Minería de textos
Text mining
Fouille de textes

Granada, a 10 de Mayo de 2016

Prof. Dr. José Pino Díaz

Universidad de Málaga, Andalucía Tech, Dep. de Economía y Administración de Empresas

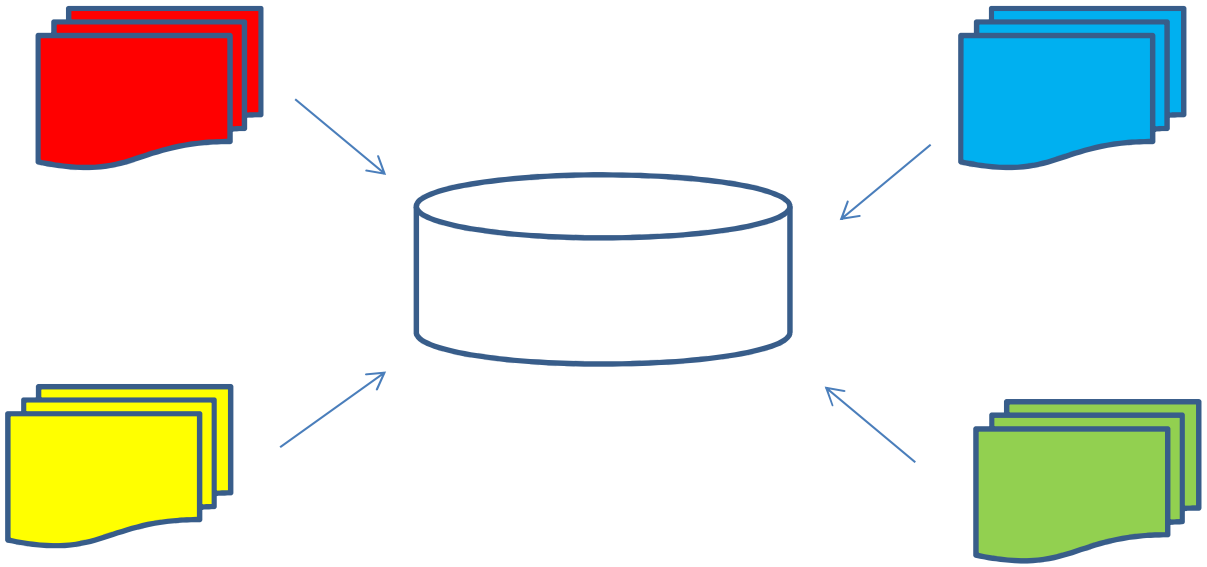
Grupos de investigación Techné (UGR) e iArtHis-Lab (UMA)

Campus de Teatinos s/n, 29071 Málaga, España

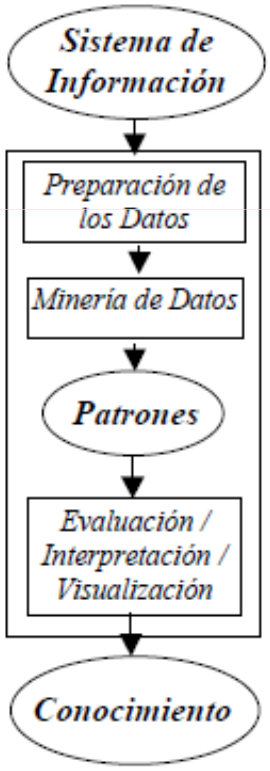
Databases

Una **base de datos** es un conjunto de información estructurada en registros y almacenada en un soporte electrónico legible por ordenador.

- Cada registro constituye una unidad autónoma de información que puede a su vez estar estructurada en diferentes campos o tipos de datos que se recogen en la base de datos
- Una base de datos se crea y mantiene de forma continuada con el objetivo de resolver necesidades de información concretas de un colectivo, una organización o el conjunto de la sociedad.



Creación de nuevo conocimiento a partir de bases de datos bibliográficas (*Knowledge Discovery in Databases, KDD*)

