Ingeniería Informática

FACULTAD DE INGENIERIA
UNIVERSIDAD NACIONAL DE JUJUY

OPTATIVA

Recuperación Avanzada de Información

Unidad N° 1

Dr. J. Federico Medrano
@jfedemedrano

Temas a desarrollar

- *Organización de la asignatura*
- *Introducción.*
- *Documentos electrónicos.*
- *Modelos de recuperación de información.*
- *Algoritmos y estructuras básicas.*
- *La recuperación de información en Internet.*
- La recuperación multilingüe.
- Sistemas hipertextuales y recuperación documental.
- La recuperación de documentos multimedia o no textuales.
- La recuperación basada en la citación.
- Sistemas de filtrado y recomendación

Organización de la asignatura

- Requisitos: tener regular la asignatura:
  - Modelo de Desarrollo de Programas y Programación Concurrente (y equivalentes según nuevo plan de estudios)
- Un encuentro virtual por semana.
- Un Trabajo Práctico por cada unidad (en grupos de 2 o 3 alumnos).
- Un cuestionario evaluativo por cada unidad (solo se puede desaprobar uno solo).
- Un trabajo final para aprobar la materia (individual).

Condiciones de la asignatura

- Para regularizar
  - Registrar el 80% de la asistencia
  - Aprobar el 80% de los TTPP
  - Aprobar el 80% de los cuestionarios o su equivalente de acuerdo a la cantidad de los mismos
  - Aprobar con un 5 o 6 el TP Final
- Para Promocionar
  - Registrar el 80% de la asistencia
  - Aprobar el 100% de los TTPP
  - Aprobar el 100% de los cuestionarios
  - Aprobar con un 7 o más el TP Final

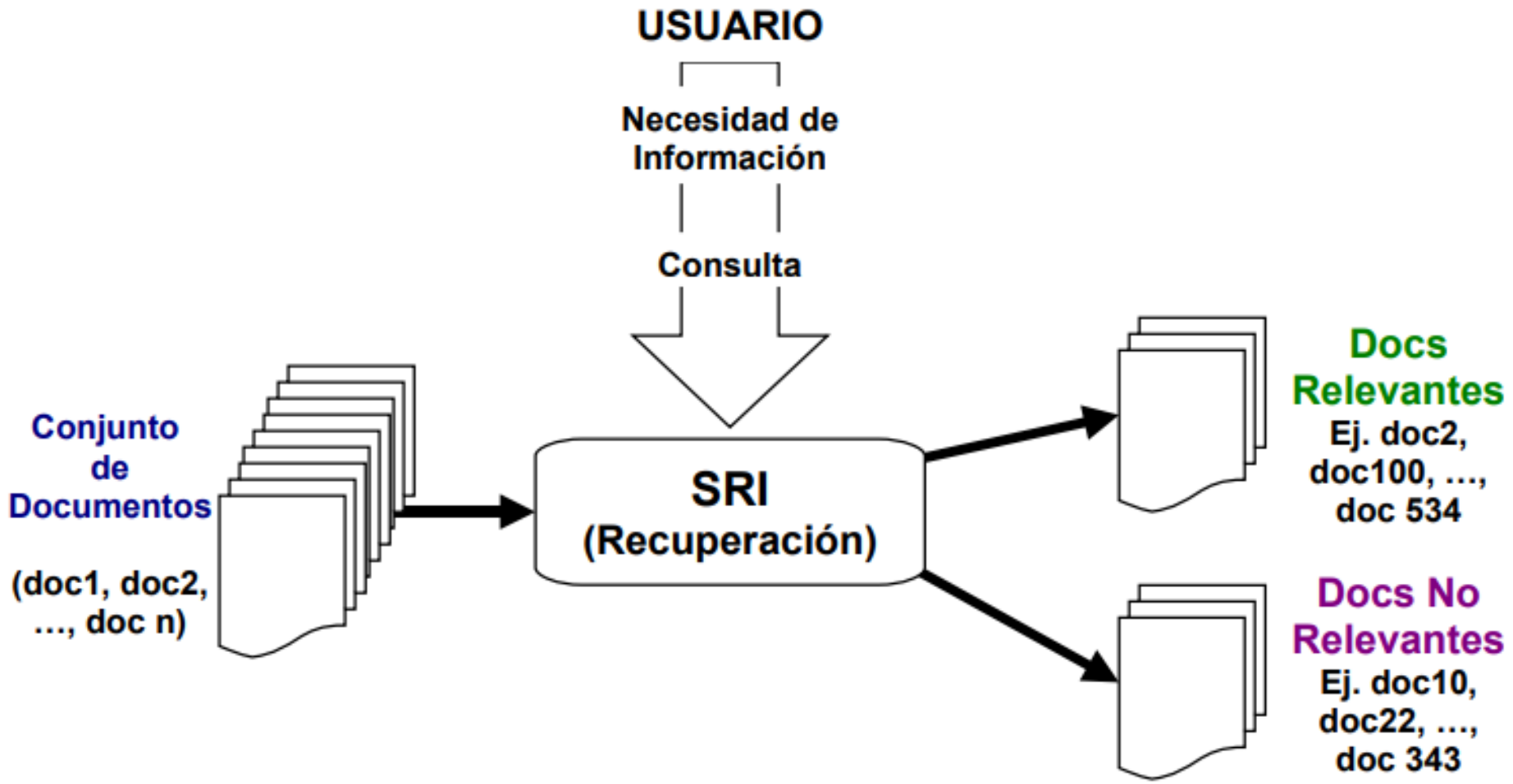
Introducción a la Recuperación de Información

