CIC-IPN

INSTITUTO POLITÉCNICO NACIONAL

CENTRO DE INVESTIGACIÓN EN COMPUTACIÓN

LABORATORIO DE LENGUAJE NATURAL Y PROCESAMIENTO DE TEXTO

Minería de texto empleando la semejanza entre estructuras semánticas

TESIS

QUE PARA OBTENER EL GRADO DE
DOCTOR EN CIENCIAS DE LA COMPUTACIÓN

P R E S E N T A

Manuel Montes y Gómez

Director de tesis: Dr. Alexander Gelbukh
Codirector: Dr. Aurelio López López

México, D.F., 2002.

A Liliana, el amor de mi vida,
por su apoyo, cariño y motivación.

Agradecimientos

Hace tres años y medio cuando inicié el doctorado, el final se veía muy lejano y el camino parecía muy complejo. Hoy que he logrado mi objetivo, el principio me parece cercano y el camino recorrido lo observo sencillo. Esta sensación, estoy seguro, la debo en gran manera al apoyo de muchas personas e instituciones. En forma especial quiero agradecer a los siguientes:

A mi asesor, el *Dr. Alexander Gelbukh*, por su apoyo incondicional, sus múltiples consejos y sus valiosas enseñanzas.

A mi coasesor, el *Dr. Aurelio López*, por toda su motivación, su enorme confianza, y su paciencia.

Al *Dr. Ricardo Baeza*, por la gran oportunidad de realizar una estancia de investigación con él en la Universidad de Chile. También por sus muchos consejos, y por todas sus experiencias compartidas.

A los doctores *Igor Bolshakov*, *Mikhail Alexandrov* y *Gregori Sidorov* del CIC, y al *Dr. Edgar Chávez* de la Universidad Michoacana por su interés en mi trabajo, sus consejos, sus críticas, pero sobre todo por su valiosa amistad.

A mis compañeros y amigos, *Sofía Galicia* y *Raúl Carrasco* del CIC, *Alberto Méndez* y *Pilar Tapia* del INAOE, y *Cesar Collazo* de la Universidad de Chile, por todo su apoyo, su compañerismo, y sus valiosas sugerencias y comentarios.

Al *CONACYT* por el apoyo económico que me otorgó durante estos tres años y medio; al programa internacional *ALPHA-Parnet* por la beca que me permitió realizar una estancia de investigación de diez meses en la Universidad de Chile; al *Centro de Investigación en Computación*, y en especial al Dr. Guzmán Arenas, por la gran oportunidad de realizar mis estudios en la mejor institución de computación del país; al *Instituto Nacional de Astrofísica, Óptica y Electrónica* y a la *Universidad de Chile* por su participación en mi preparación y su inmenso apoyo técnico y material.

Finalmente, a mi familia, mis *papás*, *hermano*, *suegros* y *cuñado*, así como a mi familia chilena, *tíos y primos Contesse Marroquín*, por su gran cariño y motivación. También a mi *esposa* por su paciencia, confianza, y por todas esas alegrías y preocupaciones que compartimos.

Abstract

Knowledge is the most valuable treasure of humankind. Most of this knowledge exists in some form of natural language such as books, journals, reports, etc. The real possession of all this knowledge depends on our capabilities for performing different tasks with texts, such as: to search for interesting texts, to compare different texts, or to summarize them.

Text mining, an emerging research area that can be roughly characterized as knowledge discovery in large text collections, is focused on this kind of tasks. It is concerned mainly with the discovery of interesting patterns such as clusters, associations, and deviations in text collections. Current methods of text mining tend to use simplistic and shallow representations of texts, e.g., keyword sets or keyword frequency vectors. On the one hand, such representations are easy to obtain from texts and easy to analyze. On the other hand, however, they restrict the knowledge discovery results to the topic level.

To obtain more useful and meaningful results, richer text representations are necessary. On the basis of this assumption, we propose **a method for doing text mining at detail level**. This method uses **conceptual graphs** for representing text content, and relies of performing some tasks on these graphs, allowing the discovery of more descriptive patterns. These tasks are mainly the following:

- **Comparison of two conceptual graphs**. Our method of comparison of conceptual graphs builds both a qualitative and a quantitative description of similarity of the graphs. It takes into account a taxonomy (synonymy and subtype/supertype relationships between the concepts and relations appearing in the conceptual graphs), thus allowing greater flexibility in approximate matching.
- **Conceptual clustering of a set of graphs**. The method we propose for clustering a set of conceptual graphs finds all *regularities* in a given set of graphs. This method, unlike the traditional cluster analysis techniques, allows not only dividing

the set of graphs into several groups, but also associating a description to each group and organizing them in a hierarchy, where the lower nodes indicate specialized regularities and the upper nodes suggest generalized regularities. We consider this clustering no only an interesting discovery, but also a kind of index of the set of graphs that permits discovering of other interesting patterns such as associations and deviations.

- **Discovery of association rules between conceptual graphs**. The method of association discovery finds rules of the form $g_i \Rightarrow g_j\,(c,s)$ in a given set of graphs. Such a rule indicates that conceptual graphs containing (i.e., specializations of) the graph g_i in c% of times also contain the more specialized graph g_j; while s% of the graphs of the collection contain the graph g_j. One interesting characteristic of such rules is that they can express associations at different levels of generalization.

- **Detection of deviations in a set of conceptual graphs**. Our method of deviation detection allows uncovering contextual deviations of the form $g_i : g_j\,(r,s)$. Such an expression indicates that for the context g_i (the set of graphs containing g_i), with coverage of s%, the graphs containing g_j are rare and represent the r% of the context. One attractive characteristic of this method is that one can detect both global and local (for a specific context) deviations.

The thesis is organized as follows. Chapter 1 defines the research problem. Chapter 2 describes previous work on text mining. Chapter 3 introduces the main features of conceptual graphs. Chapter 4 proposes a method of comparison (approximate matching) of conceptual graphs. Chapter 5 presents our methods for conceptual clustering, association discovery and deviation detection in a set of conceptual graphs. Chapter 6 shows some experimental results from the analysis of two sets of papers; one on computer science and the other on information science. Finally, chapter 7 summarizes the main contributions of the thesis and the limitations of the proposed method, as well as discusses the future work.

Contenido

Vista general de la tesis

Índice detallado de la tesis

Índice de Figuras

Capítulo 1

Introducción

En este capítulo discutimos la necesidad de crear nuevas estrategias que permitan analizar y describir el contenido de colecciones de textos. Con base en esta necesidad plantamos los objetivos de investigación de este trabajo; los cuales se enfocan en el diseño de un método de minería de texto que permita usar una representación semántica del contenido de los textos (por ejemplo, grafos conceptuales), y que en consecuencia logre descubrir patrones descriptivos del contenido de los textos que consideren información sobre entidades, acciones, atributos y sus relaciones.

Como parte final del capítulo describimos la organización de la tesis y el contenido de los capítulos subsiguientes.

Introducción

1.1 Motivación

El tesoro más valioso de la raza humana es el conocimiento. Gran parte de este conocimiento existe en forma de lenguaje natural: libros, periódicos, artículos, etcétera. La posesión real de todo este conocimiento depende de nuestra habilidad para hacer ciertas operaciones con la información, por ejemplo: buscar información interesante, comparar fuentes de información diferentes y resumir grandes conjuntos de información.

La lingüística computacional se enfoca principalmente en el diseño de los mecanismos que permitan a las computadoras entender el lenguaje natural, aunque también considera varias tareas relacionadas con el procesamiento de información textual. Algunos ejemplos de estas tareas son la búsqueda de información, la extracción de información y la *minería de texto*.

El desarrollo de los métodos para el procesamiento de información textual ha sido paralelo al desarrollo de los métodos para la comprensión del lenguaje (análisis morfológico, sintáctico y semántico). Por ello, típicamente se busca y analiza la información textual considerando únicamente el "tema" de los textos y no su contenido completo. Esta estrategia facilita el análisis de grandes conjuntos de textos, e incluso mantiene una independencia del dominio, pero limita grandemente la expresividad y la diversidad de los resultados de los sistemas de análisis de textos.

En la recuperación de información, por ejemplo, esta estrategia de análisis impide hacer búsquedas que consideren detalles del contenido de los textos que van más allá de sus temas (por ejemplo: propósitos, planes, objetivos y enfoques). Por su parte, en la minería de texto, esta estrategia impide descubrir patrones interesantes relacionados con dichos detalles del contenido de los textos.

Actualmente, buscando una solución a este problema de expresividad y diversidad de los resultados, se comienzan a usar más elementos provenientes de la lingüística computacional –comprensión del lenguaje– en las tareas de procesamiento de textos. Así pues, se empiezan a sustituir las representaciones sencillas de los textos, como las listas de palabras clave, por representaciones más completas que consideran aspectos estructurales y contextuales del contenido de los textos.

En la recuperación de información se han usado tanto representaciones sintácticas, como representaciones semánticas del contenido de los textos (Metzler *et al.*, 1984; Metzler and Haas, 1989; Schwarz, 1990; Mauldin, 1991; Girardi and Ibrahim, 1994; Chakravarthy and Haase, 1995; Myaeng, 1990; Myaeng, 1992; Myaeng and Khoo, 1992; Ellis and Lehmann, 1994; Myaeng and Khoo, 1994; Huibers, Ounis and Chevallet, 1996; López-López and Myaeng, 1996; Genest and Chein, 1997; Yang *et al*., 1992; Montes-y-Gómez *et al*., 2000), aunque su aplicación no ha sido tan definitiva y valiosa como se esperaba (Khoo, 1997, Sparck-Jones, 1999). Las principales causas de este resultado desfavorable son, entre otras, las siguientes dos:

1. Los métodos de comparación de las nuevas representaciones no son los adecuados
2. Algunas características de la búsqueda de información, por ejemplo, su naturaleza temática, la rapidez de respuesta requerida, y en muchas ocasiones la necesidad de independencia del dominio, complican la aplicación de estas nuevas representaciones.

En la minería de texto *no* se han usado representaciones que consideren algunos elementos estructurales y contextuales de los textos; ello a pesar de que tanto su objetivo, el descubrimiento de conocimiento, como algunas de sus características hacen suponer una notable mejoría en los resultados. Algunas de estas características son:

1. El descubrimiento de conocimiento es una tarea típicamente dependiente del dominio.

2. La rapidez no es un factor determinante en el proceso de descubrimiento, por el contrario, lo más importante es la *expresividad* y precisión de los resultados.
3. El proceso de descubrimiento generalmente no se realiza en un ambiente de pregunta y respuesta.

Este trabajo de investigación tiene que ver con el problema de la expresividad de los resultados de la minería de texto, y también con la oportunidad de comenzar a usar representaciones más completas del contenido de los textos en ella. Básicamente en este trabajo se plantea el uso de una representación "semántica" del contenido de los textos, y se proponen algunos métodos para el descubrimiento de patrones interesantes en un conjunto de dichas representaciones.

Así pues, este trabajo pretende definir algunas estrategias de minería de texto que mejoren la expresividad y la diversidad de los patrones descubiertos con respecto a los obtenidos usando las técnicas tradicionales.

1.2 Descripción del problema

La minería de texto es el área de investigación más reciente del procesamiento de textos. Ella se define como *el proceso de descubrimiento de patrones interesantes en una colección de textos*. Estos patrones no deben de existir explícitamente en ningún texto de la colección, y deben de surgir de relacionar el contenido de varios de ellos (Hearst, 1999; Kodratoff, 1999).

El proceso de minería de texto consiste de dos etapas principales: una etapa de *preprocesamiento* y una etapa de *descubrimiento* (Tan, 1999). En la primera etapa, los textos se transforman a algún tipo de representación estructurada o semiestructurada que facilite su posterior análisis, mientras que en la segunda etapa las representaciones intermedias se analizan con el objetivo de descubrir en ellas algunos patrones interesantes o nuevos conocimientos.

Etapa de pre-procesamiento	Tipo de representación	Tipo de descubrimientos
Categorización	Vector de temas	Nivel temático
Full-text	Secuencia de palabras	Patrones de lenguaje
Extracción de información	Tabla de datos	Relaciones entre entidades

Figura 1.1 Estado del arte de la minería de texto

Entonces, dependiendo del tipo de métodos aplicados en la etapa de preprocesamiento son el tipo de representaciones intermedias construidas, y en función de dicha representación son el tipo de métodos usados en la etapa de descubrimiento, y en consecuencia, el tipo de patrones descubiertos.

La figura 1.1 muestra las principales estrategias usadas en los actuales sistemas de minería de texto. De acuerdo con esta figura, la mayoría de los actuales de minería de texto limitan sus resultados a un nivel temático o de entidad, y por lo tanto imposibilitan el descubrimiento de cosas más detalladas como:

- *Consensos*, que por ejemplo respondan a preguntas como: ¿Cuál es la opinión mayoritaria de los mexicanos sobre el gobierno de Fox?
- *Tendencias*, que indiquen por ejemplo si han existido variaciones en la postura de Fox con respecto a la educación.
- *Desviaciones*, que identifiquen por ejemplo opiniones “raras” con respecto al desempeño de la selección mexicana de fútbol.

Una idea para mejorar la expresividad y la diversidad de los descubrimientos de los sistemas de minería de textos consiste en usar una representación del contenido de los textos más completa que las representaciones usadas actualmente. Siguiendo este enfoque, en esta investigación se propone usar los *grafos conceptuales* (Sowa, 1984; 1999) como representación del contenido de los textos.

Los grafos conceptuales se seleccionaron de entre otras representaciones (por ejemplo, árboles sintácticos, predicados lógicos y KL-ONE) por las siguientes razones:

1. Los grafos conceptuales permiten representar adecuadamente –en términos de expresividad y eficiencia notacional– la información en lenguaje natural.[1]
2. Los grafos conceptuales disponen de mecanismos formales que facilitan su manipulación, transformación y análisis (ver sección 3.2).
3. Los grafos conceptuales se usan en otras tareas afines a la minería de texto, por ejemplo en la recuperación de información y en el agrupamiento de textos.

Ahora bien, realizar la minería de texto usando los grafos conceptuales como representación del contenido de los textos involucra dos problemas diferentes. En primer lugar, la transformación de los textos en grafos conceptuales, y en segundo lugar, el análisis automático de un conjunto de grafos conceptuales.

La transformación de los textos en grafos conceptuales es un problema complejo vinculado con su análisis sintáctico y semántico (Sowa and Way, 1986; Sowa 1991). Debido a ello, todos los métodos que existen para realizar dicha transformación se enfocan en un dominio único y restringido. Algunos ejemplos son los siguientes:

- Partes de *artículos científicos* a grafos conceptuales (Myaeng and Khoo, 1994; López-López, 1995; Montes-y-Gómez, 1998; Montes-y-Gómez *et al*., 1999e).
- Partes de *expedientes médicos* a grafos conceptuales (Baud *et al.*, 1992; Rassinoux *et al.*, 1994).
- Partes de *casos legales* a grafos conceptuales (Boucier and Rajman, 1994).
- Partes de *manuales de referencia* a grafos conceptuales (Petermann, 1996).

[1] La adecuación de una representación de conocimiento a un dominio específico de aplicación se caracteriza en términos de su *expresividad* y de su *eficiencia notacional* (Woods, 1986). La expresividad indica la cantidad de elementos del dominio de aplicación que ésta puede describir, mientras que la eficiencia notacional se relaciona con la facilidad con que puede hacerse esto.

Por su parte, el análisis automático de un conjunto de grafos conceptuales orientado al descubrimiento de patrones descriptivos –nuevos conocimientos– es un problema *poco* estudiado (Mineau and Godin, 1995; Godin *et al.*, 1995; Bournaud and Ganascia, 1996; Bournaud and Ganascia, 1997; Möller, 1997). Así pues, este trabajo de investigación se enfoca en el análisis de un conjunto de grafos conceptuales –que representan el contenido de un conjunto de textos–, y en el descubrimiento de varios tipos de patrones interesantes, por ejemplo: *agrupamientos*, *asociaciones* y *desviaciones*.

1.3 Objetivos de la tesis

Objetivo general

Nosotros asumimos, al igual que muchos otros, que disponer de más y "mejor" información del contenido de los textos permitirá descubrir más y mejores conocimientos a partir de ellos. Con base en esta suposición, el objetivo general de esta investigación es diseñar un nuevo método de minería de texto apto para emplear los *grafos conceptuales* como representación del contenido de los textos, y a su vez, capaz de trasladar los descubrimientos del nivel temático a un nivel de *mayor detalle* –un nivel más descriptivo.

Objetivos específicos

Los grafos conceptuales son una representación del contenido de los textos mucho más compleja que las representaciones usadas en la minería de texto actual. Debido a esto, la mayoría de sus métodos no pueden aplicarse directamente para analizar un conjunto de grafos conceptuales, y en consecuencia es necesario diseñar nuevas técnicas para analizarlos.

A continuación se definen los objetivos específicos de este trabajo. Estos objetivos se enfocan en el diseño de las nuevas técnicas de análisis de un conjunto de grafos conceptuales.

1. Proponer un método adecuado para *comparar* dos grafos conceptuales cualesquiera.

Este método deberá describir tanto cualitativamente como cuantitativamente la semejanza entre los dos grafos. La descripción cualitativa deberá considerar las *semejanzas no-exactas* entre los dos grafos. Por ejemplo, las semejanzas a diferentes niveles de generalización.

Por su parte, la descripción cuantitativa deberá evaluar la semejanza de los grafos de acuerdo con los *intereses del usuario*.

2. Diseñar un método para *agrupar conceptualmente* un conjunto de grafos conceptuales.

Este método deberá descubrir todas las regularidades –generalizaciones comunes– implícitas en el conjunto de grafos. También permitirá visualizar dichos regularidades considerando un punto de vista específico.

3. Desarrollar los métodos para *descubrir asociaciones* y *detectar desviaciones* en un conjunto de grafos conceptuales.

Estos dos métodos deberán aprovechar el agrupamiento conceptual de los grafos, y en consecuencia permitir el descubrimiento de asociaciones y desviaciones a diferentes niveles de generalización.

Además, el método de descubrimiento de asociaciones permitirá descubrir asociaciones con distintos valores de confianza y soporte; mientras que el método de detección de desviaciones permitirá detectar no sólo los grafos raros, sino también las características comunes a dichos grafos.

1.4 Organización de la tesis

El resto del documento se organiza de la siguiente manera.

En el capítulo 2 se presenta una breve *revisión del estado del arte de la minería de texto*. En ella se discuten las principales tareas de la minería de datos, se plantea

la minería de texto como una extensión de esta última, se explican los principales métodos de la minería de texto y finalmente se mencionan algunas tendencias de investigación.

En el capítulo 3 se introducen los elementos básicos de la *teoría de grafos conceptuales*. Principalmente se presenta su terminología elemental y se describe la operación de generalización; operación básica del método de minería de texto expuesto en los capítulos subsecuentes. Además, en la parte final del capítulo se ejemplifica la representación de algunos elementos del lenguaje natural con grafos conceptuales y se describe el proceso tradicional para su transformación.

En el capítulo 4 se propone un método para la *comparación de los grafos conceptuales*. Este método tiene dos etapas principales: el casamiento de los grafos conceptuales y la medición de su semejanza. Ambas etapas se explican e ilustran detalladamente. Además, en cada caso se describe el uso de conocimiento del dominio y la posibilidad de considerar los intereses del usuario.

En el capítulo 5 se presentan algunos métodos para el *descubrimiento de patrones descriptivos* en un conjunto de grafos conceptuales. En la primera parte se presenta un método incremental para agrupar un conjunto de grafos conceptuales. Después, en la segunda parte, se describen unos métodos para descubrir asociaciones y detectar de desviaciones entre grafos conceptuales. Estos métodos utilizan el agrupamiento de los grafos como un índice de la colección, y en consecuencia descubren patrones detallados que consideran entidades, acciones, atributos y sus relaciones semánticas.

En el capítulo 6 se muestran algunos *resultados experimentales*. Estos resultados describen el análisis de dos colecciones de artículos científicos (la descripción de estas colecciones, y el proceso de construcción de sus grafos conceptuales se presenta en el apéndice A). Los resultados descritos son básicamente de dos tipos: cualitativos y cuantitativos. Los resultados cualitativos demuestran la capacidad del método propuesto para descubrir patrones interesantes a un nivel más descriptivo y completo

que el temático, y los resultados cuantitativos muestran principalmente la viabilidad del método de minería de texto propuesto.

Finalmente, en el capítulo 7 se discuten las principales *aportaciones* de esta investigación, y se mencionan algunos trabajos importantes para el futuro.

Capítulo 2

Antecedentes

En este capítulo presentamos una breve revisión del estado del arte de la minería de texto. En ella se introducen los conceptos básicos de la minería de datos tradicional, y se ilustran algunas de sus tareas principales. También se plantea el surgimiento de la minería de texto como una respuesta a la incapacidad de los métodos de minería de datos para analizar información textual. Finalmente se describen los métodos de la minería de texto actual haciendo énfasis en el tipo de patrones descubiertos, y se mencionan las principales tendencias de investigación.

Antecedentes

2.1 Minería de datos*

El origen de la minería de datos se relaciona con dos factores. Por una parte, la disponibilidad de grandes cantidades de datos almacenados electrónicamente; y por otra parte, la necesidad de transformar toda esta información en conocimiento útil para la toma de decisiones en diferentes escenarios de aplicación.

Así pues, la minería de datos es el resultado "natural" de la evolución de los sistemas de información. La figura 2.1 ilustra esta evolución. Allí se observa que primero se desarrollaron las estructuras necesarias para almacenar grandes cantidades de datos, es decir, se crearon las bases de datos. Después se diseñaron los mecanismos para administrar dichas bases de datos, por ejemplo, se implementaron varios

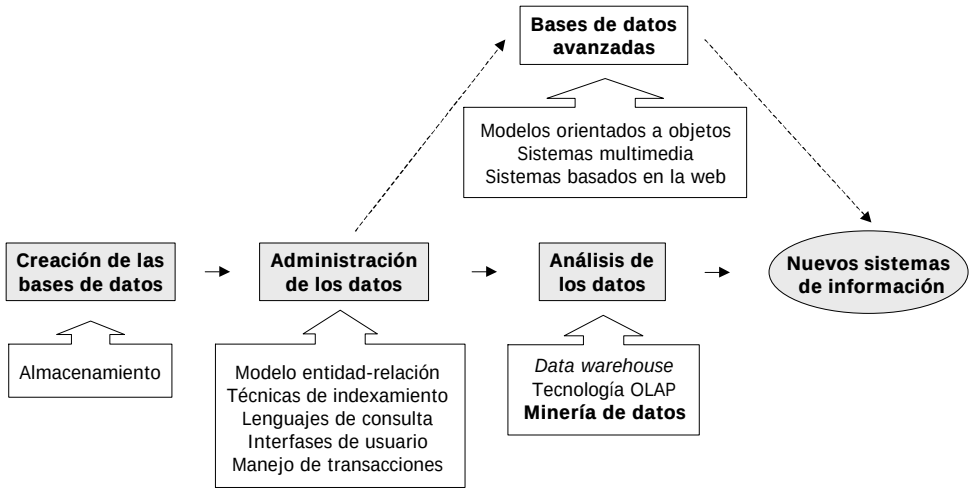


Figura 2.1 Evolución de los sistemas de información

* La revisión del estado del arte de la minería de datos se basa principalmente en las siguientes referencias (Fayyad *et al.*, 1996b; Weiss and Indurkhya, 1998; Witten and Frank, 1999; Han and Kamber, 2001);

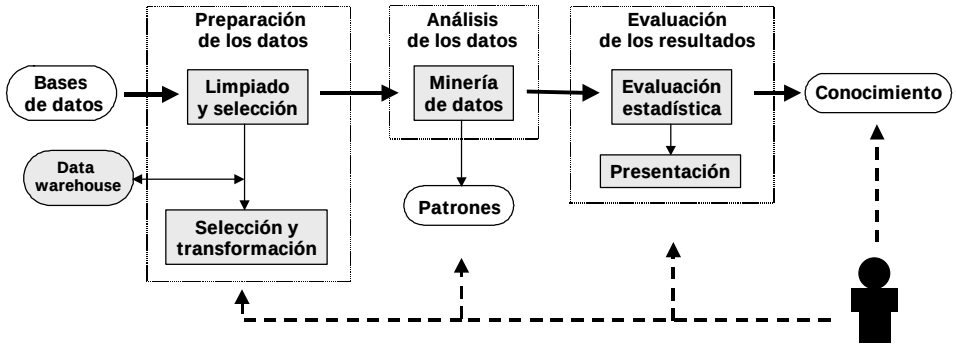


Figura 2.2 Proceso de descubrimiento de conocimiento

métodos de consulta; y ahora se desarrollan algunas herramientas para analizar y entender los datos. Un ejemplo de estas herramientas son los sistemas de *descubrimiento de conocimiento en bases de datos.*

2.1.1 Descubrimiento de conocimiento en bases de datos

El proceso de descubrimiento de conocimiento en bases de datos se ilustra en la figura 2.2. Su objetivo es identificar patrones válidos, novedosos y potencialmente útiles en grandes bases de datos.

Básicamente, el proceso de descubrimiento de conocimiento en bases de datos considera las siguientes etapas:

- **Preparación de los datos**

En esta etapa se eliminan los datos inconsistentes y se combinan distintas fuentes de datos en un solo gran almacén de datos (*data warehouse*, en inglés).

Además, en esta etapa se separaran los datos útiles (o interesantes), y se transforman en algún formato apropiado para su posterior análisis.

- **Análisis de los datos**

Esta etapa, llamada comúnmente *minería de datos*, es la parte medular del proceso de descubrimiento de conocimiento en bases de datos. Su objetivo es identificar

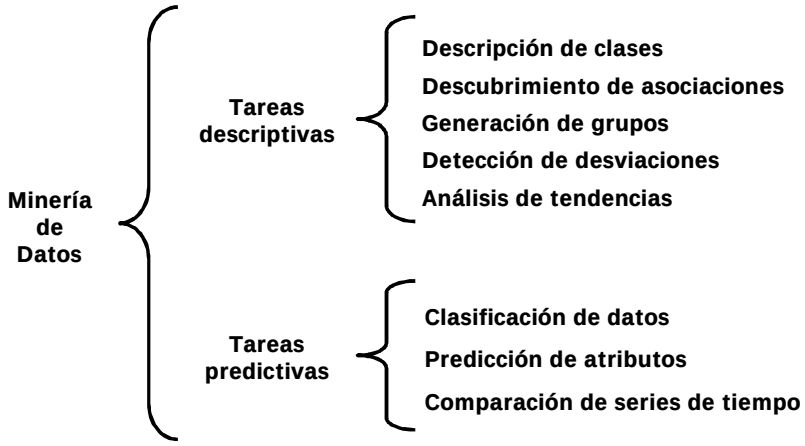


Figura 2.3 Tipos de tareas de minería de texto

distintos tipos de patrones descriptivos de los datos, por ejemplo: desviaciones, tendencias, asociaciones y grupos.

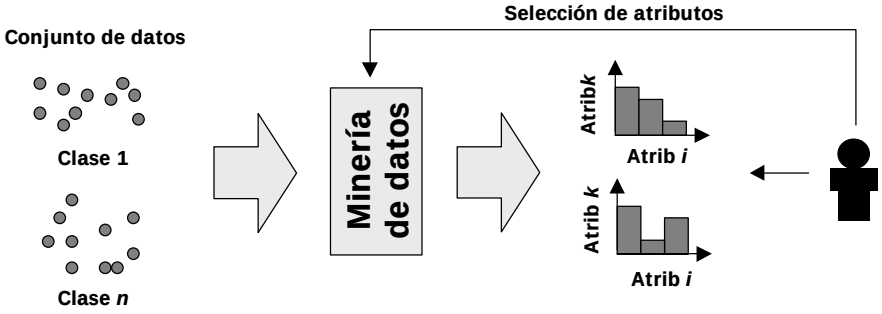
- **Evaluación de los resultados**

En esta etapa se aplican distintas medidas, principalmente estadísticas, para identificar los patrones *más interesantes*. Además se usan varias técnicas para visualizar los patrones descubiertos, y de esta forma facilitar la interacción del usuario con el sistema.

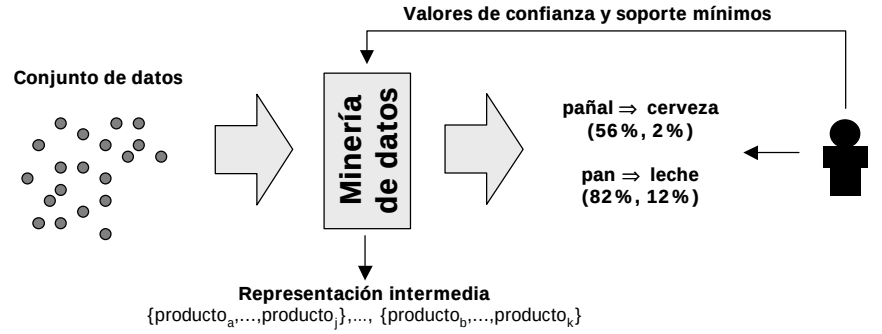
2.1.2 Tareas de minería de datos

La minería de datos, como se ha mencionado, es la etapa central del proceso de descubrimiento de conocimiento en bases de datos. En ella se realizan varias tareas que permiten identificar distintos tipos de patrones en un conjunto de datos. En general, estas tareas son de dos tipos: descriptivas y predictivas (ver figura 2.3).

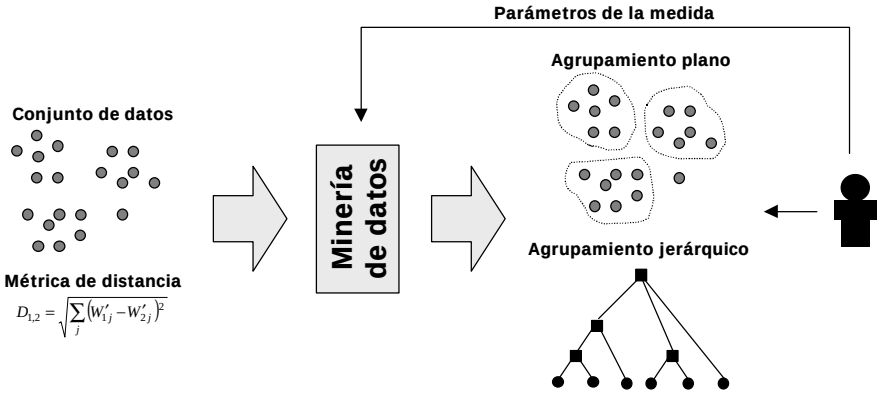
Las *tareas descriptivas* caracterizan las propiedades generales de los datos y construyen descripciones compactas de estos. Por su parte, las *tareas predictivas* hacen inferencias sobre los datos conocidos con el objetivo de predecir el comportamiento de datos nuevos.



(a) Descripción de clases



(b) Descubrimiento de asociaciones



(c) Generación de grupos

Figura 2.4. Algunas tareas de minería de datos

A continuación se describen brevemente las principales tareas de minería de datos.

- **Descripción de clases**

La descripción de clases consiste básicamente en construir una descripción resumida de los datos de una clase. Esta descripción se representa comúnmente como el caso típico de la clase, o como una gráfica (o cubo de datos) basada en un conjunto predefinido de atributos.

Básicamente, esta tarea permite visualizar adecuadamente y comparar distintas clases de datos. La figura 2.4(a) ejemplifica esta tarea.

- **Descubrimiento de asociaciones**

El descubrimiento de asociaciones consiste en encontrar las principales *reglas asociativas* entre los atributos de un conjunto de datos. Estas reglas son expresiones de la forma $A \Rightarrow B\,(confianza\,/\,soporte)$, que indican que las transacciones que tienen el conjunto de atributos *X*, un porcentaje significativo de las veces (indicado por el valor de confianza) también tienen el conjunto de atributos *Y*, y además que un porcentaje del total de las transacciones (indicado por el valor de soporte) tienen ambos conjuntos de atributos.

La figura 2.4(b) ejemplifica el tipo de asociaciones descubiertas por los sistemas de minería de datos. En este caso, las reglas asociativas corresponden a una base de datos hipotética de un supermercado.

- **Generación de grupos**

La generación de grupos es una técnica útil para la exploración de grandes conjuntos de datos. Su objetivo es dividir automáticamente un conjunto de datos –previamente no clasificados– en varios grupos “homogéneos”.

Típicamente los algoritmos de agrupamiento utilizan una medida de distancia o semejanza entre los datos en cuestión, e intentan dividir dichos datos en grupos que maximicen la semejanza entre los elementos de un mismo grupo y minimicen la semejanza entre los elementos de grupos diferentes.

Existen varias formas de representar los grupos; las más comunes son los agrupamientos planos y los agrupamientos jerárquicos. En la figura 2.4(c) se ilustra el agrupamiento de un conjunto de datos.

- **Detección de desviaciones**

Los sistemas tradicionales de análisis de datos consideran que las desviaciones son un problema, y por lo tanto buscan minimizar sus efectos. Por el contrario, los sistemas de minería de datos consideran que las desviaciones son un tipo de patrón interesante. Así pues, el objetivo de los métodos de detección de desviaciones es determinar los elementos raros –diferentes a la "norma"– dentro de un conjunto de datos.

Existen tres enfoques para detectar desviaciones en un conjunto de datos: un *enfoque estadístico*, donde se asume un modelo probabilístico para los datos, y los datos "ajenos" a este modelo son considerados desviaciones (Barnett and Lewis, 1994); un *enfoque basado en distancia*, donde los datos con un número reducido de elementos cercanos son considerados desviaciones (Knorr and Ng, 1998; Breunig, 1999); y un *enfoque basado en regularidades*, donde los elementos que se "desvían" mayormente de las características principales del conjunto son las desviaciones (Arning *et al.*, 1996).

- **Clasificación de datos**

La clasificación es el proceso de encontrar un conjunto de funciones o modelos que describan y distingan las distintas clases de datos, con el propósito de usar estos modelos para determinar la clase a la que pertenece un nuevo dato.

Los modelos (o funciones) de clasificación se construyen con base en un conjunto de entrenamiento, y pueden expresarse de diferentes formas, por ejemplo: reglas *IF-THEN*, árboles de decisión, y redes neuronales.

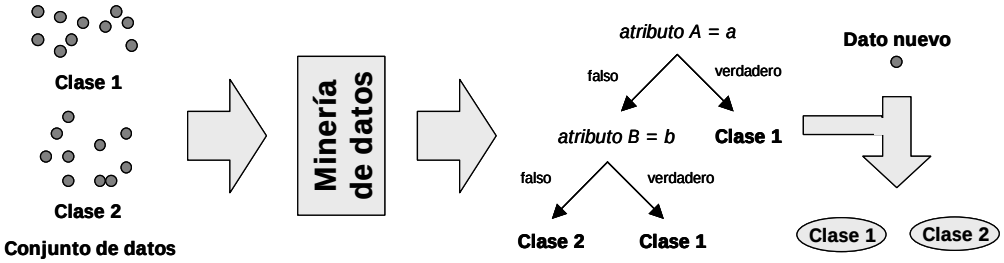


Figura 2.5 Clasificación de datos

La figura 2.5 ejemplifica un sistema de clasificación de datos. En ella se muestran un tipo común de reglas clasificación descubiertas por estos sistemas.

2.2 Minería de texto

La minería de datos se enfoca en el análisis de grandes bases de datos. Debido a ello, sus métodos consideran solamente información estructurada, principalmente numérica y booleana, y descuidan otros tipos de información. Como consecuencia de esta situación, muchos logros de la minería de datos parecen tareas muy difíciles de realizar con datos no-estructurados o semiestructurados.[1] Por ejemplo, dada una colección de textos parece muy complicado descubrir automáticamente cosas tales como:

- *Resúmenes*, que contesten a preguntas como: ¿De qué trata este documento?
- *Consensos*, que por ejemplo respondan a preguntas como ¿Cuál es el consenso de los mexicanos sobre el primer año de gobierno del presidente Fox?
- *Tendencias*, que indiquen por ejemplo si han existido variaciones en la postura de Fox con respecto al tema de la educación.
- *Desviaciones*, que identifiquen por ejemplo opiniones "raras" con respecto al desempeño de la selección mexicana de fútbol.

[1] Incluso, varios logros de los sistemas tradicionales de bases de datos parecen tareas complicadas cuando los datos son un conjunto de textos.

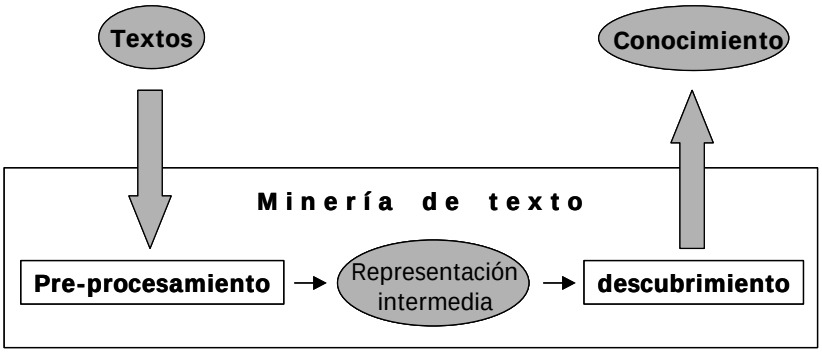


Figura 2.6 Proceso de minería de texto

- *Máximos* y *mínimos*, que permitan contestar preguntas como ¿Cuál de los países apoya más a los Estados Unidos? ¿Y cuál menos?
- *Dependencias*, que permitan identificar por ejemplo las posiciones que surgieron o desaparecieron después del ataque a Afganistán.

En este ámbito de poca exploración de la información textual, y de poca capacidad de los métodos de minería de datos para su análisis, surge la *minería de texto*.

Así pues, la minería de texto es una extensión de la minería de datos que pretende trasladar los objetivos, métodos, técnicas y logros de esta última al ámbito de la información textual (Uthurusamy, 1996; Agrawal *et al.*, 1996; Tan, 1999).

