Computación y Sistemas Vol. 9 Núm. 1, pp. 063-081

RESUMEN DE TESIS DOCTORAL

Minería de Texto empleando la Semejanza entre Estructuras Semánticas

Text Mining using Comparison of Semantic Structures

Graduated: Manuel Montes y Gómez
Centro de Investigación en Computación – IPN
Av. Juan de Dios Bátiz s/n esq. Miguel Othón de Mendizábal C. P. 07738 México D. F.
Graduado en febrero 26, 2002

Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)
Luis Enrique Erro #1, Sta. María Tonantzintla, Puebla, México.
mmontesg@inaoep.mx

Advisor: Alexander Gelbukh
Centro de Investigación en Computación - IPN
Av. Juan de Dios Bátiz s/n esq. Miguel Othón de Mendizábal C. P. 07738 México D. F.
gelbukh@cic.ipn.mx; www.Gelbukh.com

Co-Advisor Aurelio López López
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)
Luis Enrique Erro #1, Sta. María Tonantzintla, Puebla, México.
allopez@inaoep.mx

Resumen

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 realizar ciertas operaciones con la información, por ejemplo: buscarla, compararla, y resumirla. La minería de texto, una nueva área de investigación definida como descubrimiento de conocimiento en colecciones de textos, se enfoca en el análisis de grandes conjuntos de documentos. En particular, considera el descubrimiento de patrones interesantes, tales como grupos, asociaciones y desviaciones, en colecciones de textos. Los métodos actuales de minería de texto se caracterizan por usar representaciones sencillas del contenido de los documentos, por ejemplo, bolsas o vectores de palabras. Por una parte estas representaciones son fáciles de obtener y analizar, pero por otra parte restringen los patrones descubiertos a un nivel temático. Con el propósito de obtener resultados más útiles y significativos deben usarse representaciones más completas de la información. Basándonos en esta suposición se propuso un nuevo método para realizar minería de texto a nivel detalle. Este método usa los grafos conceptuales como representación del contenido de los textos, y obtiene algunos patrones descriptivos de los documentos aplicando varios tipos de operaciones sobre estos grafos.
Palabras Clave: Minería de Texto, Grafos Conceptuales, Agrupamiento Conceptual, Descubrimiento de Conocimiento.

Abstract

Knowledge is the most valuable treasure of humankind. Most of this knowledge exists in natural language format, for instance, in books, journals, reports, etc. The real possession of all this knowledge depends on our capabilities to perform different tasks with texts, such as: searching for interesting texts, comparing different documents, and summarizing them.Text mining, an emerging research area that can be roughly characterized as knowledge discovery in large text collections, is focused on automatically analyzing a set of texts. Mainly, it is concerned with the discovery of interesting patterns such as clusters, associations, and deviations from large 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 one hand, such representations are easy to obtain from texts and easy to analyze, but 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 new method for doing text mining at detail level. This method uses conceptual graphs for representing text content and relies on performing some tasks on these graphs, allowing the discovery of more descriptive patterns.
Keywords: Text Mining, Conceptual Graphs, Conceptual Clustering, Knowledge Discovery

1 Introducción

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 semánticas del contenido de los textos aunque su aplicación no ha sido tan definitiva y valiosa como se esperaba (Sparck-Jones, 1999). Las principales causas de este resultado desfavorable son, entre otras, las siguientes:

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 tesis consideró el problema de la expresividad de los resultados de la minería de texto, y también la oportunidad de comenzar a usar representaciones más completas del contenido de los textos en ella. Básicamente en esta tesis se planteó el uso de una representación "semántica" del contenido de los textos, y se propusieron algunos métodos para el descubrimiento de patrones interesantes en un conjunto de dichas representaciones. Así pues, el objetivo de este trabajo fue definir algunas estrategias de minería de texto para mejorar la expresividad y la diversidad de los patrones descubiertos con respecto a los obtenidos usando las técnicas tradicionales.

2 Minería de Texto

La minería de texto es el área de investigación más reciente del procesamiento automático de textos. Ella se define como el proceso automático 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. Entonces, dependiendo del tipo de métodos aplicados en la etapa de preprocesamiento es el tipo de representaciones intermedias construidas, y en función de dicha representación se determinan los métodos usados en la etapa de descubrimiento, y en consecuencia, el tipo de patrones descubiertos.

La figura 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 por ejemplo:

- 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.

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

Fig. 1. Estado del arte de la minería de texto

A continuación se describen brevemente los principales métodos empleados en ambas etapas de la minería de texto. Una descripción más completa del estado del arte del área, y una lista de referencias pertinentes puede consultarse en (Montes-y-Gómez, 2002; Hearst, 1999; Kodratoff, 1999; Tan, 1999).

