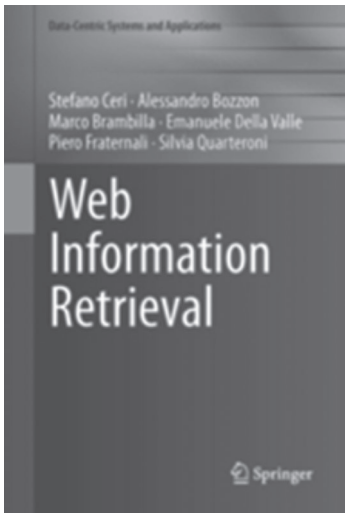

CERI, Stefano, BOZZON, Alessandro, BRAMBILLA, Marco, DELLA VALLE, Emmanuel, FRATERNALI, Piero, QUARTERONI, Silvia. *Web Information Retrieval.* Springer, 2013. 284 p. Serie Data-Centric Systems and Applications; versiones electrónica e impresa.

La recuperación de información se desarrolló en la comunidad bibliotecaria mucho antes de la proliferación de las computadoras, la cual cobró gran auge con el advenimiento de la World Wide Web y los grandes protagonistas como Google y Yahoo, a principios de este siglo. Las capacidades de búsqueda están integradas actualmente en la mayoría de los sistemas de información de diversa índole, incluyendo la tecnología móvil; por lo tanto, aprender la tecnología de recuperación de información es primordial para los estudiosos y practicantes de las disciplinas de computación y bibliotecología.

Este libro está orientado a la enseñanza de esta importante tecnología de información y fue producto de un curso en el Politécnico de Milán, Italia, sobre esta temática, por lo que surgió la necesidad de crear un libro de texto que posteriormente evolucionó en un libro muy completo sobre el tema que incluye ejercicios al final de cada capítulo. Además, ofrece diapositivas de las presentaciones e información adicional en el sitio www.search-computing.org.

La obra consta de 3 partes:

- En la primera parte se estudian los principios de la recuperación de información y la métrica clásica como la precisión y la exhaustividad, los métodos para procesar e indizar la información textual, los modelos para responder a preguntas, la clasificación y el agrupamiento de documentos y el procesamiento en lenguaje natural. Esta parte sienta las bases para entender mejor su aplicación en la Web.
- En la segunda se abordan los aspectos fundamentales de la recuperación de información en la Web y se discute la arquitectura general de los buscadores, en particular, los procesos de recopilación de páginas web y el indizado de las mismas, los métodos de análisis de enlaces, la recomendación y diversificación como dos aspectos importantes de la presentación de resultados y –por último– discute la publicidad en los buscadores, que es el principal medio por el cual éstos obtienen sus ingresos.
- En la tercera y última parte se describen los aspectos avanzados de la búsqueda en la Web, empezando por un panorama actual de la búsqueda de información en ese ambiente, la forma en que se publican los datos en esta plataforma para que sean útiles para los buscadores, la metabúsqueda, la búsqueda semántica, que va a revolucionar la forma de buscar y recuperar información, así como en el contexto de multimedios. Por último

aborda la computación humana y búsqueda colectiva, que complementan los resultados de la búsqueda mediante la interacción humana.

A continuación se describe un poco más en detalle cada una de las partes y los capítulos que las conforman.

La primera parte, *Principios de la recuperación de información,* consta de los siguientes capítulos:

- **Una introducción a la recuperación de información,** que empieza con la definición e historia de la recuperación de información y la importancia de definir la relevancia y el manejo de grandes colecciones de datos no estructurados; además, las tareas típicas como filtrado, elaboración de resúmenes, agrupamiento y categorización de documentos, sistemas de preguntas y respuestas y sistemas recomendantes. Termina con la evaluación de un sistema de recuperación de información y sus dos parámetros básicos: exhaustividad y precisión, así como la recuperación jerarquizada y las normas técnicas que se han utilizado al respecto.
- **El proceso de recuperación de información** inicia con un panorama de este proceso, la visualización lógica de documentos y el proceso de indizado. Continúa con más detalle sobre la recuperación en textos, las operaciones textuales tales como partición, análisis lexicográfico, eliminación de palabras de parada, detección de frases, truncamiento y ponderación. Posteriormente, aborda las principales leyes sobre la recuperación de textos, y por último describe las estructuras de datos y diversas técnicas para crear índices.
- **Modelos de recuperación de información** describe tres modelos clásicos de recuperación de información: booleano, espacio de vectores y probabilístico, los compara y muestra sus similitudes. En el modelo booleano, además de evaluarlo, analiza sus capacidades y limitaciones; lo mismo sucede con los modelos de espacio de vectores y probabilístico, aplicando fórmulas matemáticas para cada uno de ellos.
- **Clasificación y agrupamiento** aborda el problema de la sobrecarga de la información y su posible solución con dos técnicas automatizadas: clasificación y agrupamiento; la primera requiere de supervisión mientras la segunda no. Se describen varias técnicas de clasificación así como de agrupamiento y etiquetado de grupos. Por último, presenta varios escenarios para aplicar dichas técnicas tales como manejo de resultados de la búsqueda y agrupamiento en bases de datos.
- **Procesamiento de lenguaje natural para la búsqueda,** señala que los datos no estructurados constituyen más del 80% de los documentos digitales y están redactados en lenguaje natural, por lo que su manejo formal es muy importante. Se presentan los desafíos del procesamiento en lenguaje natural, la ambigüedad y la probabilidad de los términos; posteriormente se describen modelos para abordar este problema, los sistemas de preguntas y respuestas y la importancia de la semántica en la representación de textos, así como la jerarquización de respuestas.