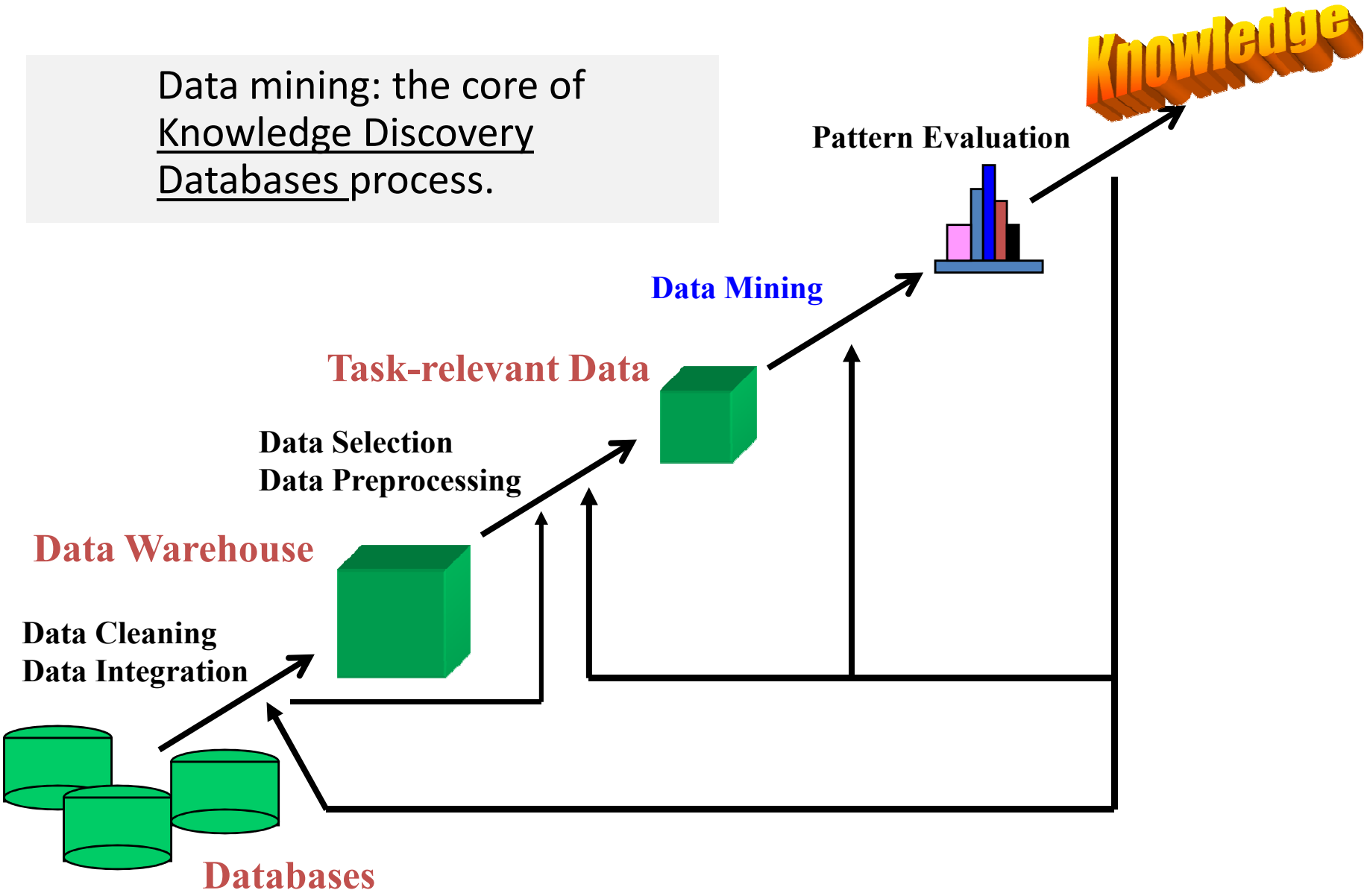


KDD

1. Determinar las fuentes de información que pueden ser útiles y dónde conseguirlas.
2. Diseñar el esquema de un almacén de datos (Data Warehouse) que consiga unificar de manera operativa toda la información recogida.
3. Implantación del almacén de datos que permita la "navegación" y visualización previa de sus datos, para discernir qué aspectos puede interesar que sean estudiados.
4. Selección, limpieza y transformación de los datos que se van a analizar. La selección incluye tanto una criba o fusión horizontal (filas) como vertical (atributos).
5. Seleccionar y aplicar el método de minería de datos apropiado.
6. Interpretación, transformación y representación de los patrones extraídos.
7. Difusión y uso del nuevo conocimiento.

Data Mining, a KDD Process

Data mining: the core of Knowledge Discovery Databases process.