2.1 Etapa de Preprocesamiento

En la etapa de preprocesamiento los textos se transforman a una representación estructurada o semiestructurada de su contenido. Estas 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.

Las representaciones intermedias más usadas en la minería de texto son básicamente de dos tipos:

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

La construcción de estas representaciones sigue diferentes estrategias. Por ejemplo, las representaciones a nivel documento se construyen típicamente usando métodos de categorización, de texto completo e indexamiento. 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.

2.2 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, una clasificación alternativa de la minería de texto 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 la minería de texto se pueden clasificar en tres enfoques: (i) descubrimientos a nivel representación, (ii) descubrimientos a nivel texto, y (iii) 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, categorización e indexamiento de los textos.

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.

Los métodos relacionados con la identificación de patrones de lenguaje se distinguen por considerar todas las palabras de los textos y mantener su orden relativo, es decir, usar representaciones de texto completo (full-text, en inglés). 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.

Por su parte los métodos relacionados con el agrupamiento de textos se caracterizan por utilizar 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-organizados. En particular estos métodos enfatizan la visualización e interpretación de los resultados. Por ejemplo, algunos emplean interfaces gráficas para analizar los agrupamientos, otros determinan una etiqueta descriptiva del contenido de cada grupo, y otros más determinan el documento representativo de cada clase. Adicionalmente, el agrupamiento de los textos se usa en el análisis exploratorio de las colecciones de textos, en la generación de resúmenes multidocumento, y en otras tareas de descubrimiento tales como la detección de asociaciones y desviaciones.

Descubrimientos a Nivel Mundo

Este enfoque considera distintas tareas, entre ellas el descubrimiento de asociaciones, la detección de desviaciones y el análisis de tendencias. Los métodos de este enfoque comparten las siguientes características: (i) emplean tanto representaciones de los textos a nivel concepto como a nivel documento; (ii) usan conocimientos de dominio, generalmente expresados en jerarquías de conceptos o conjuntos de predicados, y (iii) permiten que el usuario guíe el proceso de descubrimiento, especificando principalmente las regiones y los conceptos de mayor interés.

Entre los trabajos de descubrimiento de asociaciones destacan aquellos que plantean la detección de asociaciones temáticas no-exactas de la forma $similar(A) \Rightarrow B\ (confianza / soporte)$, y el uso de los elementos estructurados y no estructurados para la obtención de dichas relaciones.

Por su parte, los métodos de detección de desviaciones consideran la detección de los textos raros –con temática diferente al promedio– de una colección, así como la detección de los nuevos temas en una colección dinámica, por ejemplo en un flujo de noticias.

El análisis de tendencias se encarga de la descripción de la evolución de una colección de textos. Entre sus métodos destacan los siguientes dos enfoques: (i) la detección de temas de discusión con un comportamiento preestablecido, y la comparación de la temática de una colección en dos tiempos diferentes.

3 Minería de Texto usando Grafos Conceptuales

En esta sección se presentan los métodos principales de nuestro enfoque de minería de texto a nivel detalle basado en el uso de representaciones semánticas –grafos conceptuales– de los documentos. Primero se definen los criterios para la comparación de dos grafos conceptuales, y después de presentan algunos métodos para descubrir patrones interesantes (grupos, asociaciones y desviaciones) en una colección de grafos conceptuales. Los principios de la teoría de grafos conceptuales se introducen en el apéndice A, mientras que la transformación de los textos a grafos conceptuales es tratada en (Sowa and Way, 1986; Sowa, 1999).

3.1 Comparación de Grafos Conceptuales

El procedimiento general propuesto para la comparación de dos grafos conceptuales consiste de dos etapas: (i) el apareamiento de los grafos, y (ii) 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.

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 π no es necesariamente uno-a-uno y tampoco único (referirse al apéndice), 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" $\{\pi_1, \pi_2, \ldots, \pi_n\}$ tal que sus correspondientes proyecciones en G_1 y G_2 no se intercepten:

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

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 sólo 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.

Debido a que el apareamiento y la proyección de los grafos conceptuales son problemas definidos como NP-completos (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.

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).

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): (i) 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; (ii) la semejanza entre dos grafos conceptuales se relaciona con sus diferencias, entre más diferencias tengan, menos semejantes son; (iii) la mayor semejanza entre dos grafos conceptuales se obtiene cuando son idénticos, sin importar cuantos elementos comunes tengan, y (iv) 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. El valor de este coeficiente 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.