2.2.1 Proceso de minería de texto

La minería de texto se define, parafraseando la minería de datos, como el proceso de descubrimiento de patrones interesantes –y posiblemente nuevos conocimientos– en un conjunto de textos (Feldman and Dagan, 1995). La idea es que estos patrones no deben existir explícitamente en ningún texto de la colección, y deben surgir de relacionar el contenido de varios de ellos (Hearst, 1999; Kodratoff, 1999).

El proceso de minería de texto se ilustra en la figura 2.6. Este proceso consiste de dos etapas principales: una etapa de *preprocesamiento* y una etapa de *descubrimiento* (Tan, 1999).

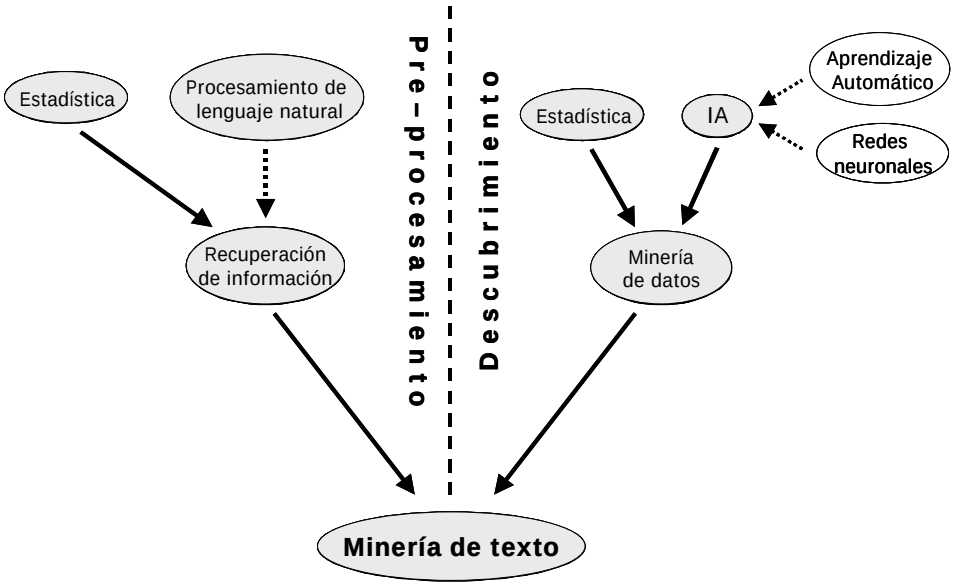


Figura 2.7 Antecedentes de la minería de texto

En la primera etapa, los textos se transforman a algún tipo de representación estructurada o semiestructurada que facilite su posterior análisis, mientras que en la segunda etapa, estas representaciones intermedias se analizan con el objetivo de descubrir en ellas algunos patrones interesantes.

La minería de texto es también un proceso *multidisciplinario* que conjuga métodos provenientes de distintas áreas (ver la figura 2.7). Por ejemplo, en la etapa de preprocesamiento se emplean algunos métodos provenientes principalmente de la recuperación de información, mientras que en la etapa de descubrimiento se usan varios métodos de la minería de datos. Estos últimos son en su mayoría de tipo estadístico, aunque también algunos incorporan técnicas provenientes del aprendizaje automático.

A continuación se describen los principales métodos empleados en ambas etapas de la minería de texto.

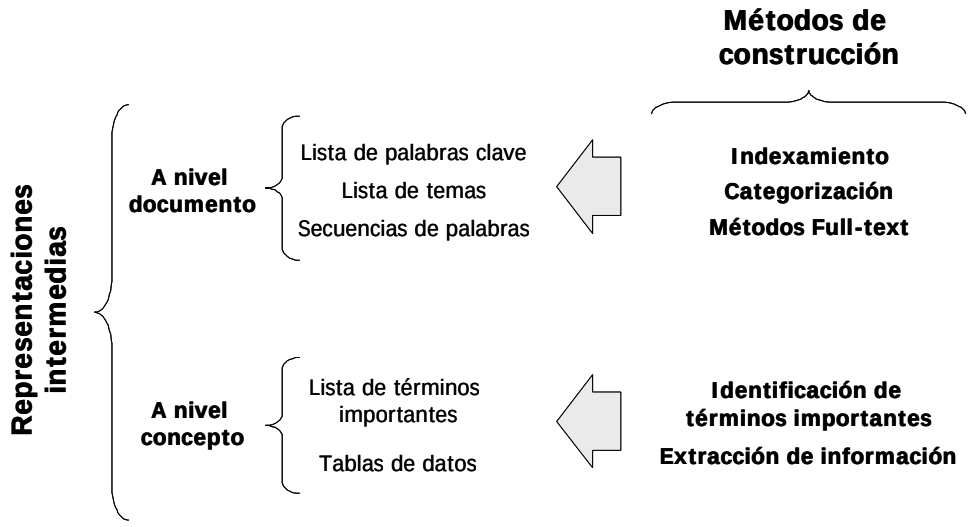


Figura 2.8 Métodos de preprocesamiento

2.2.1.1 Etapa de preprocesamiento

La etapa de preprocesamiento es la etapa del proceso de minería de texto donde se transforman los textos a una representación estructurada o semiestructurada de su contenido.

Las representaciones intermedias de los textos deben ser, por una parte, *sencillas* para facilitar el análisis de los textos, pero por otra parte, *completas* para permitir el descubrimiento de patrones interesantes, e incluso de nuevos conocimientos.

En la figura 2.8 se muestran las representaciones intermedias más usadas en la minería de texto. Estas representaciones son básicamente de dos tipos (Tan, 1999):

1. A *nivel documento*, donde cada representación se refiere a un texto diferente de la colección.
2. A *nivel concepto*, donde cada representación indica un objeto, tema o concepto interesante para el dominio específico de aplicación. [2]

[2] Un texto puede tener varios conceptos interesantes; por lo tanto también puede propiciar varias representaciones a nivel concepto.

Representaciones a nivel documento	Minería de datos	Representaciones a nivel concepto
Eliminación de palabras vacías	**Limpiado**	Eliminación de símbolos no textuales
Relevancia estadística de las palabras	**Selección / Filtrado**	Casamiento de patrones sintáctico-semánticas
Lectura de diferentes formatos de texto	**Integración**	Lectura de diferentes formatos de texto
Marcaje de palabras de diccionario de dominio	**Enriquecimiento**	Etiquetamiento sintáctico

Figura 2.9 Operaciones de preprocesamiento

La construcción de estas representaciones sigue diferentes estrategias. Por ejemplo, las representaciones a nivel documento se construyen típicamente usando métodos de *categorización*, texto completo (*full-text*, en inglés) o *indexamiento* (Feldman and Dagan, 1995; Lagus *et al.*, 1999; Merlk, 1997; Rajman and Besançon, 1997; Rajman and Besançon, 1998; Feldman *et al.*, 1997 Ahonen *et al.*, 1997a; Montes-y-Gómez *et al.*, 2001; Fujino *et al.*, 2000).

Por su parte, las representaciones a nivel concepto se obtienen básicamente aplicando métodos dependientes del dominio, tales como: la *extracción de términos importantes* y la *extracción de información* (Feldman *et al.*, 1998a; Feldman *et al.*, 1998b; Feldman *et al.*, 1998c; Nahm and Mooney, 2000; Nahm and Mooney, 2001a, Montes-y-Gómez *et al.*, 1999a; Hull, 1998; Feldman *et al.*, 1999).

En general, los métodos de preprocesamiento provienen de la recuperación de información, pero a pesar de ello comparten varias características u operaciones con los métodos de preprocesamiento de la minería de datos. Algunas de estas operaciones se enumeran en la figura 2.9.

2.2.1.2 Etapa de descubrimiento

Típicamente, los descubrimientos de minería de texto –y por consecuencia sus métodos y sus tareas– se clasifican en: descriptivos y predictivos. Sin embargo es posible clasificarlos de otras maneras. Por ejemplo, la figura 2.10 muestra una clasificación alternativa de los descubrimientos de minería de texto.

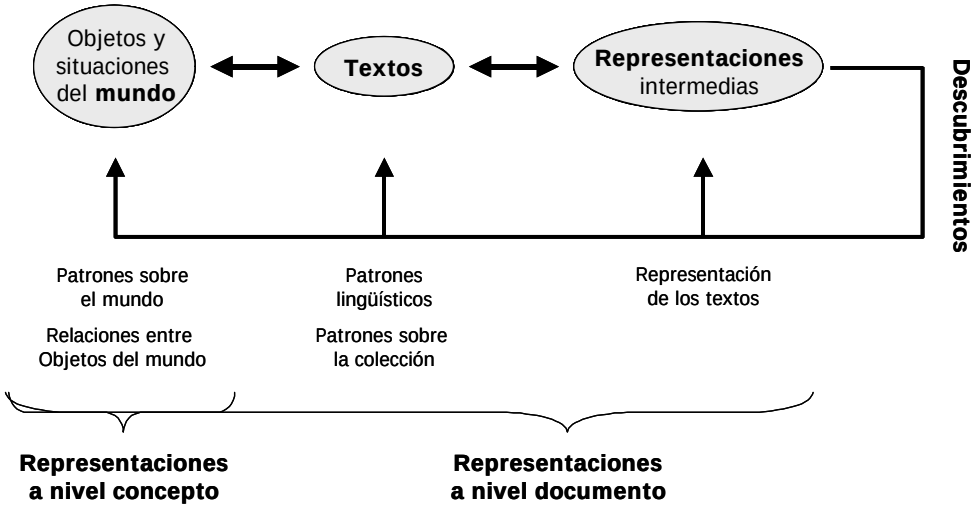


Figura 2.10 Tipos de descubrimientos de la minería de texto

En la figura 2.10 se considera que los textos son una descripción de situaciones y objetos del mundo, y que las representaciones intermedias de dichos textos –obtenidas en la etapa de preprocesamiento– son una descripción estructurada del contenido de estos últimos. Con base en esta consideración, los descubrimientos de minería de texto se pueden clasificar en los siguientes tres enfoques: descubrimientos a nivel representación, descubrimientos a nivel texto, y descubrimientos a nivel mundo.

Descubrimientos a nivel representación

Los métodos de este enfoque intentan construir o "descubrir" una representación estructurada o semiestructurada de los textos. Los más comunes se encargan de la *clasificación*, la *categorización* y el *indexamiento* de los textos (Weiss and Indurkhya, 1998; Gelfand *et al.*, 1998; Apte *et al*., 1998; Cohen and Hirsh, 1998; Perrin and Petry, 1998; Guzmán, 1998; Martínez, 1998; Weiss *et al*., 1999; Gelbukh *et al.*, 1999; Zelikovitz and Hirsh, 2000; Clifton and Cooley, 1999).

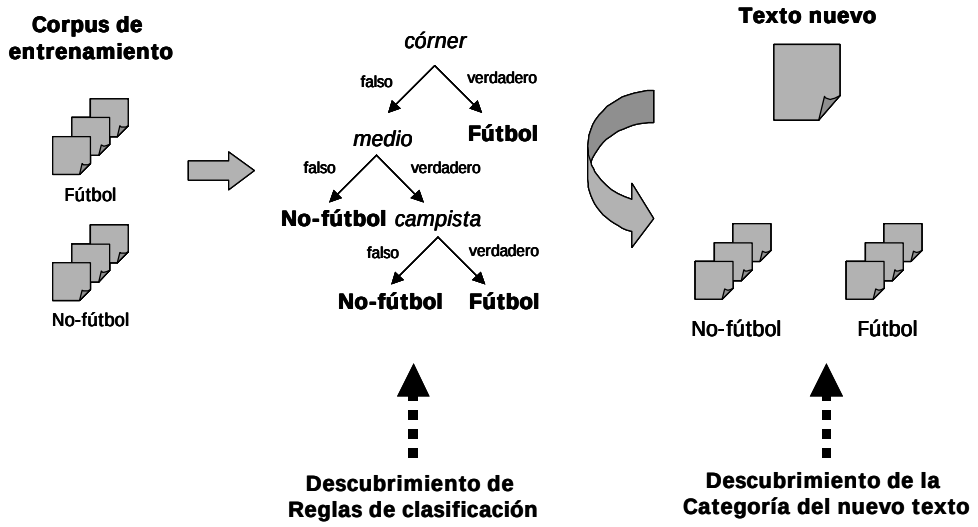


Figura 2.11 Sistema de clasificación de textos

Por ejemplo, en la figura 2.11 se ilustra un sistema de clasificación de textos. Estos sistemas descubren, a partir de un conjunto de textos conocidos, las características necesarias para clasificar un texto cualesquiera en una categoría preestablecida.

Descubrimientos a nivel texto

Los métodos de este enfoque son de dos tipos: métodos que descubren patrones de lenguaje a partir de una colección de textos, y métodos que descubren la organización "oculta" de una colección de textos.[3]

- **Identificación de patrones de lenguaje**

Los métodos de esta categoría se distinguen por dos cosas:

1. Por considerar todas las palabras de los textos y además mantener su orden relativo, es decir, usar representaciones de texto completo (*full-text*, en inglés).

[3] Las técnicas de agrupamiento también pueden hacerse sobre representaciones a nivel concepto. En tal situación los descubrimientos son a nivel mundo (Feldman *et al.*, 1998a; Feldman *et al.*, 1998c; Feldman *et al.*, 1999).

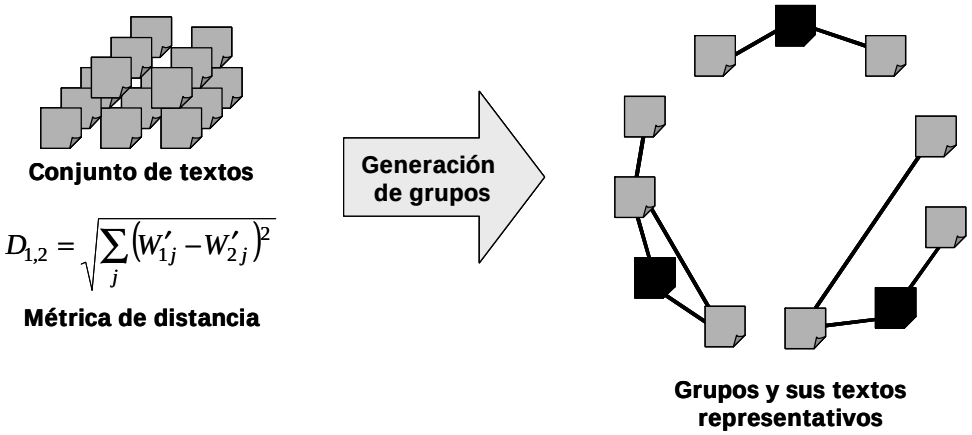


Figura 2.12 Un sistema tradicional de agrupamiento de textos

2. Por intentar aplicar directamente la mayor cantidad de técnicas provenientes de la minería de datos.

Básicamente, estos métodos detectan *secuencias frecuentes de palabras*, y en ocasiones también construyen, con base en estas secuencias, un conjunto de *reglas asociativas* que expresan combinaciones de palabras de uso común (Ahonen *et al*., 1997a; Ahonen *et al*., 1997b; Ahonen-Myka, 1999a; Ahonen-Myka, 1999b; Ahonen-Myka *et al*., 1999; Rajman and Besançon, 1997; Rajman and Besançon, 1998; Fujino *et al.*, 2000).

- **Agrupamiento de textos**

El agrupamiento de textos es una tarea ampliamente estudiada (Agrawal, 1999; Alexandrov *et al.*, 2000; Merlk, 1997; Lagus *et al.*, 1999; Larsen and Aone, 1999; Rauber and Merkl,1999). En el contexto de la minería de texto, el agrupamiento de textos tiene las siguientes características:

- Utiliza diversos tipos de métodos, desde tradicionales basados en una medida euclidiana de la distancia entre los textos, hasta sofisticados basados en redes neuronales de tipo *mapas auto organizantes*.

- Enfatiza la *visualización* e *interpretación* de los resultados. Por ejemplo, algunos métodos emplean interfaces gráficas para analizar los agrupamientos, otros determinan una etiqueta descriptiva del contenido de cada grupo, y otros mas determinan el documento representativo de cada clase (ver la figura 2.12).

Adicionalmente, el agrupamiento de los textos se usa en el *análisis exploratorio* de las colecciones de textos (Hearst, 1999), en la generación de *resúmenes multidocumento* (Larsen and Aone, 1999), y en otras tareas de descubrimiento tales como la detección de asociaciones y desviaciones (Landau *et al.*, 1998).

Descubrimientos a nivel mundo

Este enfoque considera distintas tareas, entre ellas: el descubrimiento de asociaciones, la detección de desviaciones y el análisis de tendencias. En general, los métodos de este enfoque comparten las siguientes características:

1. Emplean representaciones de los textos a *nivel concepto*, así como representaciones a *nivel documento*.
2. Usan *conocimientos de dominio*, generalmente expresados en jerarquías de conceptos o conjuntos de predicados.
3. Permiten que el *usuario guíe el proceso* de descubrimiento, especificando principalmente las regiones y los conceptos de mayor interés.

- **Descubrimiento de asociaciones**

El descubrimiento de asociaciones es la tarea más trabajada de la minería de texto (Rajman and Besançon, 1997; Feldman *et al*. 1997; Feldman *et al*. 1998b; Landau *et al*., 1998; Rajman and Besançon, 1998; Feldman and Hirsh, 1996; Lin *et al*., 1998; Montes-y-Gómez, 1999b; Singh *et al*., 1999; Nahm and Mooney, 2001b, Basu *et al.*, 2001; Montes-y-Gómez *et al.*, 2001b). Su objetivo general es descubrir reglas aso-

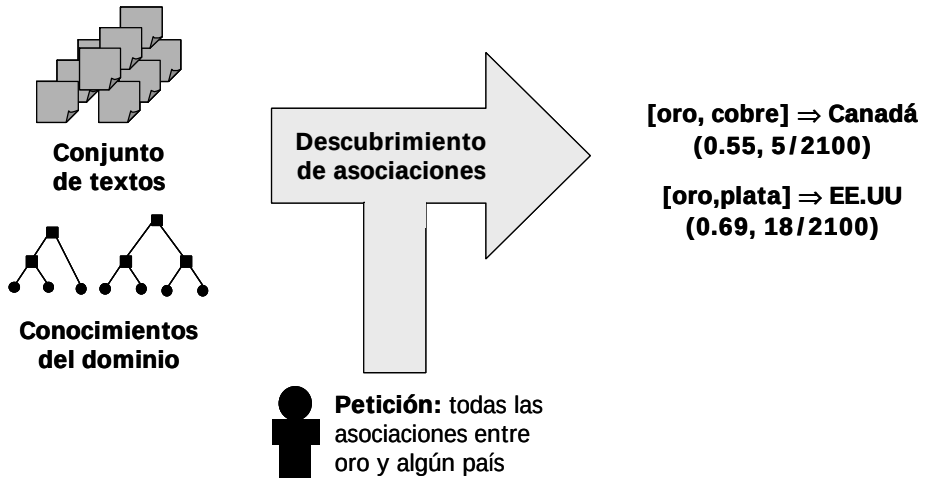


Figura 2.13 Una manera de descubrir asociaciones

ciativas de la forma $A \Rightarrow B\,(confianza / soporte)$ entre los conceptos o temas de una colección de textos.[4]

Al igual que en la minería de datos, el descubrimiento de asociaciones en una colección de textos consiste de dos etapas. En la primera etapa se generan, aplicando métodos incrementales de análisis, los conjuntos de conceptos o temas frecuentes. En la segunda etapa se construyen –infieren estadísticamente– a partir de dichos conjuntos algunas reglas asociativas.

Algunas características importantes de los métodos de descubrimiento de asociaciones en textos son las siguientes:

- Descubren asociaciones *no-exactas*, es decir, asociaciones generalizadas o asociaciones de la forma $similar(A) \Rightarrow B\,(confianza / soporte)$.
- Usan *conocimientos léxicos* para evaluar la importancia o grado de interés de las reglas asociativas.

[4] Informalmente, una regla asociativa $A \Rightarrow B$ (*confianza*/*soporte*) significa que un porcentaje de los textos de la colección (indicado por el *soporte*) menciona ambos conjuntos de conceptos ($A \cup B$); además de que una porción de los textos que menciona el conjunto de conceptos A (señalada por la *confianza*), también menciona el conjunto de conceptos B.

- Consideran tanto *elementos estructurados* (por ejemplo: autor, fecha, etc.), como *elementos no estructurados* de los textos. Estos últimos generalmente se representan por medio de un conjunto de palabras clave o tablas de datos.
- Detectan asociaciones *correlativas temporales* entre los temas de una colección.

En la figura 2.13 se muestra la manera propuesta por Feldman (Feldman and Hirsh, 1996; Feldman *et al*. 1997; Feldman *et al*. 1998b) para descubrir asociaciones en una colección de textos. Bajo este enfoque se usan representaciones a nivel documento, se considera conocimiento de dominio para hacer generalizaciones, y también una petición del usuario para activar el proceso de descubrimiento. Esta última característica permite restringir grandemente el espacio de búsqueda, y también limitar considerablemente el número de asociaciones descubiertas.

Además de ser por si mismas un tipo de patrones interesante, las reglas asociativas se usan en otras tareas. Por ejemplo se usan en la navegación de colecciones de textos (Feldman *et al*., 1997), en la clasificación de textos (Lin *et al*., 1998), y en la extracción de información (Nahm and Mooney, 2001a; Nahm and Mooney, 2001b).

- **Detección de desviaciones**

La aplicación directa de los métodos de detección de desviaciones provenientes de la minería de datos en el análisis de textos permite identificar de una forma relativamente fácil los textos raros (con una temática distinta) dentro de una colección. Este enfoque de análisis requiere de representaciones a nivel documento, y genera descubrimientos a nivel texto.

Otros métodos, propios de la minería de texto, se enfocan en la detección de los *conceptos raros* en un conjunto de textos. Algunas aplicaciones de este tipo de métodos son:

- El descubrimiento de los conceptos –temas de discusión– que presentan un comportamiento diferente a otros conceptos similares en una colección de textos (Feldman and Dagan, 1995).

- La detección de los nuevos eventos –temas de discusión– en una colección de textos que crece continuamente (Allan *et al.*, 1998)

- **Análisis de tendencias**

En términos generales, el análisis de tendencias se encarga del *análisis evolutivo* de las colecciones de textos. Entre sus métodos destacan los siguientes dos enfoques:

1. La identificación de los temas de discusión de una colección de textos que presentan un comportamiento preestablecido (Lent *et al.*,1997).
2. La comparación de la distribución temática de una colección de textos en dos tiempos diferentes (Feldman and Dagan, 1995; Montes-y-Gómez *et al.*, 1999a; Feldman *et al.*, 1998c).

Algunos de estos métodos permiten descubrir tendencias de cambio y también de estabilidad. Esto último es útil para el análisis de dominios con naturaleza cambiante, por ejemplo noticias.

2.2.2 Tendencias de investigación

La minería de texto es una nueva área de investigación del procesamiento de textos. Sus métodos, objetivos, tareas y fronteras aún no se definen completamente. Así pues, algunos de sus principales retos son:

- Establecer las *fronteras* y la manera de *importar técnicas y resultados* entre la minería de texto y otras áreas del procesamiento de textos, como por ejemplo: la extracción de información, la recuperación de información y el procesamiento estadístico de textos (Hearst 1999; Kodratoff, 1999; Feldman *et al.*, 1998a; Nahm and Mooney, 2001a).
- Aumentar la *flexibilidad* de los sistemas de minería de texto, básicamente integrando al usuario en el proceso de descubrimiento (Feldman and Hirsh, 1996; Hearst, 1999), y construyendo diferentes esquemas de análisis a partir de unir varios componentes básicos (Landau *et al.*, 1998).

- Utilizar *representaciones más completas* del contenido de los textos, que integren información estructural y contextual de su contenido, con el objetivo de aumentar la expresividad y la diversidad de los patrones descubiertos (Hearst, 1999; Tan, 1999).
- Construir métodos de preprocesamiento y descubrimiento para realizar minería de texto *multilingüe* (Tan, 1999).
- Definir algunos *métodos de postprocesamiento* encargados de validar los descubrimientos e integrar estos con otros sistemas de información (Fayyad *et al.*, 1996a).

Capítulo 3

Grafos Conceptuales

En este trabajo de investigación se propone usar los grafos conceptuales para representar el contenido de los textos. Este capítulo se enfoca en la descripción de los elementos principales de la teoría de grafos conceptuales.

Básicamente, en este capítulo presentamos la terminología elemental de los grafos conceptuales, y describimos la operación de generalización; operación fundamental para el método de minería de texto propuesto en los siguientes capítulos.

Además discutimos algunas propiedades útiles de los grafos conceptuales para la representación de lenguaje natural, y describimos el proceso tradicional para transformar un texto en grafo conceptual.

Grafos Conceptuales*

3.1 Terminología básica

3.1.1 Grafo conceptual

Un grafo conceptual es un grafo *bipartito*. Esto significa que tiene dos tipos de nodos: conceptos y relaciones conceptuales, y cada arco une solamente a un concepto con una relación conceptual.

La figura 3.1 muestra un grafo conceptual sencillo. Este grafo representa la frase "El gato Felix está sobre el sillón negro". En él se observan tres conceptos: gato Félix, sillón y negro, y dos relaciones conceptuales: sobre y atributo.

3.1.2 Concepto

Los conceptos representan entidades, acciones y atributos, y tienen un tipo conceptual y un referente. El tipo conceptual indica la clase de elemento representado por el concepto, mientras que el referente indica el elemento específico (instancia de la clase) referido por éste. Por ejemplo, el concepto [gato:Félix] de la figura 3.1 tiene el tipo gato y el referente Félix.

Figura 3.1 Un grafo conceptual sencillo

* Esta introducción a la teoría de grafos conceptuales se basa en las siguientes referencias (Sowa, 1984, 1999); más información puede obtenerse en (Way, 1992; Nagle *et al*., 1992; Sowa, 1992; Pfeiffer *et al*., 1993; Mineau *et al*., 1993; Tepfenhart *et al*., 1994; Ellis *et al*., 1995; Chein, 1996; Eklund *et al*., 1996, Lukose *et al*., 1997; Mugnier *et al*.,1998; Tepfenhart and Cyre, 1999; Ganter *et al*., 2000; Stumme, 2000).

Tipos conceptuales

Los tipos conceptuales se organizan en una jerarquía de tipos (ver por ejemplo la figura 3.2). Esta jerarquía es un ordenamiento parcialmente definido sobre el conjunto de tipos determinado por el símbolo ≤.

Entonces, dada una jerarquía de esta naturaleza, y considerando que *s*, *t* y *u* representan tres tipos conceptuales, lo siguiente puede establecerse:

- Si $s \leq t$, entonces *s* es un *subtipo* de *t*; y *t* es un *supertipo* de *s*.
- Si $s \leq t$ y $s \neq t$, entonces *s* es un *subtipo propio* de *t*, expresado como $s < t$; y *t* es un *supertipo propio* de *s*, expresado como $t > s$.
- Si *s* es un subtipo de *t* y a la vez un subtipo de *u* ($s \leq t$ y $s \leq u$), entonces *s* es un *subtipo común* de *t* y *u*.
- Si *s* es un supertipo de *t* y a la vez un supertipo de *u* ($t \leq s$ y $u \leq s$), entonces *s* es un *supertipo común* de *t* y *u*.

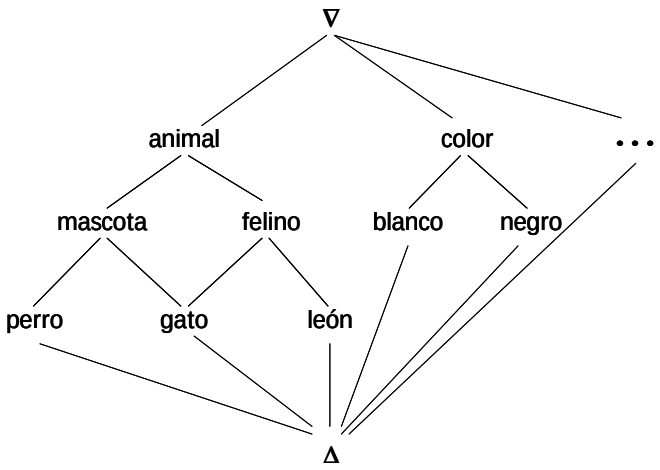


Figura 3.2 Una pequeña jerarquía de tipos

Referentes

Los referentes son de dos clases: genéricos e individuales. Los referentes genéricos se refieren a conceptos no especificados. Por ejemplo, el concepto [sillón] de la figura 3.1 significa un sillón.

Los referentes individuales funcionan como sustitutos de elementos específicos del mundo real. Por ejemplo, el concepto [gato:Félix] de la figura 3.1 es un sustituto del *gato Félix* –que existe en algún lugar.

Algunas notaciones estándares empleadas en los referentes de los conceptos se muestran en la figura 3.3.

Notación	***Significado***
[Gato]	Un gato
[Gato: *x]	un gato x
[Gato: #]	el gato
[Gato: Félix]	un gato llamado Félix (o simplemente Félix)
[Periodo: @5min]	un periodo de 5 minutos (@ indica medida)
[Gato: {*}]	unos gatos
[Gato: {*}@3]	tres gatos
[Gato: {Félix, Garfield}]	unos gatos llamados Félix y Garfield (o simplemente Félix y Garfield)

Figura 3.3. Notación de los conceptos

3.1.3 Relación conceptual

Las relaciones conceptuales señalan la manera en que los conceptos se interrelacionan. Ellas tienen un *tipo relacional*[1] y una *valencia*. El tipo relacional indica el rol "semántico" que realizan los conceptos adyacentes (conectados) a la relación, y la valencia indica el número de éstos.

[1] Igual que los tipos conceptuales, los tipos relacionales también forman un conjunto parcialmente ordenado determinado por la relación de subtipo ≤.

Algunas propiedades de las relaciones conceptuales son:

- El número de arcos que pertenecen a una relación conceptual es igual a su valencia. Una relación conceptual de valencia *n* es llamada *n*-aria, y sus arcos son numerados de 1 a *n*.
- Para cada relación conceptual *n*-aria existe una secuencia de *n* tipos conceptuales denominada la *firma* de la relación. Esta firma restringe el tipo de conceptos que pueden conectarse a cada una de los arcos pertenecientes a la relación conceptual.
- Todas las relaciones conceptuales del mismo tipo relacional tienen la misma valencia y la misma firma.

Por ejemplo, en la figura 3.1, la relación conceptual (sbr) indica que el gato Félix es quien está sobre el sillón. Esta relación es una relación binaria con firma (entidad, entidad). Por su parte, la relación (attr) indica que el sillón tiene como atributo el color negro. Esta relación también es binaria, pero su firma es (entidad, atributo).

3.2 Generalización de grafos conceptuales

Todas las operaciones de los grafos conceptuales se basan en alguna combinación de las seis *reglas canónicas de formación* (núcleo de la teoría de grafos conceptuales). Cada una de estas reglas realiza una operación básica sobre los grafos conceptuales. Por ejemplo, algunas de estas reglas los hacen más específicos, otras los generalizan, y otras únicamente cambian su forma pero los mantienen lógicamente equivalentes.

El método de minería de texto propuesto en los siguientes capítulos se fundamenta en la detección de los elementos comunes de un conjunto de grafos conceptuales, es decir, en la generalización de los grafos. Por ello, en esta sección sólo se analizan las reglas canónicas de generalización. Las demás reglas se definen y describen en (Sowa,1894, 1999).

Las reglas de generalización son dos: desrestringir y separar. La regla de desrestringir generaliza el tipo o el referente de un concepto, mientras que la regla de sepa-

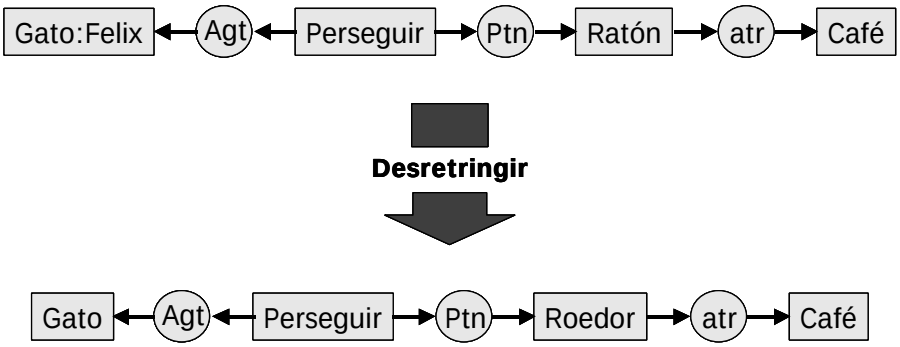


Figura 3.4 Aplicación de la regla *desrestringir*

rar divide el grafo original en dos partes tomando como base alguno de sus nodos concepto; siendo cada una de las partes resultantes una generalización del grafo original.

Desrestringir: Sea *c* un concepto del grafo *u*. Entonces el grafo *v* puede ser derivado del grafo *u* generalizando el concepto *c* tanto por tipo como por referente. La generalización por tipo reemplaza el tipo de *c* por alguno de sus supertipos, y la generalización por referente reemplaza el referente individual de *c* por un referente genérico.

La figura 3.4 ilustra la regla de desrestringir. En ella, el concepto [Gato:Félix] se generalizó a [Gato], y el concepto [Ratón] a [Roedor].

Separar: Sea *c* un concepto del grafo *u*. Entonces el grafo *v* puede ser derivado del grafo *u* haciendo una copia *d* de *c* (es decir, duplicando el concepto *c*), separando uno o varios de los arcos de las relaciones conceptuales conectadas a *c*, y conectándolos a *d*.

La figura 3.5 ejemplifica la regla de separar. En ella, el grafo conceptual original se separó en dos partes, tomando como base el concepto [Ratón]. En este caso, cada una de las partes resultantes es un grafo conceptual más general que el grafo original.

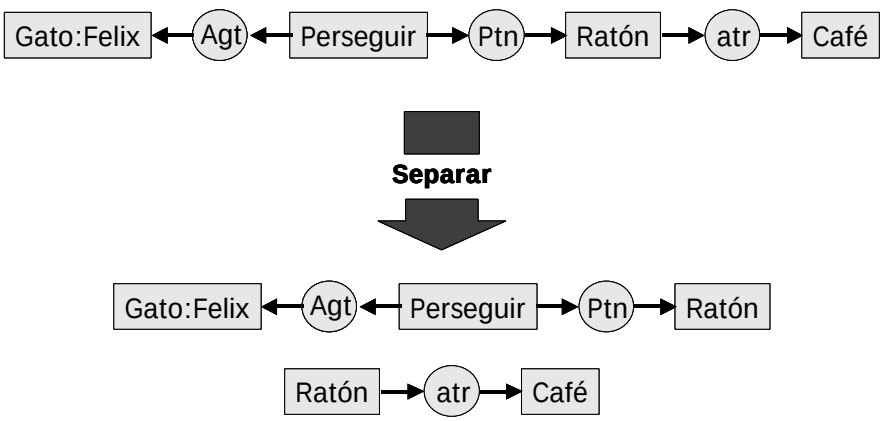


Figura 3.5 Aplicación de la regla *separar*

Ahora bien, si el grafo conceptual *v* es derivado del grafo conceptual *u* aplicando una secuencia de estas reglas, entonces *v* es una *generalización* de *u*. Esto se denota como $u \leq v$.

La operación de generalización define un ordenamiento parcial de los grafos conceptuales conocido como *jerarquía de generalización*. Entonces si *u*, *v* y *w* son grafos conceptuales de esta jerarquía, las siguientes propiedades siempre son verdaderas:

- Reflexividad: $u \leq u$.
- Transitividad: si $u \leq v$ y $v \leq w$, entonces $u \leq w$.
- Antisimetría: si $u \leq v$ y $v \leq u$, entonces $u = v$.
- Subgrafo: Si *v* es un subgrafo de *u*, entonces $u \leq v$.

Además si *v* es una generalización de *u* ($u \leq v$), entonces debe de existir un subgrafo u' inmerso en *u* que represente el grafo *v*. Este subgrafo u' es llamado *proyección* de *v* en *u*. La figura 3.6 ilustra esta proyección.

Formalmente, para dos grafos conceptuales cualesquiera *u* y *v*, siendo $u \leq v$, debe de existir un "mapeo" p: $v \rightarrow u$, donde pv es un subgrafo de *u* llamado proyección de *v* en *u*. Algunas propiedades de la proyección son:

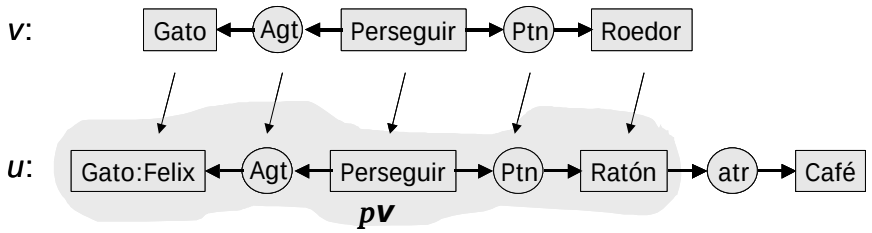


Figura 3.6 Proyección de *v* en *u*

- Para cada concepto *c* de *v*, *pc* es un concepto en *pv*, para el cual $type(pc) \leq type(c)$; y si *c* es un concepto individual, entonces también $referent(pc) = referent(c)$.[2]
- Para cada relación conceptual *r* de *v*, *pr* es una relación conceptual en *pv*, para la cual *type*(*pr*) = *type*(*r*).[3] Esto implica que si el *i*-esimo arco de *r* está conectado al concepto *c*, entonces el *i*-esimo arco de *pr* debe de estar conectado a *pc* en *pv*.

La proyección *p* no es necesariamente uno-a-uno, esto significa que dos conceptos o dos relaciones conceptuales diferentes pueden tener las mismas proyecciones (por ejemplo, los conceptos $x_1, x_2 \in v$: $x_1 \neq x_2$ pueden tener proyecciones px_1 y px_2 en *u*, tal que $px_1 = px_2$). Además, la proyección *p* tampoco es necesariamente única, es decir, un grafo *v* puede tener dos proyecciones diferentes en *u*, $p'v$ y *pv*, donde $p'v \neq pv$.

