# Un método de agrupamiento de grafos conceptuales para minería de texto*

**M. Montes-y-Gómez** y **A. Gelbukh**

Centro de Investigación en Computación (CIC), IPN, México.
mmontesg@susu.inaoep.mx, gelbukh@cic.ipn.mx

**A. López-López**

Instituto Nacional de Astrofísica, Optica y Electrónica (INAOE), México.
allopez@inaoep.mx

**R. Baeza-Yates**

Departamento de Ciencias de la Computación, Universidad de Chile, Chile.
rbaeza@dcc.uchile.cl

**Resumen.** La minería de texto, como muchas otras tareas de procesamiento de texto, se realiza usualmente sobre representaciones simples del contenido de los textos. Aquí se presenta un método para el agrupamiento conceptual de una colección de textos representados por un conjunto de grafos conceptuales, que son una representación simple pero con mayor información del contenido de los textos. Este método emplea una estrategia de aprendizaje no supervisado, que construye incrementalmente una jerarquía de los grafos conceptuales. Además este método incorpora algunas características que lo hacen atractivo para la minería de texto. Por ejemplo: considera toda la información estructural de los grafos conceptuales, emplea conocimiento del dominio, y considera los intereses del usuario.

## *1 Introducción*

Hoy en día, debido principalmente al gran valor del conocimiento y a la fácil disponibilidad de grandes conjuntos de datos, muchas instituciones consideran una de sus principales necesidades el diseño de los mecanismos que automaticen el análisis de sus datos, la extracción de información de éstos, y su conversión en conocimiento. Uno de los esfuerzos más importantes en esta dirección son los sistemas de *minería de texto.* Estos sistemas permiten analizar grandes colecciones de textos y descubrir en ellos distintos tipos de patrones interesantes (Feldman and Dagan, 1995).

Típicamente, la minería de texto se realiza en dos fases (Tan, 1999). Una fase de preprocesamiento, donde los textos son transformados a algún tipo de representación semiestructurada que permita su análisis automático, y una fase de descubrimiento, donde las representaciones intermedias son analizadas y algunos patrones interesantes, como por ejemplo: agrupamientos, asociaciones, desviaciones y/o tendencias pueden ser descubiertos.

La mayoría de los métodos actuales de minería de texto utilizan representaciones sencillas del contenido de los textos, como por ejemplo, listas de conceptos o palabras clave. Por una parte, estas representaciones se construyen y analizan fácilmente, pero por otra parte, estas representaciones limitan grandemente los tipos de patrones descubiertos.

Recientemente, en muchas aplicaciones relacionadas con el análisis de textos, existe la tendencia de utilizar representaciones más completas del contenido de los textos, es decir, representaciones que consideren más tipos de elementos textuales. En la minería de texto, por ejemplo, se cree que estas representaciones permitirán extender los tipos y mejorar la expresividad de patrones descubiertos (Hearst, 1999; Tan, 1999).

En este último contexto, en este artículo se propone un método de minería de texto que emplea *grafos conceptuales* (Sowa, 1984, Sowa, 1999) como la representación intermedia

* Trabajo hecho con apoyo parcial del CONACyT, SNI y CGEPI-IPN, México.

de los textos[1]. Bajo esta situación, dos problemas diferentes son de gran importancia: 1) la transformación de los textos en grafos conceptuales, y 2) el análisis automático de un conjunto de estos grafos.

Sobre el primer problema –la transformación de textos en grafos conceptuales– sólo mencionamos aquí que para tal transformación se emplea primero el análisis sintáctico y después los árboles sintácticos con los nodos encabezados (*heads*) se convierten en los grafos conceptuales a través de una análisis semántico. La tarea principal de este último es la recuperación de las relaciones semánticas (agente, paciente, atributo etc.) a través de las relaciones sintácticas (sujeto, objeto, modificador etc.) empleando el conocimiento jerárquico sobre las valencias semánticas de verbos, los roles típicos de objetos etc. También se recuperan marcas de referencia y correferencia a través de los artículos, resolución de la anáfora etc. También se agrega información de generalización, por ejemplo, *Fido* → [*perro*: *Fido*]. Este proceso es complejo, dependiente del dominio y basado en conocimiento (Sowa and Way, 1986). Algunos tipos de textos que se han trasformado a grafos conceptuales son: algunas partes de artículos científicos (Myaeng and Khoo, 1994; Montes-y-Gómez *et al*., 1999), de expedientes médicos (Baud *et al.*, 1992) y también de casos legales (Boucier and Rajman, 1994). Cabe mencionar que, a pesar de la complejidad del proceso de la conversión de los textos en los grafos conceptuales, se puede hacer con mayor o menor calidad y profundidad, recuperando sólo información parcial fácil –o factible– de recuperar. Entre mayor es la calidad de conversión mayor es la calidad de la minería de texto sobre sus resultados.

