Introducción a la Recuperación de información

La búsqueda de la información en fuentes de datos no estructuradas (grandes corpus documentales y la Web)

LA BÚSQUEDA DE INFORMACIÓN

GUIÓN

Recuperación de información en grandes corpus documentales no estructurados

- Arquitectura genérica de un sistema de RI
  - El modelo booleano
  - El modelo vectorial
  - El modelo probabilístico

Recuperación de información en La Web

- Directorios de búsqueda vs motores de búsqueda
- Arquitectura genérica de un sistema de RI en Web y funcionamiento básico
- La Web oculta

Bibliografía y referencias

BÚSQUEDA DE INFORMACIÓN EN CORPUS

INTRODUCCIÓN

Objetivo de los sistemas de Recuperación de Información –RI- o Information Retrieval –*IR*-

-Dada una colección de documentos (corpus documental) y una necesidad de información de un determinado usuario expresada en forma de pregunta (*query*) recuperar (extraer o listar) los documentos para resolver la necesidad de información del usuario

-Ejemplo 1:

- **Corpus:** Noticias de un periódico
- **Consulta:** Partido Barcelona-Madrid

-Ejemplo 2:

- **Corpus:** Historial médico de todos los ciudadanos de un país
- **Consulta:** Área donde viven los enfermos de legionela

BÚSQUEDA DE INFORMACIÓN EN CORPUS

INTRODUCCIÓN

Componentes básicos de un sistema de RI

- Un formalismo para representar cada uno de los documentos
- Un formalismo para representar las consultas (generalmente *keyword-based interfaces*)
- Una medida de similitud entre un documento y una consulta

Posible solución

- Dada una consulta recorrer secuencialmente toda la colección de documentos (corpus) comparando el contenido de cada documento con las palabras de la consulta.
- Emparejamiento sintáctico de patrones (*Matching* sintáctico simple)
- Solución eficiente cuando se trabaja con muy pocos documentos

BÚSQUEDA DE INFORMACIÓN EN CORPUS

INTRODUCCIÓN

Problemas derivados del *matching* sintáctico

- **Polisemia de las palabras**
  - **Ejemplo 1:** Un documento en el que aparezcan las palabras partido, Barcelona y Madrid no tiene porque ser relevante para una consulta sobre los partidos de fútbol entre el Barcelona y el Madrid.

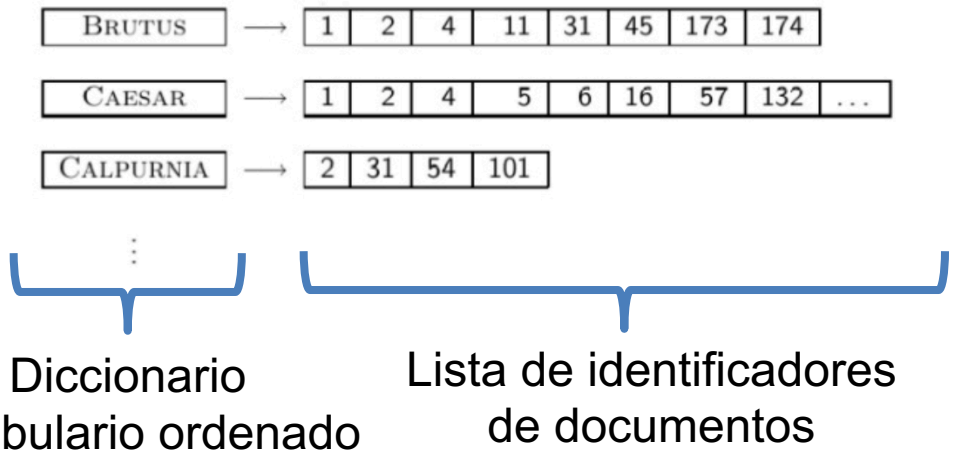
- **Diferentes representaciones del mismo concepto** (la sinonimia):
  - **Ejemplo 2:** Un documento en el que **NO aparezcan** las palabras enfermo y legionela puede ser relevante para la consulta. Por ejemplo, un documento que trate los síntomas de neumonía en un paciente o un documento que trate la enfermedad con su nombre científico *Legionella pneumophila pueden ser relevantes*.

