Minería de texto: Un nuevo reto computacional

Manuel Montes-y-Gómez

Laboratorio de Lenguaje Natural,
Centro de Investigación en Computación,
Instituto Politécnico Nacional.
Av. Juan de Dios Batís, Zacatenco, 07738 México, D.F.
mmontesg@susu.inaoep.mx

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, informes técnicos, etcétera. La posesión real de todo este conocimiento depende de nuestra habilidad para hacer ciertas operaciones con la información, por ejemplo:

- Buscar la información necesaria
- Comparar fuentes de información diferentes, y obtener conclusiones
- Manejar los textos, por ejemplo, traducirlos, editarlos, etc.

La lingüística computacional, y en particular el procesamiento automático de textos se enfocan en la solución de todos estos problemas, o este gran problema.

El objetivo de este artículo es introducir los conceptos y las tareas básicas del procesamiento de textos, haciendo énfasis en su nueva área de investigación: *la minería de texto*.

La minería de texto se enfoca en el descubrimiento de patrones interesantes y nuevos conocimientos en un conjunto de textos, es decir, su objetivo es descubrir cosas tales como tendencias, desviaciones y asociaciones entre "gran" la cantidad de información textual.

Al final del artículo se presentan las ideas básicas del sistema de minería de texto desarrollado en el Laboratorio de Lenguaje Natural y Procesamiento de Textos del Centro de Investigación en Computación del Instituto Politécnico Nacional. Este sistema se basa en el uso de grafos conceptuales para la representación del contenido de los textos, y se fundamenta en dos tareas: la comparación de dos grafos conceptuales cualesquiera y el agrupamiento conceptual de un conjunto de dichos grafos.

2 Lingüística computacional y procesamiento de textos

La lingüística computacional es la ciencia que trata de la aplicación de los métodos computacionales en el estudio del lenguaje natural (Gelbukh and Bolshakov, 1999). Esta ciencia es una combinación de dos ciencias más grandes; la lingüística, que estudia las leyes del lenguaje humano, y la inteligencia artificial, que investiga los métodos computacionales para el manejo de sistemas complejos (ver figura 1).

El problema u objetivo más importante de la lingüística computacional es la *comprensión del lenguaje*, es decir, la transformación del lenguaje hablado o escrito a una representación formal del conocimiento, como por ejemplo una red semántica.

La solución tradicional de este problema consiste en construir un procesador lingüístico constituido por diferentes módulos independientes (ver figura 2):

- El *módulo morfológico* se encarga de reconocer las palabras. Básicamente, convierte las cadenas de letras a una entrada de un diccionario, y pone las marcas de tiempo, género y número.
- El *módulo sintáctico* reconoce oraciones. Este módulo convierte las cadenas de palabras marcadas a una estructura grafica, en donde se hacen explicitas algunas relaciones entre las palabras de la oración.

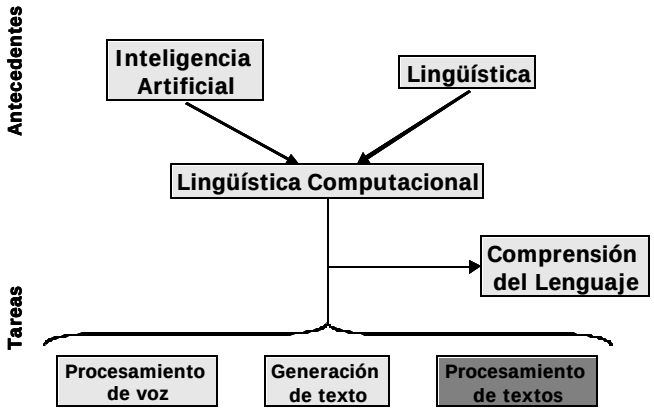

Figura 1. Antecedentes y tareas de la lingüística computacional

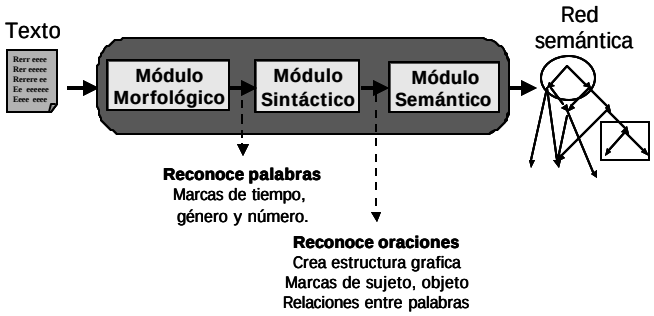

Figura 2. Procesador lingüístico

- El *módulo semántico* reconoce la estructura completa del texto y lo convierte a una "red semántica".

