UNIDEP®

Universidad del Desarrollo Profesional

unidep.mx
01800 7UNIDEP
(864-337)

Perspectivas para la integración de la minería de textos y la gestión del conocimiento

Luciana Bordoni y ***Ernesto d'Avanzo***, ENEA

Resumen

Asunto: El creciente volumen de información disponible en la web 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.

Relevancia: Los usuarios individuales y las organizaciones con responsabilidad política que utilizan la recuperación cooperativa de la información, se encuentran inmersos en el proceso de búsqueda de la información y de puesta al día del conocimiento. Por lo tanto, la gestión del conocimiento es una parte importante del buen funcionamiento de cualquier organización política.

Texto

El reto para la gestión del conocimiento (KM) es traducir el conocimiento "tácito", que es personal, difícil de formular, "en las mentes de las personas", y difícil de comunicar, en un conocimiento "explícito" que es formal, sistemático, y que puede compartirse.

Introducción

La gestión del conocimiento (KM) es una práctica empresarial relativamente nueva en la que el contenido digital en muchas formas y formatos se reúne en una arquitectura integrada que permite utilizar los datos semánticos subyacentes en el corpus como una ayuda a la comprensión estratégica y a la toma de decisiones.

La KM está destinada a servir a las prácticas empresariales, habiéndose originado en el mundo de la empresa como un método para unificar las enormes cantidades de información generada en reuniones, propuestas, presentaciones, documentos analíticos, material de enseñanza, etc. La KM la utilizan fundamentalmente las grandes organizaciones, aunque el problema de navegar en un corpus de documentos multiformatos es relevante para cualquier individuo o grupo que cree o consuma conocimiento distribuido. Ikujiro Nonaka en su artículo, 'La empresa creadora de conocimiento' sostiene que 'poner al alcance de los demás el conocimiento personal es la actividad central de la empresa creadora de conocimiento' (Nonaka, 1991).

El reto para la gestión del conocimiento (KM) es traducir el conocimiento 'tácito', que es personal, difícil de formular, 'en las mentes de las personas', y difícil de comunicar, en un conocimiento 'explícito' que es formal, sistemático, y que puede compartirse. El conocimiento se crea convirtiendo el conocimiento 'tácito' en 'explícito'. A través de la creación de conocimiento, una empresa puede traducir sus ideas en tecnologías y productos innovadores. En los últimos años, la industria, el mundo académico y los gobiernos han prestado una creciente atención a la KM.

La KM efectiva se cita con frecuencia como una capacidad clave para adquirir una ventaja competitiva en la empresa global, y la tecnología del lenguaje humano juega un papel central en la KM; mejora el funcionamiento de la organización compartiendo el conocimiento, y mediante el aprendizaje y la aplicación de la experiencia. También son importantes los avances en las técnicas de inteligencia artificial y basadas en el conocimiento para almacenar normas y modelos, así como los sistemas de información que almacenan y organizan el conocimiento (Stewart et al., 2000). Como indicación de la importancia de la KM, muchas corporaciones que tradicionalmente medían sólo los aspectos financieros del valor están empezando también a medir los valores humanos e intelectuales.

Un conjunto de tecnologías del lenguaje humano puede facilitar la KM, incluyendo la mejora de la recuperación de información, la extracción de información, el resumen, presentación y generación de documentos. Además, las tecnologías del lenguaje humano prometen mejorar el acceso humano a la información y la interacción humana. La KM puede mejorar la eficiencia de las organizaciones mediante la integración de aspectos tecnológicos con otros humanos y organizativos.

Un conjunto de tecnologías del lenguaje humano puede facilitar la KM, incluyendo la mejora de la recuperación, la extracción de información, el resumen, presentación y generación de documentos.

Este artículo describe las perspectivas de la integración de la minería de textos y las técnicas de descubrimiento de conocimiento utilizando inteligencia artificial en la interfaz con los recursos lingüísticos.

Las aplicaciones de la minería de textos se utilizan principalmente para:

- Extraer información relevante de un documento
- Agregar y comparar información automáticamente
- Clasificar y organizar documentos según su contenido
- Organizar depósitos para búsqueda y recuperación
- Clasificar textos e indizarlos en la web

Las metodologías empleadas en los campos de la inteligencia artificial y la lingüística computacional pueden mejorar las tecnologías de la minería de textos y en consecuencia la KM. En particular, nuestro enfoque considera el papel fundamental desempeñado por las frases clave y las técnicas de indización conceptual en el campo de la minería de textos.

La minería de textos frente a la extracción de información

La minería de textos es un área de creciente de interés en el campo de la KM y en particular de la minería de datos y el descubrimiento de 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 datos para encontrar conocimiento significativo a partir de datos textuales sin estructurar (Feldman et al., 1999). Se aplican las técnicas de tratamiento del lenguaje natural (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 sus palabras clave. Esta metodología también puede aplicarse a los documentos en la Web, en lo que se viene a llamar minería de textos en la Web. La minería de textos se centra en encontrar reglas de asociación útiles y significativas para los términos o palabras duplicados.

La minería de textos se centra en encontrar normas de asociación útiles y significativas para los términos o palabras duplicados y se trata de un área de creciente interés en el campo de la KM y en particular de la minería de datos y el descubrimiento de conocimiento.

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. Las principales ventajas de la tecnología de minería frente a la tradicional actividad del 'intermediario de información' son:

- 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.

