Técnicas inteligentes de recuperación y análisis de la información

JULIÁN DORADO DE LA CALLE
Departamento de Tecnologías de la Información y las Comunicaciones
Facultad de Informática – Universidade da Coruña

1. Introducción

La informática es el "Conjunto de conocimientos científicos y técnicas que hacen posible el tratamiento automático de la información por medio de ordenadores" según la definición del Diccionario de la Lengua Española (Real Academia Española, 2001). El concepto de información, en este nivel, se menciona en su sentido más amplio, como cualquier elemento que es susceptible de generarse o procesarse en Informática, obviando que existen varios niveles de organización dentro de este concepto amplio de información.

Normalmente los niveles de organización de la información se suelen representar en forma de pirámide (ver Figura 1) con cuatro niveles. De abajo hacia arriba los cuatro niveles son: datos, información (también llamada noticias para evitar utilizar la palabra información como explicación y como concepto general), conocimiento y sabiduría.

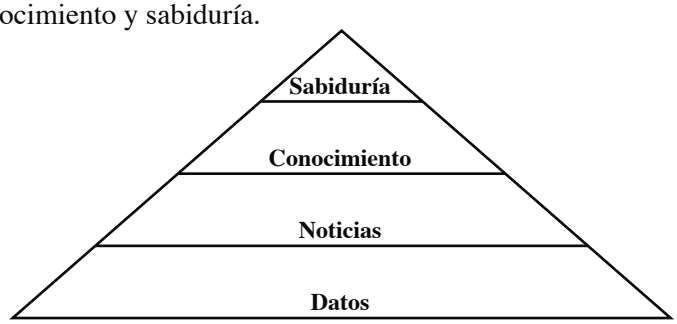

Figura 1: Pirámide de la Información

Cuando hablamos de datos, nos referimos a "tokens" o elementos individuales de información que bien pueden ser números o palabras, por ejemplo, el número 38. Dentro del nivel de información o noticias, la información toma la forma de datos unidos a un contexto que les provee de significado, siguiendo el mismo ejemplo, si decimos que 38 es una temperatura y que está medida en grados centígrados, la unión del número, la palabra y las unidades hacen progresar el dato a una noticia. El nivel de conocimiento pone en relación varias noticias para definir una regla, explicar una circunstancia o cualquier otra construcción que reconozcamos como un conocimiento. En el ejemplo anterior si decimos que una temperatura de 38ºC o más supone que una persona tiene fiebre, estamos reflejando un conocimiento médico, aunque sea un conocimiento muy básico. El último nivel, el nivel de sabiduría, resumiría y matizaría conocimientos mediante la experiencia de forma que se hace más ágil la información. Dentro del mismo ejemplo, un médico no necesita utilizar reglas como la comentada anteriormente sobre la fiebre para llegar al diagnóstico de una gripe al utilizar su experiencia previa y un contexto de apariencia general de paciente, gran mayoría de pacientes, con gripe y estación de invierno. Este último tipo de información es difícil de verbalizar y tiene que ser construida internamente por cada individuo a partir de conocimientos previamente aprendidos.

En esta estructura de la información los documentos escritos, ya sean libros, artículos o páginas web encajarían dentro del nivel de conocimiento, siendo este más o menos elaborado y estando más o menos mezclados noticias y datos.

2. Informática y el tratamiento de los distintos niveles de la información

A día de hoy, acabando el año 2010, en informática existe una amplia experiencia en la gestión de datos. Desde los primeros días de la informática se ha trabajado en el almacenamiento de datos, ya fuera en tarjetas perforadas, en los comienzos, o en los modernos discos duros de estado sólido (SSD) que se han empezado a generalizar este mismo año.

La gestión de noticias, como nivel de información, también es un tema muy trabajado, las bases de datos (BBDD) (Galindo, J. et al. 2006), desde las jerárquicas y en red, hasta las más comunes actualmente, las relacionales, permiten almacenar datos en atributos relacionados mediante registros. Es la forma más común de almacenar y gestionar información aunque ya están llegando con fuerza nuevas aproximaciones como las BBDD orientadas a objetos o los formatos de intercambio basados en lenguajes de marcado (Coombs, J. H. et al. 1995)

como XML (Goldfarb, C. F. et al. 2000), más cercanos a la realidad de internet y de la web.

Hasta el nivel de noticias, como se ha comentado, la informática actual es capaz de gestionar la información de manera eficaz y eficiente. Es a partir del nivel de conocimientos donde nos acercamos más a los sistemas de investigación que a los sistemas de uso común. Para tratar el conocimiento existen herramientas como los sistemas basados en reglas o los sistemas expertos (SSEE), pero que ya no son de uso general (no se pueden aplicar a cualquier tipo de conocimiento) y no son de desarrollo fácil (cualquier persona con estudios de informática es capaz de desarrollar una BD relacional pero no es seguro que sea capaz de construir un sistema inteligente que gestione conocimientos).

El último nivel, el de sabiduría, entra directamente dentro del ámbito de la investigación en Inteligencia Artificial (IA) y los avances, parciales y muy específicos, no son aprovechables en este momento para su uso en cualquier tipo de problema.

3. Tipos de Inteligencia Artificial

No es un tema trivial la definición de conceptos complejos, como es el caso de la inteligencia. Es muy común en este caso utilizar, en vez de una definición única, una definición en función de las características que la definen, como la capacidad de razonamiento, de comunicación, de interacción social, de resolución de problemas, etc.

