Tratamiento Automático de la Información 2

Sistemas de recuperación de información

Lic. Eduardo Pablo Giordanino
Tecnicatura Superior en Bibliotecología
Instituto de Formación Técnica Superior Nº 13
Ministerio de Educación. Ciudad de Buenos Aires

Concepto de RI

- Salton (1983): “es un campo relacionado con la estructura, análisis, organización, almacenamiento, búsqueda y recuperación de información”

- Baeza-Yates (1999): “la Recuperación de Información trata con la representación, el almacenamiento, la organización y el acceso a ítems de información”

Concepto de RI

- Croft (1987): “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”
- Korfhage (1997): “la localización y presentación a un usuario de información relevante a una necesidad de información expresada como una pregunta”

Concepto de RI

- Tolosa y Bordignon (2007): "*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*"

- SRI : Sistema de Recuperación de Información

Teoría de RI - 1

La representación de la información de los documentos se realiza asignando varios conjuntos de términos de indexación a cada documento (y no asignando clases o subclases a los documentos desde una tabla de clasificación)

Teoría de RI - 2

Las necesidades de información de los usuarios del sistema también pueden representarse mediante conjuntos de términos

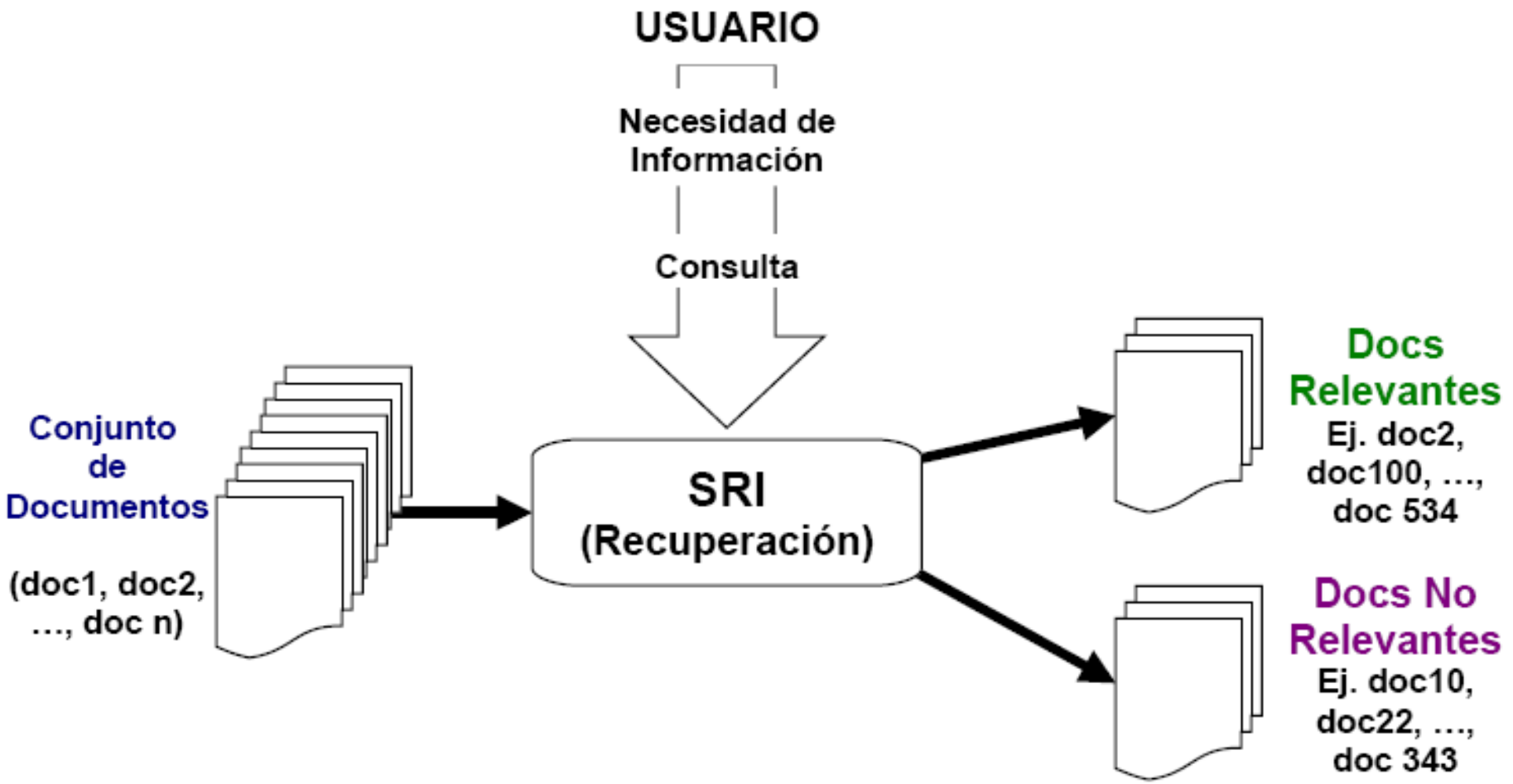
Necesidad de información = *query* = consulta = ecuación de búsqueda

