen otros caminos entre pares de nodos.

Para reducir la complejidad temporal del cálculo de la intermediación se han presentado variantes paralelas [156, 157] y el uso del índice para los caminos mínimos entre todos los pares de nodos, con complejidad $O(n^2)$ [149, 158]. Otras variantes proponen algoritmos aproximados para hacer menos costoso el cálculo de la intermediación [159-161]. A pesar de la existencia de propuestas que reducen la complejidad del cálculo de la intermediación [19, 132, 138, 156, 158], la evaluación de tal cantidad consume mucho tiempo, por ser calculada como una medida global dependiendo de las propiedades de todo el grafo. Por tal motivo, otras propuestas analizan los grafos localmente, en lugar de considerar todo el grafo en el cálculo de la intermediación.

En [162, 163] se presenta la medida intermediación ego (ego betweenness) que ajusta el cálculo de la centralidad en un grafo ego[21] o vecindad, revelando conclusiones importantes de todo el grafo desde la perspectiva local. Esta medida es sensible a variaciones de la densidad local y los enlaces entre las áreas densas [164], por eso en [163] se sugiere una variante normalizada localmente. Su cálculo es menos complejo que calcular la intermediación de todo el grafo.

Otras variantes para la medición local de la centralidad se presentan en [20, 148, 165-167]. El coeficiente de agrupamiento de las aristas (edge-clustering coefficient) [20] es una variante local que ha reemplazado la intermediación según los geodésicos en el algoritmo GN. Esta modificación tiene una complejidad $O(m)$ en grafos pequeños y $O(m^2)$ en grafos grandes, logrando resultados similares y menos costosos que los propuestos por Girvan y Newman.

[21] Redes donde uno o varios nodos son designados ego: coordinadores o facilitadores de la comunicación entre los otros, con el objetivo que el resto quede organizado en grupos de trabajo y descrubrir la red ego.

Este coeficiente sólo es aplicable a grafos no ponderados. En [168] se agrupan los objetos utilizando la intermediación según geodésicos o el coeficiente de agrupamiento de las aristas dependiendo del tamaño de los grafos.

Este índice de proximidad es medido por una partícula Browniana parcial (biased Brownian particle) que integra información local y global del grafo [148]. Cuantificar la proximidad para los pares de nodos vecinos necesita un tiempo $O(n^3)$. El algoritmo de agrupamiento Netwalk se basa en este índice. Otra variante local, centralidad de la información (information centrality), combina dos elementos en la medición de la centralidad: cuán cercano es un nodo a los otros y cuánto él media entre los otros [166]. Esencialmente mide la habilidad del grafo en la propagación de la información entre sus nodos, antes y después que un cierto nodo ha sido desactivado. En [167, 169] se generalizó a aristas. Igualmente, variantes del algoritmo GN reemplazaron la intermediación según los geodésicos por la centralidad de la información de la arista. Esta modificación es costosa, $O(m^3n)$, o $O(n^4)$ para grafos dispersos, pero más efectiva que los algoritmos anteriores especialmente cuando los grupos están muy mezclados.

1.2 Validación del agrupamiento

"El agrupamiento es un proceso subjetivo; el mismo conjunto de datos usualmente necesita ser agrupado de formas diferentes dependiendo de las aplicaciones" [170]. Esta subjetividad hace el agrupamiento difícil, y aún más su validación.

Una forma de evaluación del agrupamiento muy sencilla puede ser, por ejemplo, mediante la visualización del conjunto de datos cuando éste es pequeño y los datos son bidimensionales. Sin embargo, esta forma de evaluación puede ser difícil al intentar realizar una visualización efectiva de un conjunto de datos de alta dimensionalidad [171], aunque técnicas bien establecidas para la reducción de dimensionalidad, como el análisis de componentes principales (Principal Component Analysis; PCA) [172] y el análisis de componentes principales categórico (Categorical Principal Component Analysis; CATPCA), pueden contribuir a la evaluación de este tipo de conjuntos de datos mediante la visualiación. ¿Qué hacer cuando los datos no pueden representarse gráficamente, o es muy difícil o algunas veces imposible para un observador humano valorar el agrupamiento de los mismos; o no existe una forma simple de decidir si el resultado de un agrupamiento se ajusta a la división que deseamos de los datos?

Por otra parte, muchos algoritmos de agrupamiento varían sus resultados dependiendo de las características de los datos, por ejemplo, geometría y densidad de distribución [37]. Otros dependen fuertemente de los valores asignados a los parámetros. Por ejemplo, si hay un parámetro que controle la resolución a la cual los datos son vistos, el algoritmo produce un dendrograma en función de ese parámetro. En este caso es necesario decidir cuál nivel del dendrograma refleja mejor los grupos según las propiedades que se desea que el agrupamiento satisfaga. Algunos algoritmos necesitan que se especifique inicialmente el número de grupos a obtener, otros requieren que se especifique el número de vecinos de cada punto como un parámetro externo. Así, los resultados producidos son en función de los parámetros fijados y se hace necesario verificar cuáles se ajustan a los datos [52].

Variaciones a partir de características de los datos, diferentes técnicas de análisis de grupos y definición de parámetros para el algoritmo a aplicar, indican que una evaluación de los resultados es necesaria para medir la calidad del agrupamiento. Una práctica común, en tal sentido, es aplicar medidas de validación de grupos [38].

El procedimiento de evaluar los resultados de algoritmos de agrupamiento se conoce por validación del agrupamiento [31, 173]. Se dice medida de validación de grupos a una función que hace corresponder un número real a un agrupamiento, indicando en qué grado el agrupamiento es correcto o no [54]. Estas medidas en su mayoría son heurísticas por naturaleza para conservar una complejidad computacional plausible.

1.2.1 Clasificación de las medidas

Las medidas de evaluación del agrupamiento se clasifican en: globales y locales, subjetivas y objetivas, internas, externas y relativas, y supervisadas y no supervisadas [32, 54, 174].

Las medidas globales describen la calidad del resultado completo de un agrupamiento usando un único valor real, mientras que las locales evalúan cada grupo obtenido [54]. Las medidas objetivas miden propiedades estructurales de los resultados de los agrupamientos, por ejemplo, la separación entre los grupos y la compactación o densidad de los mismos [37]. La presencia de tales propiedades no garantiza que los resultados sean interesantes para el usuario, estas medidas carecen del enlace con los usuarios, aunque su principal atractivo es que son independientes del dominio [174]. Las medidas subjetivas evalúan considerando la usabilidad

de los grupos [38]. Las investigaciones en medidas subjetivas han sido menos intensas que las realizadas en las objetivas [175].

Una clasificación muy usada divide la validación del agrupamiento en: medidas internas, externas y relativas [31, 32]. En el Anexo 7 se muestra un esquema de esta clasificación. Otra división consiste en medidas supervisadas y no supervisadas, refiriéndose a externas e internas, respectivamente. Las medidas externas se basan en un criterio externo que es impuesto sobre los datos, por ejemplo, una estructura previamente especificada que refleje la intuición que se tenga del agrupamiento de los datos. No es posible aplicar estas medidas a situaciones del mundo real donde usualmente no está disponible una clasificación de referencia. Las medidas internas evalúan considerando solamente los resultados del agrupamiento en términos de cantidades que involucran los vectores de datos. Las medidas relativas se basan en la comparación del agrupamiento a evaluar con otros esquemas de agrupamiento o con resultados del mismo algoritmo con diferentes valores en los parámetros.

A continuación se mencionarán las principales medidas externas e internas reportadas en la literatura para la evaluación de particiones y cubrimientos; sus expresiones se muestran en el Anexo 8 y el Anexo 9, respectivamente. Las variantes para evaluar agrupamiento borroso no serán abordadas, por no ser objetivo de este trabajo.

1.2.2 Principales medidas externas

Una medida externa es la entropía [176], la cual es una función de la distribución de las clases en los grupos resultantes. La entropía total para un conjunto de grupos es calculada como la suma de las entropías de cada grupo, ponderadas con el tamaño del grupo [177]. En [178] también usan la entropía como métrica de calidad, la mejor entropía es obtenida cuando cada grupo contiene exactamente un objeto. Otras expresiones para calcular la entropía por grupos y para el resultado del agrupamiento en general se presentan en [179].

Algunas medidas externas usan las ideas de precisión (precision) y cubrimiento[22] (recall) del campo de la recuperación de información y las adaptan a la validación del agrupamiento. Precisión (Pr) y cubrimiento (Re) se calculan para un grupo *j* y una clase *i* dados, usando las

[22] En este documento se utiliza cubrimiento como traducción de la medida recall. Adicionalmente, se utiliza el término cubrimiento para nombrar una forma de la división de los objetos después de un agrupamiento duro y con solapamiento.

expresiones $\Pr(i,j)=n_{ij}/n_j$ y $\mathrm{Re}(i,j)=n_{ij}/n_i$, respectivamente; donde n_{ij} es el número de objetos de la clase i en el grupo j, n_j es el número de objetos del grupo j y n_i es el número de objetos de la clase i. La medida-*F* (*F*-measure) se obtiene calculando la media armónica de precisión y cubrimiento. Se puede variar el umbral α $(0 \leq \alpha \leq 1)$ para regular la influencia de precisión y cubrimiento en el cálculo de esta medida [180]. Un valor global, la medida-*F* global (Overall *F*-measure; OFM), se calcula usando el promedio ponderado de los valores máximos por clase de la medida-*F* sobre todos los grupos [178]. La medida-*F* intenta capturar cuanto los grupos del agrupamiento obtenido se hacen corresponder correctamente con los grupos de referencia [181]. En [182] se propuso una variante de la medida-*F* para un agrupamiento jerárquico, tomando por clase el máximo valor de la medida-*F* sobre todos los grupos a todos los niveles de la jerarquía. Variantes de precisión y cubrimiento, micro-averaged precision y micro-averaged recall, son utilizadas para evaluar el agrupamiento [183], las expresiones para su cálculo coinciden si cada objeto pertenece a sólo un grupo y la clasificación de referencia también tiene una única clasificación para cada objeto.

Las formas del cálculo de las medidas precisión y cubrimiento antes mencionadas son usualmente aplicadas a particiones. En [184] se presentan variantes del cálculo de estas medidas sobre pares de puntos y frecuentemente se usan para evaluar cubrimientos resultantes de algoritmo duros y con solapamiento.

En [181] se sugiere el uso del estadístico Kappa para evaluar la concordancia que existe entre dos particiones relativas a una partición de referencia, considerando las dos particiones obtenidas como intentos de clasificaciones siguiendo la referencia. La información mutua entre dos grupos, otra forma de evaluar, toma valores entre cero y el máximo valor de entropía de los grupos, alcanzado cuando ambos grupos son idénticos [185]. Existen variantes normalizadas, pero no se proponen expresiones para valorar globalmente los resultados.

El índice nombrado error del agrupamiento utiliza el número de asociaciones incorrectas o ausentes para medir la cercanía entre el resultado del agrupamiento y la clasificación de referencia [186]. Se dice asociación en una partición a un par de objetos que pertenezcan al mismo grupo, asociación incorrecta a aquella que existe en la partición de referencia y no en el resultado del agrupamiento, y asociación ausente a aquella que existe en el resultado del agrupamiento y no en la partición de referencia. Esta medida favorece particiones pequeñas,

por tanto una variante normalizada en el intervalo [0, 1] se presenta para proveer una menor dependencia del tamaño de ambas particiones. Siguiendo estas ideas, en [186] se redefinen cubrimiento y precisión para reflejar cuán bien el agrupamiento detectó las asociaciones entre los objetos y cuál es la precisión de las asociaciones detectadas, respectivamente. Estas variantes de medición son muy utilizadas en algoritmos que obtienen cubrimientos.

La distancia Euclidiana ha sido utilizada para medir la equivalencia estructural de dos grupos, siendo cero si los grupos son estructuralmente equivalentes y mayor si no lo son [113]. El coeficiente de correlación también permite medir la equivalencia estructural de los grupos, mediante la división de la covarianza de la representación vectorial de los objetos por el producto de su desviación estándar. Este coeficiente toma valores entre -1 y +1 (los grupos son estructuralmente equivalentes).

Existen varias medidas basadas en la distribución de pares de objetos, entre ellas: el estadístico Rand, coeficiente Jaccard, índice Folkes y Mallows, estadísticos Γ Huberts y Γ normalizado [187]. Estas medidas usualmente se trabajan utilizando las técnicas de Monte Carlo que calculan una función de densidad probabilística de los índices estadísticos definidos y tienen un alto costo computacional [31]. La evaluación se realiza comparando el agrupamiento con grupos de referencia o con la matriz de proximidad [58].

1.2.3 Principales medidas internas

Los algoritmos de agrupamiento generan una estructura espacial y se pueden definir medidas para diferentes aspectos de esta estructura, por ejemplo, densidad [188]. Existen varios trabajos encaminados al desarrollo de medidas que validan el agrupamiento de una manera no supervisada. Algunos antiguos como el índice Goodman-Kruskal que tiene una alta complejidad computacional [189], el índice C apropiado cuando los grupos tienen tamaños similares [190], los índices propuestos en [191, 192] utilizan criterios de información, seguidos por los propuestos en [61, 193]. El cálculo de la dispersión dentro del grupo y la separación entre los grupos han sido ampliamente trabajados, un ejemplo es el índice Calinski-Harabasz [194], utilizado recientemente en [195].

La cohesión de los grupos puede usarse como una medida de validación de éstos. La medida interna nombrada similitud global (overall similarity; OS) se ha utilizado para medir la cohesión basándose en la media de la similitud de los pares de objetos en un grupo [178].

Los índices para evaluar particiones generalmente se basan en alguna motivación geométrica para estimar cuán compactos y bien separados están los grupos. Un ejemplo son los índices Dunn [33] y sus generalizaciones [35]. Los índices Dunn varían en función de la medida de distancia entre grupos y el cálculo del diámetro del grupo que se utilice. Originalmente Dunn utilizó el mínimo de todas las distancias entre pares de elementos para calcular la distancia entre los grupos, y consideró el diámetro del grupo como la mayor distancia entre sus miembros [33]. Así, las medidas tienden a producir valores elevados en agrupamientos con grupos compactos y muy bien separados.

Bezdek determinó que el índice Dunn es muy sensible al ruido [35]. Por ejemplo, la distancia entre un par de grupos puede ser menor que el diámetro de un grupo. Bezdek propuso una modificación en el cálculo de la distancia entre grupos mediante la estandarización respecto al tamaño de los mismos y una nueva forma de cálculo del diámetro del grupo calculando la distancia de todos sus elementos al centro del grupo, también estandarizado por su tamaño. Esta variante obtiene mejores resultados para diferentes dominios, pero requiere la existencia de un centro de grupo, y no todos los algoritmos trabajan con prototipos, ni la estructura de todos los datos son grupos con forma esférica. Cinco generalizaciones de los índices Dunn se han propuesto para validar grupos con diferentes formas hiperesféricas y disminuir su sensibilidad al ruido [196]. Éstas abogan por definiciones apropiadas para el cálculo del diámetro de los grupos y la distancia entre los grupos, siguiendo el principio de que todos los datos deben estar explícitamente implicados en el cálculo del índice. En el Anexo 9 se aprecia que el índice Dunn provee una estructura muy general para definir índices de validación, cada combinación de distancia y tamaño de los grupos define un nuevo índice. Estos índices requieren una cantidad de tiempo considerable para su cálculo. Se desea obtener valores altos de estos índices al evaluar el resultado de un agrupamiento.

La medida Davies-Bouldin se basa en la idea que una buena partición es aquella con gran separación entre grupos y alta homogeneidad y compactación dentro de cada grupo. Esta medida es una proporción de la suma de la dispersión interna del grupo y la separación entre grupos. La dispersión dentro del grupo es relativa a los centros de éstos y la distancia entre los grupos considera la distancia entre sus centros. Una dispersión baja y una distancia grande entre grupos tienden a producir valores bajos, por tanto se desea una minimización de esta

medida [34]. Los índices Dunn y Davies-Bouldin son relativos al análisis geométrico de los grupos: típicamente centroidal y con forma esférica; elementos no presentes en todos los datos.

En [197] se generalizan los índices Dunn y Davies-Bouldin utilizando las estructuras de grafos GG, RNG y MST, donde los nodos son los objetos que fueron agrupados y las aristas pesadas entre los objetos indican la distancia que existe entre ellos. Se mostró que estas generalizaciones son más robustas a la presencia de ruido utilizando la distancia Euclidiana entre los objetos a agrupar. En [198] transforman los índices Dunn, Davies-Bouldin y el estadístico normalizado de Hubert a un marco de trabajo simbólico.

El índice I también sigue un esquema general similar a los índices Dunn, pero utiliza la distancia máxima entre grupos y adiciona, en lugar de promediar, las distancias multiplicadas por el número de grupos [195]. Por otra parte, el índice v_{SV} calcula la suma pesada de las distancias entre grupos y dentro de los grupos escogiendo la distancia mínima entre grupos y el promedio de las varianzas de los grupos como sus componentes, respectivamente [199].

Otras variantes de índices, en su mayoría con un alto costo computacional especialmente cuando el número de grupos y objetos es muy grande [200], se han propuesto en [201, 202]. Una medida interna y global es el índice de separación [54]. Valores altos de éste indican buenas particiones. Para conjuntos de datos grandes, la determinación del índice de separación es computacionalmente muy costosa porque el número de operaciones para determinar los diámetros y las distancias de los grupos depende cuadráticamente del número de datos, en su defecto es posible utilizar la medida llamada granularidad-disimilitud [203]. Esta medida es estable al evaluar la granularidad en resultados de agrupamientos basados en prototipos y existen extensiones para validar agrupamientos borrosos. El coeficiente de correlación Cohenetic permite medir el grado de similitud entre un dendrograma producido por un algoritmo jerárquico y la matriz de proximidad. Valores del índice cercanos a cero indican que hay gran similitud entre las dos matrices y la técnica de Monte Carlo puede ser usada para la validación [58].

En [36] se presenta el índice de validación SD que suma el promedio de compactación de los grupos y la separación total entre ellos. Un valor pequeño para el primer término indica grupos compactos, su valor es directamente proporcional a la dispersión. El segundo término es

sensible a la geometría de grupos e incrementa su valor al crecer el número de grupos, lo que hace que no pueda manipular adecuadamente grupos de formas arbitrarias. Además, requiere la existencia de centros de grupos. S_Dbw es otro índice que evalúa positivamente datos que formen grupos compactos y bien separados [204]. Se basa en la compactación de los grupos (medida como la varianza de los términos en cada grupo) y la densidad entre los grupos. No debe ser aplicado sobre grupos con formas no convexas, por ejemplo, anillos.

Hasta aquí se han mencionado varios índices que calculan la razón entre las distancias dentro de los grupos y las distancias entre los grupos, por ejemplo: índices Dunn, Davies-Bouldin e índice I. Otros calculan la suma pesada de esas dos distancias, por ejemplo SD, S_Dbw y v_{SV}. En [205] incluyen un análisis del diseño y funcionamiento de estos índices, y proponen modificaciones de los mismos a partir nuevas formas de cálculo de las distancias entre grupos y dentro del grupo para resolver las limitaciones encontradas.

El índice silueta es el promedio, sobre todos los grupos, del ancho de la silueta de sus puntos [206]. Dos cálculos fundamentales intervienen en la silueta de un punto: la distancia promedio entre el punto y todos los otros puntos en el grupo, y el mínimo de la distancia promedio entre el punto y los puntos en otros grupos. Valores altos del índice silueta global indican grupos más compactos y bien separados. El cálculo de este índice tiene una alta complejidad; sin embargo, investigaciones actuales lo utilizan para la validación del agrupamiento [188].

Los índices de validación: RMSSTD, SPR, RS y distancia entre dos grupos, deben ser usados simultáneamente para estimar el número de grupos existente en un conjunto de datos. Estos cuatro índices pueden ser aplicados a cada uno de los pasos de un algoritmo de agrupamiento jerárquico aglomerativo [207]. RMSSTD mide la homogeneidad de los grupos formados en cada paso del algoritmo, SPR mide la pérdida de homogeneidad después de combinar dos grupos, RS mide el grado de diferencias entre grupos y la distancia entre grupos se calcula cuando dos grupos son combinados en un paso dado. Estos índices, esencialmente RMSSTD y RS, son usados también para evaluar resultados de agrupamientos no jerárquicos. Por otra parte, en [208] proponen un método de validación basado en la distribución de los miembros de los grupos de una generación del agrupamiento a la próxima.

En [209] se propuso el método FOM y en [40] se proponen valores para sus parámetros. FOM se utiliza para estimar el número de grupos. Su principal limitante es que no puede aplicarse a

datos con condiciones experimentales diferentes; además, su puntuación disminuye cuando el número de grupos aumenta.

En [210] se presenta un método de agrupamiento jerárquico que utiliza dos medidas de validación interna durante la construcción de la jerarquía. Una de ellas compara el agrupamiento con los resultados de una hipótesis nula para todos los objetos y la otra escoge aleatoriamente un número de ejemplos y calcula fronteras estadísticas para determinar si se crearon muchos o pocos grupos.

Se han desarrollado varias medidas para medir la calidad de resultados de agrupamientos sobre datos representados gráficamente [29, 38, 127, 211-213]. Generalmente estas medidas asumen que cada objeto es un nodo del grafo y la ponderación de las aristas indica la similitud entre ellos. La no existencia de aristas representa pares altamente diferentes.

En [212] proponen una primera medida que calcula la expansión de un agrupamiento como la expansión mínima resultante de los árboles de expansión mínimos de cada subgrafo (grupo). La segunda, llamada conductancia, generaliza la primera permitiendo ponderar subconjuntos de vértices para reflejar su importancia. Así, le dan más valor a vértices con muchos vecinos similares y le restan a aquellos con pocos vecinos. Estas propuestas utilizan solamente el valor mínimo de expansión o conductancia entre todos los grupos, por tanto, pueden evaluar un agrupamiento desfavorablemente cuando la mayoría de los grupos tengan alta calidad. La calidad mínima de los grupos y el peso total de las aristas que no están cubiertas por grupos son criterios para resolver este problema, pero su optimización es costosa.

En [38] introducen las medidas conectividad parcial pesada Λ y densidad esperada ρ que también interpretan los datos como un grafo de similitud pesado [38]. Se desean valores altos de estas medidas al evaluar el agrupamiento. Sus principales desventajas son que tienen alta complejidad computacional y validan el agrupamiento considerando solamente la relación de los objetos dentro del grupo.

Una forma de medir la densidad en representaciones gráficas es mediante la conectividad interna del grupo, calculando la proporción del número de aristas dentro de él respecto al número posible de aristas. Por su parte, la cohesión se puede medir identificando cuán conectados están los miembros de un grupo a nodos que no son miembros de él, mediante el cociente: promedio de la interacción dentro del grupo entre el promedio de la interacción de

los nodos fuera del grupo. Otras medidas consideran información no-topológica adicional sobre la naturaleza de los grafos para evaluar resultados de métodos jerárquicos [12, 20, 29]. Una propuesta específica para algoritmos basados en la intermediación y limitada al más bajo nivel de la estructura de los grupos se presenta en [12]. Otra se introdujo en [20] donde se utilizan las definiciones grupo débil si la suma de todos los grados referentes a conexiones internas de los nodos que pertenecen al grupo es mayor que la suma de todos los grados de conexiones de éstos al resto del grafo y grupo fuerte si cada nodo que pertenece a él tiene más conexiones dentro del grupo que con el resto del grafo. Existen otras definiciones de grupo débil y fuerte [123]. La modularidad, utilizada para evaluar agrupamientos jerárquicos, mide la fortaleza de los grupos encontrados analizando las interconexiones antes y después del agrupamiento realizado [29, 127].

En [213] proponen el índice de tendencia del agrupamiento (clustering tendency index; IC) que parte de un grafo k-partito (asociado a los k grupos obtenidos por un algoritmo), donde dos vértices que pertenezcan a diferentes conjuntos son adyacentes si la disimilitud entre ellos es mayor que un umbral α. La idea del índice es contar el número de aristas que faltan en el grafo bipartito para que sea completo y sumar cada una de estas diferencias. También se presenta su variante normalizada. Este índice depende de la definición de un umbral de corte y no tienen en cuenta las relaciones dentro de los grupos.

Un nuevo índice de validación que mide características geométricas de los datos y la separación entre los grupos es propuesto en [214], basado en trabajos previos [215, 216]. Este índice trabaja bien para datos que contienen grupos con tamaños diferentes y cercanamente distribuidos.

El error de un agrupamiento puede definirse como la diferencia esperada entre sus etiquetas y etiquetas generadas, por ejemplo, con una distribución Gausiana. Las etiquetas son referidas a la asignación de las instancias a los grupos [188].

Algunos métodos para evaluar se basan en la estabilidad de los grupos y controlan y alteran el ruido mediante el remuestreo del conjunto de datos original [39, 52]. En [217] proponen una medida basada en la estabilidad del agrupamiento, pero tiene una alta complejidad computacional.

1.3 Consideraciones finales del capítulo

La esencia del agrupamiento y su validación no está totalmente resuelta y depende del dominio de aplicación para considerar qué aspectos son los más significativos. Por ejemplo, en dominios textuales generalmente no se sabe cuántos grupos existen y ni siquiera hay razones para considerarlos del mismo tamaño. Por tanto, al realizar agrupamientos es favorable utilizar las propiedades estructurales de los datos, reducir la complejidad computacional, manejar datos de alta dimensionalidad y tener independencia de qué similitud se utilice para comparar los objetos, entre otros elementos.

El agrupamiento utilizando la intermediación de las aristas según los geodésicos utiliza propiedades topológicas de los grafos; sin embargo, tiene una alta complejidad computacional, requiere el recálculo y el conocimiento global del grafo.

La principal desventaja de GN es su complejidad: $O(m^2n)$. La matriz de intermediación se calcula en un tiempo $O(mn)$ para grafos no ponderados, siendo $O(n^3)$ para grafos densos y $O(n^2)$ para grafos dispersos –m número de aristas y n número de nodos. El paso 3 del algoritmo GN, mostrado en el Anexo 6, recalcula la intermediación para todas las aristas restantes, es crítico y adiciona otro orden de complejidad con respecto a m, el cuál significa en el peor de los casos el aumento de la complejidad en n^2; sin embargo, este paso se considera el corazón del algoritmo.

Nuevas medidas para el cálculo de la intermediación de las aristas pudieran ser diseñadas de manera tal que capturen eficientemente la información topológica que codifica la estructura del problema, de forma tal que se evite el recálculo en el proceso divisivo, se reduzca la complejidad y se obtengan grupos que asocien mejor a los datos.

Aunque las etiquetas para la evaluación del agrupamiento disponibles en los repositorios de archivos de datos pueden ser cuestionables, su existencia permite predecir comportamientos y aproximaciones de errores de los métodos de agrupamiento a analizar. Sin embargo, en algunas situaciones prácticas las clasificaciones de referencia no están disponibles porque se requiere mucho esfuerzo humano para obtenerlas y es costoso adquirir tal información. Por tanto, en dependencia de situaciones concretas se aplican medidas internas o externas, y sus usos no son incompatibles.

Varias medidas internas se han desarrollado; sin embargo, cada medida existente no logra captar todas las propiedades estructurales deseadas al evaluar el agrupamiento, por ejemplo, densidad, cohesión, compactación, separación. Por tanto, deben seleccionarse para ser aplicadas en función de los requerimientos de una situación dada. El funcionamiento de un algoritmo de agrupamiento puede ser juzgado de manera diferente dependiendo de qué medida de validación se haya usado. Los expertos pueden considerar los mejores resultados determinados por varios índices de validación y seleccionar aquel que mejor se ajuste a sus demandas. Para obtener resultados más confiables es adecuado utilizar varias medidas; ya que cualquier nueva medida, o nuevo enfoque, puede contribuir a un mayor entendimiento del agrupamiento. Una desventaja de la aplicación de múltiples medidas es que los cálculos en su mayoría son costosos y no es posible rehusar parte del procesamiento realizado para una como un subcálculo de otras porque no todas tienen igual naturaleza.

La evaluación del agrupamiento es una tarea con presencia de incertidumbre y vaguedad, sobre todo en dominios textuales. Además, considerando la aplicación que motiva este trabajo, así como el proceso de estimación de cortes en dendrogramas resultantes de agrupamientos jerárquicos, no se cuenta con clasificaciones de referencia de los objetos. Es posible preguntarse ¿existen herramientas matemáticas que permitan medir la calidad, precisión y consistencia de grupos de diversas formas sin requerir conocimiento previo del dominio?

RST es una herramienta para modelar la incertidumbre cuando ésta se manifiesta en forma de inconsistencia [48-51, 218, 219]. Dos ventajas de RST pueden utilizarse en la validación ya que no requiere información adicional o preliminar sobre el conjunto de datos, ni suposición sobre éstos, y se usa en circunstancias caracterizadas por vaguedad e incertidumbre.

2 Intermediación diferencial y agrupamiento en grafos

Existen métodos basados en las relaciones de los objetos representados en un grafo que tienen un alto costo computacional porque utilizan mediciones de la centralidad que no captan eficientemente las propiedades topológicas que codifican la estructura del problema. En este capítulo se propone la intermediación diferencial que es una nueva medida para evaluar el grado de intermediación que tiene una arista en un grafo. Esta medida tiene propiedades que la distinguen como un elemento a considerar en los algoritmos de detección de comunidades y agrupamiento. Adicionalmente, se presenta un algoritmo de agrupamiento basado la intermediación diferencial para manejar grafos ponderados y no ponderados. Finalmente, se muestra la aplicación del algoritmo al agrupamiento de colecciones textuales representadas en grafos ponderados con la similitud coseno entre los documentos.

2.1 Breves antecedentes

Esta investigación forma parte de un proyecto[23] que tiene como objetivo evaluar las posibilidades de los métodos de análisis de redes complejas existentes para abordar problemas de minería de texto. Al comenzar el estudio se detectaron deficiencias en tales métodos, por lo que la investigación se encaminó a la búsqueda de mejores variantes para analizar estas redes.

La idea principal emergió a partir de la observación siguiente: el desarrollo de los métodos de detección de comunidades en redes complejas puede dividirse en dos etapas fundamentales. En la primera se utilizó el concepto de intermediación [133, 137, 220, 221] que dice cuán central es un nodo (arista) en un grafo y se evalúa calculando la cantidad de caminos más cortos que atraviesan el nodo (arista). Se han propuesto muchas variantes y medidas similares [21-23, 26, 29, 144-149], pero desafortunadamente los algoritmos que se han creado son costosos. Estos métodos requieren de una fase de validación de los agrupamientos donde se usa frecuentemente la modularidad [29]. En la segunda etapa, los autores prefieren optimizar directamente la modularidad y abandonan el uso de las medidas de centralidad que se usaron en la primera etapa [30]. La opinión que se tiene en esta investigación es que el abandono de las medidas de centralidad en los problemas de detección de comunidades fue prematuro.

[23] Dirigido por el Dr. Alberto Ochoa, investigador del Instituto de Cibernética, Matemática y Física de Cuba (ICIMAF).

Pueden ser diseñadas nuevas medidas de tal manera que capturen eficientemente la información topológica que codifica la estructura del problema. Los métodos de optimización de medidas de calidad son convenientes, incluso necesarios en ocasiones, pero la utilización combinada de ambas estrategias parece ser la idea más promisoria. Los resultados que se presentan en este capítulo confirman la validez de este planteamiento.

Adicionalmente, se mostrará como algunos métodos del análisis de redes complejas se pueden utilizar satisfactoriamente en problemas de agrupamiento de documentos, particularmente aquellos que emplean una de las técnicas más utilizada en los últimos años en el campo del análisis de redes sociales, biológicas y de información: la intermediación de aristas y nodos. Después de dar los pasos iniciales y mientras se realizaban revisiones bibliográficas y se aprendía qué se estaba haciendo en ese terreno, se descubrió que existían algunas dificultades. A partir de los resultados de estas observaciones se propone en este capítulo:

1. Determinar las causas de las dificultades que provocaron que muchos investigadores abandonaran el cálculo de la intermediación, en favor de técnicas de optimización de índices de validación de agrupamiento como la modularidad.
2. Estudiar la posibilidad de eliminar el paso de recálculo del algoritmo de Girvan y Newman ya que este aumenta la complejidad del método impidiendo su uso en problemas grandes.
3. Proponer una nueva medida que no tenga los problemas identificados.
4. Evaluar la posibilidad de crear un nuevo algoritmo de agrupamiento, a partir de una matriz de similitud, basado en la nueva propuesta.
5. Investigar las posibilidades de la nueva propuesta en un problema de agrupamiento de documentos que utiliza similitud coseno. Comparar los resultados con los obtenidos con otros algoritmos aplicados al agrupamiento de documentos.

2.2 La intermediación diferencial en el agrupamiento en redes complejas

Se escogió el algoritmo clásico GN como punto de partida para el desarrollo de las ideas que aquí se presentan. Aunque existen muchas mejoras de este algoritmo, los elementos fundamentales están en él. Como fue descrito en el Capítulo 1, GN se basa en la intermediación de las aristas y es proporcional al número de caminos más cortos entre pares de vértices que fluyen a través de éstas. Este algoritmo es lento, necesita tiempo $O(m^2n)$, donde m

es el número de aristas y n es el número de nodos. Como se muestra en el Anexo 6, el paso 3 es crítico y adiciona otro orden de complejidad con respecto a m, el cuál significa, en el peor de los casos, el aumento de la complejidad en n^2. Sin embargo, Newman escribió en [29]:

> "De hecho, parece que la característica más importante del algoritmo es el paso de recálculo, en lo que respecta a obtener resultados satisfactorios. Como se mencionó anteriormente, nuestros estudios indican que, una vez que uno considera seriamente la idea de utilizar la intermediación para pesar las aristas, la medida exacta empleada parece no influenciar altamente los resultados. El paso de recálculo, por otra parte, es absolutamente crucial para el funcionamiento de nuestros métodos y no existía en intentos anteriores de resolver el problema de agrupamiento por medio de algoritmos divisivos. La ausencia de este paso dio lugar en aquel entonces a resultados verdaderamente pobres, fallando al encontrar la estructura en comunidades conocida incluso en el más sencillo de los casos."

Una de las contribuciones de este trabajo es presentar una perspectiva diferente sobre el asunto anterior, concretamente, se le ha dado más importancia a las propiedades de la medida utilizada que al paso de recálculo de la intermediación. En esta sección se propone una nueva medida de intermediación que es menos sensible al recálculo. Esta medida, llamada intermediación diferencial, presenta propiedades que permiten crear mejores algoritmos de agrupamiento. Se presentará una comparación de la propuesta con resultados del cálculo de la intermediación según Girvan y Newman usando los caminos más cortos.

2.2.1 Intermediación diferencial

Una cuestión fundamental relacionada con los problemas presentados por los métodos basados en el concepto de intermediación tiene que ver con la distinción entre la noción objetiva de intermediación y lo que es posible medir de ella. Se entiende por noción objetiva a la propiedad que tienen algunas aristas de "mediar" entre comunidades con independencia de que se pueda o no medir esa capacidad –también llamada intermediación real– adecuadamente. La comparación que se ha realizado con un proceso de medición es totalmente intencional. Esta conduce a una reflexión sobre la calidad del instrumento de medición: si no se puede medir lo que se quiere, no se debe esperar que las inferencias con estas mediciones sean correctas. Al medir la intermediación de una arista, las aristas vecinas y algunas no tan vecinas pueden constituir factores de ruido. Este efecto es el que se trata de evitar en el algoritmo GN con la eliminación y el recálculo. Así, se puede decir que existe una dependencia entre las mediciones de los valores de la intermediación de las aristas. Esta dependencia es compleja y

ni siquiera el proceso de eliminación puede garantizar ciento por ciento que se arribe a una estimación correcta.

En esta investigación se hace referencia a la independencia del proceso de medición y no a la independencia de los valores reales de intermediación. Idealmente, se quiere que la habilidad para medir la intermediación de una arista no dependa de las otras aristas, ni siquiera de la topología del grafo. El análisis del segundo tipo de independencia, que sí tiene que ver mucho con la topología del grafo, permite realizar también otra reflexión importante sobre como debe ser una buena medida de intermediación. Los valores de intermediación de las aristas son independientes de los valores del resto de las aristas del grafo dados los valores de una cierta vecindad de las mismas. Esta propiedad –markoviana– permite que las mediciones puedan hacerse localmente, con los consiguientes beneficios en términos de eficiencia de los algoritmos.

Las consideraciones anteriores son el punto de partida, conceptual y metodológico, de la nueva medida de intermediación que ahora se introduce formalmente. Se emplearán las definiciones y notación establecidas en el Anexo 2 respecto a la Teoría de Grafos. Se parte de un grafo no dirigido $G=(V, E)$, donde $V=\{1,\ldots,n\}$ representa el conjunto de nodos correspondiente a cada objeto a agrupar y $E \subseteq V \times V$ es el conjunto de sus aristas indicando las relaciones entre los objetos. Siempre se considerarán los geodésicos entre pares de nodos que sean alcanzables, es decir, que estén en una misma componente conexa y una condición necesaria para obtener los geodésicos entre un par de nodos a y b que pasen por una arista $i-j$ es que a, b, i y j también sean alcanzables, pertenecientes a una misma componente conexa. La notación $\|a -_s b\|$ indica la longitud de los caminos de longitud mínima entre los nodos a y b distintos y alcanzables en el grafo y $\|a -_s b \mid i-j\|$ expresa la longitud de los caminos de longitud mínima, entre los nodos a y b distintos, que pasan por la arista $i-j$, donde a, b, i y j pertenecen a una misma componente conexa. Las definiciones que a continuación se presentarán son válidas para grafos no ponderados y ponderados, para estos últimos se puede reemplazar el término longitud por costo. Para presentar las definiciones con mayor claridad se asumirá el trabajo con grafos no ponderados, la implementación para el caso ponderado se abordará en el epígrafe 2.4.

La definición (2.1) introduce la localidad en este enfoque. Los resultados aquí alcanzados son válidos con otras definiciones de *c*-vecindad.

Definición 2.1 (*c*-vecindad). Una *c*-vecindad de la arista $i - j$ en el grafo $G=(V, E)$, con c positivo, es el subgrafo $G_{c,\, i-j}$ inducido por $V_{c,\, i-j}$, donde:

$$V_{c,i-j} = \left\{ v \in V \,\middle|\, \|v -_s i\| \le c \vee \|v -_s j\| \le c \right\} \tag{2.1}$$

Definición 2.2 (diferencial geodésico). Sean a, b, i y j nodos del grafo $G=(V, E)$, donde i y j son nodos adyacentes, y a y b son nodos diferentes. La función ζ se define como sigue:

$$\zeta(i, j \,|\, a, b) = \|a -_s b \,|\, i - j\| - \|a -_s b\| \tag{2.2}$$

La ecuación (2.2) denota el geodésico diferencial del par de nodos a y b con respecto a la arista $i - j$. Éste alcanza valores no negativos en su dominio de definición y el valor mínimo cuando el geodésico entre a y b pasa a través de $i - j$. La definición 2.3 es la clave para introducir la intermediación diferencial.

Definición 2.3 (λ-intermediación). Se llama λ-intermediación de la arista $i - j$ dado el par de nodos a y b a la expresión (2.3)

$$B_\lambda(i, j \,|\, a, b) = e^{-\lambda \zeta(i, j \,|\, a, b)} \tag{2.3}$$

donde la función $\zeta(i, j \,|\, a, b)$ sigue la definición 2.2.

Para λ positivo se cumple que $0 < B_\lambda(i, j \,|\, a, b) \le 1$ y $B_\lambda(i, j \,|\, a, b)$ alcanza el máximo cuando los caminos más cortos entre a y b pasan por $i - j$.

Definición 2.4 (Intermediación diferencial). Sea $G=(V, E)$ un grafo dado. La magnitud $DB_{\lambda,c}(i, j)$ dada por la expresión (2.4)

$$DB_{\lambda,c}(i, j) = \sum_{a,b \in V_{c,i-j}} B_\lambda(i, j \,|\, a, b) = \sum_{a,b \in V_{c,i-j}} e^{-\lambda \zeta(i, j \,|\, a, b)} \tag{2.4}$$

se llama Intermediación Diferencial de la arista $i - j$ en la *c*-vecindad $V_{c,\, i-j}$ con parámetro λ.

La concepción de la Intermediación Diferencial es más bien general, siendo en efecto, la expresión (2.4) sólo una de sus posibles implementaciones. En la medición de la intermediación de un puente se desea disminuir el efecto de sus puentes vecinos, por eso se

sustituyó el conteo de caminos por la ponderación de sus longitudes relativas. Así, al calcular la λ-intermediación se desea que la función a utilizar alcance el máximo cuando el diferencial geodésico sea cero y preferiblemente que el máximo sea uno. Las razones para seleccionar una función con esas características son: la intermediación clásica cuenta los caminos y la experiencia de su uso en diversas aplicaciones muestra que su funcionamiento es lógico, si dada una arista *i-j* el diferencial de todas las parejas de nodos *a* y *b* fuese cero (único puente entre dos grupos) entonces la intermediación diferencial sólo contará la cantidad de caminos y coincidirá con el valor del cálculo original de la intermediación, y, de esta forma, se realiza una extensión de la idea clásica que mejora sus problemas, pero siempre tomando la clásica como punto de partida. Adicionalmente, al calcular la λ-intermediación se necesita una variación monótona de la expresión de la intermediación con respecto al diferencial geodésico, en algunas variantes de cálculo para grafos ponderados se requiere, además, continuidad. Finalmente, se desea utilizar una función positiva. Estas fueron las causas por las que se escogió la función exponencial para el cálculo de la λ-intermediación –alcanza el máximo igual a uno cuando el valor de la variable es cero, es continua, monótona, positiva y asintótica a la abscisa (diferencial geodésico). Cualquier otra función que cumpla estas propiedades pudo haber sido utilizada.

2.2.2 Intermediación diferencial en la red Zachary

El objetivo de esta sección es presentar un ejemplo donde la intermediación diferencial (DB) capture más información sobre la estructura de la red que la intermediación de los caminos más cortos (GN) [55]. Se tomó como punto de referencia la famosa red del Club de Karate Zachary[24] [222]. Esta red se muestra en la Figura 2.1.

Se calcularon tres valores de intermediación para cada arista $i-j$. $DB_{0.01,\,2}(i,j)$, $DB_{0.01,\,4}(i,j)$ y $GN(i,j)$. Para este ejemplo, c=4 utiliza todo el grafo como la c-vecindad de las aristas. Las listas con los valores de intermediación fueron ordenadas decrecientemente. Además de los valores de intermediación, la Tabla 2.1 muestra la posición P, de los puentes entre comunidades acorde al orden mencionado.

[24] Disponible en UCINet (http://vlado.fmf.uni-lj.si/pub/networks/data/UciNet/UciData.htm)

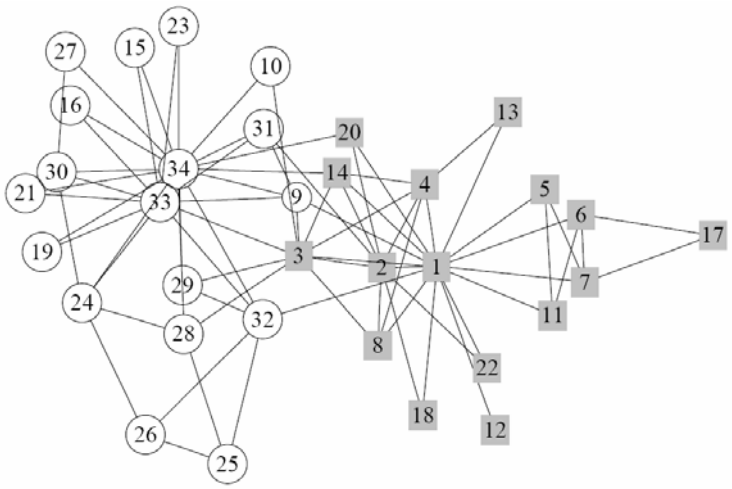


Figura 2.1 Red del Club de Karate Zachary dividida en dos comunidades, fuente [29]. Este grafo tiene 34 vértices, 78 aristas y dos comunidades conectadas por 10 puentes.

Tabla 2.1 Valores de intermediación para los puentes de la red Zachary.

Puente	*P*	DB_2	*P*	DB_4	*P*	*GN*	Puente	*P*	DB_2	*P*	DB_4	*P*	*GN*
3-9	2	396.37	13	455.82	5	41.65	3-33	22	276.03	49	290.89	60	8.33
3-29	14	317.03	37	330.83	57	10.47	3-10	23	275.51	42	317.40	66	5.15
3-28	16	313.39	39	327.21	75	2.37	2-31	24	255.78	38	329.76	74	2.54
1-9	18	301.29	44	301.29	76	1.89	20-34	25	253.83	50	273.63	24	19.49
1-32	20	295.02	48	295.02	65	5.5	14-34	28	231.93	52	249.78	35	16.61

En la Tabla 2.1 es posible observar que cuando se calcula la intermediación GN las posiciones de los puentes están esparcidas a lo largo del intervalo [1,78]. El puente con menor intermediación GN ocupa la posición 76. Además, el 60% de los puentes ocupan las últimas 18 posiciones. Esto evidencia por qué no es posible usar esta medida para descubrir la estructura de la red: para eliminar todos los puentes es necesario eliminar todas las aristas. En otras palabras, es posible encontrar puentes con altos y bajos valores de intermediación. Sólo el recálculo puede ayudar en esta situación. La lista DB muestra una situación diferente. Tomando, por ejemplo, la DB con c=2 se observa que el último puente ocupa la posición 28. Así, el problema está resuelto eliminando las 28 aristas con los mayores valores de intermediación diferencial; esta es una evidencia de que la propuesta no requiere el recálculo.

La DB con c=4 es claramente peor que con c=2 pero aún mucho mejor que el caso GN. Esto es muy conveniente considerando que el costo computacional depende del tamaño de la vecindad. El caso c=1, que no fue incluido en la Tabla 2.1, no fue tan malo como se esperaba.

Este ejemplo ha permitido ilustrar que DB es una medida que tiene mayor habilidad que GN para obtener una buena medición de la centralidad de un puente en la situación (común) en la que hay más de una arista entre un par de comunidades dadas. La intermediación DB no sólo está considerando los caminos más cortos, como lo hace la intermediación GN, sino también algunos cortos. El significado que se considera de "camino corto" es controlado por el parámetro λ. En otras palabras, no negociará valores de intermediación entre los puentes paralelos de dos comunidades. Se conoce que la intermediación GN negocia valores entre puentes paralelos, por eso el algoritmo que la usa requiere el paso de recálculo después de cada eliminación. La medida propuesta realiza una mejor estimación del valor real de intermediación de una arista sin ser significativamente afectado por sus puentes vecinos.

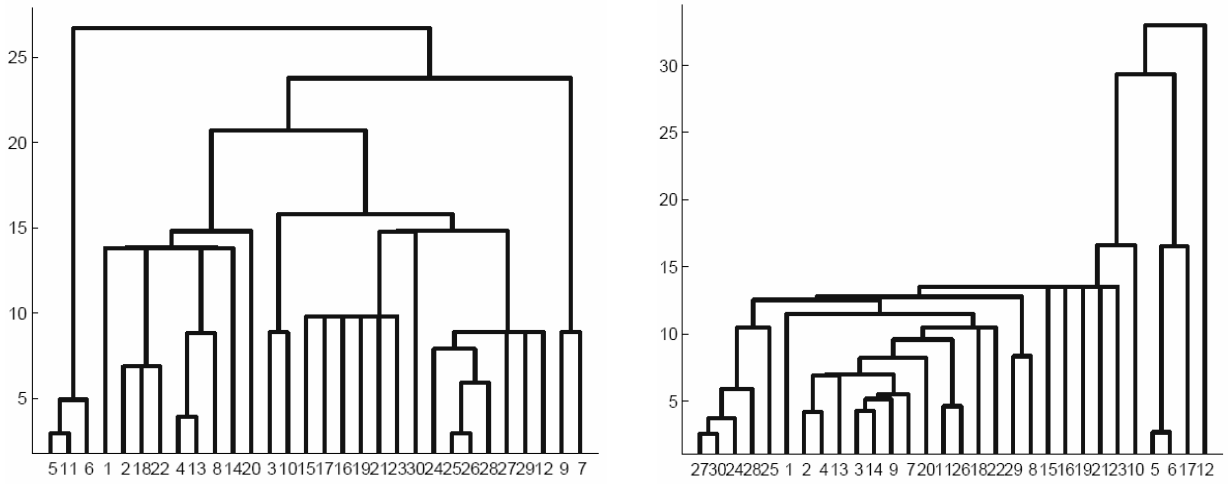


Figura 2.2 Dendrograma obtenido de la red Zachary usando las matrices de intermediación como medidas de disimilitud. Izquierda) Diferencial con c=1, λ=0.01. Derecha) Camino más corto. Los coeficientes de correlación cohenetic son 0.44 y 0.32, respectivamente.

La intermediación diferencial también puede ser considerada una medida de disimilitud topológica. Un valor alto de $DB_{\lambda,c}(i, j)$ significa que los vecinos de i están probablemente conectados a los vecinos de j a través de la arista $i - j$ (la mayoría de los caminos cortos entre pares de vértices en la c-vecindad de $i - j$ pasan a través de esta arista). Inversamente, una intermediación diferencial pequeña sugiere que los vértices tienen muchos vecinos en común. Este análisis está soportado por la intuición siguiente: cosas similares tienen cosas (los vecinos) en común. Para mostrar la afirmación anterior se calcularon los dendrogramas, con el método de enlace simple, de la red Zachary usando las matrices de intermediación como

medidas de disimilitud. La Figura 2.2 muestra que la intermediación DB capta mucho mejor el agrupamiento natural del problema que la intermediación GN.

2.3 Cálculo de la intermediación diferencial

El cálculo de la intermediación diferencial puede hacerse con varios métodos. La vecindad de una arista $i - j$ donde residen las parejas de nodos a y b diferentes, según la expresión (2.4), puede ser diferente de la vecindad donde residen los caminos que conectan estos nodos.