Finalmente, si u_1, u_2 y *v* son grafos conceptuales, y $u_1 \leq v$ y $u_2 \leq v$, entonces *v* es una *generalización común* de u_1 y u_2. El grafo conceptual *v* es la *máxima generalización común* de u_1 y u_2, si y sólo si, no existe otra generalización común v' de u_1 y u_2 ($u_1 \leq v'$ y $u_2 \leq v'$), tal que $v' \leq v$.

---

[2] La función *type*(*c*) proyecta el concepto *c* en la jerarquía de tipos conceptuales, y la función *referent*(*c*) indica el referente del concepto *c*.

[3] La función *type*(*r*) indica el tipo relacional de la relación conceptual *r*; cuando existe una jerarquía de tipos relacionales, esta función proyecta la relación conceptual *r* en dicha jerarquía.

3.3 Lenguaje natural y grafos conceptuales

En este trabajo se plantea el uso de los grafos conceptuales como representación del contenido de los textos. Esta decisión se debe entre otras cosas al potencial de los grafos conceptuales (en relación con otras representaciones de conocimiento) para representar en forma simple y directa algunos detalles finos del lenguaje.

Por ejemplo, los grafos conceptuales permiten representar dependencias contextuales que no pueden representarse con los predicados lógicos, y también permiten representar referentes, contextos y modales que son difícilmente expresados con la más usada de las representaciones de conocimiento, la red semántica *KL-one* (Sowa, 1991; Biébow and Chaty, 1993).

A continuación se ilustran varias características de los grafos conceptuales, pero principalmente se ejemplifica la representación de algunos elementos de las oraciones en lenguaje natural con grafos conceptuales. También, en la parte final de esta sección, se describe el proceso usual para transformar los textos en grafos conceptuales.

3.3.1 Correspondencias entre lenguaje y grafos conceptuales

El formalismo de los grafos conceptuales tiene un mínimo de características predefinidas –lo que lo hace suficientemente general–, pero tiene ciertas reglas que guían la transformación de lenguaje natural a grafos conceptuales.

La regla de transformación básica es que las palabras contenido, como sustantivos, verbos, adjetivos y adverbios, corresponden a nodos concepto, mientras que las palabras funcionales, como preposiciones, conjunciones y verbos auxiliares, corresponden a nodos relación.

A continuación se muestra la manera en que varios elementos de las oraciones en lenguaje natural, como por ejemplo, nombres propios, sustantivos plurales, moda-

les, tiempos, conjunciones e información sintáctica, se representan con grafos conceptuales.[4]

- Los sustantivos, verbos, adjetivos y adverbios corresponden a los tipos de los nodos concepto:

  Presidente ⇒ [Presidente]

  criticaron ⇒ [Criticar]

  electo ⇒ [Electo]

- Los nombres propios se representan como el referente de los nodos concepto. En este caso, el tipo indica la clase del objeto referenciado:

  Fox ⇒ [Presidente: Fox]

  Bellas Artes ⇒ [Edificio: Bellas Artes]

- Las referencias definidas contextualmente corresponden a un referente con el símbolo #:

  El presidente ⇒ [Presidente: #]

  - El símbolo # seguido de una variable indica una co-referencia:

    [Persona: Fox *x] [Presidente: #*x]

- Los sustantivos plurales se representan con el referente plural {*}, seguido de un indicador opcional de cantidad:

  nueve presidentes ⇒ [Presidente: {*}@9]

  - Plurales específicos –no genéricos– y parcialmente especificados se representan de la siguiente manera:

    Fox y Bush ⇒ [Presidente: {Fox, Bush}]

    Fox, Bush y otros presidentes ⇒ [Presidente: {Fox, Bush, *}]

---

[4] Por razones de simplicidad, en los ejemplos se utiliza el formato lineal de los grafos conceptuales.

- Los prefijos Coll{*} y Dist{*} se usan para indicar una interpretación colectiva y distributiva de los sustantivos plurales:

  Fox y Bush platican ⇒ [Presidente: Coll{Fox, Bush}]

- Los auxiliares como poder y deber corresponden a relaciones conceptuales como PSBL (de posibilidad) y OBLG (de obligación). Estas relaciones afectan el *contexto* que encierra el grafo de la oración o cláusula.

  El presidente Fox posiblemente vaya ⇒

  (PSBL)→[[Presidente: Fox]←(AGNT)←[ir]]

- Los tiempos de los verbos corresponden a relaciones conceptuales como (PASD) o (FUTR). Estas relaciones también afectan el *contexto* que encierra el grafo de la oración o cláusula.

  El presidente Fox fue ⇒ (PASD)→[[Presidente: Fox]←(AGNT)←[ir]]

  - Los contextos pueden tener varios niveles de anidamiento, por ejemplo la frase *el presidente Fox no debió ir* se representa como:

    (PASD)→[(NO)→[(OBLG)→[[Presidente: Fox]←(AGNT)←[ir]]]]

- El verbo tener, cuando es usado como verbo principal, puede corresponder a distintas relaciones conceptuales como PART (de parte) y POSS (de posesión):

  El presidente Fox tiene unas botas ⇒

  [Presidente: Fox]→(POSS)→[Bota: {*}]

- Las terminaciones en los lenguajes con inflexiones, y el orden de las palabras en lenguajes sin éstas corresponden a roles temáticos como AGNT (agente), PTNT (paciente), INST (instrumento), RCPT (recipiente), etc.

  Manuel le envió una petición a Fox por e-mail ⇒

  (PASD)→[[enviar] –

  (AGNT)→[Ciudadano: Manuel]

  (PTNT)→[Petición]

(RCPT)→[Presidente: Fox]

(INST)→ [e-mail]]

- La información sintáctica de las oraciones corresponde a los comentarios en los nodos concepto. Esta información es útil para tareas como traducción automática y generación de lenguaje.

Fox está visitando China ⇒

[Presidente:Fox;s]←(AGNT;subj)←[Visitar;grd]→(DEST;obj)→[País:China;n]

3.3.2 Trasformación texto ⇒ grafo conceptual

Los grafos conceptuales son un formalismo basado en lógica, y orientado a la representación del lenguaje natural. En la sección anterior se demostró su potencial para representar el lenguaje natural. En esta sección se presentan algunos métodos relacionados con la transformación automática de un texto en grafos conceptual (Sowa and Way, 1986; Fargues *et al.*, 1986; Sowa, 1988; Velardi *et al.*, 1988). Estos métodos se fundamentan en el operador de máxima unión –unificación de grafos–, y siguen una estrategia basada en la sintaxis.

En general, todos estos métodos hacen lo mismo: recorren el árbol sintáctico de la oración analizada en forma ascendente (*bottom-up*, en inglés), y unen –con base en el operador de máxima unión– los grafos canónicos asociados a cada uno de los nodos de dicho árbol. El proceso general para el análisis semántico conducido por grafos conceptuales se ilustra en la figura 3.7.

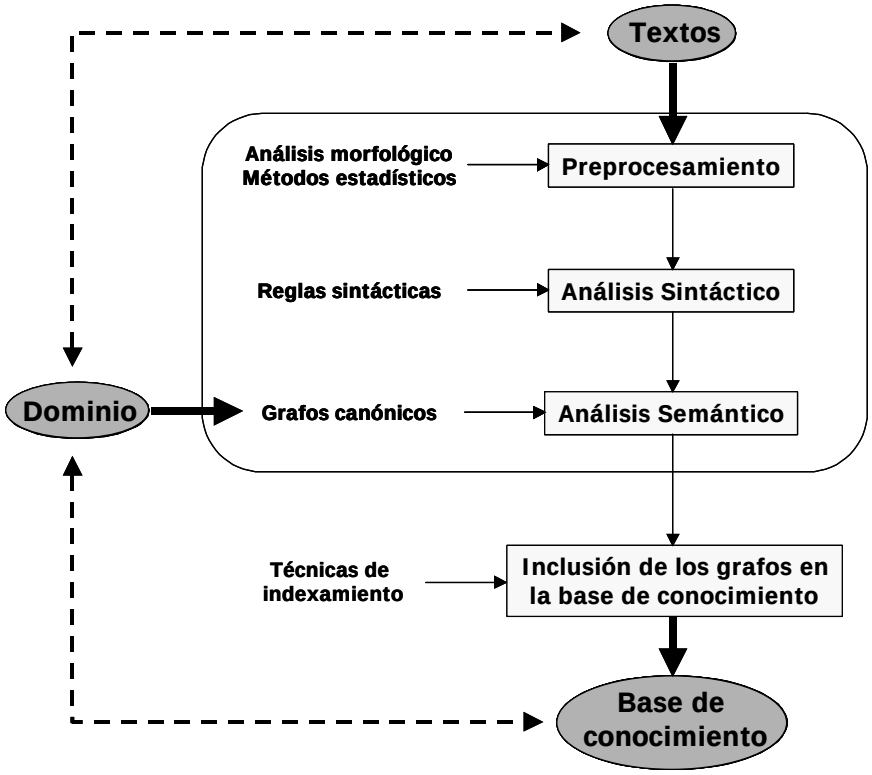


Figura 3.7 Transformación textos ⇒ grafo conceptuales

Otros métodos, menos formales, consideran solamente la transformación de textos de *dominios específicos* en grafos conceptuales, aunque también siguen una estrategia basada en la sintaxis. Algunos ejemplos son:

- Partes de *artículos científicos* a grafos conceptuales (Myaeng and Khoo, 1994; Tapia-Melchor and López-López, 1998; Montes-y-Gómez *et al*., 1999e).
- Partes de *expedientes médicos* a grafos conceptuales (Baud *et al.*, 1992; Rassinoux *et al.*, 1994).
- Partes de *casos legales* a grafos conceptuales (Boucier and Rajman, 1994).
- Partes de *manuales de referencia* a grafos conceptuales (Petermann, 1996).
- Relaciones semánticas especificas, como por ejemplo, *relaciones causales*, a grafos conceptuales (Khoo, 1995).

El proceso de construcción de los grafos conceptuales que se usan en los experimentos de este trabajo se describe con mayor detalle en el apéndice A.

Capítulo 4

Comparación de Grafos Conceptuales

Una de las operaciones básicas para el análisis "inteligente" de un conjunto de grafos conceptuales es su comparación. Así pues, en este capítulo proponemos un método flexible para comparar dos grafos conceptuales, el cual consiste de dos etapas principales: casamiento de los grafos y medición de la semejanza. En la primera etapa se construye una descripción cualitativa de la semejanza entre los grafos representada por un conjunto máximo de generalizaciones comunes compatibles. Después, en la segunda etapa se mide la semejanza de los grafos según el coeficiente de Dice, y de acuerdo con algunas características especiales de los grafos conceptuales.

A lo largo del capítulo enfatizamos la flexibilidad del método de comparación de los grafos conceptuales. Básicamente destacamos el uso de conocimiento del dominio y la participación directa del usuario.

Comparación de Grafos Conceptuales

4.1 Ámbito general del problema

Una de las operaciones básicas para el análisis automático de cualquier conjunto de objetos es, sin lugar a dudas, su *comparación*. Así pues, el desarrollo de un método de minería de texto que utilice los grafos conceptuales como la representación intermedia del contenido de los textos requiere de un método adecuado para la comparación de dos grafos conceptuales cualesquiera.

La comparación de grafos conceptuales es una tarea compleja, relacionada con problemas definidos como *NP-completos*, tales como: el apareamiento y la proyección (Myaeng and López-López, 1992; Mugnier and Chein, 1992; Mugnier, 1995). A pesar de ello existen varios métodos para la comparación de dos grafos conceptuales. La mayoría de éstos provienen de la recuperación de información y permiten comparar, en un tiempo razonable, grafos pequeños (con menos de 30 conceptos). Entre estos métodos sobresalen los siguientes dos grupos:

1. Métodos que limitan la comparación de los grafos conceptuales al problema de determinar si uno de ellos, digamos, el grafo "petición", está completamente contenido en el otro, digamos el grafo "documento" (Huibers *et al*., 1996; Ellis and Lehmann, 1994).
2. Métodos que detectan todos los elementos, conceptos y relaciones comunes de los dos grafos (Myaeng and López-López, 1992; Myaeng, 1992; Genest and Chein, 1997)..

Típicamente, los métodos del primer grupo no obtienen ninguna descripción ni ninguna medida de la semejanza entre los dos grafos conceptuales, mientras que los métodos del segundo grupo si obtienen una medida de la semejanza entre los dos

grafos conceptuales, pero describen dicha semejanza como el conjunto de todas sus generalizaciones comunes, permitiendo así información duplicada.

En general, todos estos métodos no son apropiados para la minería de texto por dos razones principales. En primer lugar, porque no construyen una descripción de la semejanza, o construyen una descripción imprecisa de ésta –con información duplicada–. En segundo lugar, porque no son suficientemente flexibles para adaptarse a los distintos escenarios de aplicación e intereses de los usuarios.

El método para la comparación de grafos conceptuales que proponemos a continuación soluciona o aminora estos problemas. Algunas de sus principales características son:

- Describe *cualitativa* y *cuantitativamente* la semejanza entre los grafos conceptuales. Esto significa que construye una descripción de la semejanza entre los grafos (expresada en forma de grafo conceptual), y además determina la medida de dicha semejanza.
- Permite visualizar la semejanza entre los grafos conceptuales desde *diferentes perspectivas*, todas ellas sin información duplicada, y también seleccionar la mejor de ellas de acuerdo con los *intereses del usuario*.
- Utiliza *conocimiento del dominio*, por ejemplo, un diccionario de sinónimos y algunas jerarquías de conceptos[1], para detectar semejanzas no exactas, pero interesantes para el usuario.

[1] Las jerarquías de conceptos no sólo describen los conceptos del dominio en cuestión, también enfatizan los intereses de descubrimiento del usuario. Esto último se logra permitiendo que el usuario establezca dichas jerarquías.

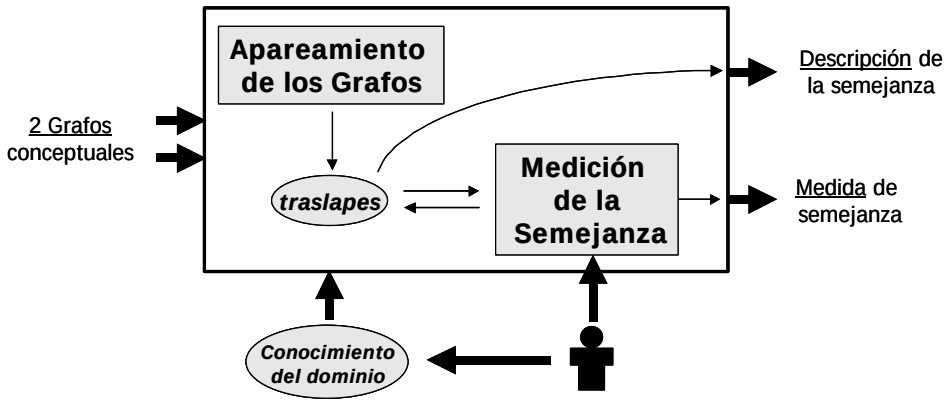


Figura 4.1 Método de comparación de grafos conceptuales

4.2 Método de comparación

El procedimiento general que proponemos para la comparación de dos grafos conceptuales se ilustra en la figura 4.1. Este procedimiento consiste de dos etapas: el apareamiento de los grafos y la medición de la semejanza.

En la primera etapa se identifican *todos* los elementos, conceptos y relaciones, comunes de ambos grafos, y se construye, a partir de estos, la o las descripciones de dicha semejanza. Estas descripciones las llamamos *traslapes*.

En la segunda etapa se calcula la medida de la semejanza de los dos grafos. Esta medida expresa la importancia relativa del traslape con respecto a los grafos conceptuales originales. Cuando se identifica más de un traslape, se calcula una medida de semejanza con respecto a cada uno. La mayor medida se considera la medida de semejanza final, y el traslape que la produce la mejor descripción de la semejanza.

En ambas etapas, la de apareamiento y la de medición, se utiliza conocimiento del dominio y se consideran los intereses del usuario. El conocimiento del dominio se expresa a través de un conjunto de *jerarquías de conceptos*. Básicamente, estas jerarquías permiten determinar semejanzas entre los conceptos de los grafos a diferentes niveles de generalización.

Por su parte, los intereses del usuario se expresan por dos medios. En primer lugar, a través de algunos *parámetros* de la medida de semejanza, por ejemplo, los que determinan la importancia relativa de las entidades, acciones y atributos. En segundo lugar, a través del conocimiento del dominio que el usuario establece libremente.

El uso de conocimiento del dominio –jerarquías de conceptos– establecido por el usuario permite enfocar la comparación de los grafos sobre algunos conceptos considerados importantes, y en consecuencia limitar la descripción de la semejanza a ciertos niveles máximos, aún interesantes, de generalización.[2] La figura 4.2 muestra una jerarquía de conceptos de este tipo.

4.2.1 Apareamiento de grafos conceptuales

Típicamente, el apareamiento de dos grafos conceptuales permite identificar *todos* sus elementos –generalizaciones– comunes. Debido a que el operador de proyección

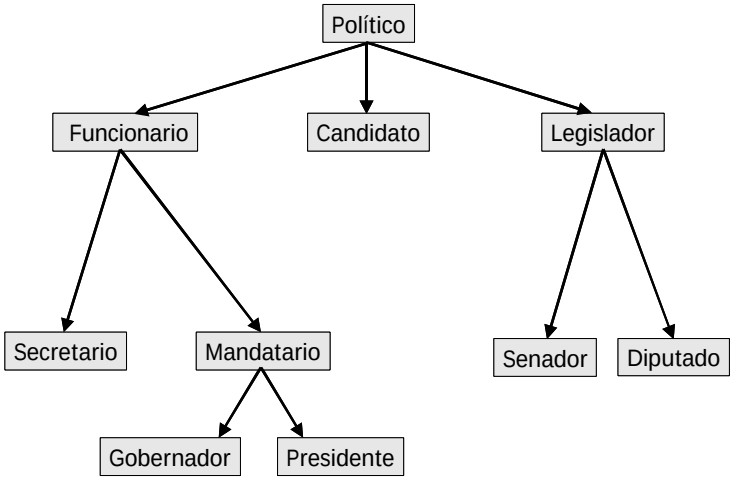


Figura 4.2 Conocimientos de dominio restringidos

[2] Este tipo de jerarquías de conceptos se emplean en la minería de datos (Srikant and Agrawal, 1995; Han and Kamber, 2001), y en la minería de texto (Feldman and Dagan, 1995). En ambos casos constituyen un medio para agregar evidencias y para enfocar el análisis sobre algunos elementos interesantes para el usuario.

p no es necesariamente uno-a-uno y tampoco único (ver sección 3.2), algunas de estas generalizaciones comunes pueden expresar información redundante o duplicada. Entonces, para lograr construir una descripción precisa de la semejanza entre dos grafos conceptuales es necesario identificar los conjuntos de generalizaciones comunes que formen una máxima generalización común *compatible*. Cada uno de estos conjuntos es lo que llamamos un *traslape*.

Un traslape lo definimos de la siguiente manera:

Definición 1. El conjunto de generalizaciones comunes $O = \{g_1, g_2, \ldots, g_n\}$ de los grafos conceptuales G_1 y G_2 es *compatible* si y solo si existe un "mapeo" $\{p_1, p_2, \ldots, p_n\}$ tal que sus correspondientes proyecciones en G_1 y G_2 no se intercepten:

$$\bigcap_{i=1}^{n} p_{G_1} g_i = \bigcap_{i=1}^{n} p_{G_2} g_i = \emptyset$$

Definición 2. El conjunto de generalizaciones comunes $O = \{g_1, g_2, \ldots, g_n\}$ de los grafos conceptuales G_1 y G_2 es *máximo* si y solo si no existe otra generalización común g de G_1 y G_2, tal que alguna de las siguientes condiciones se satisfaga:

1. $O' = \{g_1, g_2, \ldots, g_n, g\}$ es compatible.
2. $\exists i : g \leq g_i$, $g \neq g_i$, y $O = \{g_1, \ldots, g_{i-1}, g, g_{i+1}, \ldots, g_n\}$ es compatible.

Definición 3. El conjunto de generalizaciones comunes $O = \{g_1, g_2, \ldots, g_n\}$ de los grafos conceptuales G_1 y G_2 es un *traslape* si y solo si es compatible y máximo.

De acuerdo con esto, cada traslape expresa en forma *completa* y *precisa* la semejanza entre dos grafos conceptuales. Esto implica que traslapes distintos pueden indicar diferentes maneras de visualizar e interpretar dicha semejanza.

```
Procedimiento Construye_P
//Entrada: los dos grafos G1 grafos G2
1  Inicializa conjunto P = ∅
//casamiento de los nodos conceptos
2  Para cada concepto ci de G1
3    Para cada concepto cj de G2
4       Encontrar la generalización común Cij de ci y cj.
5       Si existe la generalización común Cij
6          P ← Cij
//casamiento de los nodos relación
7  Para cada relación ri de G1
8    Para cada relación rj de G2
9       Encontrar la generalización común Rij de ri y rj.
10       Si existe la generalización común Rij
11         P ← Rij
```

Figura 4.3 Algoritmo de apareamiento (*fase* 1)

4.2.1.1 Algoritmo de apareamiento

Dados dos grafos conceptuales G_1 y G_2, el objetivo del apareamiento es encontrar *todos* sus traslapes. Este apareamiento se realiza en dos fases.

En la primera fase se identifican todas las semejanzas –correspondencias– entre los dos grafos conceptuales (Myaeng and López-López, 1992; Poole and Campbell, 1995; Petermann, 1996). Este conjunto de semejanzas P expresa el producto cartesiano del conjunto de nodos concepto y del conjunto de nodos relación de ambos grafos, pero solamente considera las parejas con generalizaciones comunes no nulas (diferentes del concepto universal ⊤). El algoritmo correspondiente a esta fase se ilustra en la figura 4.3.

En la segunda fase se detectan todos los conjuntos de elementos compatibles, es decir, se construyen todos los *traslapes*. El algoritmo correspondiente a esta fase se describe en la figura 4.4. Este algoritmo es una adaptación de un algoritmo muy popular para la identificación de todos los conjuntos de elementos frecuentes en una base de datos (Agrawal *et al.*, 1994).

```
Construye los traslapes a partir de P
//Inicialmente todo concepto común es un traslape
1  Traslapes1 es conjunto de todos los conceptos de P
3  Para(k = 2; Traslapesk-1 = ∅; k++)
4    Construir Traslapesk a partir de Traslapesk-1
5    Eliminar traslapesk-1 cubiertos por Traslapesk
//Unión de todos los traslapes máximos
6  MaxTraslapes ← ⋃k Traslapesk
//Insertar las relaciones de P en los traslapes
7  Para cada relación r en P
8    Para cada traslape Oi ∈ MaxTraslapes
9      Si todos los conceptos vecinos a r existen en Oi
10       Oi ← r
```

Figura 4.4 Algoritmo de apareamiento (*fase* 2)

Inicialmente, este algoritmo considera que cada uno de los conceptos del conjunto *P* es por si solo un traslape. Después, en cada una de las subsecuentes iteraciones, los traslapes de la iteración anterior se usan como "semillas" para generar nuevos y más grandes traslapes. Al final de cada iteración se eliminan los traslapes de la iteración anterior que fueron empleados para construir los nuevos traslapes (porque estos no son máximos), y los nuevos traslapes se convierten en las "semillas" para la nueva iteración. Este proceso continúa hasta que no se encuentra ningún nuevo traslape. Al final se insertan las relaciones del conjunto *P* en los traslapes correspondientes (para que una relación pueda insertarse en un traslape todas sus conexiones deben de establecerse).

La construcción de los traslapes de tamaño *k* a partir de los traslapes de tamaño *k*-1 se realiza con la función *gen_traslape*(). Esta función toma como argumento $Traslapes_{k-1}$, el conjunto de traslapes de tamaño *k*-1 (es decir, los traslapes que incluyen *k*-1 conceptos), y devuelve como salida $Traslapes_k$, el conjunto de traslapes de tamaño *k* (los traslapes que incluyen *k* conceptos).

Cada uno de los traslapes de tamaño *k* se obtiene de la unión de dos traslapes compatibles de tamaño *k*-1. En términos generales, la función *gen_traslape*() se define de la siguiente manera.

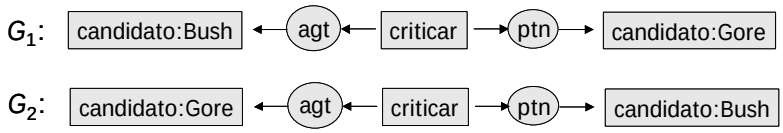


Figura 4.5 Dos grafos conceptuales

$$Traslapes'_k = \{X \cup X' \mid X, X' \in Traslapes_{k-1}, |X \cap X'| = k-2\}$$
$$Traslapes_k = \{X \in Traslapes'_k \mid X \text{ contiene } k \text{ miembros de } Traslapes_{k-1}\}$$

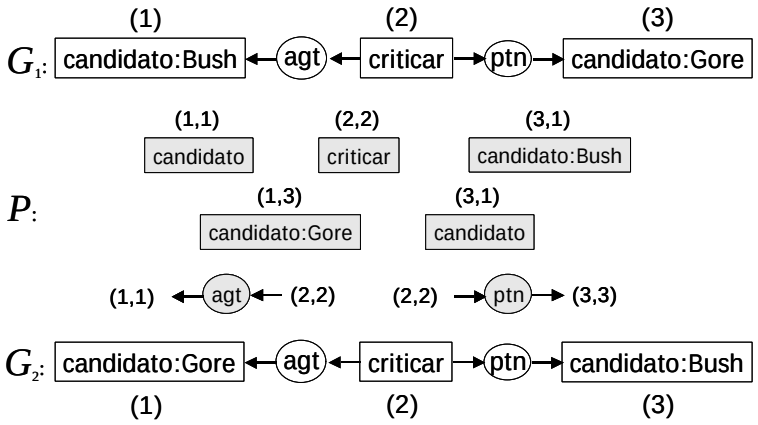
con una excepción para el caso $k = 2$, donde:

$$Traslapes_2 = \{X \cup X' \mid X, X' \in Traslapes_1, X \text{ y } X' \text{ son conceptos compatibles}\}$$

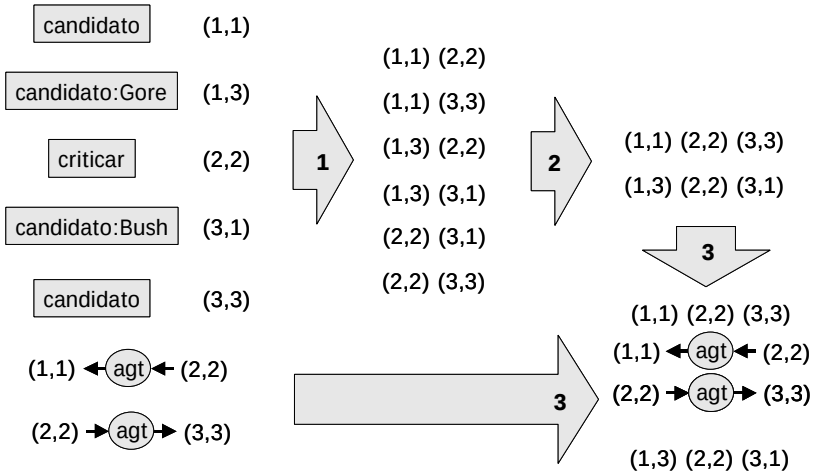
Debido a que el apareamiento y la proyección de los grafos conceptuales son problemas definidos como NP-completos (Myaeng and López-López, 1992; Mugnier and Chein, 1992; Mugnier, 1995), nuestro algoritmo es de complejidad exponencial con respecto al número de conceptos comunes de los dos grafos. Sin embargo, esto no implica ninguna limitación importante para su aplicación en la minería de texto (tal y como nosotros la pretendemos realizar), ya que los grafos que serán comparados son generalmente el resultado del análisis sintáctico superficial (shallow parsing, en inglés) de pequeñas partes descriptivas del contenido de los textos, y en consecuencia son pequeños –30 conceptos como máximo– y tienen solamente unos cuantos conceptos comunes.

Nuestro algoritmo de apareamiento es una adaptación del algoritmo APRIORI (Agrawal *et al.*, 1994), también de complejidad exponencial, pero reportado como rápido para la mayoría de los casos. Además, nuestro algoritmo realiza en forma especial el peor caso de análisis que es cuando sólo existe un traslape[3]. En tal situación se omite la segunda fase del algoritmo, y se construye el único traslape uniendo

[3] Con grafos pequeños, de menos de 10 conceptos, este caso es el más frecuente. Con grafos más grandes, este caso también ocurre aunque no es el más frecuente.



(a) Primera fase



(b) Segunda fase

Figura 4.6 Proceso de apareamiento de los grafos conceptuales

todos los elementos de *P*. Este proceso alternativo tiene una complejidad lineal con respecto al número de conceptos comunes de los dos grafos.

4.2.1.2 Ejemplo del apareamiento

Para ilustrar el recién propuesto algoritmo de apareamiento de grafos conceptuales se usarán los grafos de la figura 4.5. El primer grafo representa la frase "*Bush critica a Gore*", y el segundo grafo la frase "*Gore critica a Bush*".

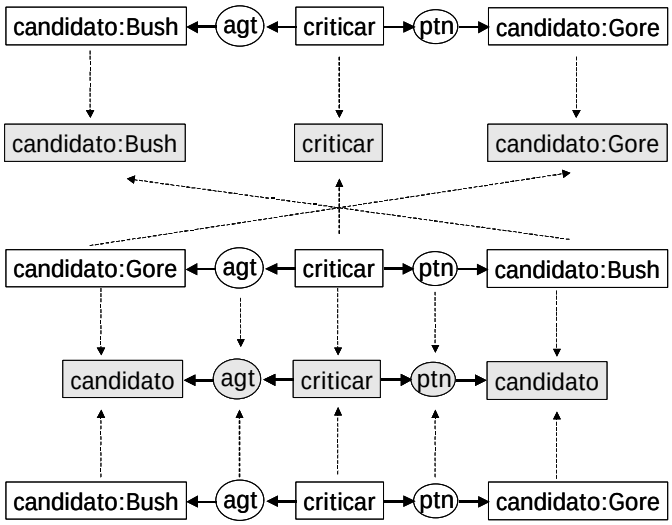


Figura 4.7 Los dos traslapes resultantes

La figura 4.6 ilustra el proceso completo de su apareamiento. Por ejemplo, la figura 4.6(a) muestra el resultado de la primera fase del apareamiento, y la figura 4.6(b) ilustra el proceso aplicado en la segunda fase.

A partir de la figura 4.6(b) se observa que la semejanza de los dos grafos puede interpretarse de dos formas diferentes, cada una de ellas asociada a un traslape distinto. Estos traslapes se muestran en la figura 4.7. El primero de ellos indica que en ambos grafos se mencionan los conceptos "*Bush*", "*Gore*" y "*criticar*", mientras que el segundo indica que en ambos grafos "*un candidato critica a otro candidato*".

La selección de uno de estos traslapes como la descripción más apropiada de la semejanza entre los dos grafos depende de los intereses específicos del usuario. Básicamente, estos intereses se modelan a través de la medida de semejanza. Así pues, el traslape que produce la mayor medida de semejanza es considerado la mejor descripción de ella.

En la siguiente sección se describe el método de cuantificación de la semejanza.

4.2.2 Medición de la semejanza

La medición de la semejanza es la segunda etapa de la comparación de los grafos conceptuales. En esta etapa se recibe como entrada los dos grafos que se comparan y el conjunto de todos sus posibles traslapes. Para cada traslape se calcula una medida de semejanza. Finalmente se entrega como resultado la mayor medida y el traslape que la produce (que es la descripción final de la semejanza).

Ahora bien, dados dos grafos conceptuales G_1 y G_2, y *uno* de sus traslapes, la medida de semejanza expresa la importancia relativa de los elementos comunes (traslape) con respecto a toda la información de los grafos originales. En general, nuestra medida tiene las siguientes características:

1. Se fundamenta en las siguientes *intuiciones básicas* (Lin, 1998):
   - La semejanza entre dos grafos conceptuales se relaciona con su traslape (elementos comunes); entre más especializado y más extenso sea éste, más semejantes son los grafos.
   - La semejanza entre dos grafos conceptuales se relaciona con sus diferencias; entre más diferencias tengan, menos semejantes son.
   - La mayor semejanza entre dos grafos conceptuales se obtiene cuando son idénticos, sin importar cuantos elementos comunes tengan.
   - La semejanza entre dos grafos conceptuales es nula cuando los grafos no tienen ningún elemento común, esto es, cuando su traslape es nulo.

2. Se basa en una medida conocida para la comparación de textos; a saber: el *coeficiente de Dice* (Rasmussen, 1992).[4]

   El coeficiente de Dice entre dos textos T_1 y T_2 se define como: $s(T_1, T_2) = 2\, t_{12} / t_1 + t_2$, donde t_i es el número de términos del texto T_i, y t_{12} es el número de términos comunes de los textos T_1 y T_2.

3. Aprovecha la estructura *bipartita* de los grafos conceptuales. La medida de semejanza se obtiene combinando dos tipos de semejanzas parciales: una *semejanza conceptual* y una *semejanza relacional*. La semejanza conceptual expresa que tan similares son las entidades, acciones y atributos mencionados en los dos grafos conceptuales; mientras que la semejanza relacional señala que tan parecidas son las interconexiones entre los conceptos comunes de ambos grafos.

4. Considera *conocimiento del dominio*. Este conocimiento se expresa en forma de un diccionario de sinónimos y algunas jerarquías de conceptos, y permite evaluar adecuadamente la contribución de las semejanzas no exactas.

5. Permite que el *usuario* establezca algunos parámetros de la medida de semejanza. Por ejemplo, la importancia relativa de las semejanzas conceptual y relacional, y la importancia relativa de las entidades, acciones y atributos. Esta característica otorga una gran flexibilidad al proceso de comparación de los grafos conceptuales.

4.2.2.1 Medida de semejanza

Dados dos grafos conceptuales G_1 y G_2, y uno de sus traslapes, denotado por O, su semejanza $0 \leq s \leq 1$ es una combinación de dos valores: una semejanza conceptual s_c y una semejanza relacional s_r.

---

[4] El coeficiente de Dice se seleccionó como la expresión base para el cálculo de la semejanza entre dos grafos conceptuales principalmente por su *simplicidad* y *normalización*, aunque también por su alta correlación con las evaluaciones manuales de la semejanza (Jiang and Conrath, 1999).

Semejanza conceptual

La semejanza conceptual $0 \leq s_c \leq 1$ depende de la cantidad de conceptos comunes de G_1 y G_2. A grandes rasgos, esta semejanza indica que tan parecidas son las entidades, acciones y atributos mencionadas en ambos grafos conceptuales.

La semejanza conceptual s_c se calcula usando una expresión análoga al coeficiente de Dice:

$$s_c(G_1, G_2) = \frac{2\sum_{c \in O} \left(weight(c) \times b\left(p_{G_1} c, p_{G_2} c\right)\right)}{\sum_{c \in G_1} weight(c) + \sum_{c \in G_2} weight(c)}$$

En esta expresión, la función $weight(c)$ indica la importancia del concepto *c* dependiendo de su tipo, y la función $b\left(p_{G_1} c, p_{G_2} c\right)$ expresa el nivel de generalización del concepto común $c \in O$ con respecto a sus proyecciones en los grafos originales $p_{G_1} c$ y $p_{G_2} c$.

La función $weight(c)$ evalúa en forma diferente los distintos tipos de conceptos. Esta función se define de la siguiente manera:

$$weight(c) = \begin{cases} w_E & \text{si } c \text{ representa una entidad} \\ w_V & \text{si } c \text{ representa una acción} \\ w_A & \text{si } c \text{ representa un atributo} \end{cases}$$

Aquí, w_E, w_V y w_A son constantes positivas que indican la importancia relativa de las entidades, acciones y atributos respectivamente. Sus valores son *asignados por el usuario* de acuerdo con sus intereses de análisis.

Condiciones	Efecto
$w_E = w_V = w_A$	***No se enfatiza nada*** Misma importancia para entidades, acciones y atributos
$w_E > w_V, w_A$	Énfasis en las semejanzas relacionadas con las ***entidades***
$w_V > w_E, w_A$	Énfasis en las semejanzas relacionadas con las ***acciones***
$w_A > w_E, w_V$	Énfasis en las semejanzas relacionadas con los ***atributos***

Figura 4.8 Evaluación de la importancia de los conceptos

La figura 4.8 describe las posibles combinaciones de valores de estas constantes y sus correspondientes efectos sobre la medida de semejanza. El caso por omisión considera $w_E = w_V = w_A$.

Por su parte, la función $b\left(p_{G_1}c, p_{G_2}c\right)$ expresa la semejanza semántica entre los conceptos originales $p_{G_1}c$ y $p_{G_2}c$ con base en una jerarquía de conceptos preestablecida. Esta función se define de la siguiente manera:[5]

$$b\left(p_{G_1}c, p_{G_2}c\right) = \begin{cases} 1 & \text{si } type\left(p_{G_1}c\right) = type\left(p_{G_2}c\right) \text{ y } referent\left(p_{G_1}c\right) = referent\left(p_{G_2}c\right) \\ \dfrac{depth}{depth+1} & \text{si } type\left(p_{G_1}c\right) = type\left(p_{G_2}c\right) \text{ y } referent\left(p_{G_1}c\right) \neq referent\left(p_{G_2}c\right) \\ \dfrac{2 \times d_c}{d_{p_{G_1}c} + d_{p_{G_2}c}} & \text{si } type\left(p_{G_1}c\right) \neq type\left(p_{G_2}c\right) \end{cases}$$

En la primera condición, los conceptos $p_{G_1}c$ y $p_{G_2}c$ son iguales, y por lo tanto $b\left(p_{G_1}c, p_{G_2}c\right) = 1$.