Sin embargo, este artículo no está relacionado con el problema de la transformación de textos en grafos conceptuales, sino se enfoca únicamente en el *agrupamiento* de un conjunto de grafos conceptuales, y por tanto, en el descubrimiento de sus principales regularidades[2].

Actualmente son conocidos dos métodos para el agrupamiento conceptual de grafos conceptuales (Mineau and Godin, 1995; Godin *et al*., 1995; Bournaud and Ganascia, 1996; Bournaud and Ganascia, 1997). El método aquí descrito se basa en estos dos métodos. Al igual que ellos, este método emplea una estrategia de aprendizaje no supervisado que construye incrementalmente el agrupamiento –jerarquía conceptual– del conjunto de grafos; pero diferente de ellos, este método incorpora algunas otras características favorables para la minería de texto. Por ejemplo, este método 1) considera toda la información estructural de los grafos conceptuales; 2) utiliza conocimientos del dominio durante la construcción de dicha jerarquía, y 3) obtiene una jerarquía conceptual que enfatiza los intereses del usuario.

El resto del artículo se organiza de la siguiente manera. La sección 2 describe brevemente nuestro método de comparación de dos grafos conceptuales. La sección 3 define formalmente la jerarquía conceptual y presentan el procedimiento incremental para su construcción. Finalmente, la sección 4 expone algunas conclusiones y discute los principales trabajos futuros.

## *2 Comparación de grafos conceptuales*

Una de las operaciones básicas para el agrupamiento de los grafos conceptuales es, sin lugar a dudas, su *comparación*. En algunos trabajos previos hemos presentado un método flexible para la comparación de dos grafos conceptuales cualesquiera (Montes-y-Gómez *et al.*, 2000; Montes-y-Gómez *et al.*, 2001). Este método utiliza un diccionario de sinónimos y algunas jerarquías de conceptos. El diccionario de sinónimos permite considerar la semejanza entre conceptos equivalentes no necesariamente iguales, mientras que las jerarquías de conceptos permiten determinar semejanzas a diferentes niveles de generalización y además enfocar la comparación de los grafos sobre los conceptos más importantes para el usuario.

En general, este método obtiene una descripción cualitativa de la semejanza entre los dos grafos, así como una medida cuantitativa de esta semejanza. La descripción de la semejanza es simplemente un traslape de los grafos, es decir, un conjunto máximo de generalizaciones comunes compatibles entre los dos grafos, mientras que la medida de semejanza es una expresión de la importancia relativa del traslape

[1] Aquí, el término "texto" se usa para indicar una parte de una oración, una oración completa, un párrafo, o incluso, aunque en el menor de los casos, un documento completo.

[2] Una regularidad es cualquier elemento común a dos o más grafos de la colección.