El primer método de cálculo utiliza la misma vecindad de la arista para determinar los caminos entre los nodos que residen en la vecindad. Básicamente la idea es la siguiente: 1) se determinan los nodos que pertenecen a la c-vecindad de la arista $i - j$; 2) se extrae el subgrafo inducido por el conjunto de nodos pertenecientes a la c-vecindad de $i - j$; y 3) se calcula la intermediación diferencial de la arista $i - j$ en este subgrafo. La técnica para determinar qué nodos pertenecen a la c-vecindad de una arista puede variar. En los experimentos realizados en esta tesis se ha utilizado la idea de incluir todos los nodos que se pueden alcanzar desde los extremos de la arista en una cantidad de pasos que no supere un umbral dado; es decir, con caminos de longitud inferior al umbral especificado, siguiendo fielmente la definición 2.1.

El segundo método busca los caminos en un grafo que contiene los subgrafos de las vecindades. En particular, este grafo pudiera ser el grafo original, lo que sin pérdida de generalidad, se va a asumir en las explicaciones de este epígrafe. La idea fundamental de este segundo método consiste en obtener toda la información necesaria para calcular la matriz de intermediación diferencial a partir de la matriz de caminos más cortos. Esta última matriz, si el grafo no es ponderado, se calcula con complejidad $O(mn)$ haciendo una búsqueda primero a lo ancho, donde m es el número de aristas y n el número de nodos. En el caso ponderado se utiliza el algoritmo de Dijkstra, que tiene complejidad $O(mn+n^2\log n)$. Por razones que serán explicadas en la sección 2.3, ahora sólo se analizará el caso no ponderado. La siguiente proposición es clave para desarrollar el método de cálculo.

Proposición 2.1. Dado $G=(V, E)$, si $i - j \in E$, entonces $\forall a \in V$

$$\zeta(i, j \mid a, j) = 0 \quad \Rightarrow \quad \zeta(i, j \mid a, i) > 0 \tag{2.5}$$

Prueba. Un camino que cumple con el antecedente de la implicación tiene que estar formado por la concatenación de cualquier geodésico p_k: $a -_s i$ con la arista $i - j$, de lo contrario no fuese mínimo. Por otra parte, $j \in p_k$ viola el antecedente por lo que $j \notin p_k$. La prueba termina notando que $\|a -_s i \mid i - j\| = \|a -_s i\|$ implica la existencia de un camino $a -_s i \mid j$ de costo mínimo lo que contradice $j \notin p_k$. □

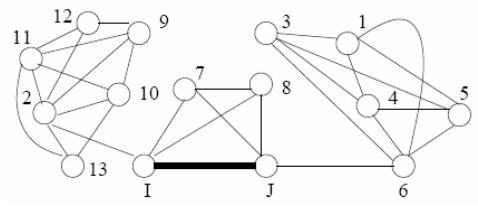


Figura 2.3 Ejemplo de vecindad de la arista $i - j$.

Según esta proposición cada arista define una tri-partición del conjunto de nodos del grafo. La igualdad $\zeta(i, j \mid a, j) = 0$ indica que el camino más corto entre a y j pasa por i; es decir, significa que el nodo a está más cerca de i que de j. Se pueden, entonces, distinguir sólo tres casos: 1) cuando a no está más cerca de ninguno de los extremos de la arista $i - j$; 2) cuando a está más cerca de i; 3) cuando a está más cerca de j. Se dirá que a es un nodo de tipo 0, 1 ó 2, respectivamente. Al analizar la contribución de una pareja de nodos a y b a la intermediación diferencial de la arista $i - j$ se tienen cuatro casos posibles: A) ambos del tipo 0; B) uno del tipo 0 y otro de otro tipo; C) ninguno del tipo 0, pero diferentes; D) iguales pero no del tipo 0. En la Figura 2.3 la pareja de nodos (7, 8) pertenece al caso A, (7, 3) al caso B, la pareja (3, 9) al caso C y las parejas $\{(1, 5), (9, 13)\}$ al caso D.

La clasificación presentada y la matriz de caminos más cortos, permiten calcular de manera fácil la intermediación diferencial de una arista. Sea *SP* una matriz, tal que *SP*(*i*,*i*)=0 y *SP*(*i*,*j*) es la longitud del camino más corto entre *i* y *j* en el grafo dado.

1. Dada la cuarteta <*a*,*b*,*i*,*j*>, la *c*-vecindad de $i - j$ se calcula como:

$$V_{c,i-j} = \{a \mid SP(a,i) \le c\} \cup \{b \mid SP(b,j) \le c\} \tag{2.6}$$

2. Determinación del tipo de nodo.

$$tipo(a) = 0 \Leftrightarrow SP(a,i) = SP(a,j) \tag{2.7}$$

$$tipo(a) = 1 \Leftrightarrow SP(a, j) = SP(a, i) + 1 \tag{2.8}$$

$$tipo(a) = 2 \Leftrightarrow SP(a, i) = SP(a, j) + 1 \tag{2.9}$$

3. Cálculo del diferencial geodésico: $\zeta(i, j \mid a, b) = M - SP(a, b) + 1$.

   a. Caso A: $M = \min(SP(a, i) + SP(b, j), SP(a, j) + SP(b, i))$

   b. Casos B y C: $M = SP(a, i) + SP(b, j)$

   donde *tipo*(*a*)=1, *tipo*(*b*)=0 o *tipo*(*a*)=0, *tipo*(*b*)=2 o *tipo*(*a*)=1, *tipo*(*b*)=2.

El caso D, aunque pudiera incluirse, en esta investigación se considera que aporta poco al cálculo de la intermediación, además su cálculo es costoso, razones que han provocado que aquí no se emplee. De hecho todos los experimentos realizados no lo incluyen. Esto quiere decir que si se tiene en cuenta fielmente a la definición de DB dada en la sección 2.2.1, el desconocimiento del caso D es equivalente a obtener un estimado de la misma. Otra forma de ver el problema es redefinir el concepto de intermediación diferencial de tal manera que este caso quede formalmente excluido. En esta tesis se va a seguir este camino.

Excluir el caso D garantiza en el peor caso un tiempo $O(n^2)$ para el cálculo de la DB de una arista, cuando la vecindad tiene n nodos. Esto es posible comprobarlo a partir del análisis de la clasificación introducida arriba. Se considera que es necesario reflexionar sobre la complejidad algorítmica de la DB y la importancia de haber eliminado el paso de recálculo.

Tanto la intermediación DB como GN calculan la matriz de los caminos más cortos en un grafo dado. Para grafos no ponderados este cálculo se realiza con una búsqueda primero a lo ancho y tiene complejidad $O(mn)$. El algoritmo GN repite este proceso m veces debido al recálculo, el cuál consume $O(m^2n)$. La matriz DB puede ser calculada en el peor de los casos en un tiempo $O(mn^2)$ usando la matriz de los caminos más cortos calculada. Así, la complejidad total es $O(mn)+O(mn^2)=O(mn^2)$. Es posible apreciar que para grafos densos, la variante propuesta gana un orden de complejidad. Sin embargo, para grafos dispersos ambos algoritmos tienen $O(n^3)$. La igualdad para el caso disperso es falsa porque las constantes son muy diferentes. El cálculo de la DB involucra m iteraciones de $O(n^2)$ operaciones de aritmética simple con la matriz de caminos más cortos, mientras que GN necesita muchas operaciones en el grafo en cada iteración.

Es posible calcular la DB en un tiempo $O(mn)$ si se fija el tamaño de la c-vecindad, $V_{c,i-j}$. Asumiendo que la vecindad tiene un máximo de K nodos, se obtiene $O(mn)+O(mK^2)=O(mn)$ o $O(n^2)$ para grafos dispersos. En este caso, los geodésicos son calculados en todo el grafo, pero la DB de cada arista se estima en su vecindad. Alternativamente, es posible calcular los geodésicos en la vecindad, consumiendo el tiempo $O(K^3)$ y $O(K^2)$ para subgrafos (vecindades) densos y dispersos, respectivamente. Por consiguiente, bajo ciertas condiciones especiales es posible alcanzar una complejidad temporal lineal: $O(m(K^3+K^2))=O(m)$ o $O(n)$ para grafos dispersos. Afortunadamente, las condiciones son naturales porque muchas redes del mundo real cumplen la propiedad small-world, que es justo lo que se necesita para estos cálculos.

2.4 Intermediación diferencial y similitud

La medida de intermediación diferencial presentada, también es útil para manejar casos cuando las aristas tienen pesos asociados. En este capítulo, es de interés resolver problemas de agrupamiento de vectores a partir de los cuales se construye una matriz de similitud. Por eso se ha seguido la propuesta presentada en [28]. Newman propuso calcular la intermediación de todas las aristas en el grafo ponderado siguiendo la expresión clásica, ignorando los pesos, y entonces dividir cada uno de los valores de intermediación por el peso de la arista correspondiente en el grafo original. Este algoritmo es sencillo, tan rápido como la versión no ponderada original (adiciona solamente la operación de división por el peso de la arista), y funciona excelentemente en la detección de comunidades incluyendo el recálculo. Esta propuesta puede ser utilizada también con la intermediación diferencial, incluso con mejores resultados que los que obtuvo Newman, ya que se heredan las buenas propiedades de la medida. Es posible aplicar otras generalizaciones para el caso de redes ponderadas, pero en este capítulo sólo se aborda este enfoque. Por tanto, se define la intermediación diferencial en un grafo pesado como sigue:

Definición 2.5 (Intermediación diferencial pesada). La intermediación diferencial de una arista en un grafo ponderado con una medida de similitud, se obtiene de la división del valor de intermediación diferencial obtenido asumiendo que el grafo es no ponderado (todos los pesos tienen valor uno) por el peso de la arista en el grafo ponderado original.

2.4.1 La intermediación diferencial y la similitud Coseno

Al aplicar la intermediación diferencial en dominios textuales, se puede utilizar una representación en grafos donde la interacción entre dos documentos (peso de las aristas) se exprese en función de cuán similares son. Zonas de valores altos de similitud implica nodos muy interconectados (alto grado). Generalmente, documentos en un mismo grupo tienen mayor similitud que documentos en grupos diferentes.

La similitud Coseno ha sido ampliamente utilizada para comparar documentos que son sometidos a un proceso de agrupamiento. Sin embargo, existen colecciones donde dos documentos pueden tratar temas diferentes y tener alta similitud Coseno. Estos documentos no son similares a sus vecinos respectivos y sólo el uso de la similitud no logra identificarlos en grupos diferentes. En tales casos, se hace necesario explotar la estructura topológica del grafo, de forma tal que se puedan detectar estructuras ocultas a partir de la representación gráfica y del uso adecuado de los enlaces entre los objetos.

En este epígrafe se ilustra la utilidad de la intermediación diferencial en la detección de los puentes entre grupos, al grupar documentos donde la similitud Coseno expresa las interrelaciones entre ellos. Para ello, se conformó una colección a partir de los artículos publicados en BioMed Central[25], donde los primeros doce artículos científicos tratan de fibrosis cística y los siete últimos abordan el tema de terapia génica. Uno de los criterios de selección fue la alta interrelación que existe entre estos temas, de forma tal que algunos artículos científicos que abordan temas diferentes tienen alta similitud Coseno. El preprocesamiento textual aplicado a esta colección coincide con el descrito en el epígrafe 2.5.

En el Anexo 12 se muestran los resultados del estudio acerca de los puentes entre grupos, tanto considerando la similitud como la intermediación diferencial. En la Tabla A12.1 aparecen cinco columnas principales. La primera se divide en cuatro subcolumnas donde P indica la posición según orden creciente de la similitud Coseno de las aristas que enlazan los documentos I y J pertenecientes a grupos diferentes, las dos columnas siguientes muestran los vértices extremos de dichas aristas puentes y finalmente V indica el valor de la similitud Coseno entre los documentos I y J. De las cuatro columnas restantes, las dos primeras

[25] BioMed Central, artículos sobre Bioinformática y Medicina. http://www.biomedcentral.com/info/about/datamining/

muestran resultados del cálculo de la intermediación diferencial pesada, mientras que las dos últimas no consideran los pesos en el cálculo. Para cada caso se realizó el estudio considerando los caminos de longitud dos y longitud tres al conformar las vecindades de las aristas. Cada una de estas columnas también se divide en cuatro subcolumnas donde se muestra la posición P en orden decreciente de la intermediación diferencial de las aristas puentes con vértices extremos I y J y el valor de la intermediación diferencial.

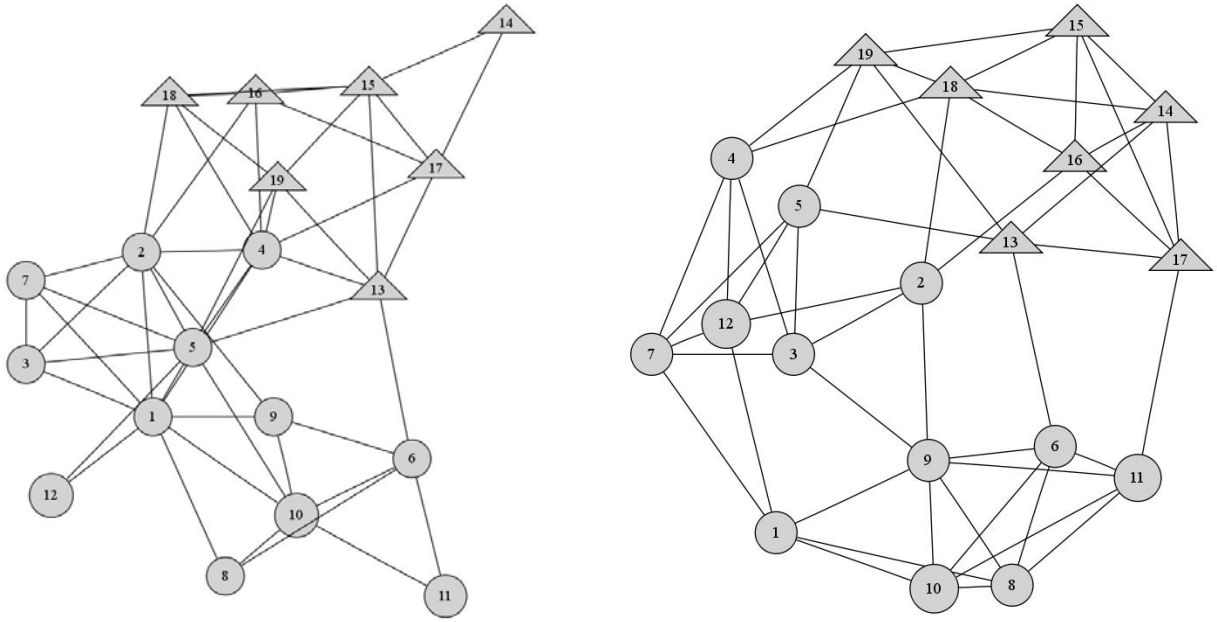


Figura 2.4 Grafo de similitud asociado a la colección perteneciente a BioMed (izquierda) Corte según umbral igual a 0.21 (derecha) Corte considerando los K vecinos más cercanos según la similitud (K=5).

Los primeros resultados que se muestran en la Tabla A12.1 partieron de la construcción del grafo de similitud a partir de los K vecinos más cercanos, y los últimos resultados, a partir del corte por umbral de similitud. En la Figura 2.4 se muestran dos de los grafos conformados, uno considerando cinco vecinos más cercanos, y el otro, realizando un corte por el umbral de similitud igual a 0.21. En la Tabla A12.1 es posible observar que sólo considerando la similitud Coseno es difícil detectar los puentes entre grupos, generalmente quedan puentes que tienen una de las últimas posiciones respecto al número total de aristas, ya sea por una u otra vía de construcción del grafo a partir de la matriz de similitud. Por tanto, en un proceso de eliminación de aristas sólo considerando similitud Coseno, sería necesario eliminar prácticamente todas las aristas para eliminar los puentes y se perdería la estructura de comunidad del grafo. Sin embargo, la intermediación diferencial logra captar propiedades

topológicas y reflejarlas en el cálculo de la centralidad de las aristas, incluso cuando se trabaja de forma no ponderada. Los grafos no ponderados reflejan en su topología la ponderación, porque al construirlos, ya sea por K vecinos más cercanos o por corte de umbral, se tuvo en cuenta la similitud Coseno entre los documentos. El cálculo de la intermediación diferencial logra ubicar a los puentes en las primeras posiciones en un orden decreciente de intermediación; por tanto, eliminando las primeras aristas es posible detectar los grupos. En la Figura A12.2 y en la Figura A12.3 se reflejan las potencialidades de la intermediación diferencial en la identificación de los puentes con el objetivo de obtener buenos agrupamientos a partir de la eliminación sucesiva de aristas con alta intermediación. Sin embargo, es posible observar en la Figura A12.1 que si sólo se utiliza la similitud Coseno, el dendrograma que se obtiene dista de la clasificación de referencia para la colección textual en estudio.

Este ejemplo ilustra la utilidad del cálculo de la intermediación diferencial en el agrupamiento de documentos relacionados según la similitud Coseno. Esta forma del cálculo de la centralidad de las aristas descubre los puentes entre grupos porque, a diferencia de la similitud Coseno, logra explotar las propiedades topológicas del grafo. Además, se observa resultados similares para el cálculo de la intermediación pesada y no pesada.

2.4.2 Un algoritmo de agrupamiento a partir de la matriz de similitud

En esta sección se propone un algoritmo de agrupamiento basado en el concepto de intermediación diferencial.

Algoritmo 1. Un algoritmo simple de agrupamiento basado en la intermediación diferencial.

1. Obtención del grafo de similitud.
2. Cálculo de la matriz de intermediación diferencial pesada.
3. Estimación de las aristas a eliminar.
4. Determinación de los núcleos del agrupamiento mediante la extracción de componentes conexas.
5. Clasificación de los nodos que no pertenecen a los núcleos.

A continuación se describe cada paso del algoritmo propuesto.

PASO 1. Existen varias formas de obtener un grafo que modele las relaciones más importantes que existen en el problema. Una forma simple es eliminar de la matriz de similitud todas las aristas que no superen un umbral de similitud dado. Otra, es asociar con

cada nodo los K vecinos más cercanos. Así, es necesario determinar un parámetro de entrada. Mientras más aristas tenga el grafo más información sobre el problema original tendrá, pero mucho mayor será el costo de procesamiento. Para realizar el corte por el umbral se puede tener en cuenta el histograma de frecuencias de las similitudes y realizar el corte de forma tal que se obtenga una matriz dispersa.

PASO 2. La matriz de intermediación diferencial pesada es cuadrada y su orden coincide con el de la matriz de adyacencia correspondiente al grafo de similitud. Para cada una de las aristas del grafo de similitud (aristas no eliminadas en el paso 1) se calcula la intermediación diferencial según la definición 2.5 y este valor se asigna a la celda correspondiente a dicha arista en la matriz de intermediación. Aquí los parámetros de entrada son: λ para regular la λ-intermediación y c para especificar el orden de la c-vecindad. En la mayoría de los experimentos se ha utilizado λ=0.01 porque garantiza una variación discreta de la λ-intermediación con respecto al diferencial geodésico (decrece aproximadamente en pasos de 0.01 cuando el diferencial geodésico se va incrementando en uno). El parámetro c por defecto es tal que siempre se toma todo el grafo como vecindad de las aristas, aquí se asume que se parte de un grafo conexo. Si el grafo no es conexo, el algoritmo es válido para cada componente conexa; por tanto, la vecindad de una arista $i - j$ incluirá por defecto todos los nodos alcanzables desde i y desde j.

PASO 3. Eliminar del grafo las P aristas con mayor valor de DB. La intermediación diferencial trata de que todos los puentes ocupen las primeras posiciones de la lista decreciente de valores de DB. La idea es reducir la cantidad de aristas no-puentes a eliminar o al menos tratar de saber cuántas pueden eliminarse sin afectar los resultados del agrupamiento.

PASO 4. Se dice núcleo del agrupamiento a todas las componentes conexas del grafo obtenido al eliminar las aristas en el paso anterior, siempre y cuando éstas cumplan un conjunto determinado de condiciones previamente fijadas. Intuitivamente, el núcleo contiene aquellos elementos de los cuales se está seguro que pertenecen a un grupo en particular. Por ejemplo, cuando se crean cubrimientos es menos probable que los elementos del núcleo sean compartidos entre grupos. Hay que tener en cuenta cuán fuertes son las condiciones exigidas para clasificar como núcleo. Éstas pueden ser: las K componentes más grandes, todas las que tengan al menos K nodos y criterios de densidad o entropía, entre otras. Si se seleccionan las K

componentes más grandes, en cierto modo se le da al algoritmo una idea del número de grupos que se espera. A veces esto es útil, pero por lo general es rechazado por la mayoría de los autores como una deficiencia de los algoritmos. Sin embargo, la condición que establece la identificación como núcleo de aquellas componentes que tengan al menos K nodos no da ninguna información con respecto al número de grupos. Esta es la condición utilizada en los experimentos realizados. En muchos problemas este valor (mínimo tamaño del núcleo) puede prefijarse de antemano sin dificultad, para estos casos el parámetro no es necesario.

PASO 5. Asociar a los núcleos detectados aquellos nodos que no fueron incluidos en alguno en el paso anterior. La unión de un núcleo con los nodos que se le asociaron determina un grupo. En el esquema que se ha utilizado, cada nodo se asocia al grupo que contiene mayor cantidad de sus vecinos en el grafo de similitud. Una variante consiste en utilizar un grafo de similitud menos restrictivo. Por ejemplo, si para obtener el grafo de similitud se realizó un corte en la matriz de similitud de 0.18, se puede hacer la clasificación en un grafo con corte 0.1. La clasificación se puede hacer de muchas maneras, pero es importante notar que se puede hacer en orden $O(n)$.

2.5 Intermediación diferencial y agrupamiento de documentos

Un caso especial de lo discutido en la sección anterior es el problema del agrupamiento de documentos cuando la relación entre éstos se codifica en una matriz de similitud. En este epígrafe se presenta un estudio empírico que muestra que el algoritmo de agrupamiento que utiliza la intermediación diferencial como un elemento esencial del mismo, tiene un comportamiento comparable a los algoritmos del estado del arte en esta área. Incluso su desempeño es superior para algunos problemas.

2.5.1 Problemas a estudiar

Se han conformado tres corpus de documentos (BioMed, Reuters, CEC2006) que permiten mostrar la aplicabilidad de la intermediación diferencial y el Algoritmo 1 propuesto que la utiliza en el agrupamiento en dominios textuales. En la Tabla 2.2 se describen estos corpus.

BioMed fue extraído de la colección de libre acceso BioMed Central. De la fuente original se seleccionaron dos tópicos relacionados con el SIDA, el primero abordando el tema desde el punto de vista social y epidemiológico, y el segundo desde el punto de vista de sus estudios

mediante la Biología Molecular. La relación que existe entre los tópicos hace que documentos que abordan a uno y a otro sean similares. Esta situación, intencionalmente provocada, influye en que la detección de los grupos no sea trivial.

Tabla 2.2 Descripción general de los corpus textuales y su clasificación de referencia.

Corpus	Cantidad de documentos	Cantidad de grupos	Distribución de documentos por grupos
BioMed	31	2	$Grupo_1$=[1..11] $Grupo_2$=[12..31]
Reuters	29	2	$Grupo_1$=[1..12] $Grupo_2$=[13..29]
CEC2006	29	2	$Grupo_1$=[1..18] $Grupo_2$=[19..29]

Reuters se formó a partir de la colección de noticias de la agencia Reuters[26]. Siguiendo el criterio anterior para la selección de los tópicos, y considerando que en esta colección es fácil identificar grupos de documentos que abordan tópicos muy diferentes, se seleccionaron aleatoriamente noticias que tratan sobre el suministro de dinero (money-supply) y el tipo de cambio de monedas (money/foreign exchange).

CEC2006 está formada por un subconjunto de los resúmenes de artículos científicos presentados en el evento CEC 2006[27]. Estos resúmenes se encuentran multiclasificados por expertos; no obstante, se identificaron documentos que muestran resultados en el área de la optimización combinatoria y numérica, y otro grupo, que incluye aquellos que presentan aplicaciones del mundo real. Tanto en la conformación de este corpus, como en los restantes ya descritos, después de identificados los tópicos, la selección de los documentos fue aleatoria.

Se seleccionó la representación espacio-vectorial (Vector Space Model; VSM) [223] para representar los documentos. El software CorpusMiner[28] [224], desarrollado como parte de esta investigación, contribuyó al preprocesamiento de las colecciones textuales y permitió la aplicación a cada colección de las cuatro etapas de la representación textual [2]: transformación del corpus, extracción de términos, reducción de la dimensionalidad, y normalización y pesado de la matriz. Finalmente, se seleccionaron los mejores 600 términos que caracterizaron cada corpus textual y se obtuvo el grafo de similitud a partir de la matriz de similitud coseno entre los documentos para cada una de las colecciones consideradas. En el

[26] Colección Reuters-21578 disponible en el sitio web de David D. Lewis http://www.research.att.com/~lewis
[27] IEEE Congress on Evolutionary Computation, suministrada por Nees Jan van Eck y Rudolf Kruse
[28] Sistema para la representación, agrupamiento y resumen de corpus textuales. Registro CENDA: 2333-2005.

epígrafe 4.3 se describe en detalles como se realiza el preprocesamiento textual en esta investigación.

2.5.2 El algoritmo propuesto y el problema BioMed

Para construir el grafo de similitud se realiza un corte en la matriz de similitud a partir de un umbral dado, estimado a partir de los histogramas de frecuencias de las similitudes e intentando obtener matrices dispersas. En el esquema de clasificación que se ha utilizado en los experimentos, cada nodo se asocia al grupo que contiene mayor cantidad de sus vecinos en el grafo de similitud. Se exige que los núcleos tengan al menos cuatro nodos, lo que es una restricción suave para la mayoría de los problemas. En esta sección se utiliza λ=0.01.

En el primer experimento se usa una *c*-vecindad de tamaño dos. Los resultados se muestran en la Figura A11.1. Cada una de los gráficos (tomados en orden de filas) corresponden a los resultados de agrupamientos utilizando grafos de similitud con 69, 87, 124 y 155 aristas (con 6, 11, 20 y 29 puentes), obtenidos con umbrales 0.2, 0.18, 0.14 y 0.12, respectivamente. Cada gráfico de la Figura A11.1 muestra curvas correspondientes a índices de validación. Se incluyó la medida externa OFM porque están disponibles las clasificaciones de referencia para cada colección. Se muestran curvas correspondientes al índice interno modularidad, ya que es uno de los más utilizados en las investigaciones en redes complejas. Se calculó la modularidad de la partición definida por las componentes conexas obtenidas después de eliminar las primeras *X* aristas (eje de las abscisas) y la modularidad del agrupamiento obtenido. A pesar de que este índice tiene gran importancia en la validación no supervisada de estos tipos de métodos de agrupamiento, en este ejemplo su desempeño no es muy bueno, al menos si se compara con los valores de la medida externa utilizada. Por tal motivo, y considerando que un resultado importante para esta tesis es el comportamiento de las medidas internas, se incorporaron las medidas precisión y calidad generalizadas que siguen exactamente la forma de la curva OFM (al menos para los dos primeros casos), lo que no ocurre con los índices de modularidad. Estas dos últimas medidas son de gran utilidad porque, sin considerar la clasificación de referencia, logran tener comportamientos similares a índices externos. Esta es una característica fundamental cuando se evalúan resultados de agrupamientos en dominios textuales, donde generalmente no se tiene conocimiento adicional del dominio. En el Capítulo 3 se presentarán éstas y otras medidas basadas en RST con tales características.

Lo primero que reflejan las gráficas es que el algoritmo se comporta bien en este problema, se alcanza 0.9 en el índice OFM, e incluso, con los umbrales 0.2 y 0.18, se llega a 0.97. En la próxima sección se muestran comparaciones con otros métodos. Los resultados son mejores en los dos primeros casos y se deterioran al disminuir el umbral. En esta investigación se considera que este problema está determinado fundamentalmente por el aumento del número de puentes y no tanto por el aumento del número de aristas.

Para un rango grande de valores de umbral [0.12, 0.2] se pueden encontrar buenos agrupamientos (altos valores de OFM). Los buenos valores forman mesetas por lo que es posible considerar un amplio rango en la determinación del número de aristas a eliminar. Por ejemplo, en el segundo caso se puede eliminar desde la decimonovena hasta la trigésimoquinta aristas y en todos estos casos se obtiene OFM igual a 0.97.

En el segundo experimento se estudió el efecto del tamaño de la *c*-vecindad en el comportamiento del algoritmo. Se utilizó el segundo grafo de similitud del ejemplo anterior pero se varió el tamaño de la *c*-vecindad: 1, 2, 3 y 4. Los resultados se observan en la Figura A11.2. Las medidas RST siguen comportándose bien. El desempeño del algoritmo es bueno al variar *c* y también ofrece un amplio margen para determinar la cantidad de aristas a eliminar.

Con las otras colecciones textuales se realizaron análisis similares al de BioMed, pero para no hacer tedioso el desarrollo de este epígrafe, se muestra el estudio comparativo en la próxima sección, donde se observa que el algoritmo se comporta bien con los otros conjuntos de datos.

2.5.3 Desempeño de otros algoritmos y estudio comparativo

En este subepígrafe se mostrará cómo se comportan, sobre las tres colecciones descritas previamente, algoritmos ampliamente utilizados en el agrupamiento de documentos y el algoritmo GN [55]. Finalmente se mostrará el estudio comparativo con el Algoritmo 1 propuesto.

Se seleccionaron los algoritmos SKWIC [64], ES [106] y sus mejoras ACONS [103] y GStar [104], el algoritmo Enlace y el algoritmo GN por ser un antecesor de la propuesta que se hace en esta investigación. El primero genera una partición en el conjunto de datos, las variantes de Estrella (ES, CONS y GStar) un cubrimiento, Enlace construye una jerarquía de los objetos de forma aglomerativa y GN divisivamente forma una jerarquía de los objetos. Se utilizaron las

implementaciones que facilita CorpusMiner para los algoritmos SKWIC y ES, las implementaciones de ACONS y GStar facilitadas por investigadores del CENATAV[29], las variantes del algoritmo Enlace (linkage) incluidas en Matlab[30] y la implementación del algoritmos GN incluida en el software GARLucene (descrito en el Capítulo 4). Todos utilizaron la similitud coseno para comparar los documentos.

Los documentos pertenecientes a las colecciones conformadas están etiquetados y es útil hacer uso de esa clasificación para comparar los resultados del agrupamiento respecto a la clasificación de referencia. Por eso, como criterios para la validación de los resultados de los algoritmos de agrupamiento se utilizaron las medidas externas precisión, cubrimiento y OFM [178], descritas en el Anexo 8.

Para cada uno de los algoritmos utilizados se buscaron las mejores configuraciones de parámetros de forma tal que se obtuvieran buenos valores para las medidas externas consideradas. Al ser ES, GStar y ACONS algoritmos que dependen del umbral de similitud β_0, se evalúa cada resultado del agrupamiento a partir de modificaciones del umbral. Se consideraron los histogramas de frecuencias para la selección de los umbrales y se refinaron los valores de los umbrales en los entornos de altos valores de las medidas de validación. Así, se seleccionó aquel corte que arrojó los mejores valores de estos índices. Adicionalmente, se cubrió un amplio espectro de valores de los umbrales, variándolos desde los pequeños que permitan un alto cubrimiento hasta obtener altos valores de precisión, con un incremento de 0.02. Por su parte, SKWIC inicia su procesamiento con la generación aleatoria de los centros de los grupos y requiere que se especifique inicialmente el número de grupos a obtener. Por tal motivo, para este algoritmo siempre se realizaron 10 corridas y se promediaron los valores resultantes. Siempre se especificó el número de grupos a obtener coincidente con la clasificación de referencia, con el objetivo de propiciar el mejor desempeño del algoritmo al suministrar los valores de los parámetros. Las condiciones de terminación en el proceso iterativo de reajuste de los grupos fueron: estabilización de los centros indicada por un 99% de

[29] Agradecimientos especiales a José E. Medina Pagola y a Airel Pérez Suárez por facilitar la documentación e implementación para utilizar los algoritmos ACONS y GStar.
[30] The Language of Technical Computing Matlab Versión 7.0

coincidencia de los mismos entre una iteración t y una iteración t+1, y se consideró 20 el número máximo de iteraciones.

El algoritmo Enlace produce jerarquías de los objetos, que varían en función del criterio para la unión de grupos que se utilice. Los criterios considerados son: simple (single), completo (complete), promedio (average), pesado (weighted) y varianza mínima (ward). Se realizaron cortes en la jerarquía especificando el número máximo de grupos a obtener según la clasificación de referencia de cada colección. Para el algoritmo GN se consideraron varias configuraciones iniciales mediante la eliminación de aristas a partir de umbrales de corte. Los umbrales de corte se calcularon a partir de los histogramas de frecuencias. Además, al igual que con las variantes Estrella, se cubrió un amplio espectro de umbrales de corte para analizar el comportamiento del algoritmo. Para cada corte, el algoritmo GN construye una jerarquía de los objetos. Se seleccionó una partición resultante de un corte en la jerarquía que alcanzara buenos valores de los índices de validación.

Al aplicar variantes del algoritmo Estrella, el máximo valor de OFM se alcanza cuando existe un equilibrio entre precisión y cubrimiento, ya que para valores pequeños de β_0 el algoritmo crea muchos grupos con alta precisión y bajo valor de la medida cubrimiento, a medida que el umbral aumenta, aumenta el tamaño de los grupos y por tanto la precisión de los mismos disminuye, pero aumenta su cubrimiento respecto a las clases de referencia. La correspondencia entre precisión, cubrimiento y OFM es variable al validar los resultados del algoritmo GN. Los umbrales utilizados para aplicar este algoritmo establecen los cortes en la matriz de similitud inicial. La Tabla A13.1 muestra la media ($\bar{x}$) y la desviación estándar (s) de las 10 corridas de SKWIC para los cuatro corpus estudiados.

BioMed. En la Tabla A13.1 se muestra el valor promedio de OFM para SKWIC (0.776). Este valor indica que los resultados de SKWIC distan de la clasificación de referencia; no obstante, este algoritmo alcanzó en la novena corrida el valor 0.868. Esta fluctuación en los resultados se debe a la inicialización aleatoria de los centros de los grupos a obtener. Es posible observar en la Tabla A13.3 que los algoritmos ES, ACONS y GStar tienen un mejor comportamiento al agrupar esta colección, alcanzando el valor 0.866 para OFM, descubriendo nueve, seis y cinco grupos, respectivamente. El algoritmo GStar tiene un comportamiento ligeramente superior porque se acerca a la cantidad de grupos a obtener según la clasificación de referencia. Se

aprecia en la Tabla A13.2 que el algoritmo Enlace logra resultados similares a SKWIC para esta colección. Por otra parte, el algoritmo GN, al formar cuatro grupos, alcanza valores de OFM (0.937) que superan los resultados del resto de los algoritmos mencionados.

Reuters. En la Tabla A13.1 se muestra la media de OFM para SKWIC (0.664), incluso el máximo alcanzado (0.722) es bajo. Una de las causas de estos resultados es que las noticias son muy pequeñas y las similitudes entre aquellas en un mismo grupo o en grupos diferentes tienen pocas diferencias; marcando la diferencia con la colección anterior que está compuesta por el texto completo de artículos científicos. Las tres variantes de Estrella logran el máximo valor de OFM (0.876) alcanzado para esta colección cuando descubren tres grupos. Es posible observar en la Tabla A13.2 que el algoritmo Enlace alcanza resultados ligeramente desfavorables para esta colección (0.822), aunque supera los resultados de SKWIC. El máximo valor de OFM se alcanzó con el criterio Completo. La Tabla A13.6 refleja que los resultados alcanzados por el algoritmo GN (0.932) supera el OFM producido por las tres variantes del algoritmo Estrella.

CEC2006. Este corpus textual está compuesto por documentos muy pequeños y existe baja similitud entre ellos, lograr un agrupamiento coincidente con la clasificación de referencia es difícil. En la Tabla A13.1 se muestra la media de OFM (0.668) y el máximo valor alcanzado por esta medida en las diez corridas (0.692). La Tabla A13.7 refleja que ES, GStar y ACONS tienen un comportamiento superior a SKWIC, logrando el máximo valor de OFM igual a 0.744. GStar y ES alcanzan el máximo cuando obtienen cuatro grupos, mientras que ACONS lo logra formando tres grupos. El mejor resultado para esta colección se muestra en la Tabla A13.2 y lo logra el algoritmo Enlace con el criterio de varianza mínima, alcanzando el valor 0.830 de OFM. El algoritmo GN arroja resultados desfavorables respecto a la clasificación de referencia para esta colección, el valor más alto de OFM obtenido es 0.672 y lo logra formando cuatro grupos. Usando el umbral de corte 0.10 se obtienen tres componentes conexas y las aristas puentes representan más del 30% del total de las aristas. No obstante, este algoritmo nunca tiene la más baja puntuación en las comparaciones establecidad.

La Figura A13.4 permite establecer la comparación entre los algoritmos estudiados y el Algoritmo 1 propuesto, respecto a los valores máximos de OFM alcanzados por los algoritmos para cada colección. BioMed y Reuters lograron valores máximos similares, CEC2006 tuvo

resultados un tanto desfavorables respecto a las dos colecciones anteriores. El algoritmo propuesto logra valores superiores de OFM, respecto a los algoritmos utilizados en el estudio comparativo, para las colecciones textuales BioMed y Reuters, y resultados comparables con el algoritmo Enlace para la colección CEC2006.

2.6 Declaración de resultados

Se declaran los siguientes resultados como los aportes más importantes de este capítulo.

1. Introducción del concepto de intermediación diferencial en el campo del análisis de redes complejas y de agrupamiento de datos.
2. Se mostró que en los algoritmos basados en el concepto de intermediación, la elección de la medida es crucial, siendo más importante que el paso de recálculo. Hasta el momento, exactamente lo contrario es ampliamente aceptado por la comunidad científica.
3. Presentación de un algoritmo de agrupamiento basado en el concepto de intermediación diferencial para manejar grafos ponderados y no ponderados.
4. Aplicación del método de la intermediación diferencial a un problema de la minería de textos.

2.7 Conclusiones parciales

Se ha presentado una nueva medida para evaluar el grado de intermediación que tiene una arista en un grafo. Esta medida tiene propiedades que la distinguen como un elemento a considerar en los algoritmos de detección de comunidades y agrupamiento en general: 1) es adecuada para redes ponderadas y no ponderadas, 2) no necesita el paso del recálculo, 3) capta mejor las propiedades topológicas y es menos sensible al ruido que las medidas de intermediación existentes, y 4) es una medida de disimilitud topológica. Esta nueva forma de medición consideró otro enfoque para el cálculo de la intermediación, a partir de la extracción eficiente de la información local presente en las redes.

Se creó un nuevo algoritmo de agrupamiento basado en el cálculo de la intermediación diferencial de las aristas a partir de una matriz de similitud entre objetos. Este algoritmo muestra las posibilidades del nuevo enfoque para el cálculo de la intermediación y la factibilidad de producir nuevos algoritmos. Este algoritmo requiere la especificación de

parámetros, aquí se ha mostrado que para algunos valores de los parámetros el algoritmo obtiene buenos resultados.

Se mostró que el algoritmo propuesto tiene un buen desempeño en problemas de agrupamiento de documentos partiendo de una matriz de similitud coseno entre los textos a agrupar. Los resultados fueron comparados con aquellos producidos por los algoritmos SKWIC, ES, GStar, ACONS, Enlace y GN. Para las tres colecciones textuales consideradas, el algoritmo basado en la intermediación diferencial obtuvo resultados comparables y en la mayoría superiores a los alcanzados por los algoritmos citados. Estos resultados se deben esencialmente a las potencialidades de la intermediación diferencial para capturar las propiedades topológicas de los grafos y determinar las aristas puentes, elementos de gran utilidad en situaciones donde la similitud coseno entre los documentos no lo permite.

3 Conjuntos aproximados para valorar agrupamientos

Las medidas existentes para validar el agrupamiento no siempre dan criterios certeros, principalmente en dominios textuales donde frecuentemente se carece de la clasificación de referencia. Cada medida no puede captar todas las propiedades deseadas del agrupamiento, y mientras más propiedades se evalúen mejor será la valoración que se tenga del mismo. En este capítulo se propone la aplicación de RST para validar el agrupamiento permitiendo medir su precisión, calidad y consistencia, sin requerir conocimiento del dominio. Se crearon nuevas medidas internas basadas en RST y un algoritmo que permite su aplicación para validar el agrupamiento. Se conformaron casos de estudio utilizando archivos de datos generales y en dominios textuales para mostrar la confiabilidad y validez de la propuesta. Finalmente, se sugiere el uso de las aproximaciones inferiores de los grupos para obtener sus documentos más representativos, mostrando otra aplicación de RST en el post-agrupamiento.

3.1 Fundamentos teóricos

La teoría de conjuntos aproximados fue introducida por Z. Pawlak en 1982 [225]. La descripción más general es que se basa en aproximar cualquier concepto, un subconjunto duro del dominio (por ejemplo, una clase de un problema de clasificación o un grupo resultante de un proceso de agrupamiento), por un par de conjuntos exactos llamados aproximación inferior y superior del concepto aproximado.

Los conjuntos aproximados consideran que a todo objeto *x* de un universo *U* está asociada una cierta cantidad de información, expresada por medio de algunos atributos que describen el objeto [226, 227]. La estructura de información básica de esta teoría es el sistema de información; par (*U, A*) donde $A = \{a_1, a_2, ..., a_m\}$ es el conjunto de atributos y *U* es un conjunto no vacío llamado universo de objetos descritos usando los atributos a_i [226][31].

Los objetos que tienen la misma descripción son inseparables (similares) con respecto al conjunto *B* de atributos considerados; $B \subseteq A$. Esta relación de inseparabilidad constituye la base matemática de la teoría, es una relación de equivalencia e induce una partición del

[31] Esta definición es independiente a la definición de sistema de información de Shannon

universo U en bloques de objetos inseparables (similares) [226]. Cualquier subconjunto X (concepto) del universo U se puede expresar en términos de estos bloques de forma exacta o aproximada. La vaguedad es una propiedad de los conceptos y puede ser atribuida a los límites del conjunto, mientras que la incertidumbre es una propiedad de los elementos del concepto y tiene que ver con la pertenencia o no a éste [228]. Cuando un concepto es vago, los elementos del universo no pueden ser identificados con certeza como elementos del concepto.

Algunas extensiones de la teoría clásica de los conjuntos aproximados no requieren que se cumpla la transitividad ni la simetría, tales como las relaciones llamadas de tolerancia o similitud. La extensión de RST clásico a relaciones de similitud R'_B acepta que objetos que no son inseparables pero sí suficientemente cercanos o similares puedan pertenecer a la misma clase [229]. En el Anexo 3 se muestran algunas de las medidas de similitud entre objetos que pueden utilizarse para formar las relaciones.

El objetivo de esta extensión de la teoría es construir relaciones de similitud R'_B a partir de relaciones de inseparabilidad, relajando las condiciones iniciales de éstas. No obstante a la flexibilización, si R_B es una relación de inseparabilidad definida en U, R'_B es una relación de similitud extendida de R_B sí y sólo sí $\forall x \in U$, $R_B(x) \subseteq R'_B(x)$ y $\forall x \in U$, $\forall y \in R'_B(x)$, $R_B(y) \subseteq R'_B(x)$, donde $R'_B(x) = \{y \in U : y\ R'_B\ x\}$; es decir, $R'_B(x)$ es la clase de similitud de x.

R'_B no tiene que ser necesariamente simétrica, aunque la mayoría de las definiciones de similitud usualmente lo son. No se impone, tampoco, que R'_B sea transitiva. El único requerimiento es la reflexividad. R'_B puede ser siempre vista como una extensión de la relación de inseparabilidad trivial R_B definida por $R_B(x)=\{x\}$, $\forall x \in U$. En RST extendida a relaciones de similitud, cada objeto puede pertenecer a más de una clase de similitud, por lo que el cubrimiento inducido por R'_B sobre U no es necesariamente una partición. En esta investigación es de interés considerar todos los atributos que caracterizan los objetos, ya que ellos fueron sometidos a un proceso de reducción de la dimensionalidad previo a la aplicación de esta teoría, así $B=A$ y se excluye B de la notación.

La aproximación de un concepto $X \subseteq U$, usando una relación de inseparabilidad R, ha sido inducida mediante los conjuntos llamados aproximaciones R-inferior y R-superior de X [230]. En las expresiones (3.1) y (3.2) se muestran sus formas de cálculo a partir de cualquier

relación reflexiva *R*'; las aproximaciones *R*'-inferior ($R'_*(X)$) y *R*'-superior ($R'^*(X)$), respectivamente [231].

$$R'_*(X) = \{x \in X : R'(x) \subseteq X\} \tag{3.1}$$

$$R'^*(X) = \bigcup_{x \in X} R'(x) \tag{3.2}$$

La expresión (3.3) muestra el cálculo de la región límite o frontera (*BN*(*X*)) de *X* para la relación *R*' considerando las expresiones (3.1) y (3.2).

$$BN(X) = R'^*(X) - R'_*(X) \tag{3.3}$$

Si el conjunto *BN* es vacío entonces el conjunto *X* es exacto respecto a la relación *R*'. En caso contrario, $BN(X) \neq \varnothing$, el conjunto *X* es inexacto o aproximado con respecto a *R*'.

Usar relaciones de similitud permite representar naturalmente varios problemas y además ofrece mayores posibilidades para la construcción de las aproximaciones; sin embargo, al trabajar en un espacio mayor, resulta más complejo computacionalmente buscar las aproximaciones relevantes [232].

3.2 Conjuntos aproximados para validar el agrupamiento

La validación del agrupamiento utilizando RST es no supervisada y permite que el cálculo común inicial de las relaciones y aproximaciones inferiores y superiores pueda ser reutilizado por varias medidas de calidad, inclusión y proximidad de conceptos.

Los objetos a agrupar están descritos por rasgos y constituyen el sistema de información que es el punto de partida para aplicar RST. Cada grupo resultante de un proceso de agrupamiento al cual dichos objetos son sometidos constituye un concepto X_i. Sólo se considera en esta investigación resultados de agrupamientos duros y deterministas, donde los conceptos forman una partición. No obstante, la teoría puede ser aplicada a la evaluación de cubrimientos.

Para cada objeto agrupado se calcula el conjunto de objetos relacionados con él, siguiendo la relación definida por la expresión (3.4), donde $s(x, y)$ retorna un valor de similitud entre los objetos *x* e *y*, y ξ es el umbral de similitud a considerar.

$$R'(x)=\{y\in U : yR'x, \text{ es decir } y \text{ está relacionado con } x \text{ si y sólo si } s(x,y)>\xi\} \quad (3.4)$$

La forma de medir la similitud y qué umbral utilizar para formar los conjuntos de relaciones depende del dominio donde fue aplicado el agrupamiento, cómo fueron descritos los objetos y qué nivel de granularidad se desea evaluar en los resultados. En el Anexo 10 se muestran posibles variantes para el cálculo del umbral. Adicionalmente, se calculan las aproximaciones inferiores y superiores de cada grupo (concepto) usando (3.1) y (3.2), respectivamente.

A partir del cálculo de las aproximaciones inferiores y superiores por grupos, se propone validar el agrupamiento y cada grupo aplicando las medidas ofrecidas por RST para evaluar los conceptos definidos sobre sistemas de información. Éstas permiten tener una noción de la proximidad de los conceptos y pueden ser aplicadas en varios esquemas de razonamiento [233]. En [234-236] se presentan trabajos previos de la autora siguiendo este enfoque.

Una medida que permite validar cada concepto es la precisión de la aproximación [237]. Un concepto aproximado X puede ser caracterizado numéricamente por el coeficiente (3.5) llamado precisión de la aproximación, donde $|Y|$ denota la cardinalidad de un conjunto Y finito y no vacío. Obviamente, $0\leq\alpha(X)\leq1$. Si $\alpha(X)=1$, X es duro (exacto), si $\alpha(X)<1$, X es aproximado (vago, inexacto), siempre respecto al conjunto de atributos considerado [238].

$$\alpha(X)=\frac{\left|R'_*(X)\right|}{\left|R'^*(X)\right|} \quad (3.5)$$

La calidad de la aproximación es otra medida que permite evaluar conceptos [237]. El coeficiente (3.6) expresa el porciento de objetos que pueden ser correctamente asignados a X. Además, $0\leq\alpha(X)\leq\gamma(X)\leq1$, y $\gamma(X)=0$ si $\alpha(X)=0$, mientras $\gamma(X)=1$ si $\alpha(X)=1$ [238].

$$\gamma(X)=\frac{\left|R'_*(X)\right|}{\left|X\right|} \quad (3.6)$$

Las medidas precisión y calidad de la aproximación están asociadas a cada concepto, por tanto, ofrecen una valoración local de cada grupo obtenido. Sin embargo, en muchos casos es necesario medir la calidad y la precisión como un todo considerando el sistema de información

y los conceptos X_1, …, X_k definidos sobre él, con k total de conceptos. El coeficiente calidad del agrupamiento[32], expresión (3.7), describe la inexactitud de los conceptos, expresando la proporción de los objetos que pueden estar correctamente asignados a los grupos en el sistema. Si ese coeficiente es uno, el sistema de información según los conceptos definidos es consistente, en otro caso es inconsistente [230].

$$\Gamma = \frac{\sum_{i=1}^{k} \left| R'_{*}(X_i) \right|}{\left| U \right|} \tag{3.7}$$

La precisión del agrupamiento[33] expresa las posibles asignaciones correctas a grupos. En la expresión (3.8) se observa que su esencia es mostrar la proporción entre la cantidad de objetos que pudieran estar bien agrupados y la cantidad de objetos que pudieran o no pertenecer a los grupos del sistema de información [239].

$$\mathrm{A} = \frac{\sum_{i=1}^{k} \left| R'_{*}(X_i) \right|}{\sum_{i=1}^{k} \left| R'^{*}(X_i) \right|} \tag{3.8}$$

Si bien las medidas calidad y precisión del agrupamiento logran medir globalmente el nivel de inconsistencia, calidad y precisión de los conceptos en un sistema de información dado, consideran que cada grupo tiene igual repercusión en la evaluación. Sin embargo, no todos los grupos deben tener igual influencia al evaluar el agrupamiento, por tanto, se desea una ponderación de los mismos. Por ejemplo, en el Anexo 8 se observa que las medidas entropía y OFM ponderan los grupos por su cardinalidad, igual sucede con las medidas conectividad parcial pesada y densidad esperada mostradas en el Anexo 9. Por tal motivo, en esta investigación, inicialmente, se obtuvieron expresiones generalizadas de precisión y calidad del agrupamiento considerando la ponderación de los grupos por su cardinalidad, calculando el peso w_i asociado al grupo X_i como $w_i = \left|X_i\right| / \left|U\right|$ [234, 235]. En trabajos posteriores se

[32] Nombrado en la literatura calidad de la aproximación de la clasificación o también calidad de la clasificación. En la literatura utilizan la palabra clasificación porque asumen que los conceptos coinciden con clases de un atributo de decisión. El uso en este trabajo es sobre los conceptos asociados a cada grupo resultante de un proceso de agrupamiento.

