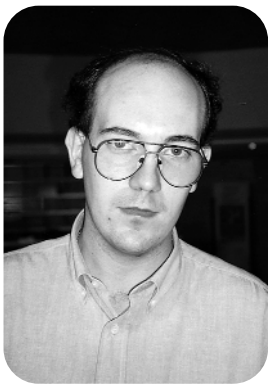
Artículo

Minería textual

Por **Ricardo Eíto Brun** y **Jose A. Senso**



Ricardo Eíto Brun, licenciado en documentación, es profesor de la Universidad Carlos III de Madrid, trabajo que compagina con el de responsable de documentación en Adecco Wwit. Ha publicado numerosos artículos sobre lenguajes de etiquetado, autor de un libro sobre XML y responsable de numerosos cursos especializados para profesionales.

Resumen: *Este artículo trata de establecer una definición de minería de textos así como delimitar su relación con otras disciplinas: recuperación textual, minería de datos y lingüística computacional. Se analizan además el impacto de la minería textual, algunas de las aplicaciones comerciales existentes en el mercado y, por último, se realiza una breve descripción de las técnicas utilizadas para desarrollar e implementar sistemas de minería de textos.*

Palabras clave: *Minería de texto, Minería de datos, Recuperación de información, Agrupación de documentos, Categorización, Clasificación de conceptos.*

Title: Text mining

Abstract: *This article attempts to establish a definition for "text mining" and, at the same time, to identify its relationship with other fields: text retrieval, data mining and computational linguistics. In addition, there is an analysis of the impact of text mining, a reference to existing commercial applications on the market and, lastly, a brief description of the techniques used for developing and implementing text mining systems.*

Keywords: *Text minig, Data mining, Information retrieval, Clustering, Categorizing, Concept classification.*

***Eíto Brun, Ricardo** y **Senso, Jose A.** "Minería textual". En: El profesional de la información, 2004, enero-febrero, v. 13, n. 1, pp. 11-27.*

Jose A. Senso, doctor en Documentación, es profesor de la Universidad de Granada desde 1998 y subdirector de la revista El profesional de la información. Sus líneas de investigación se centran en sistemas de metadatos, especialmente rdf, el desarrollo e implementación de la web semántica y Topic Maps.

Introducción

La minería textual es una de las tecnologías que, desde su formulación inicial a principios de la década de los noventa, ha tenido un mayor impacto en las actividades relacionadas con la inteligencia militar. Si bien este impacto nunca ha alcanzado el nivel de generalización de la minería de datos, los desafortunados acontecimientos del 11 de septiembre de 2001 hicieron que distintos medios prestasen atención a las tecnologías empleadas por las organizaciones policiales encargadas de luchar contra el terrorismo. Así, a partir de esa fecha podemos encontrar un mayor número de referencias al uso de la minería textual y de datos con este propósito.

En este trabajo se ofrece una visión introductoria a la "minería textual" —*text mining*—. La minería textual es una aplicación de la lingüística computacional y del procesamiento de textos que pretende facilitar la identificación y extracción de nuevo conocimiento a partir de colecciones de documentos o corpus textuales. Relacionada con la minería de datos (desde una perspectiva comercial podríamos decir que la minería textual es "la hermana pequeña" de la minería de datos), la diferencia entre estas dos aplicaciones está en que con esta última se pretende extraer conocimiento a partir de los patrones observables en grandes colecciones de datos estructurados que se almacenan en bases de datos relacionales. En el caso de la minería textual, se tomará como punto de partida para la extracción de nuevo conocimiento repositorios documentales o texto. Es decir, *información no estructurada*.

Definición de *minería textual*

En primer lugar debemos delimitar el alcance del término "minería textual" o "minería de textos", que utilizaremos como sinónimos en este documento. Acotar el significado de este término —y de la minería textual como disciplina—, no resulta fácil, ya que en ella confluyen distintas técnicas y principios teóricos desarrollados en otras disciplinas, mucho antes de que se comenzase a hablar de minería textual y de datos como tales. La minería textual recoge distintas técnicas

Artículo recibido el 2-11-03
Aceptación definitiva: 7-01-04

formuladas en el ámbito de la recuperación textual —o *text retrieval*— y la lingüística computacional.

Por otra parte, desde una perspectiva comercial, algunos fabricantes de aplicaciones para minería textual la presentan como una aplicación complementaria a la minería de datos que analiza y pretende identificar patrones en los datos almacenados en repositorios de información estructurada (bases de datos relacionales y almacenes de datos o *data warehouses*).

Si recurrimos a la literatura publicada sobre el tema, encontramos distintas definiciones. **Dan Sullivan** (2001, p. 324), autor de una de las pocas monografías dedicadas en exclusividad al tema, recoge dos de ellas: la primera define minería textual como *cualquier operación realizada para extraer y analizar textos procedentes de distintas fuentes externas con el objetivo de obtener inteligencia*. La segunda define minería textual como el *descubrimiento de información y conocimiento que anteriormente no se conocía, a partir de corpus textuales*.

Esta segunda definición coincide en líneas generales con la que quizá sea la más popular y que formuló **Marti A. Hearst** en su artículo *Untangling text data mining*. En ese texto, que se considera una lectura obligada como introducción a la minería textual, **Hearst** señala que ésta tiene como objetivo *descubrir información y conocimiento que previamente se desconocía, y que no aparecía en ninguno de los documentos analizados*. De acuerdo con esta definición, la minería textual sería un proceso con el que se pretende descubrir nueva información o conocimiento, y en el que la *información que se descubre* debe ser desconocida de antemano, incluso por los autores de los documentos que se hayan tomado como punto de partida del proceso.

En un trabajo sobre la permeabilidad de las disciplinas científicas, **Don Swanson** señaló las limitaciones derivadas del desconocimiento que los expertos en un área determinada tienen de la literatura publicada en otras áreas de conocimiento, y que pueden ser relevantes para sus temas de estudio.

Estas limitaciones se ilustran con un ejemplo sobre la migraña. **Swanson** extrajo una serie de enunciados de distintos artículos publicados por expertos en distintas áreas, con lo cual la probabilidad de que un mismo científico accediese a todos ellos resultaba remota. Los enunciados utilizados por Swanson son los siguientes:

—El estrés está relacionado con las migrañas.

—El estrés puede producir pérdidas de magnesio.

—Los bloqueos de calcio previenen a las migrañas.

—El magnesio es un bloqueante natural del calcio.

—Los niveles altos de magnesio inhiben la *SCD*.

—Etc.

A partir de estos enunciados —que recordamos proceden de una colección de artículos inconexos—, se podía deducir una relación entre las deficiencias de magnesio y las migrañas. Lo que interesa es que esta hipótesis no se encontraba documentada en ninguno de los artículos. Es decir, se trata de *nuevo conocimiento* que podía extraerse directamente a partir de un corpus de textos inicialmente inconexos. Facilitar y permitir este tipo de deducciones a partir de las conexiones ocultas existentes entre distintos textos sería el objetivo que realmente persigue la minería textual.

La minería textual como herramienta para el análisis

El objetivo de la minería textual —extraer nuevo conocimiento a partir del análisis de corpus textuales— puede resultarnos un tanto "esotérico", aún a pesar de la claridad del caso propuesto por **Swanson**. ¿Se exige a una herramienta de minería textual que extraiga las conclusiones o el nuevo conocimiento, o simplemente que facilite el análisis a un investigador humano?

«El objetivo de la minería textual es extraer nuevo conocimiento a partir del análisis de corpus textuales, pero no de deducirlo»

Inicialmente, la minería textual debe facilitar el análisis de corpus textuales que a priori nos resultarían inmanejables debido a su tamaño. Así, un investigador podrá analizar esos datos, identificar relaciones entre documentos y extraer conclusiones. **Hearst** deja claro el alcance de la minería textual, al indicar que *para hacer progresos no es necesario un análisis del texto propio de la inteligencia artificial, sino que una mezcla de análisis humano y automatizado puede dar excelentes resultados*. La autora llega incluso a definir minería textual como *el descubrimiento semi-automatizado de patrones y tendencias en grandes conjuntos de datos*.

Así pues, la minería textual tendrá como objetivo intermedio —y previo al *descubrimiento de nuevo conocimiento*— procesar y presentar la información disponible en grandes colecciones de documentos en un formato que facilite su comprensión y análisis. Con esta definición nos acercamos a una definición más pragmática para esta disciplina.

En esta misma línea, **Sullivan** señala cómo *la minería textual es el proceso de compilar, organizar y analizar grandes colecciones de documentos para apoyar en la distribución de información a los analistas y a las personas encargadas de tomar decisiones, y para descubrir relaciones entre hechos relacionados que se reparten entre distintos dominios de investigación.*

Partiendo de la tesis de **Hearst** —quizás demasiado ambiciosa— hemos llegado a una definición más práctica que nos permite situar la minería textual en un marco más próximo a su uso real y a la funcionalidad que nos ofrecen las aplicaciones software de minería textual disponibles en el mercado.

Obviamente, esto no significa que en un futuro más o menos inmediato ampliemos nuestras exigencias a las aplicaciones de minería textual, hasta el punto de que la capacidad de deducir nuevos conocimientos de forma automatizada o desatendida llegue a formar parte de la definición de *minería textual*.

Pero no queremos concluir este apartado sin recoger una definición que aparece en distintos documentos oficiales de *IBM* —fabricante de una de las herramientas comerciales de minería textual a la que nos referiremos en un apartado posterior—. En estos documentos se define la minería textual como *el proceso de extracción automática de información fundamental de textos, detección automática de temas predominantes en un conjunto de documentos y búsqueda de textos relevantes mediante consultas de grandes prestaciones y flexibilidad* (IBM, 1998, p. 50).

Esta definición comparte el pragmatismo al que nos hemos referido anteriormente. En ella se evita cualquier referencia a la posibilidad de *identificar nuevos conocimientos* a partir de documentos existentes, con lo que se aleja de la propuesta inicial de **Hearst**. En esta definición se llega a incluir entre las actividades propias de la minería textual la recuperación de información. Nos parece acertada esta definición porque recoge la funcionalidad real que, a día de hoy, implementan las aplicaciones de minería textual de las que hemos podido obtener información.