Medida de Semejanza

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

Semejanza Conceptual

La semejanza conceptual $0 \le s_c \le 1$ 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 mencionados en ambos grafos conceptuales.
La semejanza conceptual s_c se calcula usando una expresión análoga al coeficiente de Dice:

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

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

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

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

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

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

En la segunda condición, los conceptos $\pi_{G_1}c$ y $\pi_{G_2}c$ se refieren a diferentes "individuos" del mismo tipo, esto es, a diferentes instancias de la misma clase. En este caso, $\beta(\pi_{G_1}c, \pi_{G_2}c) = depth/(depth+1)$, 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 $\pi_{G_1}c$ y $\pi_{G_2}c$ tienen diferentes tipos, es decir, señalan elementos de distintas clases. En este caso, $\beta(\pi_{G_1}c, \pi_{G_2}c)$ expresa la semejanza semántica de los conceptos $type(\pi_{G_1}c)$ y $type(\pi_{G_2}c)$ en la jerarquía de conceptos preestablecida. Esta semejanza se calcula usando, una vez más, una expresión análoga al coeficiente de Dice:

$$\beta(\pi_{G_1}c, \pi_{G_2}c) = \frac{2 \times d_c}{d_{\pi_{G_1}c} + d_{\pi_{G_2}c}}$$

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

Semejanza Relacional

La semejanza relacional $0 \le s_r \le 1$ 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:

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

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:

$$s = s_c \times (a + b s_r)$$

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 $s = a s_c$.

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 $a + b = 1$.

3.2 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 es un 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 (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. 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 ($cov(h_i)$, $desc(h_i)$, $coh(h_i)$), donde:

- $cov(h_i)$, llamada cobertura de h_i, es el conjunto de grafos cubiertos por o asociados con la regularidad h_i.
- $desc(h_i)$, 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: $\forall G_i, G_j \in \text{cov}(h_i): sim(G_i, G_j) \geq coh(h_i)$.

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: $\text{cov}(h_j) \subset \text{cov}(h_i)$.
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:

$$S(h_i) = \left\{ h_j \in H \,\middle|\, h_j < h_i, \nexists h_k : h_j < h_k < h_i \right\}$$
$$P(h_i) = \left\{ h_j \in H \,\middle|\, h_i < h_j, \nexists h_k : h_i < h_k < h_j \right\}$$

Construcción de la Jerarquía Conceptual

El método propuesto emplea una estrategia de aprendizaje no supervisado que permite construir incrementalmente el agrupamiento conceptual del conjunto de grafos. Así, la incorporación de un 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 (ver la figura 2). 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 estrategia ascendente, 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 2(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:

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

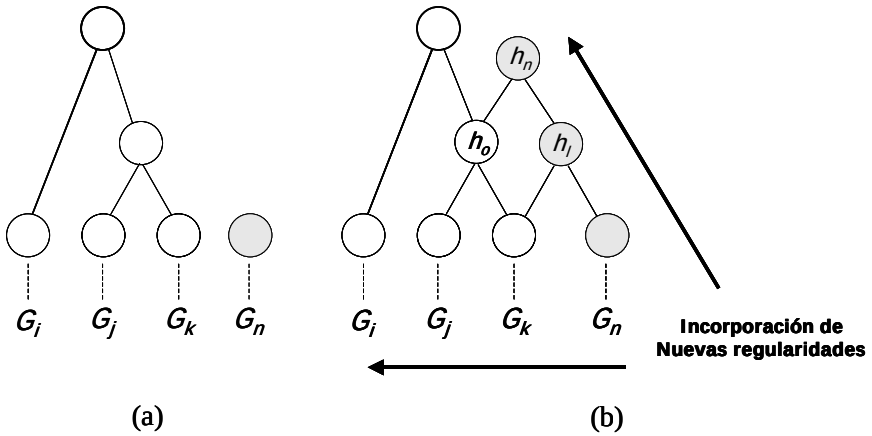


Fig. 2. Incorporación de un nuevo grafo a la jerarquía

En este caso, la función $match(G_i, G_j)$ regresa el mejor traslape de los grafos G_i y G_j; la función $sim(G_i, G_j)$ regresa la medida de semejanza de los grafos G_i y G_j; y la función $\min(coh(h_i), coh(h_j))$ 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.

3.3 Descubrimiento de Asociaciones

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.

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

Cada nodo h_i de esta jerarquía 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. Con base en la jerarquía conceptual H es posible determinar dos tipos de reglas asociativas.