[33] Nombrado en la literatura precisión de la aproximación de la clasificación o también precisión de la clasificación.

consideraron otras formas de ponderación [236, 240]. Aquí se proponen la calidad y precisión generalizadas del agrupamiento, expresiones (3.9) y (3.10), respectivamente. El peso asociado a un grupo X_i se representa por w_i, cumpliéndose las restricciones $w_i \geq 0$ y $\sum_{i=1}^{k} w_i = 1$.

$$\Gamma_G = \frac{\sum_{i=1}^{k} \left(\left| R'_*(X_i) \right| \cdot w_i \right)}{|U|} \tag{3.9}$$

$$\mathrm{A}_G = \frac{\sum_{i=1}^{k} \left(\left| R'_*(X_i) \right| \cdot w_i \right)}{\sum_{i=1}^{k} \left(\left| R'^*(X_i) \right| \cdot w_i \right)} \tag{3.10}$$

Varios criterios pueden ser empleados para ponderar los grupos y así captar mejor las propiedades deseadas; por ejemplo, similitud dentro del grupo, pertenencia de los objetos al grupo y cardinalidad del grupo.

Una forma de medir la pertenencia de un objeto a un grupo es la función de pertenencia aproximada [241]. En la expresión (3.11) se muestra que ésta cuantifica el grado de solapamiento relativo entre $R'(x)$ y el concepto al cual x pertenece. Esta función se interpreta como una estimación basada en frecuencias de la probabilidad condicional de que el objeto y pertenezca al conjunto X, dados los valores del objeto x con respecto al conjunto de atributos. El valor $\mu_X(x)$ mide el grado de inclusión del objeto x en el grupo X [50, 238].

$$\mu_X(x) = \frac{|X \cap R'(x)|}{|R'(x)|} \tag{3.11}$$

La media de la pertenencia aproximada de los objetos a cada grupo también puede ser empleada para ponderar los grupos [236, 240, 242, 243]. Su expresión se muestra en (3.12). Sin embargo, esta ponderación puede fallar en algunos casos. En el Anexo 14 se presentan ejemplos que muestran deficiencias del cálculo de la pertenencia aproximada.

$$w_i = \frac{\sum_{x \in X_i} \mu_{X_i}(x)}{|X_i|} \tag{3.12}$$

Utilizando los datos del Ejemplo A14.1 se calculó el grado de pertenencia según la expresión (3.11) y se obtuvo que $\mu_{\text{Grupo1}}(x)=\mu_{\text{Grupo2}}(x)=3/6=0.5$. Sin embargo, el objeto x tiene mayor pertenencia al segundo grupo que al primero, y la expresión no es capaz de captarlo porque esta forma de calcular la pertenencia sólo tiene en cuenta en el denominador la cardinalidad del conjunto de objetos relacionados con x.

Considerando las situaciones donde falla la expresión (3.11) se propone, como resultado de esta investigación, otra expresión para medir la pertenencia aproximada de un objeto x a un grupo X, cuantificando en qué grado la clase de similitud de x ($R'(x)$) cubre el grupo X. Las expresiones (3.13) y (3.14) muestran su forma de cálculo y la variante de ponderación que se deriva de ella, respectivamente. Esta nueva propuesta ya ha sido utilizada en [236, 242, 243] con el nombre función de compromiso aproximado.

$$\nu_X(x) = \frac{\left|X \cap R'(x)\right|}{\left|X\right|} \tag{3.13}$$

$$w_i = \frac{\sum_{x \in X_i} \nu_{X_i}(x)}{\left|X_i\right|} \tag{3.14}$$

Con esta forma de cálculo de la pertenencia, expresión (3.13), se obtiene para los datos del Ejemplo A14.1 que $\nu_{\text{Grupo1}}(x)=3/25=0.12$ y $\nu_{\text{Grupo2}}(x)=3/5=0.6$, indicando que x tiene una mayor pertenencia al segundo grupo. No obstante, existen situaciones donde esta variante falla, tal es el caso del

Ejemplo **A14.2**. Utilizando los datos de este ejemplo se calculó el grado de pertenencia según la expresión (3.13) y se obtuvo que $\nu_{\text{Grupo1}}(x)=10/20=0.5$ y $\nu_{\text{Grupo2}}(x)=5/10=0.5$. Sin embargo, el objeto x tiene mayor pertenencia al primer grupo que al segundo, y la expresión no es capaz de captarlo porque esta forma de cálculo de la pertenencia sólo tiene en cuenta en el denominador la cardinalidad del grupo. Al calcular la pertenencia utilizando la expresión (3.11) se obtienen los valores $\mu_{\text{Grupo1}}(x)=10/20=0.5$ y $\mu_{\text{Grupo2}}(x)=5/20=0.25$, indicando que el objeto x tiene una mayor pertenencia al Grupo_2.

A partir del análisis de las situaciones donde ambas variantes para el cálculo de la pertenencia de los objetos a los grupos no permiten cuantificar adecuadamente la representatividad de los

objetos en los mismos, en esta tesis se presenta otra propuesta que combina las dos formas de cálculo anteriores. La expresión (3.15) muestra la nueva forma de cálculo de la pertenencia aproximada y la expresión (3.16) su uso para ponderar los grupos.

$$\varpi_X(x) = \frac{|X \cap R'(x)|}{|X \cup R'(x)|} \tag{3.15}$$

$$w_i = \frac{\sum_{x \in X_i} \varpi_{X_i}(x)}{|X_i|} \tag{3.16}$$

Al aplicar la expresión (3.15) a los datos del Ejemplo A14.1 se obtuvo que $\varpi_{\text{Grupo1}}(x)$=3/28=0.11 y $\varpi_{\text{Grupo2}}(x)$=3/8=0.38, indicando que *x* tiene una mayor pertenencia al segundo grupo, similar a los valores alcanzados con la expresión (3.13). Al aplicar la expresión (3.15) a los datos del Ejemplo A12.2 se obtuvo que $\varpi_{\text{Grupo1}}(x)$=10/30=0.33 y $\varpi_{\text{Grupo2}}(x)$=5/25=0.2, indicando que *x* tiene una mayor pertenencia al primer grupo, similar a los valores alcanzados con la expresión (3.11). Esta última variante de cálculo de la pertenencia de los objetos a los grupos incluye en el análisis tanto la representatividad de los objetos en los grupos, como el grupo en los objetos relacionados con él.

Es útil emplear otras formas de ponderación, por ejemplo, pesar los grupos considerando su cohesión y densidad puede hacer la evaluación más precisa y cercana a la realidad. Una medida que se puede utilizar para ponderar la precisión y la calidad de los grupos es la similitud global [178], o la varianza de las similitudes entre los objetos en los grupos (en esta medida se desea una minimización). Otras variantes también se pueden considerar; por ejemplo, si el agrupamiento a evaluar fue realizado sobre objetos representados en un grafo, las propiedades de éste pueden influir en la ponderación. Las expresiones que se proponen para la evaluación tienen un aporte adicional, porque permiten evaluar integrando varios conceptos mediante el uso de medidas internas ya definidas o utilizando propiedades y relaciones entre los datos.

A partir de los conceptos de RST a aplicar y las nuevas medidas definidas, se propone el Algoritmo 2 para guiar la aplicación de RST en la validación del agrupamiento. Las entradas a este algoritmo son: la colección de objetos (sistema de información), el resultado del

agrupamiento (conceptos), la medida y umbral de similitud, y formas de ponderación de los grupos. Las salidas del algoritmo son los valores de las medidas de precisión y calidad aplicadas a los grupos y al agrupamiento en general.

Algoritmo 2. Aplicación de RST en la validación del agrupamiento.

1. Obtener las clases de similitud (3.4) para cada objeto en el sistema de información.
2. Calcular las aproximaciones inferiores (3.1) y superiores (3.2) por grupo.
3. Calcular calidad (3.5) y precisión (3.6) por grupo.
4. Calcular calidad (3.7) y precisión (3.8) del agrupamiento.
5. Para cada variante de cálculo de peso especificada:
   a. Calcular los pesos por grupos.
   b. Calcular calidad (3.9) y precisión (3.10) generalizadas del agrupamiento.

De esta forma es posible medir la vaguedad o imprecisión de cada grupo obtenido y del agrupamiento en su totalidad. Si la región límite es pequeña, entonces se obtendrán mejores resultados de las medidas utilizadas en la evaluación (valores cercanos a uno). Valores altos de las medidas indican un mejor agrupamiento.

Los resultados de los pasos 1 y 2 del Algoritmo 2 son comunes para la aplicación posterior de cualquier medida de validación basada en RST. En el cálculo de la complejidad temporal de estos dos pasos interviene la complejidad de la medida de similitud que se emplee. Por tal motivo, y considerando que el algoritmo se aplica en dominios textuales, se fija para el estudio la similitud Coseno, ampliamente utilizada en estos dominios. En el cálculo de la complejidad se considera: n número de objetos, m número de rasgos que describen dichos objetos, y k número de grupos. La complejidad temporal del primer paso es $O(mn^2)$, la que está dada por el cálculo de la matriz de similitud y la obtención del conjunto de objetos relacionado con cada objeto. La complejidad temporal del cálculo de las aproximaciones es $O(kn^2)$, ya que para el peor caso se considera que a cada grupo pertenecen todos los objetos y cada objeto está relacionado con todos. Por tanto, la complejidad de este procesamiento, previo al cálculo de las medidas basadas en RST, es $O(wn^2)$, donde $w = \max\{k, m\}$. Teniendo en cuenta que, en los problemas de minería de textos m es mucho mayor que k, se asume que la complejidad temporal es $O(mn^2)$.

3.3 Confiabilidad y validez de las medidas basadas en RST para validar el agrupamiento

La evaluación de medidas de validación del agrupamiento es una tarea ardua. Para chequear la confiabilidad y validez de los resultados se han diseñado experimentos, aplicados posteriormente a los casos de estudio que se definen, que permiten un análisis sobre la evaluación de los conjuntos aproximados como instrumento de medición.

3.3.1 Definición de casos de estudio y herramientas utilizadas

El uso de RST para validar mediante la determinación local y global de la precisión, calidad e inconsistencia de los resultados de agrupamientos es posible en diversos dominios. Es por eso que se definieron casos de estudio que recopilan archivos provenientes de diferentes áreas de aplicación. Especialmente se definió un caso de estudio en dominio textual, para mostrar la aplicabilidad de la propuesta cuando es necesario validar resultados de agrupamientos de colecciones de documentos, principal motivación de este trabajo.

El primer caso de estudio definido constituye una recopilación de archivos de datos, internacionalmente utilizados para evaluar agrupamiento y en general algoritmos de aprendizaje automático, donde aparecen objetos de diversa naturaleza descritos por rasgos en su mayoría numéricos, aunque también hay presencia de rasgos simbólicos en algunos de ellos. La selección, para este primer caso de estudio, ha incluido conjuntos de datos con la presencia o no de valores ausentes y no ha sido un requerimiento la presencia de rasgos objetivos. En la Tabla A15.1 se muestra la descripción y la fuente de cada uno de los archivos de datos que conforman el primer caso de estudio. El conjunto de archivos descritos en la Tabla A15.2 constituye el segundo caso de estudio definido. Todos conjuntos de datos constan de rasgos predictores y objetivos, por tanto existe la clasificación de referencia para cada uno de ellos. Por último, el tercer caso de estudio consta de 50 corpus textuales que han sido conformados a partir de la selección aleatoria de noticias de la colección de la agencia Reuters[34], incluyendo los tópicos que abordan las noticias.

[34] Colección Reuters-21578 disponible en el sitio web de David D. Lewis http://www.research.att.com/~lewis

La herramienta de aprendizaje automatizado Weka[35] y el sistema CorpusMiner [107] fueron utilizados para obtener los resultados experimentales a partir de los casos de estudio diseñados. Weka fue seleccionada con el objetivo de considerar los resultados de varios algoritmos de agrupamiento y utilizar las facilidades que brinda para el estudio experimental. CorpusMiner, por su parte, para poder realizar un estudio de las medidas de validación del agrupamiento en dominios textuales. La similitud utilizada en Weka para aplicar las medidas basadas en RST fue el dual de la distancia HEOM y en CorpusMiner se utilizó la similitud Coseno; en el Anexo 2 se describen ambas expresiones.

Tanto en Weka como en CorpusMiner fue necesario implementar el Algoritmo 2 y las medidas basadas en RST. Para realizar el estudio comparativo, se han seleccionado e incorporado medidas de validación que cumplen diferentes propiedades, son intuitivamente plausibles, fáciles de implementar y están incluidas en la literatura clásica, por tanto sirven como una base adecuada para la evaluación. Medidas externas incluidas: entropía [178, 179, 181], precisión (P), cubrimiento (R) y OFM [180]. Medidas internas incluidas: similitud global (OS) [178], los índices Dunn (DD) [33] y su generalización (DB) [35], la medida Davies-Bouldin (IDB) [34], la conectividad parcial pesada (CPP) y la densidad esperada (DE) [38].

Los algoritmos EM, FarthestFirst, DBSCAN, SimpleKMeans (algoritmo *k*-medias) y XMeans[36], incluidos en Weka, fueron aplicados a los casos de estudio procedentes de dominios no textuales, mientras que los algoritmos SKWIC, ES y la concatenación ES–SKWIC, incluidos en CorpusMiner, fueron aplicados a los casos de estudio en dominios textuales. CorpusMiner, además, permite el preprocesamiento y representación espacio vectorial de los documentos, detalles de esta etapa se describen en el epígrafe 4.3.

3.3.2 Diseño y aplicación de experimentos

El nuevo enfoque para validar los resultados de agrupamientos, instrumento de medición a evaluar, se basa en el Algoritmo 2, que incluye la aplicación de las medidas de RST. Las tres

[35] WEKA es una herramienta de código abierto escrita en Java. Está disponible en http://www.cs.waikato.ac.nz/~ml/weka bajo licencia pública GNU. En este trabajo ha sido utilizada como herramienta para asistir la evaluación del instrumento de medición del agrupamiento usando RST.

[36] Se han utilizados los mismos identificadores que utiliza Weka para referirse a los algoritmos

variantes de cálculo de la media de la pertenencia de los objetos a los grupos, expresiones (3.12), (3.14) y (3.16), se consideran pesos en las expresiones de calidad y precisión generalizadas. Otras dos ponderaciones que se incluyen son: la cardinalidad normalizada de los grupos y la medida similitud global por grupo. En las tablas resultantes de los experimentos se utilizó la notación siguiente: precisión aproximada (PA), calidad aproximada (CA), precisión generalizada (PGRM1, PGRM2, PGRM3, PGC y PGOS) y calidad generalizada (CGRM1, CGRM2, CGRM3, CGC y CGOS) con ponderaciones según (3.12), (3.14), (3.16), la cardinalidad normalizada y la similitud global por grupo, respectivamente.

La Figura 3.1 muestra el esquema utilizado para chequear la confiabilidad y validez del instrumento [244-246].

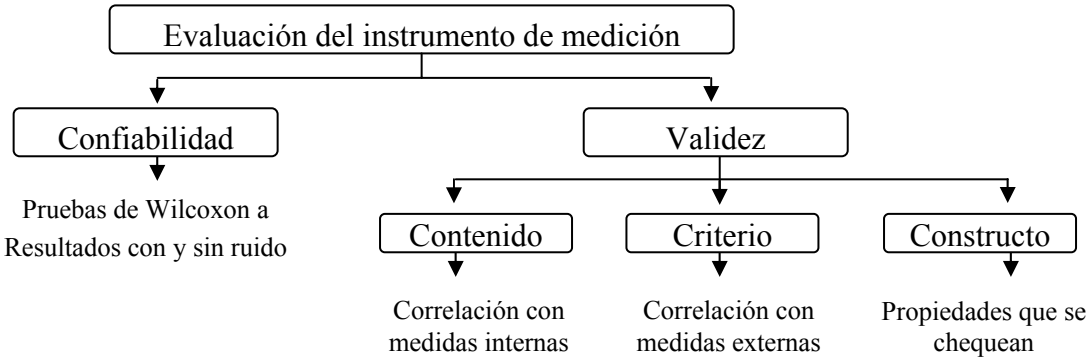


Figura 3.1 Esquema general para evaluar la propuesta de validación de agrupamiento basada en RST.

Medir confiabilidad. La confiabilidad se midió comparando los valores arrojados por el instrumento de medición a los resultados del agrupamiento con y sin ruido. Todos los archivos de datos recopilados en los dos primeros casos de estudio fueron utilizados en este experimento. Las instancias de cada archivo original de datos fueron agrupadas mediante los algoritmos incluidos en Weka: EM, FarthestFirst, DBSCAN, SimpleKMeans y XMeans. Cada resultado de agrupamientos fue evaluado con las medidas internas basadas en RST. Así, ya se tenían los valores de las medidas para cada agrupamiento aplicado a cada archivo sin ruido. Posteriormente, se introdujo ruido a cada archivo de datos. Se modificaron un 5%, 10%, 15%, 20% y 25% de los valores de los rasgos que describen los objetos en cada archivo. La alteración, también aleatoria, fue de a lo sumo un 10% del valor original del rasgo. A cada uno de los archivos con ruido introducido desde un 5% hasta un 25%, se le aplicaron los

algoritmos de agrupamiento ya mencionados y los resultados fueron evaluados utilizando las medidas internas basadas en RST. Así, ya se tienen resultados de las medidas para cada archivo de datos sin ruido y con varios niveles de ruido después de aplicado cada algoritmo de agrupamiento. La prueba no paramétrica de Wilcoxon[37] fue aplicada entre los valores de las medidas resultantes de agrupamientos con y sin ruido. En el Anexo 16 se muestran los valores de significación de esta prueba, y es posible observar que, alterando inclusive hasta un 25% de los valores de cada uno de los conjuntos de instancias, los valores de significación siempre son superiores a 0.05. Esto indica que no existen diferencias significativas entre las poblaciones comparadas horizontalmente (medidas aplicadas sobre conjuntos de datos con y sin ruido). Es importante señalar que los resultados de las medidas están incluso influenciados por la sensibilidad al ruido de cada uno de los métodos de agrupamiento utilizados. Esto evidencia que, para los datos considerados, las medidas son confiables porque produjeron valores similares en condiciones similares, el ruido no cambió los resultados.

Medir validez. Para medir la validez hay que tener evidencias de la validez de contenido, de criterio y de constructo [244]. La primera se refiere al grado en que el instrumento refleja un dominio específico del contenido de lo que mide. La segunda compara los resultados del instrumento con un criterio externo. La última, se refiere al grado en que una medición se relaciona consistentemente con otras mediciones de acuerdo con hipótesis derivadas teóricamente y que conciernen a los conceptos o constructos.

***Validez de contenido*.** Se verifica que el instrumento al menos cubre las propiedades que las medidas internas utilizadas en el estudio verifican en los datos, considerando todos los archivos recopilados en los dos primeros casos de estudio. Los objetos de cada archivo de datos fueron agrupados mediante los algoritmos ya mencionados incluidos en Weka. Se utilizaron estos algoritmos porque consideran en su mayoría prototipos y forman grupos compactos y bien separados, lográndose de esta forma que las medidas internas implementadas tengan un buen comportamiento en la evaluación. Se aplicó el método de correlación Tau b de Kendall entre los resultados de las medidas basadas en RST y las medidas internas seleccionadas. Los resultados de la correlación se muestran en el Anexo 17

[37] Se utilizó SPSS 13.0 para Windows

donde cada fila corresponde a las medidas internas referenciadas y cada columna a las medidas basadas en RST propuestas. La primera subfila corresponde al coeficiente de correlación y la segunda muestra la significación de la correlación entre cada par de medida referenciada y propuesta. No todas las medidas basadas en RST correlacionan con todas las medidas internas utilizadas en el análisis, pero se observa que las medidas basadas en RST en su conjunto logran cubrir todas las propiedades que cubren las medidas ya publicadas, para los archivos de datos considerados. Las medidas propuestas tienen un cálculo común inicial; sin embargo, las medidas citadas requieren estructuras y cálculos diversos y en la mayoría de los casos, costosos. En la Tabla A17.4 se observa que ninguna de las medidas propuestas es capaz de correlacionar con la similitud global y con la conectividad parcial pesada, incluso en el resto de las tablas del Anexo 17 pocas medidas propuestas son capaces de correlacionar con estas dos medidas internas referenciadas. Esta situación se debe esencialmente a que estas dos medidas sólo tienen en cuenta el comportamiento interno de cada grupo a evaluar pero no consideran en su análisis la integración entre los grupos y por tanto, pueden dar una valoración positiva del agrupamiento cuando en realidad no lo es. Las medidas propuestas siempre tienen en cuenta la integración de los grupos en el análisis.

Validez de criterio. El criterio externo que se ha utilizado en esta evaluación del instrumento es el resultado producido por las principales medidas externas referenciadas. Por tanto, sólo han sido posible utilizar los casos de estudio dos y tres, para ellos existe la clasificación de referencia. Los archivos del segundo caso de estudio fueron sometidos a cada uno de los algoritmos de agrupamiento seleccionados de Weka. A los resultados de los agrupamientos se les aplicaron las medidas basadas en RST y las medidas externas referenciadas. El objetivo de este experimento es verificar, para los archivos de datos considerados, que las medidas propuestas logran tener un comportamiento similar a las medidas externas, sin requerir conocimiento adicional de los datos a agrupar. Por tanto, se sugiere su utilidad en la práctica donde no existan clasificaciones de referencia. Para ello, al igual que al validar contenido, se ha aplicado el método de correlación Tau b de Kendall entre los resultados de validación de la nueva propuesta y las medidas externas referenciadas. En el Anexo 18 se muestran los resultados de la correlación. Aquí cada fila corresponde a las medidas externas referenciadas. No todas las medidas basadas en RST correlacionan con todas las medidas externas utilizadas

en el análisis, igual situación ocurrida en la comparación anterior; pero sí es posible observar que todas las medidas externas correlacionan con al menos una de las medidas basadas en RST. Esto muestra que, para los conjuntos de datos considerados, las medidas propuestas tienen un comportamiento similar a las medidas externas aplicadas sin utilizar una clasificación humana de referencia.

En el Anexo 19 aparecen las correlaciones obtenidas entre las medidas basadas en RST y las medidas externas, aplicadas a resultados de agrupamientos de las colecciones textuales del tercer caso de estudio definido. Cada medida externa queda cubierta por más de una de las medidas propuestas, así lo reflejan los valores de significación menores que 0.05 y en muchos casos menores que 0.01. Se introdujo la nueva medida, nombrada Medida-F aproximada (Rough F-measure; RFM), que imita en el contexto de RST la ya conocida y tan usada en dominios textuales: OFM. RFM calcula la media armónica de precisión y calidad generalizadas ponderadas con la tercera forma de cálculo de la pertenencia aproximada, aunque la expresión es válida para combinar cualquier forma de cálculo de precisión y calidad. Esta medida y la calidad generalizada con igual ponderación son las que arrojan las mejores correlaciones. Aunque las medidas basadas en RST han sido propuestas para validar particiones, en este estudio se han incluido resultados del algoritmo ES que crea cubrimientos, para mostrar el comportamiento y uso posible de las medidas para validar cubrimientos en dominios textuales.

Se diseñaron dos ejemplos con el objetivo validar gráficamente los agrupamientos a partir de dos dimensiones obtenidas por las técnicas PCA o CATPCA[38] y así comparar la calidad que refleja la visualización con la que reflejan las medidas basadas en RST. Los ejemplos conformados provienen de colecciones textuales. El primero, es una colección extraída de BioMed Central que contiene 31 documentos, y el segundo, constituye una colección de 60 noticias provenientes de Reuters. Ambas colecciones fueron procesadas con CorpusMiner, donde definitivamente quedaron descritas por 600 términos y se agruparon con el algoritmo ES-SKWIC. Al aplicar este algoritmo se modificaron los valores de los parámetros de forma

[38] En el SPSS hay una versión de CATPCA implementada por Data Theory Scaling System Group (DTSS), Faculty of Social and Behavioral Science, Leiden University, The Netherlands

tal que se obtuvieran agrupamientos de mala y buena calidad según la clasificación de referencia y cada resultado fue evaluado con la medida RFM. Es posible observar en el Anexo 20 la visualización en el plano de esta calidad del agrupamiento a partir de la reducción de las 600 dimensiones a sólo dos y la correspondencia que existe entre la visualización y los resultados de RFM. Esta es otra forma de evidenciar la utilidad de las medidas basadas en RST, esencialmente en dominios textuales.

Validez de constructo. Se han propuesto dos criterios principales para la evaluación del agrupamiento [37]: la compactación de los grupos y la separación entre ellos; constituyendo dos de los constructos principales en esta validación. Los elementos de la propuesta encaminados a chequear la compactación de los grupos son las expresiones que permiten la ponderación de las variantes generalizadas: las tres formas de cálculo de la pertenencia aproximada y la similitud global. Todas las variantes de precisión y calidad aproximadas chequean la separación entre los grupos, y sus variantes ponderadas logran chequear ambos constructos. Así lo muestran el Anexo 17, el Anexo 18 y el Anexo 19 con los resultados de las correlaciones realizadas entre los valores de las medidas propuestas y las medidas internas y externas seleccionadas de la literatura. Adicionalmente, las medidas propuestas permiten medir la precisión y calidad de los grupos, y el nivel de inconsistencia de los mismos, resultados que se logran por la propia definición de las medidas.

Los experimentos realizados no sólo han permitido mostrar que las medidas son un instrumento adecuado para validar el agrupamiento, sino que sugieren reducir el número de medidas a considerar en la evaluación. El Anexo 17, Anexo 18 y el Anexo 19 reflejan que la precisión y calidad ponderadas con la cardinalidad de los grupos, la similitud global y la tercera expresión de la pertenencia aproximada son las medidas que alcanzan las mejores correlaciones; indicando de esta forma que este subconjunto de medidas debe ser el más utilizado para validar agrupamientos utilizando RST.

Después de los experimentos realizados es posible preguntarse: ¿es esta nueva forma de evaluación mejor que las anteriores? No existe una medida que sea válida para todos los tipos de agrupamientos y que pueda captar todas sus propiedades, por tanto, es mejor preguntarse ¿cuáles son las ventajas de estas medidas respecto a las anteriores? Con este nuevo enfoque para la validación es posible:

- Calcular el nivel de inconsistencia, la precisión y calidad local de cada grupo, y globalmente de los grupos respecto al sistema de información dado.
- Validar agrupamientos con independencia de la forma de los grupos resultantes.
- Incluir en la validación el análisis de propiedades estructurales de los grupos.
- Brindar una variedad de medidas que requieren un cálculo inicial común para todas y el cálculo de las particularidades de ellas es poco costoso.
- Validar sin tener en consideración centros de los grupos, por tanto, se puede evaluar resultados de un mayor número de tipos de agrupamientos.
- Utilizar similitudes asimétricas entre los objetos, en aplicaciones donde sea requerido.
- Considerar elementos como pertenencia de los objetos a los grupos y cardinalidad de los grupos en el proceso de evaluación.
- Utilizar en la validación la misma función de similitud entre objetos que fue utilizada en el agrupamiento.

3.4 Valor adicional obtenido al etiquetar los grupos

En muchos casos existe un divorcio entre los resultados del agrupamiento de un conjunto de datos y la necesidad de los usuarios de recuperar etiquetas de los grupos resultantes, por lo que se requiere una etapa post-agrupamiento usualmente nombrada etiquetamiento [41]. Las etiquetas deben ser sintetizadoras, expresivas, contiguas, no redundantes, poseer consistencia jerárquica y poder discriminante [47]. Algunos de los mayores retos de esta etapa son etiquetar grupos con formas irregulares, distinguir puntos fuera de rango y extender los límites de los grupos [41]. Algunos trabajos han sido desarrollados en esta área [41-47, 78, 79, 247-249]; sin embargo, etiquetar es usualmente ignorado por los investigadores en agrupamiento y se le ha prestado menos atención a crear buenos descriptores de grupos. Una de las razones de esta situación es que el problema del agrupamiento no ha sido resuelto aún [41].

Se desea que el proceso de etiquetamiento sea no supervisado para que se pueda realizar rápidamente sin el alto costo de la intervención humana en la anotación de los datos [43]. Una forma de etiquetar es representando adecuadamente los grupos [41]. Las representaciones de los grupos pueden ser basadas en centros, puntos representativos, árboles de clasificación y

reglas. El enfoque basado en los puntos más representativos funciona mejor que el basado en centros, ya que describe los grupos con más detalles [170]. Definir los puntos más representativos desde grupos con formas arbitrarias y donde sea necesario extender la región límite es una tarea difícil.

Considerando, por un lado, que técnicas de muestreo y resumen pueden ser usadas en el etiquetamiento y las propiedades deseadas de una etiqueta, y por otro, las facilidades que brinda RST, en este trabajo se sugiere utilizar RST para etiquetar mediante la extracción de los objetos más representativos de cada grupo. No es objetivo comparar esta forma de etiquetamiento con las ya referenciadas, el propósito es mostrar una potencialidad adicional de RST en otra etapa post-agrupamiento. Etiquetar de esta forma no requiere cálculos adicionales si previamente fue utilizada RST para validar el agrupamiento.

La idea del uso de RST para etiquetar es reemplazar el concepto vago sobre representatividad por el concepto llamado aproximación inferior, porque éste incluye todos los objetos que con certeza pertenecen al grupo. Así, un objeto se dice más representativo de un grupo si él pertenece a la aproximación inferior del grupo [236].

La inclusión o no de objetos a la aproximación inferior de un grupo depende del umbral utilizado para construir las relaciones de similitud al aplicar RST. El Anexo 21 muestra un ejemplo que ilustra cómo cambia el conjunto de objetos más representativos mediante la variación del umbral de similitud. Al etiquetar grupos resultantes del agrupamiento en dominios textuales, ser flexible en cuanto a la variación del umbral es ventajoso, por ejemplo: cuando los usuarios desean tener control del tamaño del conjunto de documentos representativos, las aproximaciones inferiores establecen prioridades al enfrentarse a todo el grupo o permiten ponderar los objetos con el propósito de realizar un etiquetamiento más refinado, es deseado visualizar los resultados del agrupamiento a partir de saltos discretos de representatividad y hay un conocimiento parcial de la colección a procesar. Sin embargo, no siempre se desea la flexibilidad en cuanto a la variación del umbral, sobre todo cuando los usuarios desconocen cómo usar los parámetros, los documentos más representativos se utilizarán directamente como una etapa intermedia a un proceso de etiquetamiento más refinado, perder información tiene consecuencias graves en la toma de decisiones de los usuarios y hay un desconocimiento total de la colección a procesar.

Un problema previsible al etiquetar grandes volúmenes de datos es que los límites de los grupos pueden ser extendidos más o menos mediante la incorporación de un mayor o menor número de puntos de datos etiquetados [41]. La extensión de los límites o fronteras puede provocar la conexión de diferentes grupos y por tanto puede ser necesario combinarlos. Una forma de extender las regiones límites de los grupos utilizando RST es calculando las aproximaciones superiores para cada uno de ellos. De esta forma es posible, además, decidir combinar grupos, sobre todo al analizar resultados de agrupamientos donde fueron generados muchos grupos unitarios. Por tanto, un proceso de fusión de grupos pudiera ser posible utilizando sus aproximaciones superiores. En el Anexo 21 se muestra como las aproximaciones inferiores también contribuyen a la fusión de grupos.

3.5 Consideraciones sobre el umbral

La obtención de las clases de similitud depende de la medida de similitud y el umbral que se utilice. Tanto la validación, como el etiquetamiento, y la posible fusión utilizando RST, dependen de las aproximaciones inferiores y superiores que se obtengan, y éstas a su vez dependen de las clases de similitud. Por tanto, es fundamentel seleccionar adecuadamente la similitud a emplear y estimar correctamente el umbral a utilizar.

La selección del umbral adecuado es una tarea difícil, ésta depende del dominio donde fue aplicado el agrupamiento, cómo fueron descritos los objetos y qué nivel de granularidad se desea al validar, etiquetar y fusionar los grupos. Otro factor que puede influir es el tipo de agrupamiento empleado. Por ejemplo, considerando que se realizó un agrupamiento en dominios textuales con el algoritmo ES, se aconseja utilizar el mismo umbral empleado al agrupar, porque este algoritmo requiere que se le especifique un umbral de similitud β_0 para buscar las componentes β_0 conexas. Posiblemente para otro método de agrupamiento que requiera umbrales no sea adecuado utilizar el mismo umbral porque su semántica no se corresponda con el uso posterior que se le dará; tal es el caso del umbral de corte que se establece en los métodos basados en la intermediación. Adicionalmente, algunas veces se quiere ser más restrictivo al validar que al agrupar y se modifica el umbral.

Existe dependencia entre el umbral utilizado para validar, etiquetar y fusionar. Incluso, se propone que sea el mismo umbral para que la conformación de las clases de similitud y

aproximaciones inferiores y superiores se realice una sola vez y sea común para estas etapas post-agrupamiento basadas en RST. No obstante, algunas veces se requiere variar el umbral para regular la cardinalidad de los conjuntos que contienen los objetos más representativos.

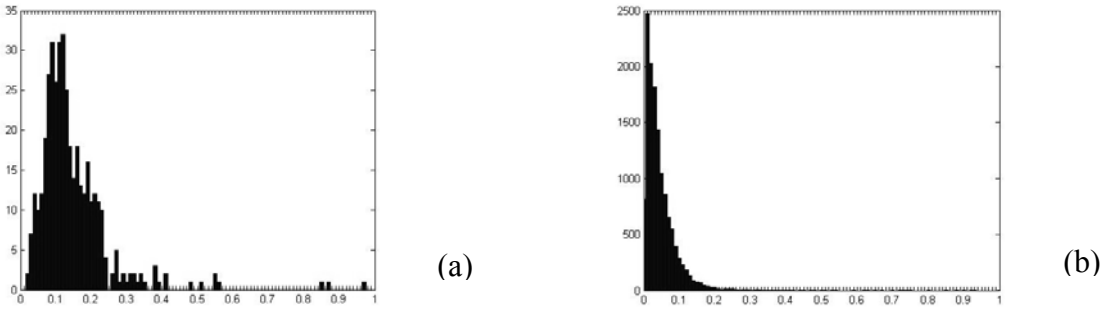


Figura 3.2 Histograma de frecuencias de similitudes coseno, procedentes de: (a) fragmento de la colección de la agencia Reuters de noticias y (b) fragmento de la colección de resúmenes de artículos presentados a CEC2006.

Varias formulaciones son posibles para la determinación y uso de umbrales. Si se conocen datos sobre la distribución estadística de la distancia o la similitud de cualquier patrón del grupo a su centro o a un patrón del mismo. Además, si se conoce la distribución estadística de cierta función de la distancia o la similitud, se podrían utilizar estadígrafos que conduzcan a la definición del umbral. La dificultad de este criterio radica en el conocimiento exacto de la distribución estadística de una distancia al grupo. Otra posibilidad es utilizar los histogramas de frecuencias de los valores de similitud o distancia entre los objetos a procesar. Por ejemplo, en la Figura 3.2 se muestran los histogramas de frecuencias de un fragmento de la colección de la agencia Reuters de noticias y de la colección de conversaciones en listas de correos 20-NewsGroups[39], donde se observa que las similitudes Coseno entre los documentos en ambas colecciones producen histogramas de frecuencias diferentes.

Otros elementos que influyen en la estimación del umbral son la variabilidad en la densidad de los grupos y la varianza y desviación estándar de las similitudes. Por otro lado, el umbral, en algunos casos, constituye una herramienta que tiene el usuario para hacer que el método se ajuste a sus requerimientos y características del problema.

En el Anexo 10 se muestran expresiones que permiten estimar umbrales a partir de una matriz de similitudes o distancias y que no requieren información adicional del conjunto de datos que

[39] Colección 20-NewsGroups disponible en http://www.ai.mit.edu/people/jrennie/20Newsgroups

se procese. Una de estas variantes es la media de las similitudes entre todos los pares posibles de objetos. La media es sensible a observaciones extremas, y en su defecto, se pudiera utilizar la media recortada para eliminar los valores fuera de rango. También se puede utilizar la media de los valores máximos de las similitudes entre cualquier par de objetos. Esta forma de cálculo puede provocar la obtención de un umbral muy alto, conduciendo a que exista coincidencia del grupo y sus aproximaciones. Esta situación puede arrojar valores de precisión y calidad cercanos a uno, cuando en realidad el resultado del agrupamiento no sea tan bueno. Por el contrario, la media de los valores mínimos de las similitudes entre cualquier par de objetos permite obtener umbrales de similitud muy bajos. De esta forma, la aproximación inferior de un grupo será mucho menor que el grupo, y éste a su vez, mucho menor que su aproximación superior. Esto provoca que se obtengan valores muy bajos de precisión y calidad cuando en realidad el resultado del agrupamiento no sea de tan baja calidad. En varios experimentos realizados como parte de esta investigación se ha utilizado la media ponderada de la media de las similitudes y la media de los máximos.

3.6 Conclusiones parciales

Las medidas basadas en RST que se proponen en la tesis para validar resultados del agrupamiento logran correlaciones altamente significativas con las principales medidas internas y externas referenciadas en la literatura, para los casos de estudio diseñados que incluyen archivos de datos provenientes de disímiles dominios y especialmente de dominios textuales. Así, se muestra que esta propuesta logra al menos captar las mismas buenas propiedades que las medidas internas incluidas en el estudio. La medición de precisión, calidad e inconsistencia tuvo comportamientos similares a las medidas externas para los casos de estudio definidos, sin requerir clasificaciones de referencia. Se mostró la confiabilidad y validez de las medidas a partir de resultados experimentales para los datos considerados.

Las medidas son independientes de la forma de los grupos resultantes, permiten la inclusión de propiedades estructurales y de medidas de otras naturalezas en la validación, posibilitan la variación de la granularidad con que se desea evaluar el agrupamiento (especificando el umbral de similitud entre los objetos) y no dependen de la existencia de centros de grupos para la evaluación, por lo que permiten validar resultados provenientes de un gran número de métodos de agrupamiento.

Las medidas precisión y calidad generalizadas con las ponderaciones: cardinalidad, similitud global y tercera expresión de la pertenencia aproximada por grupos, tienen una correlación altamente significativa con los resultados de las medidas internas y externas aplicadas. Esto evidencia que cubren una mayor diversidad de medidas para los datos considerados en los experimentos. Los experimentos muestran que las variantes generalizadas arrojan mejores resultados que las medidas de calidad y precisión ya establecidas en RST.

La propuesta tiene utilidad en otras dos formas de post-agrupamiento: el etiquetamiento de los grupos y el refinamiento mediante la fusión de grupos de los resultados de agrupamientos, mediante el uso de las aproximaciones inferiores y superiores de los grupos. Variaciones en el umbral de similitud permiten restringir o no el conjunto de objetos más representativos para caracterizar los grupos. Esta situación puede ser favorable o no en dependencia del procesamiento posterior que se realice con las aproximaciones inferiores y superiores de éstos. Los histogramas de frecuencias de las similitudes, medidas estadísticas, así como umbrales estimados para algoritmos de agrupamiento que previamente se aplicaron, deben considerarse para la estimación de los umbrales requeridos en el cálculo de las relaciones de similitud.

Todo lo anterior muestra que el empleo de RST, concretamente el Algoritmo 2 y las medidas propuestas, permite valorar los grupos y los resultados generales de agrupamientos; validando esencialmente su precisión, calidad e inconsistencia y la caracterización de los grupos mediante sus objetos más representativos y relacionados. Adicionalmente, la teoría tiene un uso potencial en el refinamiento de los resultados de agrupamientos al sugerir posible fusión entre grupos.

4 La manipulación de documentos para contribuir a la gestión de información y conocimiento

La información es cada día más creciente, heterogénea, diversa y dinámica; y constituye una fuente importante de conocimiento. Gestionar la información y el conocimiento son retos que tienen hoy las organizaciones. Enfrentarlos en dominios textuales es un desafío aún mayor, ya que se hace necesario el desarrollo de sistemas de manipulación de documentos que contribuyan a dicha gestión. Los sistemas que se encargan de recuperar, organizar y analizar de forma precisa y eficiente la información, y recomendar acciones a partir del procesamiento realizado, constituyen la principal motivación de este trabajo. En este capítulo se presenta un esquema general de aplicación que permite integrar el agrupamiento y post-agrupamiento en el desarrollo de sistemas manipuladores de documentos que contribuyen a la gestión de información y conocimiento en las organizaciones. Para ello, previamente, se comentarán los elementos principales que caracterizan estos sistemas, particularizando en la etapa de representación textual indispensable en este tipo de procesamiento y se mostrarán en detalles los sistemas SATEX y GARLucene que materializan el esquema propuesto. Finalmente, se comenta el nivel de aceptación de los usuarios respecto a la aplicación, basado en un análisis descriptivo de las encuestas donde ellos plasmaron su valoración de los sistemas.

4.1 Gestión de información y conocimiento: manipulación de documentos

La información y el conocimiento surgen de acciones humanas que interconectan señales, signos y artefactos en diversos medios. El espacio de información está dado por su codificación, abstracción y difusión [250]. Las dos primeras se refieren a la creación de categorías que faciliten la clasificación de fenómenos y la minimización del número de categorías necesarias, mientras que la tercera combina las dos primeras. Las organizaciones requieren utilizar la información no sólo para darle significado en su entorno, sino para crear nuevo conocimiento, compartirlo y tomar decisiones [250]. Sus principales incentivos para la gestión del conocimiento son [251, 252]: buscar, aportar, contribuir, diseminar, explotar y evaluar el conocimiento. Se necesita transformar información en conocimiento: estructuración de datos e información y la acción humana sobre éstos [250].

El esquema general de aplicación que aquí se propone contribuye a una forma de transformación de la información en conocimiento, porque manipula documentos mediante el agrupamiento y post-agrupamiento, revelando el orden de los datos e imponiendo patrones en los grupos descubiertos. La manipulación de documentos trabaja con conocimiento objetivo, formalizado acorde a algún esquema de codificación; por ejemplo, patente, reporte, artículo, norma. Este conocimiento se clasifica como explícito. Otras dos clasificaciones son: tácito, cuando es personal, en un contexto específico y formalizarlo es difícil, e implícito, cuando es subjetivo [250, 253].

El conocimiento explícito tiene gran importancia porque codifica aprendizaje pasado, habilita la coordinación de actividades y funciones y reduce el procesamiento de información mediante la estipulación de premisas, criterios y opiniones [250, 253]. Además, se comunica fácilmente, aunque transferirlo requiere conocimiento colateral del receptor para entenderlo y aplicarlo. Conocimiento colateral que tiene naturaleza tácita porque los expertos interpretan el significado de la nueva información y surgen incógnitas cuando tratan de usarlo. Por ejemplo, un investigador que se enfrenta al resultado de una colección de artículos científicos automáticamente agrupada, y a la determinación automática de los documentos más representativos y relacionados con cada grupo, requiere utilizar conocimiento tácito para interpretar adecuadamente los resultados y tomar decisiones en el estudio del arte en función de la recomendación recibida automáticamente. La forma de post-agrupamiento que se propone contribuye a la reducción del procesamiento de información mediante la especificación por grupos de cuáles son sus documentos más representativos, simplificándole a los usuarios la revisión de la colección.

Una de las formas de creación de conocimiento se alcanza moviéndolo desde el nivel de individuos hasta el de grupos, durante cuatro etapas del ciclo de conversión del mismo: socialización, exteriorización, combinación e internalización [254]. El esquema de aplicación propuesto contribuye a la manipulación del conocimiento partiendo del conocimiento codificado (explícito, repositorio de conocimiento compartido) y tributa a las dos últimas etapas del ciclo. Por un lado, se crea conocimiento explícito mediante el agrupamiento y valoración de los grupos textuales a partir de la recopilación del conocimiento explícito de múltiples fuentes. Así, se pone en práctica una de las formas de llevar a cabo la combinación:

producir nuevo conocimiento explícito mediante la combinación (agrupamiento y organización) del conocimiento explícito acumulado [250]. Por otro, se recomienda a los usuarios cómo enfrentarse a la colección textual, y por tanto, tienen más elementos para personificar el conocimiento explícito y aumentar sus experiencias en el tácito.

Las organizaciones deben considerar servicios de conocimiento para lograr la integración de fuentes locales (por ejemplo, intranet local, servidores de ficheros, sitios públicos), intra-redes y extra-redes [251]. Pero sólo esto no es suficiente, es necesario utilizar adecuadamente las componentes de las tecnologías de la información para desarrollar gestores de conocimiento.

Los sistemas de gestión de documentos[40] han tomado un gran auge en la actualidad [255]. Docyoument, creado por Media-style[41], encuentra y lista información relevante en varias fuentes a partir de tópicos y preguntas definidas por los usuarios, genera reportes y resúmenes de los documentos, y descubre relaciones entre contenidos. El grupo de herramientas Text Miner[42] descubre y extrae conocimiento desde textos. Otro ejemplo lo constituye Worldox[43], que permite la seguridad de los documentos, el control del acceso, búsquedas a texto completo, visualizadores de documentos en diversos formatos, archivar y guardar la historia de un documento. Autonomy[44] permite un servicio de búsqueda inteligente en Internet a partir de fuentes en 65 lenguajes, deriva el significado de las palabras en el contexto dado y genera los perfiles de los usuarios. Knexa[45] es un ejemplo de plaza de mercado electrónico para conocimiento documentado y el Instituto Kaieteur[46] identificó otros ejemplos[47]. Estos sistemas y herramientas han sido desarrollados en su inmensa mayoría por importantes compañías[48] que han invertido gran capital en la manipulación de documentos para contribuir a la gestión de información y el conocimiento.

---

[40] Sistemas de almacenamiento, sistemas de soporte de búsquedas, modelos de categorización y análisis de contenidos, ontologías y servicios de control de acceso y coordinadores de trabajo colaborativo.
[41] Creación, publicación, gestión, organización, descubrimiento y análisis de contenidos. http://www.media-style.com
[42] http://www.sas.com/technologies/analytics/datamining/textminer
[43] http://www.worldox.com/
[44] http://www.autonomy.com
[45] http://www.knexa.com
[46] Instituto para la gestión del conocimiento (Kaieteur Institute for Knowledge Management http://www.kikm.org)
[47] Knowinc.com, Keen.com, Yet2.com, iExchange.com, Saba.com, cordis.lu, IQ4Hire.com, petrocore.com y eBrainx.com
[48] Autonomy http://www.autonomy.com, ClearForest http://www.clearforest.com, IBM Intelligent Miner for Text http://www-306.ibm.com/software, LexiQuest http://www.lexiquest.com, Teragram http://www.teregram.com y SAS http://www.sas.com

La mayoría de los sistemas mencionados están más dirigidos al comercio del conocimiento en Internet que a la gestión del conocimiento en las organizaciones; las diferencias se muestran en el Anexo 22. Aquellos dirigidos a las organizaciones tienen un alto precio en el mercado internacional, debido esencialmente a los beneficios que les reportan. Ésta es una de las causas que hace que para las instituciones cubanas sea muy difícil adquirir este tipo de sistemas. Las universidades no están exentas de esta limitante y son de las organizaciones en el país que más requieren manipular grandes cantidades de documentos científicos debido a la actividad académica y científica que desarrollan.

4.2 Integración agrupamiento, evaluación y etiquetamiento para gestionar documentos

La gestión del conocimiento juega un rol importante en las universidades e institutos de investigación porque sus procesos son estables, generan y preservan valiosa información proveniente de diversos procesos, tienen acceso a importantes fuentes de información externa, poseen capital humano bien capacitado y buen desarrollo de las tecnologías de la información.

En estas organizaciones se requiere el desarrollo de sistemas gestores de información y conocimiento en función de tareas específicas que tienen que asumir y en dependencia de las características y recursos de las mismas. Algunas de estas tareas son: recomendar a investigadores y docentes cómo enfrentarse a grandes volúmenes de artículos científicos al comenzar la revisión del estado del arte de un tema de investigación, ya que toda nueva investigación comienza con una amplia revisión bibliográfica sobre el tema en cuestión, organizar materiales por equipos de estudiantes para la docencia, organizar por temáticas los artículos que le han llegado al comité científico del programa de un evento, o tener una idea de las asociaciones que existen entre los documentos recuperados y así organizarlos. Estas tareas son ejemplos de necesidades de automatización sobre todo en las universidades cubanas donde existen grandes depósitos centrales de información porque ésta se comparte y publica, y el gran contenido de trabajo hace que el tiempo sea limitado al enfrentarse a revisiones bibliográficas. Además, las tareas mencionadas, en su mayoría, tributan a la producción científica que se lleva a cabo en las universidades, actividad relevante que se refleja en un gran número de los indicadores para la medición del capital intelectual en estas organizaciones [256].