La Recuperación de Información (RI)

- El término Recuperación de Información (IR) se usa para describir el proceso de “encontrar material (generalmente documentos) de naturaleza no estructurada (generalmente texto) que satisface una necesidad de información específica dentro de grandes colecciones (generalmente almacenadas en computadoras) "

- Del inglés *Information Retrieval*

Problemática de la RI



Problemática de la RI

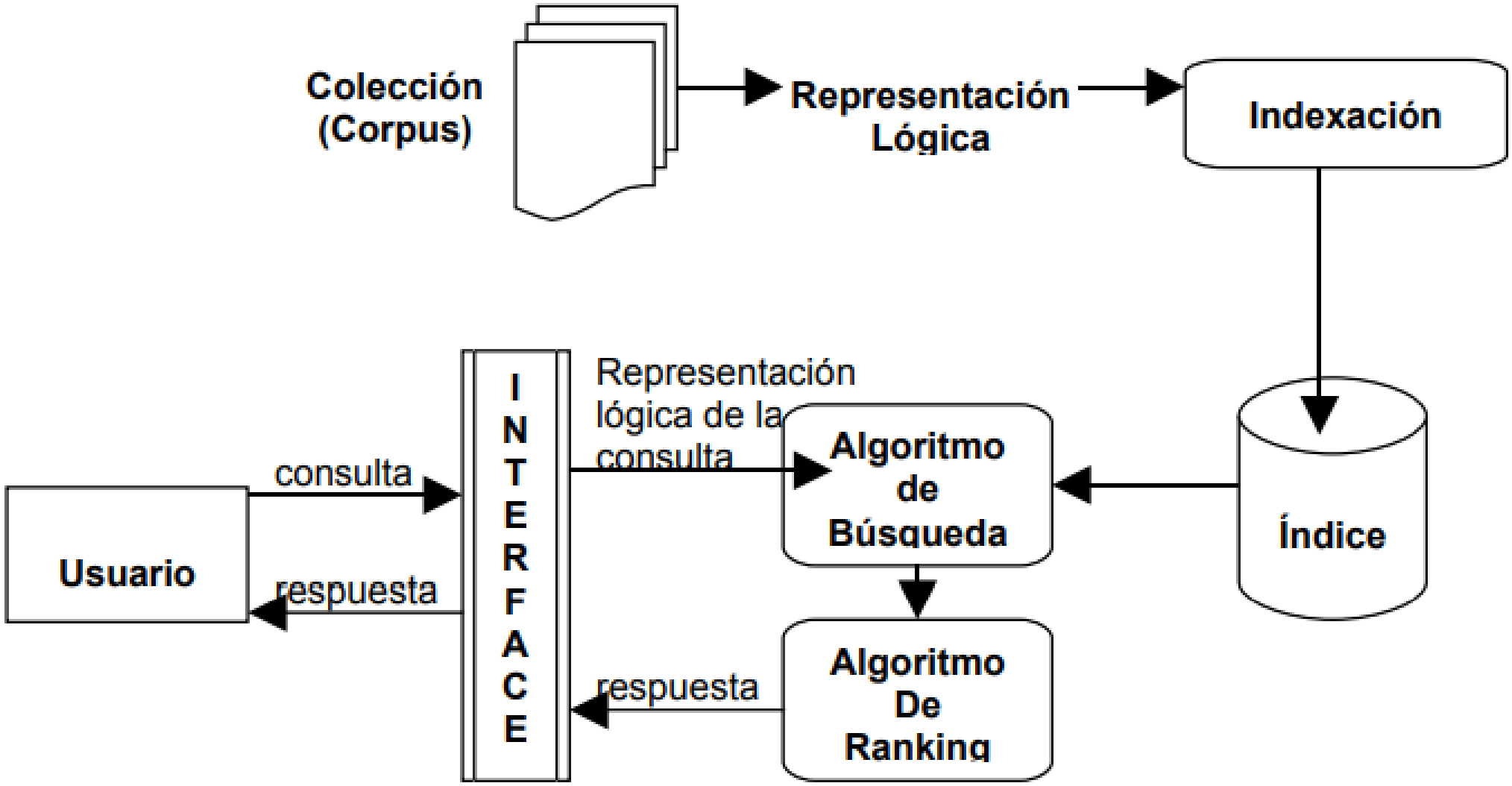
- Planteamos que la respuesta “ideal” de un SRI está formada solamente por **documentos relevantes** a la consulta, pero – en la práctica – esta no es aún alcanzable.
- El SRI recupera la mayor cantidad posible de documentos relevantes, minimizando la cantidad de documentos no relevantes (ruido) en la 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.

