Recuperación de información: un área de investigación en crecimiento

MsC. Fernando Bordignon
MsC. Gabriel Tolosa Chacón

RESUMEN

A partir de la expansión y consolidación de Internet, como medio principal de comunicación electrónica de datos, se ha puesto a disposición de casi toda la humanidad una importante cantidad de información de todo tipo. A los efectos de aprovechar este potencial de información, es necesario poseer accesos que permitan que la tarea de recuperación sea efectiva y eficiente en términos de recursos invertidos por los usuarios. Este artículo plantea cual es el objeto de estudio del área denominada «recuperación de información», en qué estado se encuentra y cuáles son sus principales líneas de trabajo.

Palabras clave: internet, recuperación de información, investigación

ABSTRACT

From the expansion and consolidation of Internet, as a main means of electronic communication of data, an important amount of information of all types has been made available for almost all humans. For the purpose of taking advantage of this information stock, it is necessary to have accesses that may allow the recovery task to be effective and efficient in terms of resources invested by the users. This article states what the object of study of the area called «information retrieval» is, in what condition it is, and what its main work lines are.

Keywords: Internet, information retrieval, research

El entorno de la recuperación de información

Históricamente, el hombre ha necesitado de medios sobre los cuales representar todo acerca del mundo que lo rodea y de reflejar –de alguna manera– su evolución. La escritura ha sido el mecanismo «tradicional» y fundamental que soporta su conocimiento en el tiempo.

Esta misma evolución ha facilitado la existencia de diferentes medios de representación de la escritura y llega hasta nuestros días. Hoy la información puede representarse digitalmente, almacenarse, y distribuirse masivamente en forma simple y rápida, a través de redes de computadoras. La digitalización abrió nuevos horizontes en las formas en que el hombre puede tratar con la información que produce.

De igual manera, el volumen de información disponible crece permanentemente y adquiere diferentes formas de representación, desde simples archivos de texto en una computadora personal o un periódico electrónico, hasta librerías digitales y espacios mucho más grandes y complejos como la web. Algunos investigadores han planteado que -desde hace varios años- existe un fenómeno denominado «sobrecarga de información» [1], debido a que el volumen y la disponibilidad hacen que los usuarios no cuenten con suficiente tiempo físico para «procesar» todo el cúmulo de medios a su alcance [2].

Resulta importante tratar con toda esa información

disponible electrónicamente para que pueda servir a diferentes personas (usuarios) en diferentes situaciones. Esto plantea un desafío interesante: hay importantes volúmenes de información y hay usuarios que se pueden beneficiar con la posibilidad de acceder a ésta, por lo tanto, ¿cómo poder unir preguntas con respuestas, necesidades de información con documentos, consultas con resultados? En las ciencias de la computación existe un área, la Recuperación de Información (Information Retrieval), que estudia y propone soluciones al escenario presentado, al plantear modelos, algoritmos y heurísticas.

La Recuperación de Información (RI) no es un área nueva, sino que se viene desarrollando desde finales de la década de 1950. Sin embargo, en la actualidad adquiere un rol más importante, debido al valor de la información. Disponer o no de la información justa en tiempo y forma puede resultar en el éxito o fracaso de una operación. De aquí la importancia de los Sistemas de Recuperación de Información (SRI) que pueden manejar de manera eficaz y eficiente -con ciertas limitaciones- estas situaciones.

¿Qué se entiende concretamente por «Recuperación de Información»? Para Ricardo Baeza-Yates y otros [3] «la Recuperación de Información trata con la representación, el almacenamiento, la organización y el acceso a ítems de información».

Años antes, Salton [4] propuso una definición amplia que plantea que el área de RI «es un campo relacionado con la estructura, análisis, organización, almacenamiento, búsqueda y recuperación de información».

En las definiciones anteriores los elementos de información son no estructurados, tales como documentos de texto libre (por ejemplo, un archivo de texto que contenga La Biblia) o semi-estructurados, como lo son las páginas web.

Croft [5] estima que la recuperación de información es **«el conjunto de tareas mediante las cuales el usuario localiza y accede a los recursos de información que son pertinentes para la resolución del problema planteado. En estas tareas desempeñan un papel fundamental los lenguajes documentales, las técnicas de resumen, la descripción del objeto documental, etc**.» Por otro lado, Korfhage [6] definió la RI como **«la localización y presentación a un usuario de información relevante a una necesidad de información expresada como una pregunta».**

Ciertamente, es un área amplia, donde se abarcan diferentes tópicos, algunos computacionales -como el almacenamiento y la organización-; y otros relacionados con el lenguaje y los usuarios, como la representación y la recuperación propiamente dicha.

Nótese que Croft y Korfhage plantean explícitamente el rol del usuario como fuente de consultas y destinatario de las respuestas. Por lo tanto, de manera más genérica, podemos plantear que la recuperación de información intenta resolver el problema de «**encontrar y rankear documentos relevantes que satisfagan la necesidad de información de un usuario, expresada en un determinado lenguaje de consulta**». Sin embargo, existe un problema que dificulta sobremanera esta tarea y consiste en poder «compatibilizar» y comparar el lenguaje en que está expresada tal necesidad de información y el lenguaje de los documentos.

La problemática de la RI

De forma general -según Baeza-Yates [3]- el problema de la RI puede ser estudiado desde dos puntos de vista: el computacional y el humano. El primer caso tiene que ver con la construcción de estructuras de datos y algoritmos eficientes que mejoren la calidad de las respuestas. El segundo caso corresponde al estudio del comportamiento y de las necesidades de los usuarios.