Automatizar este tipo de tareas requiere la integración de varias áreas del saber: el descubrimiento de conocimiento en bases de datos, la minería de datos y de textos. Esta última integra la recuperación y extracción de información, el análisis de textos, el resumen, la categorización, la clasificación, el agrupamiento, la visualización, la tecnología de bases de datos, el aprendizaje automático y la minería de datos [1, 9]. El agrupamiento y los procesos post-agrupamiento permiten organizar la información, determinar información relevante y crear nuevo conocimiento a partir de la información disponible. Uno de los objetivos del agrupamiento es mejorar la habilidad de los usuarios para acceder a la colección, donde los descriptores o etiquetas ofrecen de alguna forma conocimiento sobre la misma [44].

En este trabajo se propone un esquema general de la aplicación del agrupamiento y el post-agrupamiento para recomendar a los usuarios cómo enfrentarse a grandes colecciones textuales, mediante la organización y extracción de conocimiento de las mismas, contribuyendo a la gestión de información y conocimiento. La entrada a éste proviene del resultado de un proceso de recuperación de información [257], de un servidor o colección personal de información textual. Las salidas son los grupos homogéneos de documentos afines, los términos relevantes y los documentos más representativos de cada grupo, así como los que se relacionan con ellos y la calidad con que fueron obtenidos los grupos; garantizando el control para la evaluación de los resultados del agrupamiento. En el Anexo 23 se muestran los cuatro módulos que conforman el esquema propuesto:

Módulo 1. Recuperación de la información o especificación del corpus textual a procesar.

Módulo 2. Representación del corpus textual obtenido o fijado por el usuario.

Módulo 3. Agrupamiento de los documentos.

Módulo 4. Valoración (validación y etiquetamiento) de los grupos textuales obtenidos.

Los sistemas desarrollados que aplican este esquema general son el Sistema para el Agrupamiento, etiquetamiento y evaluación de colecciones TEXtuales (SATEX) y el Sistema para la Gestión de Artículos científicos Recuperados usando Lucene (GARLucene). El primero, parte de una colección textual especificada por los usuarios e ilustra la utilidad de RST en la validación y etiquetamiento de los grupos textuales obtenidos a partir del algoritmo

que concatena los resultados del algoritmo ES con SKWIC [107]. El segundo, utiliza las ventajas de LIUS[49] y Lucene[50] para indexar y recuperar información textual, y muestra la utilidad del agrupamiento basado en la intermediación para agrupar resultados de procesos de recuperación de información y la factibilidad de RST para validar y etiquetar los grupos textuales resultantes de este tipo de agrupamiento.

La introducción de SATEX y GARLucene en el Centro de Estudios de Informática de la Universidad Central "Marta Abreu" de Las Villas (CEI-UCLV) es parte de su estrategia de informatización, que consta de cuatro componentes principales: el portal del centro, los servicios de apoyo a la docencia de pregrado y postgrado, los servicios de apoyo a la gestión administrativa y el módulo de información científico-técnica. Este último está formado por dos componentes principales: el repositorio de información y sistemas de ayuda al procesamiento de información, orientados a facilitarles a los investigadores (profesores y estudiantes) el proceso de revisión bibliográfica, y así contribuir a una mejor administración del tiempo necesario para hacer ciencia.

4.3 Representación textual

Ambos sistemas trabajan con textos (datos no estructurados), por tanto, la representación textual es indispensable para su procesamiento posterior [258]. En esta investigación se ha seleccionado la representación espacio-vectorial (Vector Space Model; VSM) [223] por ser efectiva para representar documentos, ajustarse a otras formas de indexado y ser ampliamente reconocida en la comunidad de minería de textos. Además, es la representación que utiliza CorpusMiner[51] [107, 224, 259, 260], Lucene provee herramientas para manipularla, y se corresponde con los sistemas de información al aplicar RST. En VSM cada documento es identificado como un vector de rasgos en un espacio en el cual cada dimensión corresponde a términos indexados distintos (palabras). Un vector documento dado, en cada componente tiene un valor numérico para indicar su importancia. El segundo módulo del esquema propuesto, según se muestra en el Anexo 23, corresponde a la representación textual y está compuesto por

[49] http://sourceforge.net/projects/lius
[50] http://lucene.apache.org
[51] Sistema resultante de las etapas por las que ha transitado esta investigación

la transformación del corpus, la extracción de términos, la reducción de la dimensionalidad, y la normalización y pesado de la matriz [2].

4.3.1 Transformación del corpus

El objetivo de la transformación del corpus es convertir los ficheros de entrada en una secuencia de ítems lingüísticos (tokens[52] de palabras). En el paso subsiguiente a la extracción de términos, estos tokens serán usados para generar rasgos significativos (índices de términos).

El primer paso en la trasformación del corpus es reconocer los componentes textuales desde los diferentes formatos. Segundo, la secuencia resultante de tokens debe ser transformada. Posibles transformaciones son: convertir las letras todas a mayúsculas o todas a minúsculas, eliminar las marcas de puntuación al final de los tokens, omitir los tokens que contienen caracteres alfanuméricos, identificar los nombres de personas, localidades y organizaciones, y sustituir las contracciones y abreviaturas por sus expresiones completas [2].

4.3.2 Extracción de términos

El submódulo 2.2 se dedica a la extracción de términos. Parte de una secuencia de tokens, obtenida a partir de la transformación del corpus, y produce una secuencia de términos indexados basados en esos tokens. En este esquema de aplicación se propone que el vocabulario se cree a partir de los términos indexados resultantes de la extracción. Representar documentos del lenguaje natural por el significado de un conjunto de índices de términos es un reto, sobre todo porque la información siempre depende del contexto.

Es necesario considerar los niveles descritos en el Anexo 24 al realizar el análisis lingüístico de los textos. En esta tesis se analizan léxicamente los documentos identificando las palabras simples como rasgos [261]. Así, se explota básicamente el plano estadístico de los textos y no se considera la secuencia de aparición de las palabras en un documento (modelo bolsa de palabras; bag-of-words model) [262], aunque alguna información sintáctica puede enriquecer posteriormente los resultados. El análisis léxico es ventajoso porque la definición de los términos es independiente del lenguaje y computacionalmente muy eficiente, la representación resultante es fácil de analizar por los humanos, por lo que se logra la interpretabilidad

[52] Los tokens son cadenas de caracteres delimitadas por espacios en blanco (por ejemplo espacios, cambios de líneas, tabs).

requerida en el post-agrupamiento, con este tipo de rasgos se obtiene naturalmente el sistema de información necesario para la aplicación de RST y es independiente del dominio. Una desventaja es que cada inflexión de una palabra es un posible rasgo y el número de éstos puede ser innecesariamente grande, requiriéndose la reducción de dimensionalidad.

4.3.3 Reducción de la dimensionalidad

Es esencial controlar la dimensionalidad del espacio del vector documento cuando se utilizan las palabras como los términos a indexar. Las razones principales son: (i) la complejidad de muchos algoritmos de agrupamiento depende crucialmente del número de rasgos y reducirlo hace tratables estos algoritmos y (ii) existen palabras que son irrelevantes y producen peores resultados, por tanto, eliminarlas puede aumentar la eficiencia del agrupamiento a realizar. En esta investigación se utiliza el término reducción de dimensionalidad para abarcar técnicas que controlen la dimensionalidad del vector: selección de rasgos y reparametrización [2], o combinación de ambas técnicas [263].

La eliminación de palabras de parada o gramaticales[53] (stop word elimination) [264, 265], es una de las técnicas de selección utilizadas en este esquema de aplicación. También se utilizan métodos de filtrado para decidir cuando incluir un término en el vocabulario o no. El vocabulario final se establece seleccionando todos aquellos rasgos que su puntuación sea superior o inferior a un umbral predeterminado, o los *m* mejores rasgos, es decir, los *m* rasgos con mayor o menor puntuación acorde a la magnitud empleada. La determinación del umbral o el número de rasgos apropiado es aún un tema abierto. En esta tesis se han considerado los criterios siguientes para la evaluación de los términos: muy baja o muy alta frecuencia de aparición en los documentos siguiendo la Ley de Zpif [266], número de documentos en los cuales el término aparece [264], alta frecuencia de aparición respecto a la frecuencia inversa de documentos TFIDF [264] y expresiones I y II de calidad de términos [64]. Estos criterios y otros frecuentemente utilizados se muestran en el Anexo 25.

Las técnicas lingüísticas que reducen la dimensionalidad por reparametrización utilizadas son: la homogeneidad ortográfica (spelling) que permite convertir todas las palabras en un idioma estándar y la reducción de las palabras a su forma raíz (stemming) [264] que permite reducir la

[53] Se considera en la aplicación que la lista es suministrada, disponible en ftp://ftp.cs-cornell.edu/pub/smart/english.stop

dimensionalidad del espacio de rasgos haciendo corresponder palabras morfológicamente similares con la palabra raíz asociada [180, 267]. Técnicas como el análisis de latencia semántico (Latent Semantic Analysis)[54] [268] no han sido incluidas por su alta complejidad computacional. Otras, el uso de tesauros[55] y de ontologías [269, 270], no se incluyeron porque generalmente están asociadas a un dominio específico.

4.3.4 Normalización y pesado de la matriz

Dadas las estadísticas de frecuencias de los términos en todos los documentos, en el submódulo 2.4 se genera un vector pesado para cualquier documento basado en el vector de frecuencias de términos. Cada peso expresa la importancia de un término en un documento con respecto a su frecuencia en todos los documentos. Algunas variantes que permiten pesar los términos indexados se muestran en el Anexo 26. En esta investigación se utilizan las formas de pesado global basadas en alguna variación de la fórmula TF-IDF [64, 271] y la normalización dividiendo la frecuencia de aparición de los términos por la longitud de los documentos para abstraerse de su variedad de tamaños.

4.4 Sistema para el Agrupamiento, etiquetamiento y evaluación de colecciones TEXtuales (SATEX)

SATEX es un sistema manipulador de documentos que ha sido concebido para recomendar a los usuarios cómo enfrentarse a una colección de documentos, mediante el agrupamiento y su validación y etiquetamiento usando RST.

El sistema fue desarrollado en Borland Delphi 7.0. Su ejecutable ocupa 630 KB y requiere, que junto con éste, se encuentre la carpeta Dictionary, con los cinco diccionarios necesarios en la etapa de representación textual; la carpeta PDFBox, que contiene el convertidor PDFBox[56] de ficheros en formato pdf a documentos textos y la carpeta Util con utilitarios para la aplicación. SATEX funciona en cualquier versión de Windows donde esté instalada la máquina virtual de JAVA y el marco de trabajo de .NET, por requerimientos del PDFBox.

[54] Creado por investigadores de Bell Communications Research, Livingston, NJ. Patentado 15-09-1988. No: 07/244,349
[55] Ejemplos: Thesaurus of English Words & Phrases (ed. P. Roget); ISBN 0062720376, Webster's New World Thesaurus (ed. C. Laird); ISBN 0671519832, y Oxford American Desk Thesaurus (ed. C. Lindberg); ISBN 019512674
[56] PDFBox es una biblioteca realizada en Java para el trabajo con documentos PDF. http://www.pdfbox.org

El sistema no implementa los módulos 2 y 3 del esquema general de aplicación propuesto, porque reutiliza la representación textual y el agrupamiento de la herramienta CorpusMiner[57]. Esta herramienta requiere la especificación de diccionarios para la lematización, homogeneidad ortográfica, contracciones, abreviaturas y palabras gramaticales. Su diseño permite que sea extensible, flexible y reutilizable con facilidad. En CorpusMiner cada funcionalidad es desarrollada por múltiples algoritmos, por tanto, cada resultado puede obtenerse de distintas formas: variando el algoritmo o sus parámetros. Estas posibilidades permitieron realizar estudios donde se determinó experimentalmente que la similitud Coseno arrojó los mejores resultados, que la reducción de dimensionalidad utilizando la calidad de término II es un buen método de filtrado para obtener agrupamientos de calidad, y que los métodos concatenados obtienen grupos refinados sin requerir conocimiento previo del dominio [107, 224]. Así, es posible sugerir a los usuarios de SATEX los valores por defecto de los parámetros para la representación textual y el agrupamiento.

En el Anexo 27 se muestra la interfaz gráfica de SATEX y sus funcionalidades principales. El sistema parte de una colección de documentos en idioma inglés y formato pdf especificada por el usuario. El primer módulo sólo se encarga de conformar el corpus textual a partir de todos los ficheros especificados. El módulo 2 en SATEX tiene total correspondencia con la descripción de la representación textual que se definió en el epígrafe 4.3, y para llevarlo a acabo se invocan los métodos implementados en CorpusMiner que tributan a este módulo. La primera fase de la transformación textual convierte ficheros pdf a textos utilizando el PDFBox como una caja negra. Este es el proceso más costoso en todo el procesamiento con SATEX, por lo que es posible guardar el corpus creado para reutilizarlo en futuros agrupamientos. Las principales transformaciones incluidas en la segunda fase son: la lematización, la sustitución de términos, la homogenización ortográfica y la reexpresión de símbolos. Se sugiere el pesado de la matriz usando TF_IDF I y su normalización por frecuencias, según las expresiones del Anexo 26. La reducción de la dimensionalidad se realiza eliminando las palabras de parada, realizando la lematización[58] y seleccionando aquellos términos, en un rango de 500 a 1000, con mayor calidad para el agrupamiento.

[57] Resultado de las etapas por las que ha transitado esta investigación.
[58] La lematización se realiza consultando listas de palabras que contienen todas las variaciones morfológicas.

El tercer módulo de la aplicación propuesta se lleva a cabo en SATEX utilizando la variante de agrupamiento concatenado de CorpusMiner [107], donde se refina la salida del algoritmo ES [106] con el algoritmo SKWIC [64]. El cuarto módulo parte de las salidas del agrupamiento y es fundamental porque se evidencia la aplicabilidad de los aportes teóricos en la etapa post-agrupamiento para valorar grupos textuales, enriqueciendo la interpretabilidad de diversos resultados de agrupamientos. En SATEX se valoran los grupos textuales mediante un listado de las palabras claves [260, 272], los documentos más representativos, otros documentos relacionados con el grupo, y el resultado de medidas internas de validación. Estas últimas tres formas de valoración son posibles debido a la aplicación de RST al post-agrupamiento. Los documentos más representativos se corresponden a la aproximación inferior del grupo y los documentos relacionados a la aproximación superior del mismo.

En SATEX siempre se sugieren valores por defecto para el umbral β_0 requerido por el algoritmo ES y ese mismo umbral se utiliza para el cálculo de las relaciones de similitud al aplicar RST en el post-agrupamiento. La estimación del umbral se realiza calculando la media ponderada de la media de las similitudes y la media de los máximos, no obstante, el usuario puede seleccionar otras formas para el cálculo del umbral, según se describe en el Anexo 10. Cada umbral calculado se publica en la salida del sistema y el usuario puede ajustar independientemente aquel requerido para el agrupamiento o para aplicar RST. Así, el usuario dispone de una herramienta para hacer que el método se ajuste a sus requerimientos y características del problema. Esta flexibilidad no es ventajosa en algunos casos. Si se utilizan valores muy altos del umbral tienden a coincidir las aproximaciones inferiores y superiores con el grupo, mientras que si los valores son muy bajos, la aproximación inferior será mucho menor que el grupo y éste a su vez mucho menor que la aproximación superior; provocando valores de las medidas de calidad cercanos a uno y a cero, respectivamente.

4.5 Sistema para la Gestión de Artículos científicos Recuperados usando Lucene (GARLucene)

GARLucene es otro sistema manipulador de documentos, que al igual que SATEX, sigue el esquema general de aplicación propuesto en esta tesis, pero a diferencia de éste, organiza y caracteriza los resultados de un proceso de recuperación de información. El sistema fue desarrollado completamente en JAVA, su código es libre y multiplataforma. Sólo requiere

que el sistema operativo tenga instalado Java Runtime Enviroment (JRE). Su archivo .jar ocupa 286 KB y requiere que junto con éste se encuentren los archivos necesarios para la configuración del sistema y las bibliotecas necesarias para el proceso de creación de índices, recuperación, agrupamiento y evaluación de los grupos. GARLucene permite la creación de índices de múltiples tipos de ficheros[59] y el usuario puede especificar el visualizador a utilizar.

LIUS y Lucene se utilizaron en el primer módulo de GARLucene. El proyecto Apache Lucene desarrolló una biblioteca que permite la búsqueda y recuperación de información, sobre una indexación utilizando LIUS. Ambos han sido desarrollados sobre tecnología Java y sus fuentes se encuentran totalmente disponibles, elemento esencial para decidir utilizarlos. Lucene es multiplataforma, tiene un alto rendimiento y es escalable, permite la creación incremental de índices, los algoritmos de búsqueda son potentes, fiables y eficientes, permitiendo: ordenar resultados por relevancia, utilizar un amplio lenguaje de consulta, realizar búsquedas por campos y por rangos de fechas, ordenar por cualquier campo, y buscar mientras se actualiza el índice. LIUS es un marco de trabajo para la indexación y búsqueda de documentos. Éste funciona como una capa por encima de Lucene para organizar el trabajo de extraer datos de distintos tipos de documentos y establecer su correspondencia a campos de Lucene. LIUS permite manipular qué campos se crean, de qué tipo y qué analizadores se deben usar, entre otros. Así, ofrece más comodidad al configurar Lucene y decidir qué, cómo y dónde se indexa.

Es muy fácil utilizar LIUS. Toda la configuración para indexar se encuentra definida en un fichero XML, para hacer posteriormente las búsquedas. Dependiendo del tipo de fichero a procesar, se almacenan determinados campos en el índice para hacer posteriormente las búsquedas. LIUS le transfiere estos campos a Lucene de la forma clave-valor[60]. En el Anexo 28 se muestran los campos asociados a cada tipo de ficheros, particularmente los pdf. La división por campos aumenta la potencialidad de GARLucene al procesar artículos científicos.

El primer módulo en GARLucene se dedica esencialmente a la creación de índices y recuperación. La creación de índices se basa en LIUS, sólo hay que especificarle la dirección de los documentos a indexar y los parámetros de configuración necesarios. LIUS procesa las consultas de los usuarios y Lucene comprueba la existencia del índice y obtiene las

[59] Tipos de ficheros: Ms Word, Ms Excel, Ms PowerPoint, RTF, PDF, XML, HTML, TXT, Open Office.
[60] Clave: nombre del campo. Valor: información que se extrae del documento asociada a ese campo.

propiedades de los documentos; permitiendo la recuperación de información. El usuario puede tener más de un directorio de índices para realizar el proceso de búsqueda y recuperación. Es posible realizar múltiples tipos de búsquedas en los textos indexados[61].

GARLucene reutiliza las facilidades de Lucene para la representación textual, correspondiente al módulo 2 del esquema propuesto. La primera fase de la transformación se realiza en la creación de índices y recuperación de información. Lucene permite la representación VSM de la colección recuperada, principalmente a través de la clase StandardAnalyzer que implementa StandardFilter para normalizar los tokens extraídos, LowerCaseFilter para convertir los tokens a minúsculas y StopFilter[62] para eliminar las palabras de parada. Adicionalmente, Analyzer permite obtener las raíces de las palabras mediante heurísticas, y tratar la sinonimia y polisemia. En GARLucene se enriqueció la representación textual adicionando los métodos de filtrado para la selección de términos descritos en el subepígrafe 4.3.3. Se utilizó la clase StandardAnalyzer de la biblioteca Lucene para realizar las transformaciones que se aplican en este módulo, agregándole la eliminación de los tokens alfanuméricos y la obtención de raíces.

GARLucene aplica métodos basados en la intermediación de las aristas y por tanto representa los documentos recuperados como grafos no dirigidos y ponderados, donde cada documento es un nodo y las aristas entre ellos están ponderadas con su similitud Coseno. Inicialmente este grafo es completo, por tanto, con el objetivo de reducir el número de aristas, GARLucene calcula reiteradamente umbrales de similitud entre los documentos y elimina aquellas aristas con ponderación inferior al umbral calculado. De esta forma se facilitan los cálculos y se hacen más efectivos. El umbral de corte que se utiliza es la media ponderada de la media de las similitudes y la media de los máximos, como se muestra en el Anexo 10.

GARLucene implementa tres variantes de agrupamiento jerárquicos divisivos utilizando la intermediación de las aristas. El sistema muestra los grupos en forma de árbol jerárquico. El proceso de agrupamiento comienza por la raíz que está formada por los artículos científicos recuperados que guardan relación con la palabra o frase que originó la búsqueda. Cada grupo que el sistema dividió en varios subgrupos, es decir, que no es una hoja de la jerarquía

[61] Tipos de búsquedas: por palabra, por frase, apoyado por comodines de textos, borrosas, por proximidad, por rango, fomentando un término, usando operadores booleanos y paréntesis para agrupar expresiones, y cualquier combinación de éstas. En el manual de usuarios de GARLucene es posible acceder a los detalles de cada búsqueda.

[62] Utiliza una pequeña lista de palabras de paradas que en esta investigación fue enriquecida.

generada, puede expandirse o contraerse. Si el usuario selecciona un grupo, su contenido se visualiza en el recuadro derecho de la aplicación. Cada configuración de la jerarquía, dada por contracciones y expansiones de los grupos, puede ser evaluada como una partición. Esta evaluación se lleva a cabo calculando las medidas locales y globales basadas en RST. El sistema también ofrece un listado de los documentos más representativos de cada grupo, utilizando para ello las aproximaciones inferiores. El umbral utilizado para conformar la relación de similitud coincide con el umbral de corte utilizado en el grafo original.

4.6 Análisis de la aceptación de los sistemas por parte de los usuarios

Cuando se desarrolla un sistema, se desea tener éxito con el mismo, entonces cabe preguntarse: ¿por qué existe?, ¿por qué algunas personas deben utilizarlo?, ¿por qué regresan a él? Después de responder estas preguntas, surgen nuevas interrogantes: ¿el sistema satisface a sus desarrolladores? y ¿satisface a los usuarios? Interrogantes de este tipo están presentes aún con más fuerza cuando se desarrollan sistemas que producen resultados que conducen a un análisis influenciado por elementos subjetivos; tal es el caso de SATEX y GARLucene.

Las técnicas que se aplican en los sistemas propuestos fueron validadas en los capítulos 2 y 3, basándose tanto en experimentos de laboratorio como en situaciones del mundo real a través de grafos[63], colecciones de datos[64], textos estándares[65] y colecciones arbitrarias pero bien definidas. Por tanto, un primer nivel de evaluación, según [273], ya fue realizado. Falta cubrir otros dos niveles en la evaluación: retroalimentarse verbalmente, por escrito o mediante historias para arribar a conclusiones sobre los resultados obtenidos y realizar encuestas que contengan cuestionarios a usuarios sobre percepción, uso y valor de los resultados [274]. Se ha seleccionado la encuesta para tener retroalimentación del criterio de los usuarios respecto a los resultados; haciendo énfasis en aquellas preguntas centradas en los usuarios, aunque se incluyen algunas centradas en la aplicación [274]. Cada pregunta incluida tributa a los

[63] Laszlo Barabasi http://www.nd.edu/~networks/resources.htm, Alex Arenas http://www.etse.urv.es/~aarenas/data, UCINet http://vlado.fmf.uni-lj.si/pub/networks/data/UciNet y Mark Newman http://www-personal.umich.edu/~mejn/netdata,

[64] UCI ML Repository http://archive.ics.uci.edu/ml, UCI KDD Archives http://kdd.ics.uci.edu y Data sets for clustering techniques http://www.uni-koeln.de/themen/statistik/data.

[65] Reuters http://www.davidlewis.com/resources/testcollections/reuters21578, BioMed Central http://www.biomedcentral.com y 20-newsgroups http://www.ai.mit.edu/people/jrennie/20Newsgroups.

objetivos propuestos: investigar y explorar acerca de la calidad y utilidad de los resultados alcanzados con los sistemas, en función del criterio de los usuarios.

Considerando que el esquema de aplicación propuesto permite la manipulación de documentos para contribuir a la gestión de información y conocimiento, se han recopilado y adaptado criterios, medidas y variables para evaluar este tipo de sistemas. Se analiza en qué medida estos resultados permiten el desarrollo de capacidades[66] en la gestión del conocimiento.

Algunos métodos de evaluación para los sistemas que contribuyen a gestionar conocimiento se focalizan en beneficios internos como eficiencia, beneficios de los usuarios, crecimiento, innovación y transparencia [273]. La flexibilidad es importante, indicando el grado de ajuste a diferentes configuraciones y necesidades individuales. Además, se verifica la presencia de un lenguaje común dado por un vocabulario bien definido y buen entendimiento de conceptos y relaciones. Un elemento importante en la evaluación centrada en el usuario es la usabilidad[67]. Elementos que contribuyen a la usabilidad son la eficacia, la eficiencia y la satisfacción con las cuales los usuarios logran sus objetivos[68]. Las decisiones de diseño también pueden influir en la usabilidad, donde los principales objetivos son: facilidad de aprendizaje y de uso, flexibilidad y robustez[69]. Otros intereses al evaluar son: confiabilidad, desempeño y capacidad de soporte del sistema y sus resultados [275].

A partir de los principales intereses al evaluar sistemas similares a los propuestos y sus resultados, y adecuando el esquema propuesto en [275], se ha definido la evaluación (E) como $E=\{O, V, M, U\}$, cuyos elementos son: los objetivos de la evaluación (O), las variables de medición (V), las medidas (M) y las unidades de medición (U). En el Anexo 30 se detallan las variables de medición y las medidas que tributan a ellas. Las cinco primeras –eficacia, eficiencia, conformidad, confiabilidad, robustez– se focalizan en los resultados alcanzados y las preguntas desde la seis a la diecinueve se asocian a ellas. Las cuatro últimas –usabilidad, flexibilidad, rendimiento y capacidad de soporte– están dirigidas al funcionamiento del sistema. Éstas tienen menor representatividad en la encuesta desarrollada, esencialmente las

[66] Generar, aplicar, distribuir/compartir y almacenar conocimiento, descubrir y aprender.
[67] ISO 9241, después de: Alan Dix, Janet Finlay, Gregory Abowd, Russell Beale. Human Computer Interaction. Prentice Hall Europe. 1998. http://www.tau-web.de/hci/space/i7.html
[68] ISO TS 16071. http://www.usability-forum.com/bereiche/accessibility.shtml
[69] http://www.tau-web.de/hci/space/x12.html

preguntas de la veinte a la veinticuatro, la cuatro y la cinco, se asocian a ellas. El interés fundamental de la evaluación está encaminado a valorar el nivel de aceptación de los resultados de los sistemas por parte de los usuarios, no tanto así con el funcionamiento de éstos. Las unidades de medición varían en función de las medidas y los tipos de preguntas incluidas en la encuesta. En el Anexo 31 se presenta la encuesta elaborada, siguiendo los lineamientos trazados en [244-246].

La encuesta se aplicó a 15 profesores e investigadores del CEI-UCLV. Cada uno de los expertos seleccionados utilizó los sistemas SATEX y GARLucene con colecciones textuales conocidas de sus ramas de investigación[70], tanto heterogéneas como homogéneas. Se definieron 22 variables a partir de las preguntas realizadas en la encuesta. Aquellos usuarios que han utilizado frecuentemente los sistemas, tienen más criterios para dar una valoración de éstos. Por tanto, en el análisis descriptivo de datos realizado en el SPSS, se ponderaron los sujetos utilizando la variable que indica a si el usuario ha utilizado el sistema frecuentemente, algunas veces o casi nunca, con codificaciones 3, 2 y 1, respectivamente.

Los expertos reflejaron que fue fácil familiarizarse tanto con GARLucene como con SATEX. El 95% reflejó que usar GARLucene es relativamente fácil o fácil, respondiendo esta última el 75% de los expertos. El 100% de los encuestados respondieron que es fácil usar SATEX. Las opiniones respecto a la efectividad del uso de los parámetros para obtener mejores resultados del agrupamiento y de la valoración de los grupos fueron variadas. Posibles causas son: la especificación muy técnica de los mismos que impiden su uso, los expertos requieren utilizar más los sistemas, la correspondencia entre los valores de los parámetros y los tipos de colecciones textuales a utilizar. La varianza de las respuestas a las preguntas 7 y 14, relativas a la utilidad de la modificación de los parámetros para el agrupamiento y la valoración, respectivamente, es alta. Las respuestas a las preguntas 7 y 14 respecto a GARLucene tienen varianza 2.892 y 3.712, respectivamente. Se obtienen las varianzas 3.648 y 4.885 a partir de la valoración de SATEX, para las preguntas 7 y 14, respectivamente.

[70] Bioinformática, código genético, biología molecular, estructuras de proteínas, inteligencia artificial, selección de rasgos, minería de datos, descubrimiento de conocimiento, optimización, manejo de incertidumbre, aprendizaje automático, computación gráfica, procesos de negocios, bases de datos y redes neuronales artificiales.

La Tabla A32.1, la Tabla A32.2 y la Tabla A31.3 muestran que los expertos consideran pertinentes y adecuados los agrupamientos obtenidos, así como su validación y las facilidades para el análisis y la interpretación de los resultados brindadas por los sistemas. Los resultados que comprueban la validez del esquema de aplicación propuesto se muestran en la Tabla A32.4 y la Tabla A32.5. La primera muestra la correlación positiva que existe entre las facilidades para descubrir conocimiento, tomar decisiones y analizar la colección de documentos, y la conformidad con los resultados de los agrupamientos. La segunda muestra también correlaciones positivas, en este caso respecto a la conformidad con la valoración del agrupamiento.

4.7 Conclusiones del capítulo

El esquema general que se propone en este capítulo muestra que la integración del agrupamiento, y la evaluación y etiquetamiento de sus resultados, permite la manipulación de documentos, y con ello, contribuye a la gestión de información y conocimiento. Los resultados de las encuestas muestran que los sistemas han sido satisfactoriamente aceptados por los usuarios, sobre todo con el enfoque de la aplicación propuesta para la gestión de artículos científicos, porque manifiestan que la calidad de los resultados del agrupamiento y su valoración contribuye al descubrimiento de conocimiento y toma de decisiones y facilita el análisis de las colecciones a procesar.

SATEX y GARLucene son sólo ejemplos de sistemas que incorporan e integran los resultados teóricos y evidencian la aplicabilidad de los mismos en el agrupamiento y post-agrupamiento de documentos textuales, ya sean provenientes de una colección personal, o resultantes de la recuperación de información. El uso de CorpusMiner fue provechoso para estudiar previamente técnicas para la representación textual que permiten obtener los mejores resultados del agrupamiento y que son computacionalmente factibles.

Los sistemas desarrollados muestran, además, la flexibilidad y facilidad de integración de los resultados teóricos previamente obtenidos. La validación basada en RST se aplicó para validar los resultados local y globalmente, y como criterio de parada en la construcción de la jerarquía en GARLucene. Se evidenció la utilidad de las aproximaciones inferiores para etiquetar grupos textuales.

CONCLUSIONES Y RECOMENDACIONES

Como resultado de esta investigación se diseñó la medida Intermediación Diferencial que captura eficientemente la información topológica que codifica la estructura del problema, así como el algoritmo de agrupamiento que la utiliza eficientemente, particularmente en dominios textuales. Además, se utilizó RST para validar los resultados de agrupamientos a través de nuevas medidas basadas en esta teoría que no requieren considerar la clasificación de referencia y se mostró la utilidad de las aproximaciones inferiores y superiores para caracterizar los grupos; cumpliéndose de esta forma el objetivo general planteado, ya que:

1. Se creó la medida Intermediación Diferencial y el algoritmo de agrupamiento basado en su cálculo. Este algoritmo explota las buenas propiedades que tiene la intermediación diferencial para el agrupamiento: 1) es adecuada para grafos ponderados y no ponderados, 2) no necesita el paso del recálculo, 3) captura mejor las propiedades topológicas y es menos sensible al ruido que las medidas de intermediación anteriormente existentes, y 4) es una medida de disimilitud topológica. Esta nueva forma de medición consideró otro enfoque para el cálculo de la intermediación, a partir de la extracción eficiente de la información local presente en las redes.
2. Se mostró que el Algoritmo 1 propuesto tiene un buen desempeño en problemas de agrupamiento de documentos partiendo de una matriz de similitud coseno entre los textos a agrupar. Los resultados fueron comparados con aquellos producidos por los algoritmos SKWIC, ES, GStar, ACONS, Enlace y GN. Para las tres colecciones textuales consideradas, el algoritmo basado en la intermediación diferencial obtuvo resultados comparables y en la mayoría superiores a los alcanzados por los algoritmos citados.
3. El empleo de RST, concretamente el Algoritmo 2 y las medidas propuestas, permite valorar los grupos y los resultados generales de agrupamientos, mediante: la validación a partir de la medición de la precisión, calidad y consistencia de los grupos y el agrupamiento en general, y la caracterización de los grupos identificando sus objetos más representativos y relacionados. Adicionalmente, la teoría tiene un uso potencial en el refinamiento de los resultados de agrupamientos al sugerir posible fusión entre grupos.

4. Para los casos de estudio diseñados, especialmente en dominios textuales, las medidas internas basadas en RST que se proponen en la tesis para la validación del agrupamiento logran correlaciones altamente significativas con las principales medidas internas y externas referenciadas en la literatura. Estos resultados experimentales muestran la confiabilidad y validez de esta propuesta de validación. Los experimentos muestran que las medidas generalizadas, así como las formas de ponderación propuestas, particularmente la nueva expresión para medir la pertenencia aproximada, arrojan mejores resultados que las medidas de calidad y precisión ya establecidas en RST.
5. El esquema general de aplicación que se propuso muestra, mediante los sistemas SATEX y GARLucene desarrollados, que la integración del agrupamiento, y la evaluación y etiquetamiento de sus resultados, permite la manipulación de documentos, y con ello, contribuye a la gestión de información y conocimiento. El análisis estadístico realizado a los resultados de las encuestas aplicadas a los usuarios evidencian que la calidad del agrupamiento y su valoración, están altamente correlacionados con el descubrimiento de conocimiento y la toma de decisiones a partir de las colecciones textuales procesadas.

Derivadas del estudio realizado, así como de las conclusiones generales emanadas del mismo, se recomienda:

1. Estimar los parámetros de entrada óptimos, así como los umbrales requeridos en cada paso del Algoritmo 1 propuesto. Obtener nuevas versiones del algoritmo que permitan refinar los resultados del agrupamiento de forma tal que se puedan obtener cubrimientos.
2. Estudiar la aplicación de las medidas basadas en la Teoría de los Conjuntos Aproximados para validar resultados de agrupamientos que formen cubrimientos en los datos.
3. Mejorar los sistemas propuestos a partir de las sugerencias recogidas en las encuestas aplicadas a los usuarios. Por ejemplo, ser más flexibles en la definición de umbrales, reducir el número de medidas de validación y optimizar algunos procesamientos costosos, entre otros.
4. Estudiar el comportamiento de los algoritmos 1 y 2 en dominios textuales a partir de otras representaciones de los documentos y otras formas de cálculo de las interacciones entre ellos.

Referencias bibliográficas

1. Dixon, M., *An overview of document mining technology*. 1997: http://www.geocities.com/ResearchTriangle/Thinktank/1997/mark/writings/dixm97_dm.ps.
2. Lanquillon, C., *Enhancing Text Classification to Improve Information Filtering*, in *Research Group Neural Networks and Fuzzy Systems*. 2001, University of Magdeburg "Otto von Guericke": Magdeburg. p. 231.
3. Bueno, E. *Estado del arte y tendencias en creación y gestión del conocimiento*. in *Congreso Iberoamericano de Gestión del Conocimiento y la Tecnología (IBERGECYT 2001)*. 2001. La Habana, Cuba.
4. Canals, A., M. Boisot, and A. Cornella, *Gestión del conocimiento*. 2003, Gestión: 2000: Barcelona, España.
5. Frappaolo, C., *Knowledge Management*. 2006, West Sussex, England: Capstone Publishing Ltd. (A Wiley company).
6. Dalkir, K., *Knowledge Management in Theory and Practice*. 2005, Burlington, USA: Elsevier.
7. Swoyer, S. *Unstructured Data: Attacking a Myth*. Enterprise Systems 2007 [cited; Available from: http://esj.com/business_intelligence/article.aspx?EditorialsID=8577.
8. Passoni, L., *Gestión del conocimiento: una aplicación en departamentos académicos.* Gestión y Política Pública, 2005. XIV(1): p. 57-74.
9. Tan, A. *Text Mining: The state of the art and the challenges*. in *Proceedings of the Conference Knowledge Discovery and Data Mining (PAKDD'99): Workshop Knowledge Discovery from Advanced Databases*. 1999. Pacific Asia.
10. Kruse, R., C. Döring, and M.-J. Lesor, *Fundamentals of Fuzzy Clustering*, in *Advances in Fuzzy Clustering and its Applications*, J.V.d. Oliveira and W. Pedrycz, Editors. 2007, John Wiley and Sons: Est Sussex, England. p. 3-27.
11. Gilbert, E.W., *Pioneer maps of health and disease in England.* The Geographical Journal, 1958. 124(2): p. 172-183.
12. Wilkinson, D.M., *A method for finding communities of related genes.* Proceedings of the National Academy of Sciences (PNAS USA), 2004. 101(1): p. 5241-5248.
13. Yu, L. and S. Ramaswamy. *Verifying design modularity, hierarchy, and interaction locality using data clustering techniques*. in *Proceedings of the 45th ACM Southeast Conference*. 2007. Winston-Salem, North Carolina.
14. Jeong, H., Z. Néda, and A.L. Barabási, *Measuring preferential attachment in evolving networks.* Europhysics Letters, 2003. 61(4): p. 567-572.
15. Newman, M.E.J., *The structure of scientific collaboration networks.* Proceedings of the National Academy of Sciences (PNAS USA), 2001. 98(2): p. 404-409.

16. Bade, K. and A. Nürnberger. *Personalized hierarchical clustering*. in *Proceedings of the IEEE WIC ACM International Conference on Web Intelligence (WI 2006)*. 2006. Hong Kong, China: IEEE Computer Society.
17. Kleinberg, J. and S. Lawerence, *The structure of the Web.* Science, 2001. 294: p. 1849-1850.
18. Getoor, L. and C.P. Diehl, *Link mining: a survey.* SIGKDD Exploration Newsletter, 2005. 7(2): p. 3-12.
19. Wu, A.Y., M. Garland, and J. Han. *Mining scale-free networks using geodesic clustering*. in *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2004)*. 2004. Seattle, WA, USA: ACM Press.
20. Radicchi, F., et al., *Defining and identifying communities in networks.* Proceedings of the National Academy of Sciences (PNAS USA), 2004. 101(9): p. 2658-2663.
21. Donetti, L. and M.A. Muñoz, *Detecting network communities: a new systematic and efficient algorithm.* Journal of Statistical Mechanics: Theory and Experiment, 2004. 2004(10): p. P10012.
22. White, S. and P. Smyth. *A spectral clustering approach to finding communities in graphs*. in *Proceedings of the SIAM International Conference of Data Mining*. 2005. Newport Beach, CA, USA.
23. Pinney, J.W. and D.R. Westhead, *Betweenness-based decomposition methods for social and biological networks*, in *Interdisciplinary Statistics and Bioinformatics*, S. Barber, et al., Editors. 2006, Leeds University Press: Leeds. p. 87-90.
24. Newman, M.E.J., *Modularity and community structure in networks.* Proceedings of National Academy of Sciences (PNAS USA), 2006. 103(23): p. 8577-8582.
25. Newman, M.E.J., *Finding community structure in networks using the eigenvectors of matrices.* Physical Review E, 2006. 74: p. 036104.
26. Newman, M.E.J., *A measure of betweenness centrality based on random walks.* Social Networks, 2005. 27(1): p. 39-54.
27. Newman, M.E.J., *Detecting community structure in networks.* The European physical journal B, 2004. 38(2): p. 321-330.
28. Newman, M.E.J., *Analysis of weighted networks.* Physical Review E, 2004. 70(52): p. 056131.
29. Newman, M.E.J. and M. Girvan, *Finding and evaluating community structure in networks.* Physical Review E, 2004. 69(2): p. 026113.
30. Newman, M.E.J., *Fast algorithm for detecting community structure in networks.* Physical Review E, 2004. 69(6): p. 066133.
31. Theodoridis, S. and K. Koutroubas, *Pattern Recognition*. 1999: Academic Press.
32. Kaufman, L. and P.J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. Wiley Series in probability and mathematical statistics. 1990: John Wiley and Sons.
33. Dunn, J., *A fuzzy relative isodata process and its use in detecting compact well-separated clusters.* Journal of Cybernetics, 1974. 3: p. 32-57.

34. Davies, D.L. and D.W. Bouldin, *A cluster separation measure.* IEEE Transactions on Pattern Analysis and Machine Learning, 1979. 1(4): p. 224-227.
35. Bezdek, J. and N. Pal. *Cluster validation with generalized Dunn's indices*. in *Proceedings of the 2nd International two-stream Conference on ANNES*. 1995. Piscataway, NJ: IEEE Press.
36. Halkidi, M., M. Vazirgiannis, and Y. Batistakis. *Quality scheme assessment in the clustering process*. in *Proceedings of the 4th European Conference on Principles of Knowledge Discovery on Data (PKDD 2000)*. 2000. Lyon, France: Springer-Verlag London, UK.
37. Halkidi, M., Y. Batistakis, and M. Vazirgiannis, *On clustering validation techniques.* Journal of Intelligent Information Systems, 2001. 17(2/3): p. 107-145.
38. Stein, B., S. Meyer, and F. Wißbrock. *On clustering validity and the information need of users*. in *Proceedings of 3rd IASTED International Conference on Artificial Intelligence and Applications (AIA 2003)*. 2003. Benalmádena, Spain: ACTA Press.
39. Borgelt, C. and R. Kruse. *Finding the number of fuzzy clusters by resampling*. in *Proceedings of IEEE International Conference on Fuzzy Systems*. 2006. Vancouver, BC: IEEE Press.
40. Olex, A.L., et al. *Additional limitations of the clustering validation method figure of merit*. in *Proceedings of ACM Southeast Regional Conference*. 2007. Winston-Salem, NC, USA: ACM Press.
41. Chen, K. and L. Liu. *ClusterMap: labeling clusters in large datasets via visualization*. in *Proceedings of the ACM IEEE 13th Conference on Information and Knowledge Management (CIKM 2004)*. 2004. Washington, D.C.
42. Hasegawa, T., S. Sekine, and R. Grishman. *Discovering relations among named entities from large corpora*. in *Proceeding of the 42nd Annual Meeting on Association for Computational Linguistics (ACL 2004)*. 2004. Barcelona, Spain: Association for Computational Linguistics.
43. Jickels, T. and G. Kondrak. *Unsupervised labeling of noun clusters*. in *Proceedings of the 19th Conference of the Canadian Society for Computational Studies of Intelligence Advances in Artificial Intelligence*. 2006. Québec, Canada: Springer.
44. Treeratpituk, P. and J. Callan. *Automatically labeling hierarchical clusters*. in *Proceedings of the International Conference on Digital Government Research*. 2006. San Diego, California: ACM Press.
45. Pantel, P. and D. Ravichandran. *Automatically labeling semantic classes*. in *Proceedings of the Human Language Technology / North American Association for Computational Linguistics (HLT/NAACL-04)*. 2004. Boston, M.A.
46. Skupin, A. and C.d. Jongh. *Visualizing the ICA: a content-based approach*. in *Proceedings of the 22nd International Cartographic Conference*. 2005. A Coruña, Spain: CD-ROM.
47. Stein, B. and S. Meyer. *Topic identification: framework and application*. in *Proceedings of the 4th International Conference on Knowledge Management (I-KNOW 2004)*. 2004. Vienna, Austria.

48. Pal, S.K., V. Talwar, and P. Mitra, *Web mining in soft computing framework: relevance, state of the arte and future directions.* IEEE Transactions on Neural Networks, 2002. 13(5): p. 1163-1177.
49. Skowron, A. and J.F. Peters. *Rough sets: trends and challenges*. in *Proceedings of the 9th International Conference on Fuzzy Sets, Data Mining, and Granular Computing (RSFDGRC 2003)*. 2003. Chongqing, China: Springer.
50. Grabowski, A., *Basic properties of Rough Sets and Rough Membership Function.* Formalized Mathematics, 2004. 12(1): p. 21-28.
51. Grzymala-Busse, J.W. and S. Siddhaye. *Rough set approaches to rule induction from incomplete data*. in *Proceedings of the 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2004)*. 2004. Perugia, Italy.
52. Levine, E. and E. Domany, *Resampling method for unsupervised estimation of cluster validity.* Neural Computation, 2001. 13(11): p. 2573-2593.
53. Anderberg, M.R., *Clustering Analysis for Applications*. 1973: New York: Academic.
54. Höppner, F., et al., *Fuzzy cluster analysis: methods for classification, data analysis and image recognition.* 1999, West Sussex, England: John Wiley & Sons Ltd.
55. Girvan, M. and M.E.J. Newman, *Community structure in social and biological networks.* Proceedings of the National Academy of Sciences (PNAS USA), 2002. 99(12): p. 7821-7826.
56. Baumes, J., A. Goldberg, and M. Magdon-Ismail, *Efficient identification of overlapping communities*, in *Intelligence and Security Informatics*. 2005, Springer Berlin / Heidelberg: Berlin. p. 27-36.
57. Tong, H. and C. Faloutsos. *Center-piece subgraphs: problem definition and fast solutions*. in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2006. Philadelphia, PA, USA ACM Press.
58. Halkidi, M., Y. Batistakis, and M. Vazirgiannis. *Clustering algorithms and validity measures*. in *Proceedings of the 13th International Conference on Scientific and Statistical Database Management*. 2001: IEEE Computer Society.
59. Han, J. and M. Kamber, *Data mining: concepts and techniques*. Data Management Systems. 2001, San Francisco: Morgan Kaufmann.
60. Pedrycz, W., *Knowledge-based Clustering: from Data to Information Granules*. 2005: Wiley.
61. Jain, A.K. and R.C. Dubes, *Algorithms for clustering data*. 1988, Englewood Cliffs, NJ: Prentice Hall.
62. McQueen, J., *Some methods for classification and analysis of multivariate observations*, in *5th Berkeley Symposium on Mathematics*. 1967.
63. Xiong, H., J. Wu, and J. Chen. *K-means clustering versus validation measures: a data distribution perspective*. in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD 2006)*. 2006. Philadelphia, PA, USA: ACM Press.

64. Berry, M.W., *Survey of Text mining: Clustering, Classification, and Retrieval*. 2004, New York, USA: Springer Verlag.

65. Ng, R.T. and J. Han. *Efficient and effective clustering methods for spatial data mining*. in *Proceedings of the 20th International Conference on Very Large Data Bases*. 1994. Santiago de Chile, Chile: Morgan Kaufmann.

66. Agarwal, P.K. and N.H. Mustafa. *k-means projective clustering*. in *Proceedings of the twenty-third ACM SIGMOD-SIGAT-SIGART Symposium on Principles of database systems (PODS 2004)*. 2004. Paris, France: ACM Press.

67. Hochbaum, D.S. and D. Shmoys, *A best possible heuristic for the k-center problem.* Mathematics of Operations Research, 1985. 10(2): p. 180-184.

68. Liu, Y., et al. *An efficient clustering algorithm for small text documents*. in *Seventh International Conference on Web-Age Information Management (WAIM 2006)*. 2006. Hong Kong, China: IEEE Computer Society.

69. Torre, F.d.l. and T. Kanade. *Discriminative cluster analysis*. in *Proceedings of the 23rd International Conference on Machine Learning (ICML 2006)*. 2006. Pittsburgh, Pennsylvania: ACM Press.

70. Bordes, A., et al., *Fast kernel classifiers with online and active learning.* Journal of Machine Learning Research, 2005. 6: p. 1579-1619.

71. Bolelli, L., et al. *A clustering method for web data with multi-type interrelated components*. in *Proceedings of the 16th International Conference on World Wide Web (WWW 2007)*. 2007. Banff, Alberta, Canada: ACM Press.

72. Gower, J.C. and G.J.S. Ross, *Minimum spanning trees and single-linkage cluster analysis.* Applied Statistics, 1969. 18: p. 54-64.

73. Gotlieb, G.C. and S. Kumar, *Semantic clustering of index terms.* Journal of the Association for Computing Machinery (JACM), 1968. 15(4): p. 493-513.

74. Backer, F.B. and L.J. Hubert, *A graph-theoretic approach to goodness-of-fit in complete-link hierarchical clustering.* Journal of the American Statistical Association, 1976. 71: p. 870-878.

75. Cheng, D., et al. *A divide-and-merge methodology for clustering*. in *Proceedings of the 24th ACM SIGMOD SIGACT SIGART Symposium on Principles of Database Systems (PODS 2005)*. 2005. Baltimore, Maryland: ACM Press.

76. Cheng, D., et al., *A divide-and-merge methodology for clustering.* ACM Transaction on Database Systems (TODS), 2006. 31(4): p. 1499-1525.

77. Jonyer, I., D.J. Cook, and L.B. Holder, *Graph-based hierarchical conceptual clustering.* Journal of Machine Learning Research, 2002. 2: p. 19-43.

78. Zhang, T., R. Ramakrishnan, and M. Livny. *BIRCH: An efficient data clustering method for very large databases*. in *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 1996. Montreal, QB, Canada: ACM Press.

79. Guha, S., R. Rastogi, and K. Shim, *CURE: An efficient clustering algorithm for large databases*, in *International Conference on Management of Data*. 1998, ACM Press: Seattle, WA, USA.