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:

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

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

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

Además, $\forall h_j \in H$: $desc(h_j) < desc(h_i)$:

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

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

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

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 que debe ser eliminada.

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

- Existe otra regla asociativa implícita $g_h \Rightarrow g_l(1,\alpha)$, 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 $g_i \Rightarrow g_j(1,\beta)$ en combinación con la regla asociativa explícita $g_j \Rightarrow g_k(\gamma,\alpha)$, en donde, $g_k < g_j < g_i$.

3.4 Detección de Desviaciones

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 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%.

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

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: $S(h_i) = \{h_j \in H \mid h_j < h_i, \nexists h_k : h_j < h_k < h_i\}$, indica una partición del contexto h_i, donde 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 se establece 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: $|\text{cov}(h_j)| \geq m \times |\text{cov}(h_i)|$.

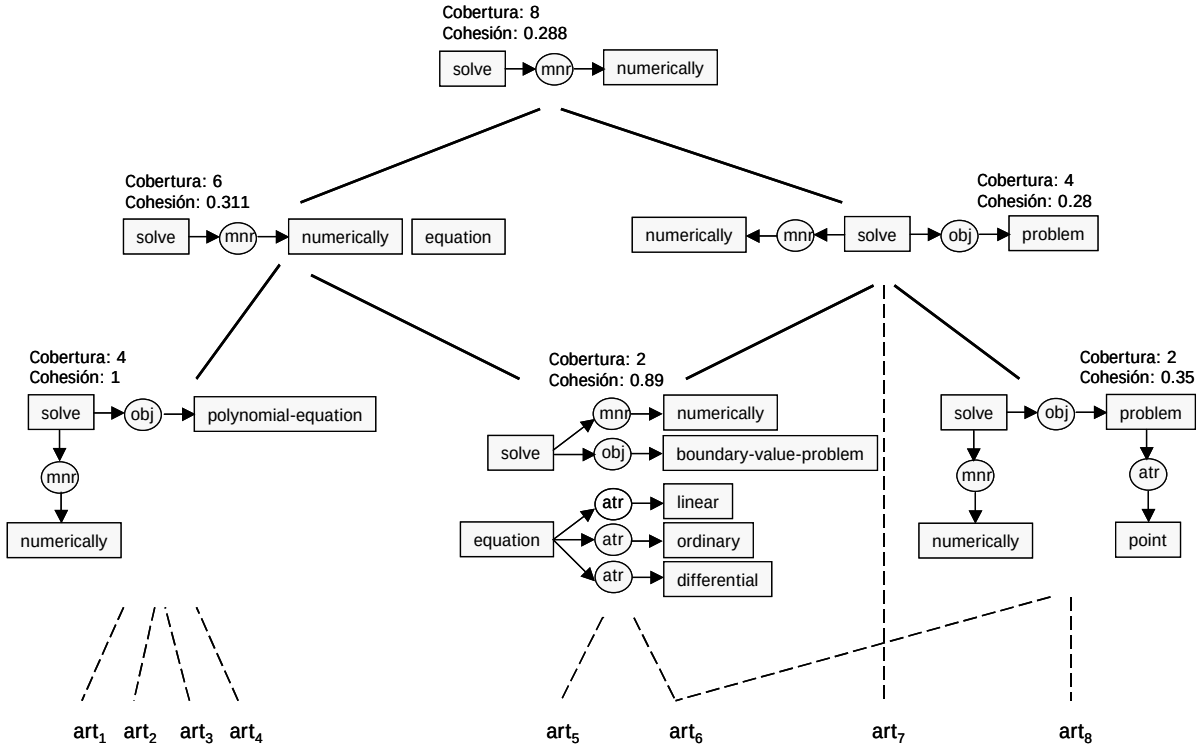
Entonces, el conjunto de características representativas del contexto h_i se define como:

$$F(h_i) = \{ desc(h_j) \mid h_j \in S(h_i), |\text{cov}(h_j)| \geq m \times |\text{cov}(h_i)| \}.$$

Grafo Conceptual Raro: El grafo conceptual $G_i \in \text{cov}(h_i)$ 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: $G_i \in \text{cov}(h_j)$.

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

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: $\forall G_i \in \text{cov}(h_k) \Rightarrow G_i \in R(h_i)$.



art$_1$ a art$_4$ – ...(numerical solution of the polynomial equation)...
art$_5$ – ...(the numerical solution of boundary value problems for linear ordinary differential equations)...
art$_6$ – ...(the numerical solution of an n-point boundary value problem for linear ordinary differential equations)...
art$_7$ - ...(the numerical solution of a thin plate heat transfer problem)...
art$_8$ - ...(the numerical solution of nonlinear two-point boundary problems by finite difference methods)...