Es, por tanto, también difícil poder emular o reproducir comportamientos inteligentes de forma computacional pero, como ya se ha comentado, cualquier sistema que intente trabajar con conocimientos o sabiduría entra dentro del campo de la informática denominado IA (Luger, G. et al. 2004). Los sistemas desarrollados dentro de este campo se pueden agrupar fundamentalmente en simbólicos, como los SSEE, y los sistemas conexionistas, como las redes de neuronas artificiales (RNA).

La inteligencia artificial simbólica se basa en emular el conocimiento de uno o varios expertos humanos en un dominio concreto. Trata de extraer el conocimiento de los expertos en un proceso también conocido como elicitación y explicitarlo en forma de reglas para poder después razonar automáticamente con ellas. Tanto el proceso de elicitación como de depuración del conocimiento extraído es

complejo y depende de los expertos escogidos y de su capacidad de hacer explícito el conocimiento o sabiduría que atesoran.

Los sistemas simbólicos más comúnmente usados son los SSEE (Fernandez Gil, A. 2010), famosos por los desarrollos hechos en diagnóstico médico aunque sin el éxito en su aplicación y generalización esperado inicialmente. Realmente, en cualquier labor en la que se desarrolle conocimiento experto se puede aplicar un SE. Un SE se compone de tres partes fundamentales, una base de conocimientos que almacena el conocimiento, normalmente en forma de reglas, un motor de inferencia, que utiliza las reglas para razonar, y una base de hechos que almacena los datos que se le suministran al sistema y las conclusiones que se van alcanzando a partir del razonamiento con las reglas.

En el ámbito de la biblioteconomía y la documentación la tarea de indización y catalogación es, normalmente, desarrollada por expertos que podían ser asistidos o monitorizados por SSEE desarrollados a partir del conocimiento de expertos reconocidos en el área.

La inteligencia artificial conexionista se basa en simular la estructura del cerebro o de los sistemas nerviosos utilizando pequeños procesadores a modo de neuronas e interconectándolos para que trabajen de forma coordinada. Estos sistemas aprenden a partir de ejemplos no necesitando expertos que dirijan su desarrollo. Su aprendizaje es similar a cómo una persona aprende a solucionar un problema estudiando casos de ejemplo e intentando después solucionar problemas similares no tratados anteriormente. Estos sistemas construyen internamente una generalización de cómo resolver los problemas y se pueden enfrentar con éxito a problemas del mismo tipo que no han visto previamente.

Los sistemas conexionistas más conocidos son las RNA (Ripley, B. D. 2008) y se utilizan para el reconocimiento de patrones en tareas de clasificación. Han sido aplicados con éxito en tareas de reconocimiento de imagen como en los OCR (Optical Character Recognition) con los que están equipados los escáner y pueden ser usados en cualquier tarea que precise del reconocimiento de regularidades o de valores atípicos en los datos. En el ámbito de la biblioteconomía y la documentación se puede utilizar para detectar hábitos de búsqueda en usuarios o para optimizar términos de búsqueda o de indización de forma automática evitando la subjetividad de los expertos humanos.

4. Herramientas para recuperar información

La aplicación de las técnicas de IA o el desarrollo de algoritmos inteligentes para el problema de la recuperación de información depende, en gran medida, de las características de la información con la que se está trabajando. En este sentido, es muy distinta la información estructurada, como las BBDD que se procesan con técnicas de data mining, de la información no estructurada o documental, a la que se le aplican técnicas de text mining o de la información en vídeo que necesita de algoritmos mucho más específicos.

En los siguientes apartados se detallarán las técnicas específicas para cada tipo de contenido.

4.1. Data mining

La minería de datos (Ye, N. 2003), más conocida por su nombre original en inglés, data mining (DM), es una disciplina que aglutina un conjunto de técnicas, tanto de IA como estadísticas, que son capaces de extraer, a partir de un conjunto de datos, información relevante que no es explícita en una revisión directa. Estas técnicas, como ya se comentaba en el caso de las RNA, que se utilizan también en DM, buscan regularidades en los datos que puedan ser de utilidad en su explotación. Es muy conocida la aplicación de las técnicas de DM en los datos de venta en supermercados, para optimizar la colocación de los productos con el objetivo de maximizar las ventas, o la búsqueda de patrones anormales en los movimientos de tarjetas de crédito para conseguir una detección temprana de usos posiblemente fraudulentos.

Una técnica muy común, aparte de las RNA, y también muy fácil de entender son las reglas de asociación. Buscan en datos del mismo tipo patrones que se repitan y los muestran en forma de reglas. En una biblioteca o librería se podría aplicar a los libros que solicita cada usuario o cliente. Con esta técnica se puede desarrollar un sistema automático de recomendación, ya que la técnica construye reglas del tipo: el lector que adquiere libros de literatura fantástica también adquiere libros de ciencia ficción o el lector que consulta libros de cocina y libros de decoración también está interesado en libros de viajes.

Hay que decir también que para conseguir buenos resultados es necesario contar con una gran cantidad de datos con los que las técnicas de DM puedan trabajar para dar una buena generalidad a las conclusiones que se obtengan. Por

este motivo, es normal utilizar el DM sobre un conjunto de datos o de BBDD que provengan de distintas fuentes. Su distinta procedencia provoca que se necesite una etapa de uniformización u organización de los datos, denominada Data Warehouse DW (Jensen, C. S. et al. 2010), que permite la utilización del DM como si se aplicase a un único conjunto de datos. Esta combinación de la etapa de DW seguida de la de DM es parte de un proceso más general denominado Knowledge Discovery on Databases (Maimon, O. et al. 2005) abreviado como KDD.

