tecnimap Murcia 2004

29

POSIBLE APLICACIÓN DE LA MINERÍA DE TEXTOS A LOS TRABAJOS DE LA COMISIÓN MINISTERIAL DE INFORMÁTICA

M.ª del Pilar Cantero Blanco
Jefa de Servicio de Sistemas Informáticos. Subdirección General de Planificación y Coordinación Informática
Ministerio de Trabajo y Asuntos Sociales

1. INTRODUCCIÓN

La Minería de Textos es una tecnología emergente cuyo objeto es la búsqueda de conocimiento en grandes colecciones de documentos no estructurados.

Aproximadamente un 80% de la información de las organizaciones está almacenada en forma textual no estructurada: informes, e-mail, actas de reuniones, legislación de interés, etc. La minería de textos opera sobre bases de datos textuales no estructuradas con el objetivo de detectar patrones no triviales e incluso información sobre el conocimiento almacenado en las mismas.

Otro aspecto en el que las tecnologías de minería de textos encuentran un prometedor área de aplicación es el de la web semántica. Este nuevo modelo de Internet pretende construir toda una estructura de metadatos, información sobre la estructura y significado de los datos almacenados, e incluirlos en los documentos de forma que sean navegables, identificables y "comprensibles" por las máquinas.

Los sistemas de text mining permiten el análisis léxico de los textos y especialmente la construcción automática de estructuras de clasificación y categorización que se codifican en forma de tesauros. No en vano tesauro proviene del griego thesaurós: tesoro. La importancia del tesauro consiste en que cada uno de sus términos, se utiliza para denotar un concepto, la unidad semántica básica, que permite expresar una idea.

Los sistemas de minería de textos pueden ayudar en la categorización de la información existente en una organización, en el filtrado de información, por ejemplo de e-mail, en la detección de información similar o relacionada con otra existente o para eliminar información duplicada.

Algunas aplicaciones han visto ya la luz. Por ejemplo, algunas empresas utilizan ya sistemas de minería de textos para identificar el contenido de los e-mails que les envían sus clientes y redirigirlos a los departamentos apropiados. En otros casos, si el sistema es capaz de identificar el contenido de una consulta frecuente en un e-mail, envía una respuesta estándar a la consulta, sin necesidad de intervención humana.

Donde quizá lleva más tiempo utilizándose esta tecnología es en el campo de la vigilancia tecnológica e inteligencia competitiva para, buceando en las bases de datos textuales, seguir la evolución de los productos de la competencia.

También, existen trabajos orientados hacia la aplicación de estas técnica en la investigación de mercados en la Web, mediante la recogida de estadísticas sobre la utilización de determinados conceptos y/o temas en la red, con el objetivo de estimar la demografía y las curvas de demanda de productos asociados a los mismos.

2. ANTECEDENTES

Ya en 1977, el sistema THOMAS ilustró cómo las palabras o las frases clave podían utilizarse para guiar a los usuarios en el descubrimiento de documentos de referencia útiles. Las frases clave son un tipo especialmente útil de información abreviada. Sin embargo, tales frases se eligen con frecuencia manualmente, bien por los autores o por indizadores profesionales. Condensan documentos en unas pocas palabras y frases, ofreciendo una descripción breve y precisa de los contenidos de un documento.

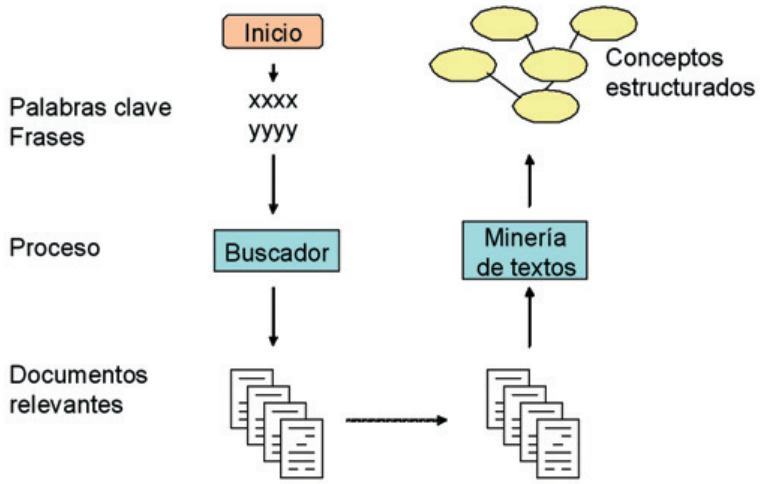
Tienen muchas aplicaciones: clasificación o agrupación de documentos, interfaces de búsqueda, motores de búsqueda y construcción de tesauros.