80. Boley, D., *Principal Direction Divisive Partitioning.* Data Mining and Knowledge Discovery, 1998. 2(4): p. 325-344.
81. Ester, M., et al. *A density-based algorithm for discovering clusters in large spatial databases with noise*. in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*. 1996. Portland, Orlando, USA: AAAI Press.
82. Hinneburg, A. and D.A. Keim. *An efficient approach to clustering in large multimedia databases with noise*. in *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*. 1998. New York, USA: AAAI Press.
83. Ankerst, M., et al. *OPTICS: Ordering points to identify the clustering structure*. in *International Conference on Management of Data*. 1996. Philadelphia, PA, USA: ACM Press.
84. Kriegel, H.-P. and M. Pfeifle. *Density-based clustering of uncertain data*. in *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD 2005)*. 2005. Chicago, Illinois, USA: ACM Press.
85. Ruiz-Shulcloper, J., E. Alba-Cabrera, and G. Sánchez-Díaz. *DGLC: a density-based global logical combinatorial clustering algorithm for large mixed incomplete data*. in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2000)*. 2000. Honolulu, Hawaii, USA: IEEE Press.
86. Qian, Y., G. Zhang, and K. Zhang. *FA\ÇADE: a fast and effective approach to the discovery of dense clusters in noisy spatial data*. in *SIGMOD '04: Proceedings of the 2004 ACM SIGMOD International Conference on Management of Data*. 2004. Paris, France: ACM Press.
87. Dourisboure, Y., F. Geraci, and M. Pellegrini. *Extraction and classification of dense communities in the web*. in *Proceedings of the 16th International Conference on World Wide Web (WWW 2007)*. 2007. Banff, Alberta, Canada: ACM Press.
88. Wang, W., J. Yang, and R.R. Muntz. *STING: a statistical information grid approach to spatial data mining*. in *Proceedings of the 23rd International Conference on Very Large Data Bases*. 1997. Athens, Greece: Morgan Kaufmann.
89. Sheikholeslami, G., S. Chatterjee, and A. Zhang. *WaveCluster: a multi-resolution clustering approach for very large spatial databases*. in *Proceedings of the 24th International Conference on Very Large Data Bases*. 1998. New York, NY, USA: Morgan Kaufmann.
90. Sheikholeslami, G., S. Chatterjee, and A. Zhang, *WaveCluster: a wavelet-based clustering approach for spatial data in very large databases.* The VLDB Journal - The International Journal on Very Large Databases, 2000. 8(3-4): p. 289-304.
91. Agrawal, R., et al. *Automatic subspace clustering of high dimensional data for data mining applications*. in *Proceedings of the ACM SIGMOD International Conference on Management of Data*. 1998. Seattle, WA, USA: ACM Press.
92. Zhao, Y., C. Zhang, and Y.-D. Shen. *Clustering high-dimensional data with low-order neighbors*. in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI 2004)*. 2004. Beijing, China: IEEE Computer Society.
93. Orlandic, R., Y. Lai, and W.G. Yee. *Clustering high-dimensional data using an efficient and effective data space reduction*. in *Proceedings of the 14th ACM*

International Conference on Information and Knowledge Management. 2005. Bremen, Germany: ACM Press.

94. Bradley, P.S., U. Fayyad, and C. Reina. *Scaling clustering algorithms to large databases*. in *Proceedigns of the 4th International Conference on Knowledge Discovery and Data Mining*. 1998. New York, USA: AAAI Press.
95. Ordonez, C. and E. Omiecinski. *FREM: fast and robust EM clustering for large data sets*. in *Proceedings of the eleventh international conference on Information and knowledge management (CIKM 2002)*. 2002. McLean, Virginia, USA: ACM Press.
96. Gaber, M.M. and P.S. Yu. *A framework for resource-aware knowledge discovery in data streams: a holistic approach with its application to clustering*. in *Proceedings of the ACM Symposium on Applied Computing (SAC 2006)*. 2006. Dijon, France: ACM Press.
97. Ester, M., et al. *Joint cluster analysis of attribute data and relationship data: the connected k-center problem*. in *Proceedings of the 6th SIAM International Conference on Data Mining*. 2006. Bethesda, Maryland.
98. Chee, B. and B. Schatz. *Document clustering using small world communities*. in *Proceedings of the 7th ACM IEEE-CS Joint International Conference on Digital Libraries*. 2007. Vancouver, BC, Canada: ACM Press.
99. Dhillon, I.S. *Co-clustering documents and words using bipartite spectral graph partitioning*. in *KDD 01*. 2001. San Francisco, CA, USA: ACM.
100. Yoo, I., X. Hu, and I.-Y. Song. *Integration of semantic-based bipartite graph representation and mutual refinement strategy for biomedical literature clustering*. in *KDD'06*. 2006. Philadelphia, Pennsylvania, USA: ACM.
101. Gao, B., et al. *Consistent bipartite graph co-partitioning for star-structured high-order heterogeneous data co-clustering*. in *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. 2005. Chicago, Illinois, USA: ACM Press.
102. Deodhar, M. and J. Ghosh. *A framework for simultaneous co-clustering and learning from complex data*. in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007)*. 2007. San Jose, California, USA: ACM Press.
103. Gago, A., J.E. Medina, and A. Pérez. *ACONS: a new algorithm for clustering*. in *Proceedings of the 12th Iberoamerican Congress on Pattern Recognition (CIARP 2007)*. 2007. Viña del Mar, Valparaiso, Chile: Springer-Verlag.
104. Pérez, A. and J.E. Medina. *A clustering algorithm based on generalized stars*. in *Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition (MLDM 2007)*. 2007. Leipzig, Germany: Springer Verlag.
105. Aslam, J., K. Pelekhov, and D. Rus. *Static and dynamic information organization with star clusters*. in *Proceedings of the ACM SIGIR SIGMIS Seventh International Conference on Information and Knowledge Management*. 1998. Bethesda, Maryland, USA: ACM Press.

106. Gil-García, R., J.M. Badía-Contelles, and A. Pons-Porrata. *Extended Star clustering algorithm*. in *Proceedings of the Iberoamerican Congress on Pattern Recognition, Speech and Image Analysis (CIARP 2003)*. 2003. Havana, Cuba: Springer.

107. Arco, L., et al., *Agrupamiento de documentos textuales mediante métodos concatenados.* Revista Iberoamericana de Inteligencia Artificial, 2006. 10(30): p. 43-53.

108. Zahn, C.T., *Graph-theoretical methods for detecting and describing gestalt clusters.* IEEE Transactions on Computers, 1971. 20(1): p. 68-86.

109. Gibson, D., J.M. Kleinberg, and P. Raghavan. *Clustering categorical data: an approach based on dynamical systems*. in *Proceedings of the 24th International Conference on Very Large Data Bases*. 1998. New York, USA: Morgan Kaufmann.

110. Karypis, G., E.-H. Han, and V. Kumar, *CHAMELEON: a hierarchical clustering algorithm using dynamic modeling.* IEEE Computer, 1999. 32(8): p. 68-75.

111. Jenssen, R., et al. *Clustering using Renyi's Entropy*. in *Proceedings of the International Joint Conference on Neural Networks*. 2003: IEEE Press.

112. Gil-García, R.J., J.M. Badía-Contelles, and A. Pons-Porrata. *A general framework for agglomerative hierarchical clustering algorithms*. in *Proceedings of 18th International Conference on Pattern Recognition (ICPR 2006)*. 2006: IEEE Computer Society.

113. Falkowski, T., J. Bartelheimer, and M. Spiliopoulou. *Mining and visualizing the evolution of subgroups in social networks*. in *Proceedings of the IEEE WIC ACM International Conference on Web Intelligence (WI 2006)*. 2006. Washington, DC, USA: IEEE Computer Society.

114. Hu, X. and D.C. Wu, *Data Mining and Predictive Modeling of Biomolecular Network from Biomedical Literature Databases.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2007. 4(2): p. 251-263.

115. Kalisky, T., et al., *Scale-free networks emerging from weighted random graphs.* Physical Review E, 2006. 73: p. 025103.

116. Fortunato, S., L.C. Freeman, and F. Menczer, *Scale-free network growth by ranking.* Physical Review Letters, 2006. 96(21): p. 218701.

117. Stumpf, M.P.H., C. Wiuf, and R.M. May, *Subnets of scale-free networks are not scale-free: sampling properties of networks.* Proceedings of the National Academy of Sciences (PNAS USA), 2005. 102(12): p. 4221-4224.

118. Xu, X., et al. *SCAN: a structural clustering algorithm for networks*. in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007)*. 2007. San Jose, California, USA: ACM Press.

119. Epter, S. and M. Krishnamoorthy. *A multiple-resolution method for edge-centric data clustering*. in *Proceedings of International Conference on Information and Knowledge Management (CIKM 1999)*. 1999. Kansas City, Missouri, USA: ACM Press.

120. Epter, S., M. Krishnamoorthy, and M. Zaki, *Clusterability detection and initial seed selection in large data sets*. 1999, Rensselaer Polytechnic Institute, Computer Science Dept.: Troy, NY 12180.

121. Cortes, C., D. Pregibon, and C. Volinsky. *Communities of interest*. in *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis*. 2001.
122. Aggarwal, C.C. and P.S. Yu. *Online analysis of community evolution in data streams*. in *Proceedings of SIAM International Data Mining Conference (SDM 2005)*. 2005. Newport Beach, California.
123. Wasserman, S. and K. Faust, *Social network analysis: methods and applications*. 1994, Cambridge: Cambridge University Press.
124. Ferrer, R. and R.V. Solé, *Patterns in syntactic dependency networks.* Physical Review E, 2004. 69(5): p. 051915.
125. Ferrer, R. and R.V. Solé, *The small world of human language.* Proceedings of the Royal Society of London B: Biological Science, 2001. 268(1482): p. 2261-2265.
126. Watts, D.J. and S.H. Strogatz, *Collective dynamics of 'small-world' networks.* Nature, 1998. 393(6684): p. 440-442.
127. Newman, M.E.J., *Mixing patterns and community structure in networks*, in *Statistical Mechanics of Complex Networks*, R. Pastor-Satorras, M. Rubi, and A. Diaz-Guilera, Editors. 2003, Springer-Verlag. p. 66-87.
128. Clauset, A., M.E.J. Newman, and C. Moore, *Finding community structure in very large networks.* Physical Review E, 2004. 70(6): p. 066111.
129. Rousseau, R. and L. Zhang, *Betweenness centrality and Q-measures in directed valued networks.* Scientometrics, 2008. 75(3): p. 575-590.
130. Grassi, R., et al., *Betweenness centrality: extremal values and strcutural properties*, in *Networks, topology and dynamics - theory and applications to economics and social systems*, A.K. Naimzada, S. Stefani, and A. Torriero, Editors. 2008, Springer: Haidelberg.
131. Wang, H., J. Martin, and P. Mieghem, *Betweenness centrality in a weighted network.* Physical Review E, 2008. 77(4): p. 046105.
132. Newman, M.E.J., *Scientific collaboration networks. I. Network construction and fundamental results.* Physical Review E, 2001. 64(1): p. 016131.
133. Newman, M.E.J., *Scientific collaboration networks. II. Shortest paths, weighted networks, and centrality.* Physical Review E, 2001. 64(1): p. 016132.
134. Newman, M.E.J., *The structure and function of complex networks.* SIAM Review, 2003. 45(2): p. 167-256.
135. Bavelas, A., *A mathematical model for group structures.* Human Organization, 1948. 7: p. 16-30.
136. Wasserman, S. and K. Faust, *Social Network Analysis*. 1994: Cambridge University Press.
137. Newman, M.E.J., *Who is the best connected scientist? A study of scientific coauthorship networks.* Physical Review E, 2001. 64(1).
138. Brandes, U., *A faster algorithm for betweenness centrality.* Journal of Mathematical Sociology, 2001. 25: p. 163-177.
139. Ahuja, R.K., T.L. Magnanti, and J.H. Orlin, *Network Flows: Theory, Algorithms, and Applications*. 1993, Englewood Cliffs, NJ: Prentice-Hall.

140. Brandes, U., et al., *On Modularity Clustering.* IEEE Transactions on Knowledge and Data Engineering, 2008. 20(2): p. 172-188.
141. Gaertler, M., R. Gorke, and D. Wagner. *Significance-driven graph clustering*. in *Proceedings of the Third International Conference on Algorithmic Aspects in Information and Management (AAIM 2007)*. 2007. Portland, OR, USA: Springer-Verlag.
142. Muff, S., F. Rao, and A. Caflisch, *Local modularity measure for network clusterizations.* Physical Review E, 2005. 72(5): p. 056107.
143. Leicht, E.A. and M.E.J. Newman, *Community structure in directed networks.* Physical Review Letters, 2008. 100(11): p. 118703.
144. Bonacich, P.F., *Power and centrality: a family of measures.* American Journal of Sociology, 1987. 92(5): p. 1170-1182.
145. Stephenson, K.A. and M. Zelen, *Rethinking centrality: methods and examples.* Social Networks, 1989. 11: p. 1-37.
146. Freeman, L.C., S.P. Borgatti, and D.R. White, *Centrality in valued graphs: A measure of betweenness based on network flow.* Social Networks, 1991. 13: p. 141-154.
147. Noh, J.D. and H. Rieger, *Random walks on complex networks.* Physical Review Letters, 2004. 92(11): p. 118701.
148. Zhou, H. and R. Lipowsky. *Network Brownian Motion: A new method to measure vertex-vertex proximity and to identify communities and subcommunities*. in *Proceedings of the International Conference on Computational Science and its Applications (ICCSA 2004)*. 2004. Perugia, Italy: Springer.
149. Rattigan, M.J., M. Maier, and D. Jensen. *Graph clustering with network structure indices*. in *Proceedings of the 24th International Conference on Machine Learning (ICML 2007)*. 2007. Corvalis, Oregon: ACM Press.
150. Brandes, U., *On variants of shortest-path betweenness centrality and their generic computation.* Social Networks, 2008. 30(2): p. 136-145.
151. Gould, R.V. and R.M. Fernández, *Structures of mediation: a formal approach to brokerage in transaction networks.* Sociological Methodology, 1989. 19: p. 89-126.
152. Friedkin, N.E., *Theoretical foundations for centrality measures.* American Journal of Sociology, 1991. 96: p. 1478-1504.
153. Borgatti, S.P. *Types of network flows and how to destabilize terrorist network*. in *Proceedings of Sunbelt International Social Networks Conference*. 2002. New Orleans.
154. Borgatti, S.P., *Centrality and network flow.* Social Networks, 2005. 27(1): p. 55-71.
155. Borgatti, S.P. and M.G. Everett, *A graph-theoretic perspective on centrality.* Social Networks, 2005.
156. Yang, Q. and S. Lonardi. *A parallel algorithm for clustering protein-protein interaction networks*. in *Proceedings of the IEEE Computational Systems Bioinformatics Conference*. 2005. Stanford, California: IEEE Press.
157. Bader, D.A. and K. Madduri. *Parallel algorithms for evaluating centrality indices in real-world networks*. in *Proceedings of the International Conference on Parallel Processing (ICPP 2006)*. 2006.

158. Rattigan, M., M. Maier, and D. Jensen. *Using structure indices for efficient approximation of network properties*. in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2006. Philadelphia, PA, USA: ACM Press.

159. Geisberger, R., P. Sanders, and D. Schultes. *Better approximation of betweenness centrality*. in *Proceedings of the Ninth Workshop on Algorithm Engineering and Experiments (ALENEX 2008)*. 2008. San Francisco, California.

160. Puzis, R., Y. Elovici, and S. Dolev, *Fast algorithm for successive computation of group betweenness centrality.* Physical Review E, 2007. 76(5): p. 056709.

161. Bader, D.A., et al., *Approximating betweenness centrality*, in *Algorithms and models for the web-graph*. 2007, Springer: Berlin. p. 124-137.

162. Freeman, L.C., *Centered graphs and the structure of ego networks.* Mathematical Social Sciences, 1982. 3: p. 291-304.

163. Everett, M. and S.P. Borgatti, *Ego network betweenness.* Social Networks, 2005. 27: p. 31-38.

164. Campbell, D.T., *Ethnocentrism of disciplines and the fish-scale model of omniscience*, in *Interdisciplinary relationships in the social science*, M. Scherif and C. Scherif, Editors. 1969, Aldine Publishing Company: Chicago. p. 328-348.

165. Latora, V. and M. Marchiori, *Efficient behavior of small-world networks.* Physical Review Letters, 2001. 87(19): p. 198701.

166. Latora, V. and M. Marchiori, *A measure of centrality based on the network efficiency.* 2004.

167. Fortunato, S., V. Latora, and M. Marchiori, *Method to find community structures based on information centrality.* Physical Review E, 2004. 70(5): p. 056104.

168. Falkowski, T., J. Bartelheimer, and M. Spiliopoulou. *Community dynamics mining*. in *Proceedings of the 14th European Conference on Information Systems*. 2006: (on CD-ROM).

169. Latora, V. and M. Marchiori, *Vulnerability and Protection of Critical Infrastructures.* 2004.

170. Jain, A.K., M.N. Murty, and P.J. Flynn, *Data clustering: a review.* ACM Computing Surveys, 1999. 31(3): p. 264-323.

171. Hansen, C.D. and C.R. Johnson, eds. *The Visualization handbook*. 2005, Elsevier Academic press.

172. Jolliffe, I.T., *Principal Component Analysis*. Springer Series in Statistics. 1986, New York: Springer.

173. Halkidi, M., Y. Batistakis, and M. Vazirgiannis, *Clustering validity checking methods: Part II.* ACM SIGMOD Record, 2002. 31(3): p. 19-27.

174. Silberschatz, A. and A. Tuzhilin, *What makes patterns interesting in knowledge discovery systems.* IEEE Transactions on Knowledge and Data Engineering, 1996. 8(6): p. 940-974.

175. Tuzhilin, A., *Usefulness, novelty, and integration of interestingness measures*, in *Handbook of Data Mining and Knowledge Discovery*, Z. Klösgen, Editor. 2002, Oxford University Press.
176. Shannon, C.E., *A mathematical theory of communications.* The Bell System Technical Journal, 1948. 27(3): p. 379-423.
177. Rosell, M., V. Kann, and J.E. Litton. *Comparing comparisons: document clustering evaluation using two manual classifications*. in *Proceedings of International Conference on Natural Language processings ICON*. 2004. Hyderabad, India.
178. Steinbach, M., G. Karypis, and V. Kumar. *A comparison of document clustering techniques*. in *Proceedings of 6th ACM SIGKDD World Text Mining Conference*. 2000. Boston: ACM Press.
179. Zhao, Y. and G. Karypis, *Criterion functions for document clustering: experiments and analysis*, in *Technical Report TR #01–40*. 2003, Department of Computer Science, University of Minnesota: Crookston, US.
180. Frakes, W.B. and R. Baeza-Yates, *Information Retrieval. Data Structure & Algorithms*. 1992, New York: Prentice Hall.
181. Rosell, M., V. Kann, and J.-E. Litton. *Comparing comparisons: document clustering evaluation using two manual classifications*. in *Proceedings of the International Conference on Natural Language Processing (ICON 2004)*. 2004. Hyderabad, India: Allied Publishers.
182. Larsen, B. and C. Aone. *Fast and effective text mining using linear-time document clustering*. in *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1999. San Diego, California, USA: ACM Press.
183. Niu, Z.-Y., D.-H. Ji, and C.-L. Tan. *Document clustering based on cluster validation*. in *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management (CIKM 2004)*. 2004. Washington, D.C., USA: ACM Press.
184. Banerjee, A., et al. *Model based overlapping clustering*. in *Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 2005. Chicago, Illinois, USA: ACM Press.
185. Xu, W. and Y. Gong, *Document clustering by concept factorization*, in *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2004, ACM Press: Sheffield, United Kingdom.
186. Roussinov, D.G. and H. Chen, *Document clustering for electronic meetings: an experimental comparison of two techniques.* Decision Support Systems, 1999. 27(1-2): p. 67-79.
187. Kuncheva, L. and S. Hadjitodorov. *Using diversity in cluster ensembles*. in *Proceedings of IEEE International Conference on Systems, Man and Cybernetics*. 2004. The Hague, The Netherlands: IEEE Press.
188. Brun, M., et al., *Model-based evaluation of clustering validation measures.* Pattern Recognition, 2007. 40: p. 807-824.
189. Goodman, L. and W. Kruskal, *Measures of associations for cross-validations.* Journal of the American Statistical Association, 1954. 49: p. 732-764.

190. Hubert, L. and J. Schultz, *Quadratic assignment as a general data-analysis strategy.* British Journal of Mathematical and Statistical Psychology., 1976. 29(2): p. 190-241.
191. Akaike, H., *A new look at the statistical model identification.* IEEE Transactions on Automatic Control, 1974. 19(6): p. 716-723.
192. Schwartz, G., *Estimation the dimension of a model.* Annals of Statistics, 1978. 6(2): p. 461-464.
193. Bock, H., *On significance tests in cluster analysis.* Journal of Classification, 1985. 2: p. 77-108.
194. Calinski, R.B. and J. Arabas, *A dendrite method for cluster analysis.* Communications in Statistics, 1974. 3: p. 1-27.
195. Maulik, U. and S. Bandyopadhyay, *Performance evaluation of some clustering algorithms and validity indices.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002. 24(12): p. 1650-1654.
196. Bezdek, J.C. and N.R. Pal, *Some new indexes of cluster validity.* IEEE Transactions on Systems, Man and Cybernetics - Part B: Cybernetics, 1998. 28(3): p. 301-315.
197. Pal, N.R. and J. Biswas, *Cluster validation using graph theoretic concepts.* Pattern Recognition, 1997. 30(6): p. 847-857.
198. Mali, K. and S. Mitra, *Clustering and its validation in a symbolic framework.* Pattern Recognition Letters, 2003. 24(14): p. 2367-2376.
199. Kim, D.J. and Y.W. Park, *A novel validity index for determination of the optimal number of clusters.* IEEE Transactions on Information and Systems, 2001. E84-D(2): p. 281-285.
200. Xie, X.L. and G. Beni, *A validity measure for fuzzy clustering.* IEEE Transactions on Pattern Analysis and Machine Learning, 1991. 13(4): p. 841-846.
201. Dave, R.N., *Validating fuzzy partitions obtained through c-shells clustering.* Pattern Recognition Letters, 1996. 17: p. 613-623.
202. Milligan, G.W. and M.C. Cooper, *An examination of procedures for determining the number of clusters in a data set.* Psychometrika, 1985. 50(2): p. 159-179.
203. Xie, Y., et al., *A new fuzzy clustering algorithm for optimally finding granular prototypes.* International Journal of Approximate Reasoning, 2005. 40(1-2): p. 109-124.
204. Halkidi, M. and M. Vazirgiannis. *Clustering validity assessment: finding the optimal partitioning of a data set*. in *Proceedings of the IEEE International Conference on Data Mining (ICDM 2001)*. 2001. California, USA: IEEE Computer Society.
205. Kim, M. and R.S. Ramakrishna, *New indices for cluster validity assessment.* Pattern Recognition Letters, 2005. 26(15): p. 2353-2363.
206. Rousseeuw, P.J., *Silhouettes: a graphical aid to the interpretation and validatio of cluster analysis.* Journal of Computational Applied Mathematics, 1987. 20: p. 53-65.
207. Sharma, S.C., *Applied Multivariate Techniques*. 1996: John Wiley and Sons.
208. Jonnalagadda, S. and R. Srinivasan, *An information theory approach for validating clusters in microarray data.* Bioinformatics, 2005. 21(21): p. 3993-3999.

209. Yeung, K.Y., D.R. Haynor, and W.L. Ruzzo, *Validating clustering for gene expression data.* Bioinformatics, 2001. 17(4): p. 309-318.

210. Har-even, M. and V.L. Brailovsky, *Probabilistic validation approach for clustering.* Pattern Recognition Letters, 1995. 16(11): p. 1189-1196.

211. Günter, S. and H. Bunke, *Validation indices for graph clustering.* Pattern Recognition Letters, 2003. 24(8): p. 1107-1113.

212. Kannan, R., S. Vempala, and A. Vetta, *On clusterings: good, bad and spectral.* Journal of the ACM (JACM), 2004. 51(3): p. 497-515.

213. Brás-Silva, H., P. Brito, and J.P.d. Costa, *A partitional clustering algorithm validated by a clustering tendency index based on graph theory.* Pattern Recognition, 2006. 39: p. 776-788.

214. Lam, B. and H. Yan, *Assessment of microarray data clustering results based on a new geometrical index for cluster validity.* Soft Computing, 2007. 11: p. 341-348.

215. Lam, B. and H. Yan. *A new cluster validity index for data with merged clusters and different densities*. in *Proceedings of the International Conference on Systems, Man and Cybernetics*. 2005. Hawaii, USA: IEEE Transactions on Systems, Man, and Cybernetics.

216. Lam, B. and H. Yan. *Cluster validity of DNA microarray data using a geometrical based index*. in *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics (ICMLC 2005)*. 2005. Guangzhou, China: IEEE Press.

217. Lange, T., et al., *Stability-based model selection*, in *Advances in neural information processing systems*, S. Becker, S. Thrun, and K. Obermayer, Editors. 2003. p. 617-624.

218. Greco, S., B. Matarazzo, and R. Slowinski, *Rough sets theory for multicriteria decision analysis.* European Journal of Operational Research, 2001. 129(1): p. 1-47.

219. Orlowska, E., ed. *Incomplete Information: Rough Sets Analysis*. 1998, Physica-Verlag: Berlin.

220. Freeman, L.C., *A set of measures of centrality based upon betweenness.* Sociometry, 1977. 40: p. 35-41.

221. Freeman, L.C., *Centrality in social networks: I. Conceptual clarification.* Social Networks, 1979. 1: p. 215-239.

222. Zachary, W.W., *An information flow model for conflict and fission in small groups.* Journal of Anthropological Research, 1977. 33(4): p. 452-473.

223. Salton, G., A. Wong, and C.S. Yang, *A vector space model for automatic text retrieval.* Communications of the ACM, 1975. 18(11): p. 613-620.

224. Arco, L., *Modelo para el agrupamiento de documentos afines y su ulterior resumen a través de la representación espacio vectorial de un corpus textual*, in *Departamento Ciencia de la Computación*. 2005, Universidad Central "Marta Abreu" de Las Villas: Santa Clara, Villa Clara. p. 132.

225. Pawlak, Z., *Rough sets.* International Journal of Computer and Information Sciences, 1982. 11(5): p. 341-356.

226. Komorowski, J., Z. Pawlak, and L. Polkowski, *Rough sets: a tutorial*, in *Rough-Fuzzy Hybridization: A New Trend in Decision Making*, S.K. Pal and A. Skowron, Editors. 1999, Springer-Verlag: Singapore. p. 3-98.
227. Bazan, J., H.S. Nguyen, and M. Szczuka, *A view on rough set concept approximations.* Fundamenta Informatica, 2004. 59(2-3): p. 107-118.
228. Pawlak, Z., et al., *Rough sets.* Communications of the ACM, 1995. 38(11): p. 89-95.
229. Slowinski, R. and D. Vanderpooten, *Similarity relation as a basis for rough approximations*, in *Advances in Machine Intelligence & Soft-Computing*, P.P. Wang, Editor. 1997. p. 17-33.
230. Pawlak, Z., *Rough Sets: Theoretical Aspects of Reasoning about Data*. 1991, Dordrecht: Academic Publishers.
231. Pawlak, Z., *Vagueness and uncertainty: a rough set perspective.* Computational Intelligence: an International Journal, 1995. 11: p. 227-232.
232. Pal, S.K. and A. Skowron, eds. *Rough Fuzzy Hybridization: A New Trend in Decision Making*. 1999, Springer.
233. Zhong, N., A. Skowron, and S. Ohsuga, eds. *Proceedings of the 7th International Workshop on New Directions in Rough Sets, Data Mining, and Granular Soft-Computing*. Lecture Notes in Computer Science. 1999, Springer-Verlag: London, UK.
234. Arco, L., R. Bello, and M. Artiles. *New clustering validity measures based on rough set theory*. in *Proceedings of International Symposium on Fuzzy and Rough Sets (ISFUROS'06)*. 2006. Santa Clara, Cuba.
235. Arco, L., R. Bello, and M.M. García. *On clustering validity measures and the rough set theory*. in *Proceedings of the Fifth Mexican International Conference on Artificial Intelligence (MICAI'06)*. 2006. Apizaco, México: IEEE Computer Society.
236. Arco, L., R. Bello, and M. Artiles. *Un nuevo enfoque del uso de los conjuntos aproximados en la solución de problemas de la minería de textos*. in *VII Conferencia Científica Internacional de la Universidad de Ciego de Ávila (UNICA2006)*. 2006. Ciego de Ávila, Cuba.
237. Komorowski, J., et al., *A Rough Set Perspective on Data and Knowledge*, in *The Handbook of Data Mining and Knowledge Discovery*, W. Klösgen and J. Zytkow, Editors. 1999, Oxford University Press.
238. Skowron, A. *Rough sets in KDD*. in *Proceedings of the Conference on Intelligent Information Processing (IIP2000)*. 2000. Beijing: Publishing House of Electronic Industry.
239. Liang, J., Z. Shi, and D. Li, *Applications of inclusion degree in rough set theory.* International Journal of Computational Cognition, 2003. 1(2): p. 67-78.
240. Caballero, Y., et al. *Nuevas medidas de la teoría de los conjuntos aproximados para la evaluación de sistemas de información en Bioinformática*. in *II Congreso Internacional de Bioinformática y Neuroinformática. XII Convención y Expo Internacional Informática'07*. 2007. La Habana, Cuba.

241. Pawlak, Z. and A. Skowron, *Rough membership functions*, in *Advances in the Dempster-Shafer Theory of Evidence*, R. Yager, M. Fedrizzi, and J. Kacprzyk, Editors. 1994, Wiley: New York. p. 251-271.
242. Caballero, Y., et al. *New measures for evaluationg decision systems using rough set theory: the application in seadonal weather forecasting*. in *Proceedings of the Third International ICSC Symposium on Information Technologies in Environmental Engineering (ITEE'07)*. 2007. Carl von Ossietzky Universität Oldenburg, Alemania: Springer Verlag.
243. Caballero, Y., *Aplicación de la teoría de los conjuntos aproximados en el proprocesamiento de los conjuntos de entrenamiento para algoritmos de aprendizaje*, in *Departamento de Ciencia de la Computación*. 2007, Universidad Central "Marta Abreu" de Las Villas: Santa Clara, Villa Clara.
244. Grau, R., C. Correa, and M. Rojas, *Metodología de la investigación*. 2 ed. 2004, Ibagué, Colombia: El Poira.
245. Hernández-Sampieri, R., *Fundamentos de la metodología de la investigación*. 1 ed. 2007, Madrid: McGraw-Hill Interamericana de España.
246. Hernández-Sampieri, R., C. Fernández-Collado, and P. Baptista-Lucio, *Metodología de la investigación*. 1 ed. 2006, México: McGraw-Hill Interamericana de México.
247. Popescul, A. and L. Ungar (2000) *Automatic labeling of document clusters*.
248. Glover, E., et al. *Inferring hierarchical descriptions*. in *Proceedings of the ACM CIKM International Conference on Information and Knowledge Management*. 2002. McLean, Las Vegas, USA: Springer-Verlag.
249. Cutting, D.R., D.R. Karger, and J.O. Pederson. *Constant interaction-time Scatter/Gather browsing of very large document collections*. in *Proceedings of ACM SIGIR 16th Annual International Conference on Research and Development in Information Retrieval*. 1993. Pittsburgh, Pennsylvania, United States.
250. Choo, C.W., B. Detlor, and D. Turnbull, *Web Work: Information Seeking and Knowledge Work on the World Wide Web*. Information Science and Knowledge Management. 2000: Klumer Academic Publishers.
251. Tiwana, A., *The Knowledge Management Toolkit*. 2000: Prentice Hall Inc.
252. Müller, R.M., M. Spiliopoulou, and H.-J. Lenz. *The influence of incentives and culture on knowledge sharing*. in *Proceedings of the 38th International Conference on System Sciences (HICSS-38 2005)*. 2005. Big Island, Hawaii, USA: IEEE Computer Society.
253. Fürnkranz, J., T. Scheffer, and M. Spiliopoulou. *Knowledge discovery in databases*. in *Proceedings of the 10th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2006)*. 2006. Berlin, Germany: Springer.
254. Nonaka, I. and H. Takeuchi, *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovations*. 1995: Oxford University Press.
255. Müller, R.M., M. Spiliopoulou, and H.-J. Lenz. *Electronic marketplaces of knowledge: Characteristics and sharing of knowledge assets*. in *Proceedings of the International Conference on Advances in Infrastructure for e-Business (SSGRR 2002)*. 2002. L'Aquila, Italy.

256. Bueno, E. *Gestión del conocimiento y capital intelectual: análisis de experiencias en la empresa española*. in *Actas del X Congreso AECA*. 1999. Zaragoza, España.

257. Manning, C.D., P. Raghan, and H. Schütze, *Introduction to Information Retrieval*. 2008: Cambridge University Press.

258. Lewis, D.D., *Representation and learning in information retrieval*, in *Department of Computer and Information Science*. 1992, University of Massachasetts: Massachusetts, USA.

259. Arco, L., et al. *CorpusMiner 2.0: Aplicación de la Inteligencia Artificial en el procesamiento textual*. in *IV Conferencia Internacional de Matemática y Computación (COMPUMAT 2007)*. 2007. Holguín, Cuba: Sociedad Cubana de Matemática y Computación.

260. Arco, L., et al. *CorpusMiner 2.0: Agrupamiento, extracción de palabras claves y obtención de la aproximación inferior y superior de los grupos textuales homogéneos*. in *II Simposio Internacional sobre Tecnologías de la Información en las Organizaciones Informacionales (SITIO 2007)*. 2007. Santa Clara, Cuba: Editorial Samuel Feijó.

261. Salton, G. and C. Buckley, *Term weighting approaches in automatic text retrieval.* Information Processing and Management, 1988. 24(5): p. 513-523.

262. Lewis, D.D. and M. Ringuette. *A comparison of two learning algorithms for text classification*. in *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval*. 1994. University of Nevada, Las Vegas.

263. Nigam, K., et al., *Text classification from labelled and unlabeled documents using EM.* Machine Learning, 2000. 39(2/3): p. 103-134.

264. Yang, Y. and J.O. Pedersen. *A comparative study on feature selection in text categorization*. in *Proceedings of the Fourteenth International Conference on Machine Learning*. 1997. San Francisco, US: Morgan Kaufmann Publishers.

265. Mladenic, D. and M. Grobelnik. *Feature selection for classification based on text hierarchy*. in *Working Notes of Learning from Text and the Web: Conference on Automatic Learning and Discovery (CONALD-98)*. 1998. Carnegie Mellon University, Pittsburgh, PA.

266. Zipf, G.K., *Human Behaviour and the Principle of Least Effort*. 1949: Addison-Wesley.

267. Porter, M.F., *An algorithm for suffix stripping.* Program, 1980. 14(3): p. 130-137.

268. Landauer, T.K. and S.T. Dumais, *A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge.* Psychological Review, 1997. 104(2): p. 211-240.

269. Schaal, M., et al. *RELFIN - Topic discovery for ontology enhancement and annotation*. in *Proceedings of The Semantic Web: Research and Applications, Second European Semantic Web Symposium (ESWC 2005)*. 2005. Heraklion, Crete, Greece: Springer.

270. Spiliopoulou, M., et al. *Evaluation of Ontology Enhancement Tools*. in *Proceedings of the Semantics, Web and Mining, Joint International Workshops (EWMF/KDO 2005)*. 2005. Porto, Portugal: Springer.

271. Manning, C. and H. Shütze, *Foundations of Statistical Natural Language Processing*. 2000: MIT Press.
272. Arco, L., D. Magdaleno, and R. Bello, *Keywords extraction in clusters of related documents.* Research in Computer Science, 2007. 27: p. 137-148.
273. Mertins, K., P. Heisig, and J. Vorbeck, eds. *Knowledge Management - Concepts and Best Practices*. 2nd edition ed. 2003, Springer-Verlag: Berlin.
274. Spiliopoulou, M., B. Berendt, and E. Menasalvas. *Tutorial Evaluation in Web Mining*. in *the 15th European Conference on Machine Learning and the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2004)*. 2004. Pisa, Italy: Springer Verlag.
275. Ebert, C., et al., *Best Practices in Software Measurement: How to Use Metrics to Improve Project and Process*. 2005: Springer.
276. Grassmann, W.K. and J.P. Tremblay, *Matemática Discreta y Lógica. Una perspectiva desde la ciencia de la computación.* 2003, España: Prentice Hall.
277. Batchelor, B., *Pattern Recognition: Ideas in Practice*. 1978, New York: Plenum Press.
278. Hand, D.J., *Discrimination and classification*. Wiley Series in Probability and Statistics. 1981: John Wiley and Sons.
279. Reed, S.K., *Pattern recognition and categorization.* Cognitive Pshychology, 1972. 3: p. 382-407.
280. Michalski, R.S., R.E. Stepp, and E. Diday, *A recent advance in data analysis: clustering objects into classes characterized by conjuntive concepts.* Progress in Pattern Recognition, 1981. 1: p. 33-56.
281. Diday, E. *Recent progress in distance and similarity measures in pattern recognition*. in *Second International Joint Conference on Pattern Recognition*. 1974. Copenhagen, Denmark.
282. Wilson, D.R. and T.R. Martínez, *Improved heterogeneous distance functions.* Journal of Artificial Intelligence Research, 1997. 6: p. 1-34.
283. Duch, W., *Similarity-based methods: a general framework for classification.* Control and Cybernetics, 2002. 29(4): p. 937-968.
284. Strehl, A., J. Ghosh, and R. Mooney. *Impact of similarity measures on Web-page clustering*. in *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI 2000): Workshop of Artificial Intelligence for Web Search*. 2000. Austin, Texas.
285. Andrade, R.F.S., J.G.V. Miranda, and T.P. Lobão, *Neighborhood properties of complex networks.* Physical Review E, 2006. 73(4): p. 046101.
286. Goh, K.I., et al., *Classification of scale free networks.* Proceedings of the National Academy of Sciences (PNAS USA), 2002. 99(20): p. 12583-12588.
287. Amaral, L.A.N. and J.M. Ottino, *Complex networks. Augmenting the framework for the study of complex systems.* The European Physical Journal B, 2004. 38: p. 147-162.
288. Lehmann, K.A., H.D. Post, and M. Kaufmann, *Hybrid graphs as a framework for the small-world effect.* Physical Review E, 2006. 73(5): p. 056108.
289. Newman, M.E.J., *Models of the Small World. A Review.* Journal of Statistical Physics, 2000. 101(3/4): p. 819-841.

290. Newman, M.E.J., *Mixing patterns in networks.* Physical Review E, 2003. 67(22): p. 026126.
291. Xu, X.-J., Z.-X. Wu, and Y.-H. Wang, *Properties of weighted complex networks.* International Journal of Modern Physics C, 2006. 17(4): p. 521-529.
292. Park, J. and M.E.J. Newman, *Solution for the properties of a clustered network.* Physical Review E, 2005. 72(2): p. 026136.
293. Gastner, M.T. and M.E.J. Newman, *The spatial structure of networks.* The European Physical Journal B, 2006. 49(2): p. 247-252.
294. Newman, M.E.J., *The structure and function of networks.* Computer Physics Communications, 2002. 147(1): p. 40-45.
295. Leicht, E.A., P. Holme, and M.E.J. Newman, *Vertex similarity in networks.* Physical Review E, 2006. 73(2): p. 026120.
296. Barrat, A., M. Barthélemy, and A. Vespignani, *Weighted evolving networks: coupling topology and weight dynamics.* Physical Review Letters, 2004. 92(22): p. 228701.
297. Newman, M.E.J. and D.J. Watts, *Renormalization group analysis of the small-world network model.* Physics Letters A, 1999. 263(4): p. 7332-7342.
298. Pool, I.S. and M. Kochen, *Contacts and influence.* Social Networks, 1978. 1(1): p. 5-51.
299. Barabási, A.L. and R. Albert, *Emergence of scaling in random networks.* Science, 1999. 286: p. 509-512.
300. Gupta, S., R.M. Anderson, and R.M. May, *Networks of sexual contacts: Implications for the pattern of spread of HIV.* AIDS, 1989. 3(12): p. 807-817.
301. Shaw, M.E., *Group structure and the behavior of individuals in small groups.* Journal of Psychology, 1954. 38: p. 139-149.
302. Sabidussi, G., *The centrality index of a graph.* Psychometrika, 1966. 31(4): p. 581-603.
303. Anthonisse, J.M., *The rush in a directed graph*. 1971, Stichting Mathematicsh Centrum, Amsterdam.
304. Scott, J., *Social Network Analysis: A Handbook*. 2nd. Edition ed. 2000, London: Sage Publications.
305. Holme, P., *Edge overload breakdown in evolving networks.* Physical Review E, 2002. 66(2): p. 036119.
306. García, M.M., *Monografía de reconocimiento de patrones*. 1999, Universidad Central "Marta Abreu" de Las Villas: Santa Clara, Villa Clara.
307. Blair, D., *Information retrieval and the philosophy of language.* The Computer Journal, 1992. 35(3): p. 200-207.
308. Sahami, M., *Using machine learning to improve informatio access*, in *Department of Computer Science*. 1998, Stanford University: Standford, USA.
309. Rijsberguen, C.J., *Information Retrieval*. 1979, London: Butterworths.
310. Salton, G. and M. McGill, *Introduction to Modern Information Retrieval*. 1983, New York, USA: McGraw-Hill.

311. Nürnberger, A., A. Klose, and R. Kruse. *Clustering of document collection to support interactive text exploration*. in *Proceedings of the 25th Annuals Conference of the Gesellschaft für Klassification. Studies in Classification, Data Analysis and Knowledge Organization. Exploratory Data Analysis in Empirical Research*. 2001. Germany.

312. Fukuhara, T., H. Takeda, and T. Nishida. *Multiple-text summarization for collective knowledge formation*. in *Proceedings of the IEEE International Conference om Systems, Man and Cybernetics*. 1999. Tokyo: IEEE Press.

Producción científica de la autora sobre el tema de la tesis

Publicaciones en libros, revistas y memorias de eventos (en orden cronológico)

1. L. Arco y otros. "Uso de árboles binarios para la selección de rasgos". Memorias del evento Informática'2004. ISBN 959-237-117-2. La Habana. 2004.
2. L. Arco y otros. "TextLynx: Analizador avanzado de textos". Memorias del evento Informática'2004. ISBN 959-237-117-2. La Habana. 2004.
3. L. Arco y otros. "TextLynx: Analizador avanzado de textos". XI Exposición Forjadores del Futuro. Facultad de Matemática, Física y Computación, Universidad Central de Las Villas. 2004. MENCIÓN.
4. L. Arco y otros. "La minería de textos: aplicaciones, ventajas, impacto y retos". COMPUMAT de las provincias centrales. Universidad Central de Las Villas. 2004.
5. L. Arco y otros. "La minería de datos y de textos: su necesidad y aplicación en la gestión empresarial". Memorias de la IV Conferencia Internacional de Ciencias Empresariales. Primer Simposio de Informática Empresarial. ISBN 952-250-159-9. Centro de Convenciones Bolivar. Santa Clara. 2004.
6. L. Arco y otros. "CorpusMiner: Herramienta para el agrupamiento de documentos". Memorias de la I Conferencia Científica de las Ciencias Informáticas. I Taller de Inteligencia Artificial. UCIENCIA 2005. ISBN: 959-16-0318-5. Universidad de las Ciencias Informáticas. La Habana. 2005.
7. L. Arco y otros. "Rough Text y su aplicación en la validación del agrupamiento de documentos". IV Congreso Nacional de Reconocimiento de Patrones. RECPAT 2006. Universidad de las Ciencias Informáticas. La Habana. 2006.
8. L. Arco y otros. "Un nuevo enfoque del uso de los conjuntos aproximados en la solución de problemas de la minería de textos". Memorias de la VII Conferencia Científica Internacional de la Universidad de Ciego de Ávila. UNICA 2006. ISBN 959-16-0473-4. Ciego de Ávila. 2006.
9. L. Arco y otros. "On clustering validity measures and the Rough Set Theory". 5th Mexican International Conference on Artificial Intelligence. MICAI 2006. Apizaco,

México. Publicación en: IEEE Computer Society. ISBN 0-7695-2722-1. p. 168-177. 2006.

10. L. Arco y otros. "Rough Text: a rough approach in text processing". Memorias de International Symposium on Fuzzy and Rough Sets. ISFUROS 2006. ISBN 959-250-308-7. Santa Clara. 2006.
11. L. Arco y otros. "New clustering validity measures based on Rough Set Theory". Memorias de International Symposium on Fuzzy and Rough Sets. ISFUROS 2006. ISBN 959-250-308-7. Santa Clara. 2006.
12. L. Arco y otros. "Agrupamiento de documentos textuales mediante métodos concatenados". Revista Iberoamericana de Inteligencia Artificial. ISSN: 1137-3601 (c) AEPIA. Vol. 10 No. 30. p. 43-53. 2006.
13. Y. Caballero, L. Arco y otros. "New measures for evaluating decision systems using Rough Set Theory". Memorias de Information Technologies in Environmental Engineering. ITEE 2007. Universidad Carl von Ossietzky University, Oldenburgo, Alemania. Publicación en: Springer. ISSN 1863-5520. Editores: Jorge Marx-Gómez, Michael Sonnenschein, Martin Müller, Heinz Welsch, Claus Rautenstrauch. Enviromental Science and Engineering. Subseries: Environmental Engineering. p. 161-173. 2007.
14. Y. Caballero, L. Arco y otros. "Nuevas medidas de la Teoría de los Conjuntos Aproximados para la Evaluación de Sistemas de Información en Bioinformática". Memorias de II Congreso Internacional de Bioinformática y Neuroinformática. XII Convención y Expo Internacional Informática'2007. ISBN 978-959-286-002-5. La Habana. 2007.
15. L. Arco y otros. "Diferentes enfoques y tendencias actuales del etiquetamiento de grupos textuales". II Simposio Internacional sobre Tecnologías de la Información en las Organizaciones Informacionales. SITIO 2007. Santa Clara. 2007.
16. L. Arco y otros. "CorpusMiner 2.0: Agrupamiento, extracción de palabras claves y obtención de la aproximación inferior y superior de los grupos textuales homogéneos". Memorias de II Simposio Internacional sobre Tecnologías de la Información en las

Organizaciones Informacionales. SITIO 2007. ISBN 978-959-250-326-7. Santa Clara. 2007.

17. L. Arco y otros. "Keywords extraction in clusters of related documents". 8th Conference on Computing. CORE 2007. MEJOR ARTÍCULO – TERCER LUGAR. Ciudad México. México. Publicación en la revista: Research in Computing Science. Advances in Computer Science and Engineering. Vol. 27. p. 137-148. ISSN 1870-4069. http://www.ipn.mx, http://www.cic.ipn.mx. 2007.
18. L. Arco y otros. "CorpusMiner 1.0: Herramienta para el agrupamiento de documentos". Revista Cubana de las Ciencias Informáticas. ISSN: 1994-1536. Vol. 1. No.2. p. 18-31. 2007.
19. L. Arco y otros. "Sistema de gestión de información científica". Forum de Ciencia y Técnica del Centro de Estudios de Informática. MENCION. Santa Clara. 2007.
20. R. Bello, L. Arco y otros. "El aprendizaje automático en la gestión del conocimiento. Una aplicación en el trabajo universitario". Evento provincial de la Conferencia Científica Internacional Universidad 2008. RELEVANTE. Santa Clara. 2007.
21. Y. Caballero, L. Arco y otros. "Nuevas medidas de la Teoría de los Conjuntos Aproximados para la evaluación de sistemas de información en Bioinformática". III Conferencia Científica de la Universidad de las Ciencias Informáticas. Uciencia 2007. Universidad de las Ciencias Informáticas. La Habana. 2007.
22. L. Arco y otros. "Sistema de Gestión de Información Científica: aplicación en el trabajo universitario". Segundo Taller Nacional de Actualización e Intercambio de Experiencias en Ciencias, Tecnologías, Gestión de Información y Gestión del Conocimiento de los Polos Científicos. INFOPOLO 2007. La Habana. 2007.
23. L. Arco y otros. "CORPUSMINER 2.0: aplicación de la Inteligencia Artificial en el procesamiento textual". Memorias de IV Conferencia internacional de matemática y computación. COMPUMAT 2007. X Congreso nacional de matemática y computación. ISBN 1728-6042. Holguín. 2007.
24. L. Arco y otros. "Rough Text Assisting Text Mining: Focus on Document Clustering Validity". Capítulo del libro: Granular Computing: At the Junction of Rough Sets and

Fuzzy Sets. Series: Studies in Fuzziness and Soft Computing, Vol. 224. Editores: Bello, R.; Falcón, R.; Pedrycz, W.; Kacprzyk, J. ISBN: 978-540-76972-9. Springer Berlin / Heidelberg. ISSN: 1434-9922 (Print) 1860-0808 (Online). p. 229-248. 2008.

25. R. Bello, L. Arco y otros. "El aprendizaje automático en la gestión del conocimiento". Universidad 2008. La Habana. 2008.

    http://e-spacio.uned.es/fez/list.php?collection_pid=bibliuned:19818.

26. A. Ochoa y L. Arco. "Differential Betweenness in Complex Networks Clustering". Memorias de XIII Iberoamerican Congress on Pattern Recognition. La Habana. Publicación en el libro Progress in Pattern Recognition, Image Analysis and Applications. Lecture Notes in Computer Science LNCS 5197. Editores: Ruiz-Shulcloper, J.; Kropatsch, W. Springer. p. 227-234. 2008.

Trabajo aprobado en evento, no se asistió por problemas de financiamiento

L. Arco y otros. "Rough Set Theory Measures for Evaluating Decision Systems". 5th International Conference on Machine Learning and Data Mining in Pattern Recognition. Leipzing, Alemania. 2007.