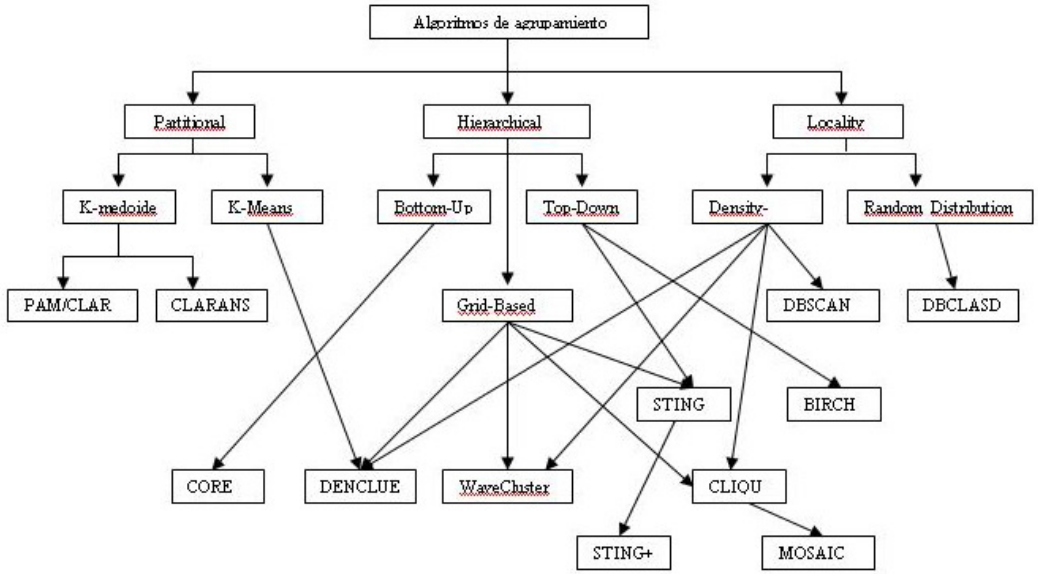


Minería de datos

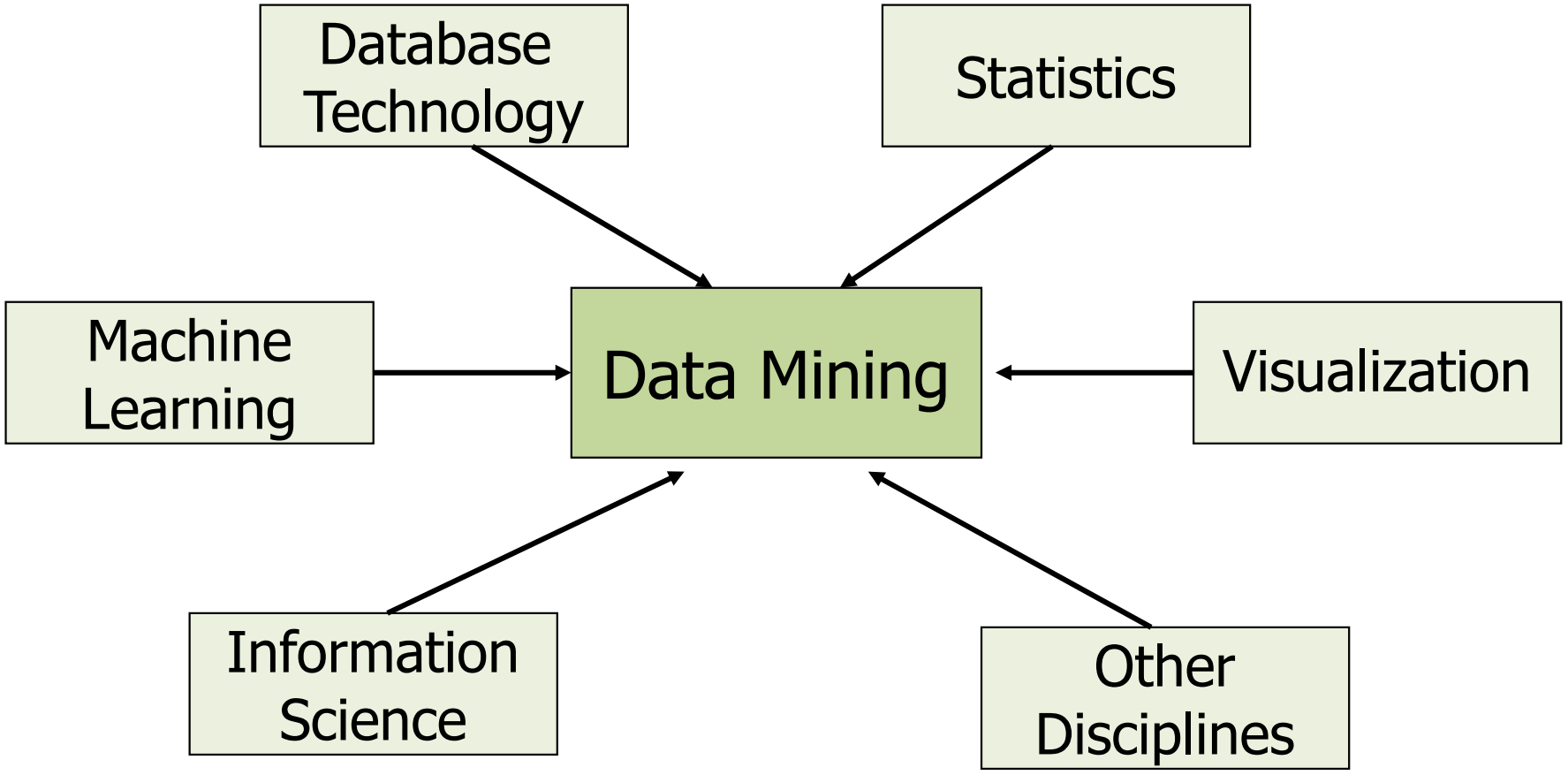
La **Minería de Datos** es la extracción dirigida de la información existente en las bases de datos con el fin de descubrir patrones, relaciones o asociaciones para generar nuevo conocimiento.

Algunos tipos de *DM*:

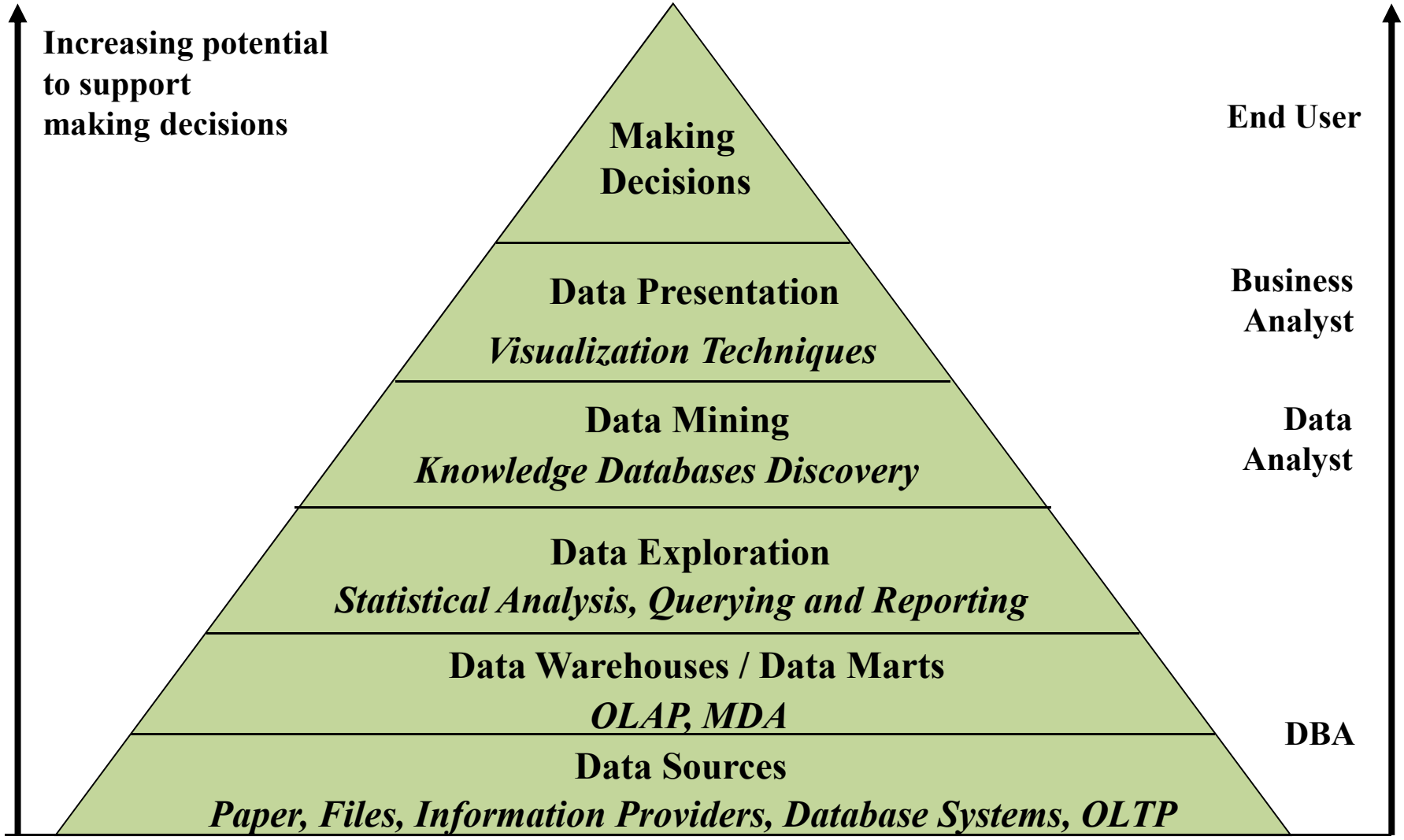
- **Web mining**
  - **Web content mining** (minería de contenido web)
  - **Web structure mining** (minería de estructura web)
  - **Web usage mining** (minería de uso web)
- **Text mining** (minería de datos textuales)
- **Spatial data mining** (minería de datos espaciales)