Otra definición recogida de un documento comercial, en este caso del fabricante de aplicaciones analíticas *SAS*, señala que *la minería textual es el proceso de investigar una gran colección de documentos en texto libre, para descubrir y usar el conocimiento disponible en la totalidad de la colección*. Esta definición resume la funcionalidad que podemos encontrar en las aplicaciones comerciales: una ayuda para facilitarnos la comprensión y la interpretación de la información recogida en grandes colecciones de documentos.

Para concluir, podríamos matizar la definición inicial de **Hearst** e incluir entre los objetivos de la minería textual la extracción y visualización de la información procedente de grandes corpus textuales en un formato que facilite su análisis y la deducción de nuevas conclusiones.

La minería textual y su relación con otras disciplinas

La minería textual recoge diferentes técnicas y planteamientos desarrollados en otras disciplinas. Existe una clara relación entre minería textual, minería de datos, recuperación de información y lingüística computacional.

Minería textual y minería de datos. En numerosas ocasiones la minería textual se presenta como una actividad complementaria a la minería de datos, si bien no ha logrado el impacto de esta última.

La minería de datos pretende obtener información a partir de los patrones y tendencias que pueden observarse en grandes volúmenes de información estructurada. Es decir, información disponible en bases de datos relacionales. Frente a esto, la minería textual busca un mismo objetivo en corpus textuales o información no estructurada.

Los principales fabricantes de aplicaciones informáticas para la minería de datos promueven la imagen de la minería textual como una disciplina complementaria a la primera, y han acoplado a sus programas diferentes módulos para la extracción y análisis de textos. Este planteamiento comercial tiene un fundamento práctico: si la mayor parte de la información que gestionan las organizaciones es textual plasmada en documentos, la minería textual resulta el complemento idóneo para los programas de minería de datos. En la literatura comercial los fabricantes presentan numerosos escenarios en los cuales la minería textual resulta un complemento clave para facilitar una mayor comprensión de las necesidades de los clientes y permitir el seguimiento de la competencia: gestión de reclamaciones, transcripciones de las llamadas registradas en los call-center, patentes, noticias de prensa, etc.

Así, existe una similitud entre minería textual y de datos, ya que ambas persiguen una misma finalidad: deducir nueva información a partir de la información ya existente. Cambiará, únicamente, el tipo de información que se toma como base del análisis: datos estructurados en el primer caso, e información no estructurada (texto) en el segundo.

Llegados a este punto es conveniente recordar que por medio de la minería textual sólo se pueden deducir relaciones que se encuentren explicitadas en los documentos que forman el corpus de trabajo y, en ningún

	Búsqueda de patrones	Búsqueda de ítems y datos	
		Nuevos	Ya conocidos
Datos no textuales	Minería de datos		Búsqueda en BB.DD.
Datos textuales	Lingüística computacional	Minería de textos	Recuperación textual

Tabla 1

caso, se podrá deducir información implícita, al carecer estos sistemas de razonamiento lógico.

Minería textual, lingüística computacional y recuperación de información. Desde el punto de vista técnico, la minería de datos recoge técnicas usadas tradicionalmente en la recuperación textual, y en la lingüística computacional. Esta influencia llega a tal punto que resulta difícil poder afirmar que la minería textual haya incorporado técnicas propias.

Hearst establece las diferencias entre minería textual, minería de datos y recuperación textual (tabla 1). En este cuadro se muestran las diferencias —y también las áreas de influencia— entre las cuatro disciplinas.

La diferencia entre minería textual y recuperación de información se encuentra en que el objetivo de ésta última es identificar los documentos relevantes para un usuario dentro de una colección. La recuperación textual parte de una representación formal de los documentos sobre los que se realizará la búsqueda, y de la formulación de las necesidades de información del usuario mediante un sistema de representación equivalente.

Sin embargo, la recuperación textual no pretende facilitar el proceso de análisis ni la extracción de nuevos conocimientos, como sí pretende la minería textual. **Hearst** hace hincapié en esta diferencia y señala cómo se ha extendido una apreciación incorrecta que iguala minería textual con sistemas de recuperación de información avanzados, o sistemas de recuperación de información adaptados para la web. Otro dato importante a tener en cuenta es que el objetivo de la recuperación de información se centra en establecer los mecanismos para satisfacer las necesidades de información de un usuario. Sin embargo, en la minería textual no sólo no existe esa necesidad sino que tampoco es del todo necesario que se cuente con una pregunta concreta que realizar al sistema.

La lingüística computacional, por su parte, agrupa una serie de técnicas para procesar textos y tratar de hacerlos comprensibles para un ordenador. La lingüística computacional permite el análisis sintáctico y gramatical de textos en formato electrónico, la alineación e identificación de correspondencias entre textos escritos en diferentes idiomas, etc. Sus principales resultados se han materializado en los sistemas de traducción automática. También en este caso, si bien la minería textual ha adoptado algunas de las técnicas desarrolladas por esta disciplina, sus objetivos son diferentes.

Usos de la minería textual

Como ya se ha comentado anteriormente, el objetivo de la minería textual es facilitar el análisis de la información disponible en grandes colecciones de documentos, y así la deducción de nuevo conocimiento. Pero, ¿cómo lograr esto?, ¿qué nos ofrece una herramienta de minería textual para lograr este propósito?

Las funciones que principalmente debería satisfacer una herramienta de minería textual, o el *output* que podemos esperar de ellas, incluiría:

—Identificar "hechos" y datos puntuales a partir del texto de los documentos, a lo que nos referiremos con el término inglés *feature extraction*.

—Agrupar documentos similares (clustering).

—Determinar el tema o temas tratados en los documentos mediante la categorización automática de los textos.

—Identificar los conceptos tratados en los documentos y crear redes de conceptos.

—Facilitar el acceso a la información repartida entre los documentos de la colección, mediante la elaboración automática de resúmenes, y la visualización de las relaciones entre los conceptos tratados en la colección.

—Visualización y navegación de colecciones de texto.

A continuación se analizarán estas funciones con más detalle.

Identificar "hechos" y datos puntuales de los documentos: *feature extraction*. Una de las aplicaciones de la minería textual es la identificación de "hechos" presentes en los documentos. Traducimos así la funcionalidad referida como "*feature extraction*" por

Los ataques terroristas del 11 de septiembre de 2001 hicieron manifiesta la capacidad de los grupos terroristas de utilizar la red internet para planificar atentados. Estudios posteriores demostraron que las organizaciones terroristas como Al Qaeda se orquestan en torno a complejas cédulas en las cuales se utiliza la Red con distintos propósitos: desde la captura de datos sobre las posibles debilidades de las infraestructuras de las instituciones o estados objeto de sus ataques, hasta la captura de financiación y la captación de nuevos adeptos.

En un artículo publicado en la revista Parameters, Timothy Thomas de la Foreign military studies office de los Estados Unidos recoge distintas muestras de sitios web utilizados o vinculados por la organización terrorista: alneda.com, assam.com, drasat.com, jehad.net y un largo etcétera.

En este nuevo contexto, la red Internet se convierte en un "campo de batalla" más en la lucha contra el terrorismo. La identificación de sitios web y el seguimiento de foros de discusión como el Muslim Hackers Club [dash]en el que según Thomas se discutía sobre sitios web norteamericanos en los que podía obtenerse datos sobre radiofrecuencias y códigos secretos utilizados por el servicio secreto norteamericano[dash] forman parte de las actividades de las unidades de inteligencia.

El proyecto TIA (Total Information Access)

Los planes de utilizar la tecnología de minería de datos en la lucha anti-terrorista se han materializado en el proyecto Total information access, TIA. Se trata de un programa con una duración inicial de cinco años dirigido desde la Defense advanced research project agency (Darpa); concretamente, desde la Information awareness office.

TIA desarrollará una serie de herramientas que faciliten el análisis de información a los agentes de inteligencia y la integración de datos disponibles en las bases de datos de distintas agencias. Las líneas de desarrollo básicas son:

—Traducción automática: para facilitar el análisis de publicaciones escritas en diferentes idiomas.

—Identificación de patrones en bases de datos: ya que la co-ocurrencia de cierta información (la petición de pasaportes, visados o permisos de trabajo, el alquiler de automóviles o la compra de productos químicos) pueden ser indicadora de una acción terrorista potencial.

La principal flaqueza de este sistema reside en el hecho de que se desconoce la fiabilidad de la combinación de estas prácticas. Si las técnicas de minería de datos han fallado en el diseño de campañas de marketing, ¿por qué no pueden fallar en la identificación de un ciudadano como un supuesto terrorista o sospechoso? Por otra parte, también se cuestiona la efectividad real de la minería de datos en la identificación de sospechosos, y el que pueda ser una herramienta útil en la lucha contra el terrorismo.

PathFinder

Otra aplicación utilizada por los Estados Unidos y por la Otan para actividades de minería de datos es PathFinder, desarrollado por la empresa PreSearch. Esta compañía cuenta con 37 años de experiencia en el desarrollo de aplicaciones para la inteligencia y la seguridad. PathFinder ha sido usada exclusivamente por agencias del gobierno norteamericano, y durante 14 años la han utilizado en exclusividad. Actualmente existe una versión comercial. La enumeración de características del programa la sitúa como una aplicación mucho más avanzadas que el resto de programas analizados en este artículo. Por ejemplo, junto a herramientas para el análisis lingüístico, PathFinder incluye funcionalidad para dibujar mapas y trabajar con datos geográficos.

Tabla 2. Ejemplos prácticos en el uso de la minería textual

Sullivan. Se trata de extraer del texto de los documentos referencias a nombres de personas, organizaciones, fechas, eventos, y las relaciones que existen entre ellas.

Por ejemplo, en un documento podrían extraerse: a) referencias a "**José María Aznar**", "intervención en Irak" y "gobierno de España" y b) relaciones entre estos conceptos, como "**José María Aznar** preside el gobierno de España" o "**José María Aznar** se manifiesta a favor de la intervención en Irak".

Esta aplicación difiere de la extracción de términos o de la indización automática tradicional, ya que no se trata de ver cuáles son los temas o las ideas que se tratan en el documento, sino de identificar personas, instituciones, eventos y la relación que existe entre ellos.

