Módulo 1

1.1 INTRODUCCIÓN

Por Óscar Cordón García

Catedrático de Universidad. Instituto Andaluz Interuniversitario en Data Science and Computational Intelligence (DasCI). Universidad de Granada

---

1. INTRODUCCIÓN A LA RECUPERACIÓN DE INFORMACIÓN

Los avances tecnológicos de los últimos sesenta años han provocado un aumento exponencial de la información. El proceso de digitalización y la transformación de documentos que se está llevando a cabo son dos claros ejemplos de la revolución de la información, la cual ha permitido su acceso a un número ilimitado de usuarios.

A ello hay que sumarle las grandes velocidades y la facilidad de distribución de información mediante las llamadas "autopistas de la información", la proliferación de las conexiones de fibra y el coste cada vez menor de los medios de almacenamiento. Todo ello nos sitúa dentro de un entorno en desarrollo de información electrónica a la que se puede acceder por medios automáticos en el marco global de la transformación digital. Otro aspecto que tenemos que considerar es la diversificación de los medios, que trae consigo una mayor cantidad de información multimedia: imagen, sonido, texto, vídeos, etc. muy extendida y fácilmente accesible hoy en día a través de Internet.

La Recuperación de Información (RI) se puede definir como el problema de la selección de información desde un mecanismo de almacenamiento en respuesta a consultas realizadas por un usuario. Los Sistemas de Recuperación de Información (SRI) son una clase de sistemas de información que tratan con bases de datos compuestas por diferentes tipos de objetos de información (documentos textuales, artículos, imágenes, audios, vídeos, etc.) y procesan las consultas de los usuarios permitiéndoles acceder a la información relevante en un intervalo de tiempo apropiado (véase la Figura 1).

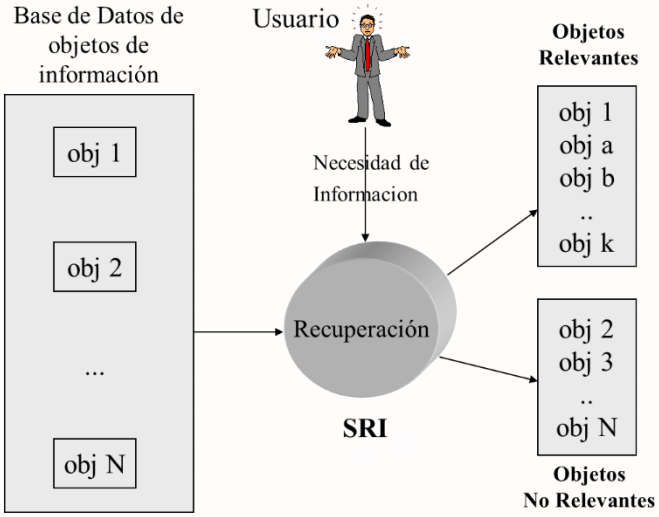

Figura 1: Proceso de recuperación de información

Un SRI permite la recuperación de la información, previamente almacenada, por medio de la realización de una serie de consultas a los objetos de información contenidos en la base de datos. Estas preguntas son sentencias formales de expresión de necesidades de información y suelen venir formuladas por medio de un lenguaje de consulta.

Un SRI debe soportar una serie de operaciones básicas sobre los objetos de información almacenados, como son: introducción de nuevos objetos, modificación de los ya almacenados y eliminación de los mismos. Debemos también contar con algún método de localización de los documentos (o con varios, generalmente) para presentárselos posteriormente al usuario, generalmente de forma ordenada con respecto a su interés para el mismo. Este proceso se resume gráficamente en la Figura 1.2. Los SRI implementan estas operaciones de varias formas distintas, lo que provoca una amplia diversidad en lo relacionado con la naturaleza de los mismos.