Si se analiza el problema de la RI desde un alto nivel de abstracción (Figura 1) podemos establecer que:

– Existe una colección de documentos que contienen información de interés (sobre uno o varios temas)

– Existen usuarios con necesidades de información, quienes las plantean al SRI en forma de una consulta (en inglés, query. En adelante, ambas palabras se utilizarán indistintamente)

– Como respuesta, el sistema retorna -de forma ideal- referencias a documentos «relevantes», es decir, aquellos que satisfacen la necesidad expresada, generalmente en forma de una lista rankeada.

Planteamos que la respuesta «ideal» de un SRI está formada solamente por documentos relevantes a la consulta, pero -en la práctica- esta no es aún alcanzable. Esto se debe a que -entre otros motivos- existe el problema de compatibilizar la expresión de la necesidad de información y el lenguaje y de los documentos. Además, hay una carga de subjetividad

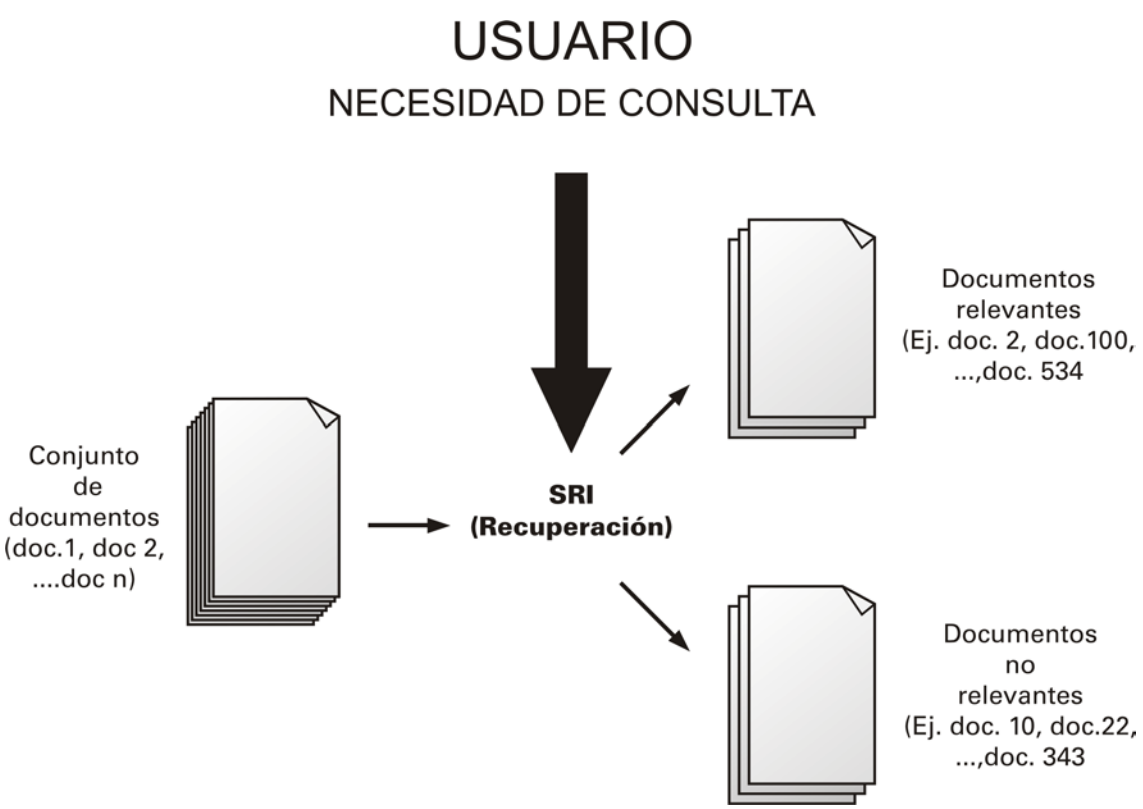

Fig. 1. La Problemática de la RI.

subyacente y depende de los usuarios. Entonces, el SRI recupera la mayor cantidad posible de documentos relevantes, minimizando la cantidad de documentos no relevantes (ruido) en la respuesta. En términos de eficiencia, se plantea la idea de ***precisión*** de la respuesta, es decir, cuando más documentos relevantes contenga el conjunto solución (para una consulta dada), más preciso será.

Para cumplir con sus objetivos, un SRI debe realizar algunas tareas básicas, las cuales se encuentran - fundamentalmente- planteadas en cuestiones computacionales, a saber:

– Representación lógica de los documentos y – opcionalmente – almacenamiento del original. Algunos sistemas solo almacenan porciones de los documentos y otros lo hacen de manera completa.

– Representación de la necesidad de información del usuario en forma de consulta.

– Evaluación de los documentos respecto de una consulta para establecer la relevancia de cada uno.

– Rankeo de los documentos considerados relevantes para formar el «conjunto solución» o respuesta.

– Retroalimentación o refinamiento de las consultas (para aumentar la calidad de la respuesta)

En la figura 2 se puede apreciar con mayor detalle la arquitectura básica de un SRI, el tratamiento de los documentos y la interacción con el usuario. Aquí se ven algunos componentes que no se habían mencionado hasta el momento.

Como podemos observar, se parte de un conjunto de documentos de texto, los cuales están compuestos por sucesiones de palabras que forman estructuras gramaticales (por ejemplo, oraciones y párrafos). Tales documentos están escritos en lenguaje natural y expresan ideas de su autor sobre un determinado tema. El conjunto de todos los documentos con los que se trata y sobre los que se deben realizar operaciones de RI se denomina **corpus**, **colección** o **base de datos textual o documental**. Para poder realizar operaciones sobre un corpus, es necesario obtener primero una **representación lógica** de todos sus documentos, la cual puede consistir en un conjunto de términos, frases u otras unidades (sintácticas o semánticas) que permitan – de alguna manera – caracterizarlos. Por ejemplo, la representación de los documentos mediante un conjunto de sus términos se la conoce como «bolsa de palabras» (bag of words).