Teoría de RI - 3

Los documentos más relevantes en relación con cada necesidad de información serán aquellos que presenten mayor grado de similitud con la necesidad de información

Los sistemas de recuperación de información son aquellos que realizan estas clases de operaciones según las tres ideas nucleares mencionadas

Problemática de la RI



Problemática de la RI

- Hay una colección de documentos con información
- Existen usuarios con necesidades de información, las que plantean al SRI en forma de consultas (*query*)
- El SRI retorna como respuesta referencias a documentos relevantes

los que considera que satisfacen la necesidad expresada, en forma de una lista ordenada [rankeada]

Problemática de la RI

- La respuesta “ideal” es díficil
- Se debe compatibilizar la expresión de la necesidad de información con el lenguaje y con los documentos de la colección
- Precisión de la respuesta (eficiencia: cuantos más documentos relevantes contenga la respuesta, más preciso será)
- Relevancia (criterio basado en similitud)
- Precisión vs. Exhaustividad (relación inversa)

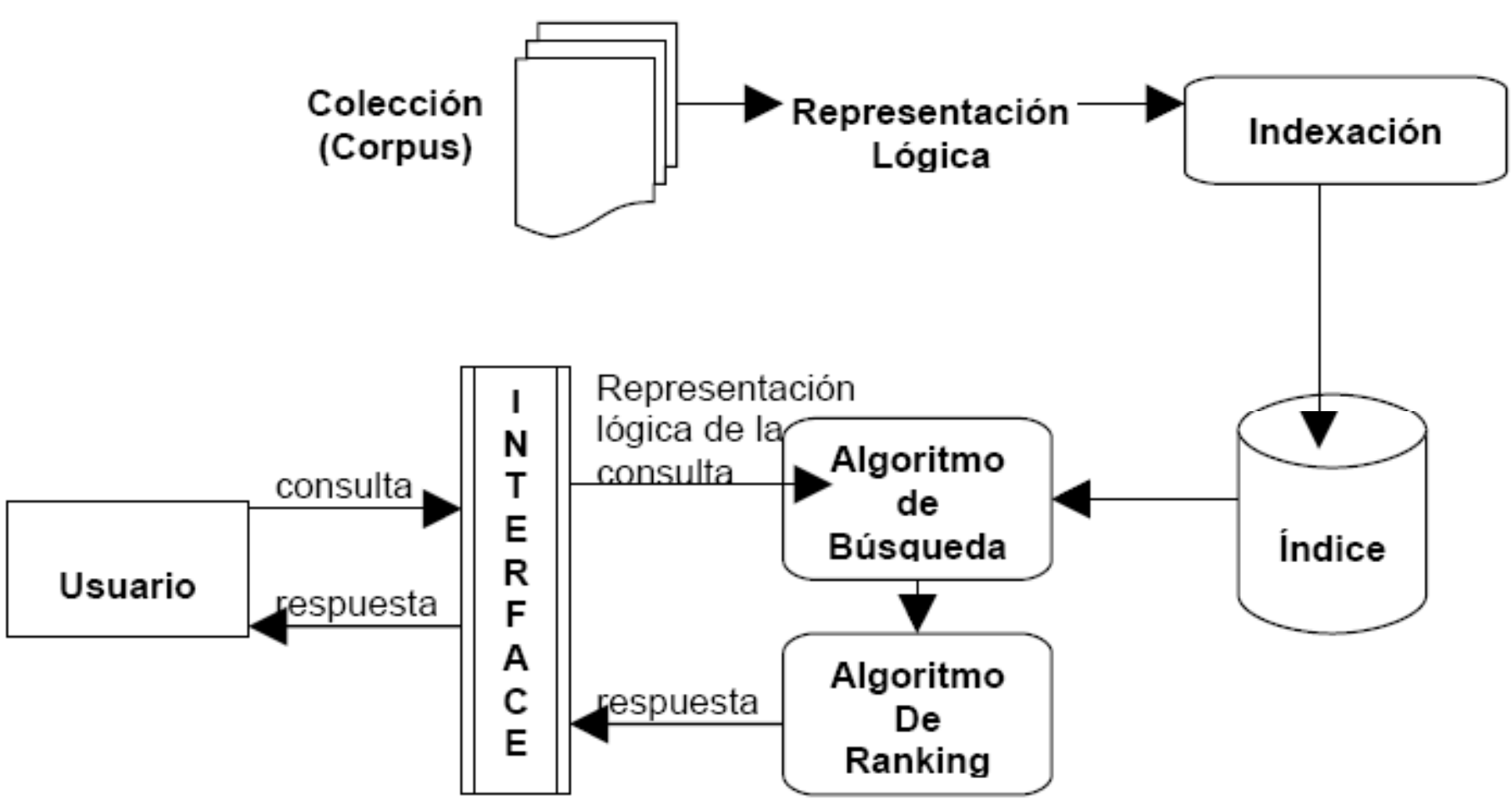
Tareas básicas de un SRI

- Representación lógica de los documentos (almacenamiento del original)
- Representación de la necesidad de información del usuario en forma de consulta
- Evaluación de los documentos recuperados desde una consulta (establecer su relevancia)
- Ranking de los documentos considerados relevantes (respuesta)
- Presentar respuesta al usuario
- Retroalimentación de consulta (para aumentar calidad de las respuestas)

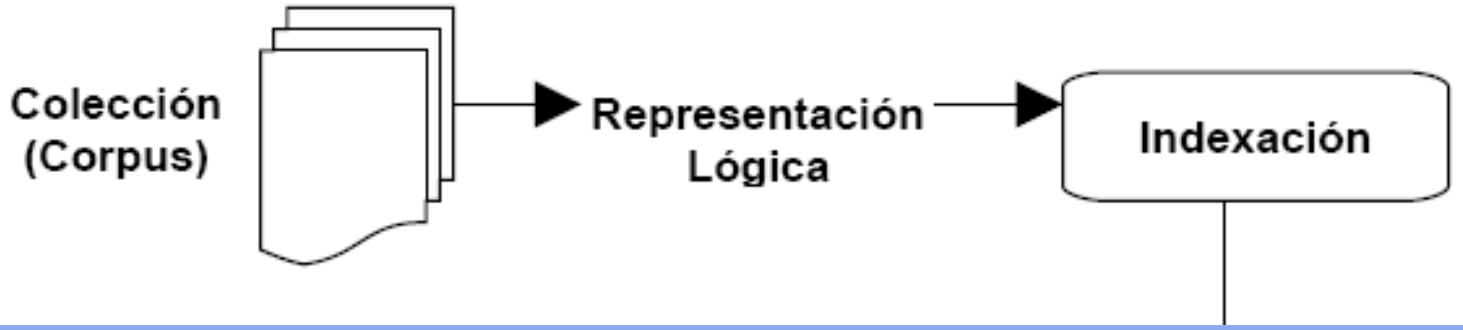
Tareas básicas de un SRI

- "Tal como represento, podré buscar; y no más"
- Necesidad de información <-=-> consulta
- *La recuperación de información, en el contexto de un SRI, consiste principalmente en la determinación de qué documentos de una colección contienen las palabras claves de una pregunta de usuario, para satisfacer la necesidad de información de ese usuario.*