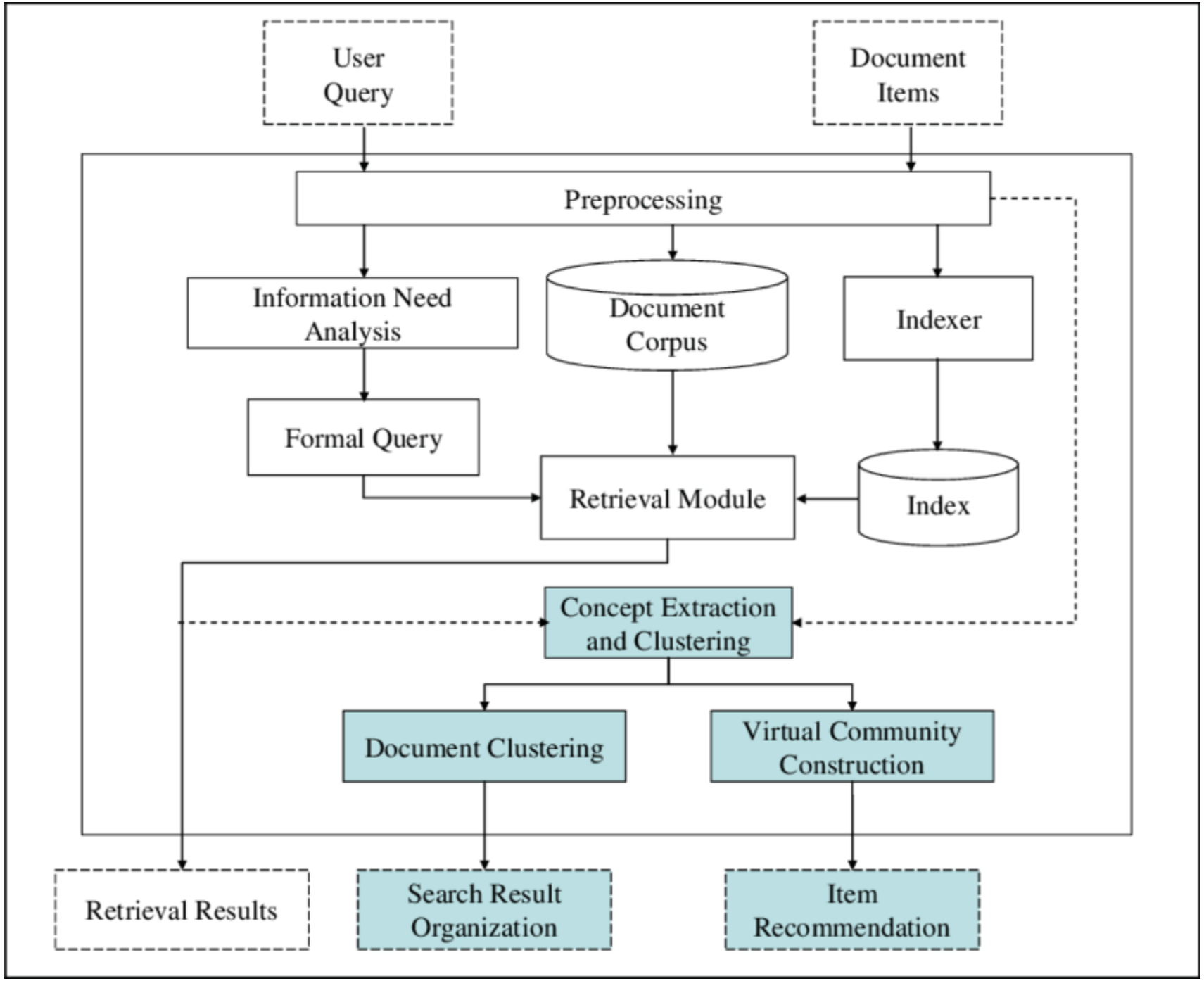
Problemática de la RI

Para cumplir con sus objetivos, un SRI debe realizar algunas tareas básicas, las cuales se encuentran 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.
- Ranking de los documentos considerados relevantes para formar el “conjunto solución” o respuesta.
- Presentación de la respuesta al usuario.
- Retroalimentación o refinamiento de las consultas (para aumentar la calidad de la respuesta)

