Web mining y obtención de información para la generación de inteligencia

Miguel Ángel Esteban

(Universidad de Zaragoza)

mesteban@unizar.es

Instituto Juan Velázquez de Velasco de Investigación en Inteligencia para la Seguridad y la Defensa

Universidad Carlos III de Madrid

22 de febrero de 2007

ELPAIS.com - El relato de Bouchar desata las risas del resto de procesados por el 11-M - España - Mozilla Firefox

Archivo Editar Ver Historial Marcadores Herramientas Ayuda

http://www.elpais.com/articulo/espana/relato/Bouchar/desata/risas/resto/procesados/11-M/elpepuesp/20070219elpepunac_7/Tes

Comenzar a usar Fire... Últimas noticias

El relato de Bouchar desata las risas del resto de procesados por el 11-M

El marroquí, uno de los presuntos autores materiales de los atentados, provoca la risa de los otros procesados al asegurar que es "un analfabeto en el uso de Internet"

ELPAIS.com - Madrid - 19/02/2007

Vota ☆☆☆☆☆ Resultado ★★★½☆ 23 votos

El marroquí Abdelmajid Bouchar, presunto autor material de la colocación de las bombas en los trenes del 11-M, ha provocado las risas e hilaridad del resto de procesados por la matanza que escuchaban su declaración desde la pecera habilitada en la sala donde se enjuicia el peor atentado islamista en la historia de Europa.

› Bouchar y Ghalyoun niegan haber estado en el piso de Leganés pese a las pruebas de ADN

Sala del Tribunal
GRAFICO - El País - 16-02-2007
Descripción del interior de la sala donde se celebra el juicio del 11-M -

A la pregunta de la fiscal Olga Sánchez sobre si había accedido a Internet para obtener información sobre la *yihad* (guerra santa), Bouchar ha dicho que eso era imposible porque "era un analfabeto en el uso de Internet", una frase que ha caído como un chiste en la sala blindada desde la que siguen el juicio los otros encausados.

Las risas han vuelto a inundar la pecera de la sala

Risas entre los procesados
VIDEO - - 19-02-2007

Esperando a adsvid.prisacom.com...

http://www.elpais.com/articulo/espana/relato/Bouchar/desata/risas/resto/procesados/11-M/elpepuesp/20070219elpepunac_7/Tes

Inicio 8 Firefox base - TextAnaly... 2 Explorador de... Mineria Textos y ... ES 11:05

Minería de la información: objetivo

- Identificar y extraer información en grandes volúmenes de datos que permanece oculta usando las técnicas clásicas de recuperación de información.

Minería de información: tipos

- ***Minería de datos (data mining)*: descubrimiento de asociaciones y patrones ocultos dentro de colecciones de datos estructurados en una o varias bases de datos relacionales.
- ***Minería de textos (text mining)*: establecimiento de asociaciones entre términos y conceptos en conjuntos de datos no estructurados (ficheros de textos, bases de datos documentales, páginas web…).

Objetivos de Web mining

- Buscar e identificar información:
  - Más relevante y específica.
  - Con relaciones entre sí.
- Crear nueva información a partir de información existente.
- Personalizar la información.
- Aprender comportamientos de usuarios web.

Tipos de Web mining

- Web usage mining
- Web content mining
- Web structure mining

Web usage mining

- Uso del data mining para analizar y descubrir modelos de navegación y de uso de los datos de una sede web o una red de sedes por los usuarios.
- Se usan los datos almacenados en el log del servidor.
- Utilidad:
  - Determinar el valor de la sede para clientes determinados.
  - Desarrollar campañas de marketing.
  - …

Web structure mining

- Análisis de la estructura de una sede web → Arquitectura de la información.
- Hay dos tipos según lo que se analiza:
  - Conexiones (hiperenlaces) → estructura de la información.
  - Estructura de datos y documentos HTML o XML.
- Usa la teoría de grafos. Representaciones gráficas.

Web content mining

- Procedimiento para descubrir información útil en la web a partir del:
  - contenido de una página web
  - o de una red de páginas extraídas por un proceso de recuperación de información
- Se trabaja con todo tipo de datos (imagen, audio, video o texto) pero hay primacía por el texto → text mining.
- Tecnologías usadas:
  - NLP (Natural Language Processing).
  - IR (Information Retrieval).

Minería de textos: prestaciones

- Mejora de la recuperación de información:
  - Refina búsqueda sobre una selección de resultados.
  - Destila el significado de un texto en una forma concisa.
  - Muestra resúmenes apropiados.
- Establece agrupaciones de textos
  - Por frecuencia y relevancia de términos y conceptos.
  - Muestra los conceptos que relacionan los textos.
- Visualiza los resultados de un modo que facilita
  - Interpretación.
  - Navegación eficiente.