La lingüística computacional se encarga de otras cosas, adicionalmente a la comprensión del lenguaje. Algunas de estas otras áreas de investigación de la lingüística computacional se muestran en la figura 2. La más grande de estas áreas, y tal vez la más importante, es el procesamiento automático de textos. El procesamiento automático de textos considera una gran diversidad de tareas (ver figura 3), desde muy simples, como la separación de palabras, hasta muy complejas como algunas tareas de *minería de texto*.

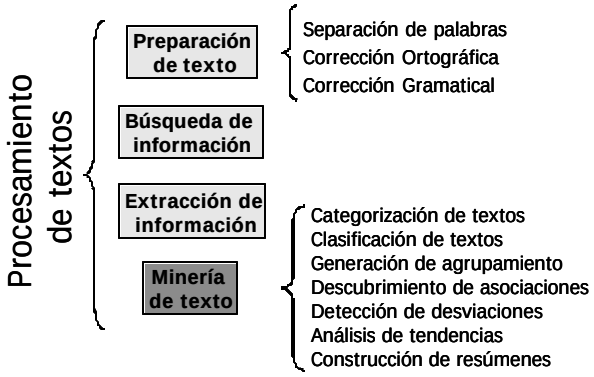

Figura 3. Tareas del procesamiento de textos

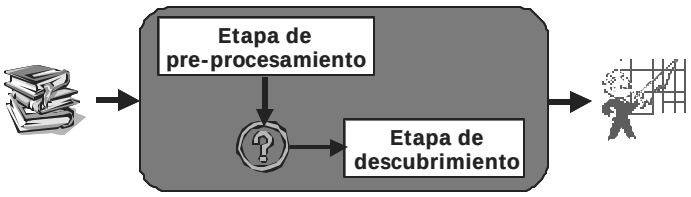

Figura 4. Proceso de minería de texto

3 Minería de texto

La minería de texto es la más reciente área de investigación del procesamiento de textos. Ella se define como el proceso de descubrimiento de patrones interesantes y nuevos conocimientos en una colección de textos, es decir, la minería de texto es el proceso encargado del descubrimiento de conocimientos que no existían explícitamente en ningún texto de la colección, pero que surgen de relacionar el contenido de varios de ellos (Hearst, 1999; Kodratoff, 1999).

Este proceso consiste de dos etapas principales: una etapa de *pre-procesamiento* y una etapa de *descubrimiento* (Tan, 1999).

En la primera etapa, los textos se transforman a algún tipo de representación estructurada o semi-estructurada 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. La figura 4 ilustra este proceso.

Dependiendo del tipo de métodos usados en la etapa de pre-procesamiento es el tipo de representación del contenido de los textos construida; y dependiendo de esta representación, es el tipo de patrones descubiertos.

La figura 5 muestra los tres tipos de estrategias empleadas en los

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 5. Estado del arte de la minería de texto

actuales sistemas de minería de texto. Como se observa, todos estos métodos limitan a un nivel temático o de entidad sus resultados, haciendo imposible descubrir cosas mas 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 educción.
- *Desviaciones*, que identifiquen por ejemplo opiniones "raras" con respecto al desempeño de la selección mexicana de fútbol.

Una idea para mejorar la expresividad y diversidad de los descubrimientos de los sistemas de minería de textos consiste en usar alguna mejor representación del contenido de los textos; por ejemplo, los grafos conceptuales (Sowa, 1984; Sowa, 1999).