Si bien es posible que esta extracción de "hechos" se apoye en diccionarios y listados de autoridad existentes, las aplicaciones de minería textual deberían ser capaces de identificar de forma desatendida aquellos fragmentos o cadenas de texto que hagan referencia a personas, organizaciones y eventos de los que no se tengan referencias previas.

Como ejemplo de la aplicación de esta funcionalidad, **Hearst** cita un proyecto de análisis bibliométrico que estudió el impacto de la investigación financiada con fondos públicos en el desarrollo tecnológico y en la redacción de patentes en norteamérica.

El estudio realizó un análisis de la bibliografía citada en una amplia colección de patentes, e identificó la procedencia de la financiación que había permitido el desarrollo de los estudios descritos en los artículos citados. Si bien no se clarifica que se hayan aplicado técnicas de minería textual en la obtención de los resultados, la autora sí señala que la extracción automática de datos resultaría una gran ayuda para identificar autores y fuentes de financiación en los artículos citados desde las patentes.

Agrupación de documentos similares o clustering. Consiste en unir documentos entre los que existe cierta similitud. La similitud se establecerá a partir de la terminología utilizada por los autores en la redacción de los textos. Esta funcionalidad de las aplicaciones de minería textual aplica una de las técnicas características de la recuperación textual: el clustering o agrupación automática de documentos.

Evitamos hablar de "clasificación automática de documentos" —si bien este término sería también correcto— por una diferencia de matiz existente entre las técnicas de clustering y la categorización automática a la que nos referiremos en un apartado posterior.

En el caso del clustering se trata de crear agrupaciones entre documentos de forma desatendida. Es de-

cir, un programa informático decidirá qué grupos va a generar a partir de la similitud que calcule entre los documentos de la colección. En el caso de la categorización automática, el programa informático deberá decidir a qué clase —dentro de un conjunto de clases predefinidas por nosotros— pertenece cada uno de los documento disponibles en una colección. En ambos casos se trata de un proceso de clasificación, por lo que no utilizaremos este término para referirnos a ninguno de ellos en concreto.

La agrupación automática tiene distintas aplicaciones, entre ellas:

—Facilitar la comprensión de una colección de documentos. Al agrupar aquellos que son similares en una misma clase, será posible obtener una visión general de los temas tratados en todos los documentos de la clase con sólo leer unos pocos de los textos incluidos en ella.

—Juzgar la relevancia de los documentos incluidos en cada grupo tras la lectura de tan sólo uno de sus representantes.

—Identificar relaciones entre documentos en una colección que previamente se desconocían.

—Identificar duplicados potenciales y documentos que — por tener una información similar— pueden no ser relevantes.

—Mejorar la organización de los resultados devueltos por un motor de indexación. El clustering supondría una mejora sustancial frente a los mecanismos de visualización de resultados actualmente generalizados, en los que cada documento recuperado se muestra de forma independiente.

Determinar el tema tratado por los documentos o categorización automática. Se trata de un proceso de clasificación automática con el que se pretende asignar un documento a una clase o tema definido con anterioridad. Un ejemplo de esta aplicación sería la asignación automática de un encabezamiento de materia o una notación de un sistema de clasificación bibliográfico a un documento.

La categorización automática parte de un entrenamiento previo del programa informático encargado de realizarla. Así, se facilitará al programa informático una serie de documentos a los que ya se ha asignado un tema o clase. De esta forma, el programa podrá analizar las características que determinan la asignación de los documentos a una u otra clase. Posteriormente, cuando se procese un nuevo documento el programa podrá "deducir" a qué clase pertenece. Esta deducción se basará en la similitud que exista entre el nuevo documento y los utilizados durante la fase de entrenamiento.

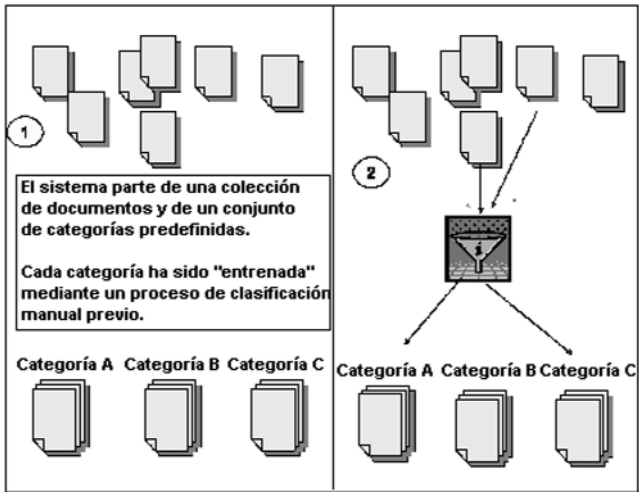


Figura 1. Ejemplo de categorización automática

En la figura 1 podemos ver como tanto en el clustering como en la categorización automática se parte de una misma base: el cálculo de las similitudes entre documentos, normalmente mediante la identificación de los términos que aparecen de forma conjunta en los documentos.

Queremos señalar que **Hearst** no incluye la categorización automática entre las técnicas propias de la minería textual, si bien recoge un ejemplo en el cual esta técnica se puede aplicar en la deducción de nuevo conocimiento: el proyecto *Darpa topic detection and tracking*, que consiste en el análisis de noticias recibidas y procesadas en orden cronológico, con el fin de detectar si alguna de ellas hace referencia por primera vez a un hecho novedoso. Es decir, se trata de identificar cuando es la primera vez en la que se hace referencia a un hecho en cuestión. La autora señala que *se puede considerar un caso de minería textual porque nos cuenta algo sobre el mundo fuera del la colección de textos procesados propiamente dicha*.

Identificar los conceptos tratados por los documentos. Esta función permitiría extraer los principales temas o ideas tratados en los documentos. No se trata de un proceso de categorización automática, ya que no se pretende asignar un documento a una clase, sino extraer un conjunto de términos que son representativos del contenido de los documentos.

Esto coincide con una de las técnicas utilizadas en la indización automática de documentos, si bien la identificación de conceptos que pretende la minería textual sería más compleja que la mera extracción de términos tal y como aparecen escritos en los textos característica de los motores de indexación automática y recuperación textual.

En el contexto de la minería textual, un concepto representaría una *idea* tratada en el documento. La capacidad de identificar un concepto se basaría en la

ocurrencia de determinados términos y combinaciones de términos en el texto del documento.

Una vez identificados los conceptos tratados por un documento, será posible identificar documentos que traten de ese mismo concepto (incluso en aquellos casos en los cuales los autores no hayan utilizado una terminología idéntica), y crear redes conceptuales a través del contenido de la colección (figura 2).

La posibilidad de crear estas redes de conceptos sería una de las principales ventajas de la minería textual. Podríamos decir que el ejemplo clásico propuesto por **Swanson** es un caso particular de esta aplicación:

—En primer lugar se extraen los principales conceptos de cada documento individual (por ejemplo, del primer documento se extraen los conceptos "estrés" y "migrañas"; del segundo "estrés" y "pérdida de magnesio", y así sucesivamente).

—A continuación se puede crear una red de conceptos o una trama que fusione los conceptos procedentes de distintos documentos que, a priori, eran inconexos. Esta red se podrá tomar como base para un proceso de análisis y obtención de conclusiones.

Esta funcionalidad estaría próxima a los intentos de creación automática de tesauros que se han realizado en el área de la documentación automatizada, a la que nos referiremos posteriormente.

Una de las aplicaciones de minería textual que describimos en un apartado posterior —*Megaputer Text Analyst*— ofrece un mecanismo para identificar los términos más representativos de una colección de documentos y las relaciones que existen entre ellos, nuevamente a partir del recuento de las ocurrencias conjuntas de los términos.

Elaboración automática de resúmenes. Entre las tareas necesarias para facilitar el análisis de grandes volúmenes de documentos, la elaboración automática

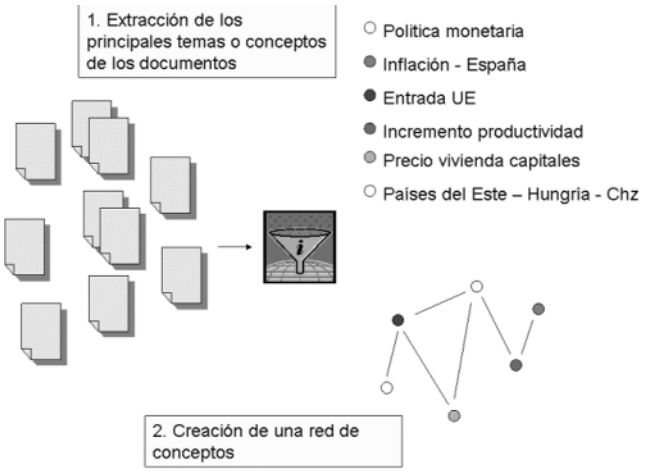


Figura 2. Extracción de conceptos y red conceptual

de resúmenes es otra de las aplicaciones y funcionalidades que caracterizan a los programas informáticos de minería textual. Las técnicas de elaboración de resúmenes por parte de programas informáticos se remontan a la década de los sesenta, y se han desarrollado distintos enfoques basados en el análisis sintáctico y en el estudio estadístico de los textos.

En el enfoque más simple —basado en la frecuencia estadística de los términos y en la ponderación de la importancia de las frases y la posición que estas ocupan en el documento— los resúmenes se generan mediante la extracción literal de frases o fragmentos del documento original, sin hacer una "reescritura" posterior.

La elaboración automática de resúmenes constituye un componente clave para facilitar el proceso de análisis, especialmente cuando éste se realiza a nivel de colección —en lugar de hacerlo a nivel de documento individual—, ya que permite lograr una de los objetivos de la documentación: ahorrar tiempo al lector.

Visualización y navegación de colecciones de texto. Hasta ahora hemos analizado las funciones que debería satisfacer cualquier herramienta de minería textual. Todas estas funciones ofrecen como resultado de su análisis listados de términos o redes de conceptos, grupos de documentos relacionados, nombres de personas, hechos, organizaciones, etc., que proceden de una colección de documentos analizada.