Arquitectura básica de un SRI

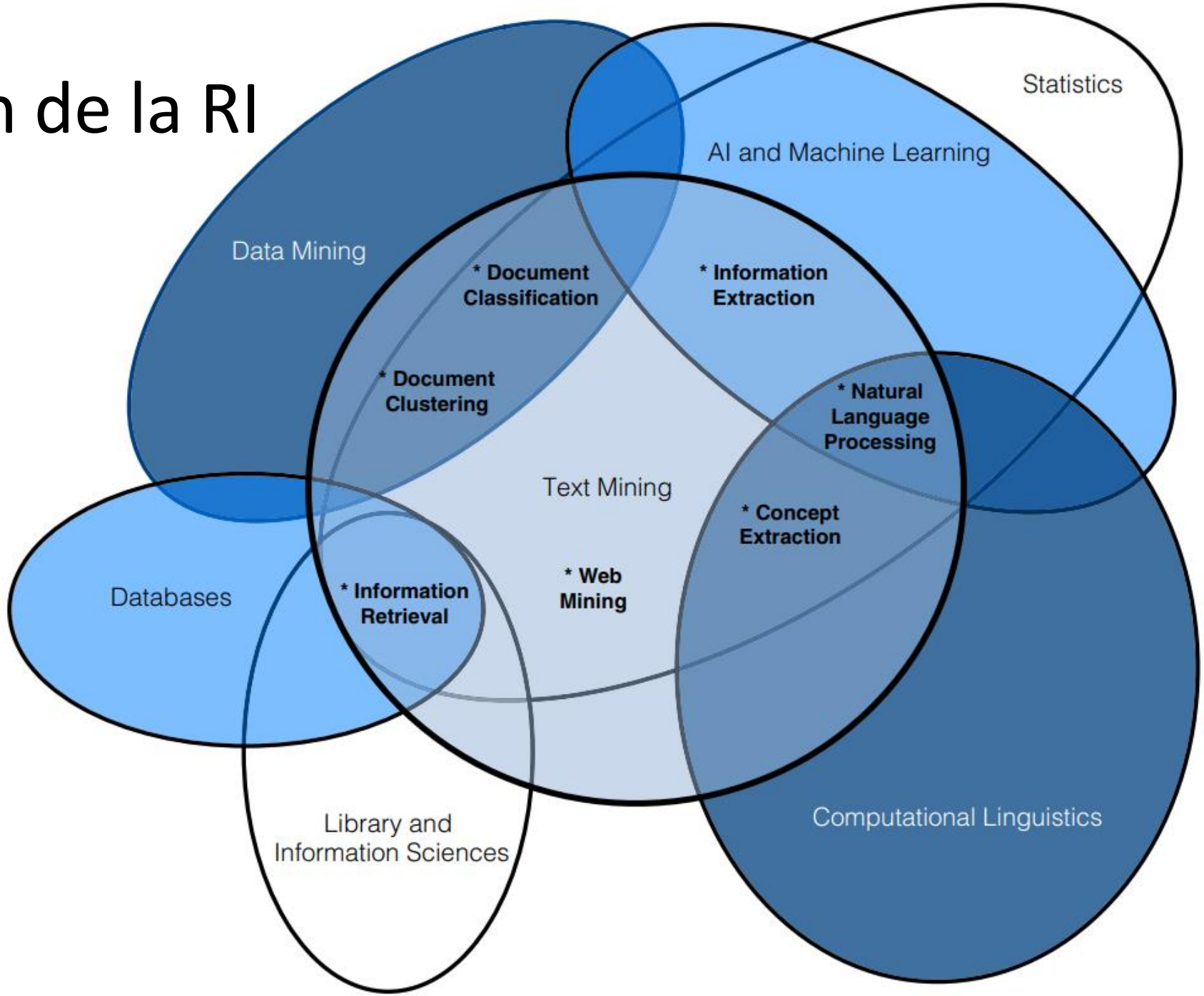



User Query
Document Items
Preprocessing
Information Need Analysis
Document Corpus
Indexer
Formal Query
Retrieval Module
Index
Concept Extraction and Clustering
Document Clustering
Virtual Community Construction
Retrieval Results
Search Result Organization
Item Recommendation