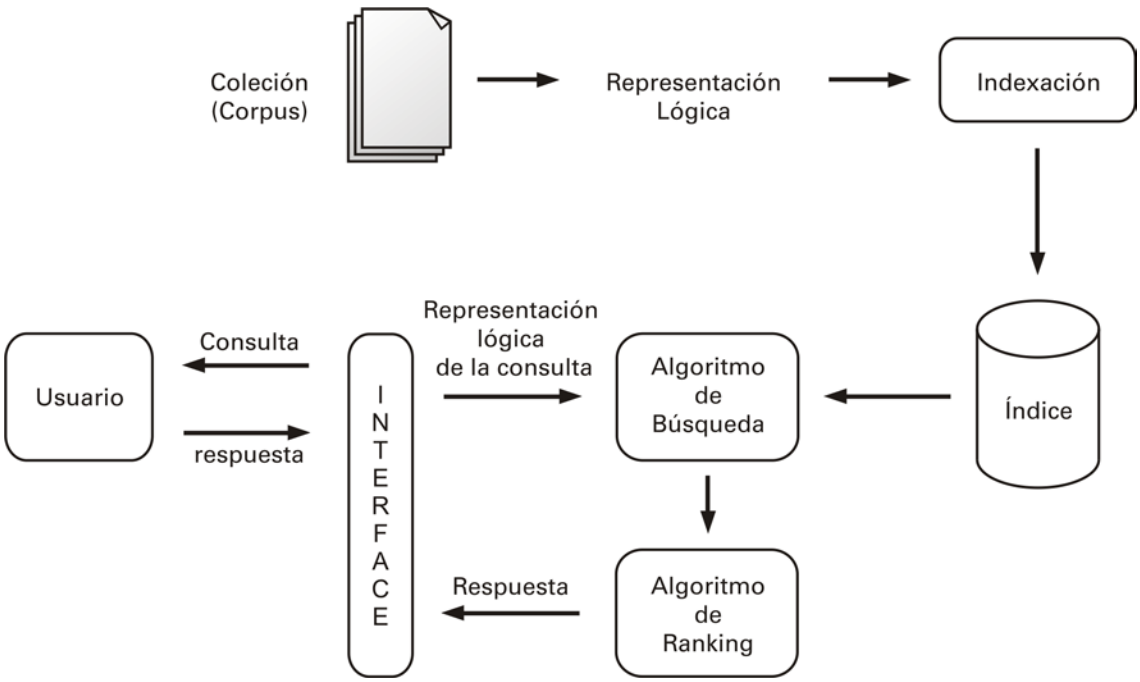

Fig. 2. Arquitectura básica de un SRI.

A partir de la representación lógica existe un proceso (**indexación**) que llevará a cabo la construcción de estructuras de datos (normalmente denominadas **índices**) que la almacene y soporte búsquedas eficientes. Es importante destacar que una vez construidos los índices, los documentos del corpus pueden ser eliminados del sistema, ya que este retornará las referencias a los mismos debido a que cuenta con la información necesaria para hacerlo. En tal caso, el usuario será el encargado de localizar el documento para consultarlo. A los sistemas que funcionan bajo este modelo se los denomina «sistemas referenciales», en contraste con los que sí almacenan y mantienen los documentos denominados «sistemas documentales» [7]. Un ejemplo de sistemas referenciales son algunos de los motores de búsqueda web, que retornan una lista de urls a los documentos, como – por ejemplo – Altavista (http://www.altavista.com/). Un caso particular es el motor de búsqueda Google (http://www.google.com/) el cual – en algunos casos – almacena en memoria caché el documento completo, el cual puede ser consultado durante cierto tiempo, incluso si ha desaparecido del sitio original.

El **algoritmo de búsqueda** acepta como entrada una expresión de consulta o query de un usuario y verificará en el índice cuáles documentos pueden satisfacerlo. Luego, un algoritmo de ranking determinará la relevancia de cada documento y retornará una lista con la respuesta. Se establece que el primer ítem de dicha lista corresponde al documento más relevante respecto a la consulta y así sucesivamente en orden decreciente.

La **interface** de usuario permite que este especifique la consulta mediante una expresión escrita en un lenguaje preestablecido y – además – sirve para mostrar las respuestas retornadas por el sistema.

Si bien hasta aquí se planteó la tarea básica de la RI y la arquitectura general de un SRI, el área es muy amplia y abarca diferentes tópicos. En general, un SRI no entrega una respuesta directa a una consulta, sino que permite localizar referencias a documentos que pueden contener información útil. Pero este es solo uno de los aspectos del área de RI en la actualidad, ya que se ha atacado el problema con una perspectiva más amplia, proponiendo y desarrollando estrategias y modelos para mejorar y aumentar la funcionalidad de los SRI. Entre otras, la RI abarca tópicos como:

– Modelos de Recuperación: La tarea de la recuperación puede ser modelada desde distintos enfoques, por ejemplo la estadística, el álgebra de boole, el álgebra de vectores, la lógica difusa, el procesamiento del lenguaje natural y demás.

– Filtrado y Ruteo: Es un área que permite la definición de perfiles de necesidades de información por parte de usuarios y ante el ingreso de nuevos documentos

al SRI, se los analiza y se lo reenvía a quienes se estima que van a ser relevantes.

– Clasificación: Aquí se realiza la rotulación automática de documentos de un corpus en base a clases previamente definidas.

– Agrupamiento (Clustering): Es una tarea similar a la clasificación pero no existen clases predefinidas. El proceso automáticamente determinará cuáles son las particiones.