Para permitir la comprensión de esta información, un componente clave en un sistema de minería textual es la interface de usuario a través del cual se va a visualizar esta información. La interface deberá mostrar los datos en un formato que haga posible su interpretación y permita al usuario moverse con facilidad entre los distintos textos analizados.

Si bien el estudio de las interfaces de las aplicaciones de minería textual queda fuera del alcance de este trabajo introductorio, en el último apartado recogemos algunos ejemplos de herramientas comerciales y de las interfaces que implementan.

Las técnicas de la minería textual

Para lograr los resultados citados en el apartado anterior la minería textual adopta una serie de técnicas procedentes de la recuperación de información y de la lingüística computacional. Estas técnicas incluyen:

—Pre-procesamiento de los documentos, que contendría la extracción de términos, eliminación de las palabras vacías y normalización de los términos restantes mediante *stemming*.

—Identificación de nombres propios. Análisis sintáctico y gramatical de los textos.

—Representación de los documentos mediante el modelo vectorial. Fórmulas para el cálculo de la similitud entre pares de documentos.

—Clustering o agrupación automática de documentos, que a su vez también toma como punto de partida la representación de los documentos según el modelo vectorial y el cálculo de similitudes.

—Categorización automática.

—Relaciones entre términos y conceptos.

Pre-procesamiento de los documentos. Esta técnica consiste en extraer las palabras utilizadas en un documento, o *segmentar el texto en distintas formas gráficas* (**Etxeberría**, p. 146). Una *forma gráfica* se define como *una secuencia de caracteres no delimitadores (en general, letras), comprendida entre dos caracteres delimitadores (espacios o signos de puntuación)* (**Etxeberría**, p. 147).

El pre-procesamiento incluye la eliminación de los signos de puntuación y la extracción de las palabras separadas entre sí por espacios en blanco o signos de puntuación (si éstos no se han eliminado en el paso previo). Para completar esta tarea, el programa informático debe convertir el documento que se va a procesar a un formato texto plano, no binario.

Una tarea habitual en el pre-procesamiento de los documentos es la eliminación de palabras vacías, carentes de significado, como son preposiciones, artículos, conjunciones, etc. Sin embargo, no todos los autores coinciden en la conveniencia de eliminar las palabras vacías.

Finalmente, como parte del pre-procesamiento se suele realizar la normalización de las palabras extraídas del documento. Esta normalización —también llamada *lematización*— consiste en dividir cada palabra en los lemas que la forman. Por ejemplo, las palabras alumno, alumna, alumnado, alumnos, etc., comparten una misma raíz léxima (alumn-) que les da el mismo significado semántico.

La lematización reduciría y representaría todas las palabras que comparten la misma raíz mediante ésta. Este proceso tiene una gran importancia de cara a hacer el recuento de las ocurrencias de una palabra y escoger así aquellos términos que son los mejores candidatos para representar el contenido del texto. Normalmente la lematización reducirá las variaciones de género y número en los sustantivos y adjetivos, así como las flexiones de los tiempos verbales. Para poder realizar esta tarea el programa informático necesitará acceso a un diccionario y a una base de conocimiento sobre las distintas flexiones, conjunciones de verbos, etc., que le permita extraer correctamente los lexemas que forman cada palabra.

Un aspecto importante en el pre-procesamiento es la identificación de los llamados "segmentos repetidos" (siguiendo la terminología utilizada por **Etxeberría**) o "frases". Es decir, secuencias de palabras que aparecen contiguas en el texto y que usadas de esta forma tienen un significado especial. Por ejemplo "marketing relacional", "instrumentos de análisis", etc. Dividir estos segmentos repetidos en los distintos términos que lo forman acarrearía una descontextualización y pérdida de significado. Normalmente, las aplicaciones de indexación y recuperación textual han prestado poca atención a este problema y han tendido a dividir los segmentos de repetición potenciales. Este enfoque es lógico: si un sistema de indexación permite formular búsquedas con operadores adyacentes (del tipo "recuperar los documentos que contengan la palabra *marketing* seguida de la palabra *relacional"*), no es necesario identificar los segmentos de repetición.

Sin embargo, en el caso de la minería textual, la extracción de estos términos compuestos sí es importante, ya que buscamos el conjunto de conceptos que representan el contenido de un documento. Identificar los segmentos repetidos que aparecen en un texto podría hacerse fácilmente teniendo acceso a un diccionario que permita identificar la categoría gramatical de cada palabra (sustantivo, adjetivo, preposición, verbo, etc.). El problema consiste en identificar qué segmentos repetidos tienen realmente una significación especial, y deberían tratarse como "términos" o "conceptos".

En cualquier texto podemos identificar un gran número de segmentos de repetición con una misma estructura sintáctica (sustantivo->adjetivo, sustantivo->preposición->sustantivo), pero de todos ellos tan sólo una mínima parte tendrán un significado especial, resultado de la unión de los dos términos. Para solucionar este problema, cabe la posibilidad de aplicar técnicas estadísticas que seleccionen únicamente aquellos segmentos de repetición que ocurren con mayor frecuencia en los documentos, o reglas heurísticas que —por ejemplo—, identificasen únicamente los segmentos de repetición que aparecen en los títulos, títulos de sección, etc., de los documentos.

Identificación de nombres propios. La extracción de nombres propios relativos a personas, organizaciones, eventos, funciones, así como cantidades monetarias y fechas es una de las principales funciones que debe satisfacer la minería textual. Además, la minería textual también debería permitirnos identificar las relaciones que existen entre estos nombres propios y constatar así "hechos" descritos en los documentos.

Un resumen de las dificultades que implica la identificación de nombres propios en el texto de los documentos lo encontramos en el informe *Extracting names from natural-language text*, de **Yael Ravin** y **Nina Wacholder**. En este informe los autores describen la herramienta *Nominator*, desarrollada como proyecto de investigación por el *T. J. Watson research center* de *IBM*. Sospechamos que los resultados de este proyecto se aplicaron en uno de los módulos de la aplicación *IBM Intelligent Miner for Text*.

El proyecto consistió en el desarrollo de un prototipo basado en reglas heurísticas, con capacidad para identificar aquellos fragmentos que pueden corresponder a un nombre propio, identificar su tipo (es decir, si se refiere a una persona, organización, lugar, etc.) y las formas alternativas que se utilizan en el texto para hacer referencia a esa misma entidad.

En *Nominator* no se utilizaron reglas sintácticas, a pesar de lo cual se obtuvieron resultados satisfactorios: los autores señalan un porcentaje del 97.8% de éxito en las pruebas realizadas con el prototipo.

Un tema más complejo en la identificación de nombres propios, es la extracción de las relaciones que existen entre los términos. En este sentido, es necesario recurrir a técnicas de parsing y análisis sintáctico de las sentencias, para identificar los verbos que sirven de nexo entre los nombres propios y tratar de deducir así posibles relaciones.

Representación de documentos mediante el modelo vectorial. Una premisa en cualquier aplicación de recuperación y tratamiento documental es la necesidad de representar el contenido de los documentos mediante un modelo. El modelo generalizado a día de hoy, tanto en los sistemas de indexación como en las aplicaciones de minería textual, es el vectorial.

En este modelo, un documento se caracteriza mediante el conjunto de términos que representan su contenido. Estos términos podrían ser términos extraídos directamente del texto completo del documento (tal y como se ha descrito al referirnos al pre-procesamiento), o descriptores asignados al documento por un documentalista o por una aplicación informática, tomados o no de un lenguaje documental externo.

Cualquiera que sea el caso, el documento se representará mediante una secuencia de términos o "componentes" que corresponden con los distintos términos utilizados para describir el contenido del documento.

Un vector es una estructura consistente en un número fijo de elementos o componentes, en la cual la posición de cada uno de ellos es significativa. En el modelo vectorial, cada documento se considera un vector, y cada término que aparece en al menos un documento, será un componente del vector.

En este método la recuperación de información se realiza mediante la comparación de la distancia que existe entre los vectores correspondientes a los documentos, y un vector utilizado para representar la ecuación de búsqueda.

Entre las ventajas del modelo vectorial —frente a otros modelos como el booleano— se encuentra el hecho de calcular la similitud entre la ecuación de búsqueda y los documentos. Esto permite realizar un ranking u ordenación de los documentos recuperados, mostrando al principio de la lista aquellos documentos que son más similares a la ecuación de búsqueda, y al final de la lista los que son menos.

«Existe una similitud entre minería textual y de datos, ya que ambas persiguen una misma finalidad. Sin embargo, cambia el tipo de información que se toma como base del análisis»

Análisis de clusters. Se trata de una *técnica que permite identificar grupos o clases de objetos similares a partir de un espacio multidimensional* (**Arenas**, 1992, p. 369). Según esta autora, su formulación se debe a **B. S. Everitt** en la obra *Cluster analysis*, publicada en 1974. En otra fuente[1] se señala cómo el término fue usado por primera vez por **Tryon** en 1939. **Jain** et al**.** definen el análisis de cluster como *la organización de una colección de patrones (normalmente representados mediante vectores o como puntos en un espacio multidimensional), en clusters o grupos en base a su similitud*. Aquellos patrones que pertenezcan a un mismo cluster o grupo serán más similares entre sí, que con los patrones que pertenecen al resto de los grupos.

El análisis de cluster consiste en una clasificación desatendida o no supervisada. Esto diferencia al análisis de cluster de las técnicas de clasificación supervisada (como el análisis discriminante), y que se aplican en la categorización automática.

En la clasificación supervisada se debe ordenar un conjunto de objetos en una serie de grupos predefinidos con anterioridad. En el caso del análisis de cluster no existirán grupos predefinidos a los que haya que asignar los objetos durante el proceso de clasificación.

En el análisis de cluster se deben tomar una serie de decisiones relativas a:

—La forma de representar a los objetos que se van a clasificar (a los que nos referiremos de ahora en adelante como "patrones").

—La fórmula que vamos a utilizar para calcular la similitud entre patrones.

—El algoritmo de clustering que se va a utilizar.

—El método utilizado para abstracción los datos o patrones incluidos en un mismo grupo, y poder así representarlos y visualizarlos.

—La evaluación del resultado del proceso de clustering.