Áreas de la RI

- Modelos de Recuperación
- Filtrado y Ruteo
- Clasificación
- Agrupamiento (Clustering)
- Sumarización
- Detección de novedades (Novelty Detection)
- Respuestas a Preguntas (Question Answering)
- Extracción de Información
- Recuperación cross-language (búsqueda multilingual)
- Búsquedas Web
- Recuperación de Información Distribuida
- Modelado de Usuarios
- Recuperación de Información Multimedia
- Desarrollo de Conjuntos (data-sets) de Prueba

Inserción de la RI



Recuperación de Información vs Recuperación de Documentos

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)

“*Documentos que contengan información biográfica de los entrenadores de los equipos de fútbol de Argentina que ganaron más torneos en los últimos 10 años*”

Recuperación de Información vs Recuperación de Documentos

	SGBD	SRI
Estructura	Información estructurada con semántica bien definida.	Información semi o no estructurada.
Recuperación	Determinística. Todo el conjunto solución es relevante para el usuario	Probabilística. Una porción de los documentos recuperados puede no ser relevante.
Consulta y Lenguaje	Especificación precisa (no hay ambigüedad). Lenguaje formal, preciso y estructurado.	Hay imprecisión en su formulación. Lenguaje natural, ambiguo y no estructurado.
Resultados	Aciertos exactos	Aciertos parciales

Documentos electrónicos

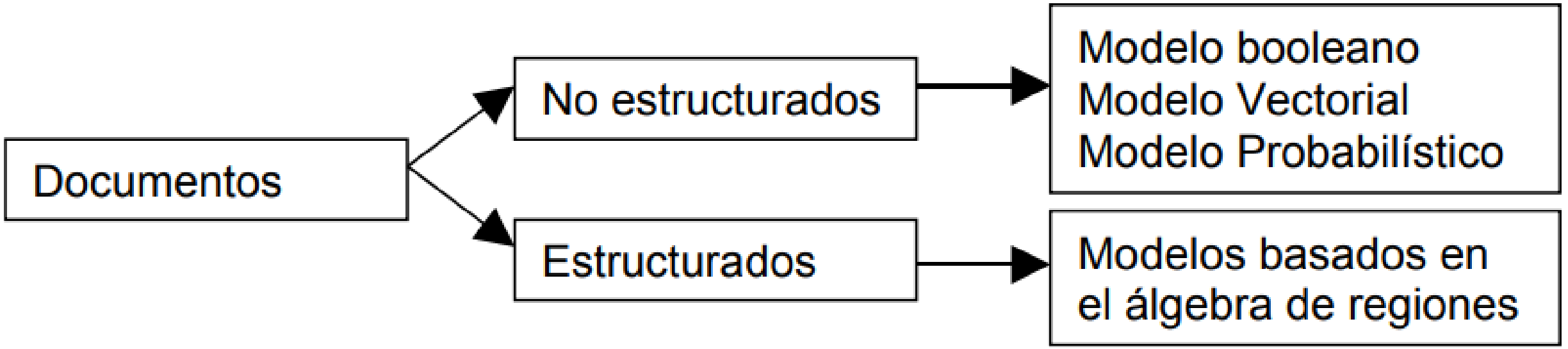
Documentos electrónicos

- Los documentos electrónicos requieren técnicas automáticas, siendo el concepto actual de RI.
- Proceso complejo que asigna automáticamente los mejores términos índice a los documentos.
- Se persigue que las consultas puedan realizarse en lenguaje natural (texto libre).
- Problemas:
  - ❑Información pobremente estructurada.
  - ❑Formatos de documentos.
  - ❑Codificación de la información.
  - ❑Problemas de detección y conversión.
  - ❑Normalización de términos (mayúsculas/minúsculas, acentos...).

Modelos de RI

Modelos de RI

- Se presenta una posible clasificación de modelos de RI – la cual no es exhaustiva – de acuerdo a características estructurales de los documentos.