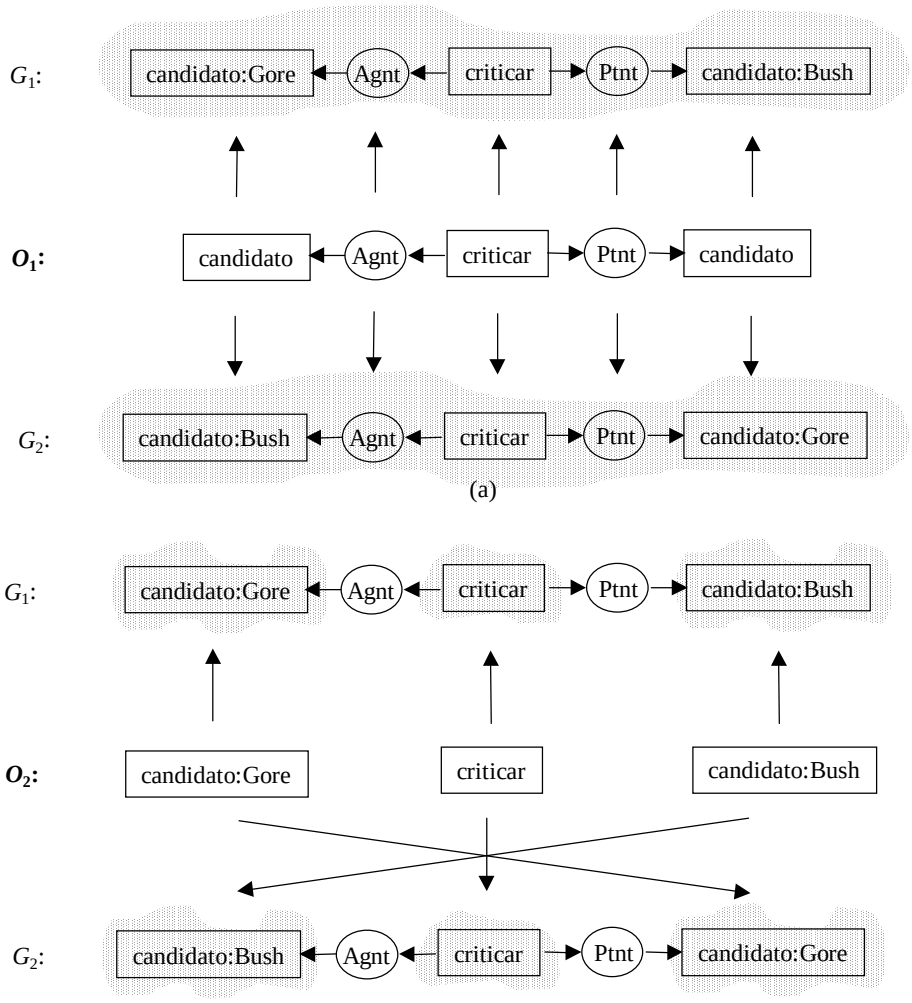


Figura 1. Apareamiento de dos grafos conceptuales

con respecto a la información contenida en los grafos conceptuales originales.[3]

Esta semejanza se calcula de acuerdo con los intereses del usuario, y se define como:

$$sim(G_i, G_j) = f(w_E, w_V, w_A, i_c, i_r).$$

Aquí, $w_E$, $w_V$ y $w_A$ indican la importancia relativa de las entidades, acciones y atributos respectivamente, mientras que $i_c$ señala la importancia de la semejanza conceptual, es decir, de la semejanza originada por los nodos concepto equivalentes, y $i_r$ señala la importancia de la semejanza relacional, esto es, la semejanza causada por las equivalencias a nivel estructural. Los valores de estos parámetros se rigen por las siguientes condiciones:

1. $w_E, w_V, w_A > 0$.
2. $i_c + i_r = 1$.

La descripción detallada de la medida de semejanza se encuentra en (Montes-y-Gómez *et al.*, 2001).

Ajustando el valor de estos parámetros es fácil adaptar el método de comparación ante los distintos intereses de los usuarios.

Por ejemplo, dados los grafos conceptuales $G_1$ y $G_2$ de la figura 1, uno de ellos representando la frase "*Gore critica a Bush*" y otro la frase "*Bush critica a Gore*", este método determina dos diferentes descripciones de su semejanza, cada una de ellas asociada a un traslape distinto. Así pues, el traslape $O_1$ indica que en ambos grafos "*un candidato critica a otro candidato*", mientras que el traslape $O_2$ indica que ambos grafos mencionan los conceptos "*Bush*", "*Gore*" y "*criticar*".

La selección del mejor traslape, esto es, de la descripción más apropiada de la semejanza

[3] Cuando existe más de un (posible) traslape entre los grafos, entonces aquel que produce la mayor medida de semejanza es seleccionado como la descripción de dicha semejanza.