– Sumarización: Área que entiende sobre técnicas de extracción de aquellas partes (palabras, frases, oraciones, párrafos) que contienen la semántica que determina la esencia de un documento.

– Detección de novedades (Novelty Detection): Se basa en la determinación de la introducción de nuevos tópicos o temas a un SRI.

– Respuestas a Preguntas (Question Answering): Consiste en hallar aquellas porciones de texto de un documento que satisfacen expresamente a una consulta, es decir, la respuesta concreta a una pregunta dada.

– Extracción de Información: Extraer aquellas porciones de texto con una alta carga semántica y establecer relaciones entre los términos o pasajes extraídos.

– Recuperación cross-language: Hallar documentos escritos en cualquier lenguaje que son relevantes a una consulta expresada en otro lenguaje (búsqueda multilingual).

– Búsquedas Web: Se refiere a los SRI que operan sobre un corpus web privado (intranet) o público (Internet). La web ha planteado nuevos desafíos al área de RI, debido a sus características particulares como – por ejemplo – dinamismo y tamaño.

– Recuperación de Información Distribuida: A diferencia de los SRI clásicos donde el corpus y las estructuras de datos que auxilian a la búsqueda están centralizadas, aquí se plantea la tarea sobre los mismos elementos pero distribuidos sobre una red de computadoras.

– Modelado de Usuarios: Esta área – a partir de la interacción de los usuarios con un SRI – estudia como se generan de forma automática perfiles que definan las necesidades de información de éstos.

– Recuperación de Información Multimedia: Más allá de que los SRI tradicionales operan sobre corpus de documentos textuales, la recuperación de información tiene que tratar con otras formas alternativas de representación como imágenes, registro de conversaciones y video.

– Desarrollo de Conjuntos (data-sets) de Prueba: A los efectos de evaluar SRI completos o nuevos métodos y técnicas es necesario disponer de juegos de prueba normalizados (corpus con preguntas y respuestas predefinidas, corpus clasificados, etc.). Esta área tiene que ver con la producción tales conjuntos, a partir de diferentes estrategias que permitan reducir la complejidad de la tarea, manejando la dificultad inherente a la carga de subjetividad existente.

¿Recuperación de información o recuperación de datos?

Muchos usuarios se encuentran familiarizados con el concepto de recuperación de datos (RD), especialmente aquellos que – a menudo – interactuan con sistemas de consulta en bases de datos relacionales o en registros de alguna naturaleza, como por ejemplo, un registro de los empleados de una organización. Sin embargo, hay diferencias significativas en los conceptos que definen que el tratamiento de las unidades (datos o información) en cada caso sean completamente diferentes.

Básicamente, existen diferencias sustanciales en cuanto a los objetos con que se trata y su representación, la especificación de las consultas y los resultados.

En el área de RD los objetos que se tratan son estructuras de datos conocidas. Su representación se basa en un formato previo definido y con un significado implícito (hay una sintaxis y semántica no ambigua) para cada elemento. Por ejemplo, una tabla en una base de datos que almacena instancias de clientes de una organización posee un conjunto de columnas que definen los atributos de todos los clientes y cada fila corresponde a uno en particular. Nótese que cada elemento (atributo) tiene un dominio conocido y su semántica está claramente establecida. Por otro lado, en el área de RI la unidad u objeto de tratamiento es básicamente un documento de texto – en general – sin estructura.

En cuanto a la especificación de las consultas, en el área de RD se cuenta con una estructura bien definida

SQL	En lenguaje natural
SELECT* FROM Clientes WHERE Localidad="Chivilcoy AND Saldo_Cuenta= 10000	Seleccionar todos los clientes de Chivilcoy que deban más de 10000 pesos (se sabe, por definición, que lo que deben es su saldo de cuenta)

dada por un lenguaje de consulta que permite su especificación de manera exacta. Las consultas no son ambiguas y consisten en un conjunto de condiciones que deben cumplir los ítems a evaluar para que la misma se satisfaga. Por ejemplo, en el modelo de bases de datos, las consultas especifican – entre otros – utilizando el lenguaje SQL (Structured Query Language) cuya semántica es precisa:

En este ejemplo, se puede ver la clara semántica de la consulta en SQL a partir de que se conoce que existe un atributo Localidad y otro Saldo_Cuenta y lo que cada uno representa. Sin embargo, esto no es tan directo ni tan simple cuando se trata de recuperar documentos en el contexto de la RI. En primer lugar, debido a que la necesidad de un usuario puede ser difícil de expresar. Por ejemplo, supóngase que se desea encontrar:

«Documentos que contengan información biográfica de los entrenadores de los equipos de fútbol de Argentina que ganaron más torneos en los últimos 10 años»

La primera dificultad consiste en construir una expresión de consulta que refleje exactamente esta necesidad de información del usuario. Especialmente, si se tiene en cuenta que para resolverla completamente quizá primero se requiera de conocer información parcial, por ejemplo, «ganaron más torneos en los últimos 10 años». ¿Qué significa «ganaron más torneos»? Esta es una situación subjetiva y – en muchos casos – el sistema debe manejar estas cuestiones, junto con ambigüedades (por ejemplo, palabras cuyo significado está determinado por el contexto) e incompletitud de la mejor manera posible. De hecho, los documentos y las expresiones de consulta se interpretan de forma que el proceso de recuperación determine un grado de similitud entre estos.