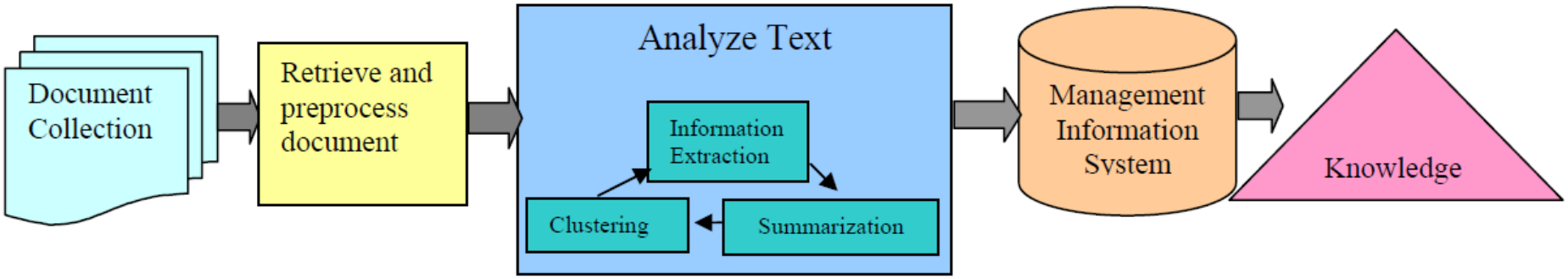
Data Mining: Confluence of Multiple Disciplines



Data mining and Making decisions



Text mining, a KDD Process



In 2001, Dow Chemicals merged with Union Carbide Corporation (UCC), requiring a massive integration of over 35,000 of UCC's reports into Dow's document management system. Dow chose ClearForest, a leading developer of text-driven business solutions, to help integrate the document collection. Using technology they had developed, ClearForest indexed the documents and identified chemical substances, products, companies, and people. This allowed Dow to add more than 80 years' worth of UCC's research to their information management system and approximately 100,000 new chemical substances to their registry. When the project was complete, it was estimated that Dow spent almost $3 million less than what they would have if they had used their own existing methods for indexing documents. Dow also reduced the time spent sorting documents by 50% and reduced data errors by 10-15%.

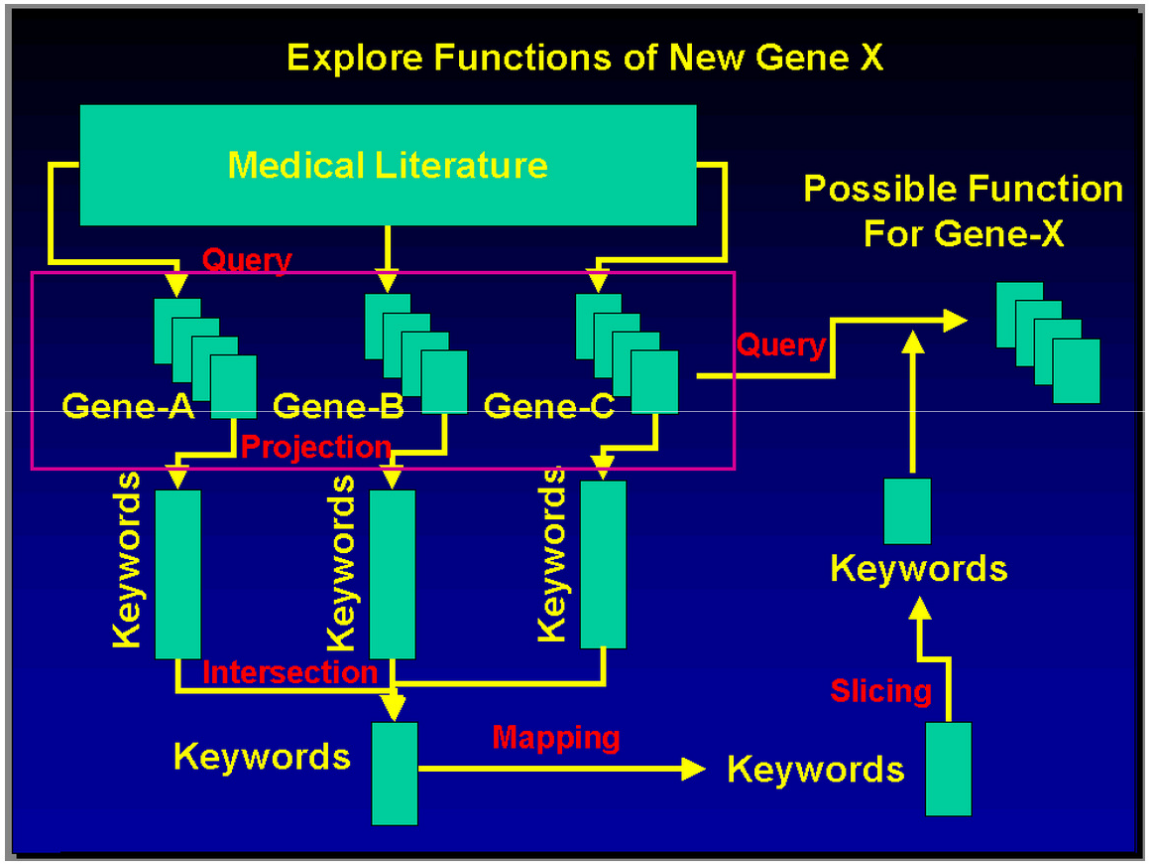
En 2001, Dow Chemicals se unió a Union Carbide Corporation (UCC). Esta unión requirió la integración de 35.000 informes de UCC en el sistema de gestión del documentos de Dow. Dow Chemicals eligió a la empresa ClearForest, líder en *text-driven business solutions*, para realizar la integración de la colección de documentos. Usando tecnología ad-hoc se identificaron sustancias químicas, productos, empresas, centros y personas. Esto permitió que Dow Chemicals agregara más de 80 años de investigación de UCC a su sistema de gestión de información y aproximadamente 100.000 nuevas sustancias químicas a su registro. Cuando el proyecto se completó, se estimó que Dow Chemicals ahorró casi 3 millones de dólares y que el tiempo empleado en clasificar los documentos se redujo un 50% y los errores de datos entre un 10-15%.