En la segunda condición, los conceptos $p_{G_1}c$ y $p_{G_2}c$ se refieren a diferentes "individuos" del mismo tipo, esto es, a diferentes instancias de la misma clase. En

[5] En esta definición, la condición $type\left(p_{G_1}c\right) = type\left(p_{G_2}c\right)$ también se satisface cuando los tipos conceptuales son *sinónimos.*

este caso, $b\left(p_{G_1}c, p_{G_2}c\right) = depth/(depth+1)$, donde *depth* indica el número de niveles de la jerarquía de conceptos dada. De acuerdo con esta asignación, la semejanza entre dos conceptos con el mismo tipo pero con diferentes referentes es siempre mayor que la semejanza entre dos conceptos con diferentes tipos.

En la tercera condición, los conceptos $p_{G_1}c$ y $p_{G_2}c$ tienen diferentes tipos, es decir, señalan elementos de distintas clases. En este caso, $b\left(p_{G_1}c, p_{G_2}c\right)$ expresa la semejanza semántica de los conceptos $type\left(p_{G_1}c\right)$ y $type\left(p_{G_2}c\right)$ en la jerarquía de conceptos preestablecida. Esta semejanza se calcula usando, una vez más, una expresión análoga al coeficiente de Dice:[6]

$$b\left(p_{G_1}c, p_{G_2}c\right) = \frac{2 \times d_c}{d_{p_{G_1}c} + d_{p_{G_2}c}}$$

En este caso, d_i es la distancia, expresada como el número de nodos, desde el concepto *i* hasta la raíz de la jerarquía de conceptos.

Semejanza relacional

La semejanza relacional $0 \leq s_r \leq 1$ indica que tan similares son las relaciones entre los conceptos comunes en ambos grafos conceptuales G_1 y G_2. En otras palabras, la semejanza relacional indica que tan parecidos son los *vecindarios* de los conceptos del traslape en los grafos conceptuales originales.

El vecindario del traslape *O* en el grafo conceptual *G*, denotado como $N_O(G)$, es el conjunto de todas las relaciones conceptuales conectadas a los conceptos comunes en el grafo *G*; esto es:

---

[6] Esta expresión se usa para la medición de la semejanza semántica de dos conceptos en una jerarquía (Wu and Palmer, 1994). Algunas evaluaciones demuestran su alta correlación con las mediciones manuales de la semejanza semántica (Lin, 1998).

$$N_O(G) = \bigcup_{c \in O} N_G(c), \text{ donde:}$$
$$N_G(c) = \{ r \mid r \text{ está conectada a } p_G c \text{ en } G \}$$

Con base en esta definición, la semejanza relacional se calcula de la siguiente manera; también análoga al coeficiente de Dice:

$$s_r(G_1, G_2) = \frac{2 \sum_{r \in O} weight_O(r)}{\sum_{r \in N_O(G_1)} weight_{G_1}(r) + \sum_{r \in N_O(G_2)} weight_{G_2}(r)}$$

En esta fórmula $weight_G(r)$ indica la importancia de la relación conceptual *r* en el grafo conceptual *G*. Este valor se calcula de acuerdo con el vecindario de *r* en *G*; así se garantiza la homogeneidad entre los pesos de los conceptos y las relaciones conceptuales.

$$weight_G(r) = \frac{\sum_{c \in N_G(r)} weight(c)}{|N_G(r)|}, \text{ donde:}$$
$$N_G(r) = \{ c \mid c \text{ está conectado a } r \text{ en } G \}$$

Semejanza total

La semejanza total se obtiene combinando la semejanza conceptual s_c y la semejanza relacional s_r. En primer lugar, esta combinación debe ser estrictamente *multiplicativa*, de tal forma que la semejanza total sea proporcional a ambos componentes. Con base en esta consideración, la semejanza total se define como: $s = s_c \times s_r$.

Sin embargo, la semejanza relacional debe tener una *importancia secundaria*, porque su existencia depende directamente de la existencia de algunos conceptos comunes, y además porque aún cuando los dos grafos no tienen ninguna relación común, cierto nivel de semejanza puede existir entre ellos.

Así, la semejanza total *s* debe ser proporcional a las semejanzas conceptual y relacional, pero puede ser diferente de cero cuando $s_r = 0$. Este comportamiento se modela *suavizando* el efecto de la semejanza relacional sobre la semejanza total:

$$s = s_c \times (a + bs_r)$$

Con esta definición, cuando no existe ninguna semejanza relacional entre los dos grafos conceptuales (es decir, cuando $s_r = 0$), la semejanza total depende exclusivamente de la semejanza conceptual, siendo $s = as_c$.

Los coeficientes *a* y *b* indican la importancia relativa de la semejanza conceptual y relacional respectivamente. Sus valores son *establecidos por el usuario* de acuerdo con sus intereses de análisis, considerando únicamente las siguientes dos condiciones: $0 < a,b < 1$ y $a + b = 1$.

La figura 4.9 describe las posibles combinaciones de valores de estos coeficientes y sus correspondientes efectos sobre la medida de semejanza. El caso por omisión considera *a* = *b*.

4.2.2.2 Ejemplo de la medición de la semejanza

La medición de la semejanza se ejemplifica usando como base los dos grafos conceptuales y los dos traslapes de la figura 4.7.

Condiciones	Efecto
a = *b*	***No se enfatiza nada*** Conceptos y relaciones tienen la misma importancia
a > *b*	Énfasis en las semejanzas a nivel ***conceptual***
b > *a*	Énfasis en las semejanzas a nivel ***estructural***

Figura 4.9 Valores de importancia de los conceptos y relaciones

Condiciones	Traslape	s_c	s_r	s
a = 0.1, **b = 0.9**	[candidato]←(agt)←[criticar]→(ptn)→[candidato]	0.86	1	0.86
$w_E = w_V = w_A = 1$	[candidato:Cárdenas] [criticar] [candidato:Fox]	1.00	0	0.10
a = 0.9, b = 0.1	[candidato]←(agt)←[criticar]→(ptn)→[candidato]	0.86	1	0.86
$w_E = w_V = w_A = 1$	[candidato:Cárdenas] [criticar] [candidato:Fox]	1.00	0	0.90
a = 0.5, b = 0.5 **w_E = 2**	[candidato]←(agt)←[criticar]→(ptn)→[candidato]	0.84	1	0.84
$w_V = w_A = 1$	[candidato:Cárdenas] [criticar] [candidato:Fox]	1.00	0	0.50

Figura 4.10 Diferentes maneras de evaluar la semejanza

Una medida de semejanza diferente se calcula con respecto a cada uno de los traslapes. La mayor de estas medidas "parciales" es la medida final de la semejanza y el traslape correspondiente, su mejor descripción.

La manera de evaluar la semejanza entre los dos grafos depende de las constantes w_E, w_V y w_A y de los coeficientes *a* y *b*; diferentes combinaciones de éstos producen diferentes medidas de semejanza, y por ende, diferentes descripciones de esta última.

La figura 4.10 muestra tres maneras diferentes de evaluar la semejanza de los dos grafos conceptuales de la figura 4.5. El primer caso enfatiza las semejanzas relacionales (estructurales, porque también los conceptos interesan); el segundo, las semejanzas conceptuales; y el tercero, las semejanzas causadas especialmente por las entidades. Para cada uno de los casos se indica la mayor medida de semejanza y su correspondiente traslape.

Capítulo 5

Análisis de un Conjunto de Grafos Conceptuales

En este capítulo se proponen algunos métodos para el análisis automático de un conjunto de grafos conceptuales. Estos métodos permiten descubrir distintos tipos de patrones en un conjunto dado de grafos conceptuales, por ejemplo grupos, asociaciones y desviaciones.

Inicialmente presentamos un método para agrupar un conjunto de grafos conceptuales. Este método no sólo divide los textos en varios grupos, también construye una descripción –expresada como un grafo conceptual– de cada grupo. Después describimos un método para descubrir asociaciones y otro más para detectar de desviaciones entre grafos conceptuales. Estos métodos utilizan el agrupamiento de los grafos como un índice del conjunto, y logran detectar patrones altamente descriptivos del conjunto de grafos (textos).

Análisis de un Conjunto de Grafos Conceptuales

5.1 Agrupamiento de grafos conceptuales

Dada una colección de textos representados por grafos conceptuales, una de las tareas más importantes para su análisis es su agrupamiento. En primer lugar, este agrupamiento permite descubrir la *estructura oculta* de la colección. En segundo lugar, este agrupamiento constituye un *resumen organizado* de la colección que facilita su visualización, su posterior análisis, y también el descubrimiento de otros tipos de patrones interesantes.

El método propuesto a continuación es un método de *agrupamiento conceptual* que, a diferencia de las técnicas tradicionales de agrupamiento, no solo permite dividir el conjunto de grafos conceptuales en varios grupos, sino también asociar una descripción a cada uno de estos grupos y organizarlos jerárquicamente de acuerdo con dichas descripciones (Michalski, 1980).

Básicamente, dado un conjunto de grafos conceptuales, nuestro método identifica todas sus *regularidades* –elementos comunes de dos o más grafos del conjunto– y construye una *jerarquía conceptual* de ellas. Por ejemplo, la figura 5.2 ilustra un

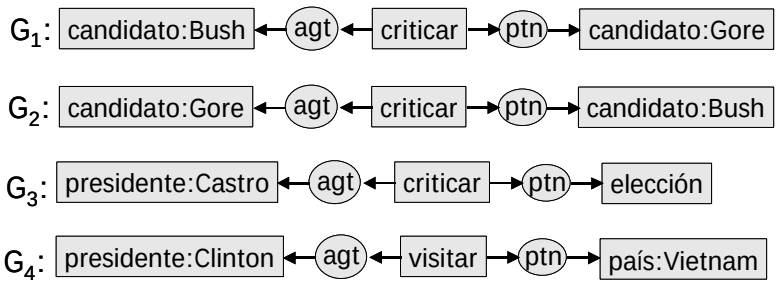


Figura 5.1 Una pequeña colección de grafos conceptuales

posible agrupamiento de los grafos de la figura 5.1.[1]

La jerarquía resultante *H* no es necesariamente un árbol o *lattice*, sino un conjunto de árboles, es decir, un bosque. Esta jerarquía es una especie de red de herencia, en donde los nodos inferiores indican regularidades especializadas y los nodos superiores sugieren regularidades generalizadas. Formalmente, cada nodo h_i de esta jerarquía se representa por una triada[2] (*cov*(h_i), *desc*(h_i), *coh*(h_i)), donde:

- *cov*(h_i), llamada cobertura de h_i, es el conjunto de grafos cubiertos por o asociados con la regularidad h_i.

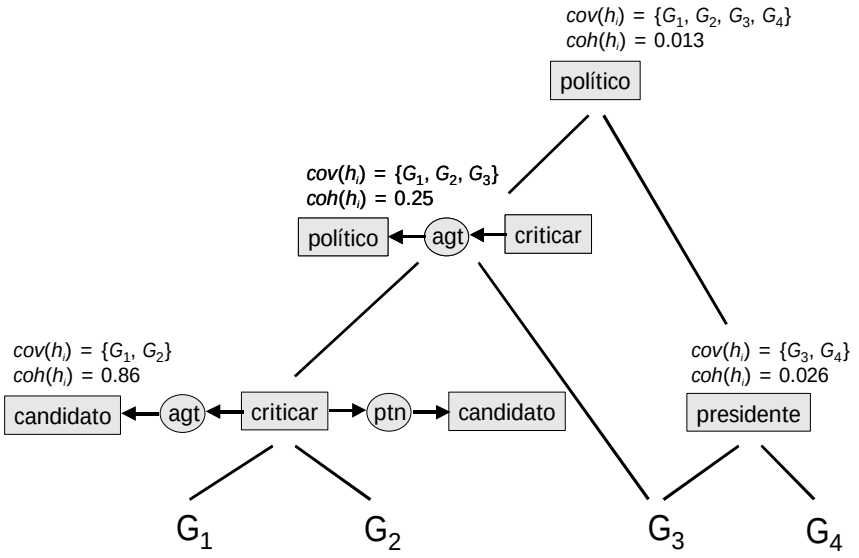


Figura 5.2 Un posible agrupamiento conceptual

[1] Este agrupamiento –jerarquía conceptual— enfatiza las semejanzas de tipo estructural, y se construyó con base en los siguientes parámetros de la medida de semejanza $w_E = w_V = w_A = 1$, y $a = 0.1$ y $b = 0.9$.

[2] Esta notación fue adaptada de (Bournaud and Ganascia, 1996), donde cada nodo se representa mediante un par (*cov*(h_i), *desc*(h_i)).

- *desc*(h_i), llamada descripción de h_i, es el conjunto de los elementos comunes de los grafos cubiertos por h_i, es decir, es el *traslape* de los grafos de *cov*(h_i). Entonces, *desc*(h_i) indica propiamente la regularidad.
- *coh*(h_i), llamada cohesión de h_i, es la semejanza mínima entre dos grafos cualesquiera de *cov*(h_i). Esto significa que para todo nodo h_i se cumple la siguiente condición: $\forall G_i, G_j \in cov(h_i),\ sim(G_i, G_j) \geq coh(h_i)$.

Dados dos nodos h_i y h_j de la jerarquía, el nodo h_j es un *descendiente* del nodo h_i, o lo que es lo mismo, el nodo h_i es un *ascendente* del nodo h_j, descrito como $h_j < h_i$, si y sólo si:

1. El nodo h_i agrupa o cubre más grafos que el nodo h_j: $cov(h_j) \subset cov(h_i)$.
2. La descripción del nodo h_i es una generalización de la descripción del nodo h_j: $desc(h_j) < desc(h_i)$.
3. La cohesión de los grafos del agrupamiento h_i es menor o igual que la cohesión de los grafos del agrupamiento h_j: $coh(h_i) \leq coh(h_j)$.

Con base en estas consideraciones, el conjunto de nodos hijos de h_i, denotado por $S(h_i)$, y el conjunto de nodos padre de h_i, denotado por $P(h_i)$, se definen de la siguiente manera:

$$S(h_i) = \{h_j \in H \mid h_j < h_i, \nexists h_k : h_j < h_k < h_i\}$$
$$P(h_i) = \{h_j \in H \mid h_i < h_j, \nexists h_k : h_i < h_k < h_j\}$$

5.1.1 Construcción de la jerarquía conceptual

La mayoría de los métodos de agrupamiento conceptual consideran solamente datos numéricos o simbólicos. Por ejemplo, el método COBWEB (Fisher, 1987) usa datos simbólicos y el método CLASSIT (Gennari *et al*., 1989) considera datos numéricos.

El agrupamiento conceptual de datos con una estructura más elaborada, como por ejemplo los grafos conceptuales, ha sido poco estudiado. Esto se debe principalmente a que su comparación es generalmente mucho más compleja que la comparación de los datos numéricos y simbólicos.

Actualmente son conocidos sólo dos métodos para el agrupamiento conceptual de un conjunto de grafos conceptuales (Mineau and Godin, 1995; Godin *et al*., 1995; Bournaud and Ganascia, 1996; Bournaud and Ganascia, 1997). Estos dos métodos construyen *incrementalmente* un espacio de generalizaciones del conjunto de grafos conceptuales. Dicho espacio refleja *todas* las semejanzas que existen entre los grafos conceptuales sin enfatizar *ningún* punto de vista específico. Por otra parte, estos métodos representan los grafos conceptuales como un conjunto de tríadas *concepto-relación-concepto*. Esta medida permite reducir un poco la complejidad de la comparación de los grafos, pero impide encontrar semejanzas descritas por un solo concepto[3], y además ocasiona la perdida de cierto tipo de información estructural.

El método propuesto a continuación se basa en estos dos métodos. Al igual que ellos, emplea una estrategia de aprendizaje no supervisado que permite construir *incrementalmente* el agrupamiento conceptual del conjunto de grafos, pero adicionalmente este método incorpora algunas características que lo hacen más atractivo para los propósitos de la minería de texto. Por ejemplo:

1. Encuentra grupos cuya descripción son *únicamente conceptos*, es decir, grupos donde la semejanza entre sus elementos es exclusivamente de tipo conceptual.
2. Considera *toda la información estructural* de los grafos conceptuales[4]. Por ejemplo, permite considerar adecuadamente relaciones *n*-arias.

---

[3] Para estos métodos las relaciones son las unidades básicas.

[4] Esta característica es una consecuencia directa del método de comparación de los grafos conceptuales descrito en el capitulo anterior.

3. Utiliza la *medida de semejanza* entre los grafos conceptuales para enfatizar los intereses del usuario durante la construcción de la jerarquía.
4. Permite hacer *agrupamientos con traslapes*, y en consecuencia considerar adecuadamente el aspecto multitemático de los textos.

Estas características permiten, entre otras cosas, mejorar la expresividad de las descripciones de grupo, y aumentar la flexibilidad del proceso de agrupamiento.

5.1.1.1 Proceso de construcción

La incorporación de un nuevo grafo G_n a la jerarquía conceptual H se realiza en dos pasos.

En el primer paso se añade a la jerarquía un nodo que cubre exclusivamente al nuevo grafo[5] (ver la figura 5.3a). Este nuevo nodo se define como ($\{G_n\}$, G_n, 1).

En el segundo paso se identifican todas las regularidades asociadas con la nueva evidencia. Estas regularidades (nuevos nodos) se añaden a la jerarquía siguiendo una

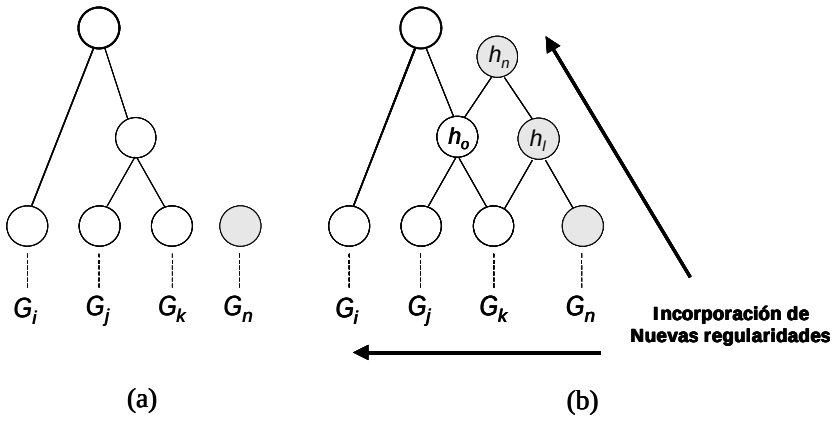


Figura 5.3 Incorporación de un nuevo grafo a la jerarquía

[5] Este nodo es parte de la jerarquía conceptual, pero no expresa ninguna regularidad, ya que solamente cubre a un grafo de la colección.

```
Procedimiento Insertar_Grafo Gn en H
1  Define nuevo nodo hn  = ({Gn}, Gn, 1)
2  Para cada nodo hi ∈ H que cubra un solo grafo
3     Identifica regularidades entre hi y hn
4  Inserta hn en la jerarquía (H ← hn)

Procedimiento Identificar_Regularidades entre hold y hlast
1  Construir nuevo nodo hnew combinando hold y hlast
2  Si hold está duplicado – igual que hnew
3     Borrar hold  (H → hold)
4     Actualizar padres de hold – conectar con hnew
5  Si hlast está duplicado – igual que hnew
6     Borrar hlast
8     Para cada nodo padre hk de hold
9        Identifica regularidades entre hk y hnew
```

Figura 5.4 Algoritmo general de agrupamiento conceptual

estrategia ascendente (*bottom-up*, en inglés), esto es, cada nodo de nivel superior se construye combinando dos nodos de niveles más bajos. Por ejemplo, el nodo h_n de la figura 5.3(b) se construye a partir de los nodos h_o y h_l. En este caso, el nodo nuevo h_n se define de la siguiente manera:

$$cov(h_n) = cov(h_o) \cup cov(h_l)$$

$$desc(h_n) = match(desc(h_o), desc(h_l))$$

$$coh(h_n) = \begin{cases} sim(desc(h_o), desc(h_l)) & \text{si } |cov(h_o)| = |cov(h_o)| = 1 \\ min(coh(h_o), coh(h_l)) & \text{otro caso} \end{cases}$$

En este caso, la función $match(G_i, G_j)$ regresa el mejor traslape de los grafos G_i y G_j (ver sección 4.2.1 del capítulo 4); la función $sim(G_i, G_j)$ regresa la medida de semejanza de los grafos G_i y G_j (ver sección 4.2.2 del capítulo 4); y la función $min(coh(h_o), coh(h_l))$ regresa la menor cohesión entre los grupos h_i y h_j.

Por otra parte, cada vez que una nueva regularidad h_n se añade a la jerarquía conceptual H, las regularidades duplicadas –redundantes– se eliminan. Las reglas de eliminación de redundancias son las siguientes:

- Si $desc(h_o) = desc(h_n)$, entonces el nodo h_o se elimina de la jerarquía.

- Si $desc(h_l) = desc(h_n)$, entonces h_l se elimina.

El algoritmo general para la inserción de un nuevo grafo conceptual a la jerarquía conceptual se describe en la figura 5.4.

5.1.1.2 Ilustración del proceso de construcción

La construcción de un agrupamiento conceptual se ilustra, paso a paso, en la figura 5.5. Allí se muestra el proceso de construcción de la jerarquía conceptual correspondiente a los grafos de la figura 5.1. Dicho proceso enfatiza la semejanza relacional entre los grafos conceptuales, y consta de tres etapas o iteraciones diferentes.

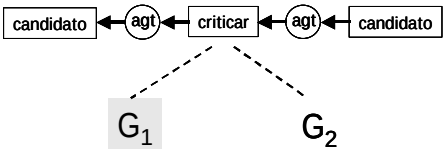
Cada una de estas iteraciones corresponde a la inserción de un nuevo grafo en la jerarquía conceptual. Los elementos resaltados representan los nodos de la iteración anterior. El resto de los elementos indican los nodos construidos en la iteración actual; de ellos, los elementos cruzados señalan las regularidades redundantes eliminadas.

A partir de esta figura, es evidente la estrecha relación que existe entre el método de comparación de grafos (referirse al capítulo 4) y la identificación de las nuevas regularidades. Esta relación define la construcción de la jerarquía conceptual como un proceso *basado en conocimiento* y *dirigido por el usuario*.

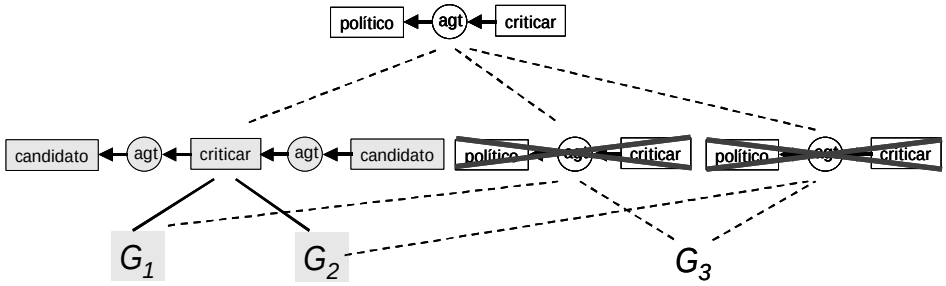
Esto último significa principalmente que el uso de distinto conocimiento del dominio –distintas jerarquías de conceptos– y el establecimiento de distintos intereses de análisis por parte del usuario –parámetros de la medida de semejanza– pueden producir *diferentes jerarquías conceptuales*.[6]

---

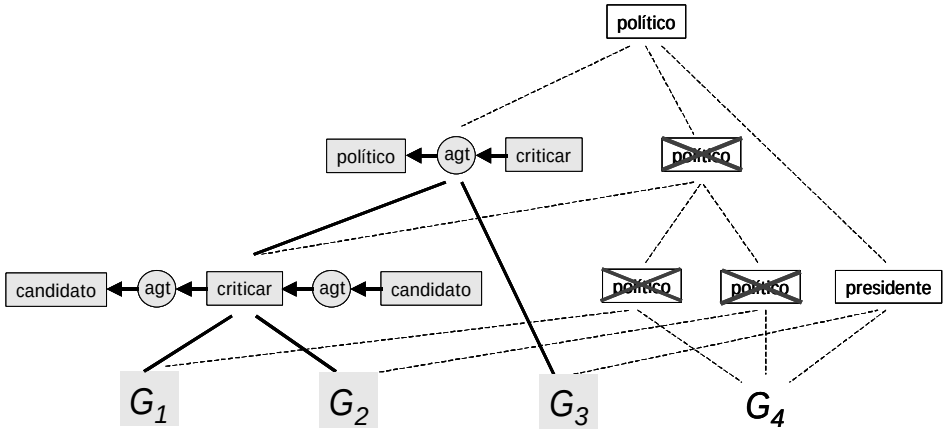
[6] En (Gibert and Córtes, 1998) se describe un método para agrupar un conjunto de objetos representados por atributos cuantitativos y cualitativos. Dicho método permite agrupar el mismo conjunto de objetos de distintas maneras dependiendo de la importancia relativa establecida por el usuario para los dos tipos de atributos.



(a) Primera iteración



(a) Segunda iteración



(c) Tercera iteración

Figura 5.5 Construcción del agrupamiento conceptual

A manera de ejemplo, en la figura 5.6 se muestra un agrupamiento conceptual diferente del mismo conjunto de grafos conceptuales. Este nuevo agrupamiento enfatiza las *semejanzas conceptuales*. Su construcción se basó en los siguientes parámetros de la medida de semejanza: $w_E = w_V = w_A = 1$ y $a = 0.1$ y $b = 0.1$.

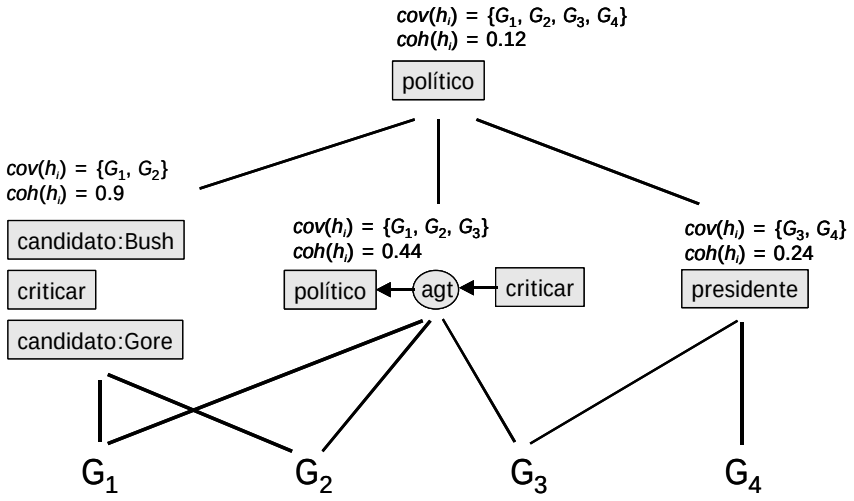


Figura 5.6 Un agrupamiento conceptual diferente

5.2 Identificación de las principales regularidades

Tal y como se explicó anteriormente, dado un conjunto de grafos conceptuales, su jerarquía conceptual *H* expresa *todas* sus regularidades de acuerdo con un punto de vista específico. En otras palabras, la jerarquía conceptual indica todas las generalizaciones comunes –cuya descripción enfatiza los intereses del usuario– de dos o más grafos de la colección.

Típicamente estas jerarquías son grandes, y por lo tanto indican muchas regularidades *no* interesantes para el usuario. Así, bajo este contexto, la identificación de las principales regularidades de una jerarquía conceptual adquiere gran relevancia.

Nosotros definimos una regularidad como interesante de acuerdo con los siguientes dos *criterios intuitivos*:

1. Entre más extensa es la regularidad, esto es, entre más grafos conceptuales cubra, más interesante es.
2. Entre mayor es la semejanza entre los grafos conceptuales cubiertos por una regularidad, más interesante es esta última.

Con base en estos dos criterios, la regularidad expresada por el nodo $h_i \in H$ es interesante si:

$$|cov(h_i)| \geq min\,cov, \text{ donde } 1 < min\,cov \leq n$$
$$coh(h_i) \geq mincoh, \text{ donde } 0 < mincoh \leq 1$$

Aquí, *mincov* y *mincoh* indican respectivamente la cobertura y la cohesión mínimas de una regularidad considerada aún interesante por el usuario.

La primera condición favorece los *grupos grandes*. Ella indica un corte transversal en la jerarquía conceptual en $|cov(h_i)| = min\,cov$, donde las regularidades por arriba de este corte son interesantes. La figura 5.7(a) ilustra esta condición.

Por su parte, la segunda condición favorece los *grupos homogéneos*, es decir, grupos que, aunque pequeños, contienen grafos conceptuales muy parecidos. Esta condición expresa un corte transversal irregular en la jerarquía conceptual en $coh(h_i) = mincoh$, donde las regularidades por abajo del corte son consideradas interesantes. La figura 5.7(b) ilustra esta condición.

La combinación de estas dos condiciones permite identificar una sección de la jerarquía conceptual *H* como interesante. La definición de dicha sección se describe en la figura 5.7(c).

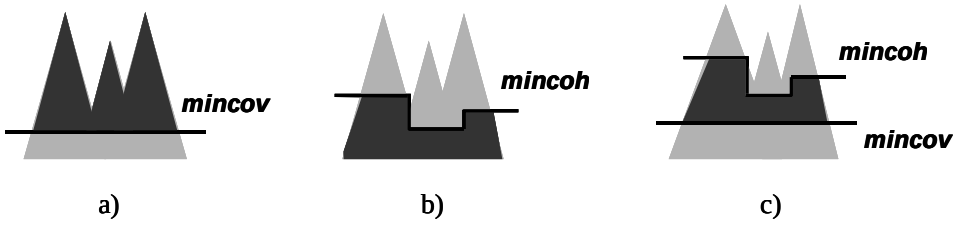


Figura 5.7 Selección de las principales regularidades

En la figura 5.8 se ejemplifica el proceso de selección de las principales regularidades de la jerarquía conceptual de la figura 5.2. Esta selección se realizó con base en los siguientes valores: $min\,cov = 3$ y $mincoh = 0.25$.

Estas condiciones indican que la descripción del nodo h_i, $desc(h_i)$, es interesante si y sólo si:

1. Este nodo cubre o agrupa al menos tres grafos de la colección, es decir, si: $|cov(h_i)| \geq 3$.

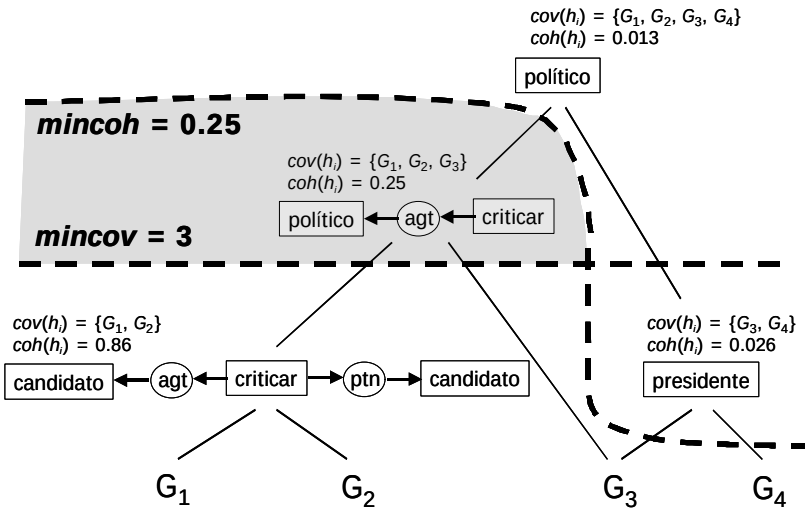


Figura 5.8 Identificación de las principales regularidades

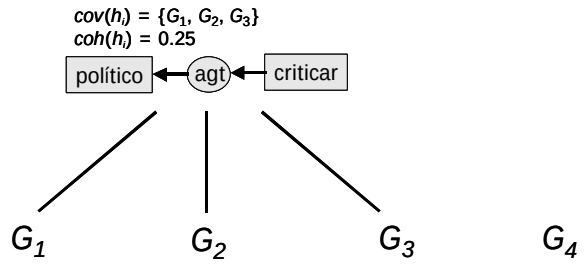


Figura 5.9 Agrupamiento conceptual reducido

```
Procedimiento Crear_Jerarquía_Reducida de H
//Parámetros mincov y mincoh
1  Para cada nodo h_i de H que cubra un solo grafo
2     Hacer copia h'_i  de h_i
3     Añadir h'_i  a la jerarquía reducida H'
4     Para cada nodo padre h_p de h_i
5         Añade_Regularidad h_p a H'

Procedimiento Añade_Regularidad h_i a H'
//Parámetros mincov y mincoh
1  Si el nodo h_i tiene más cobertura que mincov
2     Si el nodo h_i tiene más cohesión que mincoh
3         Insertar h_i en la jerarquía H'
4     Para cada nodo padre h_p de h_i
5         Añade_Regularidad h_p a H'
```

Figura 5.10 Algoritmo para construir agrupamientos reducidos

2. Los grafos cubiertos por nodo tienen al menos una semejanza de 0.25, esto es: $coh(h_i) \geq 0.25$.

Considerando solamente las regularidades interesantes de la jerarquía conceptual H se puede construir un *agrupamiento reducido* H'. La figura 5.9 muestra este agrupamiento.

El algoritmo general para identificar las principales regularidades en una jerarquía conceptual, con base en los parámetros *mincov* y *mincoh*, se describe en la figura 5.10.

5.3 Descubrimiento de asociaciones

5.3.1 Antecedentes

El descubrimiento de reglas asociativas es un problema clásico de la minería de datos (Han and Kamber, 2001). Este tipo de reglas se definen de la siguiente manera:

Dado un conjunto de transacciones, donde cada transacción es un conjunto de elementos, una *regla asociativa* es una expresión de la forma $X \Rightarrow Y$, donde X y Y son subconjuntos de elementos. Intuitivamente, estas reglas indican que las transac-

ciones que contienen el subconjunto de elementos *X* tienden a contener el subconjunto de elementos *Y*.

Por ejemplo, en una base de datos de un supermercado una posible regla asociativa es: “30% de las transacciones que contienen cerveza también contienen pañales; 2% de todas las transacciones contienen ambos elementos”. En este caso, 30% es la confianza de la regla y 2% su soporte.

Así pues, el descubrimiento de las reglas asociativas en un conjunto de transacciones se define como el problema de encontrar todas las reglas con una confianza y un soporte mayores que los valores mínimos especificados por el usuario *minconf* y *minsup* respectivamente.

Típicamente, este proceso se realiza en dos fases. En la primera fase se encuentran todas las combinaciones de elementos con un soporte mayor que *minsup*. Estas combinaciones son llamadas conjuntos de elementos frecuentes. En la segunda fase, a partir de estos conjuntos frecuentes, se generan las reglas asociativas. La idea general es que si, por ejemplo, $\{a,b\}$ y $\{a,b,c,d\}$ son conjuntos de elementos frecuentes, entonces existe la regla asociativa $\{a,b\} \Rightarrow \{c,d\}$; para la cual los valores de confianza y soporte se determinan de la siguiente manera:

$$confianza = ocurrencia\{a,b,c,d\} / ocurrencia\{a,b\}$$
$$soporte = ocurrencia\{a,b,c,d\}$$

Si la confianza de la regla es mayor o igual que el valor *minconf* predefinido por el usuario, entonces la regla es *interesante*.

5.3.2 Asociaciones entre grafos conceptuales

Emulando lo realizado en la minería de datos, nosotros definimos una asociación entre grafos conceptuales de la siguiente manera:

Dado un conjunto de grafos conceptuales $C = \{G_i\}$, donde cada grafo conceptual representa un texto diferente, una *regla asociativa* es una expresión de la forma $g_i \Rightarrow$

g_j (c/s), donde g_i es una generalización de g_j ($g_j < g_i$), c es la confianza de la regla y s es su soporte.

Básicamente, una regla de este tipo indica que los grafos conceptuales del conjunto que contienen el grafo g_i, c% de las veces también contienen el grafo más especializado g_j. Además que s% de los grafos de la colección contienen el grafo especializado g_j.

La figura 5.11 muestra dos reglas asociativas obtenidas a partir del conjunto de grafos de la figura 5.1. La primera de estas reglas indica que todos los grafos –textos– que mencionan algún funcionario, hablan en particular de algún presidente, y además que el 50% de los grafos –textos– de la colección hablan sobre presidentes.

Por su parte, la segunda regla señala que en el subconjunto de grafos –textos– que mencionan un político, un 75% se refieren a un político criticando. Esta regla también indica que un 75% de los grafos –textos– de la colección hacen referencia a un político criticando algo.

Entonces, el descubrimiento de asociaciones en un conjunto de grafos conceptuales se define como el problema de encontrar todas las reglas asociativas $g_i \Rightarrow g_j$ (c/s), tal que $c \geq minconf$ y $s \geq minsup$.

El método propuesto para descubrir asociaciones en un conjunto de grafos conceptuales se describe a continuación.

5.3.2.1 Método de descubrimiento

Básicamente, el descubrimiento de las reglas asociativas en un conjunto de grafos conceptuales $C = \{G_i\}$ se auxilia de su jerarquía conceptual H.

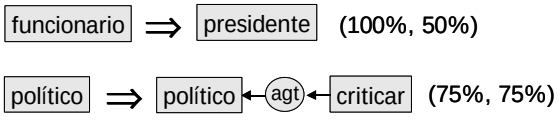


Figura 5.11 Ejemplos de reglas asociativas

Cada nodo h_i de la jerarquía conceptual *H* expresa una regularidad, cuya descripción *desc*(h_i) es una generalización común de dos o más grafos de *C*. Además, todo grafo conceptual *g* implícito en h_i, es decir, todo grafo conceptual *g* tal que: *desc*(h_i) < *g* y $\nexists h_k \in H : desc(h_i) < desc(h_k) < g$, es también una generalización común –*implícita*– del mismo subconjunto de grafos de *C*.