Fig. 3. Grupo interno de la colección B

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

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

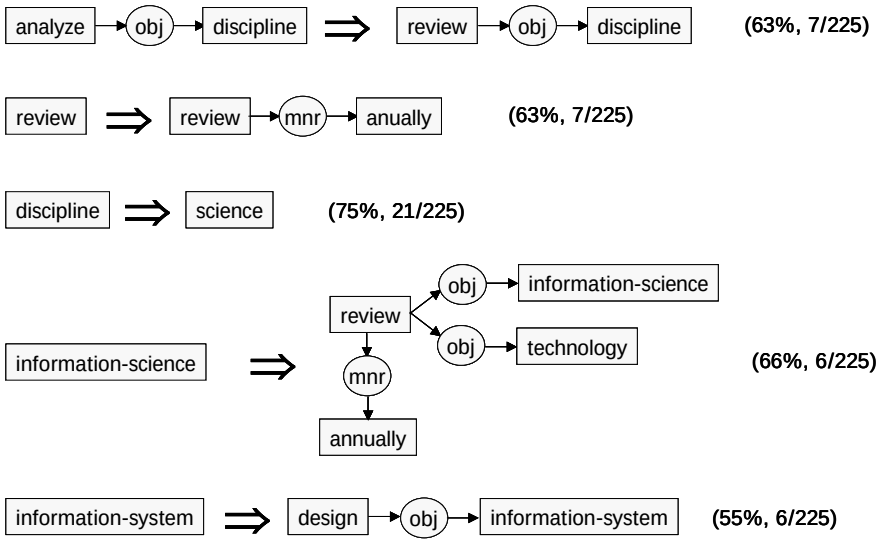
Esta definición permite encontrar todas las desviaciones contextuales –en un conjunto de grafos conceptuales con respecto a un valor predefinido de m. Muchas de estas desviaciones contienen información redundante o información implícita en otras desviaciones. Por ejemplo, si es raro que se hable de animales 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:

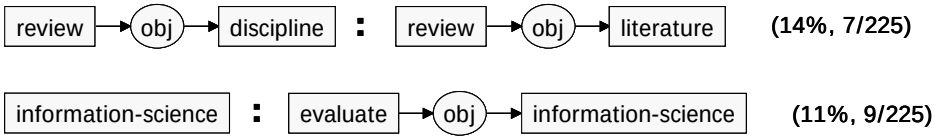
Desviación Contextual Redundante: La desviación contextual $g_i : g_k(\alpha, \beta)$ es redundante si existe otra desviación contextual $g_i : g_j(\gamma, \beta)$, con $\alpha < \gamma$, tal que g_j es una generalización de g_k. Esto implica que: $\mathrm{cov}(g_k) \subset \mathrm{cov}(g_j)$.

4 Resultados Experimentales

Nuestro método de minería de texto fue probado mediante el 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.



(a) Algunos ejemplos de asociaciones



(b) Algunos ejemplos de desviaciones

Fig. 4. Patrones descriptivos del conjunto A

Los resultados descritos a continuación son de dos tipos, cualitativos y cuantitativos. Los resultados cualitativos muestran la capacidad de nuestro método para descubrir patrones interesantes a un nivel más descriptivo y completo que el temático. Por su parte, los resultados cuantitativos demuestran la viabilidad de nuestro método de minería de texto.

4.1 Evaluación Cualitativa

Nuestro método de minería de texto permite descubrir patrones más descriptivos sobre el contenido de los textos que los métodos tradicionales. En esta sección se muestran algunos ejemplos de estos patrones; en especial se muestran algunos grupos (segmentos del agrupamiento jerárquico), asociaciones y desviaciones obtenidas a partir de los conjuntos de prueba.

