

# Text Mining

---

Agosto 2007

# ¿Qué es Text Mining?

- El proceso de descubrir y extraer patrones y relaciones significativas de una colección de textos (datos no estructurados)
- Data Mining aplicado a grandes colecciones de textos combinado con procesamiento de lenguaje natural
- NO es ni recuperación de información (information retrieval), ni recuperación de datos (data retrieval), ni limpieza de datos (data quality)
- Tampoco es data mining tradicional, ya que este se aplica a datos estructurados

Búsqueda  
(dirigida por objetivos)

Descubrimiento  
(oportunista)

Datos  
estructurados

Data  
Retrieval

Data  
Mining

Datos No  
estructurados  
(Texto)

Information  
Retrieval

Text  
Mining

- “No vivimos en una época de explosión de la información sino en una de inundación de datos”. (Hunter, 1981)
- En el 1999 se produjeron 2 exabytes de nuevos datos.
- En el 2002 se produjeron 5 exabytes de nuevos datos
- Entre 1999 y 2002 los nuevos datos almacenados crecieron un 30% anual
- El 92% de los nuevos datos está almacenado en medios magnéticos, principalmente discos rígidos

*Fuente: How Much Information? 2003*

- Los datos almacenados se duplican cada 9 meses
- Solo 10% de los datos recolectados se usan alguna vez
- 40% de los datos recolectados tienen errores

*Fayyad et al. : Summary from the KDD-03 Panel, 2003*

- El flujo de datos a través de canales electrónicos fue de casi 18 exabytes en 2002
- El 98% de ese total fue enviado y recibido a través de conexiones telefónicas (17.3 EB), la mayoría persona a persona
- La World Wide Web contenía en 2003 170 TB de datos
- El correo electrónico generó 440 PB de nuevos datos en 2003 (para el 2006 se estiman 880 PB)
- En 2002 la base de datos más grande del mundo (datos experimentales, Stanford Linear Accelerator Center) tenía 500 TB

*Fuente: How Much Information? 2003*

<b>Producción impresa mundial en 2003</b>			
<b>Tipo de medio</b>	<b>Total (escaneado)</b>	<b>Total (texto)</b>	<b>Porcentaje</b>
Libros	39 TB	0.1 TB	2.3%
Diarios	138.4 TB	0.3 TB	8.5%
Folletería mercado	52 TB	0.07 TB	3.2%
Revistas académicas	6 TB	0.03 TB	0.37%
Boletines	0.9 TB	0.006 TB	0.05%
Documentos oficina	1.397 TB	11.6 TB	85,5%
<b>Total</b>	<b>1.634 TB</b>	<b>12 TB</b>	<b>100%</b>

*Fuente: How Much Information? 2003*

# Escala de medición

Kilobyte (KB)	$10^3$ bytes	Foto de baja resolución: 100KB
Megabyte (MB)	$10^6$ bytes	Obras completas de Shakespeare: 5MB
Gigabyte (GB)	$10^9$ bytes	Obras completas de Beethoven: 20 GB
Terabyte (TB)	$10^{12}$ bytes	Colección impresa de la Biblioteca del Congreso (USA): 10 TB Base de datos del National Climatic Data Center (NOA): 400 TB
Petabyte (PB)	$10^{15}$ bytes	Todas las bibliotecas de investigación académica (USA): 2 PB 3 años de datos de EOS (2001) 1 PB
Exabyte (EB)	$10^{18}$ bytes	Medio millón de Bibliotecas del Congreso (USA) (impresos): 5EB

- Los datos no estructurados o con estructuras simples representan un volumen varios órdenes de magnitud mayor que los datos más estructurados (Oracle estima que la relación es de 90 a 10 por ciento)
- Los analistas informan que hasta un 85% de las bases de datos de clientes son texto
- Se estima que el 90% de los datos de texto de las empresas nunca se utilizan con fines de Inteligencia de Negocios
- La mayor parte de los datos se encuentran en un formato digital
- Es necesario automatizar las tareas de procesamiento de datos, especialmente las búsquedas y la extracción de información
- Parte de la automatización consiste en:
  - Aumentar la eficacia y eficiencia de los métodos automáticos
  - Aumentar el número de análisis secundarios de datos (data mining y text mining)

