Reflexiones entorno al software de gestión y el acceso a la información: Aspectos fundamentales de la teoría de la recuperación de la información en la Internet

Dra. Montserrat Sebastià Salat.
Biblioteconomía y Documentación
Universidad de Barcelona

Resumen:

Las expectativas de cambio en los sistemas de recuperación de la información se fundamentan en la teoría del conocimiento, la teoría de la recuperación de la información y la implosión de la informació generada por la Internet. Este trabajo explora el marco fundamental para el estudio de los sistemas de recuperación de la información: las formas de difusión de la información, el marco conceptual de la teoría de la información, las interacciones de las tecnologías de la información y la comunicación y las limitaciones de los sistemas de recuperación de la información (problemas esenciales del acceso a la información, y el comportamiento de los usuarios. El objetivo final es demostrar que los softwares de gestión y recuperación de la información tienen una asignatura pendiente: igualar la representación del contenido de la información frente a la representación de las necesidades de información.

Palabras clave: Teoría de la Recuperación de la Información, Software de Gestión, Sistemas de recuperación de la información, Internet, Web, Representación del conocimiento.

Abstract:

Change expectations on information retireval systems are based on knowlegde theory, information retrieval theory and information explosion generated by Internet. This paper explores the essential framework for the studying of information retrieval systems: ways for information dissemination, the conceptual frame of information theory, the interactions between information and communication technologies, and the limitations of the information retrieval systems (basic problems for information access, and user behaviour). The final goal is to show that data management and retrieval software have a matter still to be resolved: to even out the document content representations with the representations of information needs.

Keywords: Information retrieval theory, management software, information retrieval systems, Internet, Web, knowledge representation.

INTRODUCCIÓN

Inscribir esta ponencia en el perfil de la Teoría de la Recuperación de la Información significa situar el trabajo sobre tres centros de interés: (1) Las formas de difusión de la información en la actualidad, (2) La evolución histórica del *corpus* conceptual de la recuperación de la información a lo largo de los últimos cincuenta años, y (3) Las interacciones actuales entre la información disponible, la información accesible y las premisas impuestas desde las tecnologías de la información y de la comunicación.

El objetivo fundamental de muchos de los proyectos de investigación sobre las técnicas y los sistemas de recuperación de la información se basan en el desarrollo de la capacidad de creatividad para conseguir la gestión eficaz del conocimiento. El potencial existente en esta área es inmenso, tanto desde el punto de vista teórico como también desde la vertiente de aplicaciones prácticas tales como la implementación de nuevos métodos de acceso y recuperación de la información en los softwares de gestión y difusión de la información.

I.- MARCO FUNDAMENTAL PARA EL ESTUDIO DE LOS SISTEMAS DE RECUPERACIÓN DE LA INFORMACIÓN.

La Teoría de la Recuperación de la Información ha creado desde 1950 una serie de instrumentos de búsqueda que permiten evaluar los límites del acceso a la información, y a su vez ponen de manifiesto algunos de los problemas estructurales que los métodos de recuperación han volcado en los softwares de gestión. No hará falta insistir que la Internet ha asestado el golpe definitivo a la vieja tradición documental. Tradición centrada en la producción de sistemas de información estructurada y basada en la gestión de información secundaria, que ha sido explotada con subsistemas de recuperación lineal de la información. Actualmente, los sistemas de recuperación de la información se asientan sobre cuatro pilares: (1) La definición del conocimiento y el flujo de la información desarrollada desde la filosofía y la teoría del conocimiento con el análisis de los sistemas cognitivos, los procesos de producción del conocimiento y el concepto de información. (2) La gestión del conocimiento sobre la base de la disponibilidad de productos informativos –finales o intermedios que fijan los estadios de producción del conocimiento- y que han ampliado la difusión de la información desde la secundaria a la primaria, y desde la estructurada a la multimedia. (3) El marco conceptual de la Teoría de la Recuperación de la Información (TRI) dividido entre los partidarios de priorizar el análisis de contenido y los defensores de construir sistemas de recuperación de la información. (4) Las interacciones de las tecnologías de la información y de la comunicación (TICs) que están substituyendo la recuperación lineal por la recuperación interactiva de la información, así pues se impone la diversidad y la flexibilidad como principios en el diseño de los sistemas avanzados de recuperación de la información.