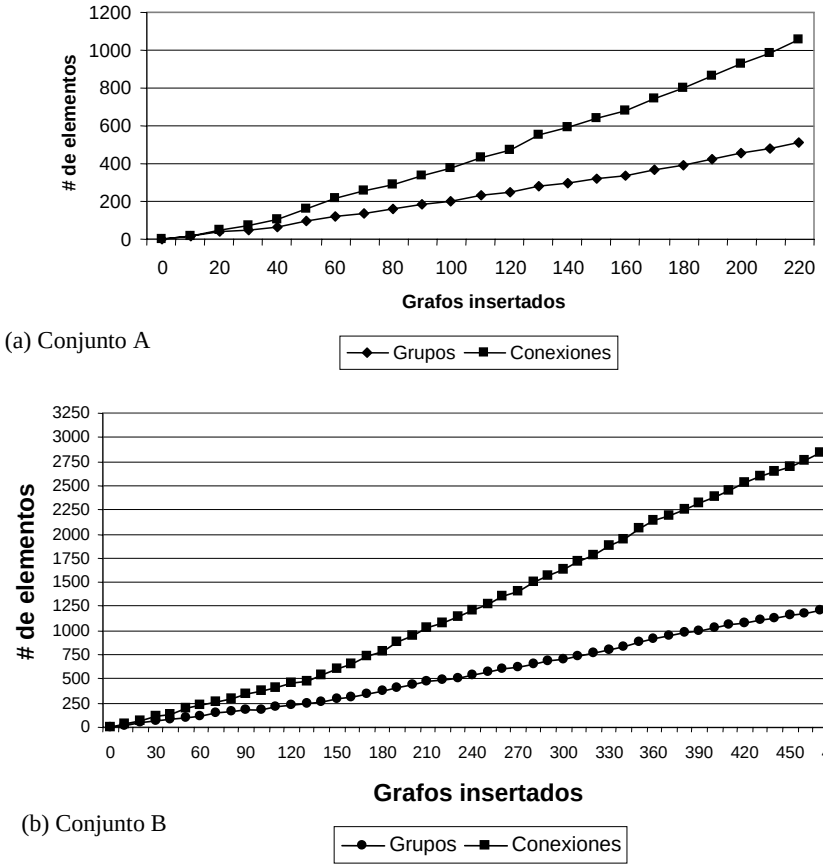
Agrupamiento Conceptual

El agrupamiento del conjunto A generó una jerarquía conceptual de 510 nodos; donde 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 3 se presenta un grupo interno de una de las jerarquías conceptuales obtenidas.

Asociaciones y Desviaciones

Los métodos de descubrimiento de asociaciones y detección de desviaciones disminuyen el problema de interpretación contextual de los patrones descubiertos. Básicamente generan reglas que preservan las relaciones semánticas (o de cualquier tipo de relación representada en los grafos) entre los conceptos participantes, y además permiten determinar patrones a diferentes niveles de generalización.

La figura 4 muestra algunos ejemplos de asociaciones y desviaciones correspondientes al conjunto de prueba A. Estos patrones indican que una parte importante de los artículos del conjunto A se enfoca en el análisis de distintas disciplinas; siendo las ciencias las más analizadas y la literatura la menos. Además, señala que estos análisis son en la mayoría de las veces revisiones anuales.



(a) Conjunto A

(b) Conjunto B

Fig. 5. Crecimiento del agrupamiento

4.2 Evaluación Cuantitativa

Crecimiento del Agrupamiento Conceptual

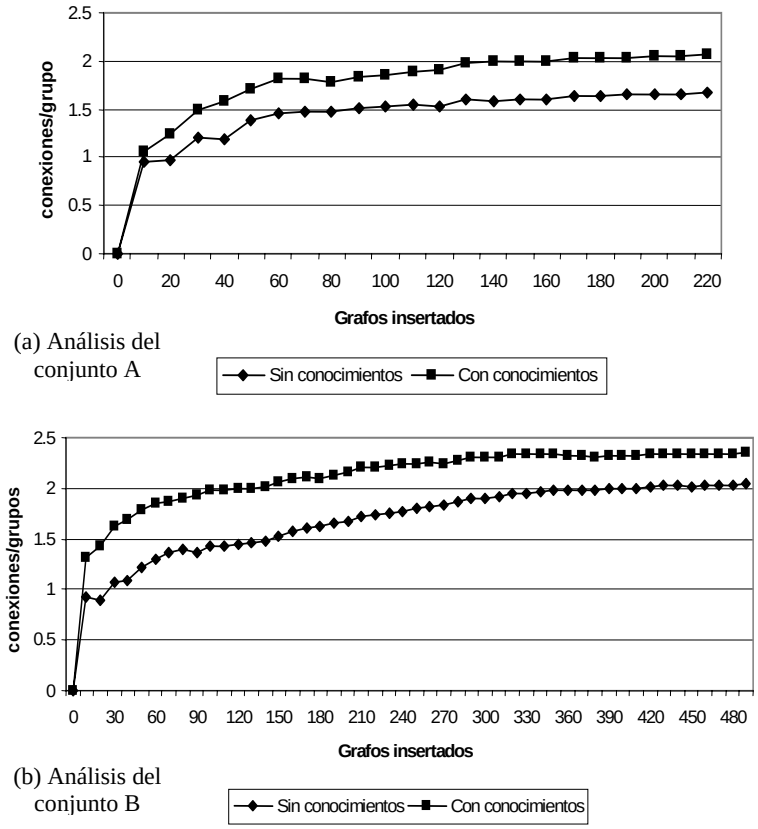
Los agrupamientos conceptuales tienen características muy interesantes para los propósitos de descubrimiento de conocimiento y minería de texto, por ejemplo, son muy descriptivos y altamente estructurados. Sin embargo, su tamaño (que puede ser exponencial con respecto al número de grafos del conjunto) limita considerablemente su aplicación.
Nuestros experimentos demostraron que el agrupamiento conceptual de un conjunto de grafos conceptuales que representan el contenido de textos es factible. Por ejemplo, en la figura 5 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. Además de que esta tendencia se mantiene cuando se emplea el conocimiento del dominio.
2. El impacto del conocimiento 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.

