1. Los sistemas de recuperación de información

Se exponen los fundamentos de los sistemas de recuperación de información y su desarrollo para internet. Se delinean las nociones básicas de metadatos y del web semántico.

1.1. Los sistemas de recuperación de información

El tratamiento y recuperación de la información ha sido una preocupación, y una actividad, que han llevado a cabo todas las sociedades. Desde una perspectiva académica, los fundamentos de esta disciplina se establecen progresivamente desde mediados del siglo XIX, cuando se establecen y se formalizan las enseñanzas superiores sobre bibliotecas, archivos y museos. Sin embargo, es necesario esperar hasta mediados del siglo XX para que se conforme la denominada *Information Science*, que traslada el centro de la actividad científica y profesional hacia diferentes sistemas de tratamiento y recuperación de información, mediados tecnológicamente, y cuyo objetivo final es satisfacer las necesidades de información de los usuarios.

La multiplicación del volumen de información científica y técnica, derivada del esfuerzo de la Segunda Guerra Mundial, trajo como consecuencia la necesidad de introducir máquinas que facilitasen el tratamiento, almacenamiento y recuperación de la información. El desarrollo del sistema SMART (*System for the Mechanical Analysis and Retrieval of Text*) por el equipo de G. Salton en la Cornell University, en la década de 1960, estableció los fundamentos de los modernos sistemas de recuperación de información, de los que los motores de búsqueda en internet son herederos. Las décadas entre 1970 y 1990 vieron un desarrollo progresivo de este tipo de sistemas de recuperación, así como la aparición de sistemas comerciales de pago, accesibles a través de redes de comunicaciones, que ofrecían principalmente acceso a bases de datos documentales de contenido científico, económico o financiero. Cuando se populariza el acceso a internet, a mediados de la década de 1990, ya existía un mercado previo y grupos de usuarios especializados en la búsqueda de información. Los motores de búsqueda en internet de la primera generación (*AltaVista*, *Lycos...*) pusieron al alcance de cualquier usuario capacidades hasta entonces limitadas a entornos cerrados y especializados. La aparición de *Google* en 1998 supuso la disponibilidad de una herramienta de fácil comprensión y uso para que cualquier usuario pudiese buscar y localizar información en internet.

Los sistemas de recuperación de información son aquellos que ofrecen al usuario funcionalidades para acceder a fuentes y recursos de información en entornos digitales, y consultar, recuperar y extraer de los mismos aquellos documentos cuyo contenido sea capaz de dar respuesta una cuestión planteada por el usuario. En muchas ocasiones, estos sistemas se presentan integrados en otros, por ejemplo como componentes de sistemas de gestión de documentos o sistemas de gestión de contenidos.

arquitectura_SRIFig. 1. Arquitectura de un sistema de recuperación de información (fuente original)

El factor clave que identifica a estos sistemas es su capacidad para procesar información textual, tanto en lo que se refiere a su adquisición y tratamiento, como en lo que respecta a la recuperación de la información contenida en el mismo. Esta información textual se recoge en documentos, que no suelen tener una estructura claramente formalizada (a diferencia de una hoja de cálculo o una tabla de base de datos, por ejemplo), y que pueden ofrecer múltiples combinaciones de contenido. La estructura funcional de un sistema de recuperación de información responde a:

- **Captura de información**: funcionalidades para capturar y almacenar documentos en diferentes formatos, para formar un corpus documental.
- **Procesamiento de información:** funcionalidades y algoritmos para generar representaciones de los documentos originales, según diferentes aproximaciones y criterios.
- **Recuperación de información**: funcionalidades y prestaciones para la formulación de expresiones o ecuaciones de búsqueda complejas, con una sintaxis propia y definida, y para la ejecución de esas expresiones contra el conjunto de representaciones resultado del procesamiento previo.
- **Salida de información:** las funcionalidades de presentación de resultados y de manipulación (filtrado, etc.), que se ofrecen al usuario con los resultados obtenidos de la ejecución de las expresiones de búsqueda.