La producción de sistemas y servicios de información mediante el uso de las actuales tecnologías de la información y de la comunicación ha permitido ampliar a tres las formas de difusión de la información. Por un lado, tenemos la ya mencionada información estructurada y operativa a partir de los años 1960 –bases de datos bibliográficas, catálogos de bibliotecas y bases de datos distribuidas-, por otro lado contamos con la información textual disponible de forma generalizada a partir de la Internet -revistas electrónicas, prensa, libros, obras de referencia y pre-publicaciones-, y por último, hemos incorporado en esta última década la información multimedia –páginas web, archivos digitales y gráficos, proyectos de bibliotecas digitales-, pasando así en menos de veinte años de la difusión de documentos secundarios a documentos primarios en texto íntegro, y después a la difusión directa de documentos multimedia. Conviene precisar que este salto cualitativo a una mayor disponibilidad de información ha ido acompañado de ventajas e inconvenientes en los sistemas de recuperación, y que en una primera etapa los métodos antiguos de indización y recuperación de la información no asumieron las alternativas de la explosión documental e informativa.

Lo que hemos llamado 'Teoría de la Recuperación de la Información' (TRI) nace a propuesta de Mooers [[i]] en 1950 bajo el nombre de *Information Retrieval* y se reconoce con las siglas *IR*. Por entonces el eje conceptual de la *IR* se configuró sobre las disputas entre dos autores: Mooers y Bar-Hillel [[ii]]. Las controversias iniciales entre ambos autores pusieron de manifiesto las limitaciones de la recuperación de la información, y dividieron el área entre los partidarios de enfatizar el análisis de contenido y los partidarios de facilitar la exhaustividad en el acceso a la información. Por una parte, existía un núcleo de autores que proponían el desarrollo del análisis de contenido a partir de los procesos de coordinación y del establecimiento de los puntos de acceso temático (encabezamientos de materias y clasificaciones) asegurando de esta manera el principio de disponibilidad de la información basado en la identificación y el

procesamiento de la información. Y por otra parte, había un grupo importante de profesionales que defendían la recuperación de la información a partir de la creación de índices temáticos exhaustivos que reflejaran el lenguaje de especialidad y las redes semánticas de cada área de conocimiento, se priorizaba de esta forma el principio de accesibilidad que se fundamenta en los sistemas y servicios de difusión de la información. Este cisma inicial prefigura todas las tendencias y controversias sucesivas en la teoría de la recuperación de la información, y subsiste con mayor o menor impacto en los subsistemas de recuperación de las distintas generaciones de softwares. Aún así, los problemas más graves no aparecerán hasta la globalización de la Internet, fundamentalmente porque la recuperación de la información pasa de ser estructurada y lineal a ser asociativa e interactiva gracias a la tecnología del hipermedia, y los servicios de la red de redes rompen con la exclusividad de los sistemas y servicios existentes basados en interfaces de usuario de primera y a lo sumo de segunda generación.

Debemos suponer, como ya hemos visto, que la Teoría de la Recuperación de la Información hoy está en una etapa expansiva: si ha cautivado el interés de la filosofía y la teoría del conocimiento, si ha llegado a delimitar un marco conceptual híbrido entre el análisis de contenido y los métodos de recuperación, y si ha favorecido en los softwares el diseño de funciones de búsqueda, la implementación de la gestión del lenguaje documental, y la integración de los servicios y de los productos de difusión de la documentación, ello significa que tiende a seguir siendo una de las áreas más activas y complejas de las llamadas ciencias de la documentación. Esta complejidad se basa en su naturaleza intradisciplinar, interdisciplinar y multidisciplinar respecto a la información, a todas las áreas del conocimiento y a la influencia de las disciplinas que le son complementarias. El reconocer que la recuperación de la información es compleja denota el conocimiento del marco fundamental de esta área más allá de la Internet y del panorama general de los servicios de información electrónica. Por consiguiente, para proponer soluciones alternativas a la situación actual de caos y de diversidad *ad infinitum* en los sistemas de recuperación debe plantarse necesariamente en la evolución de los softwares de gestión: la integración de las nuevas tendencias en la indización y en la recuperación. Y por tanto los softwares de recuperación deben incorporar: las funciones sofisticadas de recuperación, la integración de subsistemas de tratamiento lingüístico, el lenguaje natural, el acceso multilingüe, los métodos que subsanan las dificultades de compatibilidad con los usuarios, el desarrollo de interfaces inteligentes de diálogo y parametrizables, la difusión personalizada, etc.