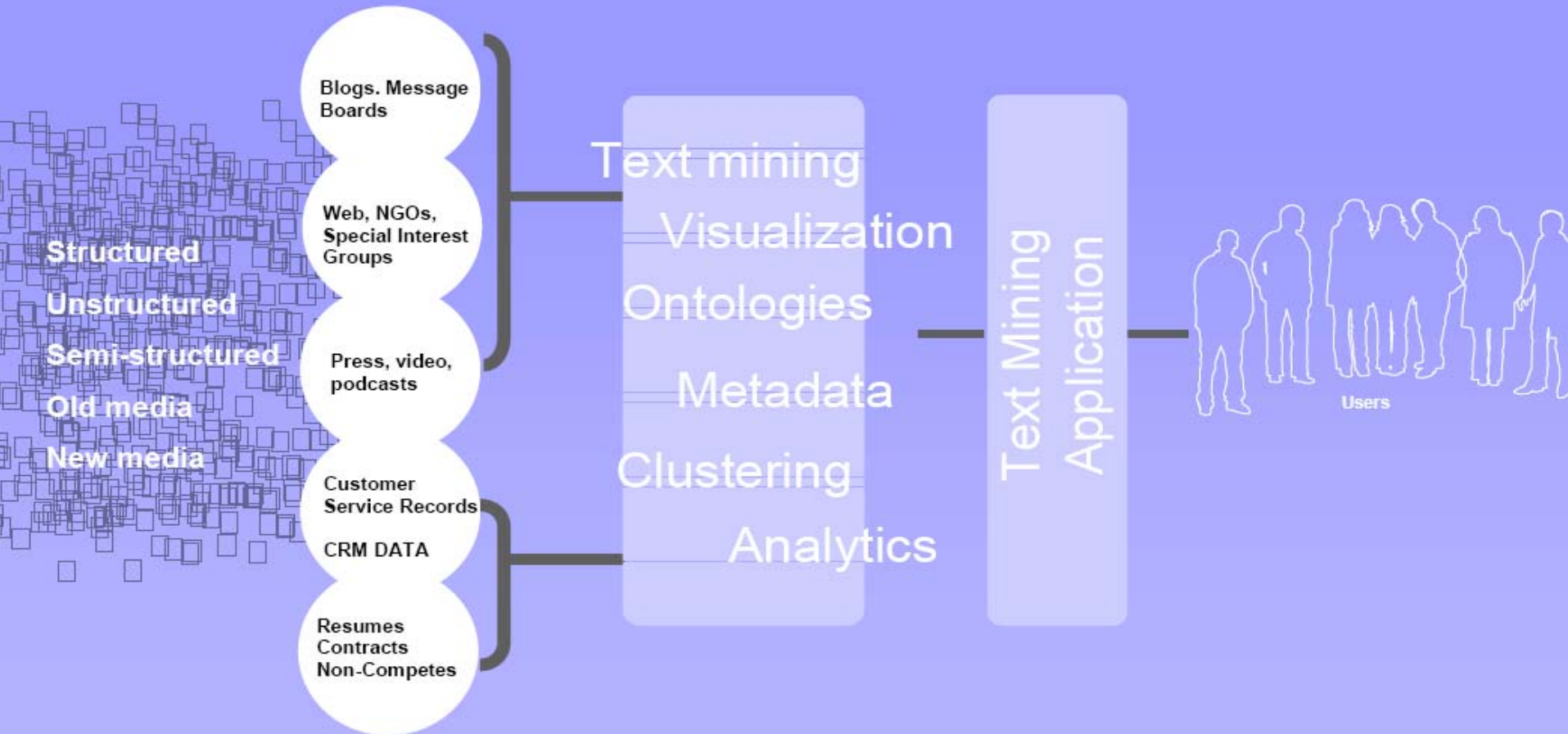
# Ejemplos de clases de textos típicos en organizaciones analizables con Text Mining

---

- Registros médicos
- Reclamos de garantías
- Registros de call centers
- Memos
- Notas
- Encuestas de texto libre
- Páginas web
- Declaraciones aduaneras o impositivas
- Feedback de clientes
- Mensajes de correo electrónico
- Reclamos de seguros
- Información de patentes
- Alertas

- Text Mining acelera el ciclo de decisiones al acortar el tiempo necesario para identificar, encontrar o descubrir información relevante
- Permite asignar más tiempo a la etapa de analizar y comprender la información descubierta
- Permite extraer valor de todos los recursos de información de una empresa

# Extracting Value from ALL Information Assets



- Supuestos:
  - Un consultor usa 11 horas semanales para recolectar y organizar la información
  - El 53% de ese tiempo lo emplea en buscar y recolectar la información
  - 303 horas por año empleadas en búsquedas
  - 15 horas promedio por búsqueda exitosa por consultor
  - Al menos el 50% de las búsquedas no son exitosas
  - Las soluciones de text mining aumentan la efectividad de la búsqueda en un 20% (10% en identificación automática de oportunidades y 10% en reducción del tiempo de búsqueda)

- *Ingresos en 2005 por consultores: U\$S 398M*
  - *Casi 400 consultores*
  - *Aprox. 8.000 búsquedas exitosas*
- *Ingresos por consultor: U\$S 1M*
  - *20 búsquedas exitosas promedio por consultor*
  - *Ingreso promedio de U\$S 49.367.4 por búsqueda*

- *ROI de Text Mining*
- *Ingresos por consultores: U\$S 497.500.000*
- *Nuevos ingresos por consultor: U\$S 1.250.000*
  - *25 búsquedas exitosas promedio por consultor*
  - *12 horas promedio por búsqueda exitosa por consultor*
- *GANANCIA NETA: U\$S 99.500.000*

- Preprocesamiento
- Parsing de los textos (análisis lingüístico y representación cuantitativa)
- Transformación o reducción de dimensionalidad
- Análisis de documentos (clustering, clasificación, etc.)
- Scoring de nuevos documentos