Por ejemplo, la figura 5.12 muestra algunas generalizaciones comunes implícitas en la jerarquía conceptual de la figura 5.2. En esta figura, los elementos resaltados corresponden a los nodos de la jerarquía conceptual original, y los demás nodos indican las generalizaciones comunes implícitas.

Con base en esta figura, es posible determinar dos tipos de reglas asociativas en una jerarquía conceptual *H*: asociaciones explícitas y asociaciones implícitas.

Asociaciones explícitas: Para cada par de nodos h_i y h_j de la jerarquía conceptual *H*, tal que $h_j < h_i$, la siguiente regla asociativa es valida:

$$desc(h_i) \Rightarrow desc(h_j) \quad \left(c = \frac{|cov(h_j)|}{|cov(h_i)|},\ s = \frac{|cov(h_j)|}{|C|} \right)$$

Asociaciones implícitas: Para todo grafo conceptual *g* implícito en h_i, las siguientes reglas asociativas son validas:

$$g \Rightarrow desc(h_i) \quad \left(c = 1, s = \frac{|cov(h_i)|}{|C|} \right)$$

- $\forall h_j \in H$: *desc*(h_j) < *desc*(h_i)

$$g \Rightarrow desc(h_j) \quad \left(c = \frac{|cov(h_j)|}{|cov(h_i)|}, s = \frac{|cov(h_j)|}{|C|} \right)$$

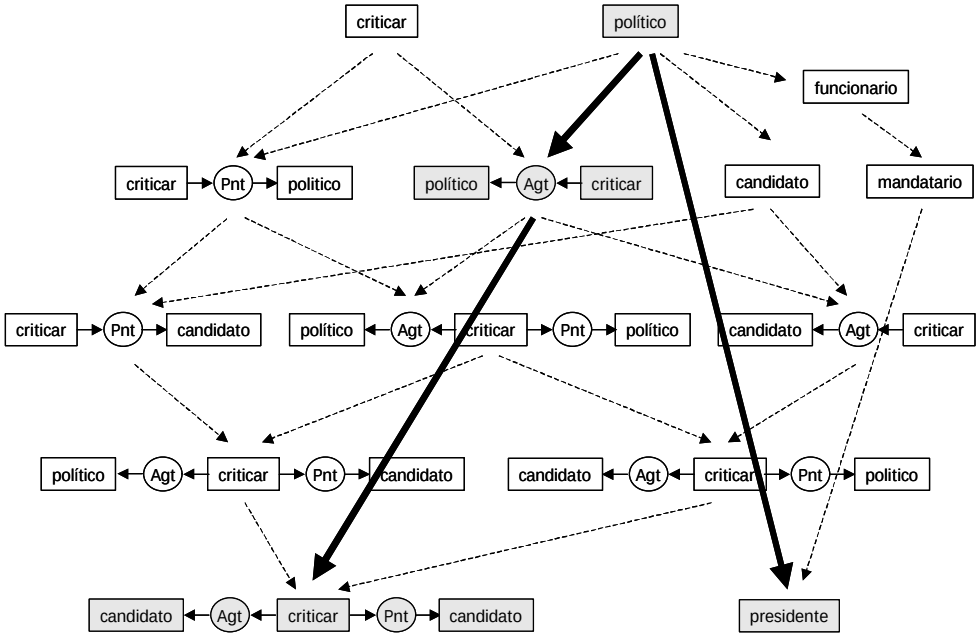


Figura 5.12 Generalizaciones comunes implícitas en *H*

- $\forall h_k \in H$: $desc(h_i) < g < desc(h_k)$

$$desc(h_k) \Rightarrow g \quad \left(c = \frac{|cov(h_i)|}{|cov(h_k)|}, s = \frac{|cov(h_i)|}{|C|} \right)$$

De acuerdo con estas definiciones es posible descubrir *todas* las reglas asociativas en un conjunto de grafos conceptuales. Usualmente, el conjunto de todas estas reglas es muy grande y contiene mucha información *redundante*. Por ejemplo, las tres asociaciones implícitas de la figura 5.13 contienen información redundante. En este caso, las dos primeras pueden ser deducidas directamente a partir de la tercera.

Así pues, es necesario eliminar, o nunca construir, las asociaciones que sean redundantes. A continuación se define una asociación implícita redundante.

Asociación implícita redundante: La regla asociativa implícita $g_i \Rightarrow g_k(1,a)$ es redundante, si y sólo si, una de las siguientes dos condiciones se satisface:

- Existe otra regla asociativa implícita $g_h \Rightarrow g_l(1,a)$, tal que g_h es una generalización de g_i ($g_i \leq g_h$), y/o g_l es una especialización de g_k ($g_l \leq g_k$).
- Existe la regla asociativa implícita $g_i \Rightarrow g_j(1,b)$ en combinación con la regla asociativa explícita $g_j \Rightarrow g_k(g,a)$, en donde, $g_k < g_j < g_i$.

En la figura 5.14 se muestran las reglas asociativas no redundantes correspondientes a la colección de la figura 5.1. Estas reglas se obtuvieron a partir de la jerarquía conceptual de la figura 5.2. Todas ellas tienen valores de confianza y soporte mayores a 0.5.

El algoritmo general para el descubrimiento de asociaciones en un conjunto de grafos conceptuales a partir de su jerarquía conceptual se describe en la figura 5.15. Este algoritmo recorre ascendentemente la jerarquía conceptual, e identifica, para cada nodo h_i de la jerarquía, las asociaciones *no redundantes* con una confianza y un soporte mayores a los valores mínimos establecidos por el usuario *minconf* y *minsup* respectivamente.

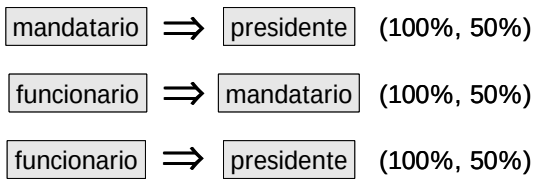


Figura 5.13 Asociaciones implícitas redundantes

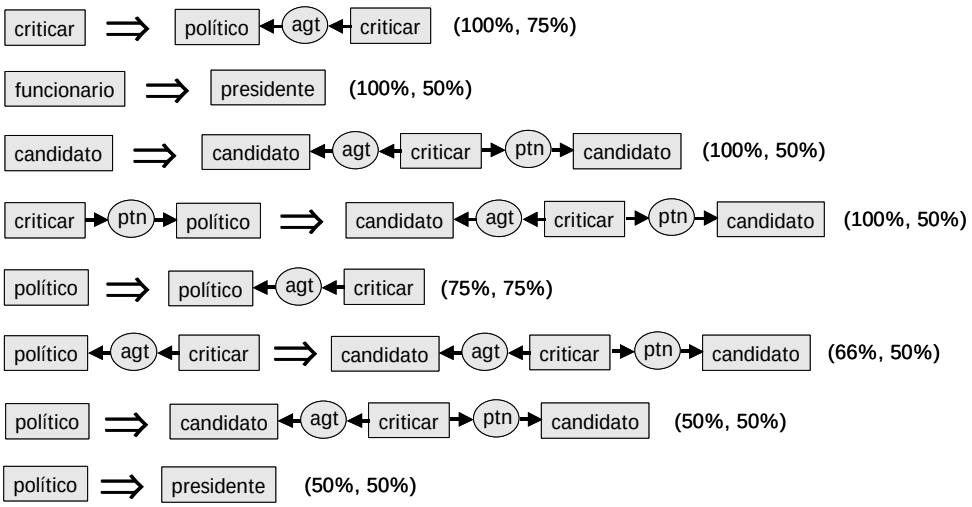


Figura 5.14 Las asociaciones del caso ejemplo

En este algoritmo, las asociaciones explícitas e implícitas son construidas por separado. Por ejemplo, las asociaciones explícitas se construyen relacionando cada nodo h_i con cada uno sus nodos antecesores; mientras que las asociaciones implícitas se construyen relacionando cada nodo h_i con los grafos conceptuales implícitos g que satisfagan la condición: $\nexists g' : h_i < g', g < g'$. Este criterio impide construir asociaciones implícitas redundantes.

5.4 Detección de desviaciones

5.4.1 Antecedentes

Los métodos estadísticos tradicionales generalmente consideran que los datos raros o *desviaciones* son una fuente de ruido. Por ello, estos métodos intentan minimizar sus efectos. Diferente a este punto de vista, algunos métodos de minería de datos se enfocan en la detección de dichas desviaciones. Estos métodos consideran que las desviaciones pueden esconder conocimientos verdaderamente inesperados e interesantes.

```
Procedimiento Descubre_Asociaciones en H
//Parámetros de entrada minconf y minsup
1  Para cada nodo h_i de la jerarquía
2    Si h_i tiene más de minsup: |cov(h_i)|/|C| ≥ minsup
3       Asociaciones_Implícitas con h_i
4       Para cada nodo padre h_p de h_i
5           Asociación_Explicita entre h_p y h_i

Procedimiento Asociaciones_Implícitas con h_i
//Parámetros de entrada h_i, minconf y minsup
1  Definir confianza: c ← 1
2  Calcular soporte: s ← |cov(h_i)|/|C|
3  Para cada concepto c_i del grafo desc(h_i)
4    Si c_i no está cubierto por ningún nodo padre de h_i
5       Construir regla "c_i → h_i (c, s)"
6  Para cada relación r_i del grafo desc(h_i)
7     Si r_i no está cubierta por ningún nodo padre de h_i
8       Construir grafo estrella g de la relación r_i
9       Construir regla "g → h_i (c, s)"

Procedimiento Asociación_Explicita entre h_p y h_i
//Parámetros de entrada h_p, h_i, minconf y minsup
1  Calcular soporte: s ← |cov(h_i)|/|C|
2  Calcular Confianza: c ← |cov(h_i)|/|cov(h_p)|
3  Si la confianza es mayor o igual que minconf
4     Construir regla "h_p → h_i (c, s)"
5     Para cada nodo padre h_k de h_p
6        Asociación_Explicita entre h_k y h_i
```

Figura 5.15 Algoritmo para el descubrimiento de asociaciones

Típicamente, los métodos para la detección de desviaciones emplean información adicional a los datos, por ejemplo: condiciones preestablecidas o restricciones de integridad (Guzmán, 1996). Solamente en algunas ocasiones estos métodos aprovechan la propia *redundancia* de los datos. Entre los métodos que aprovechan la redundancia de los datos, y que por lo tanto resultan ser los más interesantes, destacan los siguientes tres enfoques (Han and Kamber, 2001):

- **Enfoque estadístico**

Este enfoque asume una distribución o modelo de probabilidad para los datos (por ejemplo, una distribución normal), y considera una desviación todo dato que salga de dicho modelo.

Su aplicación requiere que se conozcan la distribución de los datos, algunos de sus parámetros, media y varianza generalmente, y el número de desviaciones esperadas. Una descripción amplia de este enfoque se encuentra en (Barnett and Lewis, 1994).

- **Enfoque basado en distancia**

Este enfoque considera que el objeto *o* del conjunto *C* es una desviación con parámetros *p* y *d*, si al menos una fracción *p* de los objetos de *C* está a una distancia mayor que *d* de *o*. En otras palabras, este enfoque considera que el objeto *o* es una desviación si no tiene suficientes objetos "vecinos" en una vecindad de radio *d*.

La aplicación de este enfoque requiere que el usuario establezca los parámetros *p* y *d*. Algunos algoritmos se describen en (Knorr and Ng, 1998; Breunig *et al.*, 1999).

- **Enfoque basado en regularidades**

Este enfoque detecta las desviaciones desde una perspectiva más conceptual, tratando de imitar la manera en que los humanos las detectan. Básicamente, en él se determina una descripción general del conjunto de datos (por ejemplo, un dato representativo o el promedio de los datos) y se considera que un objeto es raro si se "desvía" considerablemente de dicha descripción.

Un método basado en este enfoque se describe en (Arning *et al.*, 1996). Este método permite detectar los datos más raros en un conjunto de números. En este caso se considera que los datos raros o desviaciones son aquellos que, en conjunto, causan la mayor disimilitud entre el conjunto de números.

5.4.2 Fundamentos de nuestro método

El método propuesto a continuación es un método *basado en regularidades*. Este método se fundamenta en las siguientes consideraciones:

- Dado un conjunto de grafos conceptuales $C = \{G_i\}$, una *característica representativa* es una generalización común g_c de más de *m* grafos de la colección, siendo *m* un umbral definido por el usuario.[7] El conjunto de características representativas de *C* se detona como *F*.
- Un *grafo conceptual raro* es un grafo que no tiene ninguna característica representativa[8]. Entonces, el conjunto de grafos raros *R* se define como: $R = \{G_r \in C | \nexists g_c \in F : G_r < g_c\}$.
- Una *desviación* es un patrón descriptivo *d* de algunos grafos raros del conjunto *C*. En otras palabras, una desviación *d* es una generalización de uno o más grafos raros del conjunto *C*. Entonces, dada una desviación *d* las siguientes dos condiciones se satisfacen:
  1. $\exists G_r \in R : G_r < d$.
  2. $\nexists G \in C, \nexists g \in F : G < g \wedge G < d$.

5.4.3 Desviaciones contextuales en grafos conceptuales

Dado un conjunto de grafos conceptuales $C = \{G_i\}$, donde cada grafo conceptual representa un texto diferente, una *desviación contextual* es una expresión de la forma: $g_i : g_j (r, s)$.

En esta expresión, g_i indica un contexto y g_j expresa algunas desviaciones para dicho contexto; *r* es el grado de rareza de la desviación g_j en el contexto g_i, y *s* es el

[7] El umbral *m* indica la representatividad de una característica para ser considerada una característica representativa. Los valores más razonables varían entre 0.3 y 0.6.

[8] Si no puede determinarse ninguna característica representativa del conjunto de grafos, entonces tampoco es conceptualmente adecuado detectar alguna desviación.

soporte de dicha desviación contextual, es decir, es la representatividad del contexto g_j en el conjunto *C*.

Básicamente, esta expresión indica que: dentro del subconjunto de grafos conceptuales –textos– que contienen el grafo g_i, y que representa el *s*% del conjunto completo de grafos, los grafos –textos– que contienen el grafo g_j son *raros*; siendo éstos solamente el *r*%.

Por ejemplo, en la figura 5.16 se muestra la única desviación contextual que se detectó en el conjunto de grafos de la figura 5.1 usando $m = 0.4$. Esta desviación indica que dentro del subconjunto de grafos conceptuales –textos– en los que un político critica algo, el cual representa el 75% del conjunto completo de grafos, es raro que el presidente Castro critique las elecciones estadounidenses; solamente el 33% de este subconjunto menciona dicho suceso.

Entonces, con base en lo anterior, definimos la detección de desviaciones en un conjunto de grafos conceptuales como el problema de encontrar todas las desviaciones contextuales g_i: g_j (*r*/*s*) para un umbral *m* preestablecido por el usuario.

5.4.3.1 Método de detección

La detección de las desviaciones contextuales en un conjunto de grafos conceptuales $C = \{G_i\}$ se auxilia de su jerarquía conceptual *H*. En esta jerarquía, cada nodo h_i indica un contexto específico de *C* descrito por la regularidad *desc*(h_i) y compuesto por el conjunto de grafos *cov*(h_i).

Además, el conjunto de nodos hijo de h_i, definido como: $S(h_i) = \left\{ h_j \in H \,\middle|\, h_j < h_i, \nexists h_k : h_j < h_k < h_i \right\}$, indica una partición del contexto h_i, don-

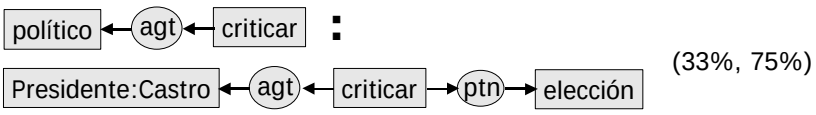


Figura 5.16 Un ejemplo de desviación contextual

de la descripción de cada uno de estos nodos hijo *desc*(h_j) expresa una característica posiblemente representativa del contexto h_i.

De acuerdo con esto último, y con base en las consideraciones expuestas en la sección 5.4.2, establecemos lo siguiente:

Característica representativa: La descripción *desc*(h_j) del nodo $h_j \in S(h_i)$ es una característica representativa del contexto h_i si:

$$\left| cov(h_j) \right| \geq m \times \left| cov(h_i) \right|$$

Entonces, el conjunto de características representativas del contexto h_i se define como: $F(h_i) = \left\{ desc(h_j) \middle| \; h_j \in S(h_i), \left| cov(h_j) \right| \geq m \times \left| cov(h_i) \right| \right\}$.

Grafo conceptual raro: El grafo conceptual $G_i \in cov(h_i)$ es un grafo raro en el contexto h_i, si y sólo si, no existe ninguna característica representativa $desc(h_j)$ en el contexto h_i tal que: $G_i \in cov(h_j)$.

Entonces, el conjunto de grafos raros del contexto h_i se define como: $R(h_i) = \{G_i \in cov(h_i) \mid \nexists g \in F(h_i) : G_i \in cov(g)\}$.

Desviación contextual: El grafo conceptual $desc(h_k)$, relacionado con el nodo $h_k < h_i$, es una desviación en el contexto h_i, si y sólo si: $\forall G_i \in cov(h_k) \Rightarrow G_i \in R(h_i)$.

En este caso, la desviación contextual puede definirse de la siguiente manera:

$$desc(h_i) : desc(h_k) \quad \left(r = \frac{\left| cov(h_k) \right|}{\left| cov(h_i) \right|}, s = \frac{\left| cov(h_i) \right|}{\left| C \right|} \right)$$

Esta definición permite encontrar *todas* las desviaciones contextuales –en un conjunto de grafos conceptuales con respecto a un valor predefinido de *m*. Muchas

de estas desviaciones contienen información redundante o información implícita en otras desviaciones. Por ejemplo, si es raro que se hable de animales en un conjunto determinado de grafos conceptuales, entonces obviamente es aún más raro que se hable de perros.

Entonces, para visualizar mejor las desviaciones es necesario eliminar las redundantes. Nosotros definimos una desviación redundante de la siguiente manera:

Desviación contextual redundante: La desviación contextual $g_i : g_k(a, b)$ es redundante si existe otra desviación contextual $g_i : g_j(g, b)$, con $a < g$, tal que g_j es una generalización de g_k. Esto implica que: $cov(g_k) \subset cov(g_j)$.

La figura 5.17 describe el algoritmo general para la detección de las desviaciones contextuales no redundantes en un conjunto de grafos conceptuales. Este algoritmo recorre, en forma descendente, todos los nodos de la jerarquía conceptual. Considera cada nodo h_i un contexto diferente, e identifica sus características representativas y sus grafos raros. Después, a partir del conjunto de grafos raros, detecta los nodos h_j descendientes de h_i cuya descripción representa una desviación para el contexto h_i.

```
Procedimiento Detecta_Desviaciones en H
//Parámetros: umbral m definido por el usuario
1  Para cada nodo h_i de la jerarquía H
2    Inicializar NO_RAROS ← ∅
5    Para cada nodo hijo h_s de h_i
6      Si |cov(h_s)| ≥ m X |cov(h_i)
7         Insertar en NO_RAROS los grafos que cubre h_s
8    Para cada nodo hijo h_s de h_i
9      Si |cov(h_s)| < m X |cov(h_i)|
10       Si el nodo h_s no cubre ningún grafo de NO_RAROS
11         Define la rareza: r ← |cov(h_s)|/|cov(h_i)|
12         Define el soporte: s ← |cov(h_i)|/|C|
13         Construye desviación "h_i:h_s (r, s)"
14       Sino entones
15         Desviaciones_Máxima de h_i con base en nodo h_s

Procedimiento Desviaciones_Máximas de h_i con base en nodo h_k
//Parámetros: h_i, h_j, conjunto NO_RAROS del contexto h_i
1  Para cada nodo hijo h_s de h_k
2    Si el nodo h_k no cubre ningún grafo de NO_RAROS
3      Define la rareza: r ← |cov(h_k)|/|cov(h_i)|
4      Define el soporte: s ← |cov(h_i)|/|C|
5      Construye desviación "h_i:h_k (r, s)"
6    Sino entonces
7      Desviaciones_Máximas de h_i con base en nodo h_k
```

Figura 5.17 Algoritmo para la detección de desviaciones

La jerarquía conceptual *H* es un conjunto de árboles, y no un solo árbol. Por ello es común que dicha jerarquía tenga varios nodos "raíz", aunque ninguno de ellos necesariamente representa la colección entera de grafos conceptuales. Para descubrir las desviaciones con respecto al conjunto completo de grafos es necesario entonces agregar un nodo raíz único a esta jerarquía conceptual. Este nodo lo definimos como: $h_r = (\{G_1,...,G_n\}, \mathsf{T}, 0)$, donde T, el concepto universal, es la generalización de cualquier grafo. Entonces, con base en este nodo se pueden obtener desviaciones contextuales de la forma T: g (a,1) que expresen desviaciones a nivel colección.

Capítulo 6

Resultados Experimentales

En este capítulo se describen los resultados del análisis de dos conjuntos de artículos científicos. El primero, denominado a partir de ahora conjunto A, se compone de 225 artículos sobre ciencias de la información; el segundo, referido como conjunto B, consiste de 495 artículos de ciencias de la computación.

Los resultados descritos a continuación son de dos tipos: cualitativos y cuantitativos. Los resultados cualitativos demuestran la capacidad de nuestro método para descubrir patrones interesantes a un nivel más descriptivo y completo que el temático. Estos resultados se ilustran con algunos ejemplos de grupos, asociaciones y desviaciones obtenidos en los conjuntos de prueba.

Por su parte, los resultados cuantitativos demuestran la viabilidad de nuestro método de minería de texto. Estos resultados describen diferentes facetas de la complejidad del análisis de los grafos. Por ejemplo: el crecimiento del agrupamiento, el tiempo de su construcción, la densidad de conexiones, etc.

Resultados Experimentales*

6.1 Resultados cualitativos

El método de minería de texto propuesto en los capítulos anteriores permite descubrir patrones más descriptivos del contenido de los textos que los métodos tradicionales. En esta sección se muestran algunos ejemplos de estos patrones; en especial se muestran algunos grupos, asociaciones y desviaciones obtenidos a partir de los dos conjuntos de prueba.

6.1.1 Agrupamiento conceptual

6.1.1.1 Descripción de los resultados

Métodos tradicionales de agrupamiento de textos

Los métodos de agrupamiento se usan frecuentemente en la exploración de grandes conjuntos de datos. Su objetivo general es dividir automáticamente un conjunto de datos –previamente no clasificados– en varios grupos "homogéneos".

En la minería de texto destacan dos enfoques de agrupamiento (ver capítulo 2). Ambos se ilustran en la figura 6.1, y se describen a continuación:

- Métodos basados en una medida o métrica de distancia entre los textos.

Estos métodos dividen la colección en varios grupos, e incluso señalan el texto más representativo de cada grupo. Son muy útiles cuando se desea conocer superficialmente la temática de una colección leyendo solamente unos cuantos textos.

* El proceso de su construcción de los grafos conceptuales de prueba se describe e ilustra detalladamente en el apéndice A. En general, estos grafos representan la intención de los artículos en cuestión, y tienen en promedio doce conceptos y diez relaciones.

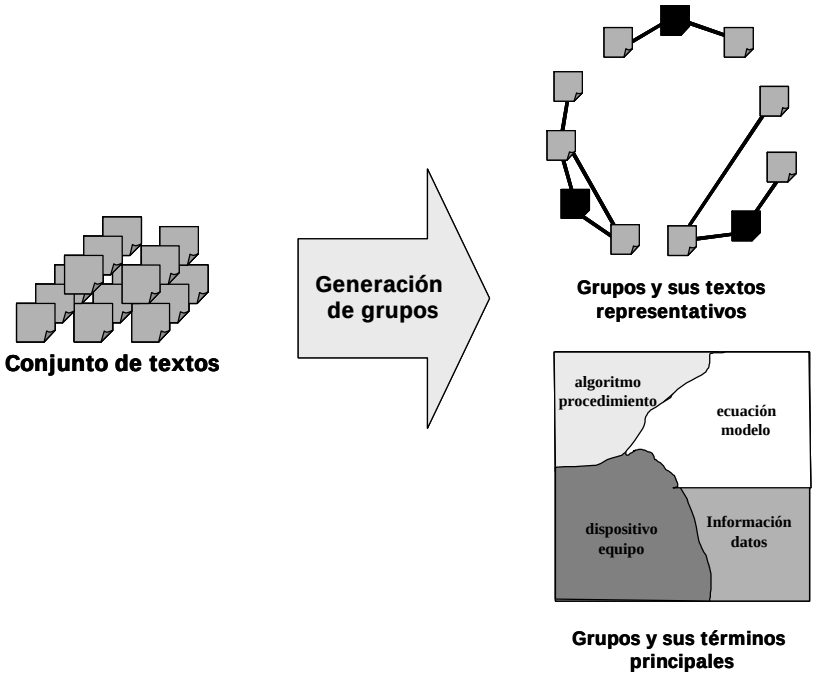


Figura 6.1 Métodos tradicionales de agrupamiento de textos

- Métodos basados en redes neuronales de tipo mapas auto-organizantes.

Estos métodos representan la colección de textos como un mapa, donde los puntos distantes representan textos muy diferentes, y los puntos cercanos textos similares. Básicamente, estos mapas dividen los textos en varias secciones o grupos con fronteras difusas, y describen cada sección con un conjunto de palabras clave.

En general, estos mapas permiten visualizar en forma directa las distintas temáticas de la colección, y además determinar el grado de pertenencia de los textos a cada una de ellas. Sin embargo, estos mapas no muestran ninguna información sobre las diferencias y las similitudes de los textos de un mismo grupo.

Nuestro método de agrupamiento conceptual

El método de agrupamiento propuesto en este trabajo no se basa en una medida de distancia o semejanza entre los textos, sino en la descripción cualitativa de dicha

semejanza; por ello el agrupamiento conceptual resultante considera *todas* las semejanzas entre los textos de la colección y no sólo las más importantes.

En términos generales, nuestro método divide en varios grupos un conjunto de textos, organiza jerárquicamente dichos grupos, y establece una descripción detallada de su contenido que considera acciones, atributos, entidades y sus relaciones semánticas.

La figura 6.2 ilustra un agrupamiento conceptual. En él se observa que los nodos superiores indican grupos grandes con descripciones generales, por ejemplo, el grupo de 25 textos que hablan sobre cáncer. También se observa que los nodos intermedios señalan subgrupos con descripciones más detalladas, por ejemplo, del grupo de textos que hablan de cáncer en estado terminal.

Entonces, gracias a su estructura jerárquica y a las descripciones detallas de los grupos, nuestro agrupamiento es más apropiado para la navegación de colecciones de textos, y también más ventajoso para el descubrimiento de otros patrones descriptivos tales como asociaciones y desviaciones.

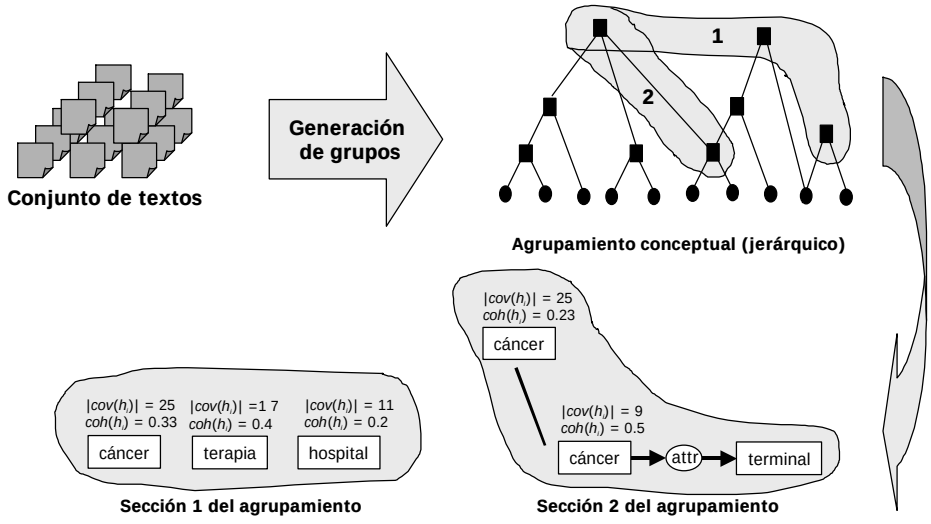


Figura 6.2 Agrupamiento conceptual de los textos

6.1.1.2 Agrupamiento de los conjuntos de prueba

El agrupamiento del conjunto A generó una jerarquía conceptual de 510 nodos; de 225 representan los artículos originales. Por su parte, el agrupamiento del conjunto B creó una jerarquía conceptual de 1272 nodos, donde 495 representan los artículos originales.

En la figura 6.3 se esboza el agrupamiento del conjunto de prueba, y se señalan sus principales nodos raíz. Por su parte, las figuras 6.4 y 6.5 presentan un acercamiento a la jerarquía conceptual e ilustran algunos grupos internos.

La figura 6.6 dibuja el agrupamiento correspondiente al conjunto de prueba B y resalta los principales nodos raíz. Las figuras 6.7 y 6.8 muestran dos grupos internos de dicho agrupamiento.

Con base en estos agrupamientos se concluye lo siguiente:

1. El agrupamiento conceptual es *altamente descriptivo*. Este tipo de agrupamiento no sólo divide los grafos conceptuales en varios grupos, también asigna una descripción del contenido a cada grupo.
2. Las descripciones de grupo van *más allá del nivel temático*. Ellas consideran conceptos que representan entidades, acciones y atributos, y también las relaciones semánticas entre estos conceptos. Un ejemplo de este tipo de descripciones es [solve]→(mnr)→[numerically], la cual indica que el grupo correspondiente integra artículos que proponen soluciones numéricas.
3. El uso de conocimientos del dominio permite encontrar más grupos (ver los resultados cuantitativos de la sección 6.2), pero además permite construir *descripciones generalizadas* de ellos. Por ejemplo, la descripción [data_structure] indica que dicho grupo se compone de artículos que hablan de matrices, conjuntos, árboles, etc.
4. El agrupamiento conceptual permite construir *grupos con traslapes*, es decir, permite clasificar en varios grupos un mismo artículo. Esta característica es im-

portante en el análisis de textos porque considera su condición multitemática. Por ejemplo, un artículo que habla sobre el diseño de algoritmos puede clasificarse con este esquema en el grupo de diseño, y a la vez en el grupo de algoritmos.