Volviendo a las características del análisis cluster, éste parte de la descripción previa de los objetos que se quieren clasificar. Esta descripción nos permitirá generar grupos de objetos con un coeficiente de similitud significativo. En el caso de la agrupación de documentos la descripción que tomaremos como base será la representación de los documentos según el modelo vectorial.

Categorización automática. Esta técnica se utiliza en la minería textual para clasificar documentos en una serie de categorías preestablecidas. Su origen se remonta a la década de los sesenta (**Maron**, 1961), si bien el auge de internet y los contenidos en formato electrónico vivido en los últimos seis años ha propiciado un aumento en el interés hacia estas técnicas.

Así, la categorización automática es aplicable en distintos procesos:

—Clasificación e indización automática de documentos.

—Filtrado de contenidos (por ejemplo, en la distribución de noticias o *newsfeeds*).

—Asignación de páginas y sitios web a listas de categorías predefinidas en portales tipo *Yahoo* o *dmoz.org*.

—Resolver la ambigüedad en palabras con polisemia.

Páginas personales y de grupos de trabajo

Página personal de Marti Hearst. Con enlaces a distintas publicaciones y artículos de la autora.
http://www.sims.berkeley.edu/~hearst/

Página personal de Wanda Pratt, profesora de la Universidad de Washington. Entre los cursos que imparte se encuentra el ICS 280 text mining (la última edición se celebró en primavera del 2001).
http://www.ischool.washington.edu/wpratt/

Página personal de Fabrizio Sebastiani, del Istituto di scienza e tecnologia dell'informazione consiglio nazionale delle ricerche area della ricerca di Pisa. Contiene numerosos artículos en texto completo escritos por el autor sobre categorización automática.
http://faure.iei.pi.cnr.it/~fabrizio/

Página personal de Haym Hirsh, profesor en Rutgers University. Incluye unos pocos artículos publicados por el autor sobre minería textual.
http://www.cs.rutgers.edu/~hirsh/

Mitre. Se trata de una organización con sedes en Bedford, Massachusetts y McLean, Virginia, fundada en 1959 como una ramificación del MIT. Se dedica a ofrecer servicios de apoyo a la administración norteamericana en temas relacionados con las tecnologías de la información y la ingeniería. Cuenta con varios "centros de investigación", y entre sus líneas de trabajo se encuentra la lingüística computacional (resumen automático, etc.) Se pueden descargar tutoriales, artículos, etc.
http://www.mitre.org

Sitio web de un grupo de trabajo sobre minería textual de la Universidad de Waikato, en Nueva Zelanda. Los documentos publicados por los miembros del grupo están disponibles en el sitio. Este grupo inició el desarrollo de una plataforma de minería textual open source.
http://www.cs.waikato.ac.nz/~nzdl/textmining/

Grupo de trabajo Computational linguistics and text mining de IBM. Está disponible el texto completo de algunos de los artículos publicados por sus miembros.
http://www.research.ibm.com/dssgrp/papers.html

Página del grupo de trabajo Multilingual text mining de Xerox. Este proyecto ya finalizó.
http://www.xrce.xerox.com/competencies/content-analysis/past-projects/dmhead/home.en.html

Páginas Web con enlaces sobre minería textual

Página con numerosos enlaces relacionados con la minería textual, recopilados por un estudiante de doctorado de la Universidad de Texas. Probablemente se trata de la recopilación de recursos más completa.
http://www.cs.utexas.edu/users/pebronia/text-mining/

Detallada recopilación de enlaces a cargo de la Office for naval research (ONR). Además de la minería textual, incluye enlaces relacionados con la transferencia tecnológica, bibliometríca, etc.
http://www.onr.navy.mil/sci_tech/special/technowatch/linkd.htm

Página con numerosos enlaces relacionados con la minería textual. Recopilados por Weiguo Fan, profesor de Virginia Tech.
http://filebox.vt.edu/users/wfan/text_mining.html

Congresos dedicados total o parcialmente a la minería textual

Sitio web del ACM Special Interest Group in Knowledge Discovery and Data Mining. Organiza la conferencia anual SIGKDD.
http://www.acm.org/sigs/sigkdd/

Text Retrieval Conference, organizada anualmente por el Nist (National Institute of Standards and Technology) y Darpa desde 1992.
http://trec.nist.gov/

Ofrece un listado de conferencias y congresos relacionados con el procesamiento del lenguaje natural y la recuperación textual.
http://www.cs.technion.ac.il/~gabr/resources/jour_conf.html#conferences

Sitios web de fabricantes de aplicaciones comerciales

http://www-4.ibm.com/software/data/iminer/fortext/

http://www.leximancer.com/overview.html

http://www.pertinence.net

http://www.eidetica.com/

http://www.xanalys.com

http://www.themis-group.com

http://www.simstat.com/wordstat.htm

http://www.textanalysis.com

http://www.textmining.com (SRA)

Revistas especializadas

Information retrieval. Revista dedicada a la recuperación de información, publicada por Kluwer Academic Press bajo de dirección de Paul Cantor, Josiane Mote y Justin Zobel.
http://www.kluweronline.com/issn/1386-4564

Journal of machine learning research, publicada por el MIT. Artículos de números anteriores están disponibles en texto completo en formato pdf.
http://www.ai.mit.edu/projects/jmlr/

Computational linguistics, publicada también por el MIT.
http://mitpress.mit.edu/journal-home.tcl?issn=08912017

Natural language engineering. Publicada por la Universidad de Cambridge.
http://titles.cambridge.org/journals/journal_catalogue.asp?mnemonic=nle

Journal of classification, publicada por la Classification Society of North America y Springer Verlag.
http://www.pitt.edu/~csna/joc.html

Data mining and knowledge discovery. Publicada por Kluwer.
http://www.kluweronline.com/issn/1384-5810

Tabla 3. Recursos web sobre minería de texto

Existen dos tipos de categorización: *single-label* y de *multilabel*. En el primero se asignará cada documento a una única categoría. En el segundo, un mismo documento podrá asignarse a más de una categoría. En ambos casos se procederá de forma similar. Así, en el caso de la categorización *multilabel* el proceso de clasificar un documento en una serie de categorías puede tratarse como problemas independientes consistentes en saber si se debe clasificar al documento en la categoría primera, segunda...

También se suele diferenciar entre procesos de categorización basados en los documentos (*document-pivoted categorization*) y procesos de categorización basados en las categorías (*category-pivoted categorization*). En el primer caso, el proceso recibe como entrada un documento que debe clasificar y un conjunto de categorías, y debe decidir cual de ellas corresponde al documento. En el segundo enfoque el proceso recibe como entrada un conjunto de documentos y una categoría, y debe decidir qué documentos pertenecen a esa categoría.

Una última distinción que se utiliza para describir los procesos de categorización diferencia entre *hard categorization* y *ranking categorization*. En el primer caso, el sistema tomará una decisión sobre si se va a clasificar un documento en cada categoría. La decisión será "verdadero" o "falso". En el segundo el sistema responderá con un valor que indicará la conveniencia o probabilidad estimada de que un documento pertenezca a una o más categorías. Estos valores probables pueden ordenarse formando un ranking, donde las categorías más probables aparecerán al comienzo de la lista y las menos probables al final. Se recomienda aplicar este segundo enfoque en aquellos casos en los que la categorización vaya a ser supervisada o validada posteriormente por una persona.

Relaciones entre términos y conceptos. Entre las técnicas utilizadas por la minería de textos se encuentra la extracción de términos o conceptos y la identificación de relaciones entre estos términos. En apartados anteriores nos hemos referido a la extracción de términos y a su ponderación para identificar aquellos que resulten más significativos del contenido de los documentos. Otras aproximaciones más complejas, como la *Latent Semantic Indexing* o el clustering, también podrían aplicarse con este propósito.

Debemos señalar que en las aproximaciones clásicas para identificar relaciones entre términos, éstas se deducen a partir de su co-ocurrencia (es decir, la ocurrencia conjunta de dos palabras en los mismos documentos o fragmentos).

Tradicionalmente estas asociaciones se han venido usando en proyectos de recuperación de información experimentales para paliar los problemas vinculados a un escaso índice de llamada. Mediante estas asociaciones entre términos se permitía al usuario recuperar documentos potencialmente relevantes, que no habían sido indexados con los mismos términos que se han utilizado en la ecuación de búsqueda. Esta idea es similar a la propuesta pionera que **Maron** y **Kuhn** hicieron en 1960 con su concepto de *recuperación aritmética* o *asociativa*.

En relación al clustering, de la misma forma que podemos agrupar documentos a partir del número de términos que comparten, sería también posible agrupar términos a partir de los documentos en los que aparecen de forma conjunta. Esta aproximación ha sido descrita por **Salton** y otros autores.

Una propuesta similar es la del llamado *thesaurus difuso*, descrito por **Miyamoto** (p. 104). Este término se ha propuesto en distintas ocasiones con distintos significados. Así, se ha recurrido a expertos en un área de conocimiento determinada para que propusiesen el grado de asociación semántica entre términos de un vocabulario especializado, aplicando técnicas de la lógica borrosa para promediar las propuestas datas por los distintos expertos (**Klir** y **Yuan**, p. 388).

Las asociaciones entre términos —como las descritas en el caso de los *thesaurus* difusos de **Klir**— son un ejemplo de cómo explotar las relaciones entre términos en el proceso de recuperación textual. En el caso de las herramientas de minería textual las asociaciones entre términos permitirían identificar y mostrar al analista conceptos relacionados, para así facilitar el análisis y la extracción de la información repartida entre distintos documentos.

Herramientas para la minería textual

En este apartado describimos brevemente algunas aplicaciones comerciales desarrolladas especialmente para la minería textual. Únicamente ha sido posible obtener una versión de evaluación de *Megaputer TextAnalyst*. Del resto de aplicaciones (*IBM Intelligent Miner for Text*, *SAS Text Miner* y *Spss LexiQuest*) a las que se hace referencia tan sólo se ha podido consultar la documentación oficial de los fabricantes.