Arquitectura de un SRI

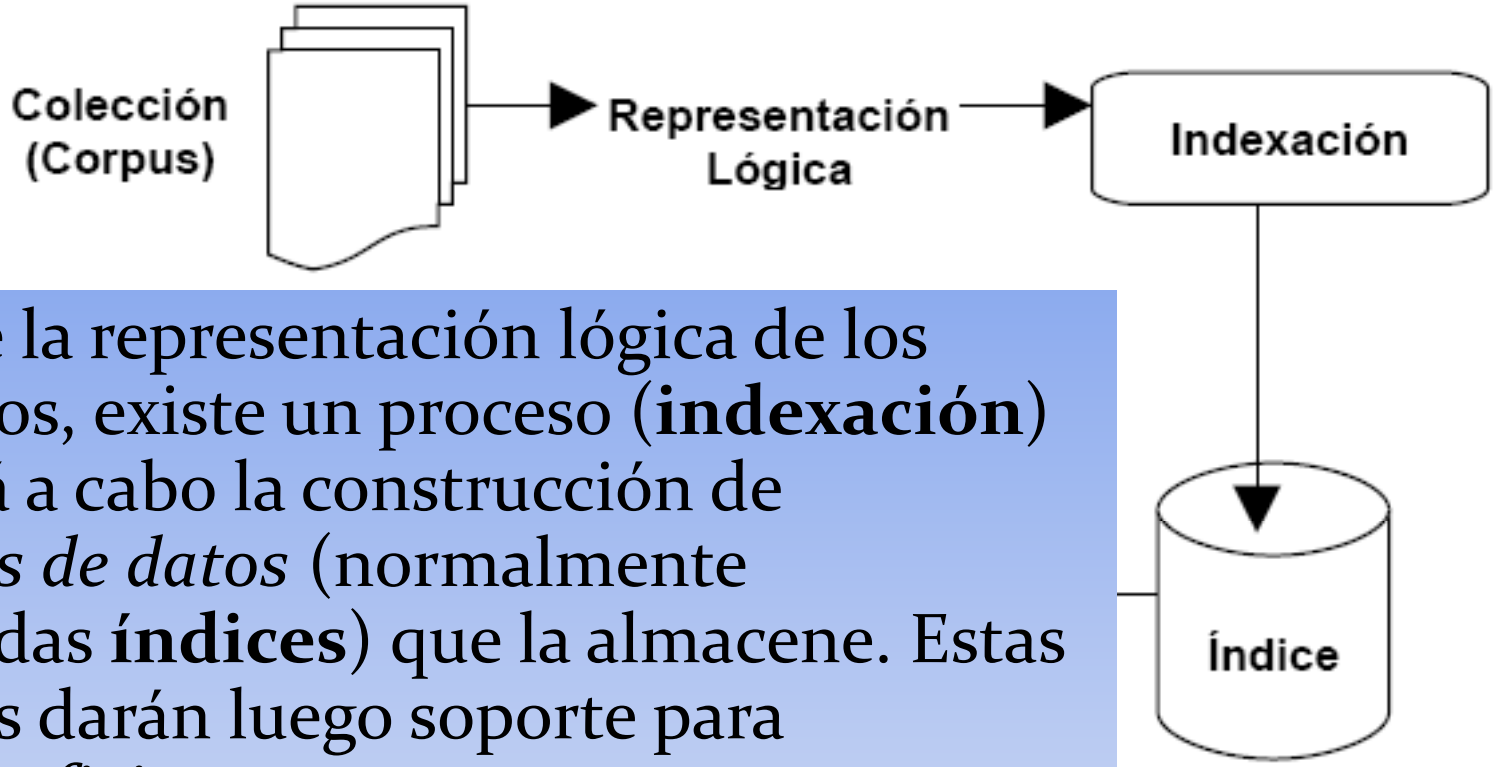


Arquitectura de un SRI



Colección: documentos de texto (sucesiones de palabras), escritos en lenguaje natural (**corpus** o **base de datos textual**). Para poder realizar operaciones sobre un corpus, es necesario primero una **representación lógica** de todos sus documentos, que puede consistir en un conjunto de términos, frases u otras unidades (sintácticas o semánticas) que permitan caracterizarlos. La representación de los documentos mediante un conjunto de sus términos se la conoce como "bolsa de palabras"