Registros de software

- Analizador avanzado de textos (TEXTLYNX) Registro: 2332-2005
- Sistema para la representación, agrupamiento y resumen de corpus textuales en idioma Inglés (CORPUSMINER) Registro: 2333-2005
- Sistema para el agrupamiento, etiquetamiento y evaluación de colecciones textuales (SATEX) Registro: 1145-2008
- Sistema para la gestión de artículos científicos recuperados usando Lucene (GARLucene) Registro: 1777-2008

Anexos

Anexo 1. Terminología

AGRID	*Advanced Grid-based Iso-Density line*
BIRCH	*Balanced Iterative Reducing and Clustering using Hierarchies*
CLARA	*Clustering LARge Applications*
CLARANS	*Clustering LARge ApplicatioNS*
CLIQUE	*CLustering In Queso*
CLONE	*Clustering with Low-Order NEighbors*
CURE	*Clustering Using REpresentatives*
DBSCAN	*Density-Based Spatial Clustering of Applications with Noise*
DENCLUE	*DENsity-based CLUstEring*
EM	*Expectation-Maximization*
FOM	*Figure of Merit*
FREM	*Fast and Robust Expectation Maximization*
Fuzzy SKWIC	*Simultaneous Soft Clustering and Term Weighting of Text Document*
GARDEN	*GAmma Region DENsity clustering in High Dimensionalities*
GG	*Gabriel Graph*
LIUS	*Lucene Index Update Search*
MST	*Minimal Spanning Tree*
OPTICS	*Ordering Points To Identify the Clustering Structure*
PAM	*Partitioning Around Medoids*
PDDP	*Principal Direction Divisive Partitioning*
RMSSTD	*Root-Mean-Square Standard Deviation*
RNG	*Relative Neighbourhood Graph*
RS	*R-Squared*
SCAN	*Structural Clustering Algorithm for Networks*
SD	*Scatter-Distance*
SKWIC	*Simultaneous Keyword Identification and Clustering of text documents*
sPDDP	*Spherical Principal Directions Divisive Partitioning*
SPR	*Semi-Partial R-Squared*
STING	*STatistical INformation Gris*
STIRR	*Sieving Through Iterated Relational Reinforcement*
SVM	*Support Vector Machines*

Anexo 2. Definiciones y notación que se asumen respecto a la Teoría de Grafos

Un **seudografo** $G=(V, E, f)$ es una terna, donde $V\neq\varnothing$ es el conjunto de nodos, E es el conjunto de aristas y f una función definida como $f: E\rightarrow V\text{x}V \cup \{\{u, v\}; u\in V, v\in V\}$.
Dos **aristas** son **adyacentes** si tienen un vértice común. Si e es una arista que asociada a los vértices u y v, u y e son **incidentes** (también lo son v y e); los vértices u y v son **adyacentes;** u y v son **extremos** de e**.**
Un seudografo $G=(V, E, f)$ se dice **no dirigido** si todas sus aristas son no dirigidas, es decir, $\forall e\in E,\ f(e)=\{u, v\}$.
Si $f(a)=f(b)=\{u, v\}$ o $f(a)=f(b)=(u, v)$ se dice que a y b son **aristas múltiples**. Un seudografo se denomina **sencillo (o simple)** si no tiene aristas múltiples.
Una arista e se dice **lazo** o **bucle** si $f(a)=(v, v)$ o $f(a)=\{v,v\}$. Se dice **grafo** a un seudografo sencillo sin bucles.

En esta tesis se utilizan grafos no dirigidos $G=(V, E, f)$. En estos casos f es una función inyectiva, por tanto se utiliza la **notación** de la forma $G=(V, E)$, porque para cada arista e, tal que $f(e)=\{u, v\}$ se representa directamente como el par $\{u, v\}\in E$ y la **notación** que se utiliza para la **arista *e*** es $\boldsymbol{u - v}$ que son sus vértices extremos.

Dado un grafo $G=(V, E)$ se llama **subgrafo** a un grafo $G'=(V', E')$ tal que $V'\subseteq V$ y $E'\subseteq E$.
Sea $G=(V, E)$ un grafo, y sea $V'\subseteq V$, el subgrafo cuyos vértices están dados por el conjunto V' y cuyas aristas son todas aquellas aristas de G incidentes solamente a vértices en V', se llama **subgrafo inducido** por V'.
Un grafo $G=(V, E)$ se denomina **bipartito** si $V=V_1\cup V_2$ con $V_1\cap V_2=\varnothing$ tales que no hayan vértices de V_1 que sean adyacentes a vértices de V_1, ni vértices de V_2 que sean adyacentes a vértices de V_2.
Un grafo se dice **ponderado** si cada una de sus aristas tiene asignada una etiqueta. Los grafos de similitud que se utilizan en esta tesis son grafos ponderados con el valor de similitud entre nodos; por ejemplo, similitud coseno entre documentos. **Notación:** G=(V, E, w), donde w: $E \rightarrow \mathbf{R}$ hace corresponder pesos reales a las aristas.
En un grafo no dirigido se llama **camino** a una sucesión de aristas, adyacentes consecutivas si son más de una.
El **vértice** v es **accesible** (o **se puede alcanzar**) desde el vértice u si existe un camino de u a v.
La relación $\Re$ definida sobre el conjunto de nodos V de un grafo no dirigido, donde $u\in V,\ v\in V$, tal que $u\ \Re\ v \Leftrightarrow$ existe un camino de u a v, se denomina **relación de camino**. Esta relación es de equivalencia. Dos vértices están en la misma clase si pueden unirse por un camino y en clases distintas en caso contrario.
Sea $G=(V, E)$ un grafo no dirigido y $V_1, V_2, \ldots, V_r$ la partición de V según $\Re$. Sea E_i $(1\leq i\leq r)$ el subconjunto de E formado por las aristas cuyos extremos están ambos en V_i. Los subgrafos $G_i=(V_i, E_i)$ se conocen como las **componentes conexas** de G. Un grafo no dirigido se dice **conexo** se tiene una única componente conexa.
Se denomina **longitud de un camino** al número de aristas del camino. Se denomina **costo de un camino** en un grafo ponderado a la suma de los pesos de las aristas del camino.
Se denomina **camino de longitud mínima** entre los nodos u y v al camino con menor número de aristas posible entre u y v. Se denomina **camino de costo mínimo** entre los nodos u y v en un grafo ponderado al camino con menor costo posible entre u y v.

Las definiciones han sido tomadas de [276].

Anexo 3. Distancias, similitudes y disimilitudes más usadas para comparar objetos

Sean los objetos O_i y O_j descritos por m rasgos, donde $O_i=(o_{i1}, \ldots, o_{im})$ y $O_j=(o_{j1}, \ldots, o_{jm})$

Distancia Euclidiana

$$D_{Euclideana}(O_i, O_j) = \sqrt{\sum_{k=1}^{m} (o_{ik} - o_{jk})^2} \tag{A3.1}$$

Distancia Minkowski [277]

$$D_{Minkowski}(O_i, O_j) = \left(\sum_{k=1}^{m} |o_{ik} - o_{jk}|^{\gamma} \right)^{\frac{1}{\gamma}} \text{ donde } \gamma \geq 1 \tag{A3.2}$$

La distancia Minkowsky es equivalente a la distancia Manhattan o city-block, y a la distancia Euclidiana cuando γ es 1 y 2, respectivamente [277]. Para los valores de $\gamma \geq 2$, la distancia Minkowsky equivale a Supermum [278, 279].

Distancia Euclidiana heterogénea (Heterogenous Euclidean – Overlap Metric; HEOM)

$$D_{HEOM}(O_i, O_j) = \sqrt{\sum_{k=1}^{m} d_{local}(o_{ik}, o_{jk})^2} \text{ , donde}$$

$$d_{local}(o_{ik}, o_{jk}) = \begin{cases} d_{Overlap}(o_{ik}, o_{jk}) & si\ k\ simbólico \\ d_{NormEuclidean}(o_{ik}, o_{jk}) & si\ k\ numérico \end{cases} \tag{A3.3}$$

$$d_{Overlap}(o_{ik}, o_{jk}) = \begin{cases} 0, & si\ o_{ik} = o_{jk} \\ 1, & en\ otro\ caso \end{cases} \quad \text{y} \quad d_{NormEuclidean}(o_{ik}, o_{jk}) = \frac{|o_{ik} - o_{jk}|}{\max_k - \min_k}$$

Distancia Camberra [280, 281]

$$D_{Camberra}(O_i, O_j) = \sum_{k=1}^{m} \frac{|o_{ik} - o_{jk}|}{|o_{ik} + o_{jk}|} \tag{A3.4}$$

Correlación de Pearson [282]

$$D_{Pearson}(O_i, O_j) = \frac{\sum_{k=1}^{m} (o_{ik} - \overline{atributo_k})(o_{jk} - \overline{atributo_k})}{\sqrt{\sum_{k=1}^{m} (o_{ik} - \overline{atributo_k})^2 \sum_{k=1}^{m} (o_{jk} - \overline{atributo_k})^2}} \tag{A3.5}$$

donde $\overline{atributo_k}$ es el valor promedio que toma el $atributo_k$ en el conjunto de datos.

Las expresiones de Chebychev, Mahalanobis, distancia de Hamming y la máxima distancia son otras variantes de cálculo de distancias entre objetos [282]. En [283] se presentan formas de medir la similitud considerando una relación asimétrica entre los objetos y teniendo en cuenta la probabilidad condicional de un objeto respecto al otro.

A partir de estudios realizados, se ha demostrado que al agrupar documentos, los coeficientes Dice, Jaccard y Coseno, han reportado los mejores resultados [180]. Una valoración del impacto de la distancia Euclidiana y los coeficientes Dice, Jaccard y Coseno en dominios textuales se presenta en [284].

Coeficiente Dice

$$S_{Dice}\left(O_i, O_j\right) = \frac{2\sum_{k=1}^{m}\left(o_{ik} \cdot o_{jk}\right)}{\sum_{k=1}^{m} o_{ik}^2 + \sum_{k=1}^{m} o_{jk}^2} \tag{A3.6}$$

Coeficiente de Jaccard

$$S_{Jaccad}\left(O_i, O_j\right) = \frac{\sum_{k=1}^{m}\left(o_{ik} \cdot o_{jk}\right)}{\sum_{k=1}^{m} o_{ik}^2 + \sum_{k=1}^{m} o_{jk}^2 - \sum_{k=1}^{m}\left(o_{ik} \cdot o_{jk}\right)} \tag{A3.7}$$

Coeficiente Coseno

$$S_{Coseno}\left(O_i, O_j\right) = \frac{\sum_{k=1}^{m}\left(o_{ik} \cdot o_{jk}\right)}{\sqrt{\sum_{k=1}^{m} o_{ik}^2 \cdot \sum_{k=1}^{m} o_{jk}^2}} \tag{A3.8}$$

Anexo 4. Propiedades de las redes

Varias propiedades estructurales han sido estudiadas en los grafos [115, 134, 285-296].

Transividad o agrupamiento. Número elevado de triángulos en el grafo –conjuntos de tres vértices donde cada uno se conecta a los dos restantes. Se cuantifica con el coeficiente de agrupamiento, que mide la densidad de los triángulos en el grafo [134].

Distribución de los grados. El grado de un vértice es el número de aristas incidentes a él. Sea p_k la razón de vértices en el grafo que tienen grado k y p_k la probabilidad que un vértice escogido aleatoria y uniformemente tenga grado k, se llama distribución del grado de los vértices al histograma que se forma al graficar p_k para un grafo dado [134].

Efecto small-world. La mayoría de los pares de vértices se conectan por un camino corto en el grafo [297, 298]. Relacionada con la propiedad de navegación [134].

Grafos scale-free. Grafos que presentan la ley de distribución del grado [299].

Elasticidad del grafo (propiedad relacionada con la distribución del grado). Esta propiedad se analiza al eliminar vértices del grafo. La topología de los grafos y el criterio de selección de los vértices a eliminar influyen en su nivel de elasticidad [134].

Patrones mixtos (assortative mixing). Se cuantifica mediante el coeficiente mixto y se utiliza **c**uando en un grafo los vértices representan distintos tipos y es interesante estudiar cómo se conectan dichos vértices mixtos [127, 300]. La correlación de los grados es un caso especial de esta propiedad, donde se considera la propiedad mixta de los vértices acorde a su grado [134].

Estructura de comunidad. Un grafo muestra estructura de comunidad cuando se observan grupos de vértices que tienen una alta densidad de las aristas entre ellos, con una baja densidad de las aristas entre grupos [134].

Centralidad. Es una propiedad estructural importante de los grafos [135, 136, 220, 221, 301-303]. No existe un consenso de qué es exactamente, sólo hay cierta conciliación sobre los procedimientos apropiados para su medición [221]. Freeman identificó tres variantes de centralidad: el grado de un vértice como índice del potencial de comunicación, su cercanía al resto de los vértices del grafo y su mediación en los caminos de comunicación [221]. Otra clasificación divide las medidas en radiales, aquellas que evalúan los caminos que comienzan

o terminan en un vértice dado, y en mediales, aquellas que cuentan el número de caminos que pasan a través de un vértice o arista dados (por ejemplo, la intermediación) [155]. Algo similar: estar cerca y colocarse o mediar entre otros [166].

Intermediación (<u>betweenness</u>). Definiciones e implementaciones de la intermediación de un vértice se presentan en [133, 137, 220, 221, 286]. La intermediación de un vértice *i* es el número de geodésicos entre otros vértices que pasan por *i* [123, 220, 304]. Para los fines de este trabajo es relevante la extensión de este concepto al caso de las aristas. Inicialmente se introdujo el concepto "rush" asociado a las aristas [303]. Girvan y Newman generalizan al caso de las aristas la propuesta de Freeman para calcular la centralidad de los vértices [220]. La expresión (A4.1) es una variante normalizada para el cálculo de la intermediación $btw(e)$ de una arista *e* [305], donde $cpath(i, j)$ es el número de caminos más cortos entre los nodos *i* y *j* del grafo y $cpath_e(i, j)$ es el número de aquellos que adicionalmente pasan por *e*. Este cociente puede ser interpretado como el rol que juega la arista *e* en la relación entre los nodos *i* y *j*.

$$btw(e) = \frac{cpath_e(i, j)}{cpath(i, j)} \qquad \text{(A4.1)}$$

Anexo 5. Algunas formas de cálculo de la distancia entre dos grupos diferentes

Sean C_i y C_j dos grupos diferentes con cardinalidades n_i y n_j, respectivamente; $\bar{x}$ y $\bar{y}$ los centros de C_i y C_j, respectivamente; y $d(x,y)$ una medida de distancia entre dos objetos cualesquiera x y y, tal que $x \in C_i$, $y \in C_j$, se definen algunas formas de cálculo de la distancia entre dos grupos como sigue [60, 61, 180, 188]:

Enlace simple (single link): Se basa en la distancia mínima entre los objetos petenecientes a los grupos C_i y C_j. El agrupamiento basado en esta distancia es uno de los más usados.

$$d_C(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y) \tag{A5.1}$$

Enlace completo (complete link): Se basa en la distancia entre los dos objetos más lejanos petenecientes a los grupos C_i y C_j.

$$d_C(C_i, C_j) = \max_{x \in C_i, y \in C_j} d(x, y) \tag{A5.2}$$

Enlace promedio (group average link): En contraste con los dos enfoques anteriores, donde la distancia se determina sobre la base de valores extremos de la función de distancia, este método considera el promedio entre todas las distancias calculadas entre todos los pares de objetos, uno de cada grupo. Todos los objetos contribuyen a la distancia entre los grupos.

$$d_C(C_i, C_j) = \frac{1}{n_i n_j} \sum_{x \in C_i, y \in C_j} d(x, y) \tag{A5.3}$$

Método de la varianza mínima (ward link): Une el par de grupos que al combinarse minimizan el incremento en el total del error cuadrático dentro de los grupos, basado en la distancia entre los centros.

Enlace de centros (centroid) y enlace promedio de centros (average to centroid): Se basan en la combinación de pares de grupos comparando los centros según expresiones en (A5.4):

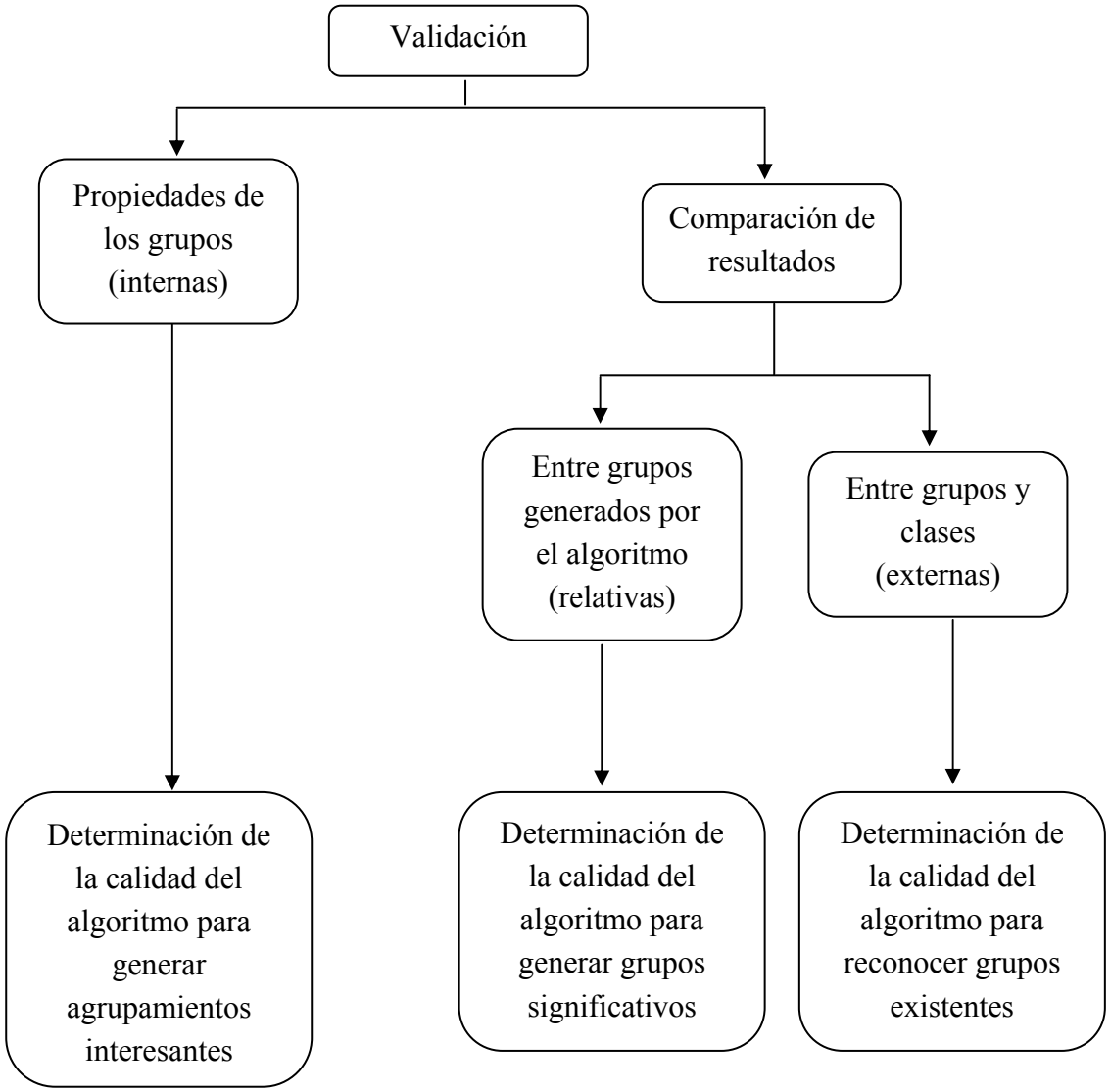
$$d_C(C_i, C_j) = d(\bar{x}, \bar{y}) \quad \text{y} \quad d_C(C_i, C_j) = \frac{1}{n_i + n_j} \left[\sum_{x \in C_i} d(x, \bar{y}) + \sum_{y \in C_j} d(y, \bar{x}) \right] \tag{A5.4}$$

Anexo 6. Algoritmo jerárquico divisivo GN, debido a Girvan y Newman

La forma general del algoritmo es la siguiente [55]:

1. Calcular los valores de intermediación para todas las aristas en el grafo.
2. Encontrar la(s) arista(s) con mayor valor(s) de intermediación y eliminarla(s).
3. Recalcular la intermediación para todas las aristas restantes.
4. Repetir desde el paso 2.

Anexo 7. Clasificación simplificada de algunas técnicas para la validación de agrupamientos[71]



[71] Tomado de 188. Brun, M., et al., *Model-based evaluation of clustering validation measures.* Pattern Recognition, 2007. **40**: p. 807-824.

Anexo 8. Algunas medidas externas para la validación del agrupamiento

Entropía [176, 178, 181] [179], donde m es el número de grupos, n_j el tamaño del grupo j, n el número total de objetos agrupados y E_j se calcula según las expresiones en (A8.2).

$$E_{CS} = \sum_{j=1}^{m} \frac{n_j * E_j}{n} \tag{A8.1}$$

Entropía de un grupo

$$E_j = -\sum_i p_{ij} \log\left(p_{ij}\right) \text{ o } E_j = -\frac{1}{\log m} \sum_{i=1}^{m} p_{ij} \log\left(p_{ij}\right) \tag{A8.2}$$

donde p_{ij} es la probabilidad que un miembro del grupo j pertenezca a la clase i.

Medida-*F* Global (Overall *F*-Measure; OFM) [178]

$$Overall\ F-Measure = \sum_{i=1}^{k} \frac{n_i}{n} \max\left\{F-Measure(i,j)\right\} \tag{A8.3}$$

donde k es el número de clases, n_i es el tamaño de la clase i, n es el número total de objetos agrupados y *F-Measure*(i, j) se calcula según la expresión siguiente. Si $\alpha = 1$, entonces OFM se nombra Purity [177].

Medida-*F* (*F*-Measure) de la clase *i* respecto al grupo *j*

$$F-Measure(i,j) = \frac{1}{\alpha\left(1/\Pr(i,j)\right) + \left(1-\alpha\right)\left(1/\mathrm{Re}(i,j)\right)} \tag{A8.4}$$

Si $\alpha = 1$ entonces *F-Measure*(i,j) coincide con precision, si $\alpha = 0$ entonces *F-Measure*(i,j) coincide con cubrimiento. $\alpha = 0.5$ significa igual peso para precisión y cubrimiento.

Micro-averaged precision y micro-averaged recall [183]

$$\text{MA - Pr} = \frac{\sum_{i=1}^{k} \alpha_i}{\sum_{i=1}^{k} \left(\alpha_i + \beta_i\right)} \quad \text{y} \quad \text{MA - Re} = \frac{\sum_{i=1}^{k} \alpha_i}{\sum_{i=1}^{k} \left(\alpha_i + \gamma_i\right)} \tag{A8.5}$$

donde α_i es el número de objetos correctamente asignados a la clase i, β_i es el número de objetos incorrectamente asignados a la clase i y γ_i es el número de objetos incorrectamente no asignados a la clase i. MA-Pr = MA-Re si cada objeto pertenece a sólo un grupo y la clasificación de referencia también tiene una clasificación única para cada objeto.

Información mutua [185]

$$MI = \sum_{i,j} p_{ij} \cdot \log_2 \frac{p_{ij}}{p_i \cdot p_j} \tag{A8.6}$$

donde p_i y p_j denotan las probabilidades que un objeto pertenezca a la clase i y al grupo j, respectivamente, y p_{ij} denota la probabilidad de que el objeto pertenezca a la clase i y al grupo j simultáneamente. Esta expresión se normaliza dividiendo por la máxima entropía.

Error del agrupamiento normalizado en el intervalo [0, 1] [186]

$$NCE = \frac{E}{A_t}, \text{ donde } A_t = A_m + A_a \tag{A8.7}$$

donde A_t es el número total de asociaciones que existen en ambas particiones sin eliminar duplicados, donde A_m es el número total de asociaciones en la partición de referencia y A_a es el número total de asociaciones en la partición resultado del agrupamiento.

Cluster Recall y Cluster Precision [186]

$$CR = A_c \big/ A_m \text{ y } CP = A_c \big/ A_a \tag{A8.8}$$

donde A_c=A_a-E_i, representa el número total de asociaciones resultantes del agrupamiento.

Rand Statistic

$$R = (a+b)/m \tag{A8.9}$$

Coeficiente de Jaccard

$$J = a/(a+b+c) \tag{A8.10}$$

Índice de Folkes y Mallows

$$FM = \left(\frac{a}{a+b} \cdot \frac{a}{a+c} \right)^{1/2} \tag{A8.11}$$

donde a es el número de pares de objetos que pertenecen al mismo grupo y a la misma clase, b es el número de aquellos pares que pertenecen al mismo grupo y a clases diferentes, c es el total de pares que pertenecen a grupos diferentes y a la misma clase, d es el número de pares de objetos que pertenecen a grupos y clases diferentes y m=a+b+c+d es el número máximo de todos los pares de objetos (es decir, m=n(n-1)/2 donde n es el número total de objetos).

Anexo 9. Algunas medidas internas para la validación del agrupamiento

Similitud global (Overall Similarity) [178]

$$OverallSimilarity(Grupo) = \frac{1}{|Grupo|^2} \sum_{O_i, O_j \in Grupo} \text{distancia}(O_i, O_j) \tag{A9.1}$$

Índices Dunn

$$I(C) = \frac{\min_{i \neq j} \{\delta(C_i, C_j)\}}{\max_{1 \leq l \leq k} \{\Delta(C_l)\}} \tag{A9.2}$$

donde $C=\{C_1, \ldots, C_k\}$ es el agrupamiento de un conjunto de objetos O, $\delta: C \times C \rightarrow \mathrm{R}$ es una medida de distancia de grupo a grupo y $\triangle: C \rightarrow \mathrm{R}$ es una medida de diámetro del grupo.

$$\delta(C_i, C_j) = \min_{x \in C_i, y \in C_j} d(x, y) \text{ y } \Delta(C_i) = \max_{x, y \in C_i} d(x, y) \tag{A9.3}$$

donde $d: C \times C \rightarrow \mathrm{R}$ es una función que mide la distancia entre los objetos de O.

Una de las propuestas de Bezdek para el cálculo de $\delta(C_i, C_j)$ y $\Delta(C_i)$

$$\delta(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i, y \in C_j} d(x, y) \text{ y } \Delta(C_i) = 2\left(\frac{\sum_{x \in C_i} d(x, c_i)}{|C_i|} \right) \tag{A9.4}$$

donde c_i es el centro del grupo C_i.

Índice Davies – Bouldin [34]

$$DB(C) = \frac{1}{k} \cdot \sum_{i=1}^{k} R_i \tag{A9.5}$$

$$R_i = \max_{\substack{j=1,\ldots,n, \\ i \neq j}} R_{ij}, \text{ donde } R_{ij} = \frac{(s(C_i) + s(C_j))}{\delta(C_i, C_j)} \text{ y } s(C_i) = \frac{1}{|C_i|} \sum_{x \in C_i} \|x - c_i\| \tag{A9.6}$$

donde $C = \{C_1, \ldots, C_k\}$ es un agrupamiento de objetos, c_i es el centro del grupo C_i, $s: C \rightarrow \mathrm{R}$ mide la dispersión dentro del grupo y $\delta: C \times C \rightarrow \mathrm{R}$ mide la distancia entre grupos.

Las medidas $\mathbf{\Lambda}$ y ρ consideran la colección de objetos como un grafo pesado $G=(V,E,w)$ con el conjunto de nodos V, aristas E y la función de peso $w: E \rightarrow [0,1]$ donde V representa los objetos y w define la similitud entre dos objetos adyacentes. Considérese $C=\{C_1,\ldots,C_k\}$ un agrupamiento de un grafo pesado $G=(V,E,w)$.

Medida de conectividad parcial pesada Λ

$$\Lambda(C) = \sum_{i=1}^{k} |C_i| \cdot \lambda_i \tag{A9.7}$$

donde λ_i designa la conectividad de las aristas pesadas de $G(C_i)$. λ de un grafo G=(V, E, w) es definida como $\min \sum_{\{u,v\} \in E'} w(u,v)$, donde $E' \subset E$ y G'=(V, $E \backslash E'$) es no conexo. λ es también designada como la capacidad de un corte mínimo de G.

Medida de densidad esperada ρ

$$\rho(C) = \sum_{i=1}^{k} \frac{|V_i|}{|V|} \cdot \frac{w(G_i)}{|V_i|^{\theta}}, \text{ donde } |V|^{\theta} = w(G) \text{ y } w(G) = |V| + \sum_{e \in E} w(e) \tag{A9.8}$$

donde θ se calcula para grafos ponderados según la expresión (A9.9).

$$w(G) = |V|^{\theta} \quad \Leftrightarrow \quad \theta = \frac{\ln(w(G))}{\ln(|V|)} \tag{A9.9}$$

Modularidad (Modularity) [29]

$$Q = \sum_{i} \left(e_{ii} - a_i^2 \right) = \mathbf{Tr}\ \mathbf{e} - \left\| \mathbf{e}^2 \right\| \tag{A9.10}$$

donde **e** es una matriz simétrica de orden k cuyo elemento e_{ij} es la razón de todas las aristas en el grafo que conectan nodos del grupo i con nodos del grupo j, $\|\mathbf{e}\|$ indica la suma de los elementos de la matriz **e** y $\mathbf{Tr}\ \mathbf{e} = \sum_i e_{ii}$ es la traza de la matriz que da la razón de aristas en el grafo que conectan nodos en el mismo grupo.

Anexo 10. Algunas variantes para el cálculo del umbral de similitud entre objetos

Algunas expresiones para el cálculo inicial del umbral [306]:

a) La media de las similitudes entre todos los pares de objetos posibles; expresión (A10.1):

$$\overline{X} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} s(O_i, O_j) \tag{A10.1}$$

b) La media de los valores máximos de las similitudes entre cualquier par de objetos; expresión (A10.2):

$$\overline{X}_{\max} = \frac{1}{n} \sum_{i=1}^{n} \max_{\substack{j=1..n \\ i \neq j}} \left\{ s(O_i, O_j) \right\} \tag{A10.2}$$

c) La media de los valores mínimos de las similitudes entre cualquier par de objetos; expresión (A10.3):

$$\overline{X}_{\min} = \frac{1}{n} \sum_{i=1}^{n} \min_{\substack{j=1..n \\ i \neq j}} \left\{ s(O_i, O_j) \right\} \tag{A10.3}$$

d) La media ponderada de la media de las similitudes y la media de los máximos; expresión (A10.4):

$$\overline{X}_w = \alpha \overline{X} + (1-\alpha) \overline{X}_{\max} \tag{A10.4}$$

La descripción de la notación utilizada es la siguiente: n es la cantidad de objetos de la colección, $s(O_i, O_j)$ es el valor de la similitud entre los vectores O_i y O_j, y α es un valor entre 0 y 1 que permite ponderar la media y la media de los máximos en la expresión (A10.4).

Anexo 11. Resultados del estudio del Algoritmo 1 con el corpus BioMed

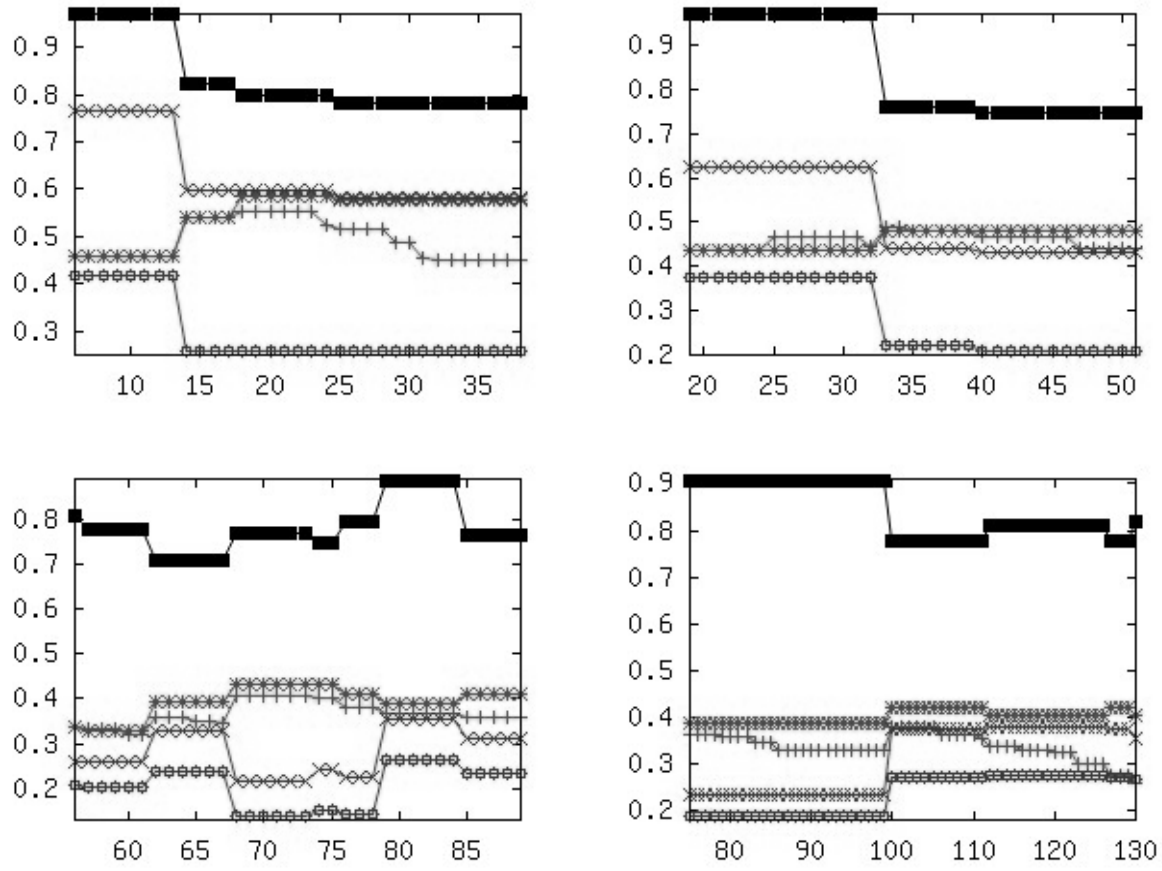


Figura A11.1 Procesamiento de BioMed con el Algoritmo 1. Las abscisas indican el número de aristas con máxima intermediación diferencial que son eliminadas. Las ordenadas indican los valores de los índices de validación internos y externos utilizados en el análisis. Valores cercanos a 1 se corresponden con un buen agrupamiento, mientras que valores cercanos a 0 indican una incorrecta división en grupos. Por orden de filas los grafos de similitud tienen 69, 87, 124 y 155 aristas, obtenidos con umbrales 0.2, 0.18, 0.14 y 0.12, respectivamente. Leyenda: '■' – OFM, '+' – 'modularidad del agrupamiento', '*' – 'modularidad de las componentes conexas', '×' – 'Precisión generalizada del agrupamiento' y 'o' – Calidad generalizada del agrupamiento.

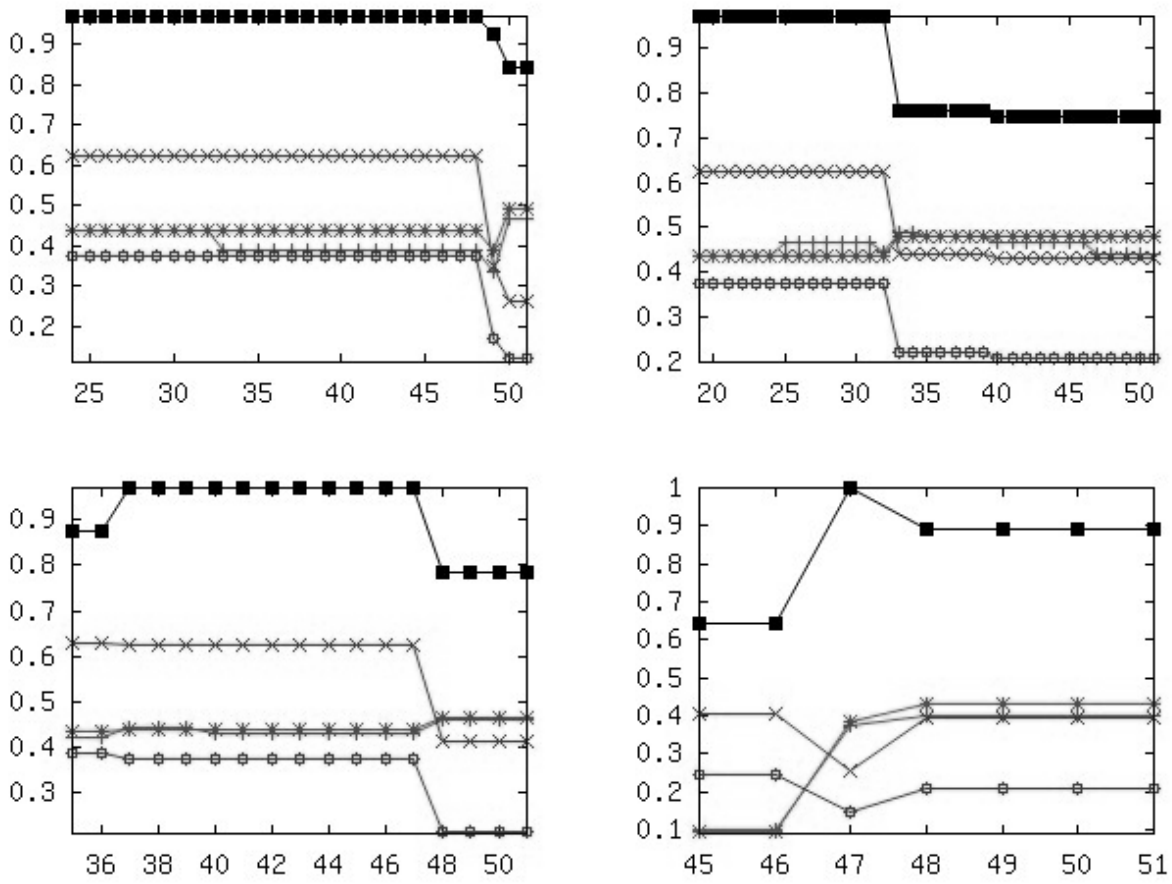


Figura A11.2 Procesamiento de BioMed con el Algoritmo 1. Las abscisas indican el número de aristas con máxima intermediación diferencial que son eliminadas. Las ordenadas indican los valores de los índices de validación internos y externos utilizados en el análisis. Valores cercanos a 1 se corresponden con un buen agrupamiento, mientras que valores cercanos a 0 indican una incorrecta división en grupos. Se usó un grafo de similitud con 87 aristas y 11 puentes obtenido con un umbral de 0.18. Por orden de filas se utilizaron vecindades de tamaño: 1, 2, 3 y 4. Leyenda: '■' – OFM, '+' – 'modularidad del agrupamiento', '*' – 'modularidad de las componentes conexas', '×' – 'Precisión generalizada del agrupamiento' y 'o' – Calidad generalizada del agrupamiento.

Anexo 12. Similitud coseno y la intermediación diferencial en la detección de puentes

Tabla A12.1 Puentes en los grafos de similitud a partir de cortes por umbral y vecinos más cercanos.

Similitud Coseno				Cálculo de la DB considerando los pesos								Cálculo de la DB sin considerar los pesos							
				Caminos de longitud 2				Caminos de longitud 3				Caminos de longitud 2				Caminos de longitud 3			
P	I	J	V	P	I	J	V	P	I	J	V	P	I	J	V	P	I	J	V
Corte con K vecinos más cercanos (K = 3). El número total de aristas es 31.																			
1	11	17	0.1243	1	11	17	153.01	1	11	17	179.29	2	6	13	20.79	2	6	13	29.53
6	4	19	0.2393	3	4	19	79.51	4	4	19	115.51	3	4	16	19.92	3	4	16	28.16
12	4	16	0.2628	4	4	16	75.81	5	4	16	107.16	4	4	19	19.03	4	4	19	27.64
25	6	13	0.3084	8	6	13	67.40	8	6	13	95.77	5	11	17	19.02	9	11	17	22.28
Corte con K vecinos más cercanos (K = 4). El número total de aristas es 39.																			
2	11	17	0.1243	2	11	17	156.57	2	11	17	187.55	2	4	16	27.91	2	4	16	31.15
13	4	19	0.2393	4	4	19	114.29	7	4	19	126.96	3	4	19	27.35	3	4	19	30.38
20	4	16	0.2628	5	4	16	106.19	8	4	16	118.52	7	11	17	19.46	7	11	17	23.31
34	6	13	0.3084	13	6	13	55.89	14	6	13	73.45	8	6	13	17.24	8	6	13	22.65
Corte con K vecinos más cercanos (K = 5). El número total de aristas es 49.																			
3	11	17	0.1243	2	11	17	168.70	2	11	17	187.07	2	2	18	25.98	3	2	19	25.99
14	4	18	0.2146	7	2	18	113.14	8	2	18	113.14	4	5	13	24.23	4	5	13	24.23
16	5	13	0.2268	8	5	13	106.84	10	5	13	106.84	5	2	16	23.98	5	2	16	23.98
17	5	19	0.2280	11	4	18	89.13	12	4	18	104.32	6	6	13	23.35	6	6	13	23.35
18	2	18	0.2297	19	5	19	77.39	17	5	19	87.55	10	11	17	20.97	7	11	17	23.25
22	4	19	0.2393	20	6	13	75.73	22	4	19	78.95	14	4	18	19.13	9	4	18	22.39
41	6	13	0.3084	24	4	19	69.28	23	6	13	75.73	17	5	19	17.65	15	5	19	19.96
45	2	16	0.3565	25	2	16	67.26	30	2	16	67.26	21	4	19	16.58	17	4	19	18.89
Corte según umbral de similitud igual a 0.21. El número total de aristas es 49.																			
2	4	18	0.2146	1	5	13	167.21	1	5	13	167.21	3	5	13	37.92	3	5	13	37.92
3	4	13	0.2190	3	4	13	140.92	3	4	13	140.92	5	4	13	30.86	8	4	13	30.86
5	4	17	0.2239	4	5	19	133.75	4	5	19	133.75	7	5	19	30.49	9	5	19	30.49
6	5	13	0.2268	10	2	18	102.21	9	4	17	109.61	10	6	13	27.55	11	6	13	27.55
7	5	19	0.2280	11	4	17	101.31	11	2	18	106.46	13	2	16	24.74	14	2	16	25.72
8	2	18	0.2297	16	4	18	90.36	17	4	18	95.15	16	2	18	23.47	15	4	17	24.54
14	4	19	0.2393	17	6	13	89.35	18	4	19	94.23	18	4	17	22.68	17	2	18	24.45
22	4	16	0.2628	19	4	19	87.43	19	6	13	89.35	19	4	19	20.92	19	4	19	22.54
39	6	13	0.3084	20	4	16	76.44	20	4	16	80.35	21	4	16	20.09	21	4	16	21.11
44	2	16	0.3565	22	2	16	69.42	24	2	16	72.16	22	4	18	19.39	23	4	18	20.41
Corte según umbral de similitud igual a 0.23. El número total de aristas es 41.																			
6	4	19	0.2393	1	4	19	119.79	1	4	19	126.58	2	2	16	34.61	3	2	16	36.56
14	4	16	0.2628	2	4	16	109.67	3	4	16	116.75	4	4	16	28.82	4	4	16	30.68
31	6	13	0.3084	4	2	16	97.10	7	2	16	102.58	5	4	19	28.66	5	4	19	30.29
36	2	16	0.3565	9	6	13	78.85	11	6	13	92.03	7	6	13	24.31	8	6	13	28.38
Corte según umbral de similitud igual a 0.25. El número total de aristas es 32.																			
5	4	16	0.2628	1	2	16	112.64	1	2	16	122.63	1	2	16	40.15	1	2	16	43.72
22	6	13	0.3084	2	4	16	108.36	3	4	16	121.51	2	4	16	28.47	6	4	16	31.93
27	2	16	0.3565	11	6	13	50.76	7	6	13	85.81	11	6	13	15.65	8	6	13	26.46
Corte según umbral de similitud igual a 0.27. El número total de aristas es 23.																			
13	6	13	0.3084	2	2	16	137.44	1	2	16	117.81	1	2	16	42	2	2	16	49
18	2	16	0.3565	5	6	13	58.36	8	6	13	29.18	6	6	13	9	5	6	13	18

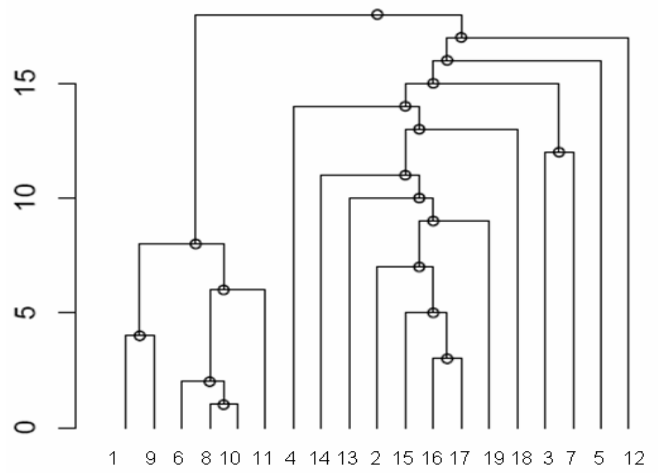


Figura A12.1 Dendrograma que se produce al eliminar aristas en orden creciente de similitud en el grafo obtenido a partir de la similitud Coseno conectando los 7 vecinos más cercanos de cada nodo.

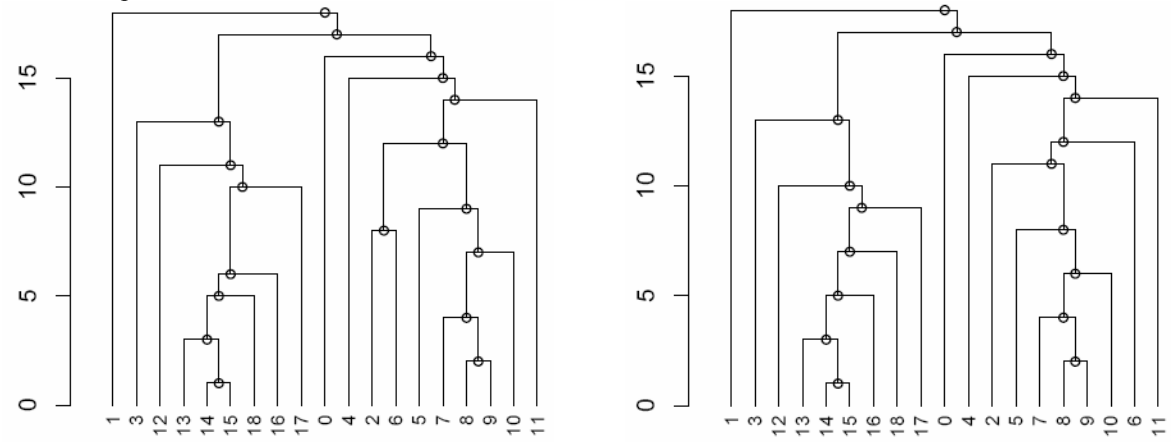


Figura A12.2 Dendrograma que se produce al eliminar aristas en orden decreciente de DB en el grafo obtenido a partir de la similitud Coseno conectando los 7 vecinos más cercanos de cada nodo. (izquierda) Cálculo de DB en grafo ponderado y cada vecindad incluyó nodos alcanzables con caminos de longitud 2 (derecha) Cálculo de DB en grafo ponderado y cada vecindad incluyó nodos alcanzables con caminos de longitud 3.

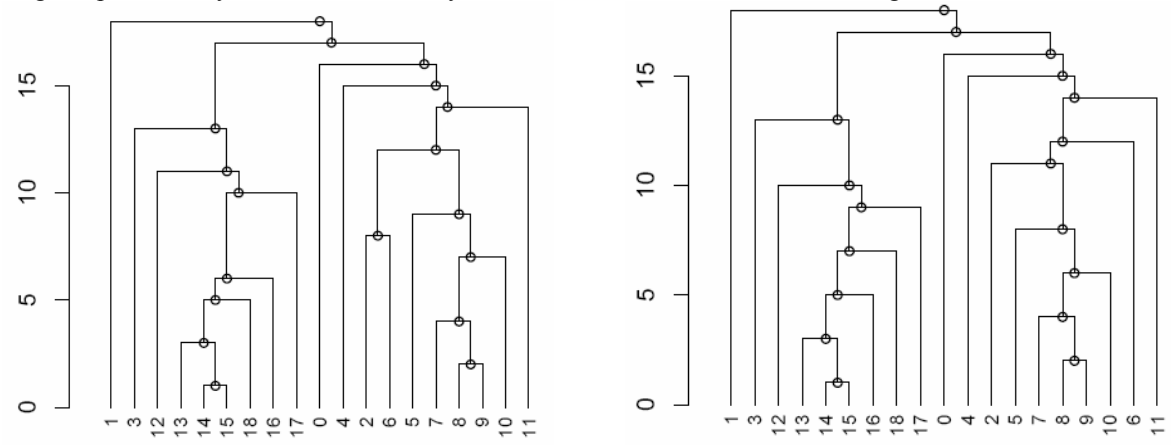


Figura A12.3 Dendrograma que se produce al eliminar aristas en orden decreciente de DB en el grafo obtenido a partir de la similitud Coseno conectando los 7 vecinos más cercanos de cada nodo. (izquierda) Cálculo de DB en grafo no ponderado y cada vecindad incluyó nodos alcanzables con caminos de longitud 2 (derecha) Cálculo de DB en grafo no ponderado y cada vecindad incluyó nodos alcanzables con caminos de longitud 3.

Anexo 13. Estudio del agrupamiento en dominios textuales

Tabla A13.1 Evaluación de los resultados del agrupamiento de las cuatro colecciones con SKWIC.