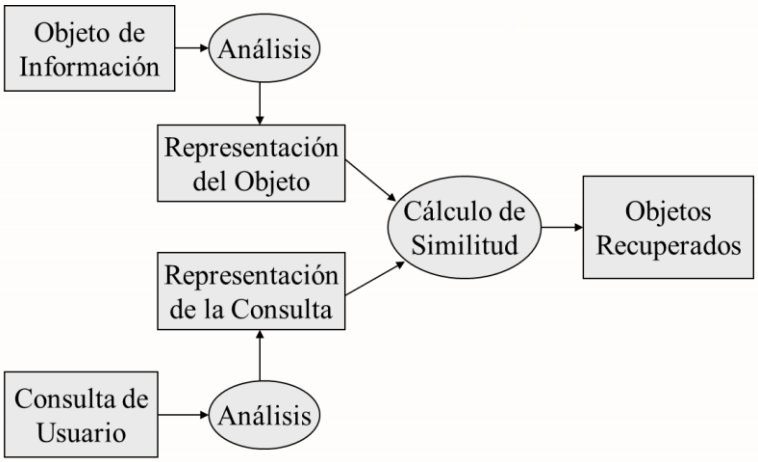

Figura 1.2: Operaciones para la recuperación de documentos

Hoy en día, los ejemplos más populares de SRIs son los motores de búsqueda en Internet, tales como *Google*, *Yahoo!* o *Bing*, pero existen mucho más tales como SRI bibliográficos que permiten realizar búsquedas de artículos científicos, como *Elsevier Scopus* o *Clarivate-Analytics Web of Knowledge*, SRI médicos, catálogos de bibliotecas, etc. Se estudiarán varios de estos SRIs en los módulos del Bloque II: "Dónde buscar información".

2. COMPONENTES DE UN SISTEMA DE RECUPERACIÓN DE INFORMACIÓN

Un SRI está formado por cuatro componentes principales: la base de datos, el subsistema de consultas, el mecanismo de emparejamiento o evaluación, y la interfaz (Figura 1.3).

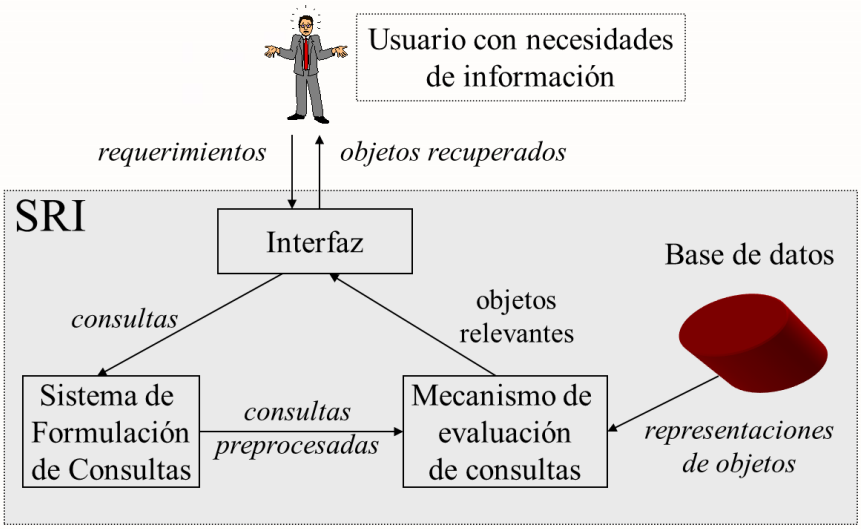

Figura 1.3: Composición genérica de un Sistema de Recuperación de Información

La **base de datos** almacena los objetos de información. Tradicionalmente se han llamado documentos y eran de naturaleza textual aunque la evolución tecnológica ha propiciado la profusión de documentos multimedia, incorporándose al texto fotografías, ilustraciones gráficas, vídeos animados, audio, etc. Estos objetos no se introducen directamente en el SRI, sino que están representados por unos elementos llamados *descriptores* obtenidos a través de un proceso de *indización*. La razón de ser de estos descriptores se basa en darle una mayor eficiencia a la base de datos, la cual será más pequeña, provocando que el tiempo de búsqueda en ella sea mucho menor. La indización más simple consiste en asociar, normalmente de forma automática, una serie de términos índice a cada objeto de información de forma que describan su contenido. Desde un punto de vista matemático, la base de datos sería una tabla o matriz de ceros y unos en la que cada columna indica las asignaciones de un determinado descriptor y

cada línea o fila representa un objeto. Existen otros mecanismos más avanzados que se estudiarán en el módulo 2 del curso "El proceso de búsqueda".