Esta solución involucra dos problemas diferentes (ver figura 6):

- La transformación de los textos en grafos conceptuales
- El análisis automático de un conjunto de grafos conceptuales

La transformación de los textos en grafos conceptuales es un problema complejo vinculado con el análisis sintáctico y semántico de los textos (Sowa and Way, 1986; Sowa 1991). Algunos tipos de textos transformados automáticamente en grafos conceptuales son:

- Algunas partes de *artículos científicos* (Myeng and Khoo, 1994;

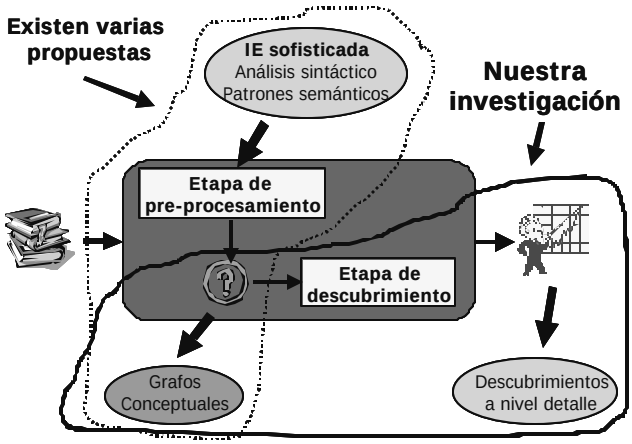

Figura 6. Nuestra propuesta de minería de texto

Montes-y-Gómez *et al*., 1999).

- Algunas partes de *expedientes médicos* (Baud *et al.*, 1992).
- Algunas partes de *casos legales* (Boucier and Rajman, 1994).

Por su parte, el análisis automático de un conjunto de grafos conceptuales orientado al descubrimiento de nuevos conocimientos es un problema poco estudiado. Solamente existen dos trabajos relacionados con el agrupamiento de grafos conceptuales (Mineau and Godin, 1995; Bournaud and Ganascia, 1996).

Así pues, este nuestra investigación, y este articulo en concreto, se enfoca en el análisis de un conjunto de grafos conceptuales, y en el descubrimiento de *agrupamientos, asociaciones* y *desviaciones* interesantes a partir de ellos.

4 Minería de texto con grafos conceptuales

A continuación se definen los grafos conceptuales y se presentan varias ideas para el análisis de un conjunto de estos grafos. Estas ideas se ilustran de manera breve y sencilla, omitiendo todo detalle matemático y de diseño. Si el lector está interesado en estos detalles debe referirse a las siguientes publicaciones (Montes-y-Gómez *et al*., .2001a; Montes-y-Gómez *et al.*, 2001b; Montes-y-Gómez *et al*.,.2001c).

4.1 Grafos Conceptuales

Los grafos conceptuales son un sistema de lógica orientado a la representación de la semántica del lenguaje natural (Sowa, 1984). Básicamente, un grafo conceptual es un grafo *bipartito* que tiene dos tipos diferentes de nodos: conceptos y relaciones conceptuales.

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, y el referente indica el elemento específico (instancia de la clase) referido por éste. Por ejemplo, para el concepto [gato:Félix], su tipo es gato y su referente es Félix.

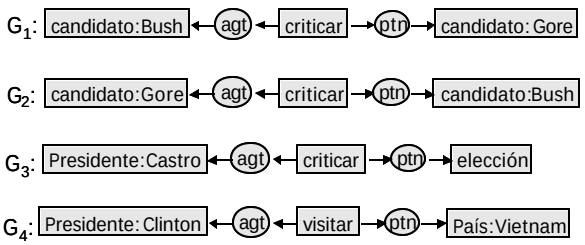

Figura 7. Un conjunto de grafos conceptuales sencillos

Las relaciones conceptuales señalan la manera en que los conceptos se inter-relacionan. 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. La figura 7 muestra un conjunto de grafos conceptuales sencillos. Todos estos grafos tienen tres conceptos y dos relaciones conceptuales. Por ejemplo, el primero de ellos representa la frase "Bush critico a Gore".

4.2 Comparación de grafos conceptuales

Una de las operaciones básicas para el agrupamiento de los grafos conceptuales es, sin lugar a dudas, su *comparación*. Algunas características de nuestro método de comparación de grafos conceptuales son:

- Utiliza *conocimiento del dominio*, básicamente 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.

- 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 con respecto a la información contenida en los grafos conceptuales originales.