Tomando ventaja de estas propiedades, las aplicaciones de la minería de textos se usan fundamentalmente para:

- Extraer información relevante de un documento (resumiendo, extrayendo lo más notable, etc.).
- Adquirir perspectivas sobre las tendencias, las relaciones entre gentes/lugares/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.

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

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 ataques terroristas, las normas pueden especificar que el sistema de extracción de información debería identificar (i) la organización terrorista que participa en el ataque, (ii) las víctimas del ataque, (iii) el tipo de ataque, y la restante información de este tipo que puede esperarse en un documento típico del área temática.

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

La mayoría de los sistemas de extracción de información se construyen manualmente para una sola área temática, lo que requiere una gran cantidad de trabajo de expertos. Por ejemplo, el mejor rendimiento en la 5ª Conferencia sobre Comprensión de Mensajes (MUC-5,1993) se obtuvo con un coste de dos años de intenso esfuerzo de programación.

Analogía entre documentos utilizando la extracción de frases clave.

Ya en 1977, el sistema THOMAS (Oddy, 1977) 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. 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 y de hojeo, motores de búsqueda y construcción de tesauros. Las frases clave se eligen con frecuencia manualmente, casi siempre por los autores de un documento pero a veces por indizadores profesionales. 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 reúnen como a los usuarios de grandes colecciones de documentos. En consecuencia, se han propuesto varias técnicas automáticas.

Se sabe, desde hace tiempo, que 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.

Un amplio conjunto de técnicas se ha aplicado al problema de la extracción de frases. Turney fue el primero en tratar la extracción como un problema de aprendizaje bajo supervisión (http://extractor.iit.ncr.ca/). El Proyecto de Biblioteca Digital de Nueva Zelanda (NZDL) (http://www.nzdl.org) ha desarrollado el sistema Kea (Frank et al., 1999) que aplica las técnicas de aprendizaje automático a la extracción automatizada de frases clave. Kea utiliza un modelo para identificar las frases de un documento que muy probablemente serán buenas 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.

Los sistemas de extracción automática de frases clave prometen incrementar la riqueza y amplitud del material accesible a la vez que se mejora la precisión y exhaustividad de la búsqueda.

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

Conclusión

Mientras que la KM es un fenómeno reciente que pretende solucionar problemas de organización y epistémicos1, la investigación en minería de textos ya ha experimentado con muchos objetivos y necesidades de KM. La minería de textos puede utilizarse como una herramienta para ayudar en la creación de conocimiento y en la extracción de información relevante. Por definición, buena parte de la KM lleva consigo la necesidad de herramientas de búsqueda eficaces e inteligentes, y cuando el depósito de la organización es grande, la contribución de la KM como herramienta de minería de textos puede ser fundamental. Los beneficios que se pueden obtener al integrar la KM con la tecnología de minería de datos parecen valiosos. Esto puede conducir a métodos que permitan a los investigadores satisfacer sus necesidades de información y conocimiento.

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 al acceso a las fuentes de información.

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 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 político, los avances en la gestión del conocimiento son bienvenidos y pueden utilizarse para ayudar a los políticos 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 políticamente 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.

Palabras clave

Extracción de información, minería de textos, gestión del conocimiento y de los contenidos

Referencias

R. Feldman, Y. Aumann, M. Fresko, O. Liphstat, B. Rosenfeld, Y. Schler, Text Mining via Information Extraction, Proceedings of PKDD¿99, 1999, págs. 165-173.

E. Frank, G. Paynter, I. Witten, C. Gutwin, y C. Nevill-Manning, Domain-specific keyphrase extraction, Proceedings of the sixteenth international joint conference on artificial intelligence, San Mateo, CA:Morgan Kaufmann, 1999.

B. Hjorland, Information seeking a subject representation. An activity-theoretical approach to information science, Westport, CT: Greenwod Press.

S. Jones, G.W. Paynter, Automatic Extraction of documents keyphrases for use in digital libraries: evaluation and applications, Journal of the American Society for Information Science and Technology, 53 (8), 2002.

S. Jones, M.S. Staveley, Phrasier: a system for interactive document retrieval using keyphrases, Proceedings of the annual international conference on research and development in information retrieval, agosto 1999.

I. Nonaka, The Knowledge Creating Company, Harvard Business Review, noviembre-diciembre, 1991.

R.N. Oddy, Information retrieval through man-machine dialogue, Journal of Documentation, 33, 1, 1977.

D.E. O¿Leary, Using AI in Knowledge Management: Knowledge Bases and Ontologies, IEEE Intelligent Systems, mayo/junio 1998.

D.E. O¿Leary, R. Studer, Knowledge Management: An Interdisciplinary Approach, IEEE Intelligent Systems, 2001.

S. Staab, Human Language Technologies for Knowledge Management, IEEE Intelligent Systems, noviembre/diciembre 2001.

K.A. Stewart, R. Baskerville, V.C. Storey, J.A. Senn, A.Raven, C. Long, Confronting the assumptions underlying the management of knowledge: an agenda for understanding and investigating knowledge management, The DATA BASE for Advances in Information Systems, otoño 2000, 31, 4.

M. Sumner, Knowledge Management: Theory and Practice, SIGCPR 1999, Nueva Orleans, EE.UU.

Recuperado de: http://libros-revistas-derecho.vlex.es/vid/integracion-mineria-textos-conocimiento-172096 10 de julio d 2011.