- Análisis exploratorio y visualización
  - Ver estadística de términos, identificar documentos similares
  - Exhibir gráficamente la relación entre términos



- Clasificación automática
  - Taxonomías – p.ej., identificar abstracts de Medline mediante jerarquías
  - Agrupamientos discretos – p. ej., identificar clusters de comentarios en un call center



- Modelización predictiva
  - Solo texto – p. ej., predecir la condición de pacientes en base a comentarios de médicos
  - Texto y datos estructurados – p. ej., predecir compras futuras en base a comentarios de productos en combinación con datos demográficos

- Churn/Attrition
  - Uso de datos de call center como predictores adicionales de churn
- Quality Management
  - Análisis de reclamos de garantías
  - Categorización automática
  - Alertas tempranas sobre tendencias
  - Análisis de encuestas
- Planificación de inventarios
  - Clasificación de SKU
- Ruteo de llamadas o pedidos en call centers o servicios de soporte al usuario
- Detección de fraudes
- Filtrado de spam

- Dow Chemicals integró los 35.000 informes de Union Carbide Corporation (UCC) (un recurso sumamente valioso) a su sistema de administración de documentos usando text mining. Se estimó que de este modo Dow ahorró 3 millones de dólares, redujo el tiempo de clasificación de los documentos en un 50% y redujo los errores en 10-15%
- American Honda Motor Co. Utiliza un sistema mixto de data y text mining para generar un sistema de alerta de defectos en las líneas de productos automotores en base a los datos obtenidos de un sistema reclamos de garantías. Esto permite corregir tempranamente defectos de producción y ahorrar millones de dólares, según Honda

- Extracción de información
- Seguimiento de tópicos
- Generación de resúmenes
- Categorización
- Clustering
- Enlace de conceptos
- Visualización de información
- Q&A

# Ejemplos de aplicaciones de técnicas de text mining

	information extraction	topic tracking	summarization	categorization	clustering	concept linkage	information visualization	question answering
<b>Medical:</b>								
FAQ's	x			x		x		x
Drug design	x				x	x		
New treatment		x				x		
<b>Business:</b>								
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			
<b>Government:</b>								
Homeland security: detecting terrorist networks	x	x			x	x	x	
Law enforcement: crime detection / prevention	x	x			x	x	x	
<b>Education:</b>								
Research on a topic		x	x	x				
Citation analysis	x				x		x	
FAQ's	x			x	x			x

Fuente: Fan et al., *Tapping into the Power of Text Mining*, C. ACM, 2005

- Número muy alto de posibles “dimensiones”
  - Todos los tipos de palabras y frases posibles en un lenguaje
- A diferencia de data mining:
  - Los registros (= documentos) no son estructuralmente idénticos
  - Los registros no son estadísticamente independientes
- Relaciones complejas y sutiles entre los conceptos de los textos
- Ambigüedad y sensibilidad al contexto
  - sinónimos
  - homónimos
  - etc.

# Técnicas de text mining ofrecidas

		Company							
		Inxight	Autonomy	Clearforest	SAS	Convera	Megaputer	SPSS	IBM
Feature	information extraction	X	X	X	X	X	X	X	X
	topic tracking		X						
	summarization	X	X			X	X		X
	categorization	X	X	X	X	X	X	X	X
	concept linkage		X	X	X				
	clustering		X			X	X		X
	information visualization	X						X	
	question answering		X				X		

Fuente: Fan et al., Tapping into the Power of Text Mining, C. ACM, 2005

- IBM Intelligent Miner for Text, TAKMI
- Semio Map (Semio Corporation)
- InXight LinguistX / ThingFinder
- Teragram
- SRA NetOwl Extractor
- SmartDiscovery, VizServer (Inxight)
- IDOL Server, Retina (Autonomy)
- ClearForest Text Analysis Suite (ClearForest)
- SAS Text Miner
- Retrieval Ware (Convera)
- TextAnalyst (Megaputer)
- LexiQuest, Clementine (SPSS)

Otras herramientas: Ver [The National Centre for Text Mining](#)  
[KDnuggets](#)

- El modo más rápido de lograr ROI de text mining es integrarlo en modelos predictivos
- Algunos aspectos fundamentales de una herramienta de text mining son:
  - Sofisticación y amplitud de sus herramientas de análisis lingüístico (stemming, tagging, extracción de entidades, manejo de sinónimos, stop lists, sublenguajes, ortografía, puntuación, etc.
  - Buenas técnicas de reducción de la dimensionalidad
  - Integración con herramientas de data mining, estadística y reporting
  - Codificación en algún sistema que permita manejar múltiples lenguajes (p. ej. Unicode)
- Es fundamental la participación en un proyecto de text mining de un consultor con conocimientos lingüísticos apropiados

# ¿Qué hace MAySA en Text Mining?

---

- Diseño de proyectos de aplicación de Text Mining
- Asesoramiento y validación de proyectos
- Desarrollo de modelos predictivos mixtos (data/text mining)
- Desarrollo de modelos mixtos de text/web mining
- Entrenamiento en técnicas y software de text mining