4.2. Text mining

Cuando la necesidad de extracción de información de interés se produce sobre información estructurada como textos o documentos, hay que aplicar técnicas de text mining (TM) (Weiss, S. M. 2005). En este caso las herramientas que se usan pueden realizar un análisis estadístico, como los algoritmos de clasificación o indización automática, o un análisis semántico o de significado, como el procesado del lenguaje natural, la realización automática de resúmenes o los análisis de opinión.

Dentro del ámbito de la recuperación de información, centrada en el área de la biomedicina, las técnicas de TM tienen una relevancia especial. La búsqueda de información específica y de documentos en esta área es enormemente compleja debido a que su juventud y rápido crecimiento hace que los distintos conceptos involucrados (genes, proteínas, mutaciones, etc.) sean denominados de forma distinta, en distintos centros de investigación o a lo largo del tiempo. Las técnicas de TM, al trabajar con la semántica, con el significado, en vez de directamente con las palabras, permiten salvar estos inconvenientes y realizar búsquedas con los niveles más altos de exactitud y exhaustividad.

Como en cualquier otro ámbito de nuestra vida como internautas, también en la realización de TM, Google puede facilitarnos la vida. Nuevas herramientas incorporadas por la interfaz de búsqueda de Google realizan de forma automática y transparente para nosotros minería de textos o TM. Funciones muy conocidas como el "Quizás quiso decir" que corrige por nosotros errores tipográficos o la traducción automática basada en ejemplos de Google Translator, para poder acceder a información en páginas escritas en otros idiomas, facilitan las búsquedas que realizamos habitualmente. Pero nuevas herramientas, como la Rueda de Búsquedas (ver Figura 2) que podemos utilizar desde la barra lateral izquierda, en el apartado Más herramientas, nos indican conceptos relacionados con

el concepto de nuestra búsqueda. Estos conceptos relacionados están basados en búsquedas similares realizadas por otros usuarios. Esta herramienta facilita encontrar resultados en las búsquedas de Google incluso aunque los términos de búsqueda que hayamos usado sean inexactos o estén alejados de los términos que conseguirían una búsqueda exitosa en el interfaz habitual de Google.

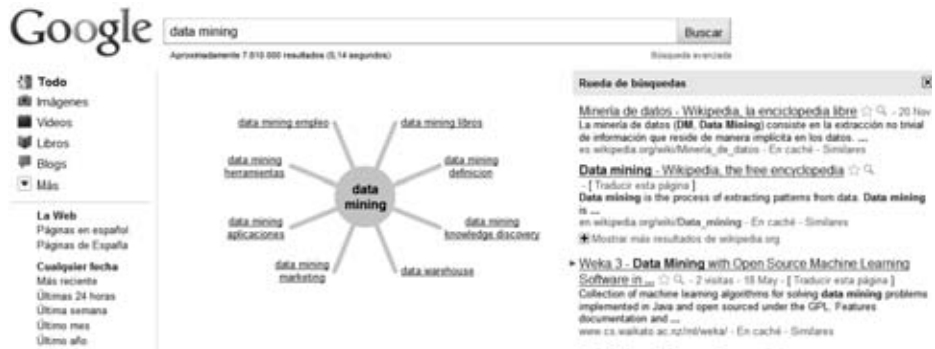

Figura 2. Rueda de búsquedas

4.3. Web mining

La aplicación de las técnicas de TM específicamente en el ámbito de la web se conoce como web mining (Scime, A. 2005) y presenta dos vertientes, el estudio de los registros de navegación de los usuarios y el estudio de los propios contenidos web. Aunque esta última vertiente, el estudio de los contenidos, es la que más se relaciona con la recuperación y análisis de la información, es también muy común el estudio de los registros en los servidores web, que todos dejamos al navegar por la red. El estudio de estos registros es muy importante para las empresas a nivel de captación de clientes o de identificación de objetivos publicitarios. Hay que tener en cuenta que la publicidad en internet es un negocio con un crecimiento exponencial en los últimos años.

En el campo del estudio de los contenidos es muy importante el significado del propio contenido de las páginas a la hora de indizar o clasificar los miles y miles de millones de páginas existentes. Es obvio que la clasificación basada en técnicas estadísticas de términos, aunque útil, no es capaz de captar los matices del idioma que permiten clasificar con exactitud cada página. El lenguaje coloquial, las abreviaturas, los distintos idiomas hacen que trabajar sólo con palabras sea muy complicado.

Es una necesidad ya perentoria integrar la semántica en el desarrollo de las páginas web de forma que las mismas páginas puedan ayudar a los buscadores y a las aplicaciones a conocer su temática. Esta necesidad está empezando a cubrirse al utilizar nuevas herramientas de almacenamiento de conocimiento, como son las ontologías que se comentan a continuación.

4.4. Ontologías

Una ontología (Tamma, V. et al. 2005) es una forma de representación del conocimiento consensuada y de base semántica. Se plasma en un grafo no dirigido en el que los términos, los conceptos del área de conocimiento, se representan como nodos y las relaciones entre términos se representan como arcos que conectan los nodos.

Las ontologías son un tipo de almacenamiento de información más estructurado que las BBDD. Almacenan información de cómo se estructura el conocimiento. Esta información, información semántica, permite integrar la información de distintas BBDD haciendo una equiparación de las tablas y atributos de cada BD con los conceptos y relaciones de la ontología. También permite realizar traducciones de conceptos o de unidades de medida conociendo el contexto de los datos que están almacenados en una cierta BBDD.