Modelo booleano

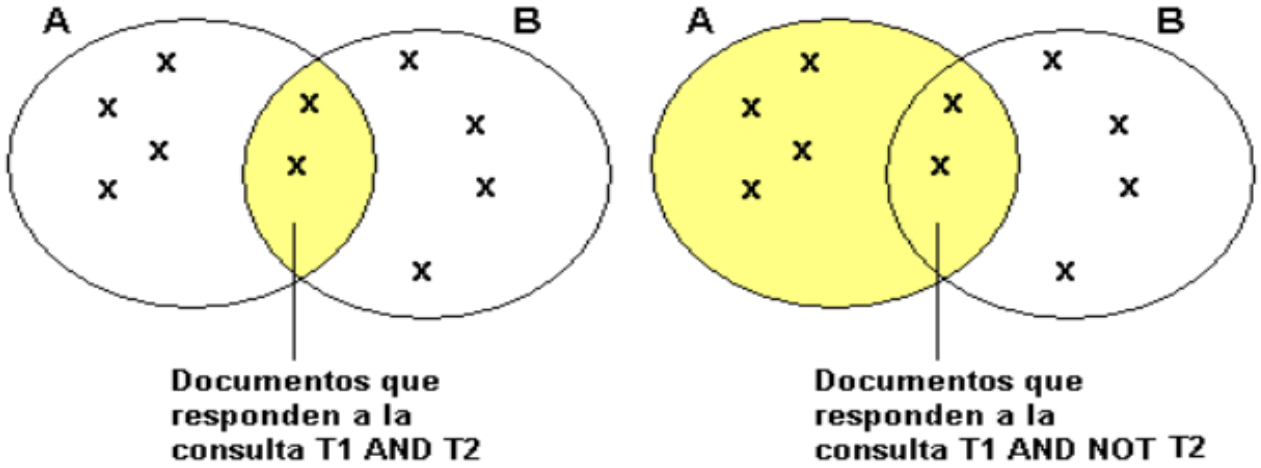
- En el modelo booleano la representación de la colección de documentos se realiza sobre una matriz binaria documento–término, donde los términos han sido extraídos manualmente o automáticamente de los documentos y representan el contenido de los mismos.

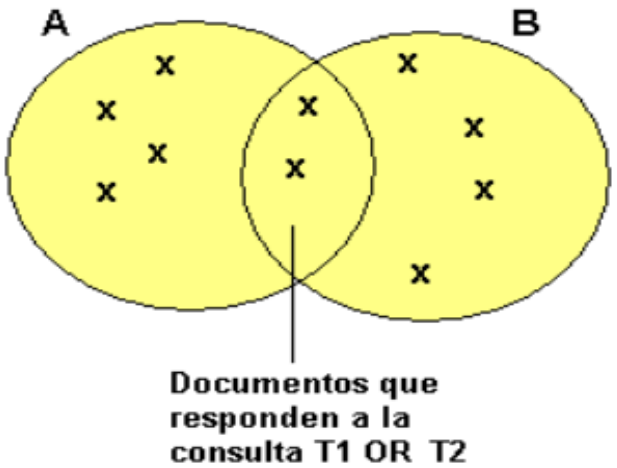
	t1	t2	t3	…	tn
d1	1	0	0		
d2	1	1	1		
d3	0	0	1		
…					
dn					

Matriz binaria documento – término

Modelo booleano

A = {Documentos que contienen el término T1}
B = {Documentos que contienen el término T2}





Modelo vectorial

- 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***.
- 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.
- 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.

Modelo vectorial

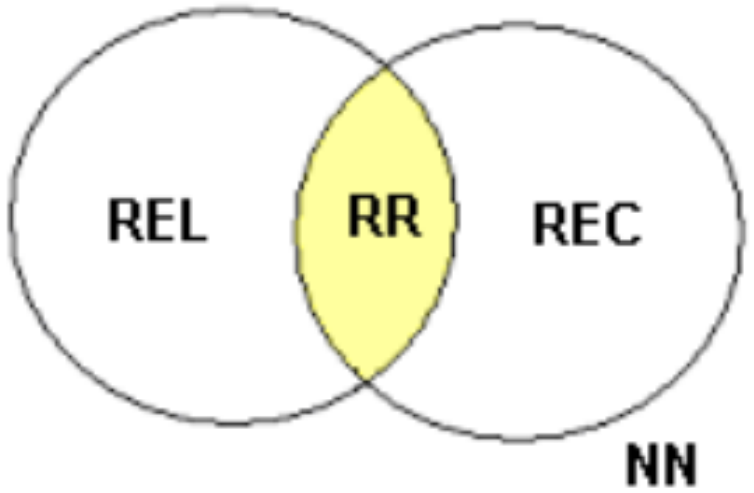
- Documento: “La República Argentina ha sido nominada para la realización del X Congreso Americano de Epidemiología en Zonas de Desastre. El evento se realizará ...”
- Consulta: “argentina congreso epidemiología”

	argentina	**...**	**congreso**	**epidemiología**	**...**
$\mathbf{d_1}$	0.5		0.3	0.2	
...					
$\mathbf{d_n}$					
Consulta	0.4		0.3	0.3	

Matriz término-documento con pesos normalizados entre 0 y 1

Modelo probabilístico

- A partir de una expresión de consulta se puede dividir una colección de N documentos 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.
- El resultado ideal de 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 performance de recuperación



Modelo para documentos estructurados

- Un modelo de recuperación de documentos estructurados utiliza la estructura de los mismos a los efectos de mejorar la performance y brindar servicios alternativos al usuario.
- 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 PDF (Portable Document Format), el XML (eXtensible Markup Language) y LATEX.