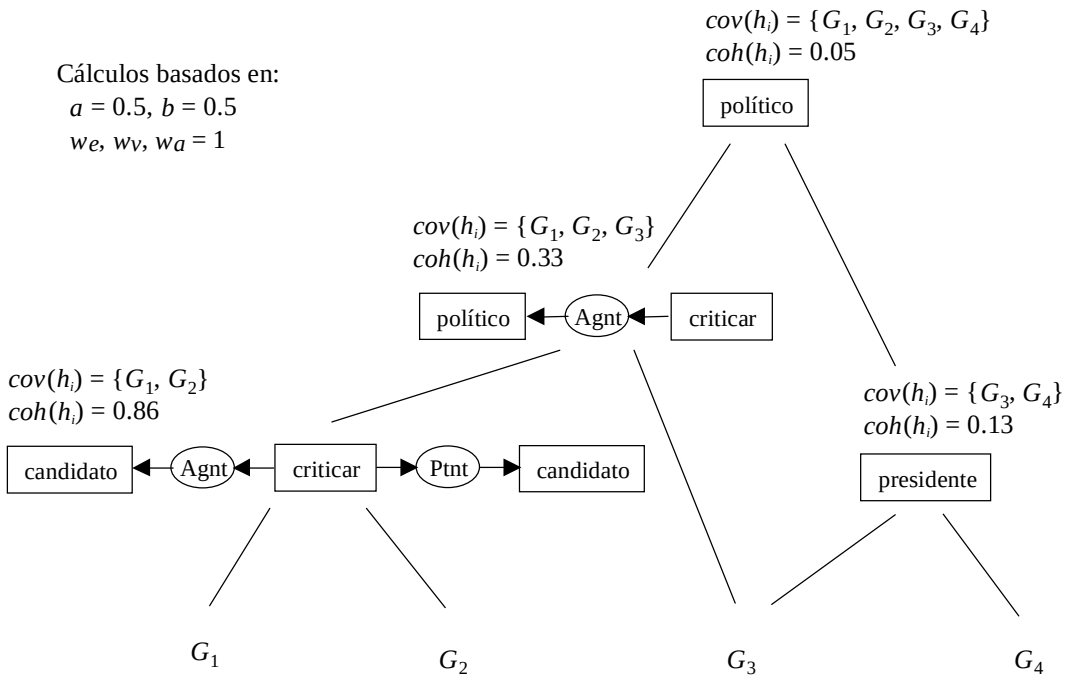


Figura 3. Agrupamiento de los grafos conceptuales.

entre los dos grafos conceptuales, depende de los intereses específicos del usuario. Por ejemplo, si la información estructural se enfatiza ($i_r > i_c$), entonces el traslape $O_1$ es la mejor descripción de la semejanza, pero si por el contrario, la información conceptual se favorece ($i_c > i_r$), entonces el traslape $O_2$ resulta la mejor descripción de la semejanza.

## *3 Agrupamiento de los grafos conceptuales*

Dado un conjunto 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 posterior análisis, y por tanto, el descubrimiento de otros tipos de patrones interesantes.

El método aquí descrito es un método de *agrupamiento conceptual* que, a diferencia de las técnicas tradicionales de agrupamiento, no sólo 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.

La jerarquía resultante no es necesariamente un árbol ni un *lattice*, sino un conjunto de árboles. Esta jerarquía es una especie de red de herencia, donde los nodos inferiores indican regularidades especializadas y los nodos superiores sugieren regularidades generalizadas. Por ejemplo, dada la pequeña colección de grafos de la figura 2, este método construye la jerarquía conceptual de la figura 3.