Otro punto fundamental de las ontologías es que su desarrollo está consensuado por un conjunto de expertos en el área, por lo que el conocimiento que atesoran está revisado y es totalmente fiable.

El gestor de ontologías más conocido y usado actualmente es Protégé, que está siendo desarrollado por el National Center for Biomedical Ontology en la Universidad de Stanford para la gestión de información en el área biomédica, pero que es utilizado en cualquier otro dominio. Normalmente las ontologías se almacenan en un formato tipo XML, aunque hay algunos específicos, como OWL (Lacy, L.W. 2005). La utilización de lenguajes de marcado garantiza su interoperabilidad en cualquier sistema informático.

Figura 3. Página web del servidor de búsqueda de información Wolfram Alpha

4.5. Web semántica

De la posibilidad de añadir semántica, por ejemplo en forma de ontologías, a las páginas web surge el concepto de web semántica (Virgilio, R. et al. 2010). Este nuevo tipo de web aprovecha las ventajas de las ontologías en cuanto al desarrollo de vocabularios estandarizados para describir conceptos de forma que las búsquedas produzcan resultados más exactos.

Además, las relaciones entre conceptos almacenadas en las ontologías permiten hacer búsquedas más exhaustivas al aprovechar relaciones de generalización y especialización; por ejemplo, si estamos buscando documentos sobre libros electrónicos y disponemos de una ontología que nos diga que: un libro electrónico es un dispositivo para lectura de documentos en formato electrónico, que el e-pub y el pdf son formatos de documento electrónico y que las tabletas y los notebooks son también dispositivos de lectura de documentos. A través de esta conexión, cuando hagamos búsquedas sobre libros electrónicos, en páginas web etiquetadas semánticamente, en los resultados también aparecerá información sobre tabletas y notebooks. Con este comportamiento el sistema nos da opciones en los resultados, nos informa sobre libros electrónicos, pero también sobre otros dispositivos que nos ofrecen una funcionalidad similar.

Un ejemplo de este nuevo tipo de web es el buscador Wolfram alpha. Los resultados de las búsquedas que ofrece Wolfram alpha (ver en Figura 3 y 4 un ejemplo de búsqueda sobre la palabra Madrid) están formados por información

relevante a los términos de búsqueda pero no son páginas web. El buscador accede a distintas páginas para recopilar información y ofrecernos un resumen de la información encontrada. Esta información está completamente actualizada al momento en que realizamos la búsqueda. Este buscador es todavía un prototipo y no es capaz de buscar información sobre cualquier término, pero nos hace vislumbrar el futuro de internet en cuanto a la web semántica.

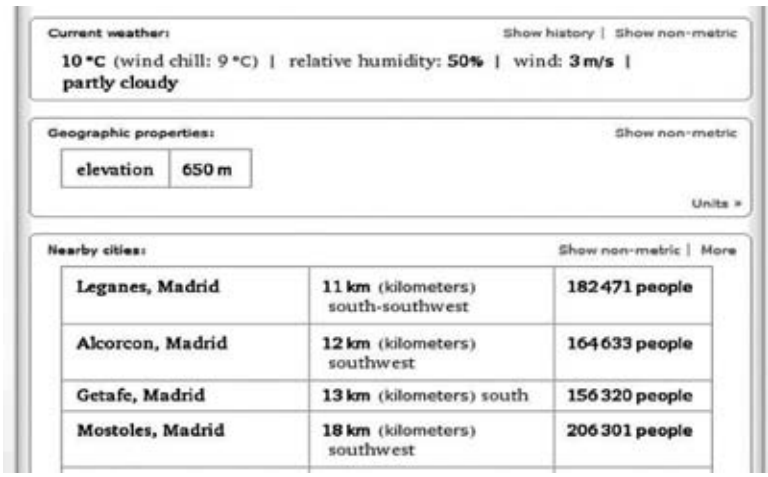

Figura 4. Continuación página Wolfram Alpha con los resultados de Madrid

4.6. Recuperación de información en Vídeo

Aparte de la información estructurada y no estructurada, incorporando números y textos, que es la que habitualmente se tiene en cuenta a la hora de recuperar información, desde hace unos años empieza a tener una importancia fundamental la información en formato vídeo (Elleithy, K. 2010). La aparición de servidores en internet para compartir vídeos, como YouTube, y la bajada continuada del precio de las cámaras y videocámaras digitales, que además pueden incorporar facilidades para subir los vídeos a internet, hacen que cada vez más personas quieran colgar sus vídeos para que los vean familiares y amigos.

También hay que tener en cuenta que los medios de comunicación se basan cada vez más en el vídeo como herramienta de transmisión de noticias.

Todo esto conlleva que la organización e indización de este material, cada vez más común y abundante, pasa a ser un tema importante en la recuperación y análisis de información. Sin embargo, estas tareas no son triviales ya que los

indicadores o palabras clave para clasificar vídeos están dentro del propio vídeo, en el sonido o las imágenes que muestran. En estos momentos, cuando los buscadores como Google nos ofrecen videos en los resultados lo hacen buscando en el texto que está cerca del vídeo, si está incluido en una página web, o en las etiquetas que le asignó el autor del vídeo cuando lo subió a un servidor como YouTube. Esto plantea ciertos problemas como puede ser que se quiera engañar al servidor asociando al vídeo etiquetas falsas o que el vídeo no quede bien indizado al no corresponder de forma exacta el contexto de una página con el contenido del propio vídeo.