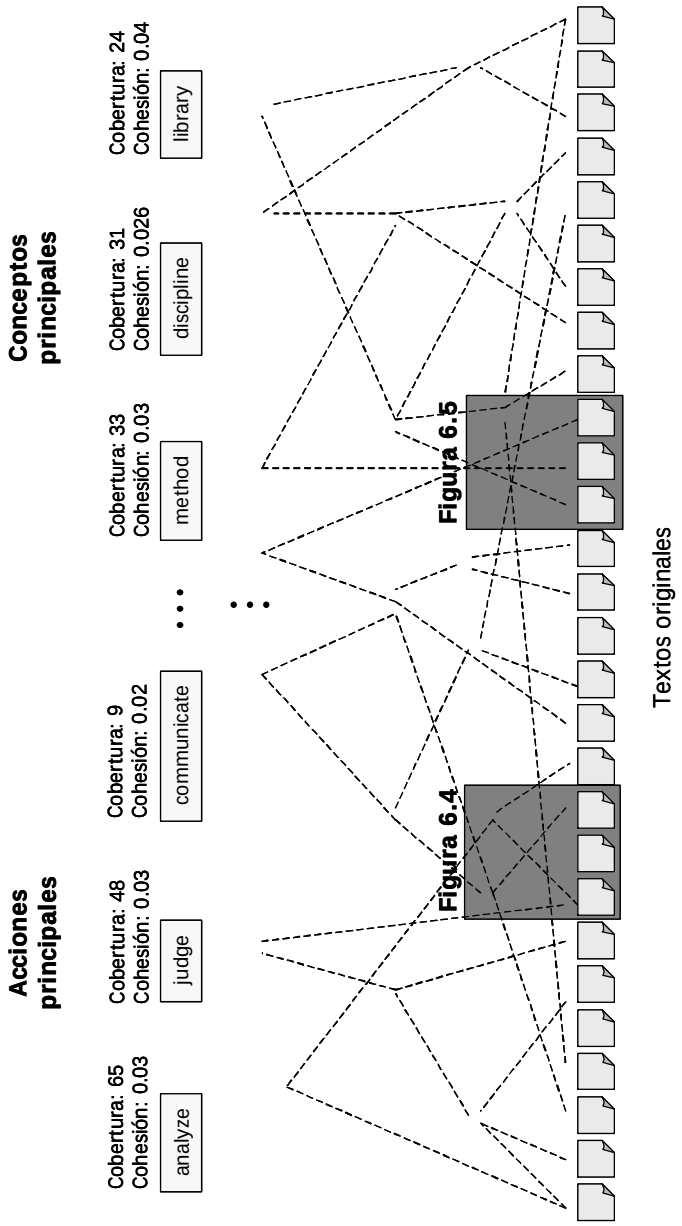


Figura 6.3 Vista General del agrupamiento del conjunto A

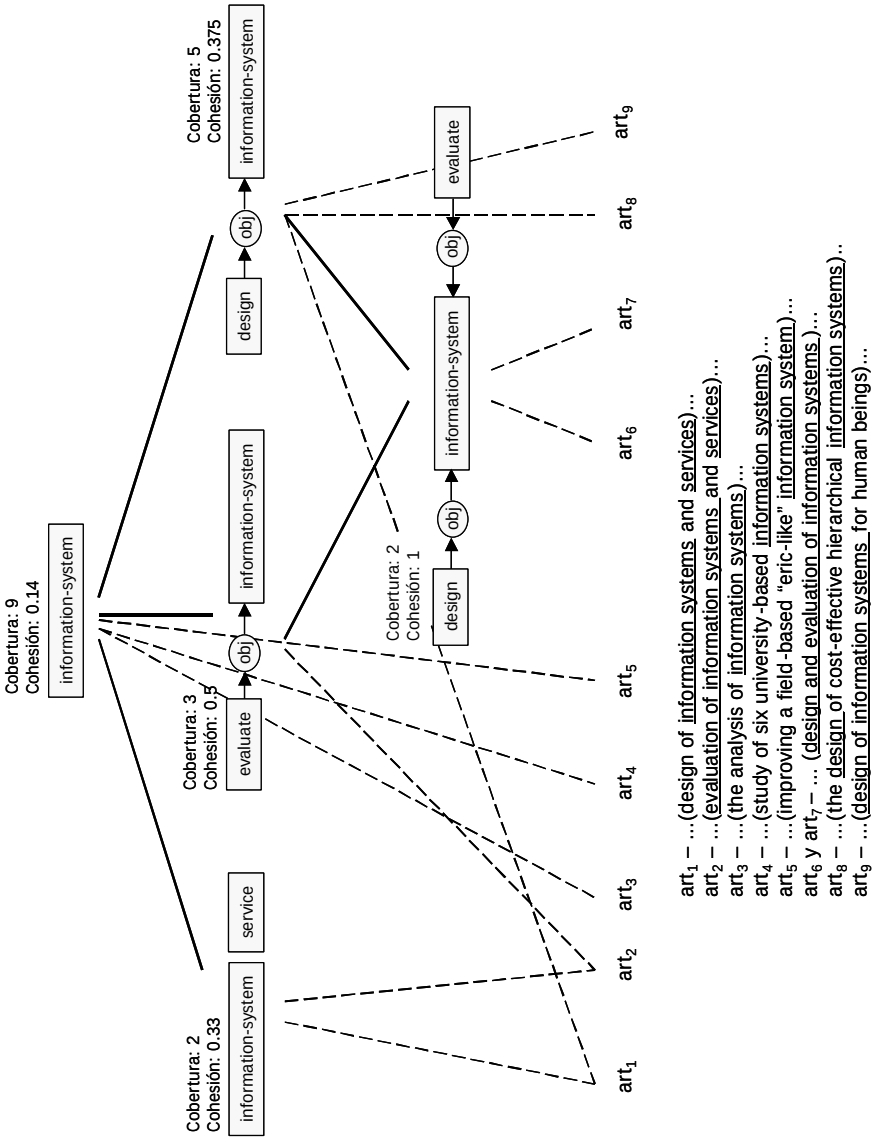


Figura 6.4 Primer acercamiento al agrupamiento del conjunto A

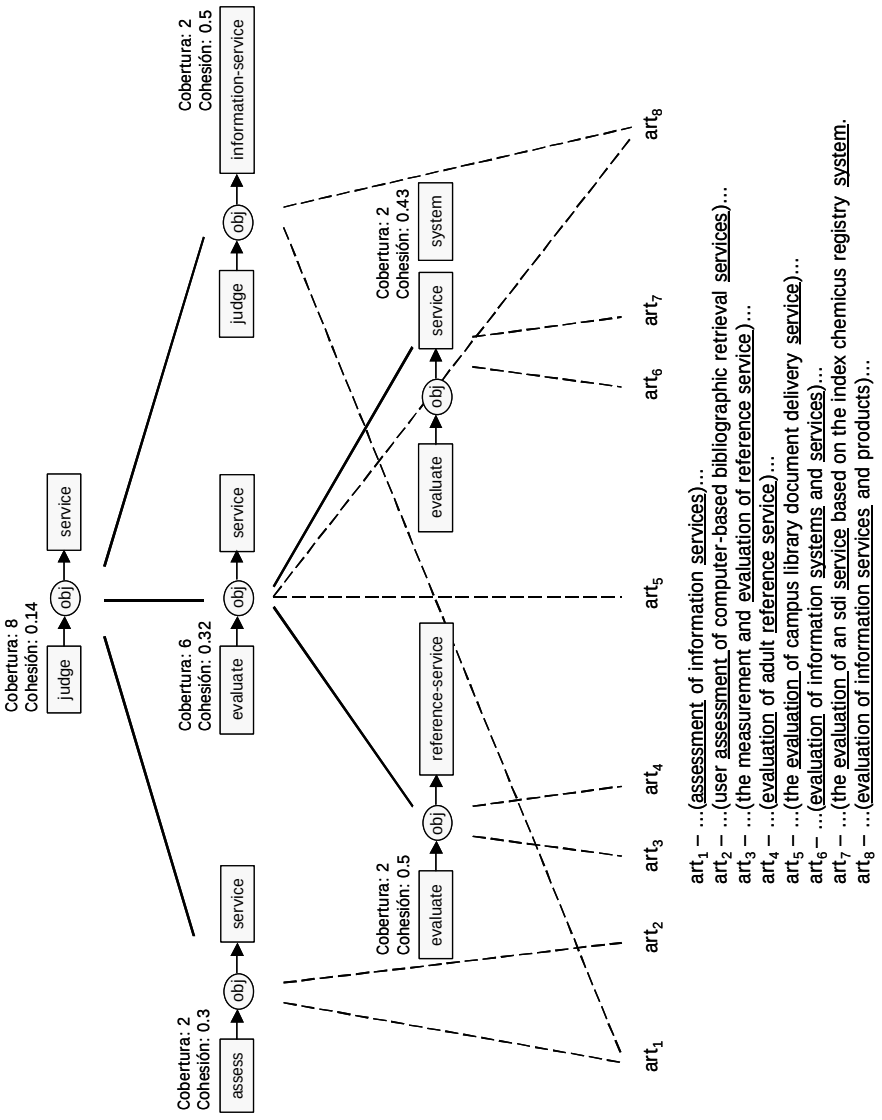


Figura 6.5 Segundo acercamiento al agrupamiento del conjunto A

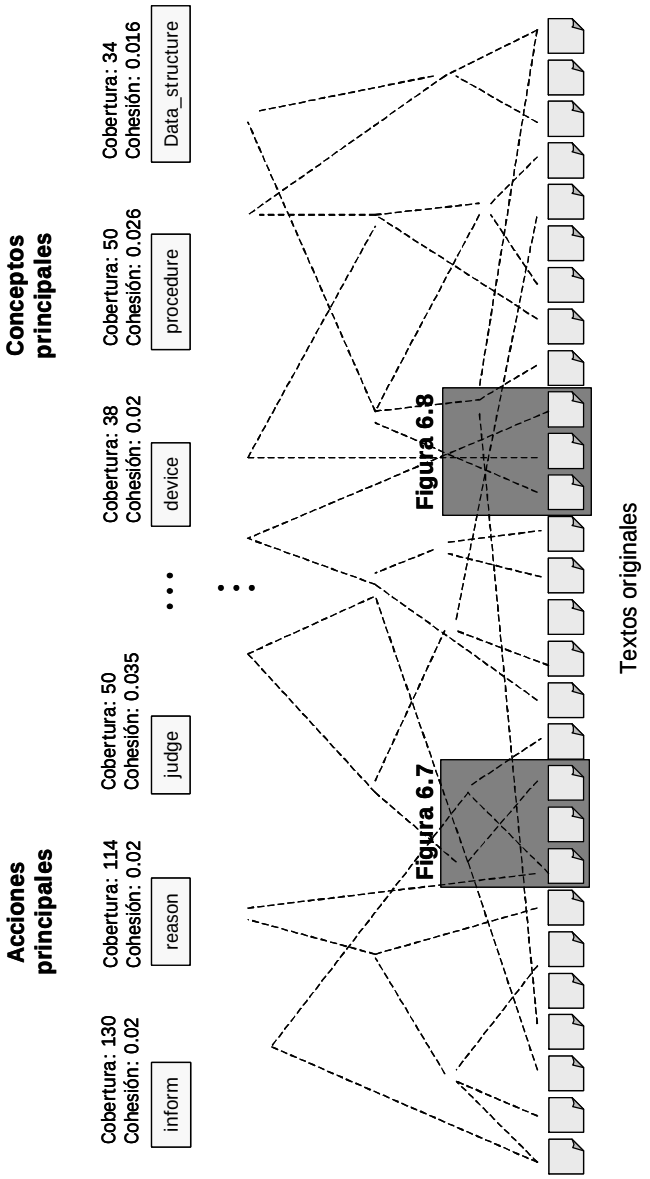


Figura 6.6 Vista General del agrupamiento del conjunto B

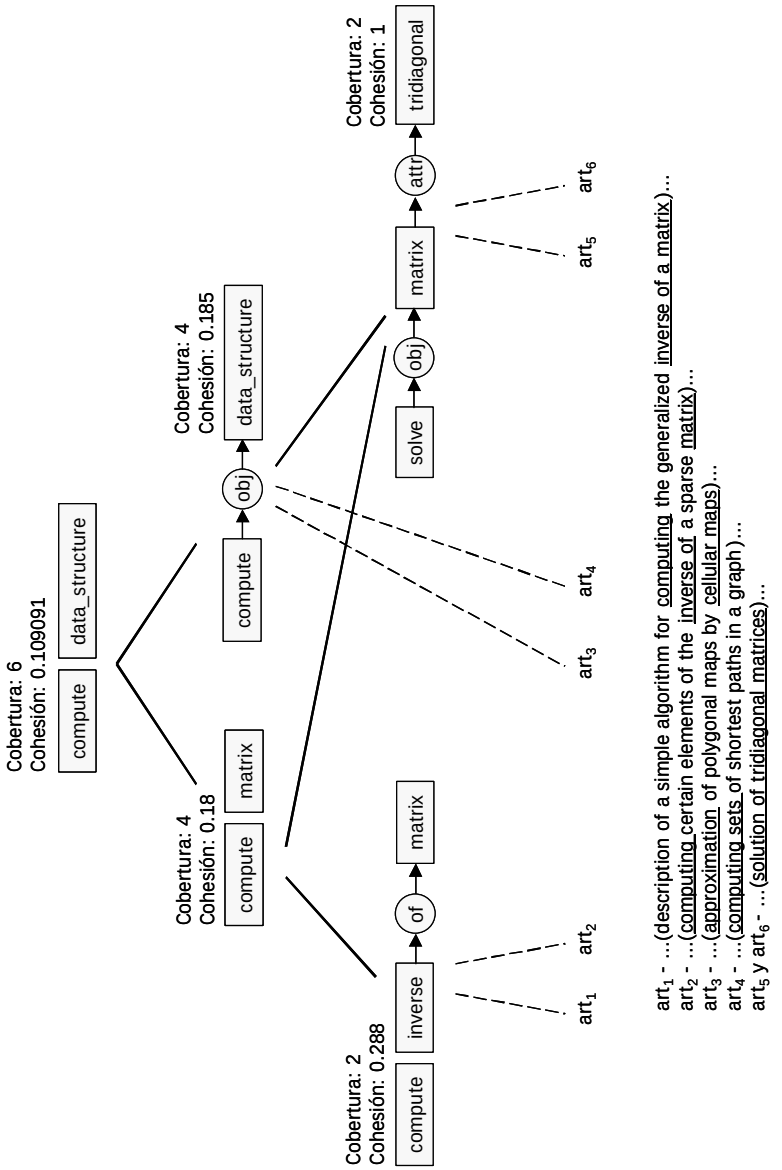


Figura 6.7 Primer acercamiento al agrupamiento del conjunto B

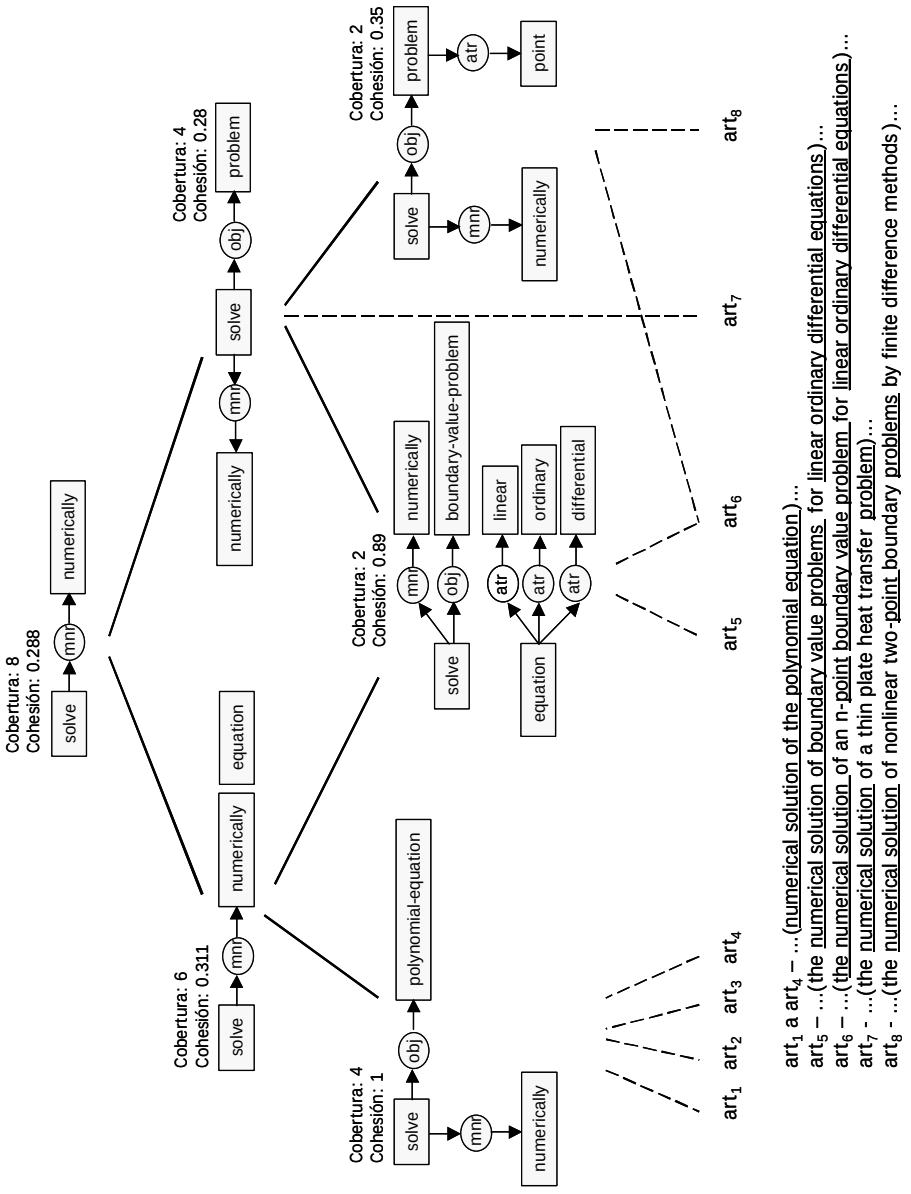


Figura 6.8 Segundo acercamiento al agrupamiento del conjunto B

6.1.2 Asociaciones y desviaciones

6.1.2.1 Descripción de los resultados

Métodos tradicionales de descubrimiento de asociaciones

El descubrimiento de asociaciones es una de las tareas más comunes de la minería de datos. Su objetivo es encontrar *reglas asociativas* de la forma $A \Rightarrow B\,(confianza / soporte)$ entre los atributos de un conjunto de datos. Un ejemplo típico de estas reglas es el siguiente:

Pañal ⇒ Cerveza (62%, 2%)

Esta regla indica que en un supermercado hipotético el 62% de las transacciones que incluyen pañales también incluyen cervezas; y además indica que el 2% del total de las transacciones incluyen ambos elementos.

En la minería de texto se intenta descubrir este mismo tipo se reglas. Allí básicamente se representan los textos como un conjunto de palabras clave y se buscan asociaciones entre éstas. Un ejemplo del tipo de reglas asociativas que se descubren con los actuales métodos de minería de texto es el siguiente:

Cáncer ⇒ Terapia (58%, 5%)

Esta regla indica que en una colección hipotética de textos médicos el 58% de los textos que hablan de algún cáncer, hablan también de una terapia; además que el 5% de los textos de la colección entera hablan sobre ambos temas.

Este tipo de reglas señalan únicamente una relación de coocurrencia de varios temas en una colección de textos, pero no indican nada sobre el contexto en que sucede dicha relación. Por ejemplo, a partir de la regla asociativa anterior es imposible determinar si los textos hablan sobre terapias para el cáncer o sobre otro tipo de terapias, por ejemplo, terapias psicológicas para los enfermos de cáncer.

Nuestro método de descubrimiento de asociaciones

El método de descubrimiento de asociaciones propuesto en este trabajo soluciona el problema contextual de las reglas asociativas. Para ello representa el contenido de los textos con grafos conceptuales en lugar de listas de palabras clave. Básicamente, este nuevo método permite descubrir reglas asociativas mucho más descriptivas, las cuales:

1. Hacen explícita la relación semántica entre los elementos participantes en la asociación.
2. Incluyen detalles sobre el contexto en el que sucede la asociación, por ejemplo, las acciones y los atributos relacionados.

Entonces, nuestro método no sólo descubre reglas asociativas como Cáncer $\Rightarrow$ Terapia (58%, 5%), también permite descubrir reglas asociativas de la forma:

[Cáncer] $\Rightarrow$ [Evaluar]$\rightarrow$(ptn)$\rightarrow$[Terapia]$\rightarrow$(for)$\rightarrow$[Cáncer] (58%, 5%)

Esta asociación indica que los textos que hablan sobre algún cáncer generalmente hablan de terapias, pero además señala que estas terapias son específicamente para el cáncer, y que en particular esta relación se presentó en un contexto de evaluación. En otras palabras, esta regla indica que cuando se habla de algún cáncer, el 68% de las veces se discute sobre evaluaciones de las terapias de cáncer.

Métodos tradicionales de detección de desviaciones

En la minería de texto, la detección de desviaciones considera dos tareas diferentes:

- Encontrar los documentos raros en un conjunto de textos.
- Encontrar los temas (palabras clave) raros en un conjunto de textos.

Los métodos que se enfocan en la primera tarea generalmente agrupan los textos de acuerdo con una métrica de distancia, y después, con base en este agrupamiento,

definen los textos aislados como desviaciones. Así pues, dada una colección de textos $C = \{\text{texto}_1, \ldots, \text{texto}_n\}$, el resultado de estos métodos es el siguiente:

$\text{texto}_i \in C$ es raro

Por su parte, los métodos que consideran la segunda tarea generalmente representan los textos como un conjunto de palabras clave (temas), y buscan entre ellas las combinaciones menos usuales. Por ejemplo, dada un conjunto hipotético de textos de medicina y una jerarquía de conceptos, estos métodos pueden detectar desviaciones como:

***P*(SIDA∈ enfermedad | terapia) es rara.**

Esto significa que entre todas las enfermedades que se consideraron (que aparecen en la jerarquía de conceptos), el SIDA es una enfermedad rara con base en el concepto de terapia. Esto significa que el concepto de terapia apareció muchas más o menos veces junto con SIDA que con el resto de las enfermedades.

En general, estos métodos son más descriptivos que los primeros porque señalan una desviación temática, y porque además sitúan esta desviación en un contexto específico.

Nuestro método de detección de desviaciones

El método para detectar desviaciones propuesto en este trabajo tiene una característica significativa: permite descubrir tanto desviaciones globales, como desviaciones locales. Esto último significa que puede detectar desviaciones con respecto a distintos subgrupos de textos de la colección.

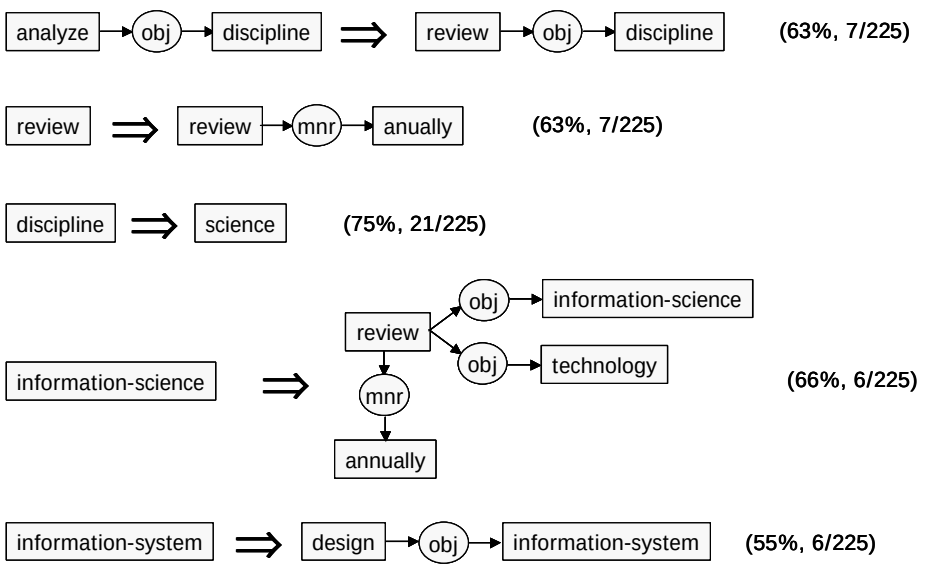
Nuestro método también mejora el nivel descriptivo de las desviaciones. Para ello representa cada desviación como una combinación de un contexto, una patrón raro, y el grado de rareza del patrón raro en dicho contexto. Además representa el

contexto y el patrón raro como grafos conceptuales que consideran acciones, atributos, entidades y sus relaciones semánticas.

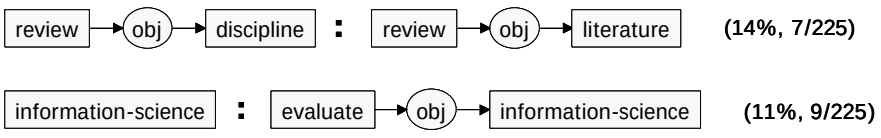
Por ejemplo, dada una colección hipotética de textos médicos, nuestro método puede descubrir desviaciones tales como:

[Cáncer]: [Cáncer]→(typ)→[Próstata] (8%, 64%)

Esta expresión indica que en el subconjunto de textos que hablan del cáncer, el cual representa un 64% de la colección, los textos que mencionan el cáncer de próstata son raros, ya que sólo significan el 8%.



(a) Algunos ejemplos de asociaciones



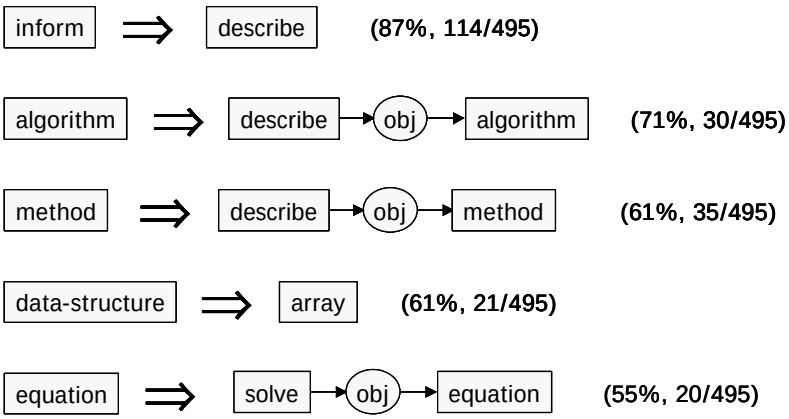
(b) Algunos ejemplos de desviaciones

Figura 6.9 Patrones descriptivos del conjunto A

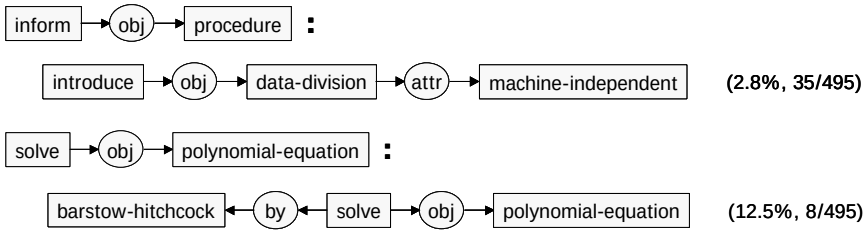
6.1.2.2 Análisis de los conjuntos de prueba

A continuación se muestran algunos patrones descubiertos en los dos conjuntos de prueba. Por ejemplo, la figura 6.9 muestra algunas asociaciones y desviaciones correspondientes al conjunto de prueba A. Estos patrones indican lo siguiente:

Una parte importante de los artículos del conjunto A se enfoca en el análisis de distintas *disciplinas*; siendo la ciencia la disciplina más analizada y la literatura la menos. Además, estos análisis son en la mayoría de las ocasiones *revisiones anuales*. Un ejemplo de esto son los artículos que realizan una revisión anual de la *ciencia de la información*.



(a) Algunos ejemplos de asociaciones



(b) Algunos ejemplos de desviaciones

Figura 6.10 Patrones descriptivos del conjunto B

En el caso de la ciencia de la información un tema importante son los *sistemas de información*. En general, los artículos que tratan este tema se concentran en el *diseño* de dichos sistemas y muy pocas veces en su evaluación.

La figura 6.10 muestra algunas asociaciones y desviaciones correspondientes al conjunto de prueba B. Estos patrones señalan lo siguiente:

Los textos del conjunto B tienen un carácter *informativo*, principalmente descriptivo. Por ejemplo, los artículos sobre métodos y algoritmos no los diseñan o evalúan, simplemente los describen.

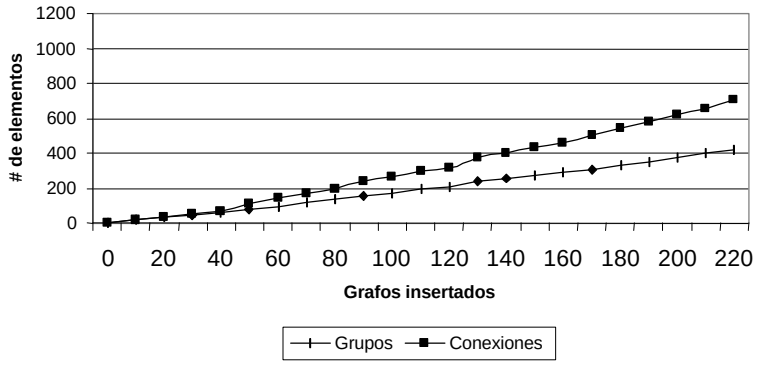
Por otra parte, este conjunto se enfoca en distintos procedimientos, principalmente en su descripción *algoritmica*. En este caso la división de datos fue el procedimiento menos estudiado.

En este conjunto también se mencionan varias estructuras de datos, principalmente *arreglos*, y se plantean varias soluciones para ecuaciones, siendo muy rara la solución de ecuaciones polinomiales por el método barstow-hitchcock.

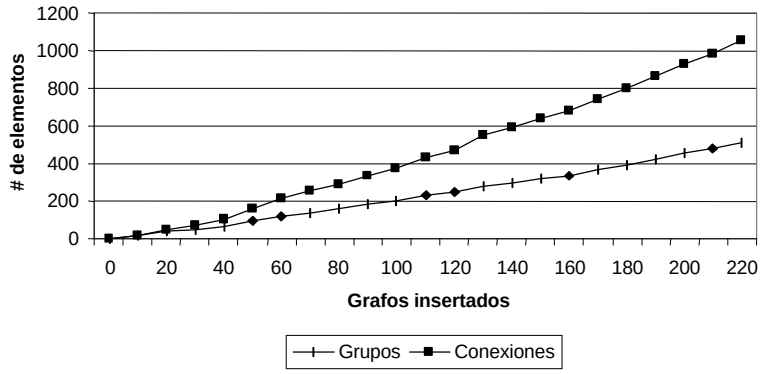
6.3 Resultados cuantitativos

6.3.1 Crecimiento del agrupamiento conceptual

Los agrupamientos conceptuales tienen características muy interesantes para los propósitos de descubrimiento de conocimiento y minería de texto, por ejemplo, son muy descriptivos y altamente estructurados. Sin embargo, su tamaño (que puede ser *exponencial* con respecto al número de grafos del conjunto) limita considerablemente su aplicación.



(a) Sin conocimiento del dominio



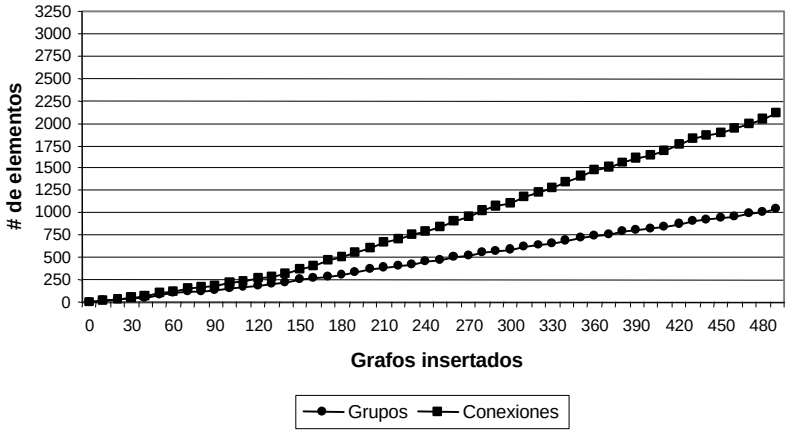
(b) Con conocimiento del dominio

Figura 6.11 Crecimiento del agrupamiento (conjunto A)

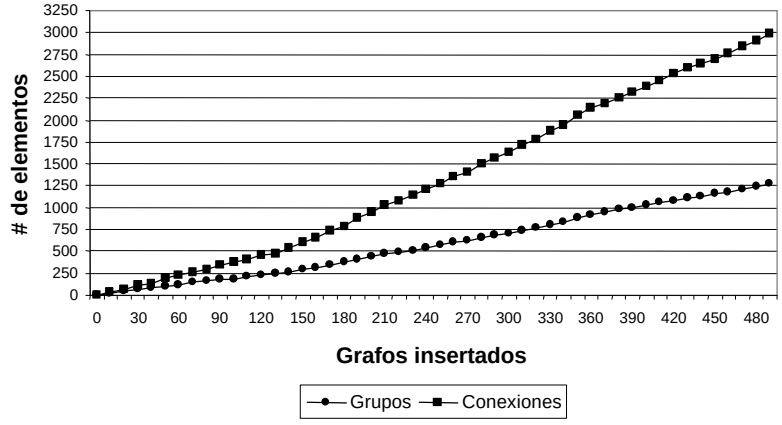
Nuestros experimentos demuestran que el agrupamiento conceptual de un conjunto de grafos conceptuales que representan el contenido de textos es *factible*. Por ejemplo, en las figuras 6.11 y 6.12 se describe el crecimiento de los agrupamientos conceptuales correspondientes a los dos conjuntos de prueba. Algunas conclusiones importantes son las siguientes:

1. El crecimiento del agrupamiento conceptual, medido en función del número de grupos y conexiones, es *casi lineal*.

   1.1. Este crecimiento casi lineal se mantiene aún cuando se emplean conocimientos del dominio.



(a) Sin conocimiento del dominio



(b) Con conocimiento del dominio

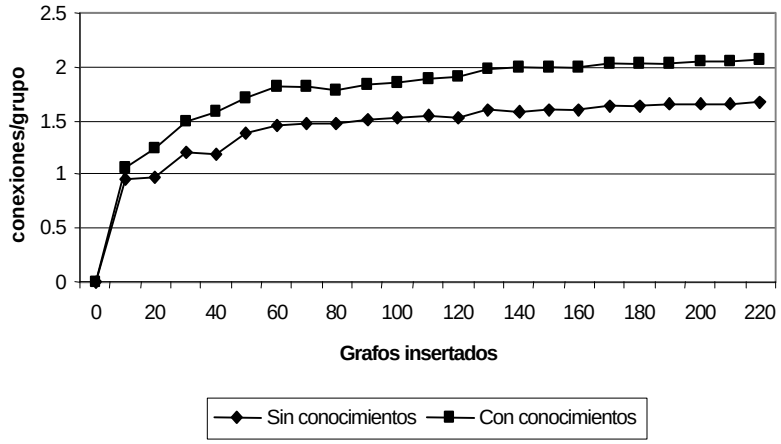
Figura 6.12 Crecimiento del agrupamiento (conjunto B)

2. El impacto de los conocimientos del dominio es mayor en las conexiones que en los grupos. Intuitivamente esto significa que se logran formar grupos más grandes y más homogéneos (más interconectados), pero no muchos más grupos.
3. Los grupos y las conexiones crecen inicialmente muy parecido, pero después, conforme se insertan más grafos en el agrupamiento, las conexiones crecen más rápidamente. Este comportamiento sucede porque en un principio, cuando el agrupamiento no existe, cada nuevo grafo genera nuevos grupos, pero después

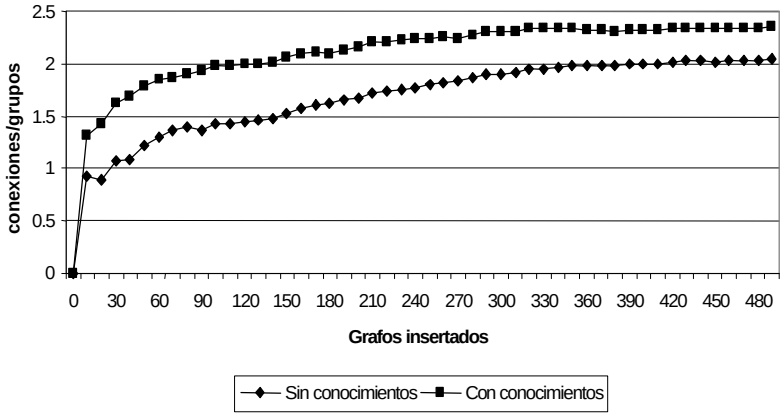
cuando el agrupamiento es mayor, cada nuevo grafo sólo se inserta en algunos grupos existentes.

6.3.2 Densidad de conexiones

Otra característica interesante de los agrupamientos conceptuales (principalmente del proceso de su construcción) es la *densidad de conexiones*, es decir, el número de conexiones por grupo.



(a) Análisis del conjunto A



(b) Análisis del conjunto B

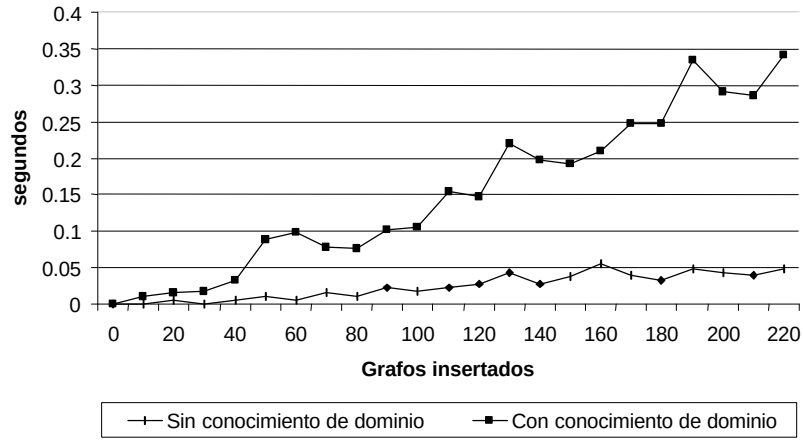
Figura 6.13 Densidad de conexiones

La figura 6.13 muestra la variación de la densidad de conexiones durante el proceso de construcción de los agrupamientos correspondientes a los dos conjuntos de prueba. Con base en esta figura deducimos lo siguiente:

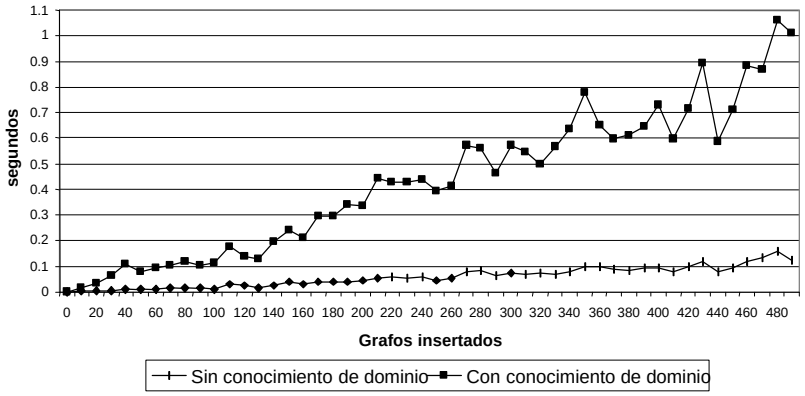
1. La densidad de conexiones se incrementa conforme se insertan los grafos conceptuales en el agrupamiento.
   1.1. La razón de aumento de la densidad de conexiones es en principio muy elevada, pero se estabiliza conforme se insertan más grafos en el agrupamiento. Este comportamiento sucede porque inicialmente, cuando el agrupamiento no existe, cada nuevo grafo genera nuevos grupos, pero después cuando el agrupamiento es mayor, cada nuevo grafo solamente se inserta o se conecta con algunos grupos existentes.
2. La densidad de conexiones aumenta cuando se usa conocimiento del dominio. Este incremento es casi constante a través de todo el proceso de construcción.

6.3.3 Tiempo de construcción

Las gráficas de crecimiento del agrupamiento y de densidad de conexiones exponen algunas ventajas del uso de conocimiento del dominio en la construcción del agrupamiento de los grafos conceptuales. Básicamente, estas gráficas muestran que este conocimiento permite encontrar grupos más grandes y más homogéneos (mejor interconectados).



(a) Análisis del conjunto A



(b) Análisis del conjunto B

Figura 6.14 Tiempo de construcción del agrupamiento

Estas ventajas tienen un costo principal: el *tiempo de construcción* del agrupamiento. En la figura 6.14 se muestran los tiempos de construcción de los agrupamientos de los conjuntos de prueba. Allí se observa que el uso de conocimiento del dominio afecta considerablemente la rapidez del análisis de los grafos.

A pesar del aumento en el tiempo de análisis de los grafos conceptuales, la construcción de su agrupamiento conceptual sigue siendo factible. Por ejemplo, el tiem-

po de inserción del grafo 495 del conjunto B en la jerarquía conceptual necesito solamente de un segundo.

Analizando la figura 6.14 se determina que el tiempo de inserción de un grafo conceptual en el agrupamiento, cuando no se usa conocimiento del dominio, es casi estable. Esto último nos permite suponer que el incremento en el tiempo de construcción cuando se usa conocimiento del dominio se origina de una mala implementación del sistema (principalmente de la jerarquía de conceptos); y que por lo tanto, un esfuerzo en esta dirección permitirá mejorar considerablemente el funcionamiento del método de minería de texto propuesto.

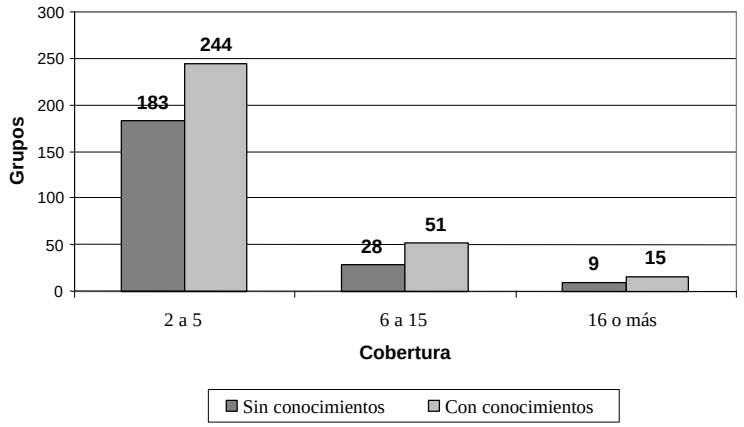
6.3.4 Atributos cuantitativos del agrupamiento conceptual

Un agrupamiento conceptual es una estructura jerárquica que consiste de grupos y conexiones. Los grupos tienen tres atributos básicos que caracterizan completamente el agrupamiento conceptual: cobertura, descripción y cohesión (referirse a la sección 5.1 del capítulo 5).

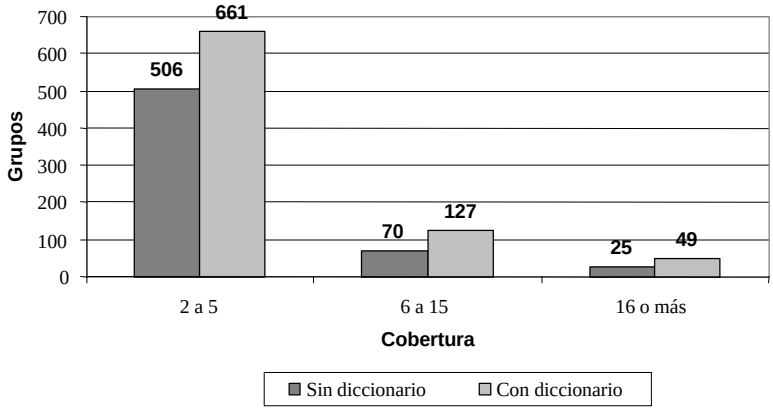
A continuación se presenta el análisis de la *cobertura* y la *cohesión* –atributos cuantitativos– de los agrupamientos correspondientes a los dos conjuntos de prueba. En la figura 6.15 se analiza el valor de cobertura de los grupos, y en la figura 6.16 el valor de cohesión. Algunas conclusiones importantes son las siguientes:

1. La mayoría de los grupos descubiertos son pequeños, y tienen una cobertura menor a 10. Este comportamiento se mantuvo aún cuando se empleo conocimiento del dominio.
2. El conocimiento del dominio incrementa, aunque no exageradamente en términos absolutos, el número total de grupos del agrupamiento conceptual.
   2.1. El impacto del conocimiento del dominio es mayor en los grupos grandes que en los pequeños. Por ejemplo, mientras los grupos con cobertura mayor a seis se duplicaron, los grupos con una cobertura menor sólo se incrementaron aproximadamente un 30%.

3. La mayoría de los grupos tienen una *cohesión baja*, es decir, están conformados por grafos que no son muy parecidos entre si. Este fenómeno es más notorio en los grupos grandes, por ejemplo, todos los grupos con más de 16 grafos tienen una cohesión menor a 0.25.
4. El conocimiento del dominio no modifica –considerablemente– la cohesión de los grupos. Incluso, contrario a nuestra intuición, se observó que el conocimiento del dominio causó una disminución en la cohesión promedio de los grupos.