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. La figura 6 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. 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.



(a) Análisis del conjunto A

(b) Análisis del conjunto B

Fig. 6. Densidad de conexiones

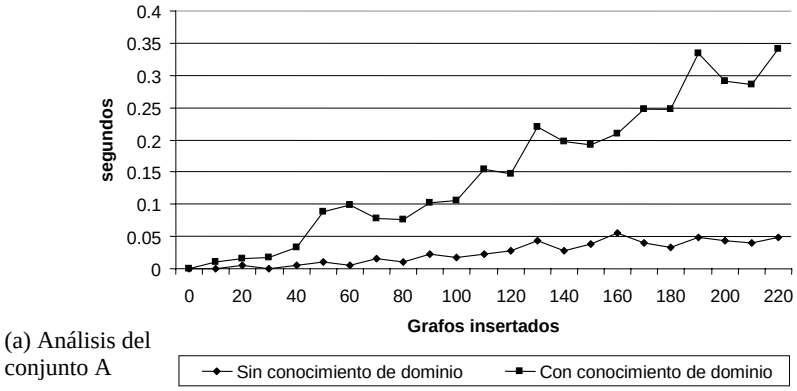
2. La densidad de conexiones aumenta cuando se usa conocimiento del dominio. Este incremento es casi constante a través de todo el proceso de construcción.

Tiempo de Construcción

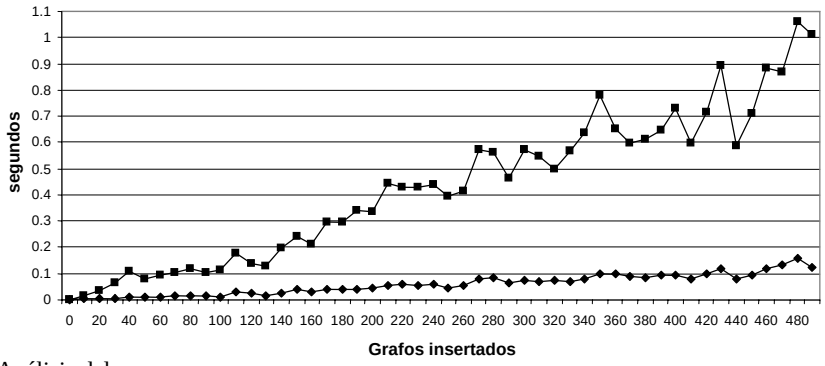
Las gráficas de crecimiento del agrupamiento y de densidad de conexiones exponen algunas ventajas del uso de conocimiento del dominio en la construcción del agrupamiento de los grafos conceptuales. Básicamente, estas gráficas muestran que este conocimiento permite encontrar grupos más grandes y más homogéneos (mejor interconectados).
Estas ventajas tienen un costo principal: el tiempo de construcción del agrupamiento. En la figura 7 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 tiempo de inserción del grafo 495 del conjunto B en la jerarquía conceptual necesito solamente de un segundo.

Analizando la figura 7 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.



(a) Análisis del conjunto A



(b) Análisis del conjunto B

Fig. 7. Tiempo de construcción del agrupamiento

5 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 esta idea, 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.
- 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 conceptuales 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.

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. Las principales limitaciones son:

- *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 jerarquías pequeñas puede ocasionar que no se detecten semejanzas posiblemente interesantes entre los grafos.

Segundo problema: La transformación automática de los textos a grafos conceptuales no es una tarea sencilla. Sus principales efectos son:

- *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, 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.
- *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.

6 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:
4. *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.
5. *Aplicar los nuestros métodos de análisis de grafos conceptuales en la minería semántica de la web*.