El **subsistema de consultas** toma la consulta escrita por el usuario, estudia su validez y la procesa para ponerla en una forma estándar que pueda ser empleada por el SRI. Para llevar a cabo esta tarea, incluye un *lenguaje de consulta* que recoge todas las reglas para generar consultas válidas y la metodología para seleccionar las partes relevantes. Un analizador sintáctico comprueba la validez de la consulta (en caso de no ser válida, el sistema devolverá al usuario un mensaje de error) y la procesa para ponerla en la misma estructura que los objetos de información de la base (*la indiza*). Después la envía al mecanismo de evaluación para determinar qué objetos se consideran relevantes para las necesidades de información que representa. Se profundizará en el estudio de los lenguajes de consulta en los módulos 2 ("El proceso de búsqueda") y 3 ("Las ecuaciones de búsqueda y los operadores").

El **mecanismo de evaluación** selecciona los objetos de información que se consideran relevantes, de entre los que forman la base de datos, de acuerdo con los criterios de nuestra consulta. Llegados a este punto, tenemos una representación del contenido de los objetos en nuestra base documental y también una representación de las consultas que queremos realizar. Este subsistema compara la representación de la consulta con las de todos los objetos para evaluar el grado en el que cada objeto satisface los requisitos expresados en la consulta y recupera aquellos documentos que son relevantes a la misma. Este grado es lo que se denomina RSV (*Retrieval Status Value o Valor de Estado de Recuperación*). Existen una gran cantidad de *modelos de recuperación de información* distintos en función de cómo se realizan estas comparaciones y varios de ellos se estudiarán en el curso en módulos como el 2 ("El proceso de búsqueda") y el 3 ("Las ecuaciones de búsqueda y los operadores").

Finalmente, la **interfaz** permite al usuario formular sus consultas al SRI y visualiza las respuestas proporcionadas por el sistema, una vez procesada su consulta. Ofrece facilidades al usuario a la hora de formular su consulta, ya que éste no tiene por qué saber exactamente el funcionamiento tanto externo como interno del sistema. En general, la respuesta del sistema ante la consulta del usuario, es decir, la lista de objetos de información que se consideran relevantes para la misma, se muestran ordenados de mayor a menor relevancia (RSV).

3. MEJORAS EN LA RECUPERACIÓN: RETROALIMENTACIÓN POR RELEVANCIA

Por muy completo que sea un SRI, por regla general tiene una carencia destacable, la *exhaustividad* (capacidad de presentar todos los objetos de información de la base de

datos que son relevantes para la necesidad de información del usuario). Los usuarios pueden recuperar algunos objetos relevantes como respuesta a sus consultas, pero casi nunca recuperan todos los objetos relevantes relacionados con las mismas. Existen casos en que esto no tiene mucha importancia para los usuarios pero en aquellas ocasiones en las que la exhaustividad es un parámetro crítico, hay algunas formas de recuperar más documentos relevantes que los que se recuperaron en un principio.

Una alternativa para resolver este problema es que el SRI ofrezca al usuario la posibilidad de modificar la consulta original mediante **retroalimentación por relevancia**. Este concepto asume que a la formulación de la consulta óptima se llega mediante una serie de interacciones entre el sistema y el propio usuario. En las operaciones de RI, la mayoría de los usuarios, que no conocen los detalles de la estructura de la base y del lenguaje de consulta del SRI, encuentran difícil formular una consulta adecuada para el propósito de recuperación que tienen. Esto sugiere que la primera operación de recuperación debe verse como un tanteo, como una ejecución de prueba cuyo objetivo es recuperar algunos elementos útiles de la base de datos. Estos elementos inicialmente recuperados pueden examinarse para determinar su relevancia y, con el conocimiento adquirido, se puede redefinir progresivamente la consulta, obteniendo una nueva y mejorada, con la aspiración de recuperar objetos de información adicionales útiles en las siguientes búsquedas.