Financiado por el Ministerio de Educación y Formación Profesional y por la Unión Europea - NextGenerationEU

logo.png

1.2. Los sistemas de recuperación en internet

Las bases del éxito en la búsqueda y recuperación de información en internet son el conocimiento de los principios básicos de la recuperación de información y de los sistemas que la hacen posible, y de las características propias de los documentos existentes en internet. Las herramientas de búsqueda en internet aplican los principios sobre tratamiento y recuperación de información textual que se han revisado en el apartado anterior, y los usuarios disponen de similares prestaciones para la recuperación, y para su consulta y filtrado. Por lo tanto, **resulta crucial que el usuario conozca los tipos de información, la variabilidad de formatos y las diferentes presentaciones que puede adoptar la información en internet.** Ello le dotará de una mayor capacidad para conocer y valorar los resultados obtenidos durante el proceso de búsqueda.

Si bien un sistema de recuperación, en su formulación clásica, trabajaba sobre corpus documentales bastante homogéneos, no puede decirse lo mismo de los sistemas de recuperación en internet. Al tratarse de un entorno abierto y cambiante, las herramientas de búsqueda ofrecen listados de resultados, que dirigen al usuario hacia el documento original. Los cambios que se producen, por la propia dinámica del web, hacen que en ocasiones esa redirección no ofrezca los resultados esperados, y que haya que completar la búsqueda mediante procesos de exploración basados en la navegación. El usuario siempre debe pensar que no es suficiente, en recuperación de información en internet, con seguir los resultados obtenidos de un motor de búsqueda: hay que explorarlos, analizarlos, valorarlos, y seleccionarlos como adecuados, o desecharlos como no pertinentes. Los sistemas de recuperación de información en el web son un medio más, una fase intermedia, no un fin.

Una cuestión que debe tenerse en cuenta cuando se busca información en internet es que, contra la extendida creencia, **no todo está disponible a través de los motores de búsqueda**, ni en Wikipedia. La puesta en línea a través de internet,desde la década de 2000, de un gran número de fuentes y recursos de información, no supuso que su contenido fuese automáticamente incorporado al contenido procesado por los motores de búsqueda. Diferentes intereses comerciales y/o limitaciones técnicas excluyen enormes volúmenes de información de la vigilancia de los motores, configurando lo que se ha dado en llamar la “***internet invisible***”.

Dark-web-deep-web.jpgFig. 2. El clásico iceberg de internet (múltiples fuentes)

En realidad, estos contenidos no son invisibles para el usuario: lo son para los motores. La noción de internet invisible se asocia a la presencia en la red de recursos de información, cuyo contenido sólo está disponible a través de los sistemas de recuperación que ofrecen los propios recursos. Esto es debido precisamente a que, a su vez, esta internet invisible se encuentra recogida en bases de datos que sólo muestran su contenido cuando son interrogadas, generando páginas web dinámicas, que evidentemente no pueden ser descubiertas y analizadas por los robots que utilizan los buscadores tradicionales. Dentro de la esta área invisible se engloban los directorios y las bases de datos especializadas, los catálogos de bibliotecas, archivos y museos, las bases de datos de prensa, etc. La conclusión lógica que se deriva de ello es que **el usuario debería conocer aquellos recursos de información especializada que resulten más adecuados para sus necesidades**. Una aproximación común es comenzar la búsqueda en un motor generalista, para completarla en recursos especializados en una segunda fase.