BÚSQUEDA DE INFORMACIÓN EN CORPUS

INTRODUCCIÓN

Problemas derivados del recorrido secuencial del corpus documental

- **Para corpus > 200 Mb requiere "demasiado" tiempo:**
  - **Índices invertidos** para acceder a los documentos
  - Estructura eficiente de almacenamiento donde a cada término (palabra) se le asocia una lista de los documentos donde aparece dicho término.

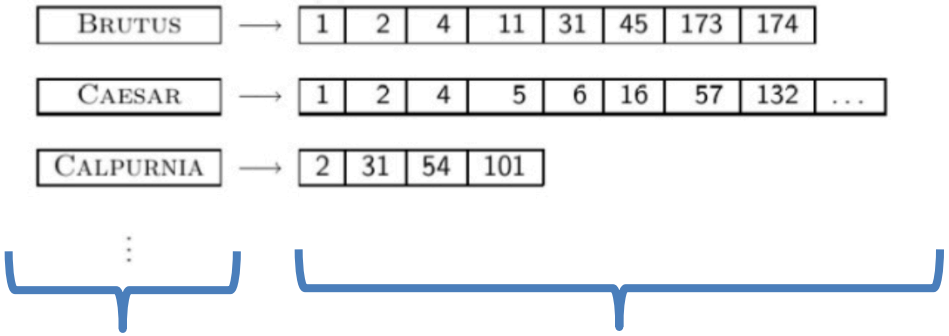


  - **Las entradas (elementos del diccionario): claves por las que se realiza el acceso**
  - **Las salidas (listas): elementos a los que se desear acceder o recuperar**

BÚSQUEDA DE INFORMACIÓN EN CORPUS

INTRODUCCIÓN

Necesidad de una medida que evalúe la relevancia entre una pregunta y un documento



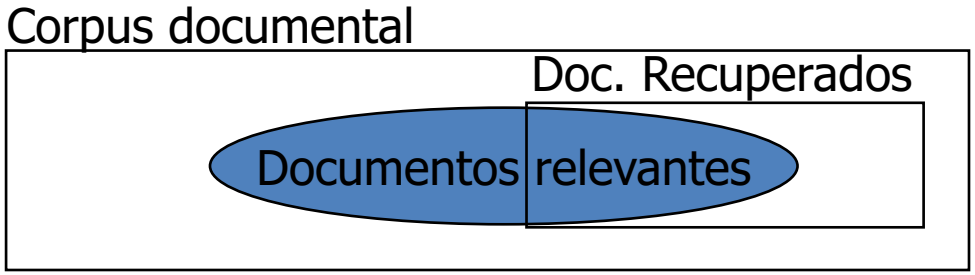
- **Dada la pregunta Caesar ¿Le proporciono toda la lista de documentos? ¿Se la proporciono en cualquier orden?**
  - Ranking de documentos. Los más relevantes en primer lugar
  - Diferentes modelos de representación de documentos y medidas de similitud entre queries y documentos:
    - » Booleano, Vectorial y Probabilístico

BÚSQUEDA DE INFORMACIÓN EN CORPUS

INTRODUCCIÓN

Como evaluar un determinado sistema de recuperación de información

- **Medidas estándar de efectividad:**
  - **Precision (P):** nº de doc. relevantes recuperados/ nº total de doc. recuperados
  - **Recall (R):** nº de doc. relevantes recuperados/ nº total de doc. Relevantes



  - **F-measurement:** Media harmónica de P y R: $\{(1+B^2)\ P\ R\}/\{B^2\ (P + R)\}$
  - **Corpus TREC (1992)**
- **Medidas de eficiencia:**
  - Menor cantidad de recursos empleados (tiempo de computación en la creación de estructuras, espacio de almacenamiento, tiempo de computación en una búsqueda, etc.)