- Calcula de acuerdo con *los intereses del usuario.*

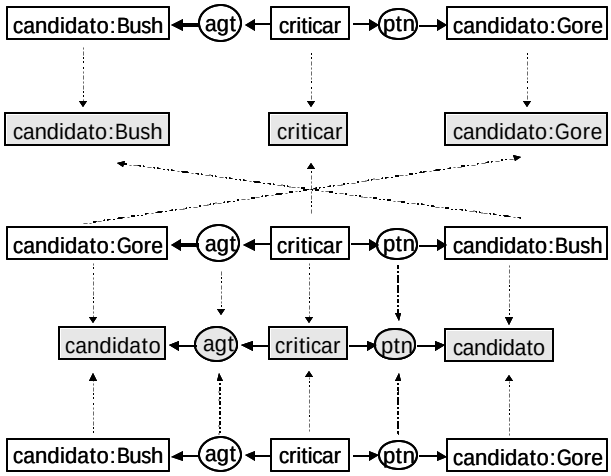

Figura 8. Comparación flexible de los grafos conceptuales

El usuario indica, mediante un conjunto de parámetros, si le interesa mas la semejanza conceptual o relacional, y también si le interesan mas la semejanza entre entidades, acciones o atributos.

La figura 8 ilustra la flexibilidad de la comparación de los grafos. Allí se muestra que la semejanza de dos grafos puede analizarse desde *diferentes perspectivas*.

4.3 Agrupamiento de los grafos conceptuales

El agrupamiento de los grafos permite descubrir la *estructura oculta* de la colección de textos, así como construir un *resumen organizado* de la colección que facilita su posterior análisis, y por tanto, el descubrimiento de otros tipos de patrones interesantes.

Nosotros usamos 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 es una especie de red de herencia, donde los nodos inferiores indican regularidades especializadas y los nodos supe-

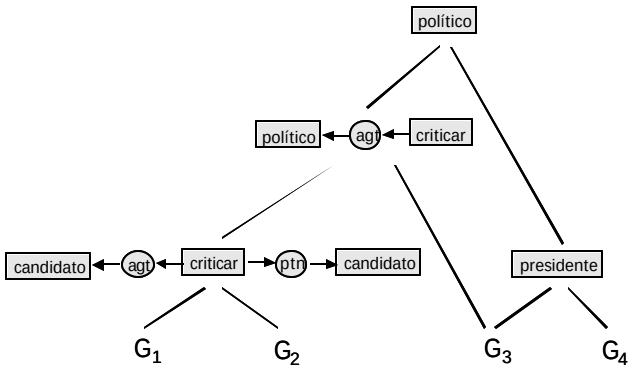

Figura 9. Agrupamiento de grafos conceptuales

riores sugieren regularidades generalizadas. La figura 9 ilustra el agrupamiento conceptual de los grafos de la figura 8.

En este agrupamiento cada nodo h_i se representa por una triada ($cov(h_i)$, $desc(h_i)$, $coh(h_i)$), 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 $cov(h_i)$.

4.3 Asociaciones y desviaciones

El descubrimiento de asociaciones y la detección de desviaciones se basan el agrupamiento conceptual de los grafos. Estos métodos consideran el agrupamiento conceptual como un *índice de la colección*, y aprovechan su estructura para localizar más fácilmente las asociaciones y las desviaciones.

Además, esta estrategia permite detectar asociaciones a diferentes niveles de generalización y desviaciones respecto a diferentes contextos (subconjuntos de la colección).

La figura 10 muestra dos reglas asociativas y una desviación contextual correspondientes al conjunto de grafos conceptuales de la figura 7, y obtenidas a partir de la jerarquía conceptual de la figura 9. Los valores entre paréntesis, para el caso de las asociaciones, corresponden a los valores de confianza y soporte, y para el caso de la desviación indican

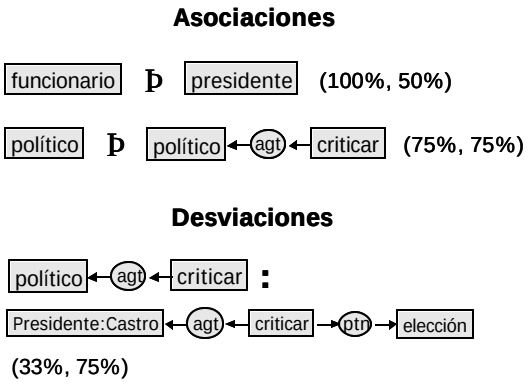

Figura 10. Asociaciones y desviaciones

los valores de rareza y soporte.

Por ejemplo, la primera asociación indica que el 100% de los grafos que mencionan a un funcionario, hablan de un presidente; además que el 50% de los grafos del conjunto hablan sobre algún presidente".

Por su parte, la desviación contextual indica que dentro de subconjunto de grafos relacionados con políticos criticando, el cual representa el 75% del conjunto de grafos, es raro que el presidente Castro critique las elecciones estadounidenses; solamente el 33% de los grafos mencionan este suceso[1].