La idea principal consiste en la elección de términos importantes ligados a ciertos objetos que previamente se han identificado como relevantes por el usuario para incluirlos o incrementar su importancia en la nueva formulación de la consulta. Análogamente, se pueden eliminar o disminuir la importancia de los términos incluidos en documentos no relevantes previamente recuperados en la nueva consulta.

4. TAREAS FUNDAMENTALES EN UN SISTEMA DE RECUPERACIÓN DE INFORMACIÓN

A la vista de lo desarrollado en las secciones anteriores y a modo de resumen, las tareas fundamentales asociadas a un SRI son las siguientes:

1. Cómo representar los objetos de información (documentos) en la base de datos.
2. Cómo representar las necesidades de información de los usuarios en forma de consultas.
3. Cómo evaluar la satisfacción de una necesidad de información (consulta) por un objeto de información.
4. Cómo presentar los resultados de la consulta al usuario.
5. Cómo refinar los resultados de una consulta previa.

5. EVALUACIÓN DE LOS SISTEMAS DE RECUPERACIÓN DE INFORMACIÓN

Un SRI puede evaluarse empleando diversos criterios tales como: ejecución eficaz (**eficacia y eficiencia**), almacenamiento correcto y recuperación efectiva (**efectividad**). La importancia relativa de estos factores debe decidirla el diseñador del sistema, y la selección de la estructura de datos y los algoritmos apropiados para su implementación dependerá de esa decisión.

La *eficacia en la ejecución* se medirá por el tiempo que toma el sistema para llevar a cabo una operación. Este parámetro ha sido siempre una preocupación principal en un SRI, especialmente desde que muchos de ellos son interactivos y un tiempo de recuperación excesivo interfiere con la utilidad del sistema, llegando a alejar a los usuarios del mismo. La *eficiencia del almacenamiento* se medirá por el número de bytes que se precisan para almacenar los objetos de información (tanto sus representaciones indizadas como el objeto en sí).

Tradicionalmente, se le ha dado mucha importancia a la *efectividad de la recuperación*, normalmente basada en la relevancia de los documentos recuperados a las necesidades reales de información del usuario, lo cual ha representado un problema ya que medir la relevancia es un proceso subjetivo y sin confianza. Las dos medidas principales son la *exhaustividad*, o habilidad del sistema para presentar todos los objetos de información relevantes, y la *precisión*, o habilidad del sistema para presentar solamente objetos relevantes en la lista de objetos de información recuperados.

REFERENCIAS BIBLIOGRÁFICAS

- **Baeza-Yates, R., Ribeiro-Neto, B**., Modern Information Retrieval: The Concepts and Technology behind Search, Addison-Wesley, 2ª edición, 2010. ISBN: 978-0321416919.

- **Cacheda, F., Fernández-Luna, J.M., Huete, J**. Recuperación de Información: Un Enfoque Práctico y Multidisciplinar. Rama, 2011. ISBN: 978-8499641126.

- **Croft, B., Metzler, D., Strohman, T.** Search Engines: Information Retrieval in Practice, Addison Wesley/Prentice Hall, 2010. ISBN: 978-0136072249.

- **Salton, G.**, Automatic Text Processing: The transformation, Analysis and Retrieval of Information by Computer, Addison-Wesley, 1989. ISBN: 978-0201122275.

- **Zhai, C.X., Massung, S.** Text Data Management and Analysis. A Practical Introduction to Information Retrieval and Text Mining. ACM & Morgan Claypool Pubs, 2016. ISBN: 978-1970001198.

Referencias Adicionales

- **Goker, J. Davies**. Searching in the 21st century. Wiley. 2009.

- **Salton, G., McGill, M.J**., An Introduction to Modern Information Retrieval, McGraw-Hill, 1983.

- **Van Rijsbergen, C.J**., Information Retrieval, Butterworth, 1979.