Formalmente, cada nodo $h_i$ de esta jerarquía se representa por una triada[4] ($cov(h_i)$, $desc(h_i)$, $coh(h_i)$), donde: $cov(h_i)$, la cobertura de $h_i$, es el conjunto de grafos cubiertos por la regularidad $h_i$; $desc(h_i)$, la descripción de $h_i$, consiste de los elementos comunes de estos grafos, es decir, es el traslape de los grafos de $cov(h_i)$; y $coh(h_i)$, la cohesión de $h_i$, indica la semejanza mínima entre dos grafos cualesquiera de $cov(h_i)$, esto es:

$$\forall G_i, G_j \in cov(h_i), sim(G_i, G_j) \geq coh(h_i)$$

En esta jerarquía, el nodo $h_i$ es considerado un antecesor del nodo $h_j$ ($h_j < h_i$), sí y sólo sí: $cov(h_j) \subset cov(h_i)$, $desc(h_j) < desc(h_i)$ y $coh(h_j) \geq coh(h_i)$.

### 3.1 Construcción de la jerarquía conceptual

Dado un conjunto de grafos conceptuales, la construcción de su jerarquía conceptual se basa

[4] Esta notación fue adaptada a partir de (Bournaud and Ganascia, 1996), donde cada nodo se representa por un par ($cov(h_i)$, $desc(h_i)$).

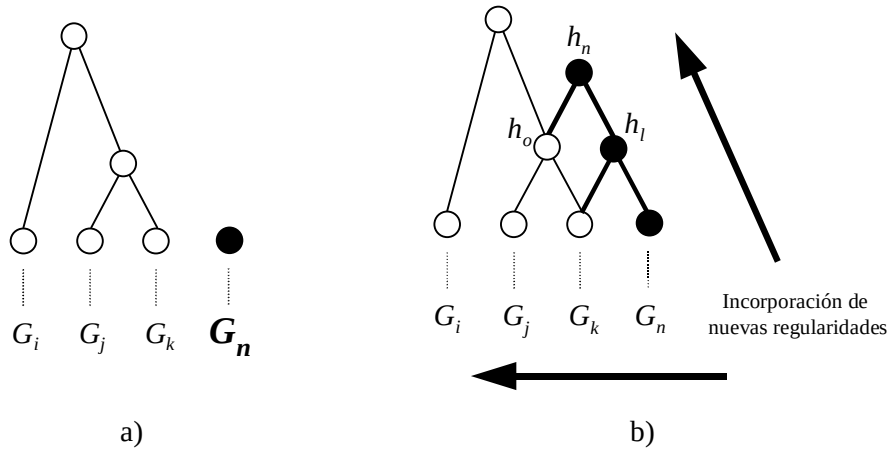


Figura 4. Incorporación de un nuevo grafo a la jerarquía.

en un método incremental. Este método considera toda la información estructural de los grafos conceptuales y además utiliza la medida de su semejanza para considerar los intereses del usuario.

En general, la incorporación de un nuevo grafo conceptual $G_n$ a la jerarquía conceptual se realiza en dos pasos (ver la figura 4). En el primer paso, un nodo que cubre exclusivamente al nuevo grafo ($\{G_n\}$, $G_n$, 1) es añadido a la jerarquía. En el segundo paso, todas las regularidades relacionadas con la nueva evidencia son identificadas. Estas nuevas regularidades (nuevos nodos) son añadidas a la jerarquía siguiendo una estrategia ascendente (*bottom-up*), esto es, cada nodo de nivel superior es construido combinando dos nodos de niveles más bajos. Por ejemplo, el nodo $h_n$ de la figura 4(b) fue construido a partir de los nodos $h_o$ y $h_l$. En este caso, el nodo $h_n$ se define como:

$$\begin{aligned}
&\mathrm{cov}(h_n) = \mathrm{cov}(h_o) \cup \mathrm{cov}(h_l) \\
&desc(h_n) = match(desc(h_o), desc(h_l)) \\
&coh(h_j) = \begin{cases} sim(desc(h_o), desc(h_l)) \\ \min(coh(h_o), coh(h_l)) \end{cases}
\end{aligned}$$

donde en la última fórmula, la primera variante se elige si:

$$|\mathrm{cov}(h_o)| = |\mathrm{cov}(h_l)| = 1$$

y la segunda en el caso contrario. Aquí, la función *match*($G_i$, $G_j$) regresa el "mejor" traslape entre $G_i$ y $G_j$, la función *sim*($G_i$,$G_j$) su medida de semejanza, y la función *min*(*coh*($h_i$), *coh*($h_l$)) la menor cohesión de los nodos $h_i$ y $h_j$.