Text mining applied

	information extraction	topic tracking	summarization	categorization	clustering	concept linkage	information visualization	question answering
Medical:								
FAQ's	x			x		x		x
Drug design	x				x	x		
New treatment		x				x		
Business:								
Competitive Analysis		x	x					
Media impact / analysis		x						
Current Awareness		x						
Intellectual property infringement	x	x			x			
Customer support for FAQ's	x			x	x			x
Social network detection							x	
Content personalization		x			x			
Government:								
Homeland security: detecting terrorist networks	x	x			x	x	x	
Law enforcement: crime detection / prevention	x	x			x	x	x	
Education:								
Research on a topic		x	x	x				
Citation analysis	x				x		x	
FAQ's	x			x	x			x

Table 3. Some examples of where text mining tools can be applied to the fields of medicine, business, government, and education.

A hypothetical text mining



Ejemplo hipotético de text mining para descubrir aspectos relacionados con un nuevo gen X. En la colección de textos biomédicos se realizan tres búsquedas sobre tres genes A, B, y C; …

En la acción final sólo se seleccionarán los documentos que contengan por lo menos una de las palabras clave que aparecen en la zona superior del ranking de palabras clave y que mencionen a los tres genes conocidos.

Text mining:
Concept linckage

Una aplicación muy popular del text mining es relatada en Hearst (*Untangling Text Data Mining,* 1999), Don Swanson intenta extraer información derivada de coleccionesde texto. Teniendo en cuenta que los expertos sólo pueden leer una pequeña parte de lo que se publica en su campo, por lo general no se dan cuenta de los nuevos desarrollos que se suceden en otros campos.

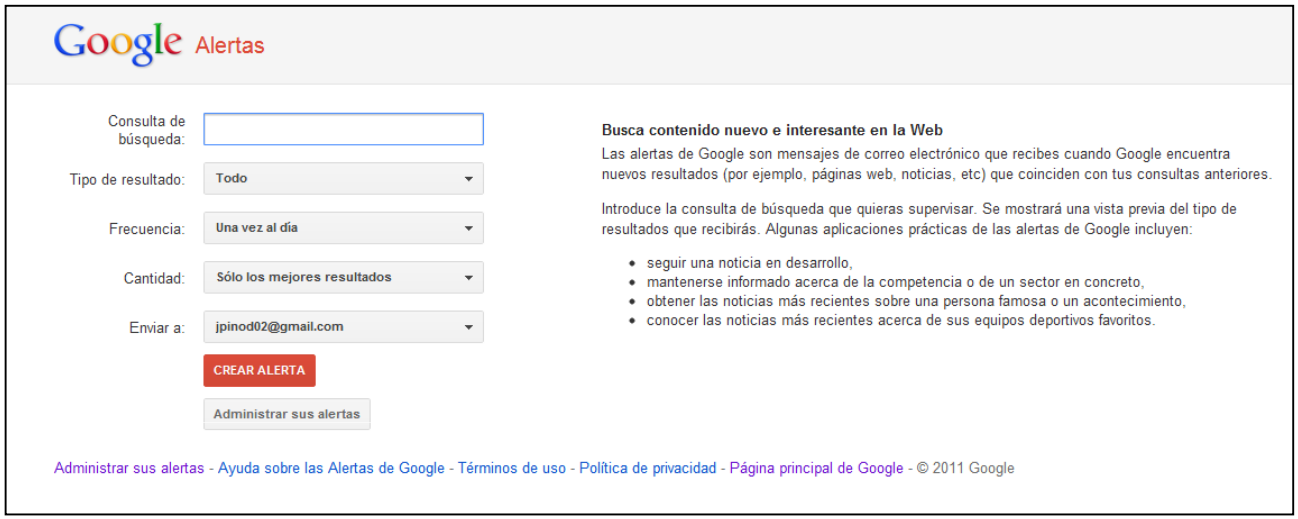
Así, Swanson ha demostrado cómo cadenas de implicaciones causales dentro de la literatura médica pueden conducir a hipótesis para enfermedades poco frecuentes, algunas de las cuales han recibido pruebas de soporte experimental. Investigando las causas de la migraña, dicho investigador extrajo varias piezas de evidencia a partir de títulos de artículos presentes en la literatura biomédica. Algunas de esas claves fueron:

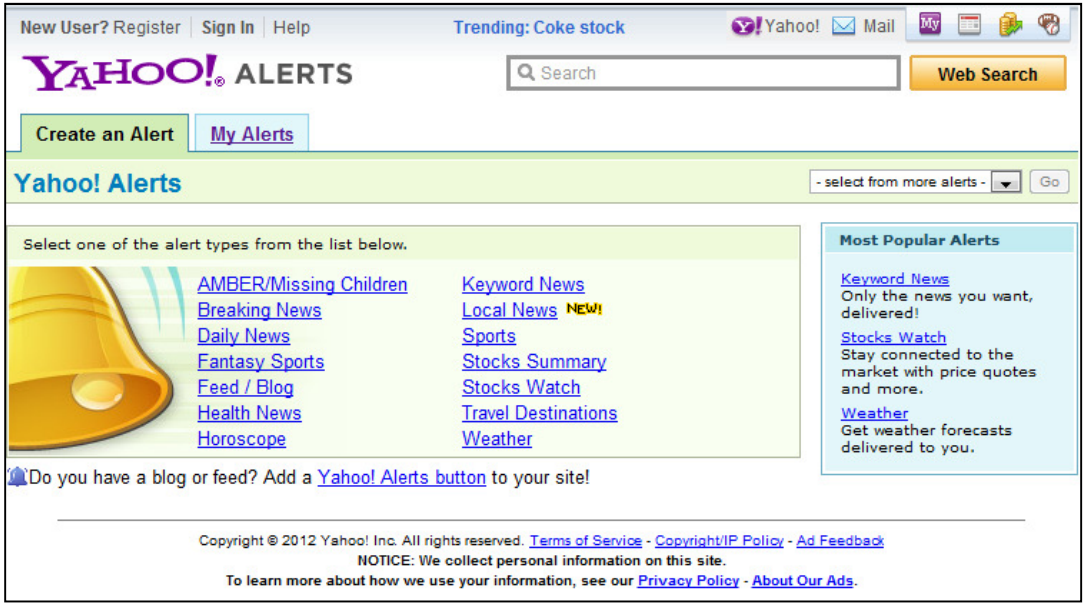
- El estrés está asociado con la migraña.
- El estrés puede conducir a la pérdida de magnesio.
- Los bloqueadores de canales de calcio previenen algunas migrañas.
- El magnesio es un bloqueador natural del canal de calcio.
- La depresión cortical diseminada (DCD) está implicada en algunas migrañas.
- Los niveles altos de magnesio inhiben la DCD.
- Los pacientes con migraña tienen una alta agregación plaquetaria.
- El magnesio puede suprimir la agregación plaquetaria.

Estas claves sugieren que la deficiencia de magnesio podría representar un papel en algunos tipos de migraña, una hipótesis que no existía en la literatura y que Swanson encontró mediante esas ligas. De acuerdo con Swanson (Swanson y otros, 1994), estudios posteriores han probado experimentalmente esta hipótesis obtenida por text mining con buenos resultados.

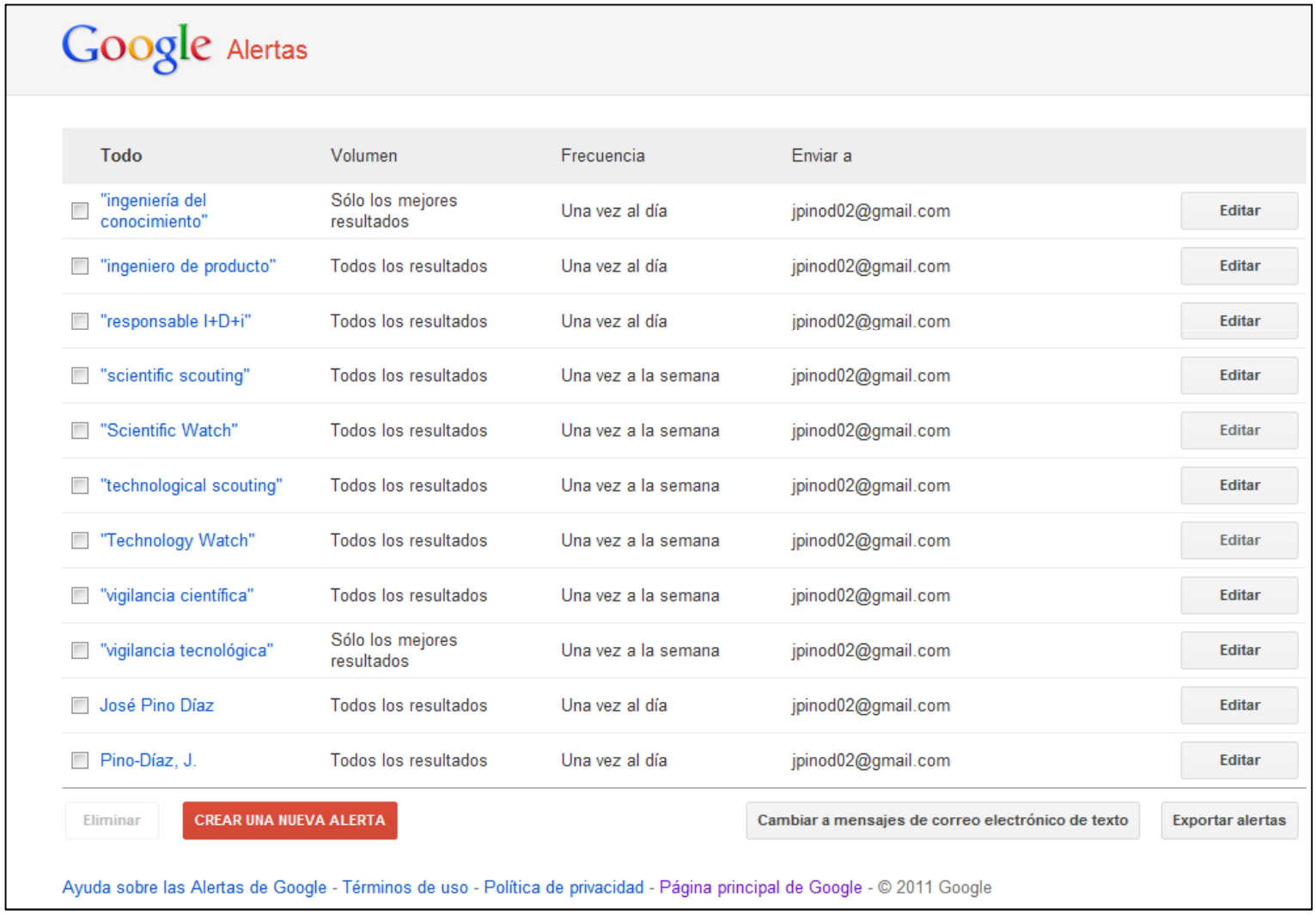
"Data mining: torturando a los datos hasta que confiesen".
http://www.uoc.edu/molina1102/esp/art/molina1102/molina1102.html