Áreas de aplicación de Web Mining

- Motores de búsqueda.
- Comercio electrónico.
- Diseño web.
- Posicionamiento web.
- Seguridad

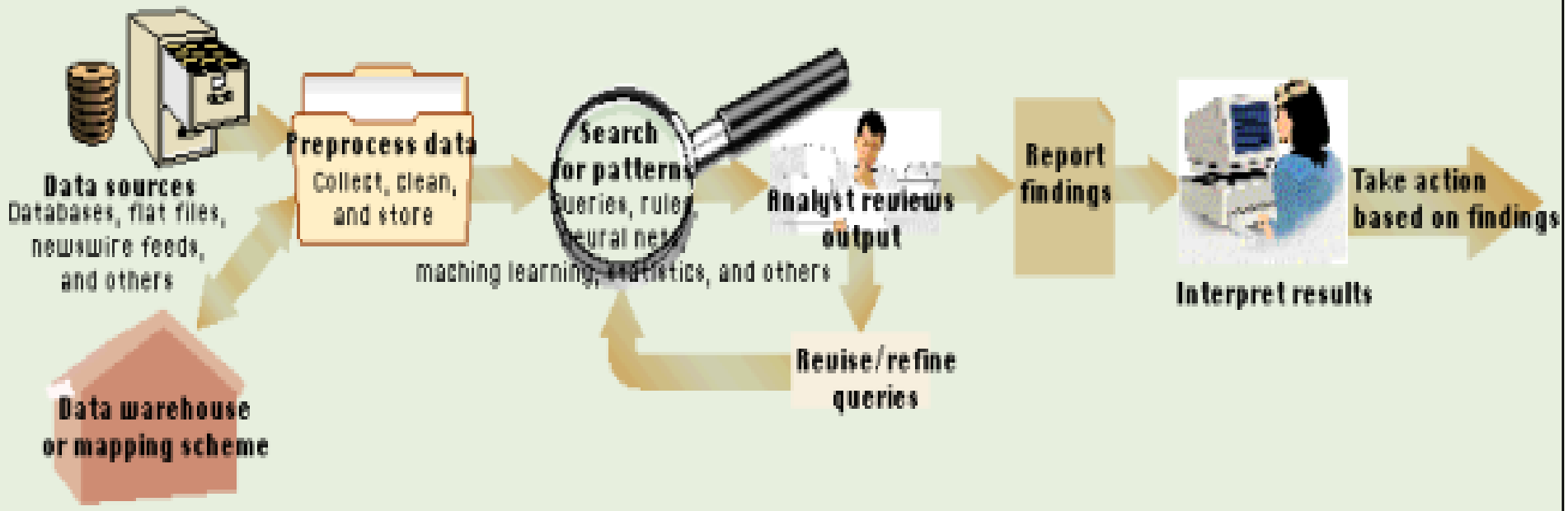
Aplicaciones de minería de textos en inteligencia para seguridad

- Prospectiva en cualquier ámbito.
  - Predicción de agentes para guerra biológica.
  - Ejército Popular de China analiza estrategia US Army.
- Seguimiento del entorno.
  - De la evolución de una situación.
  - De actividades de empresas.
  - Análisis de patentes (ej.: terrorismo tecnológico).
- Lucha contra el crimen organizado.
  - Blanqueo de dinero.
  - Detección de redes criminales: análisis de correos electrónicos y mensajes en comunidades virtuales.
  - Análisis de expedientes policiales para detectar patrones.
- Ejemplo: TIA Programa (*Total Information Awareness*).

El proceso de minería de textos

1. Adquisición de textos.
2. Normalización de los textos.
   - Usualmente en formato basado en XML.
   - Extracción de metadatos identificativos: autor, título, fecha, fuente.
3. Filtrado: identificación de textos relevantes mediante un análisis de presencia de palabras predeterminadas.
4. Condensación: extracción de términos relevantes y categorización.
5. Análisis: establecimiento de relaciones entre textos con base en los términos y categorías.
6. Visualización: uso de gráficos y diagramas.

Data-Mining Process



Minería de la información: técnica

- Uso de algoritmos de identificación y agrupación de datos relevantes:
  - Fijación de segmentos o cadenas de caracteres.
  - Representación mediante el modelo vectorial para el cálculo de similitudes.
  - Categorización y agrupación mediante análisis cluster

Las cuatro fases del Web mining

- Identificación del problema.
- Colección de datos → búsqueda.
- Preprocesamiento de los datos.
- Descubrimiento de patrones. Análisis.

Colección de datos

- Motores de búsquedas.
- Servicios de información.
- Mensajería → Lugares de encuentro.
  - Listas de correos.
  - Chats. Foros.
  - Blogs.
  - Comunidades virtuales.

Análisis de datos: soluciones

- Text Analyser de Megaputer
- SAS.
- Synthema.
- Insight Discoverer de TEMIS.

Muchas gracias por su atención

A su disposición en

mesteban@unizar.es