Además de esto, cada vez que una nueva regularidad $h_n$ se añade a la jerarquía, las regularidades duplicadas son eliminadas. Por ejemplo, si *desc*($h_o$) = *desc*($h_n$), entonces el nodo $h_o$ se elimina de la jerarquía; mientras que si *desc*($h_l$) = *desc*($h_n$), entonces $h_l$ es eliminado.

De acuerdo con esta descripción, existe una relación directa entre la identificación de las nuevas regularidades y el método de comparación de los grafos conceptuales. Esta relación define la construcción de la jerarquía conceptual como un proceso basado en conocimiento y *dirigido por el usuario*; esto implica que el uso de diferentes bases de conocimiento (jerarquías de conceptos principalmente) y el establecimiento de diferentes objetivos del usuario (parámetros de la medida de semejanza) pueden producir diferentes jerarquías conceptuales.

Por ejemplo, si se vuelven a agrupar los grafos de la figura 2, pero esta vez enfatizando las semejanzas conceptuales ($i_c > i_r$), la jerarquía conceptual resultante es distinta. Esta nueva jerarquía se muestra en la figura 5, en ella los nodos resaltados indican las diferencias con respecto a la jerarquía conceptual de la figura 3.

Algunas características interesantes de estas jerarquías son:

- Que permiten realizar un *agrupamiento con traslapes*. Por ejemplo, el grafo conceptual $G_3$ forma parte de dos grupos diferentes.
- Que no contienen regularidades duplicadas. Por ejemplo, el nodo ({G1,G2,G4}, [político], 0.05) se eliminó cuando se construyó el nodo ({G1,G2,G3,G4}, [político], 0.05).

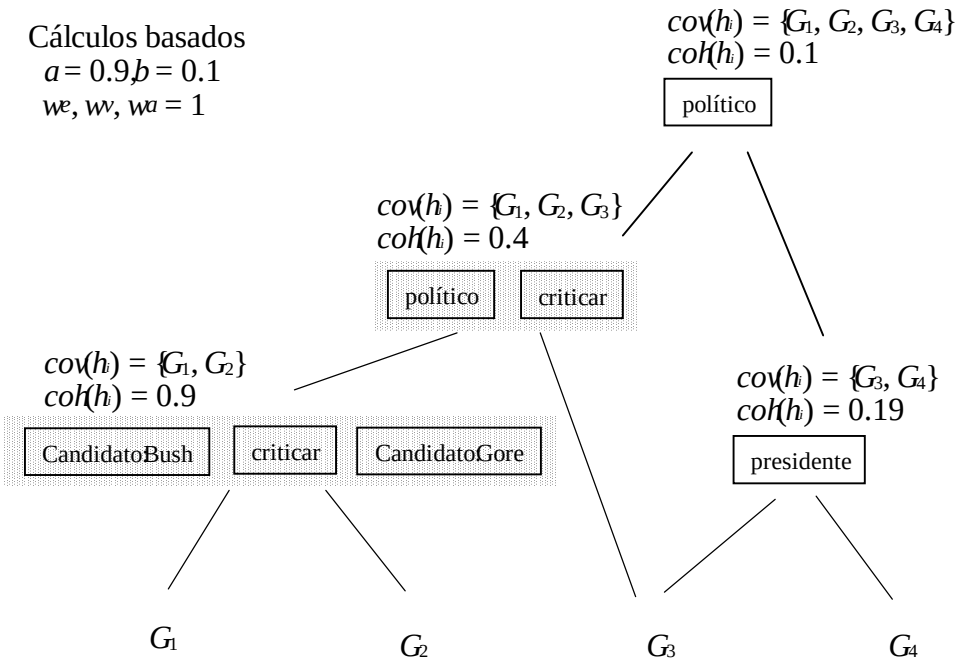


Figura 5. Una jerarquía conceptual distinta.

- Que expresan todas las regularidades de la colección de grafos, aunque estas regularidades –su descripción– enfatizan un punto de vista específico.