La asignación manual de frases clave es tediosa y lleva tiempo, requiere experiencia y puede dar resultados no coherentes, de modo que los métodos automáticos benefician tanto a los que generan y mantienen grandes colecciones de documentos como a sus usuarios. En consecuencia, se han propuesto varias técnicas automáticas.

Un amplio conjunto de técnicas se ha aplicado al problema de la extracción de frases. Así, se aplican las técnicas de aprendizaje automático a la extracción automatizada de frases clave. Las frases clave de ejemplo, generalmente las dan los autores, y una vez que se aprende un modelo para identificar frases clave a partir de documentos de prácticas, se puede utilizar para extraer frases clave de otros documentos. De este modo, las frases clave extraídas automáticamente de los textos de los documentos, pueden usarse para establecer enlaces a documentos similares y para sugerir frases de búsqueda adecuadas para los usuarios.

Esta técnica, esencial para el acceso al conocimiento y el procesado de las búsquedas, promete incrementar la riqueza y amplitud del material accesible, a la vez que se mejora la precisión y exhaustividad de la búsqueda.

Procesos de búsqueda y minería de textos



3. DESCRIPCIÓN GENERAL DE LAS TECNOLOGÍAS DE MINERIA DE TEXTOS

La minería de textos no se debe confundir con los motores de búsqueda de Internet o con capacidades avanzadas de sistemas de gestión de bases de datos. De modo análogo a la minería de los datos, que extrae información útil a partir de grandes volúmenes de datos, la minería de textos es un procedimiento aplicado a los volúmenes grandes de texto libre no estructurado. Después de que se haya realizado una búsqueda tradicional sobre documentos, recuperando por ejemplo texto completo, resúmenes, o los términos puestos en un índice, la minería de textos va más allá, explorando las relaciones complejas entre documentos.

Dado que el objetivo último de la gestión de datos y textos corporativos es ofrecer información de calidad a la dirección, cuanto más eficiente sea este proceso de Minería, mayor será en cantidad y en calidad la información disponible para soportar la toma de decisiones.

Hay tres componentes importantes de la minería de textos:

1. **Recuperación,** el paso fundamental de la minería de textos. Es la extracción de los textos relevantes para la transformación posterior.
2. **Tratamiento de la información,** la extracción de patrones de los datos recuperados obtenidos en el paso anterior. Este paso consiste en ordenar, clasificar y cuantificar el material anterior no estructurado.
3. **Integración de la Información.** Es la combinación de la salida en el ordenador del tratamiento de la información con los procesos cognoscitivos humanos.

4. APLICACIONES DE LA MINERÍA DE TEXTO

La información obtenida tras el proceso completo de minería de textos se puede utilizar para:

- Mejorar la comprensión y la importancia de la información recuperada a partir de bases de datos. La mayoría de los métodos más fiables utilizan un método iterativo par lograr una mayor comprensión de la información y precisión en la recuperación de textos.
- Identificar los elementos que intervienen en una disciplina técnica. Estos elementos pueden ser los autores, las organizaciones y las instalaciones que contribuyen al mantenimiento de dicha disciplina.
- Identificar temas técnicos, sus correlaciones, sus relaciones con la infraestructura. Se pueden categorizar frases y descubrir las relaciones o las interacciones que no serían encontradas cuando se leen por separado.
- Extrapolación de ideas a diversas disciplinas relacionadas.
- Utilización de técnicas que dan como resultado el avance en el campo de las tecnologías. Se pueden utilizar unos indicadores que recogen la información sobre el estado del ciclo vital de la tecnología. Estos indicadores, denominados de innovación, se pueden generar como ayudas para demostrar el nivel de la madurez de la tecnología.
- La inteligencia tecnológica competitiva (ITC) es otro uso de las técnicas que se utilizan en la minería de textos. Se desarrolló ampliamente en los años 90 debido a que las compañías, las universidades y las agencias estatales tenían la necesidad de saber qué capacidades tenían otras organizaciones para desarrollar una tecnología particular. El análisis de ITC se puede realizar para obtener un análisis de mercado. Mediante resultados gráficos y tablas se puede demostrar qué compañía tiene más fuerza en un campo tecnológico.