Para realizar una indización inteligente y automática de vídeos hay que tener en cuenta tres tipos de datos que podemos aislar dentro del vídeo:

Por un lado está el audio que incluyen muchos vídeos. Este se puede procesar igual que el texto de las páginas web para localizar términos relevantes. Para tratar esta información es necesario aplicar técnicas de Reconocimiento automático del lenguaje natural, que es otra área dentro de la IA. Estas técnicas han avanzado mucho en los últimos años, como podemos comprobar en algunos sistemas de atención telefónica, y presentan un rendimiento adecuado a este objetivo.

Otra de las fuentes de información que se pueden utilizar en los vídeos es el texto que aparece como subtítulo y en carteles o similares. Para recuperar esta información se pueden utilizar OCR (Bunke, H. et al. 1997), técnica ya comentada y que es de uso común desde que se empezó a generalizar el uso de escáneres. En este caso, su utilización plantea el inconveniente de que no se sabe en qué imágenes del vídeo aparecerá el texto ni en qué posición. Es por tanto la aplicación de OCR un proceso, de momento, lento por la cantidad de procesado que necesita pero que puede obtener una información muy importante a partir del vídeo.

Por último, la información más valiosa sería la de reconocer lugares o personajes que aparezcan en el vídeo. Esta tecnología también está disponible aunque, como en el caso del OCR, tampoco funciona en tiempo real, por lo que el vídeo debe ser procesado durante un tiempo para poder extraer esta información.

En estos momentos estas técnicas de recuperación de información a partir de vídeos están siendo utilizadas para documentación en cadenas de televisión, que son los centros que disponen de la infraestructura y los medios para aplicar las técnicas de procesado necesarias a la ingente cantidad de minutos de vídeo que se generan diariamente.

Figura 5. Tipos de elementos reconocidos por Google Googles

Muy relacionada con estas técnicas de reconocimiento de información en vídeo y como ejemplo de lo que el futuro inmediato nos puede deparar en este ámbito me gustaría mencionar la aplicación Googles de Google, que está accesible en su servidor en la sección de Labs. Esta aplicación, orientada a su uso en teléfonos y dispositivos móviles, permite tomar una fotografía y procesarla para reconocer distintos tipos de elementos (ver Figura 5) como si fuera un OCR avanzado. Estos elementos son muy variados, como por ejemplo, monumentos, obras de arte, títulos de libros o marcas comerciales. Una vez reconocido un elemento, muestra resultados de búsqueda de Google como si nosotros mismos hubiéramos escrito el término en el interfaz habitual de búsqueda del servidor Google.

Así es capaz de reconocer una atracción turística en la foto, como la Torre Eiffel y en resultados nos mostrará enlaces a páginas con información de ese monumento. En este caso al ser un monumento ampliamente conocido puede no parecer tan útil ya que sería más rápido escribir directamente el nombre en el cuadro de búsqueda, pero si nos encontramos ante un edificio desconocido puede que sí que nos devuelva información que no podíamos haber encontrado de otra manera. Otro ejemplo útil que se puede ver en la página de información de Google Googles es obtener la foto de la carta de un restaurante en otro idioma, la aplicación reconoce el texto como si fuera un OCR y nos permite traducirlo a cualquier idioma o buscar información sobre las recetas de los platos.

Está claro que sistemas como estos, con el adecuado desarrollo en los próximos años, pueden llevar el ámbito del reconocimiento de información de imágenes y vídeo al ámbito doméstico permitiendo indizar y categorizar automática y rápidamente cualquier material videográfico que se genere en el planeta.

5. Agregadores

Como ya todos sabemos, es muy probable que la información que busquemos resida en alguna página web, sin embargo, es cada vez más habitual que la

información que queremos no esté en una sola página web o en un solo formato (por ejemplo sobre una noticia de un evento deportivo, podemos querer leer una crónica desde un periódico pero también ver un resumen desde una cadena de televisión), incluso podríamos estar buscando una información que todavía no está disponible (información sobre una película o un videojuego que se estrenará en unos días o semanas). En todos estos casos el usuario debería estar navegando en varias páginas web y repitiendo las mismas búsquedas en distintos días para encontrar la información que busca.

Para dar solución a estas situaciones se han desarrollado las herramientas o sistemas de agregación, que pueden estar orientados a varios objetivos.

5.1. Agregadores de información

Existen dos visiones de la agregación de información en Internet, la agregación puede ser de contenidos sobre un mismo tema o la agregación puede ser de información que se va generando en distintos momentos de tiempo.

La función de agregación de contenidos la realizan actualmente los buscadores al integrar en sus resultados distintas fuentes de datos con información relevante a nuestras búsquedas. Empieza a ser habitual que los resultados se estructuren en unos primeros elementos que proceden de fuentes que podemos denominar de confianza, como Wikipedia, periódicos o instituciones públicas, y otro conjunto con resultados de fuentes más diversas.

La agregación temporal de datos se suele realizar mediante el formato RSS (Really Simple Syndication) que sigue el estándar XML, aunque existen otros menos usados como Atom (Hammersley, B. 2005). Utilizando RSS se puede enviar información actualizada, normalmente desde una fuente web, a usuarios que se han suscrito al contenido de esas páginas. Esto permite que, en vez de que el usuario esté comprobando periódicamente si se ha publicado una información en una web, sea la propia web la que informe al usuario de la publicación de una nueva información, haciendo más fácil el enterarse de las nuevas noticias relevantes para el usuario. Estas noticias se pueden leer en un navegador web, pero también se pueden leer desde programas específicos.

En el caso de la agregación temporal de contenidos, Google tiene también algo que aportar, en este caso mediante su herramienta en Tiempo real (ver Figura 6), disponible desde el panel izquierdo de su interfaz habitual de búsqueda.