En un sistema de RD los resultados consisten en el conjunto completo de elementos que satisfacen todas las condiciones del query. Como la consulta no admite errores, el resultado es exacto, ni uno más, ni uno menos. Y el orden de aparición es simplemente casual (a menos que específicamente se desee ordenar por alguna columna), pero en todos los casos este orden es irrelevante respecto de la consulta y no significa nada, es decir, no se puede implementar sistema de rankeo alguno. En el área de RI, aparece el concepto de relevancia y la salida (respuesta) se encuentra confeccionada de acuerdo a algún criterio que evalúa la «similitud» que existe entre la consulta y cada documento. Por lo tanto, el resultado es un ranking (que no es sinónimo de «orden», tal como se lo entiende habitualmente en RD), donde la primera posición corresponde al documento más relevante a la consulta y así decrece sucesivamente. El proceso de recuperación de información puede retornar documentos que no sean relevantes para el usuario, es decir, que el conjunto de respuesta no es exacto.

Otros autores también establecieron las diferencias entre ambos conceptos: Grossman y otros [8] claramente muestran la diferencia cuando enuncian que **«la recuperación de información es encontrar documentos relevantes, no encontrar simples correspondencias a unos patrones de bits».** Nótese la diferencia sustancial que existe en tratar de encontrar documentos «relevantes» a una consulta o – simplemente – encontrar aquellos donde «coinciden» patrones de términos o se cumplen ciertas condiciones. En el caso de la RD, la tarea es relativamente sencilla, mientras que en área de RI es extremadamente compleja y no existe aún una solución definitiva al problema.

La interacción del usuario con el SRI

La tarea de recuperar información puede ser planteada de diversas formas, de acuerdo a cómo el usuario interactúa con el sistema o bien qué facilidades éste le brinda. Básicamente, la tarea se la puede dividir en:

1) Recuperación inmediata: El usuario plantea su necesidad de información y – a continuación – obtiene referencias a los documentos que el sistema evalúa como relevantes. Existen dos modalidades:

a) Búsqueda (propiamente dicha) o recuperación «*ad-hoc*», donde el usuario formula una consulta en un lenguaje y el sistema la evalúa y responde. En este caso, el usuario tiene suficiente comprensión de su necesidad y sabe cómo expresar una consulta

al sistema. Un ejemplo clásico son los buscadores de Internet como Google (http://www.google.com), Altavista (http://www.altavista.com) o AllTheWeb (http://www.alltheweb.com).

b) *Navegación o browsing:* En este caso, el usuario utiliza un enfoque diferente al anterior. El sistema ofrece una interface con temas donde el usuario «navega» por dicha estructura y obtiene referencias a documentos a relacionados. Esto facilita la búsqueda a usuarios que no pueden definir claramente cómo comenzar con su consulta e – inclusive – van definiendo su necesidad a medida que observan diferentes documentos. Es este enfoque no se formula consulta explícita. Un ejemplo típico es el proyecto Open Directory (http://www.dmoz.org).

En ambos casos, la colección es relativamente estática, es decir, se parte de un conjunto de documentos y la aparición de nuevos no es muy significativa. Por otro lado, las consultas son las que se van modificando ya que este proceso es proactivo por parte del usuario.

2) Recuperación diferida: El usuario especifica sus necesidades y el sistema entregará de forma continua los nuevos documentos que le lleguen y concuerden con esta. Esta modalidad recibe el nombre de **filtrado y ruteo** y la necesidad del usuario – generalmente – define un «perfil» (profile) de los documentos buscados. Nótese que un «perfil» es – de alguna forma – un query y puede ser tratado como tal. Cada vez que un nuevo documento arriba al sistema se compara con el perfil y – si es relevante – se envía al usuario. Un ejemplo, es el servicio provisto por la empresa Indigo Stream Technologies denominado GoogleAlert (http://www.googlealert.com/).

En esta modalidad la consulta es relativamente estática (corresponde al profile) y el usuario tiene un rol pasivo. El dinamismo está dado por la aparición de nuevos documentos y es lo que determina mas resultados para el usuario.

En algunos casos, se plantea que documentos y consultas son objetos de la misma clase por lo que estos enfoques son – de alguna manera – visiones diferentes de una misma problemática. Bajo este punto de vista, documentos y consultas se pueden intercambiar. Sin embargo, esto no es siempre posible debido al tratamiento que se aplica a cada uno en diferentes sistemas. Algunos sistemas representan queries y documentos de diferente manera. Es más, existe una diferencia obvia en cuanto a la longitud

El concepto de relevancia

Como mencionamos, la recuperación de información intenta resolver el problema de encontrar documentos relevantes que satisfagan la necesidad de información de un usuario. Sin embargo, se ha planteado la dificultad para llevar a cabo esta tarea debido a la imposibilidad de expresar exactamente tal necesidad. Además, la noción de relevancia es un juicio subjetivo [9] y depende de diferentes factores relacionados más cercanamente con el usuario. La relevancia de un documento respecto de un query se refiere a cuánto el primero responde al segundo. De igual manera, luego el usuario evalúa cuánto, es decir, en qué medida, se satisface su necesidad de información [6].

Es por ello, que se plantea la relevancia como similitud, de manera de poder comparar documentos con consultas y – bajo ciertos criterios – definir una medida de distancia entre ambos. Por lo tanto, se puede plantear la idea de que «un documento es relevante a una consulta si son similares», donde la medida de similitud puede estar basada en diferentes criterios (coincidencias de términos, significado de estos, frecuencia de aparición de términos y distribución del vocabulario, entre otros).

Martínez Méndez y otros [10] resaltan la dificultad para determinar la relevancia o no de un documento respecto de una consulta. Plantean – por ejemplo – que dos personas pueden juzgar un mismo documento de diferente manera y que es difícil establecer los criterios para la evaluación de la relevancia. Finalmente, mencionan la idea de relevancia parcial, es decir, cuando solo una parte del documento se considera relevante.

Por otro lado, como el query no describe exactamente la necesidad de información del usuario, algunos autores [6] definen el concepto de «pertinencia», donde se incluyen las restricciones impuestas por el SRI. Este concepto está relacionado con la utilidad del documento para el usuario [10], de acuerdo a la necesidad de información original que guió su búsqueda, independientemente si es en parte o todo el documento.