	BioMed		Reuters		NewsGroups		CEC2006	
	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
Pr	.823	.043	.701	.108	.776	.108	.696	.118
Re	.845	.066	.748	.136	.647	.077	.900	.090
OFM	.776	.097	.664	.058	.649	.115	.668	.020

Tabla A13.2 Evaluación de los resultados del agrupamiento de las cuatro colecciones con Enlace, donde los criterios son: (1) simple, (2) completo, (3) promedio, (4) pesado, y varianza mínima (5).

	BioMed			Reuters			NewsGroups			CEC2006		
	Pr	Re	OFM	Pr	Re	OFM	Pr	Re	OFM	Pr	Re	OFM
1	.775	.968	.681	1	.931	.665	.649	.976	.414	.778	.966	.680
2	.862	.774	**.778**	.837	.828	**.822**	.764	.867	.739	.685	.793	.663
3	.840	.710	.710	.554	.931	.660	.649	.982	.415	1	.966	.673
4	.840	.710	.710	.686	.862	.648	.645	.830	.576	1	.966	.673
5	.840	.710	.710	1	.793	.774	1	.682	**.760**	.850	.828	**.830**

A continuación se muestran los valores de precisión (Pr), cubrimiento (Re) y medida-F global (OFM) para las tres colecciones y los algoritmos que constituyen variantes del algoritmo Estrella, que dependen de β_0, y el algoritmo GN, que depende del umbral de corte. Para obtener los mejores resultados se variaron los umbrales considerando los histogramas de frecuencias de las similitudes. En las tablas siguientes se muestran los resultados en la vecindad de los mejores valores obtenidos, mientras que en las gráficas siguientes, se ofrece un espectro mayor de valores siguiendo una modificación constante de los umbrales, aunque incluyendo el resultado donde se obtuvo el máximo valor. En todas las gráficas mostradas las abscisas indican valores de los umbrales y las ordenadas indican los valores de las medidas de validación.

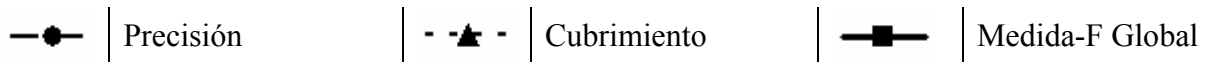


Tabla A13.3 Evaluación de los resultados del agrupamiento de BioMed con GStar, el mayor valor de OFM lo obtiene al conformar cinco grupos. ES y ACONS también alcanzan OFM igual a 0.866, pero conformando nueve y seis grupos, respectivamente.

β_0	0.13	0.14	0.145	0.15	0.155	**0.16**	**0.165**	0.17	0.175	0.18
Pr	.875	.918	.941	.941	.921	.968	.968	.968	.968	.965
Re	.871	.839	.839	.839	.710	.839	.839	.677	.677	.677
OFM	.822	.839	.852	.852	.762	**.866**	**.866**	.780	.780	.762

Tabla A13.4 Evaluación de los resultados del agrupamiento de BioMed con GN.

Corte	0.13	0.14	0.15	0.155	0.16	0.165	**0.17**	0.175	0.18	0.20
Pr	1	.918	1	1	.941	1	.945	.924	.867	.840
Re	.774	.774	.807	.807	.807	.807	.936	.903	.807	.710
OFM	.854	.827	.878	.862	.862	.862	**.937**	.905	.816	.710

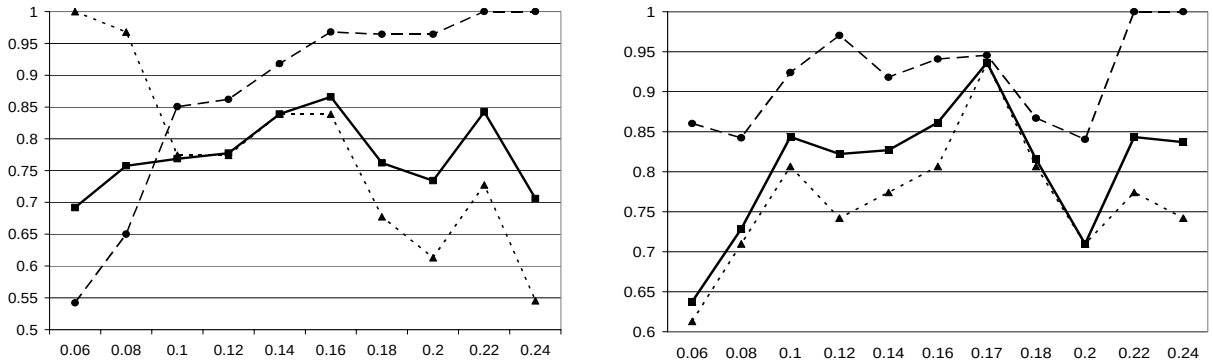


Figura A13.1 Curvas de los resultados del agrupamiento de BioMed con GStar (izquierda) y GN (derecha).

Tabla A13.5 Evaluación de los resultados del agrupamiento de Reuters con ACONS, el mayor valor de OFM lo obtiene al conformar tres grupos. ES y GStar también alcanzan OFM igual a 0.876, conformando tres grupos.

β_0	0.12	0.14	0.15	0.155	0.16	**0.165**	0.17	0.175	0.18	0.20
Pr	.637	.744	.744	.764	.781	.801	.896	.948	.966	1
Re	.966	1	1	1	.966	.966	.931	.862	.793	.552
OFM	.767	.853	.853	.866	.863	**.876**	.857	.820	.806	.605

Tabla A13.6 Evaluación de los resultados del agrupamiento de Reuters con GN.

Corte	0.16	0.165	0.17	0.175	**0.18**	**0.185**	0.19	0.195	0.20	0.22
Pr	1	1	1	1	1	1	1	1	1	1
Re	.552	.621	.738	.724	.759	.759	.621	.621	.586	.586
OFM	.657	.676	.738	.659	**.822**	**.822**	.720	.737	.737	.706

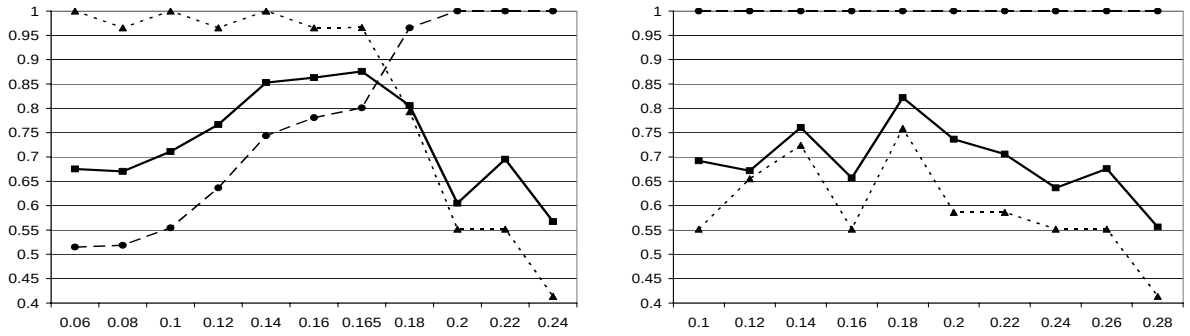


Figura A13.2 Curvas de los resultados del agrupamiento de Reuters con ACONS (izquierda) y GN (derecha).

Tabla A13.7 Evaluación de los resultados del agrupamiento de CEC2006 con ACONS, el mayor valor de OFM lo obtiene al conformar tres grupos. ES y GStar también alcanzan OFM igual a 0.744, obteniendo cuatro grupos.

β_0	0.02	0.40	0.06	0.07	**0.075**	0.08	0.085	0.09	0.10	0.12
Pr	.529	.535	.580	.607	.678	.662	.689	.683	.887	1
Re	1	.966	.931	.862	.862	.793	.793	.759	.621	.517
OFM	.684	.680	.708	.703	**.744**	.714	.734	.714	.641	.628

Tabla A13.8 Evaluación de los resultados del agrupamiento de CEC2006 con GN.

Corte	0.07	0.075	0.08	0.085	0.09	0.095	0.10	**0.105**	0.11	0.12
Pr	1	1	1	1	1	1	1	1	1	1
Re	.655	.621	.724	.655	.552	.621	.897	.724	.690	.483
OFM	.596	.600	.595	.620	.591	.634	.672	**.658**	.612	.550

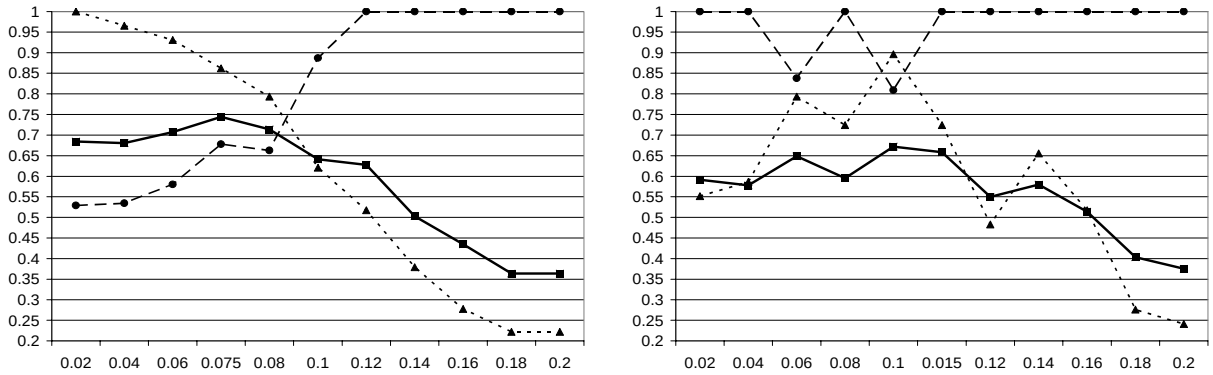


Figura A13.3 Curvas de los resultados del agrupamiento de CEC2006 con ACONS (izquierda) y GN (derecha).

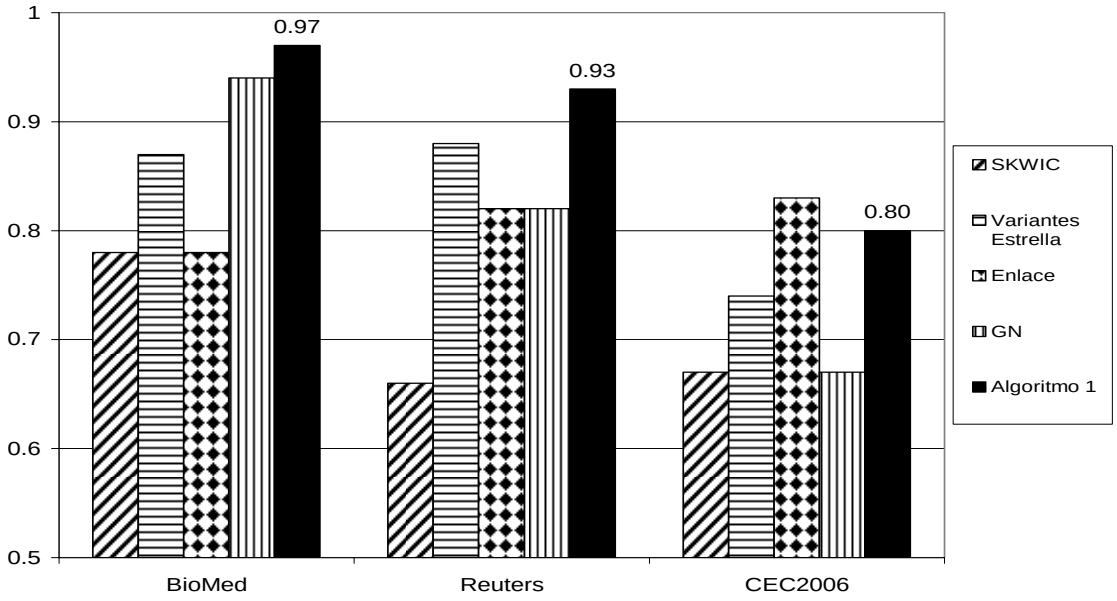
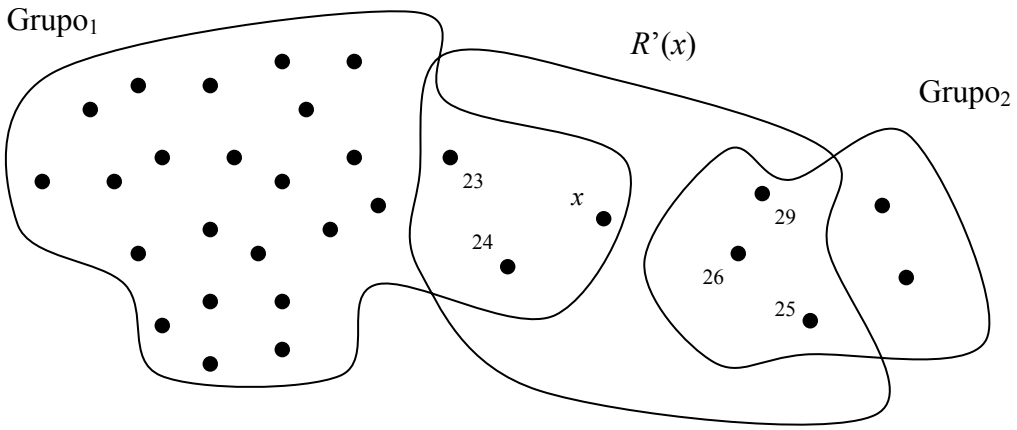


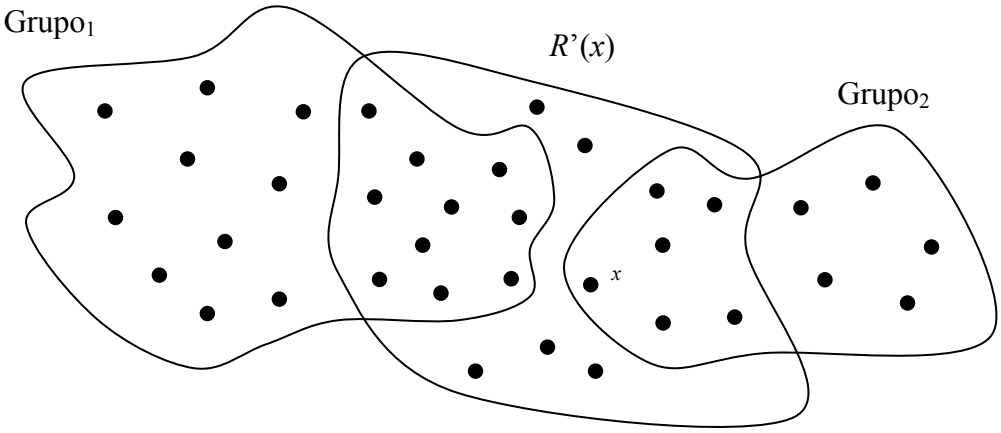
Figura A13.4 Mejores valores de OFM (ordenadas) obtenidos para cada algoritmo de agrupamiento (abscisas).

Anexo 14. Ejemplos de la pertenencia de los objetos a los grupos

Ejemplo A14.1 Supóngase que se agruparon automáticamente 30 objetos en los grupos $Grupo_1=\{1, \ldots, 24, x\}$ y $Grupo_2=\{26, \ldots, 29\}$. Sin embargo, se conoce por una clasificación de referencia que $Grupo_1=\{1, \ldots, 24\}$ y $Grupo_2=\{26, \ldots, 29, x\}$ son los grupos correctos. Se conoce además que $R'(x)=\{x, 23, 24, 25, 26, 29\}$.



Ejemplo A14.2 Supóngase que se agruparon 35 objetos en varios grupos, tal que el $Grupo_1=\{1, \ldots, 20\}$, el $Grupo_2=\{26, \ldots, 34, x\}$ y los objetos del 21 al 25 fueron agrupados en otros grupos. Sin embargo, se conoce por una clasificación de referencia que el objeto x debe pertenecer al $Grupo_1$. Se conoce además que $R'(x)=\{x, 11, \ldots, 29\}$.



Anexo 15. Descripción de los archivos utilizados para evaluar las medidas basadas en RST

Tabla A15.1 Descripción de los archivos sin clasificación de referencia.

No.	Nombre del archivo	Cantidad de instancias	Cantidad de rasgos		Valores ausentes
			Numéricos	Simbólicos	
Conjuntos de datos para técnicas de agrupamiento publicados por la Universidad de Köln http://www.uni-koeln.de/themen/statistik/data/cluster					
1	Achieve	25	4	0	No
2	Birth	70	2	0	No
3	Dentitio	66	8	0	No
4	Milk	25	4	0	No
5	Nutrient	27	5	0	No
Conjunto de datos para técnicas de aprendizaje automático y descubrimiento del conocimiento en bases de datos, publicados por la Universidad de California, Irvine. http://archive.ics.uci.edu/ml y http://kdd.ics.uci.edu					
6	AutoPrice	159	16	0	No
7	Baskball	96	5	0	No
8	Bodyfat	252	15	0	No
9	Bolts	40	8	0	No
10	Colesterol*	297	9	5	No
11	Cleveland*	297	8	6	No
12	Cloud*	108	4	0	No
13	Cpu*	209	7	0	No
14	Detroit	13	14	0	No
15	EchoMonths	130	7	3	Si
16	Elusage	55	2	1	No
17	Fishcatch	158	6	2	Si
18	Fruitfly	125	3	2	No
19	Gascons	27	5	0	No
20	Housing	506	13	1	No
21	Longley	16	7	0	No
22	Lowbwt	189	5	5	No
23	Mbagrade	61	2	1	No
24	Pbc*	312	12	7	Si
25	Pollution	60	16	0	No
26	PwLinear	200	11	0	No
27	Quake	2178	4	0	No
28	Schlyote	37	5	1	No
29	Sleep	62	8	0	Si
30	Strike	625	6	1	No
31	Veteran	137	4	4	No
32	Vineyard*	52	3	0	No
Archivos de datos del libro DATA por Andrews y Herzberg http://lib.stat.cmu.edu/datasets/Andrews/					
33	T05.1a	73	34	0	No
34	T05.1b	296	8	0	No
35	T06.1a	40	9	0	No
36	T06.1b	40	9	0	No
37	T06.2	125	12	0	No

38	T08.1	190	3	0	No
39	T09.1	109	3	2	No
40	T10.1	72	12	0	No
41	T11.1	235	12	0	No
42	T12.1	39	12	0	No
43	T14.1	732	8	0	No
44	T15.1	135	61	0	No
45	T16.1	60	12	0	No
46	T17.1	127	14	0	No
47	T21.1	33	7	0	No
48	T28.1	39	8	0	No
49	T28.2	53	8	0	No
50	T28.3	122	8	0	No
51	T30.1	19	6	0	No
52	T33.1	100	5	0	No
53	T33.2	11	13	0	No
54	T35.1	53	12	0	No
55	T35.2	52	6	0	No
56	T36.1	145	5	0	No
57	T38.1	209	9	0	Si
58	T40.2	18	5	0	Si
59	T41.1	10	14	0	No
60	T44.1	77	6	0	No
61	T47.1	96	0	24	No
62	T47.2	96	0	24	No
63	T48.1a	500	16	0	No
64	T48.1b	500	16	0	No
65	T48.1c	500	16	0	No
66	T48.1d	500	16	0	No
67	T48.1e	500	16	0	No
68	T48.1f	500	16	0	No
69	T48.1g	679	16	0	No
70	T48.3a	499	16	0	No
71	T48.3b	500	16	0	No
72	T48.3c	543	16	0	No
73	T49.1	30	7	0	Si
74	T50.1	64	16	0	No
75	T53.1	302	9	0	Si
76	T59.1	42	8	0	No
77	T60.1	95	4	0	No
78	T62.1	48	5	0	No
79	T64.1	20	12	0	No
80	T65.1	34	12	0	No
81	T65.2	34	12	0	No
82	T65.3	34	12	0	No
83	T65.4	41	8	0	No
84	T67.1	47	8	0	No
85	T70.1	32	10	0	No

Tabla A15.2 Descripción de los archivos con clasificación de referencia.

No.	Nombre del archivo	Cantidad de instancias	Cantidad de clases	Cantidad de rasgos		Valores ausentes
				Numéricos	Simbólicos	
Conjunto de datos para técnicas de aprendizaje automático y descubrimiento del conocimiento en bases de datos, publicados por la Universidad de California, Irvine. http://archive.ics.uci.edu/ml y http://kdd.ics.uci.edu						
1	Balance-scale	625	3	4	0	No
2	Credit-g	1000	2	7	13	No
3	Diabetes	768	2	8	0	No
4	Echocardiogram*	61	2	8	3	No
5	Glass	214	7	9	0	No
6	Heart-statlog	270	2	7	6	No
7	Ionosphere	351	2	34	0	No
8	Iris	150	3	4	0	No
9	Letter*	20000	26	16	0	No
10	Segment	2310	7	19	0	No
11	Sonar	208	2	60	0	No
12	Vehicle	946	4	18	0	No
13	Vowel*	990	11	10	0	No
14	WaveForm*	5000	3	40	0	No
Conjunto de datos para validar algoritmos sobre agrupamiento y series de tiempo, publicados por el Dr. Eamonn Keogh http://www.cs.ucr.edu/~eamonn/time_series_data/#Control_chart						
15	50Words*	564	15	254	0	No
16	Adiac*	411	10	176	0	No
17	Beef	60	5	254	0	No
18	Cbf	930	3	128	0	No
19	Coffee	56	2	255	0	No
20	CoverType*	731	7	54	0	No
21	Ecg200	200	2	96	0	No
22	FaceFour	111	4	254	0	No
23	Gun-Point	200	2	150	0	No
24	Lighting7	143	7	254	0	No
25	Monk2	432	2	6	0	No
26	OliveOil	60	4	255	0	No
27	Swedishleaf	1125	15	128	0	No
28	SyntheticControl*	600	6	60	0	No
29	Trace	199	4	254	0	No
30	Two-Patterns*	1244	4	128	0	No
31	Wafer*	144	2	152	0	No
32	Wbcd	699	2	9	0	No
33	Yoga*	1088	2	254	0	No
34	Zoo	101	7	16	0	No

* Archivos de datos modificados para facilitar su procesamiento

Anexo 16. Comparación de las medidas aplicadas sobre archivos de datos con y sin ruido

Algoritmo	%	Umbral	PA	CA	PGRM1	PGRM2	PGRM3	PGC	PGOS	CGRM1	CGRM2	CGRM3	CGC	CGOS
DBSCAN	5	.274	.215	.278	.339	.192	.192	.260	.217	.217	.217	.205	.227	.244
	10	.664	.502	.502	.526	.550	.550	.478	.590	.550	.391	.351	.433	.498
	15	.455	.653	.619	.821	.614	.639	.550	.821	.794	.455	.476	.351	.794
	20	.498	.679	.744	.821	.715	.848	.681	.794	.958	.986	.903	.478	.821
	25	.498	.903	.931	.903	.876	.958	.986	.958	.931	.614	.614	.741	.732
EM	5	.715	.212	.267	.260	.314	.295	.243	.212	.295	.244	.244	.306	.314
	10	.958	.858	.943	.691	.470	.764	.102	.644	.284	.102	.132	.647	.520
	15	.741	.717	.868	.881	.627	.654	.778	.654	.881	.476	.520	.872	.765
	20	.639	.601	.557	.794	.709	.737	.778	.601	.627	.035	.058	.968	.717
	25	.958	.658	.687	.639	.768	.741	.546	.664	.768	.394	.289	.314	.351
FarthestFirst	5	.375	.925	.730	1	1	.881	.826	.654	.936	.778	.737	.778	.296
	10	.434	.438	.326	.295	.469	.376	.326	.455	.296	.421	.478	.408	.332
	15	.639	.975	.875	.823	.911	.940	.959	.550	.601	.394	.394	.959	.313
	20	.931	.925	.975	.794	.601	.737	.877	.575	.823	.391	.391	.679	.601
	25	.852	1	1	.940	.970	.970	.970	.940	.970	.940	.940	1	.852
SimpleKMeans	5	.259	.638	.683	.502	.823	.681	.557	.958	.852	.709	.970	.557	.986
	10	.259	.586	.554	.852	.687	.765	.777	.590	.478	.355	.204	.711	.566
	15	.455	.968	.936	.689	.852	.876	.687	.931	.881	.911	.848	.658	.903
	20	.498	.546	.560	.768	.664	.848	.687	.931	.715	.689	.794	.737	.639
	25	.455	.654	.681	.498	.768	.664	.550	.664	.566	.394	.339	.478	.741
Xmeans	5	.244	.463	.352	.520	.184	.147	.557	.433	.398	.010	.015	.777	.520
	10	.498	.460	.394	.687	.658	.629	.723	.455	.334	.044	.062	.723	.394
	15	.322	.796	.877	.601	.654	.823	.523	.681	.687	.741	.614	.356	.715
	20	.614	.215	.179	.523	.159	.212	.605	.145	.266	.351	.351	.679	.159
	25	.768	.744	.528	.709	.681	.737	.744	.654	.502	.232	.247	.777	.520

Anexo 17. Correlaciones entre medidas basadas en RST e internas referenciadas

Tabla A17.1 Valores de significación de la prueba Kolmogorov-Smirnov para verificar distribución normal en las medidas basadas en RST aplicadas a resultados de agrupamientos a más de 100 conjuntos de datos.

Significación	PA	CA	PGRM1	PGRM2	PGRM3	PGC	PGOS	CGRM1	CGRM2	CGRM3	CGC	CGOS
DBSCAN	.123	.129	.134	.125	.132	.140	.124	.179	.027	.027	.914	.204
EM	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
FarthestFirst	.161	.005	.130	.175	.199	.020	.198	.014	.073	.066	.463	.050
SimpleKMeans	.262	.070	.256	.199	.211	.238	.198	.148	.009	.006	.838	.107
XMeans	.088	.107	.104	.060	.067	.101	.099	.097	.033	.024	.570	.422

Tabla A17.2 Valores de significación de la prueba Kolmogorov-Smirnov para verificar distribución normal en las medidas internas aplicadas a resultados de agrupamientos a más de 100 conjuntos de datos.

Significación	OS	DD	DB	IDB	CPP	DE
DBSCAN	.040	.000	.014	.419	.000	.094
EM	.020	.000	.157	.018	.000	.190
FarthestFirst	.106	.000	.000	.883	.000	.012
SimpleKMeans	.037	.000	.001	.316	.000	.067
XMeans	.039	.000	.062	.597	.000	.237

La Tabla A17.1y la Tabla A17.2 reflejan que varias significaciones son inferiores a 0.05 por lo que se rechaza la hipótesis fundamental de la prueba Kolmogorov-Smirnov y no se garantiza que los datos sigan una distribución Normal. Por tal motivo, para determinar la presencia o no de correlaciones entre las medidas se utiliza la prueba no paramétrica Tau b de Kendall.

Se considera al interpretar los resultados desde la Tabla A17.3 hasta la Tabla A17.7 que al aplicar todas las medidas basadas en RST, la medida similitud global, el índice Dunn y su variante con las expresiones de Bezdek, la conectividad parcial pesada y la densidad esperada es deseable obtener valores tan altos como sea posible, mientras que en la medida Davies-Bouldin se desean resultados opuestos.

Tabla A17.3 Correlaciones entre valores de medidas basadas en RST e internas, aplicadas a resultados de agrupamientos con el algoritmo DBSCAN.

Tau b de Kendall		PA	CA	PGRM1	PGRM2	PGRM3	PGC	PGOS	CGRM1	CGRM2	CGRM3	CGC	CGOS
OS	Correl.	.120	.120	.124	.120	.128(*)	.123	.123	.119	.108	.109	.123	.111
	Sig.	.061	.062	.052	.061	.046	.054	.055	.064	.091	.087	.054	.081
DD	Correl.	.349(**)	.352(**)	.353(**)	.350(**)	.352(**)	.350(**)	.350(**)	.353(**)	.431(**)	.433(**)	.282(**)	.287(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
DB	Correl.	.316(**)	.309(**)	.301(**)	.319(**)	.307(**)	.312(**)	.316(**)	.299(**)	.260(**)	.263(**)	.344(**)	.356(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
IDB	Correl.	-.358(**)	-.354(**)	-.346(**)	-.359(**)	-.349(**)	-.356(**)	-.359(**)	-.343(**)	-.354(**)	-.355(**)	-.385(**)	-.386(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
CPP	Correl.	.169	.119	.169	.175	.174	.163	.169	.146	-.049	-.041	.202(*)	.134(*)
	Sig.	.074	.208	.073	.064	.065	.085	.073	.124	.605	.663	.032	.015
DE	Correl.	.496(**)	.490(**)	.497(**)	.498(**)	.504(**)	.507(**)	.496(**)	.491(**)	.586(**)	.599(**)	.474(**)	.491(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

Tabla A17.4 Correlaciones entre valores de medidas basadas en RST e internas, aplicadas a resultados de agrupamientos con el algoritmo EM.

Tau b de Kendall		PA	CA	PGRM1	PGRM2	PGRM3	PGC	PGOS	CGRM1	CGRM2	CGRM3	CGC	CGOS
OS	Correl.	-.138	-.162	-.132	-.135	-.128	-.143	-.136	-.152	.028	.022	-.141	-.142
	Sig.	.146	.086	.163	.153	.176	.132	.150	.108	.768	.821	.137	.133
DD	Correl.	.359(**)	.359(**)	.339(**)	.376(**)	.359(**)	.369(**)	.367(**)	.359(**)	.373(**)	.373(**)	.372(**)	.361(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
DB	Correl.	.340(**)	.280(**)	.271(**)	.349(**)	.287(**)	.373(**)	.350(**)	.265(**)	.274(**)	.266(**)	.402(**)	.374(**)
	Sig.	.000	.000	.001	.001	.001	.000	.000	.000	.000	.000	.000	.000
IDB	Correl.	-.411(**)	-.383(**)	-.393(**)	-.405(**)	-.392(**)	-.392(**)	-.410(**)	-.378(**)	-.396(**)	-.392(**)	-.421(**)	-.440(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
CPP	Correl.	-.120	-.136	-.127	-.129	-.137	-.126	-.122	-.133	-.128	-.123	-.110	-.103
	Sig.	.207	.150	.179	.172	.148	.183	.197	.160	.176	.196	.247	.279
DE	Correl.	.536(**)	.509(**)	.521(**)	.528(**)	.513(**)	.528(**)	.538(**)	.519(**)	.528(**)	.526(**)	.519(**)	.526(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

Tabla A17.5 Correlaciones entre valores de medidas basadas en RST e internas, aplicadas a resultados de agrupamientos con el algoritmo FarthestFirst.

Tau b de Kendall		PA	CA	PGRM1	PGRM2	PGRM3	PGC	PGOS	CGRM1	CGRM2	CGRM3	CGC	CGOS
OS	Correl.	.203(**)	.197(**)	.204(**)	.193(**)	.201(**)	.198(**)	.198(**)	.199(**)	.319(**)	.314(**)	.183(**)	.147(*)
	Sig.	.002	.002	.001	.003	.002	.002	.002	.002	.000	.000	.004	.021
DD	Correl.	.353(**)	.360(**)	.350(**)	.353(**)	.352(**)	.353(**)	.353(**)	.358(**)	.304(**)	.310(**)	.269(**)	.278(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
DB	Correl.	.515(**)	.511(**)	.512(**)	.514(**)	.508(**)	.510(**)	.515(**)	.512(**)	.502(**)	.509(**)	.493(**)	.483(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
IDB	Correl.	-.508(**)	-.506(**)	-.511(**)	-.498(**)	-.497(**)	-.505(**)	-.505(**)	-.509(**)	-.533(**)	-.533(**)	-.468(**)	-.433(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
CPP	Correl.	.044	.053	.053	.058	.065	.060	.046	.063	-.062	-.045	.094	.050
	Sig.	.497	.411	.406	.366	.315	.351	.474	.327	.328	.478	.142	.432
DE	Correl.	.525(**)	.530(**)	.526(**)	.526(**)	.524(**)	.527(**)	.526(**)	.529(**)	.616(**)	.623(**)	.426(**)	.500(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

Tabla A17.6 Correlaciones entre valores de medidas basadas en RST e internas, aplicadas a resultados de agrupamientos con el algoritmo SimpleKMeans.

Tau b de Kendall		PA	CA	PGRM1	PGRM2	PGRM3	PGC	PGOS	CGRM1	CGRM2	CGRM3	CGC	CGOS
OS	Correl.	.131	.122	.139	.136	.143	.140	.133	.133	.217(*)	.214(*)	.163	.098
	Sig.	.166	.199	.141	.151	.130	.139	.160	.159	.021	.023	.085	.302
DD	Correl.	.387(**)	.390(**)	.388(**)	.383(**)	.385(**)	.383(**)	.383(**)	.387(**)	.418(**)	.419(**)	.305(**)	.321(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
DB	Correl.	.402(**)	.396(**)	.386(**)	.413(**)	.403(**)	.395(**)	.407(**)	.393(**)	.367(**)	.375(**)	.423(**)	.416(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
IDB	Correl.	-.420(**)	-.425(**)	-.417(**)	-.416(**)	-.419(**)	-.416(**)	-.420(**)	-.429(**)	-.527(**)	-.529(**)	-.408(**)	-.381(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
CPP	Correl.	.172	.129	.173	.176	.176	.166	.172	.152	-.057	-.050	.211(*)	.132
	Sig.	.068	.174	.067	.063	.063	.079	.068	.107	.548	.599	.025	.165
DE	Correl.	.515(**)	.513(**)	.521(**)	.523(**)	.527(**)	.521(**)	.518(**)	.522(**)	.609(**)	.624(**)	.499(**)	.507(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

Tabla A17.7 Correlaciones entre valores de medidas basadas en RST e internas, aplicadas a resultados de agrupamientos con el algoritmo XMeans.

Tau b de Kendall		PA	CA	PGRM1	PGRM2	PGRM3	PGC	PGOS	CGRM1	CGRM2	CGRM3	CGC	CGOS
OS	Correl.	.056	.052	.059	.043	.048	.058	.057	.058	.155(*)	.146(*)	.062	.026
	Sig.	.388	.426	.366	.515	.460	.371	.385	.371	.016	.024	.339	.691
DD	Correl.	.469(**)	.473(**)	.459(**)	.473(**)	.469(**)	.449(**)	.463(**)	.469(**)	.471(**)	.479(**)	.377(**)	.425(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
DB	Correl.	.414(**)	.406(**)	.390(**)	.439(**)	.422(**)	.397(**)	.418(**)	.397(**)	.381(**)	.387(**)	.387(**)	.446(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
IDB	Correl.	-.514(**)	-.515(**)	-.502(**)	-.511(**)	-.508(**)	-.491(**)	-.507(**)	-.515(**)	-.577(**)	-.576(**)	-.434(**)	-.459(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
CPP	Correl.	.172	.129	.173	.176	.176	.166	.172	.152	-.057	-.050	.211(*)	.132
	Sig.	.068	.174	.067	.063	.063	.079	.068	.107	.548	.599	.025	.165
DE	Correl.	.531(**)	.525(**)	.539(**)	.542(**)	.546(**)	.538(**)	.530(**)	.532(**)	.600(**)	.604(**)	.477(**)	.510(**)
	Sig.	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000

Anexo 18. Correlaciones entre medidas basadas en RST y externas referenciadas

Tabla A18.1 Valores de significación de la prueba Kolmogorov-Smirnov con corrección de Lilliefors para verificar distribución normal en las medidas externas aplicadas a resultados de agrupamientos.

Algoritmos	Significación de la prueba Kolmogorov-Smirnov con corrección de Lilliefors			
	E	P	R	OFM
DBSCAN	.044	.200	.200	.200
EM	.011	.200	.200	.200
FarthestFirst	.003	.200	.075	.200
SimpleKMeans	.017	.200	.200	.192
XMeans	.025	.200	.200	.097

Tabla A18.2 Valores de significación de la prueba Shapiro-Wilk para verificar distribución normal en las medidas externas aplicadas a resultados de agrupamientos.

Algoritmos	Significación de la prueba Shapiro-Wilk			
	E	P	R	OFM
DBSCAN	.216	.832	.690	.252
EM	.151	.819	.529	.588
FarthestFirst	.024	.281	.255	.221
SimpleKMeans	.194	.858	.472	.213
XMeans	.368	.205	.695	.063

La Tabla A18.1 y la Tabla A18.2 reflejan que algunas significaciones correspondientes a valores de la medida Entropía son inferiores a 0.05 por lo que se rechaza la hipótesis fundamental de las pruebas aplicadas y no se garantiza que la variable correspondiente a los valores de esta medida siga la distribución Normal. Adicionalemente, los resultados que se muestran en la Tabla A17.1 también reflejan que no todas las medidas basadas en RST siguen una distribución Normal. Por tal motivo, para determinar la presencia o no de correlaciones entre las medidas externas y las basadas en RST se utiliza la prueba no paramétrica Tau b de Kendall.

Se considera al interpretar los resultados desde la Tabla A18.3 hasta la Tabla A18.7 que al aplicar todas las medidas basadas en RST, las medidas OFM, precisión y cubrimiento es deseable obtener valores tan altos como sea posible, mientras que al aplicar la entropía se desean resultados opuestos.

Tabla A18.3 Correlaciones entre valores de medidas basadas en RST y externas, aplicadas a resultados de agrupamientos con el algoritmo DBSCAN.

Tau b de Kendall		PA	CA	PGRM1	PGRM2	PGRM3	PGC	PGOS	CGRM1	CGRM2	CGRM3	CGC	CGOS
E	Correl.	-.498(**)	-.472(**)	-.478(**)	-.505(**)	-.506(**)	-.495(**)	-.495(**)	-.461(**)	-.419(*)	-.425(*)	-.579(**)	-.570(**)
	Sig.	.004	.006	.006	.003	.003	.004	.004	.008	.017	.015	.001	.001
P	Correl.	.324	.330	.348	.346	.340	.310	.325	.316	.395(*)	.395(*)	.305	.270
	Sig.	.071	.065	.051	.053	.057	.085	.069	.078	.025	.025	.090	.135
R	Correl.	.348	.352(*)	.363(*)	.346	.348	.348	.350(*)	.348	.461(**)	.456(**)	.404(*)	.352(*)
	Sig.	.051	.048	.041	.053	.051	.051	.050	.051	.008	.009	.022	.048
OFM	Correl.	.358(*)	.349	.349	.377(*)	.367(*)	.356(*)	.357(*)	.330	.365(*)	.363(*)	.418(*)	.393(*)
	Sig.	.044	.050	.050	.034	.039	.045	.045	.065	.040	.041	.017	.026

Tabla A18.4 Correlaciones entre valores de medidas basadas en RST y externas, aplicadas a resultados de agrupamientos con el algoritmo EM.

Tau b de Kendall		PA	CA	PGRM1	PGRM2	PGRM3	PGC	PGOS	CGRM1	CGRM2	CGRM3	CGC	CGOS
E	Correl.	-.343(**)	-.302(*)	-.306(*)	-.286(*)	-.250(*)	-.343(**)	-.339(**)	-.302(*)	-.214	-.226	-.419(**)	-.415(**)
	Sig.	.006	.015	.014	.021	.044	.006	.006	.015	.086	.069	.001	.001
P	Correl.	.157	.149	.137	.101	.081	.165	.145	.125	.044	.056	.226	.206
	Sig.	.206	.230	.270	.417	.517	.184	.243	.315	.721	.650	.069	.098
R	Correl.	.310(*)	.278(*)	.290(*)	.262(*)	.258(*)	.302(*)	.315(*)	.278(*)	.262(*)	.266(*)	.331(**)	.327(**)
	Sig.	.013	.025	.020	.035	.038	.015	.011	.025	.035	.032	.008	.009
OFM	Correl.	.315(*)	.298(*)	.294(*)	.274(*)	.246(*)	.339(**)	.319(*)	.282(*)	.242	.246(*)	.399(**)	.379(**)
	Sig.	.011	.016	.018	.027	.048	.006	.010	.023	.052	.048	.001	.002

Tabla A18.5 Correlaciones entre valores de medidas basadas en RST y externas, aplicadas a resultados de agrupamientos con el algoritmo FarthestFirst.

Tau b de Kendall		PA	CA	PGRM1	PGRM2	PGRM3	PGC	PGOS	CGRM1	CGRM2	CGRM3	CGC	CGOS
E	Correl.	-.096	-.076	-.104	-.096	-.092	-.063	-.100	-.080	-.093	-.101	-.085	-.460(**)
	Sig.	.444	.547	.406	.444	.464	.614	.425	.526	.456	.417	.496	.000
P	Correl.	.166	.113	.171	.185	.191	.118	.168	.124	.408(*)	.409(*)	.074	.379(*)
	Sig.	.363	.540	.350	.310	.296	.521	.358	.498	.021	.020	.686	.032
R	Correl.	.418(**)	.414(**)	.434(**)	.410(**)	.414(**)	.434(**)	.430(**)	.426(**)	.127	.127	.503(**)	.321(**)
	Sig.	.001	.001	.001	.001	.001	.001	.001	.001	.307	.307	.000	.010
OFM	Correl.	.183	.115	.185	.200	.203	.142	.183	.134	.358(*)	.358(*)	.187	.542(**)
	Sig.	.316	.530	.310	.272	.266	.438	.315	.463	.044	.044	.306	.001

Tabla A18.6 Correlaciones entre valores de medidas basadas en RST y externas, aplicadas a resultados de agrupamientos con el algoritmo SimpleKMeans.

Tau b de Kendall		PA	CA	PGRM1	PGRM2	PGRM3	PGC	PGOS	CGRM1	CGRM2	CGRM3	CGC	CGOS
E	Correl.	-.341(**)	-.329(**)	-.325(**)	-.337(**)	-.341(**)	-.333(**)	-.337(**)	-.337(**)	-.333(**)	-.333(**)	-.353(**)	-.361(**)
	Sig.	.006	.008	.009	.007	.006	.007	.007	.007	.007	.007	.005	.004
P	Correl.	.252(*)	.248(*)	.244(*)	.248(*)	.252(*)	.244(*)	.248(*)	.256(*)	.301(*)	.293(*)	.240	.240
	Sig.	.043	.046	.050	.046	.043	.050	.046	.039	.016	.019	.054	.054
R	Correl.	.271(*)	.283(*)	.287(*)	.259(*)	.287(*)	.271(*)	.267(*)	.291(*)	.347(**)	.339(**)	.283(*)	.267(*)
	Sig.	.030	.023	.021	.038	.021	.030	.032	.020	.005	.006	.023	.032
OFM	Correl.	.272(*)	.276(*)	.272(*)	.268(*)	.272(*)	.272(*)	.260(*)	.285(*)	.297(*)	.289(*)	.301(*)	.293(*)
	Sig.	.029	.026	.029	.031	.029	.029	.036	.022	.017	.020	.016	.019

Tabla A18.7 Correlaciones entre valores de medidas basadas en RST y externas, aplicadas a resultados de agrupamientos con el algoritmo XMeans.

Tau b de Kendall		PA	CA	PGRM1	PGRM2	PGRM3	PGC	PGOS	CGRM1	CGRM2	CGRM3	CGC	CGOS
E	Correl.	-.351(**)	-.299(*)	-.316(*)	-.325(*)	-.333(**)	-.329(**)	-.355(**)	-.316(*)	-.329(**)	-.329(**)	-.363(**)	-.398(**)
	Sig.	.006	.018	.012	.010	.008	.009	.005	.012	.009	.009	.004	.002
P	Correl.	.390(*)	.405(*)	.398(*)	.394(*)	.407(*)	.390(*)	.391(*)	.401(*)	.517(**)	.516(**)	.376(*)	.377(*)
	Sig.	.030	.024	.027	.028	.023	.030	.030	.025	.003	.003	.037	.037
R	Correl.	.265(*)	.273(*)	.273(*)	.213	.239	.226	.243	.273(*)	.295(*)	.295(*)	.295(*)	.286(*)
	Sig.	.037	.031	.031	.092	.059	.074	.055	.031	.020	.020	.020	.024
OFM	Correl.	.316(*)	.316(*)	.299(*)	.316(*)	.325(*)	.303(*)	.312(*)	.316(*)	.338(**)	.338(**)	.363(**)	.372(**)
	Sig.	.012	.012	.018	.012	.010	.017	.014	.012	.008	.008	.004	.003

Anexo 19. Correlaciones entre medidas basadas en RST y externas en dominios textuales

Tabla A19.1 Correlaciones entre valores de medidas basadas en RST y externas, aplicadas a resultados de agrupamientos con ES, SKWIC y ES – SKWIC.

Alg.	Tau b de Kendall*		PA	CA	PGRM1	PGRM2	PGRM3	PGC	CGRM1	CGRM2	CGRM3	CGC	RFM
ES	E	Correl.	-.056	-.074	-.054	-.111	-.054	-.005	-.282(**)	-.273(**)	-.282(**)	-.370(**)	-.279(**)
		Sig.	.574	.455	.585	.264	.585	.960	.004	.006	.004	.000	.005
	P	Correl.	.005	-.008	-.004	.053	-.004	-.019	.269(**)	.244(*)	.269(**)	.450(**)	.265(**)
		Sig.	.959	.938	.972	.605	.972	.856	.009	.017	.009	.000	.010
	R	Correl.	.046	.025	.064	.018	.064	.106	.013	-.002	.013	-.001	.013
		Sig.	.639	.802	.513	.854	.513	.280	.894	.980	.894	.993	.894
	OFM	Correl.	.096	.049	.105	.082	.105	.122	.199(*)	.167	.105	.237(*)	.199(*)
		Sig.	.327	.621	.287	.407	.287	.215	.042	.088	.287	.015	.042
SKWIC	E	Correl.	-.339(**)	-.332(**)	-.350(**)	-.294(**)	-.350(**)	-.352(**)	-.464(**)	-.407(**)	-.464(**)	-.381(**)	-.465(**)
		Sig.	.001	.001	.000	.003	.000	.000	.000	.000	.000	.000	.000
	P	Correl.	.329(*)	.301(*)	.348(*)	.283(*)	.348(*)	.354(*)	.354(*)	.227	.354(*)	.300(*)	.376(**)
		Sig.	.020	.033	.013	.046	.013	.012	.012	.112	.012	.034	.007
	R	Correl.	.345(**)	.322(**)	.347(**)	.255(**)	.347(**)	.371(**)	.470(**)	.420(**)	.470(**)	.430(**)	.469(**)
		Sig.	.000	.001	.000	.009	.000	.000	.000	.000	.000	.000	.000
	OFM	Correl.	.306(**)	.286(**)	.311(**)	.225(*)	.311(**)	.332(**)	.459(**)	.402(**)	.459(**)	.386(**)	.451(**)
		Sig.	.002	.003	.002	.021	.002	.001	.000	.000	.000	.000	.000
ES - SKWIC	E	Correl.	-.071	-.046	-.096	.011	-.096	-.132	-.269(**)	-.125	-.269(**)	-.203(*)	-.273(**)
		Sig.	.466	.639	.327	.913	.327	.178	.006	.201	.006	.037	.005
	P	Correl.	.012	-.008	.052	-.058	.052	.098	.291(**)	.133	.291(**)	.178	.296(**)
		Sig.	.900	.933	.598	.552	.598	.319	.003	.173	.003	.069	.002
	R	Correl.	.206(*)	.200(*)	.230(*)	.057	.230(*)	.309(**)	.320(**)	.237(*)	.320(**)	.293(**)	.320(**)
		Sig.	.036	.042	.019	.558	.019	.002	.001	.015	.001	.003	.001
	OFM	Correl.	.099	.084	.137	-.032	.137	.206(*)	.333(**)	.198(*)	.333(**)	.257(**)	.333(**)
		Sig.	.311	.393	.162	.744	.162	.036	.001	.042	.001	.008	.001

* Se utilizó una prueba no paramétrica para determinar las correlaciones porque Kolmogorov-Smirnov con corrección de Lilliefors y Shapiro-Wilk indicaron que los datos no tienen distribución normal; por ejemplo, los resultados de RFM dio significación 0.001 para ambas pruebas, lo que evidencia que se rechaza la hipótesis fundamental.

Anexo 20. Ejemplos de validación gráfica de los agrupamientos mediante el uso de PCA

Ejemplo A20.1 Visualización de agrupamientos y medidas basadas en RST (BioMed)

Colección extraída de BioMed Central que contiene 31 documentos, los 20 primeros son artículos científicos de investigaciones sobre el SIDA y los 11 últimos abordan esta enfermedad desde el punto de vista de prevención y epidemiología.

Intencionalmente se configuraron los parámetros para el algoritmo ES-SKWIC de forma tal que se obtuviera un agrupamiento de baja calidad y otro de alta calidad, obteniéndose ocho y dos grupos, respectivamente. Al aplicar las medidas basadas en RST se obtuvo precisión 0.26, calidad 0.41 y RFM 0.3182 para el primer agrupamiento y precisión 0.87, calidad 0.93 y RFM 0.9063 para el segundo. La Figura A20.1 muestra la visualización de ambos agrupamientos utilizando dos factores obtenidos por PCA.

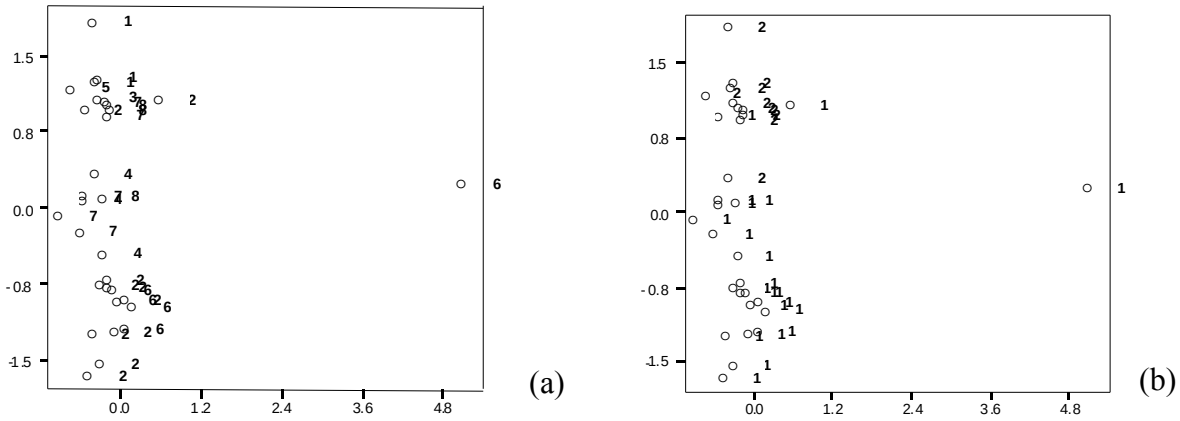


Figura A20.1 (a) Visualización de los ocho grupos utilizando dos factores obtenidos por PCA. (b) Visualización de los dos grupos utilizando dos factores obtenidos por PCA. Total de varianza explicada por PCA=17.24%. Factor 1 (F1) abscisas y Factor 2 (F2) ordenadas.