Este breve panorama del marco fundamental de los ejes que confluyen en la TRI nos parece que pone en evidencia la necesidad de crear un espacio profesional para la investigación. Pensamos que es preciso sensibilizar sobre la liberación de tiempo profesional para que los especialistas y responsables en la recuperación de la información y los profesionales que gestionan servicios de información investiguen la mejora de los softwares de gestión. Las nuevas generaciones de softwares deben contar no sólo con los métodos de recuperación, sino que deben incorporar las premisas de las formas de difusión de la información, no es lo mismo la recuperación de la información fundamentada en registros bibliográficos que texto completo o bien en documentos multimedia. La recuperación de la información ha generado su propio *corpus*, sin el cual los especialistas en sistemas de recuperación no podrían interrelacionar la calidad de la indización con la pertinencia en la recuperación de la información. La aparición de nuevas tendencias en las tecnologías de la información y de la comunicación puede suponer una transformación radical de los sistemas y métodos de recuperación de la información.

II.- LIMITACIONES DE LOS SISTEMAS DE RECUPERACIÓN DE LA INFORMACIÓN.

Los límites de los sistemas de recuperación de la información se sitúan en tres escenarios distintos y a la vez complementarios: (1) Los problemas esenciales de la recuperación de la información, (2) La evolución del comportamiento de los usuarios respecto a la Internet y los sistemas de acceso a los recursos de información.

En 1993 Lancaster y Warner [[iii]] condensan los problemas fundamentales de la recuperación de la información a partir de tres cuestiones que aún no han obtenido respuestas eficaces: (1) ¿Cómo se puede conseguir equiparar la actual producción de información frente a las necesidades de información? (2) ¿Cómo pueden los servicios de información analizar la documentación y representar el contenido informativo asegurando una global y perfecta representación del perfil informativo del documento? (3) ¿Cómo puede igualarse en eficacia la representación del contenido de la documentación con la representación de la búsqueda de información, para a su vez garantizar una recuperación completamente satisfactoria?
Algunas de las posibles respuestas a las cuestiones mencionadas por Lancaster y Warner podemos hallarlas en el desarrollo teórico de la TRI. La teoría de la recuperación de la información se ha planteado sus límites operativos a partir de la implosión de la llamada 'sociedad de la información'. Sin embargo, existen otras limitaciones estructurales concretas, que arrancan del génesis de la TRI y que fueron caracterizadas por Swanson [[iv]] como los 'principios de la impotencia' (PI) –a los cuáles nosotros preferimos denominar principios de imprecisión- en la recuperación de la información. Dichos principios se apoyan en la premisa según la cual la producción del conocimiento es compleja y diversa, y por lo tanto resulta fácil de argumentar la incapacidad para representar de forma exacta la naturaleza del propio conocimiento por parte de los intermediarios de la producción del conocimiento: los especialistas en documentación e información. A partir de este postulado se deducen nueve principios que avalan la incapacidad /impotencia en la representación del conocimiento humano, y refuerzan a su vez la imprecisión en la recuperación de la información.

Principios de Impotencia / Imprecisión (PI):