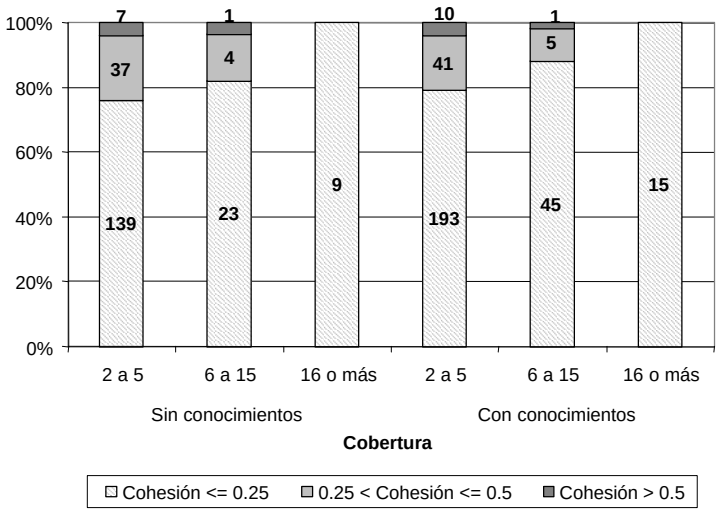


(a) Análisis del conjunto A

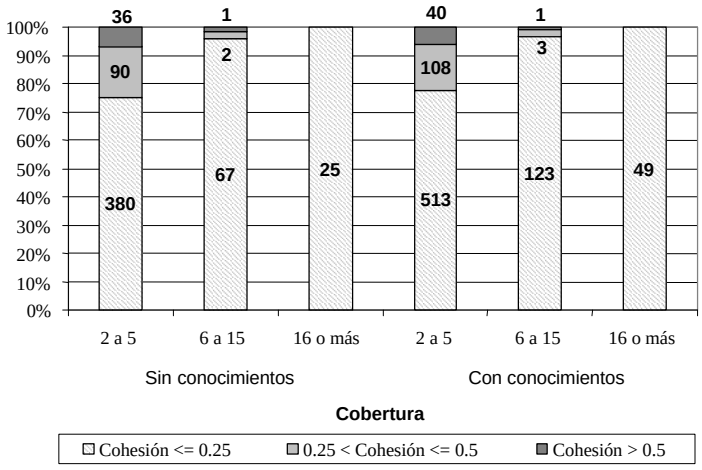


(b) Análisis del conjunto B

Figura 6.15 Cobertura de los grupos



(a) Análisis del conjunto A



(b) Análisis del conjunto B

Figura 6.16 Cohesión por nivel de cobertura

Las dos observaciones relacionadas con la cohesión de los grupos tienen una justificación. En primer lugar, la construcción del agrupamiento conceptual no se basa en una medida de distancia o semejanza entre los grafos, sino en la descripción cualitativa de dicha semejanza. Así pues, los grupos no se componen necesariamente de grafos muy parecidos, sino de grafos que, aún siendo muy diferentes, comparten algunos elementos.

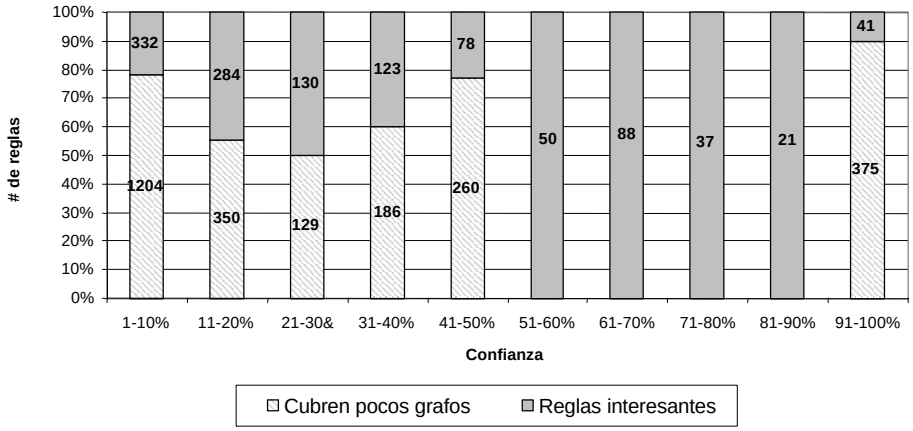
En segundo lugar, el conocimiento del dominio incrementa la semejanza entre los grafos de un mismo grupo, pero además, y en la mayoría de las ocasiones, permite detectar semejanzas pequeñas antes no consideradas, y por lo tanto construir nuevos grupos, que aunque pocos homogéneos, expresan algún concepto frecuente en el conjunto de grafos.

6.3.5 Asociaciones y desviaciones

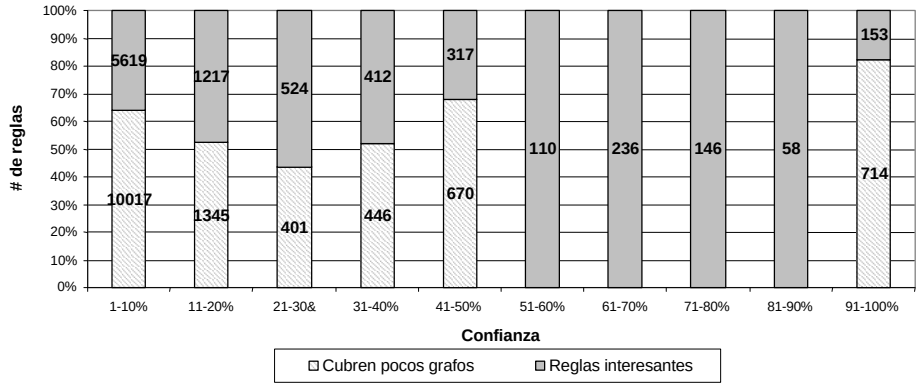
En la sección 6.2 se presentaron algunas asociaciones y desviaciones descubiertas en los dos conjuntos de prueba. En esta sección se describen algunas características cuantitativas de estos patrones.

Por ejemplo, en la figura 6.17 se analizan los valores de confianza y soporte de las asociaciones descubiertas, y en la figura 6.18 se estudia la influencia de la selección del umbral *m* en la detección de las desviaciones. Algunas conclusiones son las siguientes:

1. El número de posibles reglas asociativas es muy grande; su *crecimiento es exponencial* con respecto al número de grafos de la colección.
   1.1. Solamente muy pocas de estas reglas son interesantes, es decir, sólo pocas reglas tienen una confianza alta y un soporte aceptable.
2. Las reglas con una confianza baja, en su mayoría también tuvieron un soporte muy pequeño. Por su parte, las reglas con una confianza alta tuvieron siempre un soporte aceptable. Estas últimas reglas, que son las más interesantes, creemos que se originan de la parte media y alta del agrupamiento conceptual.



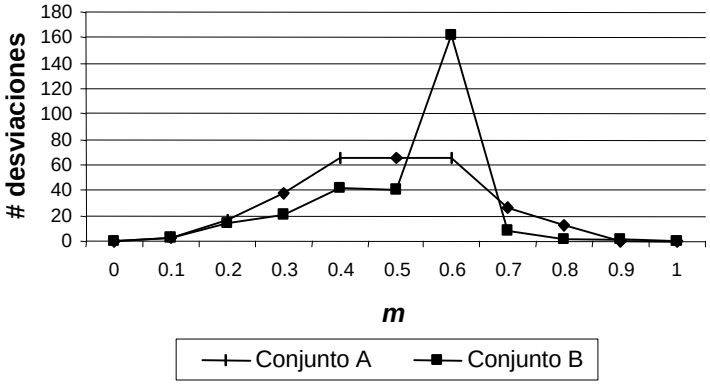
(a) Conjunto A



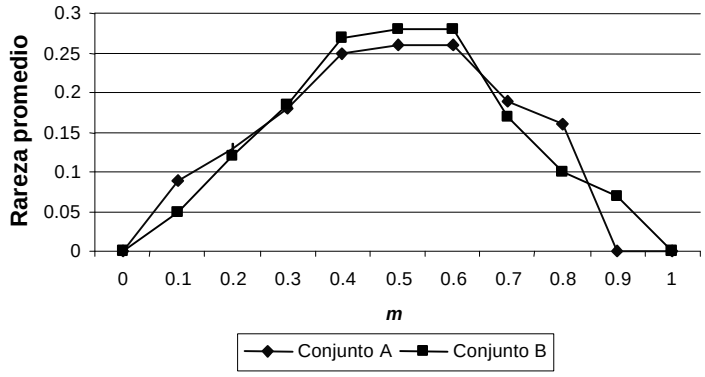
(b) Conjunto B

Figura 6.17 Análisis cuantitativo de las reglas asociativas

3. Muchas reglas con una confianza de 100% no son interesantes porque tienen soporte muy bajo. Estas reglas se obtuvieron de los nodos más bajos del agrupamiento.
4. El conocimiento del dominio no tiene un efecto notorio en las características cuantitativas de las asociaciones (esto no se ilustra en ninguna figura 6.17).



(a) *m* y el número de desviaciones



(b) *m* y la rareza de las desviaciones

Figura 6.18 Análisis cuantitativo de las desviaciones

5. La detección de desviaciones es altamente dependiente del parámetro *m.*

   5.1. Los valores medios de *m* producen muchas desviaciones, pero que indican patrones no muy raros, por su parte los valores medios de *m* permiten encontrar sólo pocas pero más raras desviaciones.

6. De acuerdo con las gráficas de la figura 6.18, y con base en los resultados cualitativos, consideramos que los valores mas adecuados de *m* son: $0.6 < m \leq 0.8$.

Capítulo 7

Conclusiones

Este capítulo se organiza de la siguiente manera. En primer lugar discutimos brevemente la aportación general de nuestra investigación. En segundo lugar describimos detalladamente varias aportaciones específicas, haciendo énfasis en las características principales de los métodos de comparación de grafos conceptuales, de agrupamiento conceptual, de descubrimiento de asociaciones y de detección de desviaciones. En la tercera parte señalamos algunas limitaciones de nuestro método y finalmente, en la cuarta parte, listamos algunos trabajos futuros interesantes para realizar.

Conclusiones

La mayoría de los métodos actuales de minería de texto utilizan representaciones sencillas del contenido de los textos, por ejemplo listas de palabras clave o tablas de datos. Estas representaciones son relativamente fáciles de construir a partir de los textos, pero impiden representar varios detalles de su contenido. Como consecuencia, los resultados de estos sistemas, es decir, los patrones que con ellos se descubren, son *poco descriptivos* y de *nivel temático*.

Una idea generalizada para mejorar la expresividad de los resultados de los métodos de minería de texto consiste en emplear representaciones de los textos más complejas que las palabras clave, es decir, representaciones que consideren más tipos de elementos textuales. Siguiendo este idea, en este trabajo propusimos un método para hacer *minería de texto a nivel detalle*. Este método tiene la capacidad de usar grafos conceptuales para representar el contenido de los textos, y el potencial para trasladar los resultados, es decir, los patrones descubiertos, del actual nivel temático a un nivel mucho más descriptivo.

Algunas contribuciones importantes de esta investigación son las siguientes:

- Se planteó, por primera vez, el uso una "representación semántica", en especifico *grafos conceptuales*, en las tareas de minería de texto.[1]
- Se demostró que el uso de los grafos conceptuales, y en general de las representaciones semánticas, en la minería de texto es *factible*, pero sobre todo, beneficioso para *mejorar el nivel descriptivo de resultados*.
- Se diseñó una nueva aproximación para realizar minería de texto. Para ello se adaptaron algunos métodos de *comparación* y *agrupamiento* de grafos conceptua-

[1] Anteriormente, los grafos conceptuales sólo se usaban como representación del contenido de los textos en la recuperación de información.

les para las tareas propias de minería de texto; y se diseñaron nuevas estrategias para *descubrir asociaciones* y *detectar desviaciones* en un conjunto de grafos conceptuales.

Así pues, esta investigación contribuyó al estado del arte de diversas áreas del conocimiento, entre las que destacan la minería de texto y la teoría de grafos conceptuales.

A continuación se describen las principales contribuciones específicas, trabajos futuros, y limitaciones del método de minería de texto propuesto en este trabajo de investigación.

7.1 Contribuciones específicas

- **Un método flexible para la comparación de grafos conceptuales**

Actualmente existen varios métodos para comparar dos grafos conceptuales (referirse a la sección 4.1 del capítulo 4). Estos métodos, que en su mayoría provienen de la recuperación de información, *no* son apropiados para la minería de texto por dos razones:

1. Porque no construyen una descripción adecuada de la semejanza entre los grafos.
2. Porque no son suficientemente flexibles para adaptarse a distintos intereses de análisis de los usuarios.

El método propuesto en este trabajo aminora estos problemas. Algunas de sus características son:

- Describe *cualitativa* y *cuantitativamente* la semejanza entre los dos grafos conceptuales.
- Utiliza más adecuadamente la *información estructural* de los grafos conceptuales (sin necesidad de separar las relaciones *n*-arias en un conjunto de relaciones bina-

rias), permitiendo entre otras cosas identificar semejanzas exclusivamente entre los conceptos cuando así ocurren.

- Detecta semejanzas *parciales* (por partes) y semejanzas a diferentes *niveles de generalización* entre los dos grafos conceptuales.
- Permite visualizar las semejanzas entre los dos grafos desde *diferentes perspectivas*, y además seleccionar la mejor de ellas de acuerdo con los *intereses del usuario*.

En general, el método propuesto realiza la comparación de los grafos en dos etapas: una etapa de casamiento y una de medición de la semejanza. A continuación se discuten las principales contribuciones de cada etapa.

Casamiento de los grafos

Nosotros describimos la semejanza entre dos grafos conceptuales por medio de un conjunto de *traslapes*, donde cada traslape es un conjunto máximo de generalizaciones comunes compatibles. Esta característica es una gran diferencia con otros métodos, porque éstos generalmente describen la semejanza entre los dos grafos a través del conjunto de todas sus generalizaciones comunes, permitiendo así información duplicada.

La construcción de los traslapes también incluye algunos aspectos novedosos. En ella se emplearon *técnicas de minería de datos*, en especifico, un algoritmo muy usado para detectar los conjuntos de elementos frecuentes en una base de datos. La aplicación de este algoritmo no disminuye la complejidad del casamiento de los grafos, que sigue siendo exponencial, pero sí disminuye el número de comparaciones realizadas. Esto último acelera ligeramente el proceso de casamiento de los grafos conceptuales.

Finalmente, otro aspecto importante del método de casamiento es la posibilidad de *detectar diferentes traslapes* para un mismo par de grafos conceptuales. Cada uno

de estos traslapes representa una manera diferente, pero precisa, de visualizar la semejanza entre dichos grafos.

Medición de la semejanza

La medida propuesta para la semejanza entre dos grafos conceptuales se basa en el *coeficiente de Dice*, que es una medida tradicional de la semejanza, pero también incorpora aspectos relacionados con la estructura bipartita de los grafos conceptuales. Básicamente, nuestra medida de semejanza tiene las siguientes propiedades:

- Extiende el coeficiente de Dice, de tal forma que considere la pérdida de información causada por *el casamiento no-exacto* entre dos conceptos.
- Considera *distintos pesos* para los diferentes tipos de conceptos (entidades, acciones y atributos); más aún permite que el usuario establezca estos pesos de acuerdo con sus intereses de análisis.
- Plantea la semejanza total como una *combinación de una semejanza conceptual y una semejanza estructural*; también permite que el usuario establezca, de acuerdo con sus intereses, la importancia relativa de estos dos tipos de semejanzas parciales.

Finalmente, la medida de semejanza permite seleccionar uno de los traslapes, el que produce la mayor medida, como la descripción final de la semejanza entre los dos grafos conceptuales.

- **Un método para agrupar *conceptualmente* un conjunto de grafos conceptuales**

Los grafos conceptuales son una representación de la información muy usada en los sistemas basados en conocimiento. Allí se desarrollaron previamente dos métodos para agrupar conceptualmente de un conjunto de grafos conceptuales.

Al igual que estos métodos, nuestro método emplea una estrategia de aprendizaje no supervisado que construye *incrementalmente* el agrupamiento conceptual, pero

adicionalmente incorpora algunas características que lo hacen más atractivo para los propósitos de la minería de texto. Por ejemplo:

1. Maneja adecuadamente la *información estructural* de los grafos conceptuales (sin necesidad de separar las relaciones *n*-arias en un conjunto de relaciones binarias), lo que permite no sólo construir mejores descripciones de grupo, también descubrir más grupos.
2. Emplea *conocimiento del dominio establecido por el usuario*; lo que permite construir grupos con descripciones generalizadas, y a la vez enfocar el agrupamiento en los conceptos más interesantes para el usuario.
3. Utiliza la *medida de semejanza* entre los grafos conceptuales para enfatizar los intereses del usuario durante la construcción del agrupamiento. Esta característica permite agrupar el mismo conjunto de grafos conceptuales de distintas maneras, o desde distintos puntos de vista.

Algunas otras características distintivas de este método de agrupamiento conceptual son las siguientes:

- El agrupamiento resultante *no depende del orden de inserción* de los grafos. Esta característica es una diferencia significativa con respecto a los métodos tradicionales de análisis incremental.
- El agrupamiento resultante *no contiene información redundante*, es decir, no existen dos nodos con la misma descripción. Esto es muy importante porque mantiene el crecimiento del agrupamiento casi lineal, a pesar que teóricamente pueda llegar a ser exponencial.
- El agrupamiento resultante es *un agrupamiento con traslapes*, es decir, un grafo conceptual puede pertenecer a más de un grupo. Esta característica permite considerar adecuadamente el aspecto multitemático de los textos.

Agrupamiento conceptual reducido

A pesar de que el agrupamiento conceptual de los grafos expresa solamente sus regularidades (elementos comunes a dos o más grafos de la colección), generalmente este agrupamiento es demasiado grande para ser considerado una síntesis o resumen de dicho conjunto.

Con base en esto último se propuso un método para identificar las principales regularidades de un conjunto de grafos conceptuales, y para construir un *agrupamiento reducido* a partir de ellas. La construcción de este nuevo agrupamiento tiene las siguientes características:

- Es un proceso *dirigido por el usuario*. En él, el usuario establece los lineamientos generales –por ejemplo mínima cobertura y mínima cohesión– para definir un grupo como interesante.
- Es un proceso *rápido* que aprovecha el agrupamiento completo de los grafos conceptuales para localizar fácilmente los grupos más interesantes.

Así pues, nuestro método de agrupamiento reducido es interesante porque permite construir un *resumen jerárquico* y *personalizado* de un conjunto dado de grafos.

- **Un método para descubrir asociaciones en un conjunto de grafos conceptuales**

Los métodos conocidos de descubrimiento de asociaciones consideran solamente datos estructurados. Por ejemplo, en la minería de texto el descubrimiento de asociaciones se realiza sobre listas de palabras clave o sobre los conceptos de una tabla con información extraída de un conjunto de textos.

En este trabajo se propuso un método para descubrir asociaciones en un conjunto de textos representados con grafos conceptuales (ya no datos estructurados, sino semiestructurados). Este método tiene las siguientes características sobresalientes:

1. Construye reglas asociativas que consideran *información estructural* del contenido de los textos, por ejemplo, algunas relaciones semánticas entre los conceptos. Esto implica una gran *mejora del nivel descriptivo* de las reglas asociativas.
2. Aprovecha el *agrupamiento conceptual* de los grafos para identificar y construir las reglas asociativas. Básicamente utiliza el agrupamiento jerárquico de los grafos como un índice –construido previamente– de la colección; lo que facilita y agiliza el descubrimiento de las asociaciones.
3. Descubre asociaciones a diferentes *niveles de generalización*.

- **Un método para detectar desviaciones en un conjunto de grafos conceptuales**

La mayoría de los métodos para la detección automática de desviaciones en conjuntos de datos son de tipo estadístico o basados en distancia (ver secciones 2.1.2 y 5.4.1). A diferencia de ellos, el método aquí propuesto se basa en el concepto de *regularidad*, y por lo tanto, detecta las desviaciones desde una *perspectiva conceptual*.

Básicamente, este método considera que todo elemento común a varios grafos de la colección es una característica representativa, y define como raro cualquier grafo que no tiene ninguna de estas características. Algunas de sus características más novedosas son las siguientes:

1. Permite detectar los grafos raros de un conjunto dado, y además sus patrones descriptivos. Estos patrones, que llamamos *desviaciones*, señalan las características comunes y exclusivas de los grafos raros. Así pues, las desviaciones son una manera *resumida* de expresar las rarezas del conjunto de grafos.
2. Considera *desviaciones locales*, es decir, desviaciones con respecto a diferentes contextos –subconjuntos– de la colección de grafos conceptuales. Esta característica permite visualizar las desviaciones desde varias *perspectivas*, y también detectar desviaciones a diferentes *niveles de generalización*.

3. Aprovecha el *agrupamiento conceptual* de los grafos para identificar los patrones raros. Esto no sólo facilita y agiliza la detección de las desviaciones, también permite la detección de desviaciones locales.

7.2 Limitaciones del método

El método de minería de texto propuesto en este trabajo tiene dos problemas que limitan considerablemente su aplicación. Estos problemas y sus limitaciones relacionadas se describen a continuación.

Primer problema: El casamiento de los grafos conceptuales es *exponencial* con respecto al número de conceptos comunes entre los dos grafos.

Principales limitaciones:

- Análisis de *grafos conceptuales relativamente pequeños*, con unas cuantas decenas de nodos concepto.

Esta limitación indica que nuestro método de minería de texto es más adecuado para analizar grafos conceptuales que representen algunas partes de los textos con un significado especial (por ejemplo, descripciones de eventos u opiniones sobre algún tema) o los detalles más importantes de su contenido, que para analizar grafos que intenten representar completamente el contenido de los textos.

- El uso de *jerarquías de conceptos relativamente pequeñas*.

Esta limitación se origina por el siguiente efecto: entre más grande es la jerarquía de conceptos, más correspondencias –elementos comunes– entre los grafos pueden detectarse, y por lo tanto, mayor es la complejidad del análisis.

Una consecuencia importante de esta limitación es la pérdida de información, es decir, el uso de jerarquía pequeñas puede ocasionar que no se detecten semejanzas posiblemente interesantes entre los grafos.[2]

Segundo problema: La transformación automática de los textos en grafos conceptuales no es una tarea sencilla.

Principales limitaciones:

- Análisis de *textos cortos* o sólo de *algunas de sus partes*.

Esta limitación es una consecuencia directa de los problemas de los métodos actuales de procesamiento de textos (por ejemplo, métodos de análisis sintáctico y semántico). Básicamente, ella implica que nuestro método de minería de texto es más adecuado para el análisis de textos cortos o de algunas partes de los textos con un significado especial.

- El análisis de textos de *un solo dominio*.

La transformación de un texto en grafo conceptual, como todo proceso que involucra el análisis semántico de los textos, requiere de cierto conocimiento del dominio. Esto último significa que es necesario un considerable esfuerzo humano para trasladar el mecanismo de transformación de los textos en grafos conceptuales, y por ende nuestro método de minería de texto, de un dominio a otro.[3]

---

[2] Los experimentos demuestran que el uso de jerarquías de conceptos grandes y generales producen muchos resultados no interesantes para el usuario, mientras que el uso de jerarquías pequeñas sólo implica la perdida de muy poca información interesante. Mas bien estas jerarquías actúan como un filtro de descubrimientos no interesantes.

[3] Esta limitación no es exclusiva de nuestro método, en general la construcción de representaciones a nivel concepto (véase el capítulo 2) es dependiente del dominio, y también dependiente del propio tipo de concepto a extraerse.

7.3 Rumbos de investigación posterior

En este trabajo propusimos un esquema general para hacer minería de texto usando grafos conceptuales, aunque nuestros esfuerzos se concentraron en la etapa de descubrimiento. Por ello, gran parte del trabajo futuro que se presenta a continuación considera el desarrollo de las demás etapas del proceso de minería de texto usando grafos conceptuales.

1. Desarrollar un método para transformar los textos en grafos conceptuales.

   Este método deberá ser *flexible*, de tal forma que permita transformar textos de distintos dominios a grafos conceptuales sin la necesidad de un gran esfuerzo humano. También deberá ser *adaptivo*, de tal forma que aprenda las distintas maneras de comunicar la información que se desea extraer y convertir a grafo conceptual.

2. Diseñar otros métodos para descubrir más patrones descriptivos en un conjunto de grafos conceptuales.

   Estos métodos deberán considerar varias tareas de descubrimiento que complementen las actuales, por ejemplo: el análisis de tendencias, la detección de contradicciones y la clasificación de textos.

3. Desarrollar varios mecanismos de postprocesamiento.

   En este sentido deberán crearse algunos criterios para evaluar el nivel de utilidad de los patrones descubiertos, y también algunas interfaces para visualizar e interpretar dichos resultados.

Otras líneas de investigación que se desprenden de este trabajo consideran el uso de los métodos propuestos en este trabajo en otras tareas de procesamiento de textos. Por ejemplo:

- Aplicar el método de comparación de grafos conceptuales en la *búsqueda de información* para manejar adecuadamente consultas complejas que consideren detalles del contenido de los textos.

- Aplicar los nuestros métodos de análisis de grafos conceptuales en la *minería semántica de la web*.

Lista de Publicaciones

Artículos en revistas

[1] Manuel Montes y Gómez, Alexander Gelbukh, Aurelio López-López. *Mining the news: trends, associations, and deviations*. **Computación y Sistemas**, Vol 5, No. 1, Julio-Septiembre 2001, ISSN 1405-5546, pp.14-25.

[2] Manuel Montes y Gómez, Alexander Gelbukh, Aurelio López-López, Ricardo Baeza-Yates. *Un Método de Agrupamiento de Grafos Conceptuales para Minería de Texto*. **Procesamiento de Lenguaje Natural**, Vol. 27, Septiembre 2001, ISSN 1135-5948, pp. 115-122.

Capítulos en Libros

[3] Manuel Montes y Gómez, Alexander Gelbukh, Aurelio López-López. *Document intentions expressed in titles. Extraction, representation, and possible use*. In: A. de Albornoz Bueno *et al*. (eds.), **Selected Works 1997-1998**, Instituto Politécnico Nacional, Centro de Investigación en Computación, ISBN 970-18-3427-5, 1999, pp. 322-330. (Una versión revisada de la publicación 5)

[4] Manuel Montes y Gómez, Alexander Gelbukh, Aurelio López-López. *Information Retrieval with the Extra-Topical Information extracted from Document Titles*, In: A. Guzmán-Arenas and F. M. Menchaca-García (Eds.), **Selected Papers 1999**, Instituto Politécnico Nacional, Centro de Investigación en Computación, ISBN 970-18-4250-2, 2000, pp.147-152. (Una versión revisada de la publicación 8)

Congresos Internacionales

[5] Aurelio López-López and Manuel Montes-y-Gómez, *Nominalization in Titles: A Way to Extract Document Details*. Proc. of the **Simposium Internacional de Computación CIC'98**, November 11 - 13, 1998, ISBN 970-18-1916-0, Mexico City, pp. 396-404.

[6] Manuel Montes y Gómez, Aurelio López-López, Alexander Gelbukh. *Text Mining as a Social Thermometer*. Proccedings of the Text Mining Workshop at the **16th International Joint Conference on Artificial Intelligence IJCAI'99**, Stockholm, Sweden, July 31 – August 6, 1999, pp. 103-107.

[7] Manuel Montes y Gómez, Aurelio López-López, Alexander Gelbukh. *Extraction of Document Intentions from Titles*. Proccedings of the Text Mining Workshop at the **16th International Joint Conference on Artificial Intelligence IJCAI'99**, Stockholm, Sweden, July 31 – August 6, 1999, pp. 101-102.

[8] Manuel Montes-y-Gómez, Aurelio López-López, Alexander Gel-bukh. *Document Title Patterns in Information Retrieval*. In Václav Matoušek et al. (Eds.). Proc. of the **2nd International Workshop on Text, Speech and Dialogue TSD-99**, Plzen, Czech Republic, September 13-17, 1999. Lecture Notes in Artificial Intelligence 1692, ISBN 3-540-66494-7, **Springer-Verlag**, pp. 364–367.

[9] Manuel Montes-y-Gómez, Alexander Gelbukh, Aurelio López-López. *Detecting the Dependencies of a Peak News Topic*. Proccedings of the **Simposium Internacional de Computación CIC'99**, November 15-19, 1999, CIC, IPN, Mexico City, ISBN 970-18-3697-9, pp. 360-366.

[10] Manuel Montes-y-Gómez, Alexander Gelbukh, Aurelio López-López. *Comparison of Conceptual Graphs*. Proc. In: O. Cairo, L.E. Sucar, F.J. Cantu (eds.) MICAI 2000: Advances in Artificial Intelligence. MICAI-2000, **1st Mexican International Conference on Artificial Intelligence**, Acapulco, Mexico, April 2000. Lecture Notes in Artificial Intelligence N 1793, ISSN 0302-9743, **Springer-Verlag**, pp. 548-556.

[11] Manuel Montes-y-Gómez, Aurelio López-López, Alexander Gelbukh. *Information Retrieval with Conceptual Graph Matching*. Proc. of the **11th International Conference and Workshop on Data-base and Expert Systems Applications**, **DEXA-2000**, Greenwich, England, Septem-ber 4-8, 2000. Lecture Notes in Computer Science N 1873, ISSN 0302-9743, **Springer-Verlag**, pp. 312–321.

[12] Manuel Montes-y-Gómez, Aurelio López-López, Alexander Gelbukh. *Finding Correlative Associations among News Topics*, Proc. of the **Conference on Intelligent Text Processing and Computational Linguistics, CICLing-2001**, February 18–24, 2001, Mexico City. Lecture Notes in Computer Science N 2004, ISBN 3-540-41687-0, **Springer-Verlag**, pp. 521–522.

[13] Manuel Montes-y-Gómez, Alexander Gelbukh, Aurelio López-López, Ricardo Baeza-Yates. *Flexible Comparison of Conceptual Graphs*. Proc. of the **12th International Conference on Database and Expert Systems Applications, DEXA 2001**, September 2001, Munich, Germany. Lecture Notes in Computer Science 2113. ISBN 3-540-42527-6, **Springer-Verlag**, pp. 102-111.

[14] Manuel Montes-y-Gómez, Alexander Gelbukh, Aurelio López-López. *A Statistical Approach to the Discovery of Ephemeral Associations among News Topics*. Proc. of the **12th International Conference on Database and Expert Systems Applications, DEXA 2001**, September 2001, Munich, Germany. Lecture Notes in Computer Science 2113. ISBN 3-540-42527-6, **Springer-Verlag**, pp. 491-500.

[15] Manuel Montes-y-Gómez, Alexander Gelbukh, Aurelio López-López. *Discovering Ephemeral Associations among news topics*. Proc. of the Workshop of Adaptive Text Mining, at the **17th International Joint Conference on Artificial Intelligence IJCAI'2001**, Seattle, WA, August 2001, pp. 25 – 30.

[16] Manuel Montes-y-Gómez, Alexander Gelbukh, Aurelio López-López. *Discovering Association Rules in Semi-structured Data Sets*. Proc. of the Workshop on Knowledge Discovery from Distributed, Dynamic, Heterogeneous, Autonomous Data and Knowledge Sources, at the **17th International Joint Conference on Artificial Intelligence IJCAI'2001**, Seattle, WA, August 2001, pp. 26 – 31.

[17] Manuel Montes-y-Gómez, Alexander Gelbukh, Aurelio López López, Ricardo Baeza-Yates. *Text Mining with Conceptual Graphs*. Proc. of the Mini Symposium on Natural Language Processing and Knowledge Engineering, at the **2001 IEEE International Conference on Systems, Man & Cybernetics**, October 2001, Tucson, Arizona, USA. ISSN 0-7803-7089-9.

[18] Manuel Montes y Gómez, *Minería de Texto: Un Nuevo Reto Computacional*, Memoria del **3er Taller Internacional de Minería de Datos MINDAT-2001**, Octubre 2001, Universidad Panamericana, Ciudad de México. CD-Rom ISBN 970-18-6971-2.

[19] Manuel Montes y Gómez, Alexander Gelbukh, Aurelio López López. *Detección de los patrones raros en un conjunto de datos semiestructurados*. Memorias del **Congreso Internacional de Computación CIC'2001**, Noviembre 2001, México, D.F. ISBN 970-18-7054-9, pp. 250-260.

Congresos Nacionales

[20] Manuel Montes-y-Gómez, Aurelio López-López, Alexander Gelbukh. *El reto de la minería de texto a nivel detalle*. Memorias del **II Encuentro de Investigación del Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)**, Noviembre 2001, Tonanzintla, Puebla. pp. 225-228.

Informes Técnicos

[21] Manuel Montes-y-Gómez, *Minería de Texto: Estado del Arte y Aplicaciones*. Informe Técnico, Serie Verde, No. 26, August 1999. **CIC-IPN**, ISBN 970-18-3516-6.

[22] Manuel Montes-y-Gómez, Aurelio López-López, Alexander Gelbukh, Grigory Sidorov, Adolfo Guzmán-Arenas. *Text mining: new techniques and applications*. Informe Técnico, Serie Azul, No. 34, August 1999. **CIC-IPN**, ISBN 970-18-3512-3.

[23] Manuel Montes-y-Gómez, Aurelio López-López, Alexander Gelbukh. *Information retrieval with conceptual graph matching*. Informe Técnico, Serie Azul, No 78, August 2000. **CIC-IPN**, ISBN 970-18-5430-6.

[24] Manuel Montes-y-Gómez, Alexander Gelbukh, Aurelio López-López, Ricardo Baeza-Yates. *Flexible Comparison of Conceptual Graphs*. Informe Técnico, Serie Azul, No. 102, Junio 2001. **CIC-IPN**, ISBN 970-18-6976-1.

[25] Manuel Montes-y-Gómez, Alexander Gelbukh, Aurelio López-López. *A Statistical Approach to the Discovery of Ephemeral Associations among News Topics*. Informe Técnico, Serie Azul, No. 103, Junio 2001. **CIC-IPN**, ISBN 970-18-6977-X.

[26] Manuel Montes-y-Gómez, Alexander Gelbukh, Aurelio López-López, Ricado Baeza-Yates. *Minería de Texto empleando Grafos Conceptuales*. Informe Técnico, Serie Azul, No. 105, Junio 2001. **CIC-IPN**, ISBN 970-18-6979-6.

Otras Pláticas (*no publicadas*)

[27] *Minería de Texto*, Seminario de Sistemas de Toma de Decisiones, Data Warehouse y Minería de Datos, **Centro de Investigación en Computación CIC-IPN**, México, D.F., Mayo 1999.

[28] *Mesa redonda: Minería de Datos*, **Foros "Computación de la Teoría a la Práctica"**. México, D.F., Mayo 1999,

[29] *Panorama General de la Minería de Texto*, **1er. Taller de Minería de Datos CIC MIDAT-99**, Centro de Investigación en Computación (CIC-IPN), México, D.F., Julio 1999.

[30] *Minería de Texto empleando la Semejanza entre las Estructuras Semánticas*, Seminario de Investigación, **Centro de Investigación en Computación CIC-IPN**, México, D.F., Mayo 2000.

[31] *Minería de Texto empleando Grafos Conceptuales*, Seminario de Investigación, Departamento de Ciencias de la Computación, Facultad de Ciencias Físicas y Matemáticas, **Universidad de Chile**, Santiago, Chile, Julio 2000.

[32] *Descubrimiento de conocimiento en conjuntos de datos semiestructurados*, Seminario de Investigación, **Centro de Investigación en Computación CIC-IPN**, México, D.F., Octubre 2001.

Referencias

[1] Aggrawal and Yu (1999), Data Mining Techniques for Associations, Clustering and Classification, 3rd Pacific-Asia Conference PAKDD on Methodologies for Knowledge Discovery and Data Mining, Lecture Notes in Artificial Intelligence 1574, Springer 1999.

[2] Agrawal, Rakesh and Srikant (1994), Fast Algorithms for Mining Associattion Rules, Proc. of the 20th. VLDB Conference, Santiago de Chile, 1994.

[3] Agrawal, Arning, Bollinger, Mehta, Shafer, Srikant (1996), The Quest Data Mining System, Proc. of the 2nd International Conference on Knowledge Discovery in Databases and Data Mining, Portland, Oregon, August, 1996.

[4] Agrawal, Bayardo Jr. and Srikant (1999), Athena: Mining-based Interactive Management of Text Databases, IBM Research Report RJ10153, July 1999.

[5] Ahonen-Myka, Heinonen, Klemettinen, and Verkamo (1997a), Appliying Data Mining Techniques in Text Analysis, Report C-1997-23, Department of Computer Science, University of Helsinki, 1997.

[6] Ahonen, Heinonen, Klemettinen, and Verkamo (1997b), Mining in the Phrasal Frontier, Proc. of the 1st Conference on Principles of Knowledge Discovery and data Mining (PKDD'97), Lecture Notes in Artificial Intelligence 1263, Springer 1997.

[7] Ahonen-Myka (1999a), Finding All Frequent Maximal Sequences in Text, Proc. of the 16th International Conference on Machine Learning ICML-99, Workshop on Machine Learning in Text Data Analysis, Ljubljana 1999.

[8] Ahonen-Myka (1999b), Knowledge Discovery in Document by Extracting Frequent Word Sequences, Invited article for the special issue of Library Trends on Knowledge Discovery in Databases, 1999.

[9] Ahonen-Myka, Heinonen, Klemettinen, and Verkamo (1999), Finding Co-occurring Text Phrases by Combining Sequence and Frequent Set Discovery, Proc. of 16th International Joint Conference on Artificial Intelligence IJCAI-99, Workshop on Text Mining: Foundations, Techniques and Applications, Stockholm, Sweden, August 1999.

[10] Alexandrov, Gelbukh and Makagonov (2000), On Metrics for Keyword-Based Document Selection and Classification, Proc. of the Conference on Intelligent Text Processing and Computational Linguistics CICLing-2000, Mexico City, Mexico, February 2000.