Es posible observar en la Figura A20.2 que, según factores de PCA, resulta por ejemplo, que los grupos 1, 4 y 8 son internamente consistentes, los grupos 3 y 5 son unitarios y sin embargo similares. Existe mucho solapamiento entre diferentes parejas. Todo concuerda con bajos valores de las nuevas medidas. CATPCA produce resultados aproximadamente similares a PCA aunque transpuestos. Con independencia de esto, logra visualizar mejor las malas agrupaciones cuando los parámetros determinan ocho grupos.

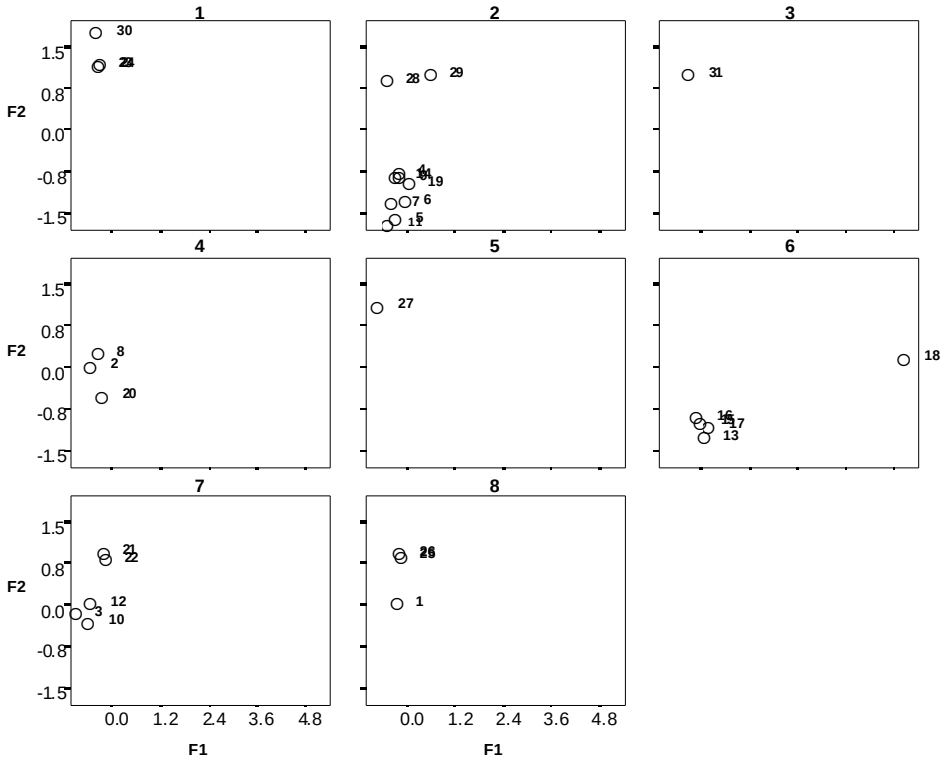


Figura A20.2 Visualización de los ocho grupos por separado (BioMed).

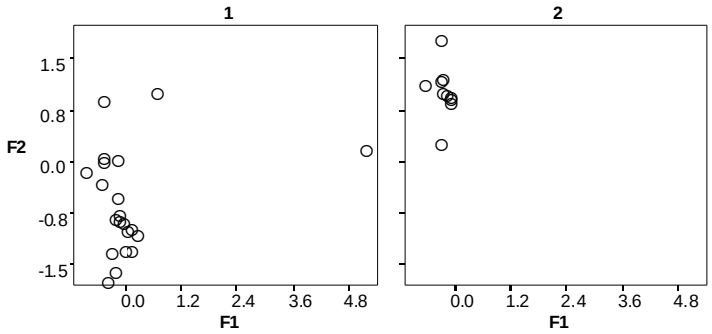


Figura A20.3 Visualización de los dos grupos por separado (BioMed).

En la Figura A20.3 se observa que PCA logra definitivamente graficar la separación de dos grupos en términos, sobre todo de la segunda componente principal, y ratifica visualmente la buena calidad del agrupamiento.

Ejemplo A20.2 Visualización de agrupamientos y medidas basadas en RST (Reuters)

Colección de 60 noticias provenientes de Reuters, donde las 30 primeras abordan la producción y comercialización de café y las 30 últimas se refieren al cacao.

Intencionalmente se configuraron los parámetros para el algoritmo ES-SKWIC de forma tal que se obtuviera un agrupamiento de baja calidad y otro de alta calidad, obteniéndose ocho y cuatro grupos, respectivamente. Al aplicar las medidas basadas en RST se obtuvo precisión 0.6216, calidad 0.7667 y RFM 0.6866 para el primer agrupamiento y precisión 0.9355, calidad 0.9667 y RFM 0.9508 para el segundo. La Figura A20.4 muestra la visualización de ambos agrupamientos utilizando dos factores obtenidos por PCA.

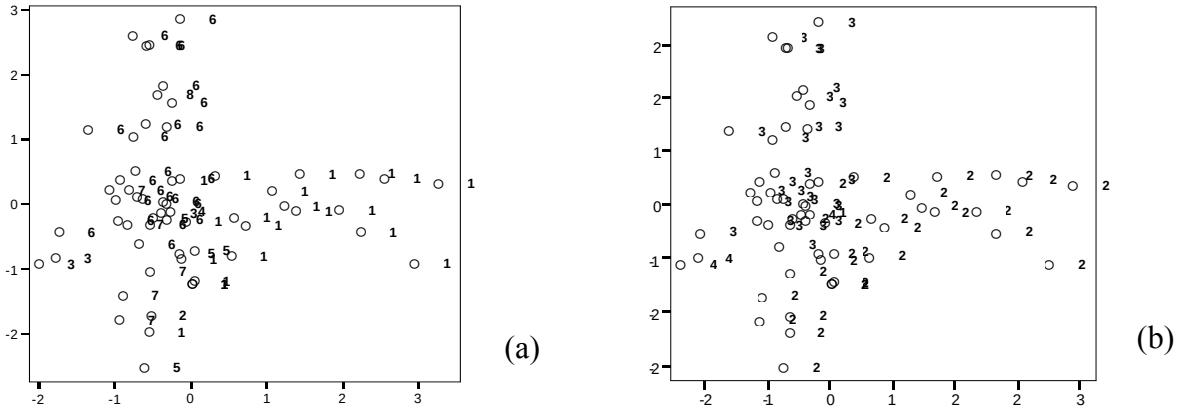


Figura A20.4 (a) Visualización de los ocho grupos utilizando dos factores obtenidos por PCA. (b) Visualización de los cuatro grupos utilizando dos factores obtenidos por PCA. Total de varianza explicada por PCA=9.2%. Factor 1 (F1) abscisas y Factor 2 (F2) ordenadas.

Es posible observar en la Figura A20.5 que, según factores de PCA, resulta por ejemplo, que los grupos 1 y 7 son internamente consistentes, los grupos 2 y 3 son unitarios, y sin embargo similares. La visualización concuerda con bajos valores de las nuevas medidas.

La visualización de los resultados de agrupamientos que se muestra en la Figura A20.6, distingue los grupos 2 y 3 internamente consistentes. Sólo dos y tres documentos pertenecen a los grupos 1 y 4, respectivamente. La visualización ratifica la buena calidad del agrupamiento.

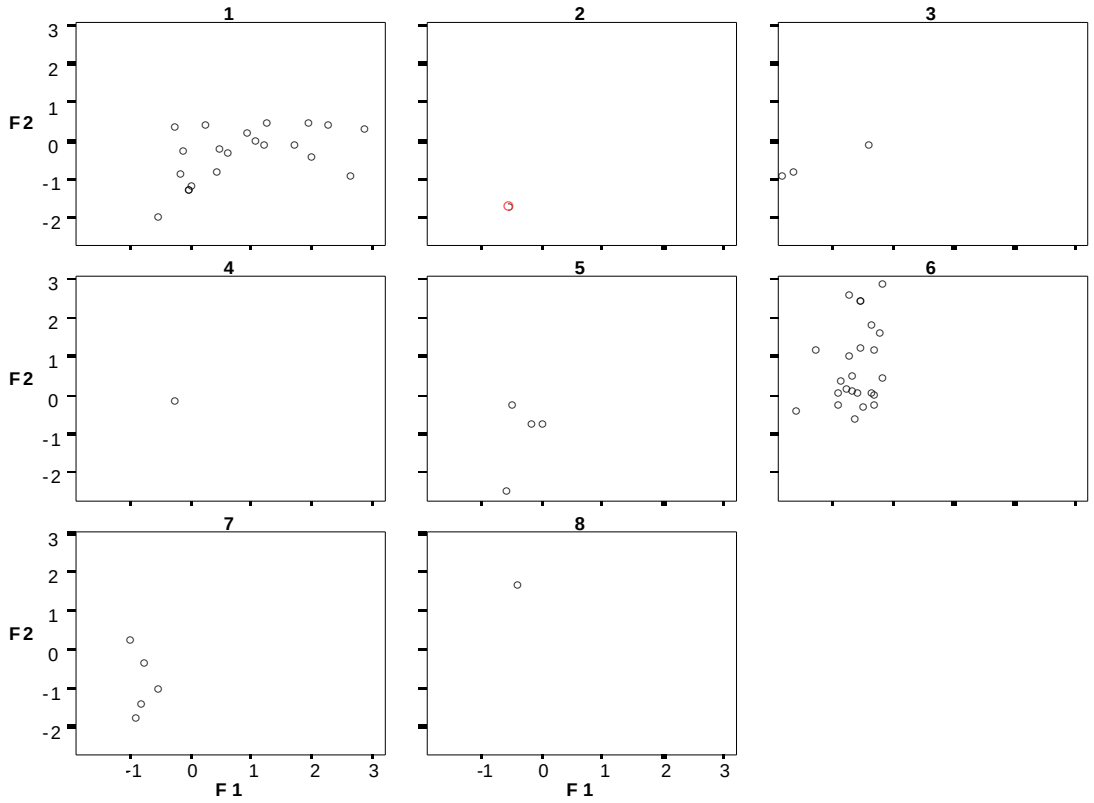


Figura A20.5 Visualización de los ocho grupos por separado (Reuters).

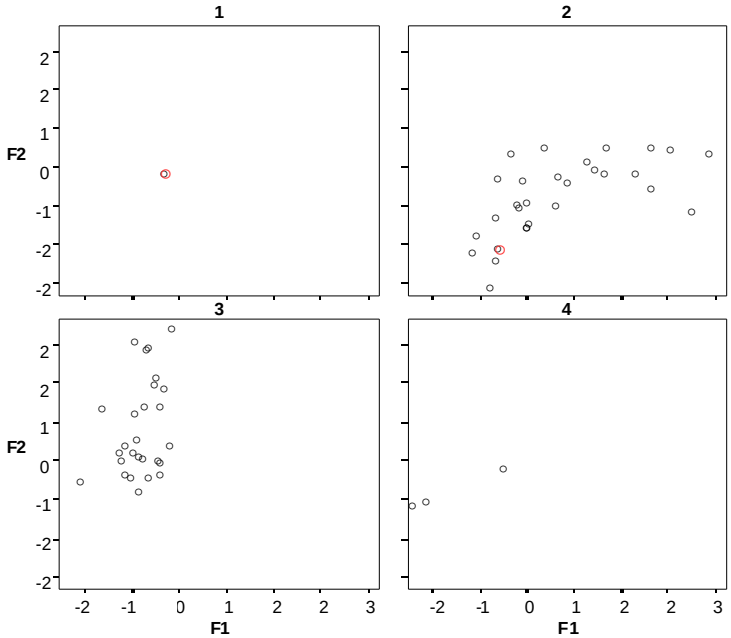


Figura A20.6 Visualización de los cuatro grupos por separado (Reuters).

Anexo 21. RST para extraer documentos más representativos y refinar agrupamientos

RST para extraer documentos más representativos. Se agruparon los artículos de un corpus conformado a partir de la colección BioMed. La Tabla A21.1 y la Tabla A21.2 muestran los conjuntos de documentos para etiquetar los grupos. Se obtienen subconjuntos más específicos con valores de umbral más bajos y menos específicos para mayores valores del umbral de similitud. Cuando decrece el umbral, las aproximaciones inferiores se hacen más pequeñas.

Tabla A21.1 Fragmento de la extracción de los documentos más representativos en los grupos 7 y 8 mediante el cálculo de las aproximaciones inferiores, variando del umbral de similitud al construir las relaciones.

Umbral	$Grupo_7$ (Diabetes Mellitus) = {32, 35, 34, 37, 39, 40, 42, 43, 44, 47, 49, 51, 54, 56, 57, 58, 59, 60, 61, 64}	$Grupo_8$ (Fibrosis cística) = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 28}
0.22	{35, 57, 58, 61, 64}	{1, 3, 4, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 28}
0.20	{35, 57, 58, 61, 64}	{1, 3, 4, 6, 8, 9, 11, 14, 15, 28}
0.18	{35, 57, 58, 64}	{1, 3, 4, 6, 8, 9, 11, 15}
0.16	{57, 58, 64}	{3, 4, 6, 9, 15}
0.14	{57, 58}	{4, 6, 9, 15}
0.12	{57, 58}	{4, 9}

Tabla A21.2 Fragmento de la extracción de los documentos más representativos en los grupos 9 y 12 mediante el cálculo de las aproximaciones inferiores, variando del umbral de similitud al construir las relaciones.

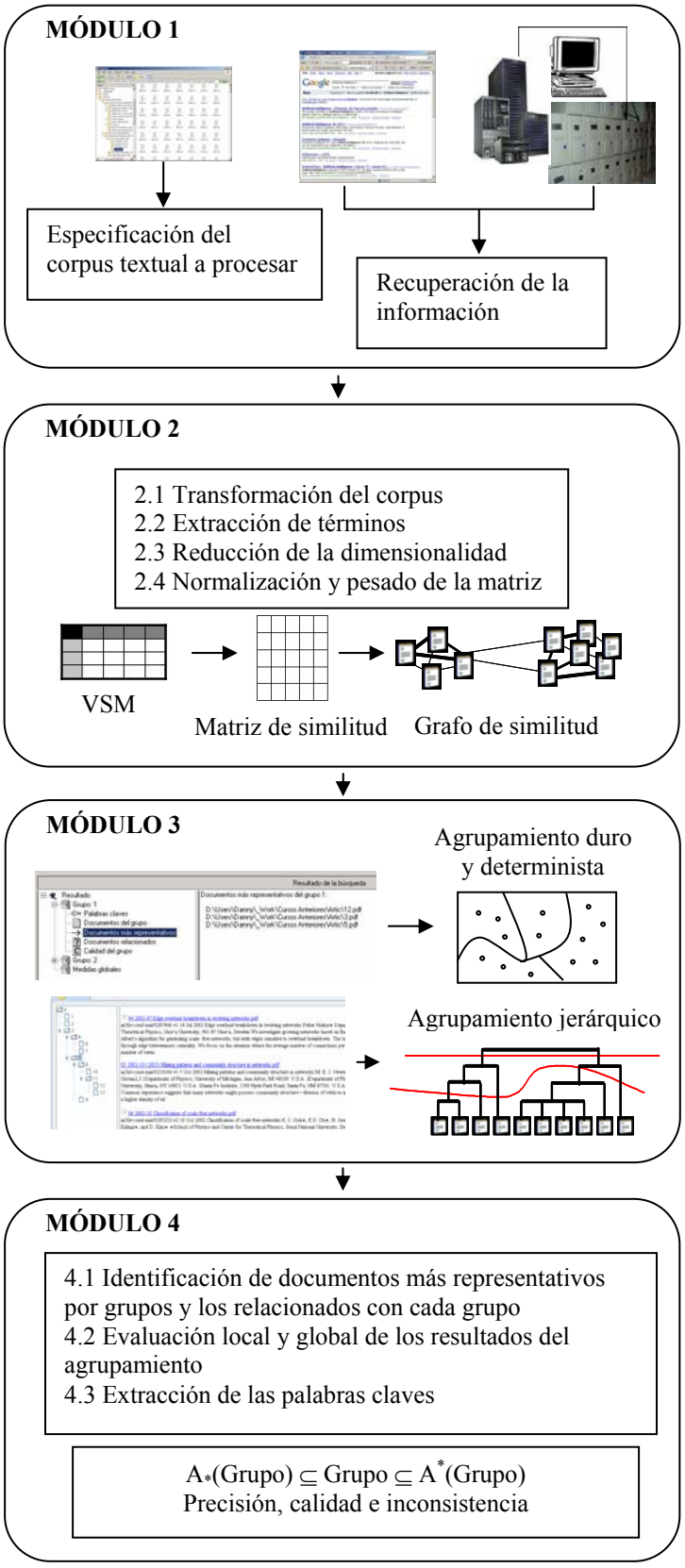
Umbral	$Grupo_9$ (Cáncer de pulmón) = {10, 98, 29, 100, 102, 103, 105, 106, 109, 110, 112, 113, 114, 119, 120}	$Grupo_{12}$ (SIDA) = {32, 71, 72, 73, 75, 79, 81, 83, 82, 84, 85}
0.22	{98, 29, 100, 102, 105, 106, 109, 110, 112, 113}	{79, 83, 82, 84, 85}
0.20	{98, 29, 100, 102, 105, 109, 110, 112, 113}	{79, 83, 84, 85}
0.18	{98, 100, 102, 105, 109, 110, 112, 113}	{79, 83, 84, 85}
0.16	{100, 102, 105, 110}	{79, 84, 85}
0.14	{100, 102, 105, 110}	{85}
0.12	{105}	{85}

Potencialidades de RST para refinar agrupamientos. Supóngase que como resultado del agrupamiento de 40 objetos se obtuvieron los grupos 1, 2, 3 y 4. Sea $Grupo_1$={1, 2, 3, 4} y $Grupo_2$={5, 6}, donde $A_*(Grupo_1)$={1, 2}, $A^*(Grupo_1)$={1, 2, 3, 4, 5, 6}, $A_*(Grupo_2)=\varnothing$ y $A^*(Grupo_2)$={3, 4, 5, 6} son las respectivas aproximaciones inferiores y superiores de estos dos grupos. Todos los elementos del $Grupo_2$ están relacionados con elementos del $Grupo_1$ (la aproximación inferior del $Grupo_2$ es vacía). Existe alta coincidencia entre las aproximaciones superiores de ambos grupos, $A^*(Grupo_1)\cap A^*(Grupo_2)$={3, 4, 5, 6}, representa que existe 2/3 de coincidencia con la aproximación superior del $Grupo_1$ y coincidencia total con el $Grupo_2$. Los resultados aplicando RST indican que el $Grupo_2$ debe ser combinado con el $Grupo_1$.

Anexo 22. Gestión del conocimiento en organizaciones vs el mercado del conocimiento en Internet

Sistemas de gestión del conocimiento dentro de las organizaciones	**Mercado del conocimiento en Internet**
En un sistema interno y cerrado	En un entorno abierto
Equipado con un mecanismo de incentivos	Donde la motivación es ganar dinero
Diseñado para pomover el establecimiento de la reputación y usualmente estimular reciprocidad	Donde la reputación es usada para el control de la calidad y para adquirir poder de negociación con configuraciones de precio
Los participantes conocen cada una de las otras personas	Los participantes usualmente tienen solamente seudónimos
O al menos se conocen como roles	Y en algunos casos, conocer a cada uno de los otros, es irrelevante para la participación en el negocio
Hay relaciones institucionales entre los participantes	No hay relaciones institucionales entre los participantes
El dinero no es la motivación primaria para participar	Los expertos contribuyen al propósito directo de ganar dinero

Anexo 23. Esquema general de la aplicación



Anexo 24. Enfoques lingüísticos para analizar significados respecto al contexto

Estos enfoques se dividen en cinco niveles [2]:

1. Nivel de grafema: Análisis sobre un nivel de sub-palabra, comúnmente concerniente a las letras.
2. Nivel léxico: Análisis concerniente a palabras individuales.

Los dos primeros niveles operan solamente con un plano estadístico sobre el texto, es decir básicamente sobre frecuencias de combinaciones de términos, que pueden ser letras o palabras.

3. Nivel sintáctico: Análisis concerniente a la estructura de oraciones.
4. Nivel semántico: Análisis relativo al significado de palabras y frases.
5. Nivel pragmático: Análisis relativo al significado, tanto dependiente del contexto como independiente del contexto (por ejemplo, aplicaciones específicas, contextos).

Los niveles más altos del análisis de textos existen para tratar de capturar mayor contenido semántico a través de la explotación de la cantidad creciente de información contextual tal como la estructura de las oraciones, párrafos o documentos; sin embargo, estos niveles implican un procesamiento computacional muy costoso.

En [307] se comenta que puede ser razonable asumir la representación de textos más compleja teniendo en cuenta que niveles más altos de análisis textual deben tender a obtener herramientas más efectivas. Pero, mientras más compleja sea la definición de términos indexados, más compleja será la representación de los textos, y la dimensionalidad de los rasgos crecerá correspondientemente. Por tanto, escoger un nivel adecuado sobre el cual basar la definición de los términos es siempre un equilibrio entre la expresividad semántica y la complejidad de la representación. En la mayoría de las aplicaciones de la minería de textos, la definición de términos simples es dominante. Típicamente, se enfoca el análisis de textos a partir de los dos primeros niveles, y a su vez, alguna información sintáctica puede ser también incorporada.

Anexo 25. Algunas medidas de calidad de términos

Umbral de frecuencia de términos y Ley de Zpif. Eliminar los términos que tienen o muy alta o muy baja frecuencia de aparición [308], a partir de un cálculo adecuado del umbral y del estudio de la Ley de Zipf [266]. Términos que raramente aparecen en una colección de documentos tendrán poco poder discriminante y pueden ser eliminados [309]. En contraste, términos con frecuencia de aparición alta se asumen que son comunes y que tampoco tienen poder discriminante[72].

Umbral de frecuencia de documentos. Teniendo en cuenta que $n(t)$ es el número de documentos en los cuales el término t aparece al menos una vez, una heurística simple de selección es excluir todos los términos desde el vocabulario cuya frecuencia de documentos es menor que algún umbral, ya que términos que ocurren en sólo muy pocos documentos improbablemente llevan información que permita distinguir los grupos textuales y tienden a ser ruidosos [264]. Además, usar la ocurrencia de términos infrecuentes no es confiable estadísticamente. Al eliminar estos términos se mantiene el poder discriminante y se mejora la efectividad del agrupamiento y clasificación textual.

Frecuencia inversa de documento y TFIDF. La importancia de los términos se asume inversamente proporcional al número de documentos en los cuales el término particular aparece. Después de eliminar las palabras de parada, la importancia de un término se incrementa con su frecuencia de uso. Combinando estas ideas se formuló la medida frecuencia del término / frecuencia inversa de documentos (tfidf).

$$\text{tfidf}(t)=tf(t)\cdot idf(t),\ \text{donde}\ idf(t)=\log\frac{n}{n(t)} \qquad \text{(A25.1)}$$

Una combinación similar de frecuencia de términos y frecuencia inversa de documentos es se utiliza usualmente para asignar pesos a los términos [261].

[72] Términos con alta frecuencia de aparición pueden formar parte de la lista de palabras de parada automáticamente construida desde la colección de documentos. En esta tesis se considera que la lista es suministrada.

Razón de señal a ruido. Medir el poder discriminante que transmite cada término, basado en el ruido R(*t*), como la entropía de la distribución de la probabilidad del término *t* entre los documentos [310]:

$$\text{SNR}(t)=\log tf(t) - \text{R}(t),\ \text{R}(t) = -\sum_{j=1}^{n} \text{P}(d_j,t)\log \text{P}(d_j,t)\ \text{ y }\ \text{P}(d_j,t) = \frac{tf_{d_j}(t)}{tf(t)} \qquad \text{(A25.2)}$$

Entropía. Calcular la entropía como una medida de importancia, según Lochbaum y Streeter en 1989 [311]:

$$\text{Entropía}(t) = 1 + \frac{1}{\ln(n)} \sum_{i=1}^{n} p_i(t) \cdot \ln(p_i(t)) \quad \text{donde } p_i(t) = \frac{tf_{d_i}(t)}{\sum_{j=1}^{n} tf_{d_j}(t)} \qquad \text{(A25.3)}$$

Calidad de términos. Medir la calidad de los términos según las expresiones q_0 y q_1, la segunda constituye una variante de la primera donde n_1 es el número de documentos en los cuales *t* ocurre al menos una vez [64].

$$q_0(t) = \sum_{j=1}^{n} \left(tf_{d\,j}(t)\right)^2 - \frac{1}{n}\left[\sum_{j=1}^{n} tf_{d\,j}(t)\right]^2 \text{ y } q_1(t) = \sum_{j=1}^{n_1} \left(tf_{d\,j}(t)\right)^2 - \frac{1}{n_1}\left[\sum_{j=1}^{n_1} tf_{d\,j}(t)\right]^2 \qquad \text{(A25.4)}$$

Skewness y Kurtosis. Calcular la parcialidad de los términos mediante la combinación de Skewness y Kurtosis según P(*t*)=w_1·Skewness(*t*)+w_2·Kurtosis(*t*), donde w_1 y w_2 son pesos positivos y *s* es la desviación estándar de la ocurrencia del término *t* en la colección de documentos [312].

$$\text{Skewness}(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{\left(tf_{d_i}(t) - \frac{tf(t)}{n}\right)^3}{s^3} \text{ y } \text{Kurtosis}(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{\left(tf_{d_i}(t) - \frac{tf(t)}{n}\right)^4}{s^4} - 3 \qquad \text{(A25.5)}$$

Anexo 26. Normalización y pesado de la matriz

En [261] se define el peso de un término t en un documento d según (A26.1), permitiendo generar un vector pesado $\mathbf{d}=(w(d,t_1),\ldots, w(d,t_m))^T$ para cualquier documento d basado en el vector de frecuencia de términos $\mathbf{d}_{tf}=(tf_d(t_1),\ldots, tf_d(t_m))^T$.

$$w(d,t)=\frac{w_{local}(d,t)w_{global}(t)}{w_{normalización}(d)} \qquad \text{(A26.1)}$$

Componente local. El factor de peso local $w_{local}(d,t)$ refleja la importancia del término t en un documento d. Se obtiene aplicando una función de transformación a la frecuencia del término t en el documento d; por ejemplo, identidad, indicador binario y normalización según la mayor frecuencia de términos en el documento d.

Componente global. El factor de peso global $w_{global}(t)$ tiene en cuenta la importancia del término t en la colección de documentos. Se utiliza la frecuencia inversa de documentos.

Componente de normalización. El factor de normalización $w_{normalización}(d)$ típicamente se calcula siguiendo la normalización coseno y por la suma de los componentes del vector pesado de un documento, permitiendo la abstracción de las longitudes de los documentos.

La mayoría de las formas de pesado se basa en alguna variación de la fórmula TF-IDF. La idea de una expresión TF-IDF es que el peso de los términos deba reflejar la importancia relativa de un término en un documento con respecto a los otros términos en el documento. Algunas variantes para el cálculo de TF-IDF se muestran a continuación [64, 271].

$$w(d,t)=tf_d(t)\left(1+\log_2\left(\frac{n}{n(t)}\right)\right) \qquad \text{(A26.2)}$$

Una modificación de la expresión (A26.2) se muestra en (A26.3), teniéndose en cuenta la cantidad de veces que ocurren en un documento aquellos términos que más aparecen, donde $\max_k tf_d(k)$ representa el número de ocurrencias que tiene la palabra que más aparece en d.

$$w(d,t)=\frac{tf_d(t)}{\max_k tf_d(k)}\left(1+\log_2\left(\frac{n}{n(t)}\right)\right) \qquad \text{(A26.3)}$$

Otra forma para el cálculo de TF-IDF es la expresión (A26.4).

$$w(d,t) = \begin{cases} (1 + tf_d(t_i)) \log_2(\frac{n}{n(t_i)}), & si\ tf_d(t_i) \geq 1 \\ 0 \qquad\qquad\qquad\qquad , & si\ tf_d(t_i) = 0 \end{cases} \tag{A26.4}$$

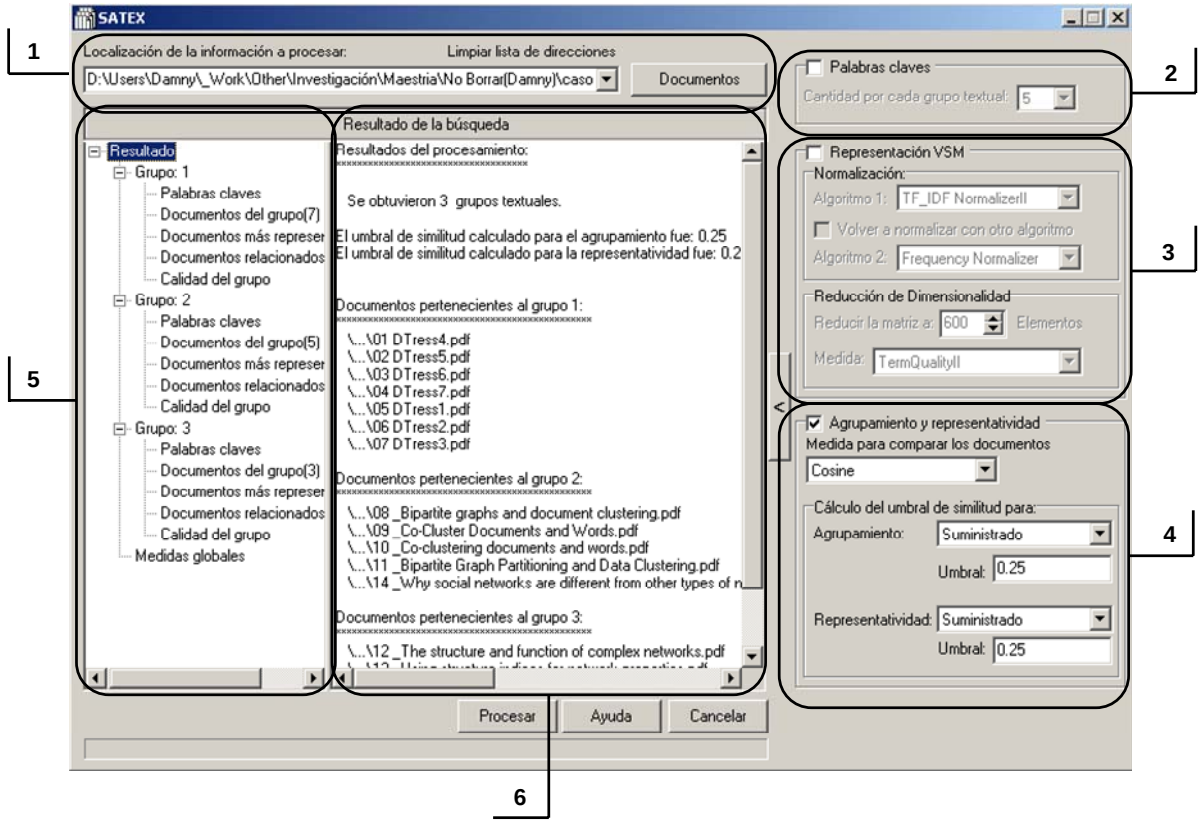
En (A26.5) se muestra una expresión para el cálculo de TF-IDF que tiene como objetivo obtener pesos en el intervalo [0,1] y considerar la componente de normalización.

$$w(d,t) = \frac{\text{tfidf}(d,t)}{\sqrt{\sum_{j=1}^{m} \left(\text{tfidf}(d,t_j)\right)^2}} \tag{A26.5}$$

La expresión (A26.6) es una variante desglosada de (A26.5), donde el numerador de este coeficiente considera la frecuencia de ocurrencia del término *t* en *d* y la discriminación del término IDF, mientras que el denominador permite la estandarización para eliminar la influencia de la longitud del documento.

$$w(d,t) = \frac{tf_d(t)\log_2\left(\frac{n}{n(t)}\right)}{\sqrt{\sum_{j=1}^{m}\left(tf_d(t_j)\log_2\left(\frac{n}{n(t_j)}\right)\right)^2}} \tag{A26.6}$$

Anexo 27. Descripción general de la interfaz de usuario de SATEX



1. Seleccionar la colección de documentos que se desea procesar.
2. Especificar el número de palabras claves que se desean obtener por cada grupo textual.
3. Parámetros asociados a la normalización y pesado de la representación VSM y la reducción de la dimensionalidad mediante la selección de mejores términos.
4. Parámetros asociados al cálculo de la similitud entre documentos y umbrales de similitud para el agrupamiento y post-agrupamiento.
5. Árbol con los resultados del procesamiento. En la raíz se hace referencia a los resultados generales y cada nodo corresponde a un grupo textual obtenido, permitiendo acceder a los documentos del grupo, conjunto de documentos más representativos y relacionados, palabras claves y evaluación del grupo.
6. Descripción detallada de los resultados del procesamiento en correspondencia con el nodo que se haya seleccionado en el árbol de resultados.

Anexo 28. Descripción general de la interfaz de usuarios de GARLucene

Figura A28.1 Menú principal del sistema GARLucene. (1) Realización de búsquedas sobre los documentos indexados, (2) Especificación de l alocalización, creación y actualización de los índices, (3) Selección del algoritmo de agrupamiento y las medidas de validación a utilizar y (4) Ayuda del sistema.

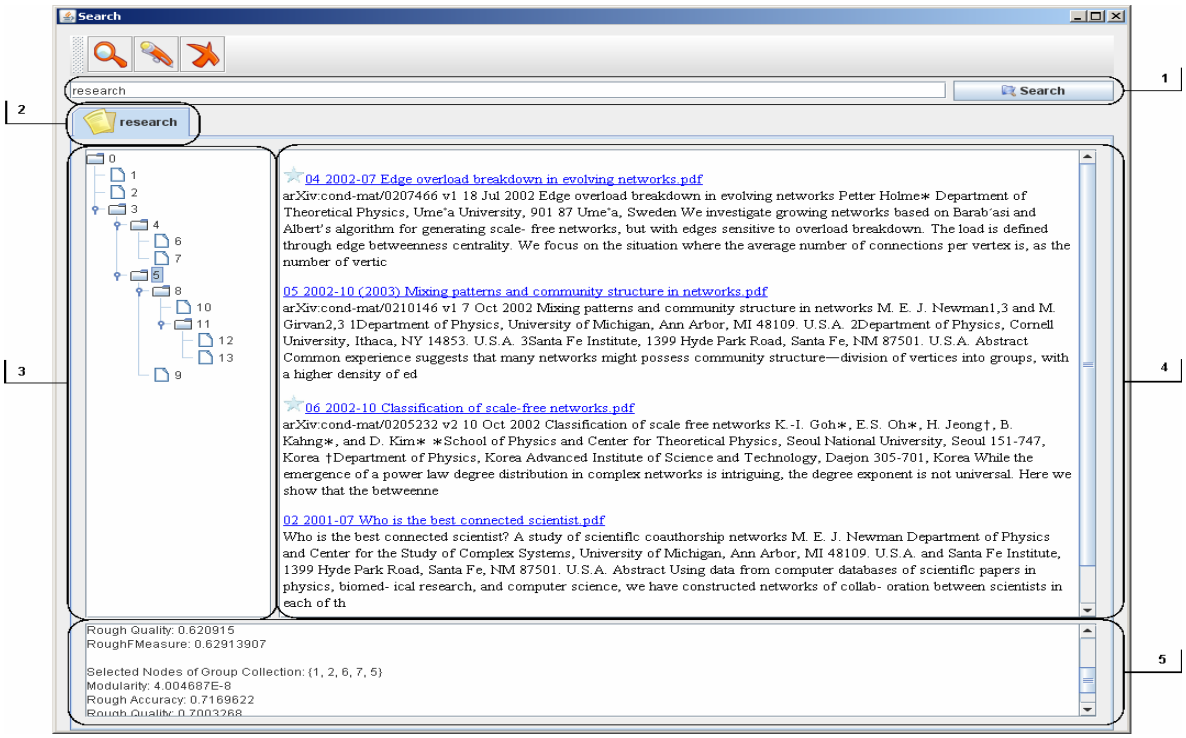


Figura A28.2 Ventana de GARLucene con los resultados agrupados de un proceso de recuperación de información.

1. Especificación de la palabra, frase o expresión regular para realizar la búsqueda.
2. Resultado de una búsqueda.
3. Jerarquía de documentos.
4. Enlace a cada documento perteneciente al grupo seleccionado y su resumen. Se marcan con estrellas los documentos más representativos y se muestra al final la evaluación del grupo.
5. Resultados globales de las búsquedas realizadas.

Anexo 29. Campos por tipos de ficheros en configuración XML de LIUS

```
<msWord setBoost="1.2">
        <indexer class="lius.index.msword.WordIndexer">
                <mime>application/msword</mime>
        </indexer>
        <fields>
                <luceneField name="fullText" get="content" type="Text"/>
        </fields>
</msWord>
<msPowerPoint setBoost="0.2">
        <indexer class="lius.index.powerpoint.PPTIndexer">
                <mime>application/vnd.ms-powerpoint</mime>
        </indexer>
        <fields>
                <luceneField name="fullText" get="content" type="Text"/>
        </fields>
</msPowerPoint>
<html setBoost="1.4">
        <indexer class="lius.index.html.JTidyHtmlIndexer">
                <mime>text/html</mime>
                <mime>application/x-asp</mime>
        </indexer>
        <fields>
                <luceneField name="fullText" get="content" type="Text"/>
        </fields>
        <!--
        <indexer class="lius.index.html.NekoHtmlIndexer">
                <mime>text/html</mime>
        </indexer>
        <fields>
                <luceneField name="fullText" xpathSelect="//*" type="Text" ocurSep="|" />
        </fields>
</html>
<rtf setBoost="1.5">
        <indexer class="lius.index.rtf.RTFIndexer">
                <mime>application/rtf</mime>
        </indexer>
        <fields>
                <luceneField name="fullText" get="content" type="Text"/>
        </fields>
</rtf>
<pdf setBoost="1.6">
        <indexer class="lius.index.pdf.PdfIndexer">
                <mime>application/pdf</mime>
        </indexer>
        <fields>
                <luceneField name="fullText" get="content" type="Text"/>
                <!--<luceneField name="title" get="title" type="Text" />
                <luceneField name="author" get="author" type="Text" />
                <luceneField name="creator" get="creator" type="Text" />
                <luceneField name="summary" get="summary" type="Text" />
                <luceneField name="keywords" get="keywords" type="Text" />
                <luceneField name="producer" get="producer" type="Text" />
                <luceneField name="subject" get="subject" type="Text" />
                <luceneField name="trapped" get="trapped" type="Text" />
                <luceneField name="creationDate" get="creationDate" type="DateToString" />
                <luceneField name="modDate" get="modDate" type="DateToString" />-->
        </fields>
</pdf>
<txt setBoost="0.1">
        <indexer class="lius.index.txt.TXTIndexer">
                <mime>text/plain</mime>
        </indexer>
        <fields>
                <luceneField name="fullText" get="content" type="Text"/>
        </fields>
</txt>
```

Anexo 30. Descripción de las variables a medir para evaluar los sistemas

<table>
<tr><td colspan="3">Variables dirigidas a valorar los resultados de los sistemas</td></tr>
<tr><td>1</td><td>Eficacia</td><td>La exactitud, precisión, veracidad, e integridad con las cuales los usuarios pueden alcanzar los objetivos en las colecciones textuales propias de su entorno de aplicación.</td></tr>
<tr><td>2</td><td>Eficiencia</td><td>Los recursos gastados con relación a la precisión e integridad de los objetivos logrados.</td></tr>
<tr><td>3</td><td>Conformidad</td><td>El confort, bienestar y aceptabilidad del uso del sistema y sus resultados para los usuarios y otros que se nutren de su uso.</td></tr>
<tr><td>4</td><td>Confiabilidad</td><td>La exactitud de los resultados y la capacidad de predicción del sistema.</td></tr>
<tr><td>5</td><td>Robustez</td><td>El nivel de soporte provisto al usuario en la determinación de logros exitosos y la evaluación de los objetivos.</td></tr>
<tr><td colspan="3">Las medidas que tributan a éstas son:
Nivel de satisfacción con la división en grupos.
Interpretabilidad del agrupamiento y su valoración.
Grado de contribución de los resultados a la extracción de conocimiento y toma de decisiones.
Utilidad de los conjuntos de documentos más representativos y relacionados a los grupos.
Influencia de los parámetros en la calidad de los resultados.
Confiabilidad de las medidas de evaluación.
Tiempo consumido para obtener los resultados.</td></tr>
</table>

<table>
<tr><td colspan="3">Variables dirigidas a valorar el funcionamiento de los sistemas</td></tr>
<tr><td>6</td><td>Usabilidad</td><td>La operatividad y facilidad de formación y uso, mediante la evaluación de la capacidad y generalidad de las funciones.</td></tr>
<tr><td>7</td><td>Flexibilidad</td><td>La multiplicidad de formas en las cuales usuarios y sistema pueden intercambiar información.</td></tr>
<tr><td>8</td><td>Rendimiento</td><td>La velocidad de procesamiento, tiempo de respuesta y consumo de recursos.</td></tr>
<tr><td>9</td><td>Capacidad de soporte</td><td>La extensibilidad, adaptabilidad, capacidad de configuración, compatibilidad, facilidad de instalación y localización de los problemas.</td></tr>
<tr><td colspan="3">Las medidas que tributan a estas variables son:
Formas de intercambio de información.
Velocidad de procesamiento.
Facilidad de uso
Extensibilidad y adaptabilidad de los sistemas.</td></tr>
</table>

Anexo 31. Encuesta a los usuarios de los sistemas

Esta encuesta se realiza con el propósito de conocer las opiniones que se tienen acerca del sistema y así poder evaluar al mismo siguiendo los criterios de sus usuarios. Los objetivos de la evaluación son investigar y explorar acerca de la calidad y utilidad de los resultados alcanzados con el sistema. Las principales variables de medición son la eficacia, eficiencia, conformidad, confiabilidad y robustez; aunque algunos elementos de usabilidad, flexibilidad, rendimiento y capacidad de soporte serán medidos. Por ello le pediría que fuera tan amable de contestar las preguntas que contribuyen a medir las variables y objetivos trazados. No le tomará más de 20 minutos. Le pedimos que conteste este cuestionario con la mayor sinceridad posible, después de haber probado el sistema con varios conjuntos de documentos. No hay respuestas incorrectas ni correctas. Lea las preguntas cuidadosamente. Las preguntas con variantes de respuestas encabezadas con "O" permiten una única respuesta y aquellas con variantes de respuestas encabezadas con "□" son de selección múltiple.

1. Ha utilizado el sistema:
   O Frecuentemente O Algunas veces O Casi nunca

2. ¿Qué tipo de colecciones de documentos ha utilizado?
   □ Conocidas □ Homogéneas □ Desconocidas □ Heterogéneas □ Desconozco características

3. Temas abordados en las colecciones de documentos utilizadas:
   _____
   _____

4. La familiarización con el sistema fue:
   O Fácil O Relativamente fácil O Aceptable O Relativamente difícil O Difícil

5. Aprender a usar los parámetros fue:
   O Fácil O Aceptable O Difícil O Nunca los uso

6. ¿Se encuentra satisfecho con la división de los documentos en grupos?
   O Definitivamente sí O Casi siempre O Algunas veces O Casi nunca O Definitivamente no

7. ¿Obtiene mejor división de los documentos en grupos al modificar los parámetros para el agrupamiento?
   O Definitivamente sí O Casi siempre O Algunas veces O Casi nunca O Definitivamente no
   O Nunca modifico los parámetros

8. ¿Cuáles son los parámetros que más modifica? _____
   _____

9. La presentación de los resultados del agrupamiento le permiten:
   O Alta interpretabilidad O Adecuada interpretabilidad OBaja interpretabilidad

10. ¿Descubre conocimiento con los resultados del agrupamiento?
    O Definitivamente sí O Casi siempre O Algunas veces O Casi nunca O Definitivamente no

11. ¿Los resultados del agrupamiento contribuyen a la toma de decisiones?
    O Definitivamente sí O Casi siempre O Algunas veces O Casi nunca O Definitivamente no

12. ¿Los documentos que arroja el sistema son a su juicio los más importantes del grupo?
    O Definitivamente sí O Casi siempre O Algunas veces O Casi nunca O Definitivamente no
    Justifique:_____
    _____

13. ¿Los documentos relacionados con cada grupo dan idea de interrelación entre los grupos?
    O Definitivamente sí O Casi siempre O Algunas veces O Casi nunca O Definitivamente no
    Justifique:_____
    _____

14. ¿Obtiene mejores resultados de los conjuntos de documentos más representativos y documentos relacionados por grupos al variar los valores de los parámetros?
    O Definitivamente sí O Casi siempre O Algunas veces O Casi nunca O Definitivamente no
    O Nunca modifico los parámetros
    Justifique:_____
    _____

15. ¿Descubre conocimiento con la caracterización que se realiza de cada grupo de documentos resultante del proceso de agrupamiento?
    O Definitivamente sí O Casi siempre O Algunas veces O Casi nunca O Definitivamente no

16. Las medidas de evaluación intentan dar una valoración de la calidad del agrupamiento realizado. Valores cercanos a 1 indican un mejor agrupamiento. ¿Considera que existe una correspondencia entre la calidad reflejada por las medidas y la calidad que a su juicio tiene la división en grupos?
    a. globalmente: O Sí O En alguna medida O No O Desconozco
    b. por grupo: O Sí O En alguna medida O No O Desconozco

17. ¿Los resultados de la caracterización que se realiza de cada grupo contribuye a la toma de decisiones?
    O Definitivamente sí O Casi siempre O Algunas veces O Casi nunca O Definitivamente no

18. ¿Los resultados le muestran nuevos enfoques e ideas del conjunto de documentos que ha procesado?
    O Definitivamente sí O Casi siempre O Algunas veces O Casi nunca O Definitivamente no

19. ¿El procesamiento automático con el sistema le facilita el análisis de la colección de documentos?
    O Definitivamente sí O Casi siempre O Algunas veces O Casi nunca O Definitivamente no

20. Interpretar los resultados del sistema es:
    O Muy fácil O Fácil O Posible O Difícil O Muy difícil

21. Las salidas del procesamiento que realiza el sistema se obtienen:
    O Muy rápido O Rápido O Tiempo adecuado O Lento O Muy lento

22. El sistema ofrece una ayuda:
    O Muy útil O Adecuada O Poco útil

23. ¿El ambiente del sistema le permite utilizarlo con facilidad?
    O Definitivamente sí O En alguna medida O Definitivamente no

24. ¿Se ofrecen alternativas para la operación del sistema?
    O Definitivamente sí O En alguna medida O Definitivamente no

Anexo 32. Resultados del análisis realizado a partir de criterios recogidos en las encuestas

Tabla A32.1 Conformidad de los usuarios con el agrupamiento y su valoración: (3) algunas veces, (4) casi siempre y (5) definitivamente sí.

	Pregunta 6		Pregunta 12		Pregunta 13	
	SATEX	GARLucene	SATEX	GARLucene	SATEX	GARLucene
3	0	5	14.3	0	0	15
4	14.3	45	57.1	17.6	14.3	65
5	85.7	50	28.6	82.4	85.7	20

Tabla A32.2 Conformidad de los usuarios con los resultados de la validación: (2) en alguna medida y (3) sí.

	Pregunta 16 (local)*		Pregunta 16 (global)	
	SATEX	GARLucene	SATEX	GARLucene
2	8.3	11.1	21.4	20
3	91.7	88.9	78.6	80

Tabla A32.3 Satisfacción de los usuarios respecto a las facilidades para el análisis e interpretación de los resultados brindados por los sistemas: (3) algunas veces, (4) casi siempre y (5) definitivamente sí.

	Pregunta 18		Pregunta 19	
	SATEX	GARLucene	SATEX	GARLucene
3	0	15	0	5.6
4	50	55	28.6	38.9
5	50	30	71.4	55.6

Tabla A32.4 Correlaciones entre conformidad con el agrupamiento y el descubrimiento de conocimiento, toma de decisiones y facilidades para el análisis de la colección de documentos.

Pregunta 6	Tau-b de Kendall	Pregunta 10	Pregunta 11	Pregunta 19
SATEX	Coef. Correlación	.881**	1.000**	.645*
	Significación	.000	.000	.013
GARLucene	Coef. Correlación	.604**	.606**	.540*
	Significación	.008	.005	.020

Tabla A32.5 Correlaciones entre conformidad con la valoración del agrupamiento y el descubrimiento de conocimiento, toma de decisiones y facilidades para el análisis de la colección de documentos.

Sistema	Preguntas	Tau-b de Kendall	Pregunta 15	Pregunta 17	Pregunta 19
SATEX	Pregunta 12	Coef. Correlación	.806**	.638*	.636*
		Significación	.000	.014	.014
	Pregunta 13	Coef. Correlación	.906**	.970**	.645*
		Significación	.000	.000	.013
GARLucene	Pregunta 12	Coef. Correlación	.467*	.411*	.573*
		Significación	.026	.048	.013
	Pregunta 13	Coef. Correlación	.676**	.716**	.593*
		Significación	.004	.003	.023

* Se mostró el porciento válido de las respuestas a esta variable, porque para ambos sistemas hubo dos valores ausentes.