Mediante esta opción se puede ver, en tiempo real y en escala temporal, las noticias que se están produciendo o que se han producido en las últimas horas o días, permitiéndonos reconstruir el flujo de documentos sobre una determinada información.

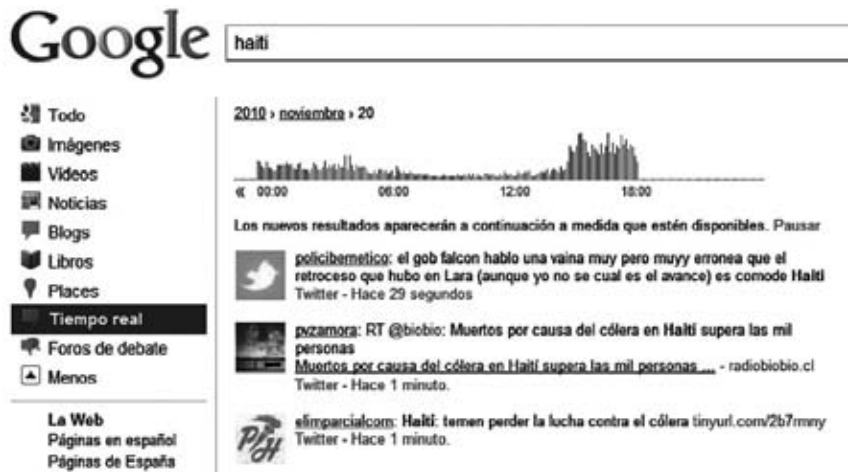

Figura 6. Agregación de noticias en tiempo real de Google

5.2. Agentes inteligentes

Volviendo al ámbito de la IA, un agente inteligente (Chen, Z. et al. 2002) es un programa, normalmente con un tamaño y habilidades reducidas, que percibe su entorno, colabora con otros agentes, procesa datos y responde o genera informes. La gran potencialidad de los agentes inteligentes es que, aunque su funcionalidad sea limitada, puede colaborar con otros agentes consiguiendo comportamientos emergentes que pueden ser de gran complejidad. Una analogía válida en este caso es el mundo de los insectos que viven en colonias, aunque una hormiga individual no es capaz de realizar tareas muy complejas, una colonia de hormigas, trabajando conjuntamente y de forma distribuida, es capaz de construir y mantener una estructura tan compleja como es el hormiguero en el que desarrollan su sociedad.

En el ámbito de Internet los agentes más comúnmente utilizados son también llamados robots y se emplean en los buscadores para indizar y categorizar todas las páginas y recursos web de forma automática y continua. Son necesarios para encontrar nuevas páginas y actualizar la información de páginas ya existentes que son modificadas.

Otro uso común de los agentes inteligentes son las aplicaciones de inteligencia empresarial (o Business Intelligence) (Moss, L. T. et al. 2003). Estos sistemas avisan de las nuevas informaciones que aparecen en internet y que son relevantes para la empresa. Por tanto, la primera parte de este tipo de sistemas es un conjunto de agentes que constantemente acceden a distintas fuentes de información, definidas previamente por el usuario, para comprobar la aparición de nuevas informaciones relevantes. Si estas se producen, los agentes acceden a las páginas y extraen los datos o información importantes para construir un informe que se hace llegar al usuario mediante una fuente en formato RSS o similar. De hecho, estos sistemas palian la no disponibilidad de fuentes web en recursos relevantes para el usuario. Nótese también que los agentes inteligentes necesitan, en este caso, comprender la semántica del recurso para poder discriminar en la información de las novedades que se van produciendo.

5.3. Agregadores de servicios

Además de la agregación de contenidos o de información ya comentada, muy recientemente está surgiendo un nuevo tipo de agregación, denominada agregación de servicios, que se refiere a la agregación de los servicios que proveen distintas páginas web. Las páginas que utilizan estos servicios que se pueden agregar se denominan, en inglés, mashups (Yee, R. 2008), no teniendo una traducción comúnmente aceptada en castellano.

Figura 7. Mashup con la información de nivel de embalses

Figura 8. Resultados de la búsqueda del grupo musical Sidonie

El servicio web que se utiliza para construir el mashup proviene, normalmente, de una página web de uso común, como por ejemplo, Google Maps, Facebook, Twiter, Flicker o YouTube, entre otras. El servicio de estas páginas es incrustado en una nueva página creada para dar un servicio diferente, de forma que se obtiene una funcionalidad distinta del servicio original. En la Figura 7 se puede ver un mashup de ejemplo que muestra una página con información sobre el nivel de agua de los embalses españoles pero, en vez de mostrar los datos como una simple tabla, los visualiza en su posición sobre un mapa de Google Maps. Nótese que para que se considere una página como mashup se tiene que utilizar información de la propia página, con el servicio que se engancha. En este ejemplo, no se añade un mapa sin más para que se acceda a Google Maps. El mapa se muestra, con marcadores verdes, que representan los embalses y, pulsando encima de cada marcador, se accede a los datos de nivel de agua, proporcionados por la web mashup.

Realmente las propias búsquedas dentro de Google ofrecen resultados como si la página del buscador fuera un mashup. Como se puede observar en la búsqueda de la Figura 8 sobre el grupo musical Sidonie, desde la página de resultados se puede acceder a la YouTube, pero no a la página principal, sino a la correspondiente a los vídeos del grupo, pero también a otras aplicaciones como la enciclopedia libre, Wikipedia, la emisora de música por internet con funciones de red social Lastfm para poder oír sus canciones o a servidores de imágenes.