Sin embargo – y a pesar de las dificultades para determinarla – el concepto genérico de relevancia es aceptado ampliamente por la comunidad de RI para evaluar la respuesta de un SRI respecto de una consulta de un usuario, la cual – como ya mencionamos – surge a partir de una necesidad de información.

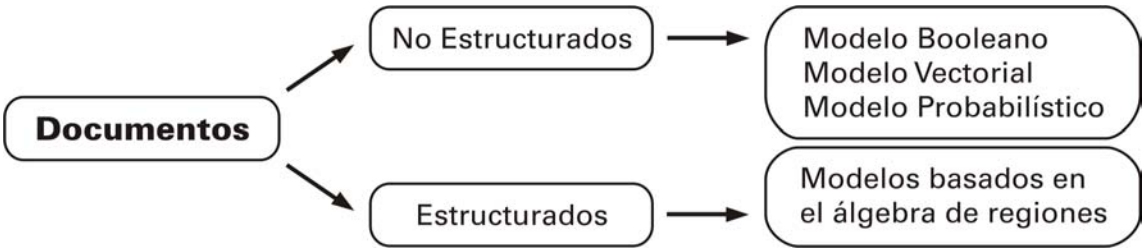

Fig. 3. Arquitectura básica de un SRI.

Modelos de RI

Los SRI toman un conjunto de documentos (colección) para procesar y luego poder responder consultas. De forma básica, podemos clasificar los documentos en estructurados y no estructurados. Los primeros son aquellos en los que se pueden reconocer elementos estructurales con una semántica bien definida, mientras que los segundos corresponden a texto libre, sin formato. La diferencia fundamental de un SRI que procese documentos estructurados se encuentra en que puede extraer información adicional al contenido textual, la cual utiliza en la etapa de recuperación para facilitar la tarea y aumentar las prestaciones.

A partir de lo expresado anteriormente en la figura 3 se presenta una posible clasificación de modelos de RI – la cual no es exhaustiva – de acuerdo a características estructurales de los documentos. A continuación se describen – de forma somera – los modelos clásicos y el álgebra de regiones.

a) Modelo booleano: En el modelo booleano la representación de la colección de documentos se realiza sobre una matriz binaria documento–término, donde los términos han sido extraídos manualmente o automáticamente de los documentos y representan el contenido de los mismos.

Las consultas se arman con términos vinculados por operadores lógicos (AND, OR, NOT) y los resultados son referencias a documentos donde cuya representación satisface las restricciones lógicas de la expresión de búsqueda. En el modelo original no hay ranking de relevancia sobre el conjunto de respuestas a una consulta, todos los documentos poseen la misma relevancia.

Si bien es el primer modelo desarrollado y aún se lo utiliza, no es el preferido por los ingenieros de software para sus desarrollos. Existen diversos puntos en contra que hacen que cada día se lo utilice menos y – además – se han desarrollado algunas extensiones, bajo el nombre modelo booleano extendido [11] [12], que tratan de mejorar algunos puntos débiles.

b) *Modelo Vectorial:* Este modelo fue planteado y desarrollado por Gerard Salton [13] y – originalmente – se implementó en un SRI llamado SMART. Aunque el modelo posee más de treinta años, actualmente se sigue utilizando debido a su buena performance en la recuperación de documentos.

Conceptualmente, este modelo utiliza una matriz documento–término que contiene el vocabulario de la colección de referencia y los documentos existentes. En la intersección de un término *t* y un documento *d* se almacena un valor numérico de importancia del término *t* en el documento *d*; tal valor representa su *poder de* discriminación. Así, cada documento puede ser visto como un vector que pertenece a un espacio n-dimensional, donde *n* es la cantidad de términos que componen el vocabulario de la colección. En teoría, los documentos que contengan términos similares estarán a muy poca distancia entre sí sobre tal espacio. De igual forma se trata a la consulta, es un documento más y se la mapea sobre el espacio de documentos. Luego, a partir de una consulta dada es posible devolver una lista de documentos ordenados por distancia (los más relevantes primero). Para calcular la semejanza entre el vector consulta y los vectores que representan los documentos se utilizan diferentes fórmulas de distancia, siendo la más común la del coseno.

Obsérvese el siguiente ejemplo donde se representa a un documento d y a una consulta c:

Documento: «La República Argentina ha sido nominada para la realización del X Congreso Americano de Epidemiología en Zonas de Desastre. El evento se realizará ...»

Consulta: «argentina congreso epidemiología»

Matriz término-documento con pesos y 1 normalizados entre 0

...	Argentina	...	Congreso	Epidemiología	...
d_1	0.5	-	0.3	0.2	-
...	-	-	-	-	-
d_n	-	-	-	-	-
Consulta	0.4	-	0.3	0.3	-

c) Modelo probabilístico: Fue propuesto por Robertson y Spark-Jones [14] con el objetivo de representar el proceso de recuperación de información desde el punto de vista de las probabilidades. A partir de una expresión de consulta se puede dividir una colección de N documentos (figura 4) en cuatro subconjuntos distintos: REL conjunto de documentos relevantes, REC conjunto de documentos recuperados, RR conjunto de documentos relevantes recuperados y NN el conjunto de documentos no relevantes no recuperados.

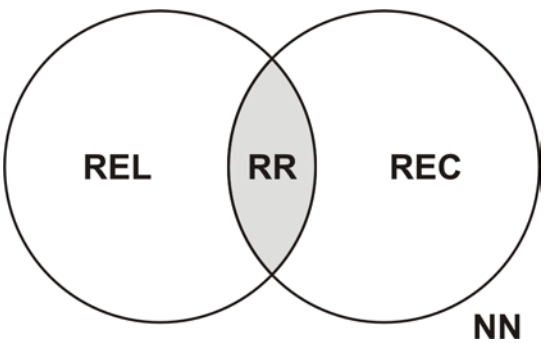

Fig. 4. División de la colección.