5. Conclusiones y trabajo futuro

5.1 Importancia de la lingüística computacional

Sin lugar a dudas, la lingüística computacional en su conjunto enfrenta uno de los más grandes retos de la ciencia computacional: lograr que las computadoras sean nuestros verdaderos ayudantes en la ocupación principal de la raza humana, pensar y comunicar.

Además, las tareas de la lingüística computacional tienen una gran *utilidad practica inmediata*, ya que se relacionan con: la toma de deci-

[1] Este valor (33%) es muy alto para ser considerado una desviación en un caso real, pero es un valor adecuado si se considera que el ejemplo consiste en una colección de cuatro grafos.

siones, la búsqueda e intercambio de conocimiento, y toda clase de operaciones relacionadas con la publicación y uso de los documentos.

Así pues, sin temor a equivocarnos, podemos decir que los países que disponen de buenas herramientas para el análisis y generación de textos tienen, en nuestro mundo competitivo, una gran ventaja económica, tecnológica y hasta militar sobre los demás países. Desgraciadamente, la mayoría de los logros actuales de la lingüística computacional se orientan al inglés.

5.2 Nuestra contribución en minería de texto

Actualmente, la mayoría de los sistemas de minería de texto emplean *representaciones sencillas* del contenido de los textos, por ejemplo, listas de conceptos o palabras clave. Esto facilita el análisis de los textos, pero a la vez limita la variedad, expresividad y utilidad de los patrones descubiertos. Básicamente, estas representaciones *limitan a un nivel temático* los patrones descubiertos.

Nuestra investigación se enfocó principalmente en la solución de este problema. En ella se planteó el uso de una representación más completa del contenido de los textos (una representación que también considera la información estructural de los textos), y se propuso un método de minería de texto capaz de analizar estas representaciones, pero además también capaz de trasladar los descubrimientos del actual nivel temático a un nivel de *mayor detalle*, un *nivel conceptual*.

En específico, en este trabajo se planteó el uso de los *grafos conceptuales* como la representación del contenido de los textos, y se diseñaron algunos métodos, previamente poco estudiados, para las siguientes tareas:

- La *comparación* de dos grafos conceptuales.
- El *agrupamiento conceptual* de un conjunto de grafos conceptuales.
- El *descubrimiento de asociaciones* entre grafos conceptuales.
- La *detección de 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, la minería de datos y la teoría de grafos conceptuales.

4.3 Investigación posterior

El trabajo futuro que se desprende de esta investigación involucra, entre otros, los siguientes aspectos:

- *Corto plazo*: Diseño de otras tareas de minería de texto, por ejemplo, el análisis de tendencias.
- *Mediano plazo*: Desarrollo de un método flexible para la transformación texto → grafo conceptual, y la aplicación del método de comparación de los grafos conceptuales en la búsqueda de información.
- *Largo plazo*: Aplicación de estos métodos en la minería semántica de la web.

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] Gelbukh and Bolshakov (1999), Avances en Análisis Automático de Textos. Proc. Foro: Computación, de la Teoría a la Práctica. IPN, Mexico City, May 26 – 28, 1999.

[5] 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.

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

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

[8] 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.

[9] Manuel Montes-y-Gómez, Alexander Gelbukh, Aurelio López-López, Ricardo Baeza-Yates (2001a). Flexible Comparison of Conceptual Graphs. To appear in the Proc. of DEXA-2001, edited by Springer Verlag.

[10] Manuel Montes y Gómez, Alexander Gelbukh, Aurelio López-

López, Ricardo Baeza-Yates (2001b). Un Método de Agrupamiento de Grafos Conceptuales para Minería de Texto. Aceptado para publicación en "Procesamiento de Lenguaje Natural", revista editada por la Sociedad Española de Lenguaje Natural, ISSN 1135-5948, Vol. 27, Septiembre 2001.

[11] Manuel Montes-y-Gómez, Alexander Gelbukh, Aurelio López-López (2001c). Discovering Association Rules in Semi-structured Data Sets. To appear in the Proc. of the Workshop on Knowledge Discovery from Distributed, Dynamic, Heterogeneous, Autonomous Data and Knowledge Sources, International Joint Conference on Artificial Intelligence (IJCAI'2001).

[12] 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.

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

[14] 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.

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

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

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