Es muy probable que esta retroalimentación entre servidores web que están abiertos a colaboraciones entre ellos haga surgir nuevas aplicaciones interesantes que aún están por descubrir.

6. Gestión de Conocimiento

Hasta ahora hemos asumido que cualquier tipo de información de la que disponemos en una organización está estructurada (como en BD, por ejemplo) o está en formato documental (como libros o páginas web). Esto nos permite indizarla, categorizarla o procesarla con herramientas informáticas más o menos comunes, desde BBDD a técnicas de IA o de análisis de imagen. Sin embargo en las organizaciones modernas ha surgido nuevas necesidades de gestión y búsqueda de información que son resueltas con nuevos mecanismos dependiendo de si estamos mirando hacia dentro (gestión de conocimiento o GC) o hacia fuera (inteligencia empresarial) de la propia organización. Así, la GC nos ayudará a gestionar el conocimiento que ya está disponible o se está generando en la propia organización, mientras que la inteligencia empresarial nos ayudará a descubrir, lo antes posible, la información relevante para nuestro negocio que esté accesible en el exterior, normalmente a través de internet, para ganar una ventaja competitiva.

Sobre inteligencia empresarial ya hemos hecho un comentario breve en el punto de agentes inteligentes y se puede considerar una forma especializada de recuperación de información con unas restricciones temporales estrictas. En este apartado vamos a comentar el otro tipo de herramienta, la que se orienta hacia la parte interna de la propia empresa, la gestión de conocimiento.

En primer lugar, quería exponer una posible definición posible de GC. La GC (Loshin, D. 2001) es cualquier proceso o práctica para crear, adquirir, capturar, compartir y usar conocimiento, cualquiera que sea el lugar donde resida, para mejorar el aprendizaje y el desempeño en las organizaciones.

Está claro que la GC surge de la necesidad de las empresas de almacenar y mantener el conocimiento que poseen o adquieren sus trabajadores, con el objetivo de que este bien intangible que es el conocimiento resida también en la empresa a la vez que en sus trabajadores, y para evitar perderlo cuando algún trabajador deje la empresa. Estas técnicas son cada vez más importantes debido a que cada vez más las empresas producen y venden conocimiento, al evolucionar la industria hacia la sociedad del conocimiento, dejando la producción o la manufactura de bienes a otros productores, sobre todo en Asia. En esta situación,

los activos de la empresa no son máquinas de alta tecnología, ni otros bienes tangibles. Los activos son los conocimientos, el "know how" y la capacidad de innovación.

Las técnicas de GC difieren bastante de las técnicas de IA comentadas hasta ahora. No se basan en la informatización total del conocimiento, en que el ordenador lo asuma y sea capaz de razonar con el. La GC es una iniciativa en una dirección distinta. Ha desarrollado un conjunto de técnicas y herramientas con el propósito de asegurar que las personas correctas de una organización tengan el conocimiento correcto en el momento adecuado. Se puede ver más bien como un directorio inteligente o un catálogo de recursos que permite almacenar de forma organizada y acceder de forma eficiente a todo el saber de la organización.

Por tanto, en vez de técnicas orientadas a extraer o emular el conocimiento, como los SSEE y las RNA, desde la IA, en GC lo que se utilizan son técnicas de organización, como la documentación indizada de informes realizados, el directorio de personal ordenado por habilidades o conocimientos o los listados de preguntas frecuentes para el apoyo a las nuevas incorporaciones a la empresa, entre otros.

Un sistema de GC normalmente cuenta con al menos estos cuatro elementos:

- Una base de datos documental, que permita una gestión profesional de los fondos documentales de la empresa, entendidos estos en un sentido amplio. En esta BBDD es necesario almacenar documentación obvia, como libros y manuales, pero también documentación sobre proyectos realizados, cursos celebrados, relación con proveedores y clientes, y cualquier otra documentación que pueda ser relevante para el funcionamiento interno.
- Listados de buenas y malas prácticas, que permiten estructurar la experiencia de los usuarios en la resolución de problemas o en la manera en que se han enfrentado distintos proyectos de forma que se pueda amortizar esta experiencia en trabajos futuros.
- Listado de preguntas frecuentes, la cual es una herramienta muy común en el ambiente informático y muy fácil de realizar haciendo una estadística de las preguntas más realizadas en cada departamento junto con sus respuestas correctas. Si el personal de la empresa se acostumbra a acudir primero a este listado antes de preguntar a otros compañeros se ahorra tiempo, se garantiza siempre la misma respuesta correcta y se agiliza el trabajo.
- Páginas amarillas, que almacenan una lista del personal, tanto de la empresa como personas relacionadas (por ejemplo clientes y proveedores). En este

directorio de personas se enlaza, además de la información habitual de teléfono, correo electrónico, número de despacho, etc., información sobre las habilidades, tipo de proyectos que realiza o en los que es experto, tipo de programas informáticos que domina o puede enseñar, etc. Este directorio, aparte de conseguir orientar al personal de la empresa sobre con quién pueden hablar sobre un determinado tema, también permite seleccionar personas para afrontar nuevos proyectos o para tratar con ciertos clientes.

Para desarrollar un sistema de GC no existen, de momento, muchas herramientas específicas, pero hay herramientas de gestión de contenidos que pueden hacer un muy buen papel como sistemas de GC. Hablando de una forma genérica un Wiki se puede personalizar para que haga las cuatro funciones antes mencionadas, incluso cualquier gestor de información personal (en inglés, Personal Information Manager o PIM), como Microsoft Outlook, que puede gestionar información de contactos, tareas, correos electrónicos y realiza búsquedas cruzadas en toda esta información también se puede utilizar en este ámbito.