Modelo para documentos estructurados

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

Estructuras de datos en RI

Matriz documento–término

Sobre las filas se representan los documentos dj y las columnas corresponden a los términos ti del vocabulario. La intersección (d_j, t_i) corresponde al peso o ponderación que se le asigna a t_i en d_j. En el caso más simple puede ser un 0 o un 1, denotando ausencia o presencia del término en el documento o bien un valor surgido de un criterio de ponderación como – por ejemplo – TF-IDF

	t_1	t_2	t_3	…	t_n
d_1	$w_{1,1}$	$w_{1,2}$	$w_{1,3}$		$w_{1,n}$
d_2	$w_{2,1}$	$w_{2,2}$	$w_{2,3}$		$w_{2,n}$
d_3	$w_{3,1}$	$w_{3,2}$	$w_{3,3}$		$w_{3,n}$
d_4	$w_{4,1}$	$w_{4,2}$	$w_{4,3}$		$w_{4,n}$
d_5	$w_{5,1}$	$w_{5,2}$	$w_{5,3}$		$w_{5,n}$
…					
d_n	$w_{n,1}$	$w_{n,2}$	$w_{n,3}$		$w_{n,n}$

Fichero invertido

Partiendo de la matriz documento–término se construye un índice que almacena su representación. La estructura de índice básica es la denominada “archivo invertido”. En su forma más simple, es un conjunto de términos donde cada uno tiene asociada una lista de los identificadores de documentos donde cada término aparece

	t_1	t_2	t_3	…	t_n
d_1	$w_{1,1}$	$w_{1,2}$	$w_{1,3}$		$w_{1,n}$
d_2	$w_{2,1}$	$w_{2,2}$	$w_{2,3}$		$w_{2,n}$
d_3	$w_{3,1}$	$w_{3,2}$	$w_{3,3}$		$w_{3,n}$
d_4	$w_{4,1}$	$w_{4,2}$	$w_{4,3}$		$w_{4,n}$
d_5	$w_{5,1}$	$w_{5,2}$	$w_{5,3}$		$w_{5,n}$
…					
d_n	$w_{n,1}$	$w_{n,2}$	$w_{n,3}$		$w_{n,n}$

	d_1	d_2	d_3	…	d_n
t_1	$w_{1,1}$	$w_{1,2}$	$w_{1,3}$		$w_{1,n}$
t_2	$w_{2,1}$	$w_{2,2}$	$w_{2,3}$		$w_{2,n}$
t_3	$w_{3,1}$	$w_{3,2}$	$w_{3,3}$		$w_{3,n}$
t_4	$w_{4,1}$	$w_{4,2}$	$w_{4,3}$		$w_{4,n}$
t_5	$w_{5,1}$	$w_{5,2}$	$w_{5,3}$		$w_{5,n}$
…					
t_n	$w_{n,1}$	$w_{n,2}$	$w_{n,3}$		$w_{n,n}$

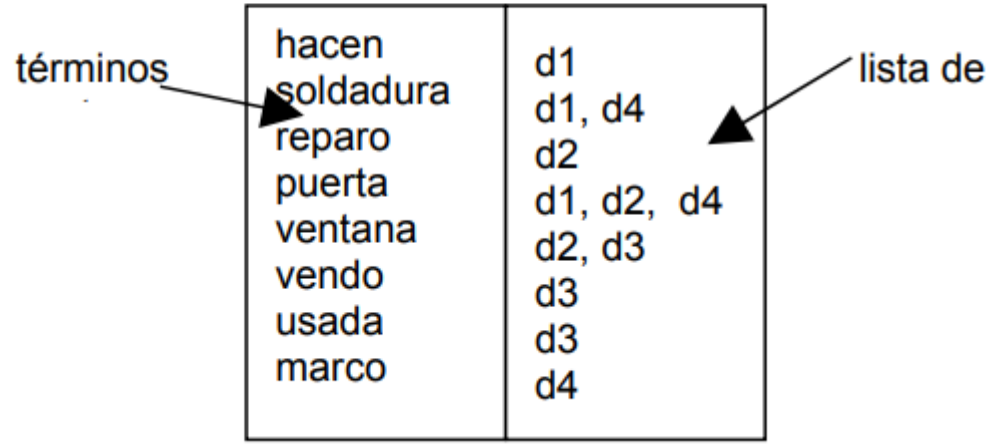
Fichero invertido

El archivo invertido que soporta tal representación consta de dos partes: la primera es un término y la segunda la lista de documentos donde éste aparece, denominada lista de posteo (posting list). Nótese que el conjunto de todos los términos corresponde al vocabulario de la colección y – por supuesto – no tiene repetidos.