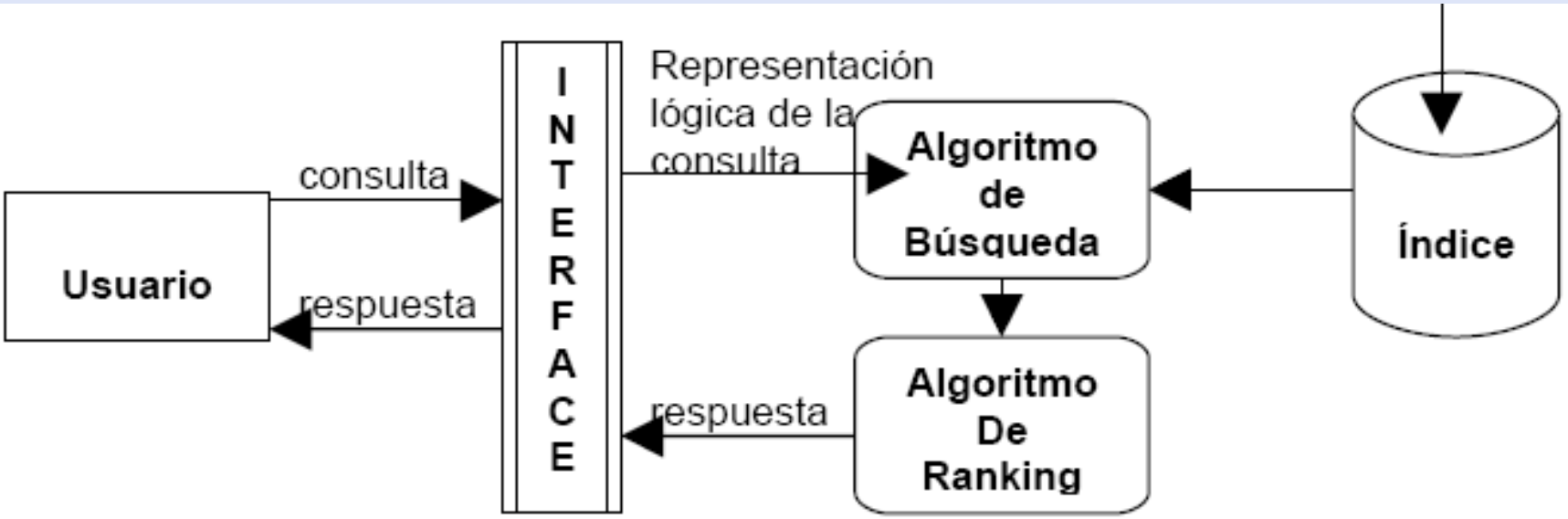
Arquitectura de un SRI



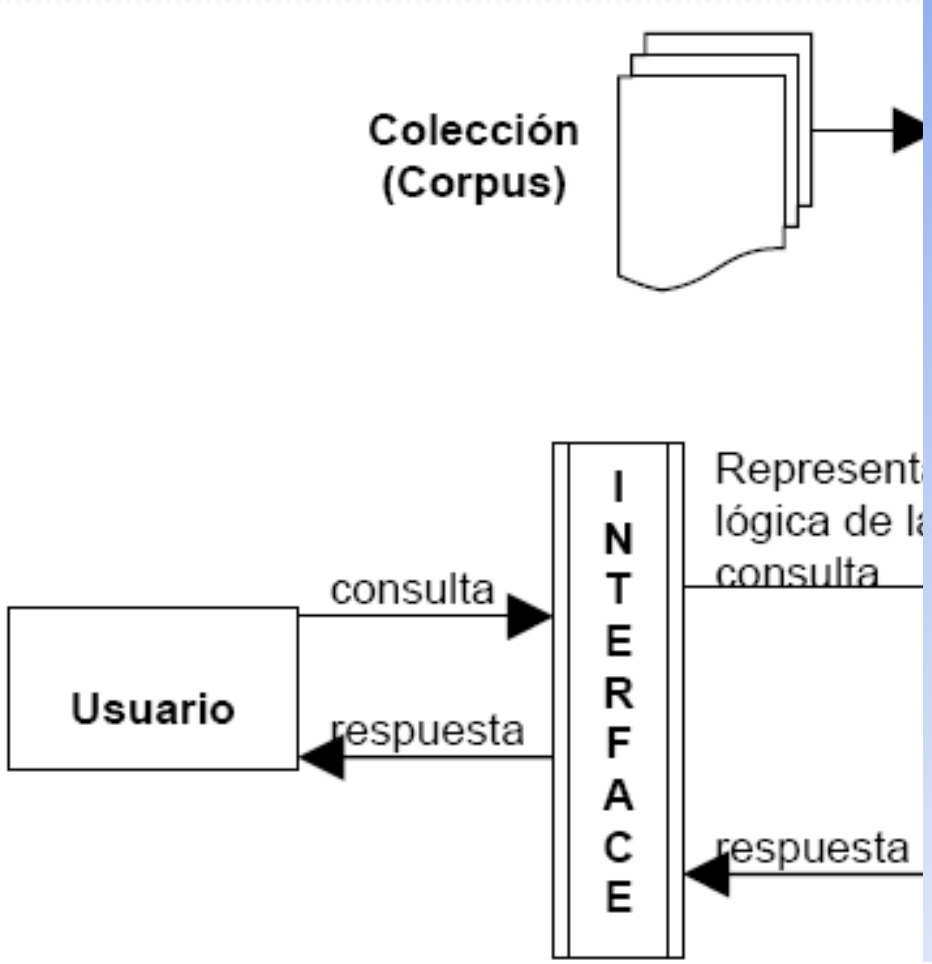
- A partir de la representación lógica de los documentos, existe un proceso (**indexación**) que llevará a cabo la construcción de *estructuras de datos* (normalmente denominadas **índices**) que la almacene. Estas estructuras darán luego soporte para búsquedas eficientes.
- Es importante destacar que una vez construidos los índices, los documentos del corpus pueden ser eliminados del sistema ya que éste retornará las referencias a los mismos debido a que cuenta con la información necesaria para hacerlo (sistemas referenciales vs. sistemas documentales. Ej: Google)