En el mercado podemos encontrar multitud de aplicaciones de este tipo. **Ah-Hwee Tan** cita algunas aplicaciones en un estudio comparativo presentado en 1999. Una enumeración exhaustiva la encontramos en el artículo *Text Mining Tools on the Internet* de **Jan van Gemert**. La descripción la limitamos a cuatro aplicaciones ya que el único propósito es ilustrar cómo distintos fabricantes han adoptado e implementado las técnicas descritas en los apartados anteriores, y delimitar así con una visión real de las ofertas tecnológi-

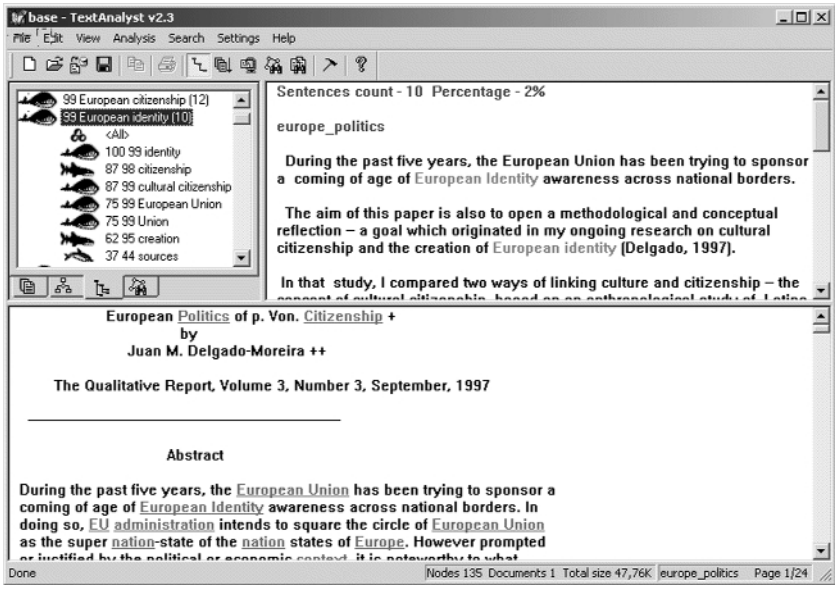


Figura 3. Megaputer TextAnalyst 2.3

cas disponibles en el mercado el concepto de minería textual.

El primer caso —*TextAnalyzer*— consiste en una aplicación desarrollada exclusivamente para la minería textual. *IBM Intelligent Miner for Text* quizá sea la única aplicación desarrollada por uno de los grandes fabricantes de software a nivel mundial. En el caso de *Spss LexiQuests* y *SAS Text Miner* se trata de dos programas adquiridos por fabricantes de entornos de análisis y minería de datos tradicionales —*SAS* y *Spss*— con el propósito de integrarlas en sus plataformas analíticas.

Spss adquirió *LexiQuest* —hasta entonces una empresa independiente— el 22 de febrero del año 2002 para completar su plataforma de minería de datos con tecnología textual. Por otra parte, *SAS* anunció la disponibilidad de *Text Miner* el 31 de mayo de 2002. También se trató de una adquisición. Concretamente *SAS* adquirió tecnología de la empresa *Inxight* (una empresa "derivada" de un proyecto de *Xerox* y que sigue trabajando de forma independiente a *SAS*).

La publicación de estas dos aplicaciones hace poco más de un año por parte de empresas líderes en el mercado de aplicaciones analíticas es una muestra de la importancia que se presta a esta tecnología, y una señal de la demanda que puede existir de este tipo de aplicaciones.

Megaputer TextAnalyst 2.3. Es una de las aplicaciones comerciales para minería textual a las que hace referencia **Dan Sullivan** en su libro *Document warehousing and text mining*.

La aplicación funciona de la siguiente forma: a partir de un texto o colección de textos en formato ascii o rtf se identifican los principales términos. A cada término se le asigna una ponderación que representa en qué medida es significativo en la colección y en el texto procesado. Los términos pueden estar formados por dos o más palabras. Para identificar términos formados por varias palabras la aplicación también se basa en la frecuencia con la que aparecen juntas. No se aplican reglas de tipo sintáctico o gramatical.

A su vez, a cada par de términos se les asocia un valor que representa la relación que existe entre ellos. Con esta información se forma una "red semántica", que se visualiza mediante una estructura jerárquica. Cada concepto representa un nodo en el árbol. Sus nodos hijo representan los conceptos con los que está relacionado.

La figura 3 muestra la interface gráfica utilizada por *Megaputer* para mostrar las relaciones entre los conceptos. En la parte superior izquierda de la ventana se listan los conceptos con su ponderación (escrita a la izquierda del nombre).

Al desplegar un término aparecen indicados los conceptos con los que está relacionado. En la imagen del ejemplo, el concepto *European identity* está relacionado con *identity*, *citizenship*, *European Union*, *creation*, etc. Los dos números que aparecen a la izquierda de cada término relacionado representan, respectivamente, el valor asignado por el programa a la relación entre los dos términos y el valor asignado al término hijo en el documento (y que es independiente de la relación).

Esta jerarquía permite identificar conceptos o términos relacionados con un término tomado como punto de partida. La relación jerárquica cuenta con múltiples niveles.

Para facilitar la identificación de los fragmentos del texto en los que aparece cada término, al hacer clic sobre cada nodo del árbol el panel de la derecha mostrará los fragmentos en los que éste aparece. Es posible descubrir únicamente aquellos fragmentos en los que aparezca el término junto con sus términos "padre" en la jerarquía. Es decir, los fragmentos en los cuales los dos términos relacionados co-ocurren.

Finalmente, el panel inferior recoge el texto completo del documento. Si en el panel superior derecho identificamos una frase que puede ser relevante, al hacer clic en ella podremos ver —en el panel inferior— la frase contextualizada en la totalidad del documento.

La aplicación también ofrece un sistema de búsqueda a texto completo. La consulta no sólo ofrece como resultado los fragmentos en los que ocurre el término buscado, sino que también muestra un listado de los conceptos o términos relacionados con los de búsqueda en un formato jerárquico similar al que se utiliza para recorrer la red semántica.

Esta funcionalidad resulta útil, ya que podemos recorrer una gran colección de textos con mayor facilidad, viendo aquellos conceptos que pueden resultar más relevantes, pero: ¿realmente se trata de una aplicación con un comportamiento inteligente, o que facilita el descubrimiento de nuevo conocimiento en la línea promulgada por **Hearst**?

Para resolver esta duda hemos realizado una prueba con una colección de noticias sobre la Guerra de Irak recogidas del sitio web *BBC Mundo*. Un primer problema ha sido la identificación del idioma, ya que el sistema no ha reconocido las palabras vacías en castellano.

Sin embargo, en muchos casos la identificación de conceptos no ha sido todo lo inteligente que podríamos esperar. Por ejemplo, el hecho de que **Rigoberta Menchú** opine sobre la guerra no queda plasmado en el árbol de conceptos generado por la aplicación. Obviamente, si el término no aparece con la frecuencia necesaria, la aplicación lo descarta y se pierde esta información en la red semántica. Sí aparece, por ejemplo, **Koichiro Matssusa** —director general de la *Unesco*—, relacionado con el término "saqueos", o la empresa *Bechtel* con el término "reconstrucción" (figura 4).

El problema de las palabras vacías se puede solucionar añadiendo un diccionario de palabras no analizables a la aplicación. Tras añadir este diccionario, la relación entre los conceptos identificados mejora ostensiblemente.

Podemos decir que esta aplicación supone una interesante ayuda para facilitar la lectura de textos procedentes de distintas fuentes, contrastar información, y sintetizar múltiples documentos.

Sin embargo, no ofrece todas las funciones propias de una aplicación de minería textual, como serían la categorización de documentos, el análisis cluster, o la extracción de hechos. Por ejemplo, el sistema identifica términos relacionados a partir de su índice de co-ocurrencia, pero no es capaz de identificar la relación real que existe entre ellos, incluso si esta está presente en el texto analizado.

Como resumen, podemos decir que *TextAnalyzer* actúa como un índice inverso de una base de datos documental que facilita la navegación entre los fragmentos indexados y la identificación de conceptos o términos relacionados a partir de la co-ocurrencia.

IBM Intelligent Miner 2.3.1. Esta es otra aplicación comercial descrita en el libro de **Sullivan**. Para analizar las funciones de este programa hemos consultado el libro de **Sullivan** y la documentación oficial de *IBM* disponible en su sitio web. No ha sido posible obtener una copia de evaluación para esta aplicación.

Intelligent Miner for Text es una herramienta complementaria a *Intelligent Miner for Data*, si bien ambas funcionan de forma independiente.

La aplicación consiste en una serie de ficheros ejecutables en entornos *Windows* y *Unix* desde línea de comandos. *Intelligent Miner* es un buen ejemplo de aplicación comercial que soporta las principales técnicas características de la minería textual y que hemos descrito en los apartados anteriores.

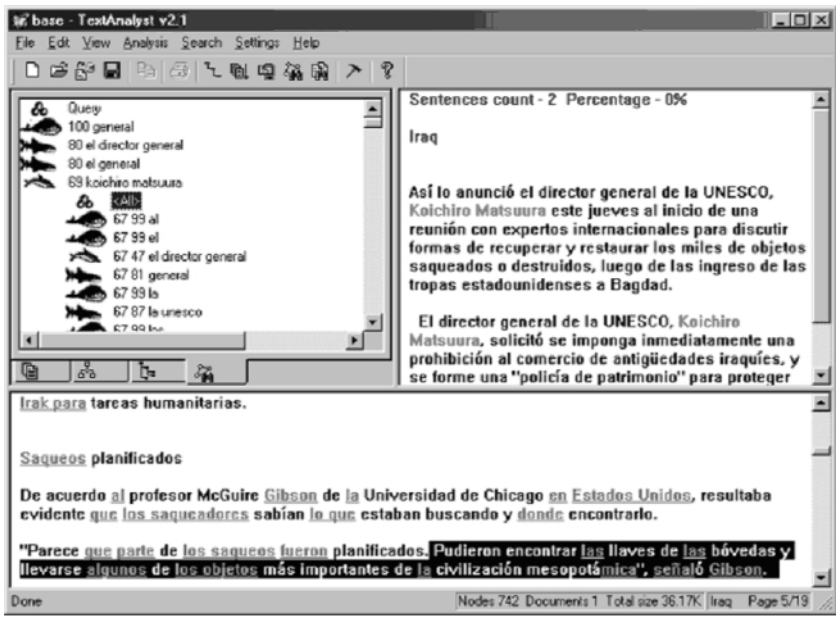


Figura 4

La aplicación incluye las siguientes utilidades:

—*Language Identification*: para identificar el idioma de un documento.