Referencias

1. **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.
2. **Kodratoff (1999)**, Knowledge Discovery in Texts: A Definition and Applications, Proc. of the 11th International Symposium on Foundations of Intelligent Systems (ISMIS-99), 1999.
3. **Lin (1998)**, An Information-Theoretic Definition of Similarity, Proc. of the International Conference on Machine Learning, Madison, Wisconsin, 1998.
4. **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.
5. **Montes y Gómez (2002),** Minería de texto empleando la Semejanza entre Estructuras Semánticas. Tesis de Doctorado, Centro de Investigación en Computación, Instituto Politécnico Nacional, México, Febrero 2002.
6. **Mugnier (1995)**, On generalization/specialization for conceptual graphs, Journal of Experimental and Theoretical Artificial Intelligence, Vol. 7, 1995.
7. **Sowa (1984)**, Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley, reading, M.A., 1984.
8. **Sowa (1999)**, Knowledge Representation: Logical, Philosophical and Computational Foundations, Thomson Learning, 1999.
9. **Sowa and Way (1986)**, Implementing a semantic interpreter using conceptual graphs, IBM Journal of Research and Development 30:1, January, 1986.
10. **Sparck-Jones (1999),** What is the Role of NLP in Text Retrieval?, In Strzalkowski Ed., Natural Language Information Retrieval, Kluwer Academic Publishers, 1999.
11. **Tan (1999)**, Text Mining: The state of the art and challenges, Proc. of the Workshop Knowledge Discovery from advanced Databases PAKDDD-99, Abril 1999.

Apéndice A. Grafos Conceptuales

A.1 Terminología Básica

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 (Sowa, 1984).
Por ejemplo, el grafo [gato:Felix]→(sobre) →[sillón] →(attr)→[negro] 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.

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] tiene el tipo gato y el referente Félix.

Tipos conceptuales: Los tipos conceptuales se organizan en una jerarquía de tipos. 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*.

Referentes: Los referentes son de dos clases: genéricos e individuales. Los referentes genéricos se refieren a conceptos no especificados. Por ejemplo, el concepto [sillón] significa un sillón. Por su parte, los referentes individuales funcionan como sustitutos de elementos específicos del mundo real. Por ejemplo, el concepto [gato:Félix] es un sustituto del gato Félix –que existe en algún lugar.

Relación conceptual: Las relaciones conceptuales señalan la manera en que los conceptos se interrelacionan. Ellas tienen un tipo relacional 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.

A.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 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 este apéndice sólo se analizan las reglas canónicas de generalización.

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 separar 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.
- 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*.

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

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

La proyección π 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 πx_1 y πx_2 en *u*, tal que $\pi x_1 = \pi x_2$). Además, la proyección π tampoco es necesariamente única, es decir, un grafo *v* puede tener dos proyecciones diferentes en *u*, $\pi' v$ y πv, donde $\pi' v \neq \pi v$.

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 $v' \leq v$.

Manuel Montes y Gómez. *Recibió, con mención honorífica, el grado de Doctor en Ciencias de la Computación (2002) del Centro de Investigación en Computación del IPN, México. Actualmente es Investigador Titular de la Coordinación de Ciencias Computacionales del Instituto Nacional de Astrofísica, Óptica y Electrónica. También es miembro del Sistema Nacional de Investigadores de México. Su principal interés de investigación es el procesamiento automático de textos; área en la que ha publicado más de cincuenta artículos científicos en revistas y congresos internacionales, y ha dirigido dos tesis de postgrado.*

Alexander Gelbukh. *Recibió, con mención honorífica, su grado de Maestro en Ciencias Matemáticas (1990) de la Universidad Estatal "Lomonósov" de Moscú, Rusia, y su grado de Doctor en Ciencias de la Computación (1995) del VINITI, Rusia. Actualmente es Profesor-Investigador del Centro de Investigación en Computación del IPN, jefe del Laboratorio de Lenguaje Natural y Procesamiento de Texto. Es miembro de la Academia Mexicana de Ciencias y del Sistema Nacional de Investigadores (SNI) de México, autor de alrededor de 300 publicaciones en la lingüística computacional y procesamiento automático de texto, recuperación de información y áreas afines; véase www.Gelbukh.com.*

Aurelio López López. *Es investigador de la Coordinación de Ciencias Computacionales del INAOE, en Tonantzintla, Puebla, México. Obtuvo el grado de Doctor en Ciencias de la Computación y la Información (1995) de la Universidad de Syracuse, Nueva York, EUA.. Sus intereses de investigación incluyen la representación del conocimiento, la recuperación y extracción de información, el procesamiento de lenguaje natural y la minería de texto.*