Arquitectura de un SRI

- El **algoritmo de búsqueda** acepta como entrada una expresión de consulta (*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. El primer ítem de dicha lista corresponde al documento más relevante y así sucesivamente en orden decreciente



Arquitectura de un SRI



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

Tareas del usuario - RI

TAREA : RECUPERAR INFORMACIÓN

1) Recuperación inmediata

El usuario plantea su necesidad de información y recibe referencias a los documentos:

a) **Búsqueda propiamente dicha o recuperación *ad-hoc*:** el usuario formula una consulta en un lenguaje y el sistema la evalúa y responde. Ejemplo típico: los buscadores de Internet como Google o Yahoo.

b) **Navegación o *browsing***: El sistema ofrece una interfase con temas donde el usuario *navega* por dicha estructura y obtiene referencias a los documentos relacionados. Facilita la búsqueda a usuarios que no pueden definir claramente cómo comenzar su consulta y van definiendo su necesidad a medida que observan diferentes categorías. Ejemplo: Open Directory (http://www.dmoz.org/)

Tareas del usuario - RI

TAREA : RECUPERAR INFORMACIÓN

2) Recuperación diferida: El usuario especifica sus necesidades y el sistema entregará de forma continua los nuevos documentos que le lleguen y concuerden con ésta.

Se la llama **filtrado y ruteo**. La necesidad del usuario define un "perfil" (*profile*) de los documentos buscados.

Un "perfil" es un *query*. Equivale al DSI y una forma moderna es el RSS. Cada vez que un nuevo documento llega al sistema se compara con el perfil y -si es relevante– se envía al usuario (Google Alertas)

- En esta modalidad la consulta es relativamente estática y el usuario tiene un rol pasivo. La dinámica está dada por la aparición de nuevos documentos.

Tareas del usuario - FRBR

Según el modelo FRBR (1997), los usuarios :

- *encontrar* entidades que correspondan a los criterios de búsqueda establecidos. Localizar una entidad en una base de datos, como resultado de una búsqueda que utiliza un atributo o relación de la entidad.
- *identificar* una entidad. Confirmar que la entidad descrita corresponde a la entidad buscada o distinguir entre dos o más entidades con características similares.
- *seleccionar* una entidad adecuada para las necesidades del usuario. Elegir una entidad que satisfaga las necesidades del usuario respecto del contenido, formato físico, etc., o rechazar una entidad no adecuada para las necesidades del usuario.
- *adquirir* u obtener acceso a la entidad descrita. Adquirir una entidad a través de la compra, préstamo, etc., o acceder electrónicamente a una entidad a través de una conexión en línea a un ordenador remoto.

Tareas del usuario - FRBR

Recordemos que el modelo FRBR…

define entidades, que forman tres grupos conceptuales, y las relaciones entre esas distintas entidades. Los tres grupos de entidades definidos son:

- **Grupo 1: Obras.** Creación intelectual o artística que se describen en los registros bibliográficos: *obra*, *expresión*, *manifestación* e *ítem*.
- **Grupo 2: Autores**. Incluyen *persona* (un individuo) , *entidad corporativa* (una organización o grupo de individuos y/o organizaciones), y *familia*.
- **Grupo 3: Materias** de las obras. El grupo incluye: *concepto*, *objeto*, *acontecimiento* y *lugar*.

Tareas del usuario

Relaciones del FRBR

Las relaciones se utilizan como vehículo para establecer el vínculo entre una entidad y otra, así como medio para ayudar al usuario a "navegar" por el universo.