—*Topic Categorization*: para la clasificación automática de documentos en categorías predefinidas.

—*Feature extraction*: para extraer nombres de personas, lugares, organizaciones, etc., de los documentos y también relaciones entre ellas.

—*Clustering*: para agrupar automáticamente los documentos, pudiéndose aplicar un clustering binario y jerárquico.

—*Summarizer*: para extraer los fragmentos más significativos de un documento y obtener un resumen de manera automática.

Estas cinco utilidades reciben el nombre de herramientas *Text Analysis*. También forma parte de *Intelligent Miner for Text* un motor de indexación —*Text Search Engine*—, un indexador de sitios web remotos *Web Crawler* y un indexador de intranets llamado *Net-Question Solution*.

El módulo de identificación de idiomas está preparado para reconocer distintos idiomas. Sin embargo las herramientas de análisis de textos (*clustering*, categorización y *feature extraction*) únicamente están preparadas para trabajar con inglés. Esto no significa que no se puedan utilizar con documentos en otros idiomas. Simplemente, los resultados no serán los óptimos ni tendrán la misma exactitud y precisión que si se trabaja con una colección de documentos en inglés.

Sobre *Feature Extraction* únicamente señalaremos que se apoya en una serie de diccionarios de autoridades y en unas reglas heurísticas. Estas reglas se basan en el contexto o los patrones en los que aparecen las palabras que pueden ser consideradas nombres propios. A partir de la regla más simple (toda palabra que empiece con una mayúscula es potencialmente un nombre propio salvo si se encuentra al principio de la frase, en cuyo caso podrá serlo o no), se comprueban reglas adicionales, como las palabras que la preceden o la siguen y que pueden indicar cual es su tipo (persona, lugar, organización, cantidades monetarias, fechas, etc.).

Otra característica destacable es la capacidad de identificar qué nombres propios —de los identificados en el texto—, pueden corresponder a una misma entidad. Por ejemplo, el sistema identificaría los nombres propios **José María Aznar** y Sr. **Aznar** como referentes a una misma persona, y optaría por uno de ellos como su forma "canónica" o autorizada. Se define la forma canónica de un nombre como "el nombre más explícito y menos ambiguo construido a partir de las distintas variantes que aparecen en un documento".

La segunda característica importante de esta herramienta, es la capacidad de identificar relaciones entre los nombres propios. Las relaciones citadas en la documentación son: "edad", "status profesional", "hace", "dependencia", "relación familiar", "origen", "similitud", etc. Estas relaciones se pueden deducir a partir de los verbos y del contexto que circunda a los términos (por ejemplo, la presencia de los términos "fabricar" o "en la fabricación de" indicaría un caso de correspondencia con "hacer"). Las relaciones se expresan mediante tripletas del tipo <objeto1> <relación> <objeto2>.

En relación al funcionamiento de las herramientas que forman *IBM Intelligent Miner*, ya hemos señalado que se ejecutan mediante línea de comandos. Los documentos procesados deben estar en formato ascii o html. El resultado de su ejecución es un documento en texto plano con marcas.

Sobre la herramienta *Summarization* —para realizar resúmenes automáticos—, en la documentación de *IBM* se señala cómo la herramienta seleccionará aquellas frases o fragmentos:

—Que contengan a los términos con una ponderación en el documento superior a la que tienen en el total de la colección, que aparezcan más de una vez, y

—también se dará más peso a las frases más próximas al principio y al fin de cada párrafo.

En la realización de resúmenes se combina tanto las técnicas estadísticas y la frecuencia de los términos como la posición que éstos ocupan en el documento procesado.

Las distintas herramientas que conforman *Intelligent Miner for Text* están dirigidas a programadores e integraciones de soluciones para minería textual. Como podemos ver, en muchas ocasiones será necesario utilizar de forma conjunta más de una herramienta, e integrarlas en una aplicación consistente y fácil de usar para un usuario final. Por ejemplo, *Summarization* se suele utilizar tras haber ejecutado previamente *Feature Extraction*, ya que el resultado obtenido por ésta puede resultar útil para identificar las frases o fragmentos relevantes para hacer el resumen.

SAS Text Miner. Esta aplicación incorpora las funcionalidades características de la minería textual, y que hemos descrito en los apartados anteriores. Entre ellas:

—Capacidad de procesar documentos en distintos formatos (pdf, ascii, html, *Microsoft Office*, etc.), y extraer términos simples y compuestos (formados por más de una palabra), eliminar palabras vacías y reducir las palabras a lexemas. El sistema está optimizado para los idiomas inglés, francés y alemán.

—Identificar la función gramatical de la palabra en el texto con el objeto de evitar posibles ambigüedades (*part-of-speech tagging*).

—En el caso del idioma alemán, posibilidad para descomponer palabras en los distintos "lexemas" que la forman.

—Representación de documentos mediante un vector de términos ponderados según su frecuencia estadística.

—Identificación de nombres propios (*feature extraction)*.

—Agrupación automática (clustering) de documentos, aplicando algoritmos de clustering jerárquico y difuso.

—Categorización automática de documentos.

Todas estas funciones están disponibles a través de una interface de consulta que permite ver simultáneamente documentos, términos, conceptos y clusters, y recorrer y analizar la colección de documentos a través de estos listados.

Sin embargo, en la documentación consultada no se cita de forma explícita ni el resumen automático ni la creación de relaciones entre términos a partir de su co-ocurrencia. Según estos documentos, *SAS Text Miner* considera las tareas de categorización y agrupación automática como el principal objetivo de la minería textual; las actividades previas (identificación de términos simples y compuestos, extracción de nombres propios, ponderación, etc.) son tareas de pre-procesamiento, que deben realizarse previamente para hacer posible la categorización y la clasificación.

Spss LixiQuest. En este caso se trata de tres productos: *LexiQuest Mine*, *LexiQuest Categorize* y *LexiQuest Guide*. Los dos primeros ofrecen funcionalidad propia de la minería textual. El último consiste en tecnología de indexación y búsqueda de documentos en intranets e internet.

LexiQuest Mine se presenta como una herramienta cuyo propósito es "automatizar el proceso de leer documentos para así descubrir su contenido". En la presentación de esta herramienta se recurre al hecho de que, de un documento de treinta páginas, tan sólo tres párrafos pueden resultarnos relevantes. El problema se acentúa cuando tenemos que leer un gran número de documentos.

LexiQuest Mine ofrece la posibilidad de procesar un gran número de documentos e identificar los términos y nombres propios que aparecen en ellos, así como mantener información sobre términos relacionados a partir de su co-ocurrencia.

LexiQuest Categorize ofrece la función de categorización automática de documentos a partir de un entrenamiento previo (no se trata de clustering).

Conclusiones. Minería textual y recuperación de información: ¿diferencias reales?

Tras analizar la forma de trabajar de algunas de las aplicaciones que se presentan como sistemas de minería textual nos surge la duda de si realmente existen diferencias significativas entre estas aplicaciones y las de búsqueda y recuperación de información tradicionales.

Los fabricantes de motores de indexación tradicionales (*Verity*, *Autonomy*, *Fulcrum*, etc.) —claramente posicionados en el mercado de gestión documental y recuperación textual para internet e intranet— vienen ofreciendo desde hace bastante tiempo funciones para la categorización automática de documentos y la generación de clusters y resúmenes. La función "buscar más documentos similares" que estamos acostumbrados a utilizar en los motores de búsqueda para la web se basan en el mismo modelo de agrupación automática y cálculo de similitudes que la técnica de generación de clusters.

Por otra parte, tanto las aplicaciones de minería textual como los indexadores parten de un mismo modelo para representar los documentos que procesan: el vectorial y la ponderación de los términos para identificar aquellos que resulten más significativos para representar el contenido de los documentos.

El hecho de partir de un mismo modelo de representación de los documentos y compartir funcionalidades críticas dificulta el trazar una línea divisoria clara entre minería textual y recuperación de información avanzada. Podríamos decir que la principal diferencia entre estas dos corrientes de la informática documental consiste en:

1. La minería textual hace explícitas algunas de las características utilizadas tradicionalmente por los sistemas de recuperación de información, pero que permanecían ocultas para los usuarios de estos sistemas.

Así, la navegación entre términos que ofrecen aplicaciones como *Text Analyzer* de *Megaputer* podría considerarse como una exteriorización de los índices inversos utilizados por las aplicaciones de recuperación textual. En la misma línea, la visualización de clusters sería el resultado de mostrar la lógica que se encuentra detrás de la popular función de búsqueda "ver documentos similares".

La función de generar resúmenes de forma automática —que hemos citado como característica de las aplicaciones de minería textual—, también forma parte de los sistemas de recuperación de información. Por ejemplo, la mayoría de los indexadores ofrecen una síntesis de cada documento recuperado consistente en los fragmentos que el sistema juzga más representativos.

2. Otra diferencia entre minería textual y sistemas de recuperación de información la encontramos en las interfaces que usamos para interactuar con las colecciones de documentos.

Las aplicaciones de minería textual ofrecen una interface que facilita el análisis y la lectura de los datos extraídos de distintos documentos. Es posible recuperación y visualización fragmentos de distintos documentos, ver las relaciones entre términos de una forma explícita, etc.

Frente a esto, las interfaces de los sistemas de recuperación de información están más centradas en el documento (y no en la colección o en el conocimiento que éstos contienen), con interfaces excesivamente rudimentarios que únicamente permiten la interacción basada en el patrón "formulación de la búsqueda->lista de resultados->visualización individual de cada documento recuperado".

3. Por último, recordar algo comentado con anterioridad en este texto, y es que el objetivo de la recuperación de información se centra en establecer los mecanismos para satisfacer las necesidades de información de un usuario. Sin embargo, en la minería textual no sólo no tiene por qué existir esa necesidad sino que tampoco es necesario formular una pregunta concreta al sistema.