## *4 Conclusiones*

En este artículo se presentó un método para el agrupamiento conceptual de una colección de textos representados por un conjunto de grafos conceptuales. Este método emplea una estrategia tradicional de aprendizaje no supervisado que construye incrementalmente una jerarquía de los grafos conceptuales. Adicionalmente, este método incorpora algunas características atractivas para la minería de texto. Por ejemplo, este método considera toda la información estructural de los grafos conceptuales, emplea conocimientos del dominio, y también permite enfatizar los intereses del usuario durante la construcción de la jerarquía conceptual. En resumen, el agrupamiento de los grafos conceptuales es un proceso basado en conocimiento y dirigido por el usuario.

La jerarquía conceptual resultante es un descubrimiento interesante por dos razones principales. En primer lugar, porque indica la *estructura oculta* de la colección, y en segundo lugar, porque constituye un *resumen organizado* de la colección que facilita su visualización y también su análisis posterior.

## *5 Trabajo futuro*

El trabajo futuro consiste de dos tareas principales. En primer lugar, la *experimentación* formal de estas ideas, donde se planea analizar una colección de títulos de artículos científicos representados con grafos conceptuales. Este análisis permitirá, entre otras cosas, medir el crecimiento de la jerarquía, estudiar la influencia de los intereses del usuario durante su construcción, y además determinar una estrategia para construir un agrupamiento reducido personalizado.

En segundo lugar, el descubrimiento de otros patrones interesantes del conjunto de grafos. En este caso se planea utilizar la jerarquía conceptual como un índice de la colección, y con base en ella descubrir algunas asociaciones y desviaciones interesantes.

## *Referencias*

1. Baud, Rassinoux and Scherrer (1992), Natural Language Processing and Semantical Representation of Medical Texts, Meth Inform Med 31:117-25, 1992.
2. Bourcier and Rajman (1994), Interactional Semantics for Legal Case-Based Knowledge, à paraître fin 1994.
3. Bournaud and Ganascia (1996), Conceptual Clustering of Complex Objects: A Generalization Space based Approach,

Lecture Notes in Artificial Intelligence 954, Springer, 1996.

4. Bournaud and Ganascia (1997), Accounting for Domain Knowledge in the Construction of a Generalization Space, Lectures Notes in AI (1257), Springer-Verlag, 1997.
5. Feldman and Dagan (1995), Knowledge Discovery in Textual databases (KDT), Proc. of the 1st International Conference on Knowledge discovery (KDD_95), 1995.
6. Godin, Mineau and Missaoui (1995), Incremental Structuring of Knowledge Bases, International KRUSE Symposium, August 11-13, Santa Cruz, California, 1995.
7. Hearst (1999), Untangling Text Data Mining, To appears in Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, 1999.
8. Mineau and Godin (1995), Automatic Structuring of Knowledge Bases by Conceptual Clustering, IEEE Transactions on Knowledge and Data Engineering, 7(5), 1995.
9. Montes-y-Gómez, López-López and Gelbukh (1999), Extraction of Document Intentions from Titles, Proc. of the Workshop on Text Mining: Foundations, Techniques and Applications, IJCAI-99, Sweden, 1999.
10. Montes-y-Gómez, Gelbukh and López-López (2000), Comparison of Conceptual Graphs, Lecture Notes in Artificial Intelligence 1793, Springer 2000.
11. Montes-y-Gómez, Gelbukh, López-López and Baeza-Yates (2001), Flexible Comparison of Conceptual Graphs, submitted to DEXA-2001.
12. Myaeng and Khoo (1994), Linguistic Processing of Text for a Large-Scale Conceptual Information Retrieval System, Lecture Notes in AI 835, Springer-Verlag 1994.
13. Sowa (1984), Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley, Reading, M.A., 1984.
14. Sowa (1999), Knowledge Representation: Logical, Philosophical and Computational Foundations, First Edition, Thompson Learning, 1999.
15. Sowa and Way (1986), Implementing a semantic interpreter using conceptual graphs, IBM Journal of Research and Development 30:1, January, 1986.
16. Tan (1999), Text Mining: The State of the Art and the Challenges, Proc. of the Workshop on Knowledge Discovery from Advanced Databases PAKDD' 99, April 1999.