- **Relaciones entre entidades del grupo 1 (OBRAS)**
- Una *obra* "se realiza mediante" la *expresión* (una *expresión* "es una realización de" una *obra*).
- Una *expresión* se "materializa mediante" una *manifestación* (una *manifestación* es la materialización de una *expresión*)
- Una *manifestación* "es ejemplificada por" un *item* (un *item* es un ejemplo de una *manifestación*)

Tareas del usuario

Relaciones del FRBR

Relaciones entre entidades del grupo 2 (AUTORES)

- Una *obra* es creada por una *persona* o una *entidad corporativa*
- Una *expresión* es realizada por una *persona* o una *entidad corporativa*
- Una *manifestación es producida por una* persona o *una* entidad corporativa
- Un *ítem* es poseído por una *persona* o una *entidad corporativa*

Tareas del usuario

Relaciones del FRBR

Relaciones entre entidades del grupo 3 (MATERIAS)

- Las entidades de los tres grupos están vinculadas a la entidad *obra* por una relación de materia.
- La relación "tiene como materia" indica que cualquiera de las entidades del modelo, incluida la propia obra, puede ser la materia de otra obra.
- La relación indica que una *obra* puede tratar sobre un *concepto*, un *objeto*, un *acontecimiento* o *lugar*; puede tratar sobre una *persona* o *entidad corporativa*; puede tratar sobre una *expresión*, una *manifestación* o un *ítem*; puede tratar sobre otra *obra*.

Modelos de datos

- **MODELO DE ENTIDAD-RELACIÓN**
- El modelo entidad-relación es una herramienta para modelar los datos de un sistema de información (SI).

- Un Sistema de Información (SI) es un conjunto organizado de elementos para el tratamiento y la administración de datos y de información.
- Los elementos pueden ser: personas, datos, actividades, recursos (materiales).

Modelos de datos

- En *Informática*, un sistema de información es cualquier sistema o subsistema de equipo de telecomunicaciones o computacional interconectados y que se utilicen para obtener, almacenar, manipular, administrar, mover, controlar, desplegar, intercambiar, transmitir o recibir voz y/o datos, e incluye tanto los programas de computación como el *hardware*.

- Las bases de datos son la implementación de un modelo de datos.

Modelos de datos

- Una estructura de datos está formada por una colección o conjunto de datos organizados de tal manera que tengan asociado un grupo de operaciones que permita administrarlos.

- Los lenguajes de programación permiten manejar tipos de datos estructurados ya predefinidos, como los arreglos o los registros (estructuras de datos y tipos de datos estructurados son sinónimos).

- Ejemplo de estructuras de datos: los números enteros. Los números se componen de un grupo de dígitos, que tienen asociadas algunas operaciones: sumar, restar, dividir, etc.

Bases de datos (RD vs RI)

Recuperación de datos VS Recuperación de información

En bases de datos de ofimática siempre "se sabe" qué se quiere

EJEMPLO:

Valor que asume la variable X de la entidad Y, previamente conocida

¿Cuál es el Total de ventas del mes de Junio de 2010 de las sucursales del partido de Almirante Brown?

- VALOR: Total de ventas
- VARIABLE: junio de 2010
- ENTIDAD CONOCIDA: Partido de Almirante Brown

Bases de datos (RD vs RI)

Recuperación de datos VS Recuperación de información

Valor que asume la variable X de la entidad Y, previamente conocida

Consulta en SQL (*Structured Query Language*)

```
SELECT *
  FROM prestamo
  WHERE importe > 1000
   AND localidad = “Almirante Brown”
```

Consulta en lenguaje natural

Seleccionar todos los atributos de préstamo
para los préstamos de más de mil pesos de la localidad AB

Se conocen los atributos PRESTAMO, IMPORTE, LOCALIDAD, y lo que cada uno representa.

La consulta no admite errores (el resultado es exacto)

Bases de datos (RD vs RI)

Recuperación de datos VS Recuperación de información

En bases de datos de documental no siempre "se sabe" qué se quiere, ni si habrá entidades que puedan satisfacer los pedidos

EJEMPLO:

Qué entidades, desconocidas, satisfacen la condición X (o el conjunto X1 X2 … Xn)

Documentos con biografía de los directores técnicos de los equipos de fútbol de Argentina que ganaron más torneos en las últimas dos décadas

Nota: dificultad para construir una expresión de consulta, aumentada por información parcial (torneos en dos décadas), subjetividad (ganaron más torneos)

Bases de datos (RD vs RI)