Estas similitudes y diferencias nos permiten considerar a la minería textual como una evolución de los sistemas de recuperación de información tradicionales, con un cambio en su alcance y objetivo. Quizás la principal diferencia entre estas dos aproximaciones sea el mayor énfasis de la minería textual en la identificación de nombres propios, conceptos y en la identificación de las relaciones que existen entre ellos.

Por otra parte, un aspecto importante a favor de la minería textual es que con ella se están materializando (en forma de aplicaciones software comerciales y proyectos en empresas e instituciones), muchas técnicas desarrolladas en el marco de la recuperación de información años atrás, y que hasta hace poco tiempo habían únicamente tenían un interés académico. El hecho de que las herramientas de clustering aún no estén generalizadas en las aplicaciones comerciales de recuperación textual a pesar de que su desarrollo teórico se remonta a los años 70 es una muestra de esta situación.

Como resumen, podemos afirmar que las herramientas de minería textual ofrecen una ayuda importante en el proceso de acceder e interpretación la información disponibles en los documentos. Probablemente, en un futuro se complete la función actual con mayores capacidades deductivas por parte de las aplicaciones informáticas. Esto exigirá una mayor capacidad de los ordenadores para interpretar el lenguaje natural (y no sólo procesarlo, como pueden hacer a día de hoy). A la espera de que esto ocurra, los primeros resultados que podemos apreciar en las aplicaciones citadas en el apartado anterior pueden juzgarse satisfactorios.

Notas

1. Concretamente en el *StatSoft electronic textbook*, disponible en: *http://www.statsoft.com/textbook*

Bibliografía sobre el tema

Berry, Michael W. *Survey of text mining: clustering, classification, and retrieval*. Septiembre, 2003.

Chang, George; **Healey Marcus J.** (eds.). *Mining the world wide web - an information search approach-*. [Dordrecht]: Kluwer, junio, 2001. 192 p.

Franke, J.; **Nakhaeizadeh, G.**; **Renz, I.** (eds.). *Text mining: theoretical aspects and applications*. Heidelberg: Physica Springer-Verlag, 2003.

Halliman, Charles. *Business intelligence using smart techniques: environmental scanning using text mining and competitor analysis using scenarios and manual simulation*. 2001, 223 p.

Thuraisingham, Bhavani M. *Web data mining and applications in business intelligence and counter-terrorism*. [s.l.]: CRC Press, junio, 2003, 592 p.

Tramullas Saz, Jesús. "Perspectivas en recuperación y explotación de información electrónica: el data mining". En: *Scire*, 1997, v. 3, n. 2, pp. 75-83.

Bibliografía utilizada

Arenas, Lourdes; **Moral, Anselmo del**. "Automatic indexing of documents". En: *Nuevas tendencias en inteligencia artificial*. Bilbao: Universidad de Deusto, 1992, pp. 355-367.

Ananyan, Sergei; **Kharlamov, Alexander**. Automated analysis of natural language texts. White paper de Megaputer. Consultado en: 27-12-03. *http://www.megaputer.com/tech/wp/tm.php3*

Baeza-Yates, Ricardo; **Ribeiro-Neto, Berthier**. *Modern information retrieval*. Harlow: Addison-Wesley , 1999, 514 p.

Berry, Michael W.; **Drmac, Zlatko** ; **Jessup, Elizabetch R**. "Matrices, vector spaces and information retrieval". En: *Siam Review*, abril, 1999, v. 41, n. 2, pp. 335-362. Consultado en: 27-12-03. *http://www.siam.org/journals/sirev/41-2/34703.html*

Chakrabarti, Soumen. *Mining the Web: Discovering Knowledge From Hypertext Data*. Amsterdam: Morgan Kaufmann, 2003, xviii, 345 p.

Cutting, Douglas R. [et al.]. "Scatter/gather: a cluster-based approach to browsing large document collections". En: *15th Annual International Sirgi 92*. Consultado en: 27-12-03. *http://citeseer.nj.nec.com/cutting92scattergather.html*

Darpa information access office web site. Consultado en: 27-12-03. *http://www.darpa.mil/iao*

Deerwester, Scott [et al.]. "Indexing by latent semantic analysis". En: *Journal of the American Society for Information Science*, 1990, v. 41, n. 6, pp. 391-407. Consultado en: 27-12-03. *http://lsa.colorado.edu/papers/JASIS.lsi.90.pdf*

Etxeberría Murgiondo, Juan [et al.]. *Análisis de datos y textos*. Madrid: Ra-ma, 1995, xi, 372 p.

Kent, Allan; **Lancour, Harold** (ed.). *Encyclopedia or library and information science*. New York: Marcel Dekker, 1968-1989.

FAS (Federation of American Scientist) web site. Consultado en: 27-12-03.*http://www.fas.org*

Gemert, Jan Van. "Text mining tools on the internet: an overview". En: *Isis Technical Report Series*, septiembre, 2000, v. 23. Consultado en: 27-12-03. *http://www.ai.mit.edu/people/jimmylin/papers/Gemert00.pdf*

Hearst, Marti. "Untangling text data mining". En: *Proceedings of ACL'99: the 37th annual meeting of the Association For Computational Linguistics*, junio, 1999. Consultado en: 27-12-03.
http://www.sims.berkeley.edu/~hearst/papers/acl99/acl99-tdm.html

IBM. *Intelligent miner for text: guía de iniciación versión 2.3*. 2ª ed. [S.l.]: IBM, diciembre, 1998. vii, 53 p. Publicación número SH 10-9238-01.

IBM. *Text analysis tools: intelligent miner for text version 2.3.1*. [S.l.]: IBM, junio 2000, viii, 92 p. Publicación número SH 12-6370-01.

Frakes, William B.; **Baeza-Yates, Ricardo** (eds.). *Information retrieval: data structures & algorithms*. New Jersey: Prentice Hall, 1992, viii, 504 p.

Jain, A. K.; **Murty, M. N.**; **Flynn, P. J**. "Data clustering: a review". En: *ACM Computing Surveys*, septiembre, 1999, v. 31, n. 3, pp. 265-323. Consultado en: 27-12-03.
http://citeseer.nj.nec.com/jain99data.html

Klir, George J.; **Yuan, Bo**. *Fuzzy sets and fuzzy logic: theory and applications*. New Jersey: Prentice Hall, 1995, xv, 574 p.

Landauer, Thomas K.; **Foltz, Peter W.**; **Lahan, Darrell**. "An introduction to latent semantic analysis". En: *Discourse Processes*, 1998, n. 25, pp. 259-284. Consultado en: 27-12-03.
http://lsa.colorado.edu/papers/dp1.LSAintro.pdf

Lucas, Marty. *Mining in textual mountains.* [Entrevista a **Marti Hearst** realizada el 18 de noviembre de 1999]. Consultado en: 27-12-03.
http://mappa.mundi.net/trip-m/hearst

Maron, M. E.; **Kuhns, J. L**. "On relevance, probabilistic indexing and information retrieval". En: *Journal of the ACM*, 1960, v. 7, n. 3, pp. 216-244.

Miyamoto, Sadaaki. *Fuzzy sets in information retrieval and cluster analysis*. Dordrecht : Kluwer Academic Publishers , 1990, x, 258 p.

National strategy for combating terrorism. Febrero 2003. Consultado en: 27-12-03.
http://www.whitehouse.gov/news/releases/2003/02/counter_terrorism/counter_terrorism_strategy.pdf

Pao, Miranda Lee. *Concepts of information retrieval*. Englewood, Colorado: Libraries Unlimited, 1990.

Ravin, Yael. *Extracting names from natural-language text: IBM research report*. Almaden: IBM research division, 04/10/1997 (RC 20338 digital libraries), 30 p.

Rijsbergen, C. J. Van. *Information retrieval*. 2ª ed. London [etc.]: Butterworths, 1979 (reimp. 1980).

Salton, Gerard; **McGill, Michael J.** *Introduction to modern information retrieval*. New York [etc.]: McGraw Hill, 1983.

SAS Institute Inc. Getting started with SAS text miner software, release 8.2. Pubcode: 58859. Consultado en: 27-12-03.
http://www.sas.com

SAS text miner: distilling textual data for competitive business advantage: a SAS white paper. Consultado en: 27-12-03.
http://www.sas.com

"SAS signs text mining alliance with Inxight". En: *eWeek*, 19 de enero de 2002. Consultado en: 27-12-03.
http://www.eweek.com

Sebastiani, Fabrizio. "Machine learning in automated text categorization". En: *ACM Computing Surveys*, marzo, 2002, v. 34, n. 1, pp. 1-47.

StatSoft textbook. Consultado en: 27-21-03
http://www.statsoft.com/textbook

Sullivan, Dan. *Document warehousing and text mining*. New York [etc.]: Wiley Computer Publishing, 2001, xviii, 542 p.

Swanson, Don R. "Assessing a gap in the biomedical literature: Magnesium deficiency and neurologic disease". En: *Neuroscince Research Communications*, 1994, n. 15, pp. 1-9.

Swanson, Don R. "An interactive system for finding complementary literatures: a stimulus to scientific discovery". En: *Artificial Intelligence*, 1997, n. 91, pp. 183-203.

Swanson, Don R. "Two medical literatures that are logically but not bibliographically connected". En: *Journal of the American Society for Information Science*, 1987, v. 38, n. 4, pp. 228-233.

Tan, Ah-Hwee. "Text mining: the state of the art and the challenges". En *Proceedings Pakdd'99 workshop on knowledge discovery from advanced databases*, abril, 1999, pp. 71-76. Consultado en: 27-12-03.
http://textmining.krdl.org.sg/people/ahhwee/

Thomas, Timothy L. "Al qaeda and the internet: the danger of 'cyberplanning'". En: *Parameters*, primavera, 2003. Consultado en: 27-12-03.
http://fmso.leavenworth.army.mil/fmsopubs/ISSUES/alqaedainternet.htm

Watkins, D. S. *Fundamentals of matrix computations*. New York: John Wiley & Sons, 1991.

Yang, Y.; **Pedersen, J. O**. "A comparative study on feature selection in text categorization". En: *Proceedings of the fourteenth international conference on machine learning*, 1997.

Ricardo Eíto Brun
reito@bib.uc3m.es
Jose A. Senso
jsenso@ugr.es