El resultado ideal de a una consulta se da cuando el conjunto REL es igual REC. Como resulta difícil lograrlo en primera intención, el usuario genera una descripción probabilística del conjunto REL y a través de sucesivas interacciones con el SRI se trata de mejorar la perfomance de recuperación. Dado que una recuperación no es inmediata dado que involucra varias interacciones con el usuario y que estudios han demostrado que su perfomance es inferior al modelo vectorial, su uso es bastante limitado.

d) Modelos para documentos estructurados: Los modelos clásicos responden a consultas, buscando sobre una estructura de datos que representa el contenido de los documentos de una colección, únicamente como listas de términos significativos. Un modelo de recuperación de documentos estructurados utiliza la estructura de los mismos a los efectos de mejorar la perfomance y brindar servicios alternativos al usuario (por ejemplo, uso de memoria visual, recuperación de elementos multimedia, mayor precisión sobre el ámbito de la consulta y demás).

La estructura de los documentos a indexar está dada por marcas o etiquetas, siendo los estándares más utilizados el SGML (Standard General Markup Language), el HTML (HyperText Markup Language), el XML (eXtensible Markup Language) y LATEX.

Al poseer la descripción de parte de la estructura de un documento es posible generar un grafo sobre el que se navegue y se respondan consultas de distinto tipo, por ejemplo:

– Por estructura: ¿Cuáles son las secciones del segundo capítulo?

– Por metadatos o campos: Documentos de «Editorial UNLu» editados en 1998

– Por contenido: Término «agua» en títulos de secciones

– Por elementos multimedia: Imágenes cercanas a párrafos que contengan Bush

Para Baeza-Yates existen dos modelos en esta categoría «nodos proximales» [15] y «listas no superpuestas» [16]. Ambos modelos se basan en almacenar las ocurrencias de los términos a indexar en estructuras de datos diferentes, según aparezcan en algún elemento de estructura (región) o en otro como capítulos, secciones, subsecciones y demás. En general, las regiones de una misma estructura de datos no poseen superposición, pero regiones en diferentes estructuras sí se pueden superponer. Los tipos de consultas soportados son simples:

· Seleccione una región que contenga una palabra determinada.

· Seleccione una región X que no que no contenga una región Y

· Seleccione una región contenida en otra región

Sobre una estructura tipo libro un ejemplo de consulta sería:

[subsección[+] CONTIENE «'tambo»]

Como respuesta el SRI buscaría subsecciones y sub-subsecciones que contengan el término «tambo».

Cabe mencionar que algunos motores de búsqueda de Internet ya utilizan ciertos elementos de la estructura de un documento – por ejemplo, los títulos – a los efectos de realizar tareas de rankeo, resumen automático, clasificación y otras.

La expansión de estos lenguajes de demarcación, especialmente en servicios sobre Internet, hacen que se generen y publiquen cada vez más documentos semiestructurados. Es necesario – entonces – desarrollar técnicas que aprovechen el valor agregado de los nuevos documentos. Si bien – en la actualidad – estas no se encuentran tan desarrolladas como los modelos tradicionales, consideramos su evolución como una cuestión importante en el área de RI, especialmente a partir de investigaciones con enfoques diferentes que abordan la problemática [17] [18] [19].

La RI en la era de la web

Con la aparición de la web surgieron nuevos desafíos para resolver en el área de recuperación de información debido – principalmente – a sus características y su tamaño. La web puede ser vista como un gran repositorio de información, completamente distribuido sobre Internet y accesible por gran cantidad de usuarios. Por sus orígenes como un espacio público existen millones de organizaciones y usuarios particulares que incorporan, quitan o modifican contenido continuamente, por lo que su estructura no es estática.

Su contenido no respeta estándares de calidad, ni estilos ni organización. Como medio de publicación de información de naturaleza diversa se ha convertido en un servicio de permanente crecimiento. Una de las características de la información publicada en la web es su dinamismo, dado que pueden variar en el tiempo tanto los contenidos como su ubicación [20] [21].

El tamaño de la web es imposible de medir exactamente y muy difícil de estimar. Sin embargo, se calcula que son decenas de terabytes de información, y crece permanentemente. Está formada por documentos de diferente naturaleza y formato, desde páginas HTML hasta archivos de imágenes pasando por gran cantidad de formatos estándar y propietarios, no solamente con contenido textual, sino también con contenido multimedial.

La búsqueda de información en la web es una práctica común para los usuarios de Internet y los sistemas de recuperación de información web (conocidos como motores de búsqueda) se han convertido en herramientas indispensables para los usuarios. Su arquitectura y modo de operación se basan en poder recolectar mediante un mecanismo adecuado los documentos existentes en los sitios web. Una vez obtenidos, se llevan a cabo tareas de procesamiento que permiten extraer términos significativos contenidos dentro de los mismos, junto con otra información, a los efectos de construir estructuras de datos (índices) que permitan realizar búsquedas de manera eficiente. Luego, a partir de una consulta realizada por un usuario, un motor de búsqueda extraerá de los índices las referencias que satisfagan la consulta y se retornará una respuesta rankeada por diversos criterios al usuario. El modo de funcionamiento de los diferentes motores de búsqueda puede diferir en diversas implementaciones de los mecanismos de recolección de datos, los métodos de indexación y los algoritmos de búsqueda y rankeo.