Si queremos un software más específico para desarrollar un sistema de GC hay dos ejemplos que se pueden mencionar. Uno es Microsoft SharePoint que permite desarrollar un portal web en el que los usuarios registrados pueden interaccionar mediante herramientas sociales (como correo, chat, etc.), compartir o depositar archivos (organizándolos en categorías), permite definir proyectos y tareas, compartir documentos y generar listas (para preguntas frecuentes, buenas y malas prácticas, etc.), entre otras posibilidades, como se puede ver en la Figura 9.

Figura 9. Capacidades de GC del programa SharePoint de Microsoft

Otro software muy interesante es Moodle, muy utilizado en entornos universitario como plataforma de tele-enseñanza, por sus capacidades de personalización y por basarse en una iniciativa de software libre. Moodle también se puede usar para desarrollar un sistema de GC con las características ya mencionadas. Probablemente con algo más de trabajo que si se usa SharePoint pero con más posibilidades de adaptación a las necesidades concretas de cada organización.

7. Conclusiones

En este artículo he querido comentar distintas técnicas (como las de IA) y aproximaciones (como los distintos tipos de minerías o la GC) que nos permiten trabajar con la parte superior de la pirámide de la información, tanto el conocimiento como la sabiduría.

Estas técnicas y aproximaciones se aplican sobre distintos tipos de información estructurada y documental que son usadas, cada vez más, sobre fuentes que están disponibles en internet a través de distintos accesos web. Esa hegemonía de la web como depositaria de la información hace que muchos de los ejemplos de herramientas vengan de la mano de la empresa Google que es, al menos, pionera de muchos de los sistemas y herramientas de búsqueda y recuperación de información.

El futuro, por lo menos a corto y medio plazo, de la recuperación de información pasa por la aplicación de aproximaciones semánticas, como las basadas en ontologías, que van a permitir mejorar la exactitud y exhaustividad a la vez que permitirán evitar los problemas que plantean los sinónimos, las expresiones coloquiales (o variantes del lenguaje como las de los mensajes cortos de texto) o los idiomas.

Siendo el tema de este artículo muy actual, estoy seguro que dentro de unos años será posible volver a utilizar el mismo título, escribiendo un artículo totalmente distinto, pero planteando los nuevos avances en las distintas áreas mencionadas. Por tanto, espero que esta hipotética situación me ofrezca de nuevo la oportunidad de disfrutar impartiendo una nueva conferencia como la que ha sido el origen de este capítulo, para poder plasmarla en un nuevo capítulo de un nuevo libro.

8 Bibliografía

- Bunke, H.; Wang, P. S. P. (1997) HandBook of Character Recognition and Document Image Analysis. World Scientific.

- Chen, Z.; Ichalkaranje, N. (2002) Intelligent Agents and Their Applications. Springer Press.

- Coombs, J. H.; Renear, A. H.; DeRose, S. J. (1995) Markup Systems and the Future of Scholarly Text Processing. Detroit, (USA) MIT Press.

- Elleithy, K. (2010) Advanced Techniques in Computer Science and Software Engineering. Springer Press.

- Fernandez Gil, A. (2010) Sistemas Expertos: Representación e inferencia. Problemas resueltos. Madrid (España). Universidad Rey Juan Carlos.

- Fogel, D. B. (2006) Evolutionary Computation: Toward a new philosophy of machine intelligence (3rd ed.). Wiley-IEEE Press.

- Galindo, J.; Urrutia, A.; Piattini, M. (2006) Fuzzy Databases: Modeling, Design and Implementation . Hershey, USA. Idea Group Publishing.

- Goldfarb, C. F.; Prescod, P. (2000) XML Handbook. New Jersey (USA). Prentice Hall.

- Hammersley, B. (2005) Developing Feeds with RSS and Atom. O'Reilly Media.

- Jensen, C. S.; Padersen, T. B.; Thomsen, C.; Ozsu, M. T. (2010) Multidimensional Databases anda Data Warehousing. Morgan Y Claypool Publishers.

- Lacy, L.W. (2005) OWL Representing Information Usingthe Web. Trafford Publissing.

- Loshin, D. (2001) Enterprise Knowledge Management. Elsevier Press.

- Luger, G.; Sttubblefield, W. (2004) Artificial Intelligence: Structures and Strategies for Complex Problem Solving. The Benjamin/Cummings Publishing Company, Inc.

- Maimon, O.; Rokach, L. (2005) Descomposition Methodology for Knowledge Discovery and Data Mining. World Scientific.

- Moss, L. T.; Afre, S. (2003) Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications. Addison-Wesley.

- Real Academia Española (2001). Diccionario de la Lengua Española. Madrid, España. Espasa-Calpe.

- Ripley, B. D. (2008) Pattern Recognition and Neural Networks. Cambridge University Press.

- Scime, A. (2005) Web Mining: applications and techniques. IGI Global Press.

- Tamma, V.; Cranefield, S.; Finin, T. W.; Willmott, S. (2005) Ontologies for agents: Theory and experiences. Birkhäuser Basel.

- Virgilio, R.; Gianchiglia, F.; Tanca, L. (2010) Semantic Web Information Management. Springer Press.

- Weiss, S. M. (2005) Text mining: predictive methods for analyzing unstructured information. Springer Press.

- Ye, N. (2003) The handbook of data mining. CRC Press.

- Yee, R. (2008) Pro web 2.0 Mashups: Remixing Data and Web Services. Apress.