BÚSQUEDA DE INFORMACIÓN EN CORPUS

INTRODUCCIÓN

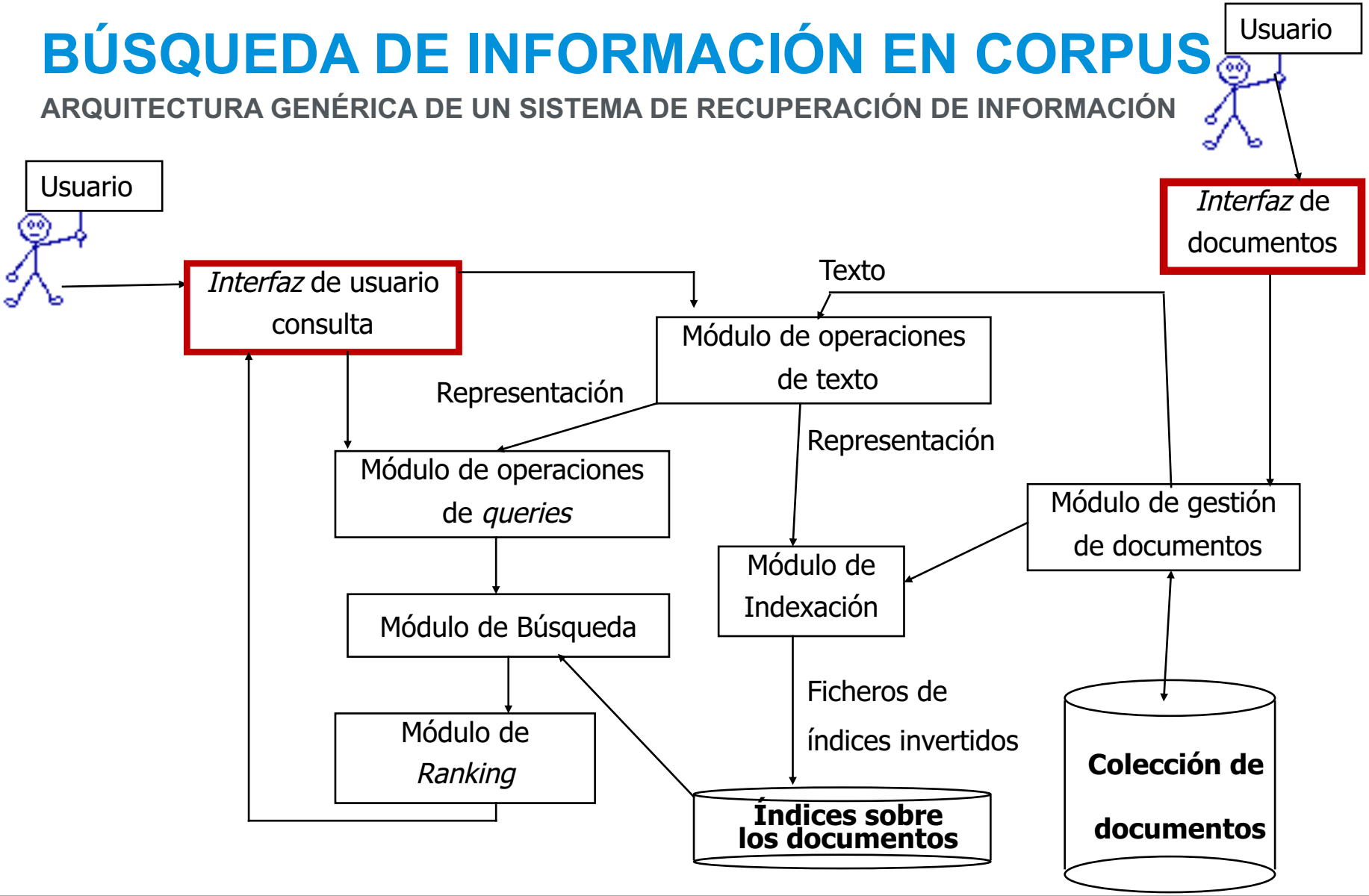
RI no sólo trata modelos para representar documentos y consultas (queries) sino también:

- Métodos de almacenamiento de documentos
- Métodos de compresión para reducir el espacio que ocupa el almacenamiento de los documentos y los índices:
  - Fundamentalmente se distingue entre técnicas de compresión que permiten la búsqueda sobre texto comprimido y entre las que no (las primeras son más eficientes)

- Métodos de indexación de documentos
- Métodos de presentación (Ranking)
- Otros:
  - Clasificación automática de documentos (*text-classification*)
  - Agrupamiento de documentos similares (*clustering*) una búsqueda, etc.)