Material complementario

- Búsqueda y recuperación de información en la web: qué ha pasado y qué podemos esperar en el futuro (2011)

Financiado por el Ministerio de Educación y Formación Profesional y por la Unión Europea - NextGenerationEU

logo.png

1.3. Metadatos

Los **metadatos** han sido definidos ampliamente como "*datos que describen otros datos*" o "*datos sobre los datos*". En el contexto de los procesos de búsqueda en internet, el término se utiliza para hacer referencia a **conjuntos de datos etiquetados, incorporados al código de las páginas web, o a la cabecera de diferentes formatos de documentos, que incluyen información sobre autoría, título, fechas de relevancia, contenidos, descripción y otros elementos similares**. Aunque es creciente el número de páginas web que van incorporando conjuntos de metadatos, en un gran volumen de contenido de internet no se han producido ni publicado estos metadatos. La descripción más precisa de los contenidos usando metadatos se lleva a cabo en el proceso de edición y publicación de contenidos, por parte de los creadores de los mismos. Esto explica que los sistemas de gestión de contenidos hayan ido incorporando funcionalidades de este tipo. Al igual que en el caso de la web semántica, los metadatos son para máquinas.

El esquema de metadatos más conocido es Dublin Core, un conjunto de quince elementos que están diseñados para describir cualquier tipo de información digital. Por su sencillez, se ha convertido en el esquema de referencia y en el formato de intercambio de infomación entre un buen número de servicios y productos de información digital.

dc.jpgFig. 3. Los 15 elementos del estándar Dublin Core (fuente original)

Los motores de búsqueda están preparados para identificar los conjuntos de metadatos, y para darles más importancia cuando proceden a generar la representación de una página web en sus índices. Evidentemente, esto supone que, a igualdad de contenido, una página con metadatos será mejor considerada por un motor para ofrecerla como respuesta a los usuarios, que la que no los lleve. Como puede imaginarse, esto tiene amplias repercusiones. La más conocida de ellas ha sido la aparición de técnicas de posicionamiento de resultados, o de optimización para motores de búsqueda (en inglés conocidas con el acrónimo *SEO*), merced a las cuales se intenta asegurar una posición mejor en los listados de respuestas de los motores de búsqueda. A la hora de analizar los metadatos, hay que tener en cuenta que también son ampliamente usados como herramienta de marketing en internet, por lo que en ocasiones puede suceder que un resultado, al ser evaluado, pueda resultar decepcionante. Los motores aprovechan los metadatos que encuentran, pero no pueden evaluar su validez.

Financiado por el Ministerio de Educación y Formación Profesional y por la Unión Europea - NextGenerationEU

logo.png

1.4. El web semántico

A comienzos del siglo XXI Tim Berners-Lee y otros investigadores propusieron avanzar en la organización y recuperación de la información en el web adoptando un conjunto de métodos y técnicas que se englobaron bajo la expresión “web semántico”. En realidad, **este esfuerzo está dirigido a estructurar contenidos de todo tipo de recursos y documentos web de acuerdo a un conjunto de estándares**. La idea básica es **incorporar datos etiquetados que sirvan para describir el contenido, el significado y la relación entre los diferentes elementos** etiquetados.

Los usuarios de estos datos estructurados no son los usuarios finales: los destinatarios son las máquinas. El web semántico está diseñado para que sean las máquinas, mediante procesos de interoperabilidad, las que lleven a cabo la tareas de identificar, relacionar y presentar la información. Precisamente una de las razones de la propuesta es superar las limitaciones que se encuentran los motores de búsqueda al procesar documentos con información textual poco estructurada y sin descripción normalizada.

mariamoliner.pngFig. 4. María Moliner etiquetada semánticamente en Wikidata (fuente original).

Los proyectos de web semántico se están centrando en etiquetar y hacer interoperables grandes silos de datos etiquetados, como los catálogos, y en la creación de ontologías, como recursos de descripción de entidades. A través de estas técnicas es posible relacionar informaciones de manera automática, integrándola desde diferentes recursos, y eliminando las posibles inconsistencias o confusiones entre las entidades. Sin embargo, y en lo que concierne a la búsqueda de información en internet, su uso aún no se ha hecho común, y los motores sólo hacen un aprovechamiento limitado, principalmente de los etiquetados de metadatos.

Financiado por el Ministerio de Educación y Formación Profesional y por la Unión Europea - NextGenerationEU

logo.png