[11] Allan, Papka and Lavrenko (1998), On-line new Event Detection and Tracking, Proc. of the 21st ACM-SIGIR International Conference on Research and Develepement in Information Retrieval, August 1998.

[12] Apte, Damerau and Weiss (1998), Text Mining with Decision Rules and Decision Trees, Conference on Automated Learning and discovery, june11-13, 1998.

[13] Arning, Agrawal and Raghavan (1996), A Linear Method for Deviation Detection in Large Databases, Proc. of the 2nd International Conference on Knowledge Discovery in Databases and Data Mining, 1996.

[14] Barnett and Lewis (1994), Outliers in Statistical Data, New York: John Wiley & Sons, 1994.

[15] Baud, Rassinoux and Scherrer (1992), Natural language processing and semantical representation of medical texts, Meth Inform Med 31:117-25, 1992

[16] Biébow and Chaty (1993), A Comparison between Conceptual Graphs and KL-ONE, Conceptual Graphs for knowledge Representation, First International Conference on Conceptual Structures, ICCS´93, 1993.

[17] Bourcier and Rajman (1994), Interactional Semantics for Legal Case-Based Knowledge, à paraître fin 1994.

[18] Bournaud and Ganascia (1996), Conceptual Clustering of Complex Objects: A Generalization Space based Approach, Lecture Notes in Artificial Intelligence 954, Springer, 1996.

[19] Bournaud and Ganascia (1997), Accounting for Domain Knowledge in the Construction of a Generalization Space, Lectures Notes in AI (1257), Springer-Verlag, 1997.

[20] Breunig, Kriegel, Ng and Sander (1999), OPTICS-OF: Identifying Local Outliers, Proceedings of the PKDD-1999, Lecture Notes in Artificial Intelligence 1704, Springer, 1999.

[21] Chakravarthy and Haase (1995), NETSERF: Using Semantic Knowledge to Find Internet Information Archives, Proc. of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, 1995.

[22] Chein, ed. (1996), Revue d'Intelligence artificielle, Special Issue on Conceptual Graphs, vol. 10, no. 1, 1996.

[23] Clifton and Cooley (1999), TopCat: Data Mining for Topic Identification in a Text Coupus, Proceedings of the PKDD-1999, Lecture Notes in Artificial Intelligence 1704, Springer, 1999.

[24] Cohen and Hirsh (1998), Joins that Generalize: Text Clasification using WHIRL, Proc. of the Fourth International Conference on Knowledge Discovery and Data Mining, 1998.

[25] Eklund, Ellis and Mann, eds. (1996), Conceptual Structures: Knowledge Representation as Interlingua, Lecture Notes in AI 1115, Springer-Verlag, Berlin, 1996.

[26] Ellis and Lehmann (1994), Exploiting the Induced Order on Type-Labeled Graphs for fast Knowledge Retrieval, Conceptual Structures: Current Practices, William M. Tepfenhart, Judith P. Dick and John F. Sowa Eds., Lecture Notes in Artificial Intelligence 835, Springer-Verlag 1994.

[27] Ellis, Levinson, Rich and Sowa, eds. (1995), Conceptual Structures: Applications, Implementation, and Theory, Lecture Notes in AI 954, Springer-Verlag, Berlin, 1995.

[28] Fargues, Landau, Dogourd and Catach (1986), Conceptual Graphs for Semantics and Knowledge Processing, IBM Journal of Research and Development, 30:1, 70-70-89, 1986.

[29] Fayyad, Piatetsky-Shapiro and Smyth (1996a), Knowledge Discovery and Data Mining: Towards a Unifying Framwork, Proc. of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, Agust 2-4, 1996.

[30] Fayyad, Piatetsky-Shapiro, Smyth and Uthurusamy (1996b), Advances in Knowledge Discovery and Data Mining, Cambridge, MA: MIT Press, 1996.

[31] Feldman and Dagan (1995), Knowledge Discovery in Textual databases (KDT), Proc. of the 1st International Conference on Knowledge discovery (KDD_95), pp.112-117, Montreal, 1995.

[32] Feldman and Hirsh (1996), Mining Associations in Text in the Presence of Background Knowledge, Proc. of the 2nd International Conference on Knowledge Discovery (KDD-96), pp. 343-346, Portland, 1996.

[33] Feldman, Klösgen, Yehuda, Kedar and Reznikov (1997), Pattern Based Browsing in Document Collections, Proc. of the 1st Conference on Principles of Knowledge Discovery and data Mining (PKDD'97), Lecture Notes in AI, Springer Verlag, Norway, 1997.

[34] Feldman, Fresko, Hirsh, Aumann, Liphstat, Schler, Rajman (1998a), Knowledge Management: A Text Mining Approach, Proc. of the 2nd Int. Conf. on Practical Aspects of Knowledge Management (PAKM98), 9.1-9.10, Basel, Switzerland, October29-30, 1998.

[35] Feldman, Fresko, Kinar, Lindell, Liphstat, Rajman, Schler, Zamir (1998b), Text Mining at the Term Level, Proc. of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98), Nantes, France, September 23-26, 1998.

[36] Feldman, Aumann, Zilberstein, Ben-Yehuda (1998c), Trend Graphs: Visualazing the Evolution of Concept Relationships in Large Document Collections, Proc. of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98), Lecture Notes in Artificial Intelligence 1510, September 1998.

[37] Feldman, Aumann, Fresko, Lipshtat, Rosenfeld, Schler (1999), Text Mining via Information Extraction, Proceedings of the PKDD-1999.

[38] Fisher (1987), Knowledge Acquisition via Incremental Conceptual Clustering, Machine Learning, 2, 1987.

[39] Fujino, Arimura and Arikawa (2000), Discovering Unordered Phrase Association Patterns for Text Mining, Proc. of the 4th Pacific-Asia Conference on Kwokledge Discovery and Data Mining PAKDD-2000, Lecture Notes in Artificial Intelligence 1805, Springer 2000.

[40] Ganter and Mineau, eds. (2000), Conceptual Structures: Logical, Linguistic, and Computational Issues, Lecture Notes in AI 1867, Springer-Verlag, Berlin, 2000.

[41] Gelbukh, Sidorov and Guzmán-Arenas (1999), A Method of Describing Document Contents through Topic Selection, Proc. of the Workshop on Text Mining: Foundations, Techniques and Applications, Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99), Stockholm, Sweden, August 1999.

[42] Gelfand, Wulfekuhler, Punch III (1998), Automated Concept Extraction from Plain Text, Conference on Automated Learning and discovery, 1998.

[43] Genest and Chein (1997), An Experiment in Document Retrieval using Conceptual Graphs, Conceptual structures: Fulfilling Peirce´s Dream. Lecture Notes in artificial Intelligence 1257, Springer 1997.

[44] Gennari, Langley and Fisher (1989), Models of Incremental Concept Formation, Artificial Intelligence, 40, 1989.

[45] Gibert and Córtes (1998), Clustering based on Rules and Knowledge Discovery in Ill-Structured Domains, Computación y Sistemas, Vol. 1, Num. 4, 1998.

[46] Girardi and Ibrahim (1994), A Similarity Measure for Retrieving Software Artifacts, Proc. of the sixth Internaltional Conference on Software Engineering and Knowledge Engineering (SEKE-94), Latvia, June 1994.

[47] Godin, Mineau and Missaoui (1995), Incremental Structuring of Knowledge Bases, International KRUSE Symposium, August 11-13, Santa Cruz, California, 1995.

[48] Guzmán (1996), Uso y Diseño de Mineros de Datos, Soluciones Avanzadas, Num. 34, 1996.

[49] Guzmán (1998), Finding the main Themes in a Spanish Document, Expert Systems with Applications, Vol. 14, pp 139-148, 1998.

[50] Han and Kamber (2001), Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2001.

[51] Hearst (1999), Untangling Text Data Mining, Proc. of ACL'99: The 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999.

[52] Huibers, Ounis and Chevallet (1996), Conceptual Graph Aboutness, Conceptual Structures: Knowledge Representation as Interlingua. Peter W. Elkland, Gerard Ellis, Graham Mann Eds., Lecture Notes in Artificial Intelligence, Springer, 1996.

[53] Hull (1998), Text Mining the Web: Extracting Chemical Compound Names, Conference on Automated Learning and discovery, june11-13, 1998.

[54] Jiang and Conrath (1999), From Object Comparison to Semantic Similarity, Proc. of the Pacific Association for Computational Linguistics PACLING-99, Waterloo, Canada, 1999.

[55] Khoo (1995), Automatic Identification of Causal Relations in Text and their Use for Improving Precision in Information Retrieval, Unpublished PhD Dissertation, Syracuse University, 1995.

[56] Khoo (1997), The Use of Relation Matching in Information Retrieval, LIBRES: Library and Information Science Research, Electronic Journal ISSN 1058-6768, Volume 7 Issue 2; 1997.

[57] Knorr and Ng (1998), Algorithms for Mining Distance-based Outliers in Large Datasets, Proc. of the International Conference on Very Large Data Bases (VLDB'98), Newport Beach, CA, 1997.

[58] Kodratoff (1999), Knowledge Discovery in Texts: A Definition and Applications, Proc. of the 11th International Symposium on Foundations of Intelligent Systems (ISMIS-99), 1999.

[59] Lagus, K., Honkela, T., Kaski, S., and Kohonen, T. (1999), WEBSOM for Textual Data Mining, Artificial Intelligence Review, volume 13, issue 5/6, pages 345-364, December 1999.

[60] Landau, Feldman, Aumann, Fresko, Lindell, Liphstat, Zamir (1998), TextVis: An Integrated Visual Environment for Text Mining, Proc. of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery, PKDD-1998.

[61] Larsen and Aone (1999), Discovering Topic Hierarchies through Document Clustering: Use of NLP-based Features and their Effectiveness, Proc. of the Workshop on Text Mining: Foundations, Techniques and Applications, Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99), Stockholm, Sweden, August 1999.

[62] Lent, Agrawal, Srikant (1997), Discovering Trends in Text Databases, Proc. of the 3rd Int'l Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, California, August 1997.

[63] Lin (1998), An Information-Theoretic Definition of Similarity, Proc. of the International Conference on Machine Learning, Madison, Wisconsin, 1998.

[64] Lin, Shin, Chen, Ho, Ko and Huang (1998), Extracting Classification Knowledge of Internet Documents with Mining Term Associations: A semantic Approach, Proceedings of SIGIR'98, Melbourne, Australia, 1998.

[65] López-López (1995), Beyond Topicality: Exploting the Metadiscourse of Abstracts to Retrive Documents with Analogous Features, Unpublished PhD Dissertation, Syracuse University, 1995.

[66] López-López and Myaeng (1996), Extending the Capabilities of Retrieval Systems by a Two Level Representation of Content, Proc. of the 1st Australian Document Computing Symposium, 1996.

[67] Lukose, Delugach, Keeler, Searle and Sowa, eds. (1997), Conceptual Structures: Fulfilling Peirce's Dream, Lecture Notes in AI 1257, Springer-Verlag, Berlin, 1997.

[68] Martínez, Beltrán, Guzmán and Ruiz Shulcloper (1998), CLASITEX+: A Tool for knowledge Discovery from Texts, Proc. of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98), Lecture Notes in Artificial Intelligence 1510, September 1998.

[69] Mauldin (1991), Conceptual Information Retrieval - A Case Study in Adaptive Partial Parsing, Kluwer, Boston, 1991.

[70] Merkl (1997), Exploration of Document Collections with Self-Organizing Maps: A Novel Approach to Similarity Representation, Proc. of the 1st Conference on Principles of Knowledge Discovery and data Mining (PKDD'97), Norway, 1997.

[71] Metzler, Noreault, Richey and Heidorn (1984), Dependency Parsing for Information Retrieval, Research and Development in Information Retrieval, Proc. of the Third Joint BCS and ACM symposium, 1984.

[72] Metzler and Haas (1989), The Constituent Object Parser: Syntactic Structure Matching for Information Retrieval, ACM Transactions on Information Systems, 7(3), 1989.

[73] Michalski (1980), Knowledge Acquisition thorough Conceptual Clustering: A Theoretical Framework and Algorithm for Partitioning Data into Conjunctive Concepts, International Journal of Policy Analysis and Information Systems, Vol. 4, 1980.

[74] Mineau, Moulin and Sowa, eds. (1993), Conceptual Graphs for Knowledge Representation, Lecture Notes in AI 699, Springer-Verlag, Berlin, 1993.

[75] Mineau and Godin (1995), Automatic Structuring of Knowledge Bases by Conceptual Clustering, IEEE Transactions on Knowledge and Data Engineering, 7(5), 1995.

[76] Möller (1997), CLASSITALL: Incremental and Unsupervised Learning in the Dia-MoLE Framework,Workshop Notes of the ECML / MLnet Workshop on Empirical Learning of Natural Language Processing Tasks, April 26, 1997.

[77] Montes-yGómez (1998), Extracción de Información de los Títulos de los Documentos, Tesis de Maestría, INAOE, 1998-

[78] Montes-y-Gómez, López-López and Gelbukh (1999a), Text Mining as a Social Thermometer, Proc. of the Workshop on Text Mining: Foundations, Techniques and Applications, Sixteenth International Joint Conference on Artificial Intelligence (IJCAI-99), Stockholm, Sweden, August 1999.

[79] Montes-y-Gómez, Gelbukh and López-López (1999b), Detecting the Dependencies of a Peak News Topic, Memorias del Congreso Internacional de Computación CIC-99, México D.F., Noviembre 1999.

[80] Montes-y-Gómez, López-López, Gelbukh, Sidorov and Guzmán-Arenas (1999c), Text Mining: New Techniques and Applications, Informe Técnico, No. 34, Serie Azul, Centro de Investigación en Computación, IPN, Agosto 1999.

[81] Montes-y-Gómez, López-López and Gelbukh (1999d), Document Title Patterns in Information Retrieval, Proc. of the Workshop on Text, Speech and Dialogue TDS'99, Plzen, Czech Republic, September 1999. Lecture Notes in Artificial Intelligence, Springer 1999.

[82] Montes-y-Gómez, Gelbukh, López-López (1999e). Document intentions expressed in titles. Extraction, representation, and possible use, Selected Works 1997-1998, CIC-IPN, 1999.

[83] Montes-y-Gómez, López-López, Gelbukh (2000), Information Retrieval with Conceptual Graph Matching, Proc. of 11th International Conference on Database and Expert Systems Applications DEXA-2000, London, UK, September 2000.

[84] Montes y Gómez, Gelbukh, López-López (2001), Mining the news: trends, associations, and deviations. Accepted for Computación y Sistemas, Ibero American Journal of Computing, ISSN 1405-5546.

[85] Montes-y-Gómez, Gelbukh, López-López (2001b). A Statistical Approach to the Discovery of Ephemeral Associations *among News Topics*. Proc. DEXA 2001, 12th International Conference on Database and Expert Systems Applications. September 2001, Munich, Germany. Lecture Notes in Computer Science 2113. ISBN 3-540-42527-6, Springer-Verlag, pp. 491-500.

[86] Mugnier and Chein (1992), Polynomial Algorithms for Projection and Matching, Proc. of thee 7th Conceptual Graphs Workshop, Las Cruces, NM, 1992.

[87] Mugnier (1995), On generalization/specialization for conceptual graphs, Journal of Experimental and Theoretical Artificial Intelligence, Vol. 7, 1995.

[88] Mugnier and Chein, eds. (1998), Conceptual Structures: Theory, Tools, and Applications, Lecture Notes in AI 1453, Springer-Verlag, Berlin, 1998.

[89] Myaeng (1990), Conceptual Graph Matching as a Plausible Inferencing Technique for Text Retrieval, Proc. of the 5th Conceptual structures Workshop, Boston, MA, 1990.

[90] Myaeng, (1992), Using Conceptual graphs for Information Retrieval: A Framework for Adequate Representation and Flexible Inferencing, Proc. of Symposium on Document Analysis and Information Retrieval, Las Vegas, March 1992.

[91] Myaeng and Khoo (1992), On Uncertainty Handling in Plausible Reasoning with Conceptual Graphs, Proc. of the 7th Conceptual Graphs Workshop, Las Cruces, NM, 1992.

[92] Myaeng and López-López (1992), Conceptual Graph Matching: a Flexible Algorithm and Experiments, Journal of Experimental and Theoretical Artificial Intelligence, Vol. 4, 1992, pp. 107-126.

[93] Myaeng and Khoo (1994), Linguistic Processing of Text for a Large-Scale Conceptual Information Retrieval System, Lecture Notes in Artificial Intelligence 835, Springer-Verlag 1994.

[94] Nagle, Nagle, Gerholz and Eklund, eds. (1992), Conceptual Structures: Current Research and Practice, Ellis Horwood, New York, 1992.

[95] Nahm and Mooney (2000), Using Information Extraction to Aid the Discovery of Prediction Rules from Text, Proc. of Workshop on Text Mining, KDD-2000, Boston, MA, 2000.

[96] Nahm and Mooney (2001a), A Mutually Beneficial Integration of Data Mining and Information Extraction, Proc. of the Seventeenth Conference of Artificial Intelligence, AAAI-2000, Austin, TX, 2001.

[97] Nahm and Mooney (2001b), Mining Soft-Matching Rules from Textual Data, to appear in the Proc. of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-01), Seatle, WA, 2001.

[98] Perrin and Petry (1998), Contextual Text Representation for Unsupervised Knowledge Discovery in Texts, 2nd Pacific-Asia Conference PAKDD'98 on Research and Development in Knowledge Discovery and Data Mining, Lecture Notes in Artificial Intelligence 1394, Springer 1998.

[99] Petermann (1996), Natural Language Text Processing and the Maximal Join Operator, Proc. of the International Conference on Conceptual Structures 96, Lecture Notes in Artificial Intelligence 954, Springer, 1996.

[100] Pfeiffer and Nagle, eds. (1993), Conceptual Structures: Theory and Implementation, Lecture Notes in AI 754, Springer-Verlag, Berlin, 1993.

[101] Poole and Campbell (1995), A Novel Algorithm for Matching Conceptual and Related Graphs, 3rd. Conference on Conceptual Structures ICCS'95, Santa Cruz, CA, USA, August 1995.

[102] Rajman and Besançon (1997), Text Mining: Natural Language Techniques and Text Mining Applications, Proc. of the 7th IFIP 2.6 Working Conference on Database Semantics (DS-7), Chapam & Hall IFIP Proceedings serie. Leysin, Switzerland, Oct 7-10, 1997.

[103] Rajman and Besançon (1998), Text Mining - Knowledge Extraction from Unstructured Textual Data, 6th Conference of International Federation of Classification Societies (IFCS-98), 473-480, Rome, July 21-24, 1998.

[104] Rasmussen (1992), Clustering Algorithms, In Frakes and Baeza-Yates Eds., Information Retrieval: Data Structures & Algorithms, Pentice Hall, 1992.

[105] Rassinoux, Baud and Scherrer (1994), A Multilingual Analyser of Medical Texts, Proc. of the Second International Conference on Conceptual Structures, ICCS-94.

[106] Rauber and Merkl (1999), Mining Text Archives Creating Readable Maps to Structure and Describe Document Collections, Proceedings of the PKDD-1999, Lecture Notes in Artificial Intelligence 1704, Springer, 1999.

[107] Schwarz (1990), Automatic Syntactic Analysis of Free Text, Journal of the American Society for Information Scinece, 41(6), 1990.

[108] Sowa (1984), Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley, reading, M.A., 1984.

[109] Sowa (1988), Using a Lexicon of Canonical Graphs in a Semantic Interpreter, In Relational Models of the Lexicon, edited by Martha Evens, Cambridge University Press, p. 113-137, 1988.

[110] Sowa (1991), Towards the expressive power of natural languages, in J. F. Sowa, ed., Principles of Semantic Networks, Morgan Kaufmann Publishers, San Mateo, CA, 1991.

[111] Sowa, ed. (1992), Knowledge-Based Systems, Special Issue on Conceptual Graphs, vol. 5, no. 3, September 1992.

[112] Sowa (1999), Knowledge Representation: Logical, Philosophical and Computational Foundations, 1st edition, Thomson Learning, 1999.

[113] Sowa (2000), Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA.

[114] Sowa and Way (1986), Implementing a semantic interpreter using conceptual graphs, IBM Journal of Research and Development 30:1, January, 1986.

[115] Sparck-Jones (1999), What is the Role of NLP in Text Retrieval?, In Strzalkowski Ed., Natural Language Information Retrieval, Kluwer Academic Publishers, 1999.

[116] Srikant and Agrawal (1995), Mining Generalized Association Rules, Proc. of the 21st VLDB Conference, Zurich Switzerland, 1995.

[117] Stumme, ed. (2000), Working with Conceptual Structures: Contributions to ICCS'2000, Shaker-Verlag, 2000.

[118] Tan (1999), Text Mining: The state of the art and challenges, Proc. of the Workshop Knowledge Discovery from advanced Databases PAKDDD-99, Abril 1999.

[119] Tapia-Melchor (1997), Extracción de Información en Documentos en la WWW para Análisis Detallado, Tesis de maestría, Electrónica, INAOE 1997.

[120] Tapia-Melchor and López-López (1998), Automatic Information Extraction from Documents in WWW, Memorias del Séptimo Congreso Internacional de Electrónica, Comunicaciones y Computadoras, CONIELECOMP 98, pag. 287-291, Febrero, 1998.

[121] Tepfenhart, Dick and Sowa, eds. (1994), Conceptual Structures: Current Practice, Lecture Notes in AI 835, Springer-Verlag, 1994.

[122] Tepfenhart and Cyre, eds. (1999), Conceptual Structures: Standards and Practices, Lecture Notes in AI 1640, Springer-Verlag, 1999.

[123] Uthurusamy (1996), From Data Mining to Knowledge Discovery: Current Chanllenges and Future Directions, In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth and Ramasamy Uthurusamy, Advances in Knowledge Discovery and Data Mining, Cambridge, MA: MIT Press, 1996.

[124] Velardi, Pazienza and De' Giovanetti (1988), Conceptual Graphs for the Analysis and Generation of Sentences, IBM Journal of Research and Development, 32:2, 251-267, 1998.

[125] Way (1991), Knowledge Representation and Metaphor, Kluwer Academic Publishers, 1991.

[126] Way, ed. (1992), Journal of Experimental and Theoretical Artificial Intelligence (JETAI), Special Issue on Conceptual Graphs, vol. 4, no. 2, 1992.

[127] Weiss and Indurkhya (1998), Predictive Data Mining: A Practical Guide, Morgan Kaufmann Publishers, Inc., 1998.

[128] Weiss, Apte, Damerau, Johnson, Oles, Goetz and Hampp (1999), Maximizing Text-Mining Performance, IEEE Intelligent Systems, July/August 1999.

[129] Witten and Frank (1999), Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kauffmann Publishers, 1999.

[130] Woods (1986), Important Issues in knowledge Representation, Proceedings of the IEEE, vol. 74, No. 10, October 1986.

[131] Wu and Palmer (1994), Verb Semantics and Lexical Selection. Proc. of the 32nd Annual Meeting of the Associations for Computational Linguistics, Las Cruces, New Mexico, 1994.

[132] Yang, Choi and Oh (1992), CGMA: A Novel Conceptual Graph Matching Algorithm, Proc. of thee 7th Conceptual Graphs Workshop, Las Cruces, NM, 1992.

[133] Zelikovitz and Hirsh (2000), Improving Short-Text Classification using Unlabeled Background Knowledge to Assess Document Similarity, Proceedings of the Seventeenth International Conference on Machine Learning (ICML-2000). Morgan Kaufmann Publishers 2000.

Apéndice A

Construcción de los grafos conceptuales de prueba

El método de minería de texto propuesto en este trabajo se experimentó analizando dos conjuntos de artículos científicos. En este apéndice se describe el proceso de construcción de los grafos conceptuales que representan el contenido de dichos artículos.

Además, con la propósito de ilustrar este proceso de construcción, en la parte final del apéndice se muestran algunos ejemplos de artículos científicos y sus correspondientes grafos conceptuales.

Construcción de los grafos conceptuales de prueba

A.1 Método de construcción de los GCs

A.1.1 Antecedentes del método

Los sistemas de recuperación de información representan generalmente el contenido de los textos con un conjunto de palabras clave. Esta estrategia facilita el proceso de búsqueda, pero limita grandemente la precisión de sus resultados.

Actualmente, en busca de una solución a este problema, algunos métodos de recuperación de información emplean representaciones más completas del contenido de los textos, es decir, representaciones que permiten considerar detalles de su contenido tales como propósitos, planes, objetivos, y otras características que van más allá de sus temas.

Basándonos en esta idea propusimos un método de *recuperación de información a dos niveles* (López-López, 1995; Tapia-Melchor, 1997; Montes-y-Gómez, 1998). Este método se ilustra en la figura A.1. En él, la búsqueda de información se realiza de la siguiente manera:

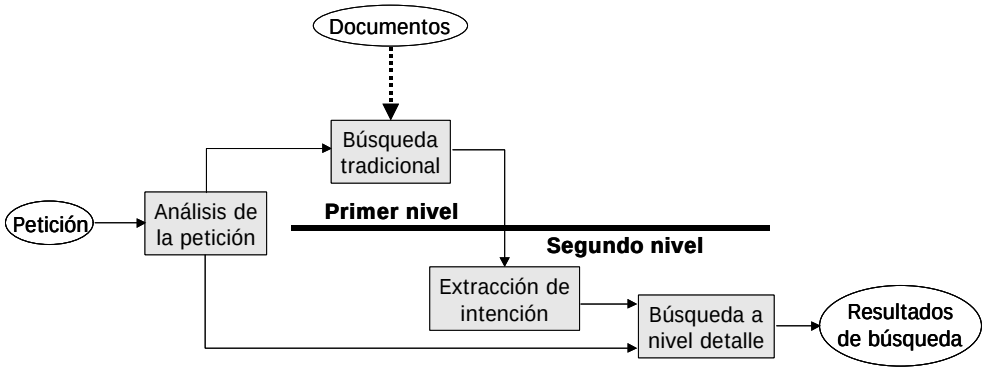


Figura A.1 Recuperación de información a dos niveles

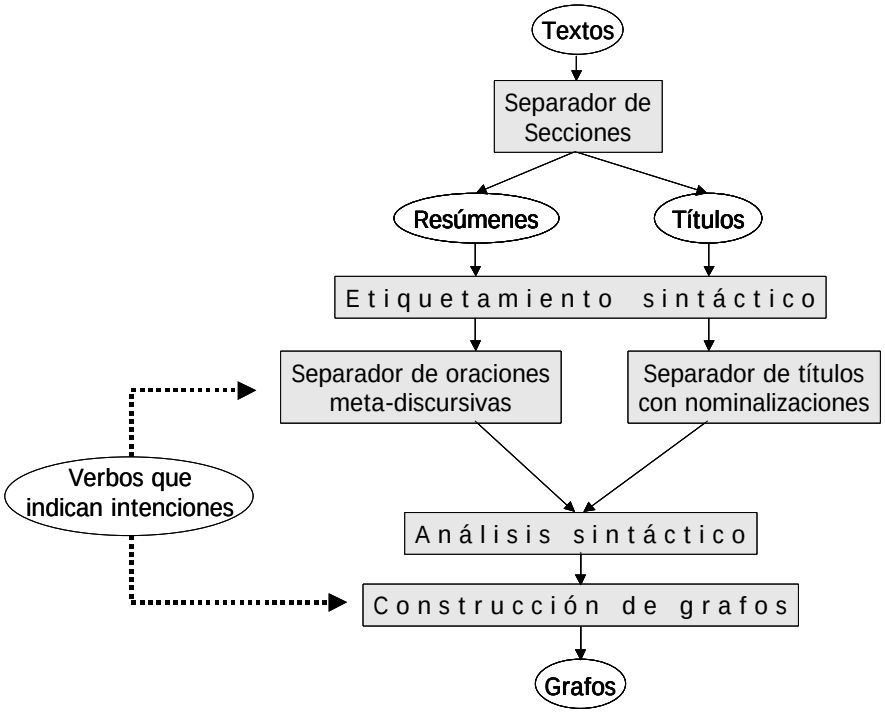


Figura A.2 Construcción de los grafos conceptuales

Primero se hace una búsqueda tradicional con palabras clave. El resultado de este primer nivel de análisis es un conjunto de textos cuyo tema coincide con el tema de interés expresado en la petición. Después se construye un grafo conceptual de cada uno de estos textos. Estos grafos expresan algunos detalles del contenido de los textos; principalmente su intención. Finalmente se realiza una búsqueda a nivel detalle, donde se comparan los grafos conceptuales de los textos con el grafo conceptual de la petición. Así, el resultado de este segundo nivel de análisis es un conjunto de textos que no sólo tratan el tema buscado; también lo hacen de la manera deseada.

La evaluación de este método consideró la búsqueda de información en colecciones de artículos científicos. Por ello, el segundo nivel se compone de un módulo que extrae la intención de cada artículo, y luego la transforma en grafo conceptual (ver figura A.1). En la siguiente sección se describe el funcionamiento de este módulo.

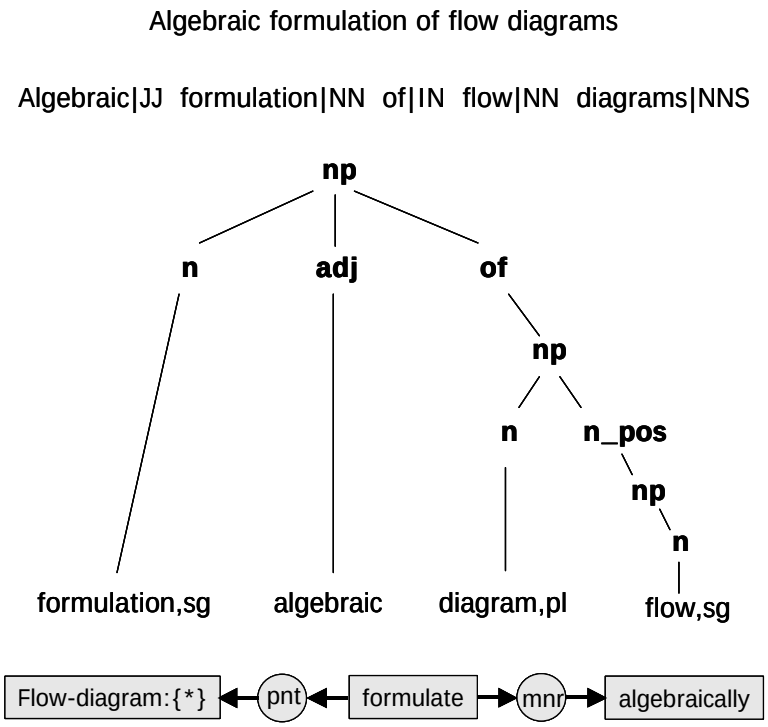


Figura A.3 Transformación de un título en grafo conceptual

A.1.2 Extracción de la intención

Todos los artículos científicos tiene alguna intención. Por ejemplo, algunos describen algo, otros lo introducen, y otros más lo evalúan. Estas intenciones se expresan en varias formas, pero nuestro método considera sólo las siguientes dos:

1. **Oraciones de los títulos con nominalizaciones.** Los siguientes títulos son dos ejemplos. En ellos se resalta la intención identificada.

   An introduction to a Machine-Independent Data Division.

   Evaluation of Jacobi symbol.

2. **Oraciones metadiscursivas de los resúmenes**. Las siguientes oraciones son dos ejemplos. Aquí también se resalta la intención identificada.

   This paper introduces the conceptual clustering.

An automatic procedure for resolving semantic problems is suggested.

El proceso de extracción de la intención de un artículo científico, y de su transformación en grafo conceptual se ilustra en la figura A.2. Este proceso tiene tres etapas principales: una etapa de preprocesamiento donde se separaran las oraciones interesantes –que expresan una intención– de los títulos y los resúmenes; una etapa de análisis sintáctico donde se "estructuran" estas intenciones, y se reconocen su sujeto, objeto y otros atributos; finalmente, una etapa de análisis semántico donde se construye el grafo conceptual representante del artículo.

La figura A.3 ilustra el proceso de transformación del título de un artículo científico en un grafo conceptual. En ella se muestran cuatro diferentes estados: el título original, el título con etiquetas de parte de la oración, el árbol sintáctico (donde se hacen explicitas las relaciones sintácticas entre las palabras de la oración), y el grafo conceptual resultante.

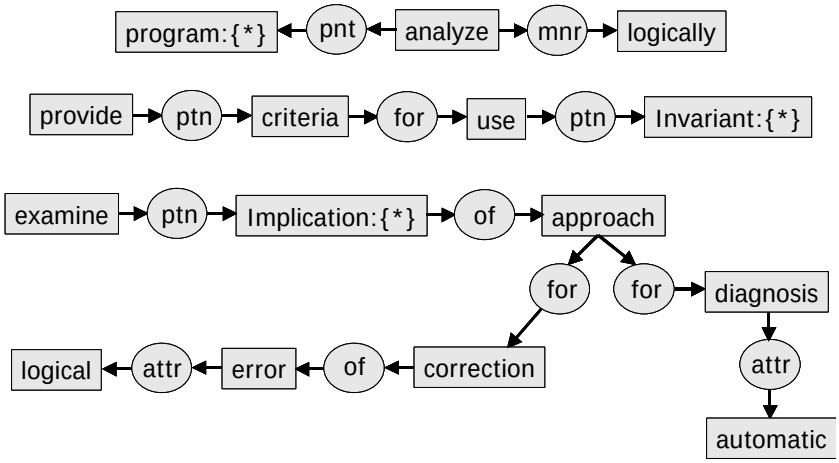
A.2 Ejemplos de los grafos conceptuales de prueba

En esta sección se muestran algunos ejemplos de los grafos conceptuales de prueba. Estos ejemplos tienen dos elementos principales: la descripción textual del artículo (en este caso solamente se muestra el título y el resumen), y el grafo conceptual correspondiente.

Ejemplo 1: Un artículo de ciencias de la computación

Logical Analysis of Programs

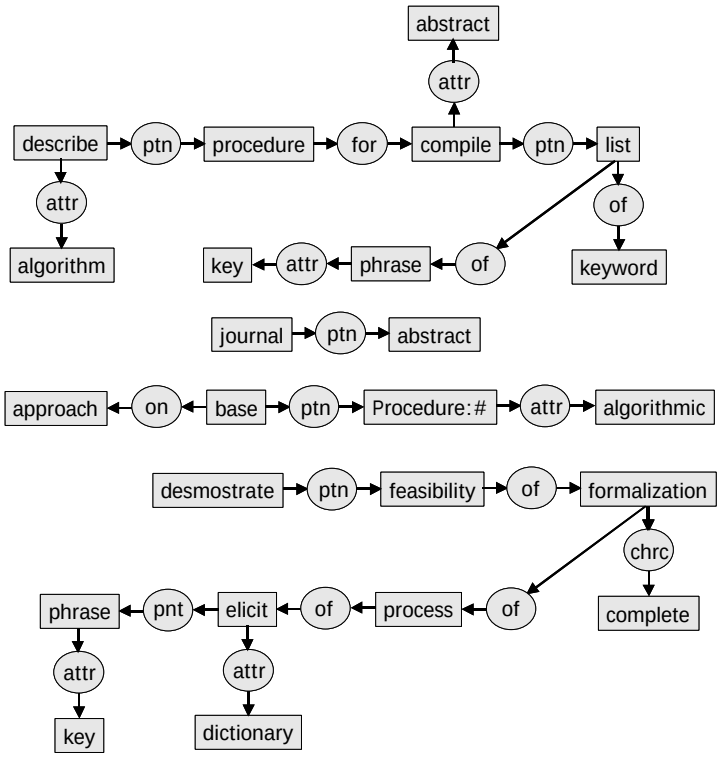
The first part of the paper is devoted to techniques for the automatic generation of invariants. The second part provides criteria for using the invariants to check simultaneously for correctness (including termination) or incorrectness. A third part examines the implications of the approach for the automatic diagnosis and correction of logical errors.



Ejemplo 2: Un artículo de ciencias de la información

Algorithmic Procedure for Compiling a List of Keywords and Key Phrases by the Abstracts in "Fizika" Abstract Journal.

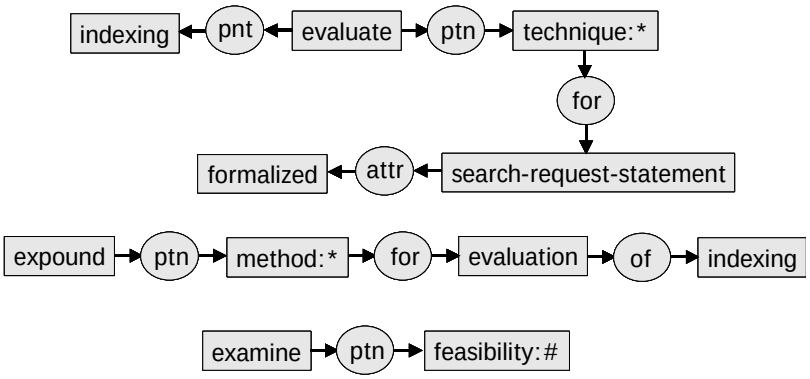
The algorithmic procedure is based on a linguistic approach; it serves to isolate key phrases from the abstracts of the "Fizika" abstract journal, recognizing set phrases with the aid of linguistic rules.. The feasibility is demonstrated of a completed formalization of the process of eliciting key phrases for a descriptor dictionary.



Ejemplo 3: Un artículo de ciencias de la computación

Evaluation of Indexing and a Technique for Formalized Search Request Statement.

A method for evaluation of indexing is expounded . The feasibility is examined of using marked documents instead of requests, called the "beacon method". A M-algorithm for formalized statement of search requests is described and exemplified by an information retrieval system in the nitrogen industry.



Ejemplo 4: Un artículo de ciencias de la información (en este caso sólo fue posible extraer información del título)

Comparisons of Four Types of Lexical Indicators of Content

An experiment was conducted to determine which of four types of lexical indicators of content could be utilized best by subjects to determine relevant from irrelevant documents and to answer a set of 100 questions. The results indicate that there were no major differences between the groups using complete text and abstracts to select relevant documents, but the group utilizing the complete text obtained a significantly higher score on the examination.