5. LA RELACIÓN ENTRE LA MINERÍA DE TEXTOS Y LA GESTIÓN DEL CONOCIMIENTO

El creciente volumen de información disponible plantea nuevos problemas y retos para la recuperación de la información. Los motores de búsqueda pueden desempeñar un papel esencial

en la viabilidad de los sistemas de información basados en Internet, siempre que existan aplicaciones que puedan analizar y evaluar la relevancia de la información para el usuario. Nuevos enfoques, basados en la integración de la minería de textos con la gestión del conocimiento, pueden ofrecer mejores soluciones a la gestión de la información.

Los usuarios individuales y las organizaciones públicas que utilizan la recuperación cooperativa de la información, se encuentran inmersos en complejos procesos de búsqueda de la información y de puesta al día del conocimiento. Por ello, la minería de textos es un área de creciente interés en el campo del descubrimiento del conocimiento.

Un problema creciente al que se enfrentan las grandes empresas e instituciones públicas es el descubrimiento de nuevo conocimiento y su gestión. Los avances recientes en este campo incluyen la aplicación de técnicas de minería de textos para encontrar conocimiento significativo a partir de datos textuales sin estructurar. Se aplican las técnicas de tratamiento del lenguaje natural (Natural Language Processing - NLP) para extraer información útil a partir de una amplia colección de textos de documentos almacenados. Se extraen de los documentos los términos duplicados y las entidades de mayor nivel y se utilizan como palabras clave. La minería de textos se centra en encontrar reglas de asociación útiles y significativas para los términos o palabras duplicados.

Una de las áreas principales de aplicación de la minería de textos es la recogida y condensación de hechos como una base de ayuda a la toma de decisiones.

6. ANÁLISIS DE LA APLICABILIDAD DE LA MINERÍA DE TEXTOS A LOS TRABAJOS DE LA COMISIÓN MINISTERIAL DE INFORMÁTICA

En la Subdirección General de Planificación y Coordinación Informática del MTAS, se analizan datos sobre la tramitación de expedientes. Así, frente a la tradicional actividad del "intermediario de información" se podrían aplicar las principales ventajas de la tecnología de minería de textos a la búsqueda de información para el tratamiento de los expedientes que se gestionan en la Comisión Ministerial de Informática. Dichas ventajas se podrían resumir en los siguientes puntos:

- La capacidad de procesar rápidamente grandes cantidades de datos textuales, lo que no puede ser llevado a cabo eficazmente por lectores humanos.
- La "objetividad" y capacidad de personalización del proceso.
- La posibilidad de automatizar las laboriosas tareas de rutina, dejando las tareas más exigentes para los lectores humanos.

Teniendo en cuenta estas ventajas, las aplicaciones de la minería de textos se podrían aplicar a los expedientes de carácter informático, fundamentalmente, para:

- Extraer información relevante de un documento (resumiendo, extrayendo lo más notable, etc.).
- Adquirir perspectivas sobre las tendencias, las relaciones entre equipos/proyectos/organizaciones, etc. agregando y comparando automáticamente la información extraída de documentos de un cierto tipo.
- Clasificar y organizar documentos según su contenido; es decir, preseleccionar automáticamente grupos de documentos con un tema específico y asignarlos a la persona adecuada.

- Organizar depósitos de meta-información relacionada con documentos para la búsqueda y recuperación.
- Recuperar documentos basándose en varios tipos de información sobre el contenido del documento.

Esta lista de actividades muestra que las principales áreas de aplicación de las tecnologías de minería de textos abarcarían dos aspectos: (1) el descubrimiento de conocimiento y (2) la extracción de información.