(1) La necesidad de información no puede ser traducida de forma exacta a una estrategia de búsqueda porque es imposible abarcar el contexto total del conocimiento. Por lo tanto, la precisión y satisfacción en la respuesta a una consulta falla, en primer lugar, por la falta de precisión en el planteamiento y formalización de la búsqueda.
(2) La dificultad que conlleva la creación de funciones automáticas que traduzcan de forma adecuada las nociones y los conceptos que definen una búsqueda, sitúa a la recuperación de la información en el ámbito de la formulación de hipótesis y/o conjeturas, porque el proceso de acceso al conocimiento no siempre puede reducirse a reglas.
(3) Un documento y una información pueden ser considerados relevantes o irrelevantes no por ellos mismos, sino porque existan y se establezcan asociaciones con el contenido informativo de otros documentos.
(4) En el proceso de recuperación de la información de ningún modo puede garantizarse que todos los documentos recuperados serán relevantes, porque en la práctica no podemos verificar si todos los documentos relevantes producidos han sido analizados.
(5) Los analistas documentales y los indizadores no pueden reproducir siempre la esencia del conocimiento humano y condensarlo en el proceso de la indización.
(6) La proporción elevada de términos, basada en los estudios estadísticos de relevancia, no asegura la eficacia en los procesos de indización y consiguientemente en la recuperación de la información.
(7) La habilidad en soportar la interacción entre usuario-sistema por parte de un sistema de recuperación de la información no puede evaluarse sobre la base de la eficacia de situaciones individuales.
(8) Se puede ser muy eficaz en el proceso de representación del conocimiento mediante el establecimiento de todas las opciones conceptuales asegurando así la relevancia de la información disponible, o bien se puede conseguir un alto nivel de efectividad en la funciones automáticas de búsqueda, pero no ambas cosas a la vez con los sistemas disponibles.
(9) Del PI anterior se deduce que la efectividad completa de la indización automática y de la recuperación adscrita a ella no es posible. El problema es pues que la representación del conocimiento humano no se reduce a normas ni a la posible manipulación y explotación en bases de datos.

De esta manera, hemos de reconocer que los problemas de impotencia/imprecisión de la representación y de la recuperación del conocimiento son primordialmente problemas conceptuales situados en los *intra muros* de la epistemología de las ciencias de la documentación y paralelos a los existentes en muchas otras áreas del conocimiento.

El impacto que generaron en la cultura profesional los PI provocó una interesante propuesta que complementa el estudio de los límites de la recuperación de la información, son los llamados 'principios de fertilidad' (PF) [[v]]. Esta propuesta se basa en el estudio del impacto que la recuperación de la información recibe desde el exterior por parte del entorno responsable de la producción del conocimiento. Las distintas ramas de la ciencia y las diversas actividades humanas actúan en los principales procesos de creación del conocimiento sobre la base de: la especialización, el aislamiento y la simultaneidad. Por esta razón las principales dificultades de la TRI son de carácter intrínseco a los procesos de análisis de la información, pero también existen dificultades que atañen a cuestiones relacionadas con la cantidad y diversidad de la producción de conocimiento.

Principios de fertilidad (PF) en la Teoría de la Recuperación de la Información:

(1) La literatura de las diversas especialidades científicas se desarrolla de manera independiente una de las otras, a pesar la las influencias inherentes en la producción de conocimiento. Además, no existen las conexiones lógicas entre las diversas literaturas producidas.
(2) A causa de la incomunicación entre las especialidades científicas las distintas literaturas tienden al aislamiento y por tanto no incorporan el bagaje de unas y otras. Por lo tanto, no solamente se excluyen sino que ni se conocen.
(3) Sin embargo, el uso de los sistemas de información electrónica permite detectar una tímida conexión entre diversas disciplinas que supone el inicio de la corrección en las prácticas parametrizadas en los dos principios anteriores. El rasgo esencial de esta inflexión en el aislamiento y la simultaneidad se ha producido aparentemente por las fáciles condiciones de acceso a la información que ofrecen los nuevos sistemas y servicios. Y induce a pensar en la posibilidad del contacto y la convergencia entre las literaturas especializadas hasta ahora aisladas y simultáneas.

Estos 'principios de fertilidad' (PF) que ponen de manifiesto la fragmentación de la producción de conocimiento también son causa y efecto en el *corpus* teórico de la TRI. La fragmentación y el aislamiento de los procesos cognitivos inciden en la expresión del contenido informativo de la literatura especializada, siendo responsables, en gran medida, de las limitaciones extrínsecas de los procesos de análisis y recuperación de la información.