La segunda parte, *Recuperación de información en la web,* consta de los siguientes capítulos:

- **Buscadores.** Se discuten los desafíos en el diseño e implementación de los buscadores que respondan a preguntas basadas en palabras clave y que arrojen sitios web relevantes como resultado de la búsqueda. Se reseña una breve historia de los buscadores, su arquitectura y componentes, el proceso de recopilación de sitios web, la resolución de URLs, la eliminación de duplicados, distribución de resultados, entre otros temas. Continúa con el indizado y sus diferentes tipos y termina con el uso de técnica de "caching" para optimizar las búsquedas.
- **Análisis de enlaces.** Resalta la importancia de la jerarquización de resultados con base en diversas técnicas para medir la relevancia de las páginas web, conociendo su estructura de hiperenlances. Describe el gráfico web que representa la estructura antes mencionada, la jerarquización con base en enlaces, el concepto de PageRank, desarrollado por Google, y el sistema HITS (búsqueda de tópicos inducidos por hipertexto). Termina recalcando la necesidad de utilizar el análisis de enlaces.
- **Recomendación y diversificación en la Web.** La efectividad de la búsqueda en la Web puede ser afectada por la existencia de recursos similares, solicitudes ambiguas y los diversos objetivos de los usuarios ante la misma búsqueda, por lo que los sistemas de recomendación y técnicas de diversificación pueden ser la solución. Se describen los factores que producen sobrecarga de información, los sistemas de recomendación como definición de perfiles de usuarios, recomendación con base en contenidos y filtrado colaborativo. Por último, describe la diversificación de resultados con el fin de reducir la redundancia, su cobertura, criterios de diversidad, el balance entre ésta y la relevancia, y la diversificación en varios dominios.
- **Publicidad en la búsqueda.** Este capítulo provee los conceptos fundamentales de la publicidad en línea y la publicidad en la búsqueda, describiendo las principales estrategias, tales como publicidad de marca y mercadotecnia directa. Presenta la terminología básica y los modelos económicos al respecto, en particular las subastas.

La tercera y última parte, *Aspectos avanzados de la búsqueda en la web*, consta de los siguientes capítulos:

- **Publicación de datos en la Web.** Presenta un panorama histórico de las opciones que han aparecido para publicar datos en la Web, las características de la Web profunda, las interfaces de aplicaciones (API), microformatos, el RDF (Resource Description Format) y los datos enlazados. Concluye comparándolos entre sí y atisbando un poco hacia el futuro.
- **Metabúsqueda y búsqueda en dominios múltiples.** Analiza el potencial de los sistemas basados en la tecnología de integración de datos que permiten el uso de interfaces sencillas para tipos específicos de recursos.

Describe la metabúsqueda y la búsqueda en dominios múltiples, el uso de identificadores de objetos (OID), el problema de asignación de atributos a objetos, la búsqueda exploratoria y la visualización de datos.

- **Búsqueda semántica.** Describe la búsqueda semántica basada en significados, los diversos modelos semánticos y su construcción, los recursos y preguntas desde el punto de vista del sistema y el usuario, el proceso que han adoptado los buscadores semánticos y algunas soluciones comerciales y académicas
- **Búsqueda de multimedios.** Se discuten los desafíos que afrontan los sistemas de multimedios, sus requerimientos y aplicaciones en diversas áreas de la actividad humana, la adquisición, normalización, indizado y consulta de contenidos, la arquitectura de un sistema de información de multimedios, el proceso de búsqueda, el uso de metadatos. Termina describiendo algunos proyectos de investigación y sistemas comerciales al respecto.
- **El proceso de búsqueda e interfaces.** Este capítulo describe el proceso de búsqueda de información y presenta los principales modelos al respecto, así como los componentes de las interfaces para los usuarios en lo que se refiere a la especificación de preguntas y presentación de resultados utilizados por los principales buscadores. Termina con la descripción de búsqueda por facetas.
- **Computación humana y búsqueda colectiva.** Describe esta disciplina que intenta armonizar la contribución de humanos y computadoras para la resolución de problemas complejos, mediante la asignación distribuida de actividades a una comunidad. Analiza diversas aplicaciones como juegos con propósito, uso colectivo de recursos (crowdsourcing), recopilación de datos. Posteriormente, analiza el marco conceptual y las fases que lo constituyen, con algunos ejemplos. Termina abordando diversas aplicaciones, desafíos y proyectos específicos, así como algunos aspectos que quedan por resolver.

Para concluir, este libro describe de una forma clara y concisa los diversos aspectos de la recuperación de información en la Web, aun cuando en algunas secciones aborda temas teóricos y de alto contenido matemático. Es una obra didáctica y a la vez cuenta con una amplia bibliografía y recursos adicionales que permitirán al lector especializarse en este tema tan importante y tener las bases para desarrollar sistemas de recuperación de información útiles y eficientes.

Federico Turnbull Muñoz
Subdirección de Servicios de Información Especializada
Dirección General de Bibliotecas - UNAM