La técnica de minería de textos, se podría aplicar en la Comisión Ministerial de Informática para la extracción de información relevante y el descubrimiento de conocimiento en los siguientes ámbitos:

1. Un sistema de extracción de información busca información específica en un documento, según normas predefinidas. Las normas son específicas de un área temática dada. Por ejemplo, si el área temática son las noticias sobre últimas tendencias tecnológicas, las normas pueden especificar que el sistema de extracción de información deba identificar:
   - las empresas implicadas
   - posicionamiento de dichas empresas en el mercado
   - los servicios que ofrecen dichas empresas
   - tipos de productos más relevantes

   y la restante información de este tipo, que puede encontrarse en un documento típico del área temática.
2. La técnica de minería de textos, se podría aplicar en la Comisión Ministerial de Informática para la extracción de información y conocimiento que pueda desprenderse de los documentos de las memorias justificativas de las distintas Unidades Organizativas. Así, resultaría de interés, saber qué servicios de carácter informático para el rediseño de la web y bajo qué características, se han gestionado en otras Unidades pertenecientes al MTAS.
3. En el área de la gestión de proyectos informáticos podría ayudar a encontrar las similitudes y diferencias planteadas en la solución. Sin nos centramos en la adquisición, en distintas Unidades, de determinados productos, se podría conocer qué funcionalidades se pueden obtener con dichos productos, lo cual resulta imprescindible para realizar un análisis detallado que clarifique su funcionamiento, en conjunto, y evitar, así, posibles repeticiones.
4. Con respecto a la instalación de diversas redes informáticas, mediante la utilización de la técnica de minería de textos, se podría obtener resultados beneficiosos y rápidos aportando nuevos datos que mejoraran las deficiencias que integran los sistemas antiguos.

Un sistema de extracción de información busca información específica en un documento, según normas predefinidas (específicas del tema). Así, tales sistemas se construyen, por lo general, manualmente para una sola área temática, lo que requiere una gran cantidad de trabajo por parte de expertos.

7. CONCLUSIONES

La minería de textos puede utilizarse como una herramienta eficaz de gestión del conocimiento que apoya la extracción de información relevante a partir de grandes cantidades de datos textuales no estructurados.

El desarrollo de aplicaciones de gestión del conocimiento viene apoyado por un conjunto de tecnologías que ya están maduras. La rápida difusión de las tecnologías de redes y telecomunicaciones contribuye a facilitar el acceso a las fuentes de información. El aumento de la potencia de los ordenadores y la disponibilidad de software más inteligente de gestión de bases de datos permite el procesado rápido, y la adaptación de técnicas de inteligencia artificial a problemas más estructurados que proporciona la lógica necesaria a los sistemas.

Los sistemas de gestión del conocimiento no pueden, evidentemente, sustituir a los seres humanos en las tareas de análisis de la información, pero pueden brindar una ayuda importante a la hora de reducir algunas de las actividades de recogida y tratamiento de la información que consumen mucho tiempo, y de ese modo permitir a los usuarios que tomen decisiones más informadas.

Desde el punto de vista del gestor público, los avances en la gestión del conocimiento son bienvenidos y pueden utilizarse para ayudar a procesar grandes cantidades de información. Vistas como una extensión natural del campo general de la tecnología de la información y la comunicación, tales aplicaciones contribuyen – casi por definición – a la construcción de una sociedad basada en el conocimiento.

También, heredan los principales problemas relevantes del campo de las TIC, tales como las normas, los derechos de autor y la seguridad. Por otro lado, no se deberían despreciar las cuestiones de normalización en las tecnologías relacionadas con las TIC, incluyendo cómo explotan las organizaciones la información y el conocimiento adquiridos, y qué mecanismos de seguridad existen para los individuos cuyos datos personales han sido procesados y almacenados en un sistema con tales capacidades

La minería de textos, aún en sus albores, es un ataque desde otro ángulo al problema común: encontrar la información relevante y sobrevivir a la infoxicación vigente. Aún, es difícil saber por donde vendrá la solución definitiva. Probablemente, será producto de un tratamiento multidisciplinar.