Otro rasgo esencial de las limitaciones de la recuperación de la información consiste en el escenario que ofrece la evolución del comportamiento de los usuarios respecto a la Internet y los sistemas de acceso a los recursos de información. El análisis de los usuarios de la Internet se basa en las tres dimensiones básicas de los estudios de usuarios: (1) Las necesidades de información (actividad de los usuarios, uso de la información, tipo de información utilizada); (2) El comportamiento y el manejo de las técnicas documentales (tiempo dedicado a la búsqueda de información, métodos y técnicas utilizadas, fuentes de información, etc.); (3) La satisfacción de las necesidades de información en función de los criterios de calidad (accesibilidad, rapidez, exhaustividad, personalización de la información, etc.). Pero, resulta fácil imaginar que los nuevos sistemas de acceso propuestos desde la Internet, han variado notablemente las necesidades de información y el comportamiento de los usuarios. Dentro de esta evolución, tres son los ejes que dibujan los cambios:

(1) La Internet ofrece al usuario un acceso directo a la información. Así pues, en una situación óptima el usuario puede acceder por si mismo a los recursos de información que él mismo defina en su búsqueda. El inconveniente consiste en la desaparición de los intermediarios físicos (profesionales) y en que el usuario está solo ante toda la información disponible y las interfaces creadas para la gestión de la recuperación.
(2) La Internet facilita el acceso inmediato a una gran cantidad de información y de documentación con una cobertura geográfica internacional. Por lo tanto, el usuario percibe la Internet como un

sistema de información gigantesco frente a la limitación de los sistemas de información tradicionales. La desventaja radica en que el usuario puede perderse entre el volumen inmenso de información.

(3) El usuario de la Internet se convierte en el responsable de su búsqueda informativa y de todos los procesos de obtención de la información. De esta manera, el usuario pasa de ser un usuario final de la información a un usuario profesional de la información, porque establece una conducta dependiente no sólo de la información, sino también de los métodos y técnicas de acceso a la información.

En resumen, la Internet ofrece a los usuarios una mayor autonomía y una participación directa en la formulación de las necesidades de información, pero en contraposición la Internet plantea a los usuarios problemas de orientación, de exhaustividad, de organización, de garantía mínima sobre la calidad de la información, de capacidad para definir la búsqueda, de tiempo disponible para dedicarlo a la recuperación de la información, etc. Esta evolución entre el acceso a la información y la conducta de los usuarios ha situado el total de las ventajas e inconvenientes de la Internet casi en tablas, y ha favorecido la aparición de estrategías correctoras en forma de sistemas avanzados de recuperación de la información (agentes inteligentes) y de servicios personalizados de difusión de la información (sistemas *push* y otras aplicaciones).

III.- LA RECUPERACIÓN DE LA INFORMACIÓN Y LOS SOFTWARES DE GESTIÓN

Los softwares documentales con relación a la recuperación de la información actualmente deben cumplir una doble función [[vi]]:

(1) Gestionar los aspectos documentales internos de cada sistema (gestión de thesaurus, métodos de búsqueda avanzados, difusión selectiva de la información).

(2) Incorporar una nueva oferta que la Internet ha consagrado como esencial: el entorno Web.

En este breve espacio es del todo imposible abordar con amplitud todos los aspectos de los sistemas de recuperación de información. Pero, es importante tener presente que al diseñar e instalar un sistema de recuperación de la información no puede obviarse los efectos negativos que se desprenden del marco conceptual de la TRI, y asimismo deben ser, o bien asumidas, o bien suavizadas las limitaciones intrínsecas y extrínsecas de los procesos de representación y de recuperación del conocimiento que hemos expuesto anteriormente.

El mercado de la indústria de la información electrónica y los paquetes de programas disponibles señala las tendencias siguientes en la gestión del conocimiento y en la recuperación de la información:

(1) Casi todos los sistemas de gestión documental y gestión bibliotecaria han incorporado la doble función: gestión de la recuperación de la información de cada sistema, y el módulo Web.

(2) La generalización del entorno Web tiende a envolver todas las funciones documentales, y en estos momentos es una tecnología usada como herramienta de recuperación y difusión de la información en todos los paquetes de programas con módulo Web. Pero, frente a las prestaciones desarrolladas de cara al usuario faltan las prestaciones en aquellos subsistemas de gestión interna de la información. Las presiones profesionales hacen prever que las interfaces inteligentes de la última generación de programas de gestión están siendo diseñadas sobre el entorno web.