d1 = {Se hacen soldadura de puertas}
d2 = {Reparo puertas y ventanas}
d3 = {Vendo ventanas usadas}
d4 = {Soldadura de puertas y marcos}

	t_1	t_2	t_3	…	t_n
d_1	$w_{1,1}$	$w_{1,2}$	$w_{1,3}$		$w_{1,n}$
d_2	$w_{2,1}$	$w_{2,2}$	$w_{2,3}$		$w_{2,n}$
d_3	$w_{3,1}$	$w_{3,2}$	$w_{3,3}$		$w_{3,n}$
d_4	$w_{4,1}$	$w_{4,2}$	$w_{4,3}$		$w_{4,n}$
d_5	$w_{5,1}$	$w_{5,2}$	$w_{5,3}$		$w_{5,n}$
…					
d_n	$w_{n,1}$	$w_{n,2}$	$w_{n,3}$		$w_{n,n}$

	d_1	d_2	d_3	…	d_n
t_1	$w_{1,1}$	$w_{1,2}$	$w_{1,3}$		$w_{1,n}$
t_2	$w_{2,1}$	$w_{2,2}$	$w_{2,3}$		$w_{2,n}$
t_3	$w_{3,1}$	$w_{3,2}$	$w_{3,3}$		$w_{3,n}$
t_4	$w_{4,1}$	$w_{4,2}$	$w_{4,3}$		$w_{4,n}$
t_5	$w_{5,1}$	$w_{5,2}$	$w_{5,3}$		$w_{5,n}$
…					
t_n	$w_{n,1}$	$w_{n,2}$	$w_{n,3}$		$w_{n,n}$



Fichero invertido con frecuencia

Una alternativa es armar la lista de posteo con información acerca de la frecuencia del término t_i en cada documento d_j. En este caso, cada ítem de la *posting list* es un par ordenado (d_j, t_f), donde d_j es el documento y t_f es la frecuencia del término t_i en éste.

t_1	$(d_1, 2)$ $(d_3, 4)$ $(d_7, 8)$
t_2	$(d_2, 5)$ $(d_7, 3)$
t_3	$(d_4, 1)$
…	
t_{n-1}	$(d_1, 1)$ $(d_2, 2)$ $(d_4, 5)$
t	(d 3) (d 1)

En este ejemplo, el término t_1 aparece dos veces en el documento d_1, cuatro veces en d_3 y ocho veces en d_7.

Fichero invertido posicional

Un archivo invertido posicional ó índice posicional es una estructura de datos que incluye las posiciones o desplazamientos donde cada término ocurre, con respecto al inicio del documento.

Esta estructura almacena información de localización de los términos, lo que permite – entre otras – búsquedas por cercanía de términos y soporta búsquedas por frases

t_1	(d_1: 2, 10) (d_3: 4, 23, 54, 101) (d_7: 16, 54, 122)
t_2	(d_2, 50, 178) (d_7, 20, 62)
t_3	(d_4, 100)
…	
t_{n-1}	(d_1, 44) (d_2, 211) (d_4, 251)
t	(d 32) (d 101)

Fichero invertido posicional

- Como se puede apreciar, la frecuencia de un término en un documento corresponde a la cantidad de posiciones donde aparece. Este tipo de estructura de datos permite soportar consultas por proximidad, por ejemplo:

❑Q1 = {red adyacente computadora}

❑Q2 = {red cerca-de computadora}

❑Q3 = {red a-n-términos computadora}

❑Q4 = {red misma-oración computadora}

La RI en la web

La RI en la web

- La web puede ser vista como un gran repositorio de información, completamente distribuido sobre Internet y accesible por gran cantidad de usuarios.
- Su estructura no es estática.
- Su contenido no respeta estándares de calidad, ni estilos ni organización.
- Una de las características de la información publicada en la web es su dinámica, dado que pueden variar en el tiempo tanto los contenidos como su ubicación.
- El tamaño de la web es imposible de medir exactamente y muy difícil de estimar.
- 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 RI en la web

- 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, 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.