Text mining:
Topic tracking

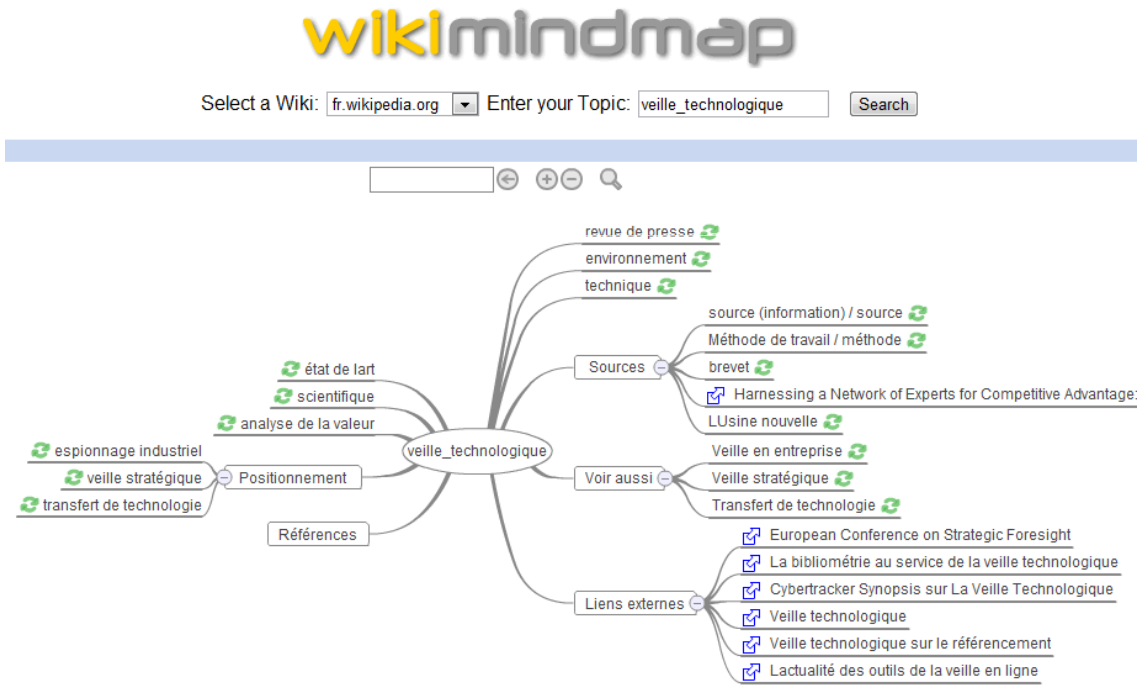


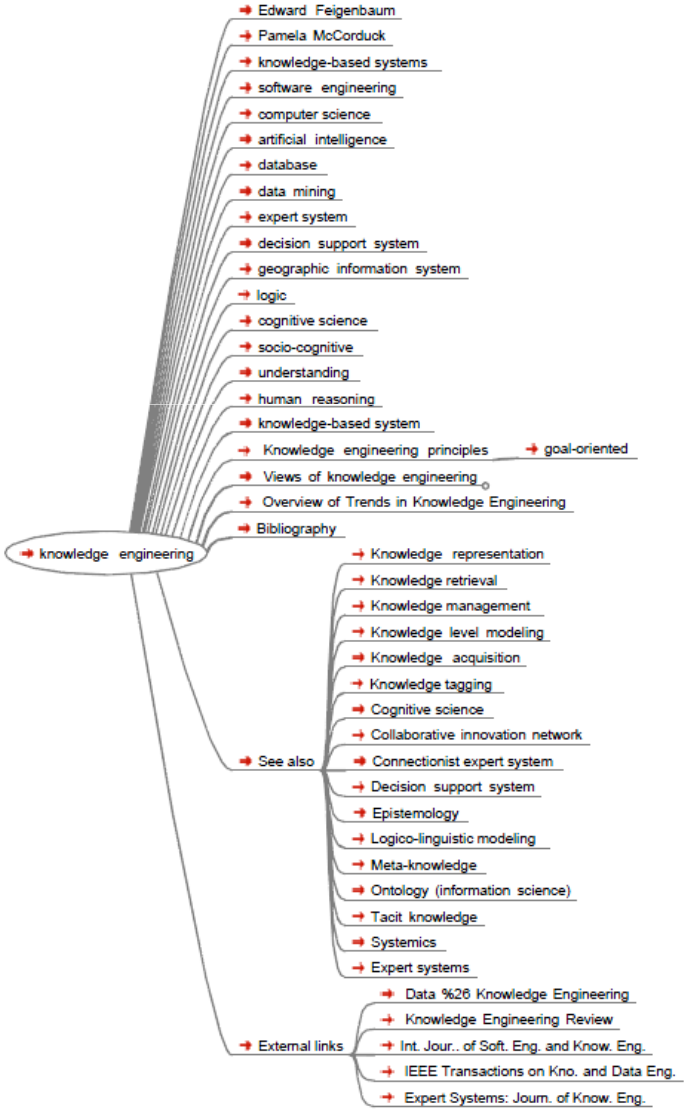


Text mining:
Topic tracking

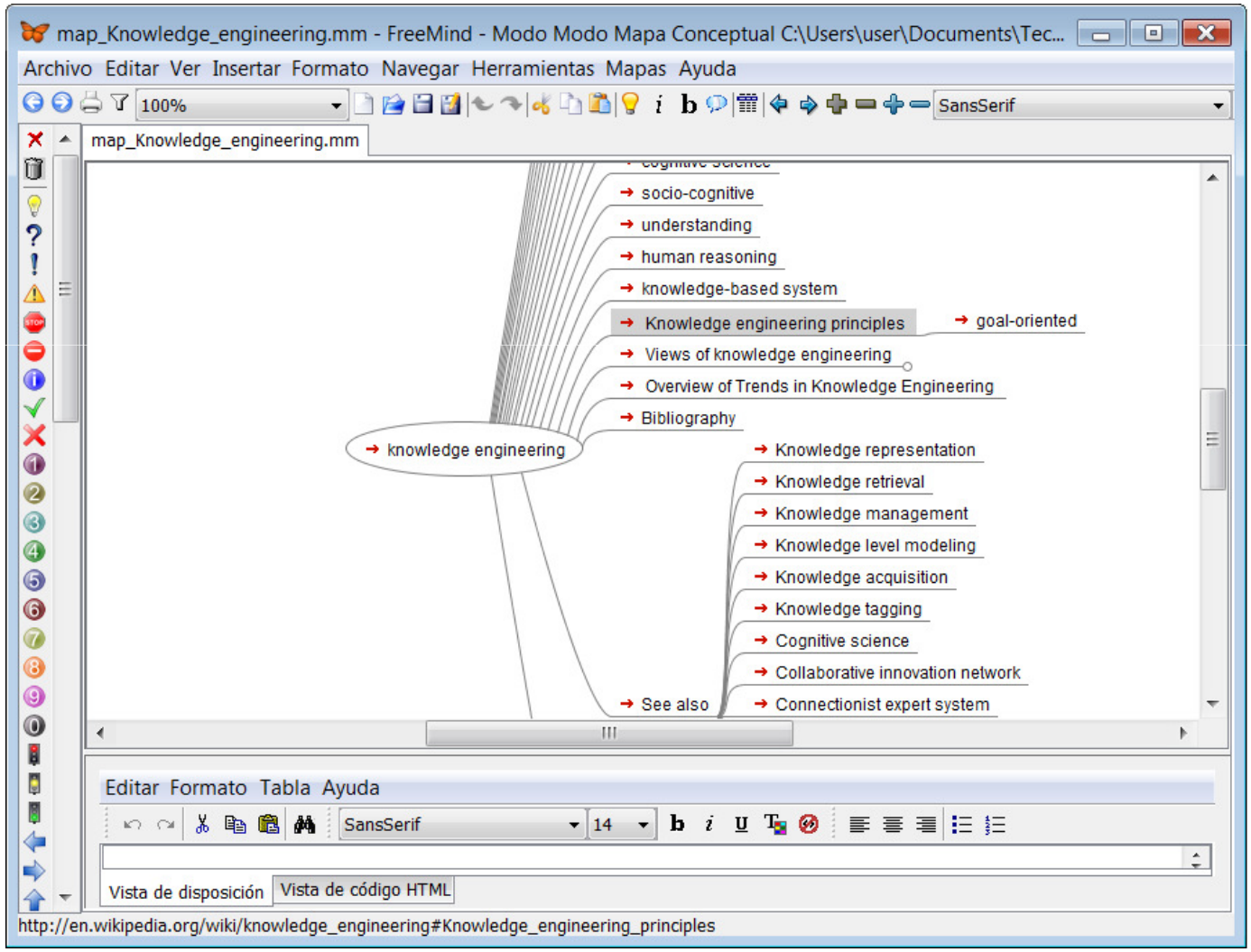


Text mining:
Information visualization





Text mining:
Information visualization



R.TeMiS [R Text Mining Solution]

R.TeMiS

Une approche intégrée et libre de l'analyse statistique de données textuelles

À propos | Téléchargement & Installation | Fonctionnalités | Exportation de corpus depuis Factiva | Publications

À propos

R.TeMiS [R Text Mining Solution] est un environnement graphique de travail sous R permettant de créer, manipuler et analyser des corpus de textes. Il a été conçu pour limiter les effets de « boîte noire », souvent inhérents aux logiciels de statistique lexicale, et favoriser la réflexivité dans l'usage sociologique des données textuelles.

L'architecture statistique de l'environnement **R.TeMiS** est fournie par le paquet tm développé par Ingo Feinerer (Feinerer, 2008; 2011; Feinerer, Hornik & Meyer, 2008). Celui-ci a été complété par d'autres paquets classiques de R comme ca pour la représentation des analyses factorielles des correspondances (Nenadic & Greenacre, 2007). Enfin des paquets spécifiques ont été développés pour faciliter l'usage de **R.TeMiS** dans le domaine des études sur les médias, par exemple pour la gestion des corpus constitués depuis la base de données d'articles de presse Factiva.

Afin de faciliter l'usage de **R.TeMiS** aux néo-utilisateurs de R, le développement d'un

Search

Présentation

Ce site est destiné à faciliter la diffusion et le développement de R.TeMiS, une solution intégrée et libre pour l'analyse statistique de données textuelles sous R. R.TeMiS a été développé par Milan Bouchet-Valat et Gilles Bastin. This blog is the home page for the development and diffusion of R.TeMiS, an integrated and open-source solution for textual data analysis with R.

Nouvelles de R.TeMiS

- Article paru dans le BMS 31/03/2014
- Présentation de R.TeMiS au Congrès de l'AFS à Nantes

¿Preguntas?… ¡Comentarios!...

Muchas Gracias

Prof. Dr. José Pino Díaz

Universidad de Málaga, Andalucía Tech, Departamento de Historia del Arte,

Grupos de investigación Techné (UGR) e iArtHis-Lab (UMA),

Campus de Teatinos s/n, 29071 Málaga, España