Sin embargo, esta tarea no es sencilla y se ha convertido en un desafío para los SRI debido a las características propias de la web. Baeza-Yates [3] plantean que hay desafíos de dos tipos:

a) Respecto de los datos

– Distribuidos: La web es un sistema distribuido, donde cada proveedor de información publica su información en computadoras pertenecientes a redes conectadas a Internet, sin una estructura ó topología predefinida.

– Volátiles: El dinamismo del sistema hace que exista información nueva a cada momento ó bien que cambie su contenido ó inclusive desaparezca otra que se encontraba disponible.

– No estructurados y redundantes: Básicamente, la web está formada de páginas HTML, las cuales no cuentan con una estructura única ni fija. Además, mucho del contenido se encuentra duplicado (por ejemplo, espejado).

– Calidad: En general, la calidad de la información publicada en la web es altamente variable, tanto en escritura como en actualización (existe información que puede considerarse obsoleta), e inclusive existe información con errores sintácticos, ortográficos y demás.

– Heterogeneidad: La información se puede encontrar publicada en diferentes tipos de medios (texto, audio, gráficos) con diferentes formatos para cada uno de éstos. Además, hay que contemplar los diferentes idiomas y diferentes alfabetos (por ejemplo, árabe ó chino).

a) Respecto de los usuarios.

– Especificación de la consulta: Los usuarios encuentran dificultades para precisar – en el lenguaje de consulta – su necesidad de información.

– Manejo de las respuestas: Cuando un usuario realiza una consulta se ve sobrecargado de respuestas, siendo una parte irrelevante.

Estas características – sumadas al tamaño de la web – imponen restricciones a las herramientas de búsqueda en cuanto a la cobertura y acceso a los documentos, exigiendo cada vez mayores recursos computacionales (espacio de almacenamiento, ancho de banda de las redes, ciclos de CPU) y diferentes estrategias para mejorar la calidad de las respuestas.

Referencias

1) Maes, P. Agents that Reduce Work and Information Overload. Communications of the ACM, 37(7): 30-40. 1994.

2) Carlson, C. Information overload, retrieval strategies and Internet user empowerment. Haddon, Leslie, Eds. Proceedings The Good, the Bad and the Irrelevant (COST 269), Helsinki (Finland). 1(1): 169-173, 2003.

3) Baeza-Yates, R. y Ribeiro-Neto, B. Modern Information Retrieval. ACM Press. Addison Wesley. 1999.

4) Salton, G. Y Mc Gill, M.J. Introduction to Modern Information Retrieval. New York. Mc Graw-Hill Computer Series. 1983.

5) Croft, W.B. Approaches to intelligent information retrieval. Information Proccesing & Management, 23(4): 249-254. 1987.

6) Korfhage, R. R. Information Storage and Retrieval. New York. Wiley Computer Publishing. 1997.

7) Peña, R., Baeza-Yates, R., Rodriguez, J.V. Gestión Digital de la Información. Alfaomega Grupo Editor. 2003.

8) Grossman, D. y Frieder, O. Information Retrieval. Algorithms and Heuristics. Kluwer Academic Publishers. 1998.

9) Van Rijsbergen, C.J. Information Retrieval. Department of Computing Science. University of Glasgow. 1979.

10) Martinez Mendez, F.J. y Rodriguez Muñoz, J.V. Reflexiones sobre la Evaluación de los Sistemas de Recuperación de Información: Necesidad, Utilidad y Viabilidad». Anales de Documentación, 7:153-170. 2004.

11) Waller, W. G. y Kraft, D. H. «A mathematical model for a weighted Boolean retrieval system». Information Processing and Management, 15(5):235-245. 1979.

12) Salton, G.; Fox, E.A. y Wu, H. Extended Boolean information retrieval. Communications of the ACM, 26(11):1022-1036. Noviembre, 1983.

13) Salton, G. (editor). The SMART Retrieval System – Experiments in Automatic Document Processing. Prentice Hall Inc. Englewood Cliffs, NJ. 1971.

14) Robertson, S.E y Spark-Jones, K. Relevance Weighting of Search terms. Journal of Documentation. 33:126-148. 1976.

15) Navarro, G y Baeza-Yates, R. A. Language for queries on structure and contents of textual databases. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York. ACM Press. 93-101. 1995.

16) Burkowski, F. Retrieval activities in a database consisting of heterogeneous

collections of structured texts. Belkin, N., Ingwersen, P., Pejtersen, A. M., and Fox, E., editors, Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York. ACM Press. 112–125. 1992.

17) Egnor, D. y Lord, R. Structured information retrieval using XML. Proceedings of the ACM SIGIR 2000 Workshop on XML and Information Retrieval. 2000.

18) Ogilvie, P. y Callan, J. Language Models and Structured Document Retrieval. Proceedings of the first INEX workshop. 2003.

19) Raghavan, S. Y Garcia-Molina, H. Integrating diverse information management systems: A brief survey. IEEE Data Engineering Bulletin, 24(4):44-52, 2001.

20) Brewington, B. E. y Cybenko Thayer, G. How Dynamic is the Web?. Proceedings of the Ninth International World Wide Web Conference. 2000.

21) Lawrence, S. y Giles, L. Accessibility and Distribution of Information on the Web. Nature, 400(6740): 107-109. 1999.

Recibido: 26 de noviembre del 2006.
Aprobado en su forma definitiva: 4 de mayo del 2007.

MsC. Fernando Bordignon
Departamento de Ciencias Básicas de la Universidad Nacional de Luján. Laboratoris de Redes de Datos. Argentina.
Sitio:
<http://www.tyr.un/u.edu.ar>

MsC. Gabriel Tolosa Chacón
Departamento de Ciencias Básicas de la Universidad Nacional de Luján.
Laboratoris de Redes de Datos. Argentina.
Sitio:
<http://www.tyr.un/u.edu.ar>