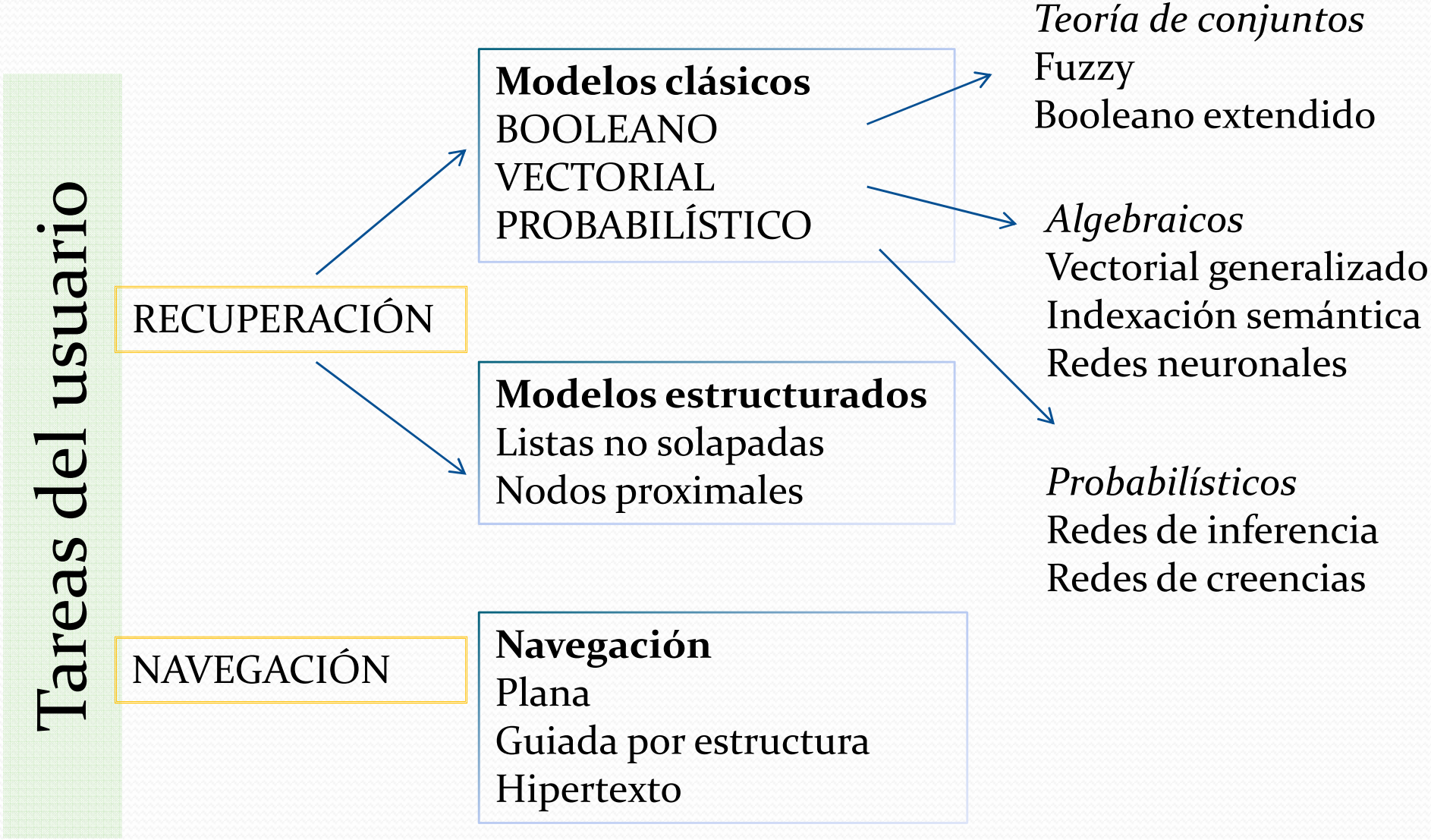
SGB

- Información estructurada
- Recuperación determinista (todo el conjunto de la solución es relevante)
- Consulta específica
- Lenguaje formal y estructurado
- Aciertos exactos

SRI

- Información semi o no estructurada
- Recuperación probabílistica (una porción de los documentos recuperados puede no ser relevante)
- Consulta imprecisa
- Lenguaje natural y no estructurado
- Aciertos parciales

Modelos de Recuperación de información



Modelos de Recuperación de información

RECUPERACIÓN
Modelos clásicos BOOLEANO VECTORIAL PROBABILÍSTICO

http://groups.google.com.ar/group/tai-2?hl=es

Unidad 1. La recuperación de información. Teoría, esquema.
Los sistemas de recuperación de información (SRI).
Tecnología y documentación. Las bases de datos.
Los modelos de recuperación de información