(3) En un porcentaje que no supera el 15% de los paquetes de programas existentes en el mercado, los constructores han incorporado un módulo de tratamiento lingüístico que refuerza el módulo de gestión de thesaurus (muy denostado en algunos paquetes de programas por su falta de integración con los restantes subsistemas de tratamiento de la información).

(4) El módulo de gestión de thesaurus ha sido incorporado en un porcentaje que supera el 50% de los paquetes de programas existentes en el mercado, y a su vez han incorporado todos ellos la visualización del thesaurus (alfabética, sistemática por dominios, y permutada).

(5) La gestión de la información y la recuperación de la información en grandes organizaciones e instituciones favorecen los proyectos intranet con el apoyo del entorno Web.
(6) El módulo Web instalado como tecnología que apoya la recuperación de la información en la totalidad de los paquetes de programas puede incorporar, dependiendo de cada constructor, las características funcionales siguientes:
   - Truncamientos (derecha, izquierda, central y simultáneos).
   - Máscaras.
   - Operadores de proximidad.
   - Operadores de cadenas de caracteres (*searching substring*).
   - Ponderación de la pertinencia de la información a partir de los criterios establecidos en cada contexto por el usuario según las funciones semánticas definidas en el módulo de tratamiento lingüístico.
   - La potencia de los sistemas con álgebra booleana ha sido criticada dado que la teoría de conjuntos necesita de un nivel de conceptualización que no todos los usuarios están en disposición de asumir.
   - Desarrollo completo de la tecnología hipermedia.
   - La gestión del texto completo ha creado motores de indización que permiten funciones de recuperación por proximidad, semejanza, tratamiento lingüístico (morfología, fonética, sintaxis y semántica) gracias a la creación de diccionarios, y cálculos de pertinencia.
   - Gestión del multilingüismo sobre la base de la duplicación de índices.
   - Difusión selectiva de la información por perfil, y por funciones de servicio *push*.

La Internet y su entorno Web han favorecido la popularización del acceso a la información mediante los sistemas de recuperación de la información (básicos y avanzados), también han fomentado la competitividad entre los constructores de programas, y sobretodo han puesto en evidencia la relación entre la representación del conocimiento (indización) y la recuperación de la información. Evaluar los sistemas de recuperación de la información en la Internet y en los entornos web en función de las encuestas de calidad (*benchmark*) permite conocer –sobre todo si se elaboran unos criterios precisos- como se están integrando en los softwares de gestión las investigaciones lingüísticas y las premisas de la TRI [[vii]].

CONCLUSIÓN

Los investigadores de la recuperación de la información tienen ante sí un reto: crear sistemas que permitan explotar la producción del conocimiento igualando la representación del contenido de la información frente a la representación de las necesidades de información.

Notas

[i] Mooers, C.N. Coding, Information Retrieval, and Rapid Selector. *American Documentation,* 1(4), 1950, 225-229.

[ii] Mooers, C.N. Comments on the Paper by Bar-Hillel. *American Documentation,* 8(2), 1957, 114-116.
Bar-Hillel, Y. A Logician's Reaction to Recent Theorizing on Information Search Systems. *American Documentation*, 8(2), 1957, 103-113.

[iii] Lancaster, F.W., A.J. Warner. *Information Retrieval Today.* Arlington: Information Resources Press, 1993.

[iv] Swanson, D.R. Historical Note: Information Retrieval and the Future of an Illusion. *Journal of the American Society for Information Science*, 39(2), 1988, 92-98.

[v] Swanson, D.R. Historical Note: Information Retrieval and the Future of an Illusion. *Journal of the American Society for Information Science*, 39(2), 1988, 92-98.

[vi] Esta parte del trabajo corresponde a los primeros datos que la autora tiene como parte de un grupo de investigación sobre sistemas de recuperación y software de gestión de la Universidad de Barcelona, y las premisas ofrecidas en esta ponencia son el resumen del primer nivel de análisis llevado a cabo sobre la base de la realización de encuestas de calidad.

[vii] Wiggins, R., J.A. Matthews. *Plateaus, Peaks, and Promises: The Infonortics '98 Search Engines Conference*. 1998. http://www.infotoday.com/searcher/jun/story4.htm (visitado el 19 de enero de 1999).