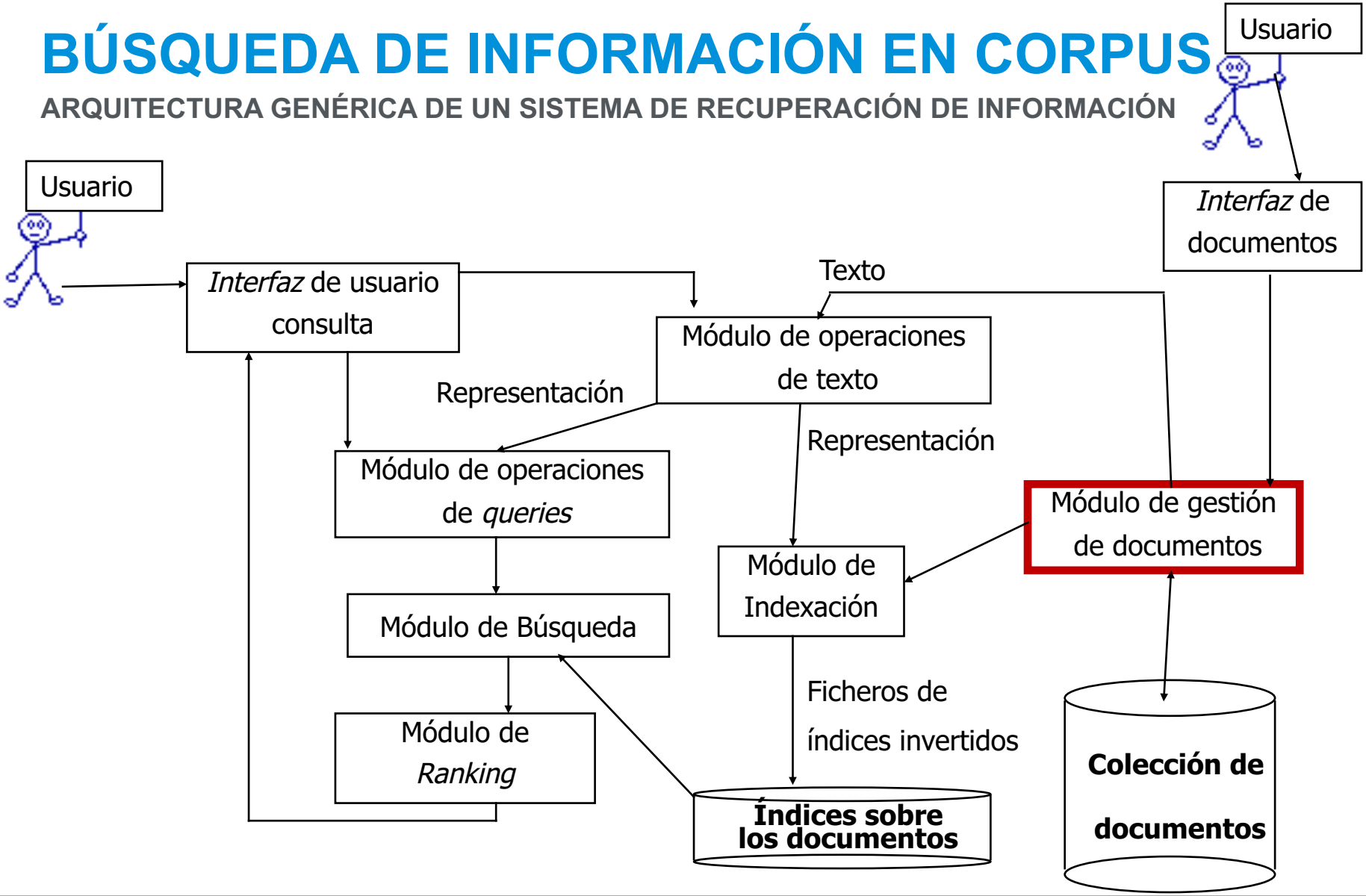
BÚSQUEDA DE INFORMACIÓN EN CORPUS

ARQUITECTURA GENÉRICA DE UN SISTEMA DE RECUPERACIÓN DE INFORMACIÓN



BÚSQUEDA DE INFORMACIÓN EN CORPUS

ARQUITECTURA GENÉRICA DE UN SISTEMA DE RECUPERACIÓN DE INFORMACIÓN



BÚSQUEDA DE INFORMACIÓN EN CORPUS

ARQUITECTURA GENÉRICA DE UN SISTEMA DE RECUPERACIÓN DE INFORMACIÓN

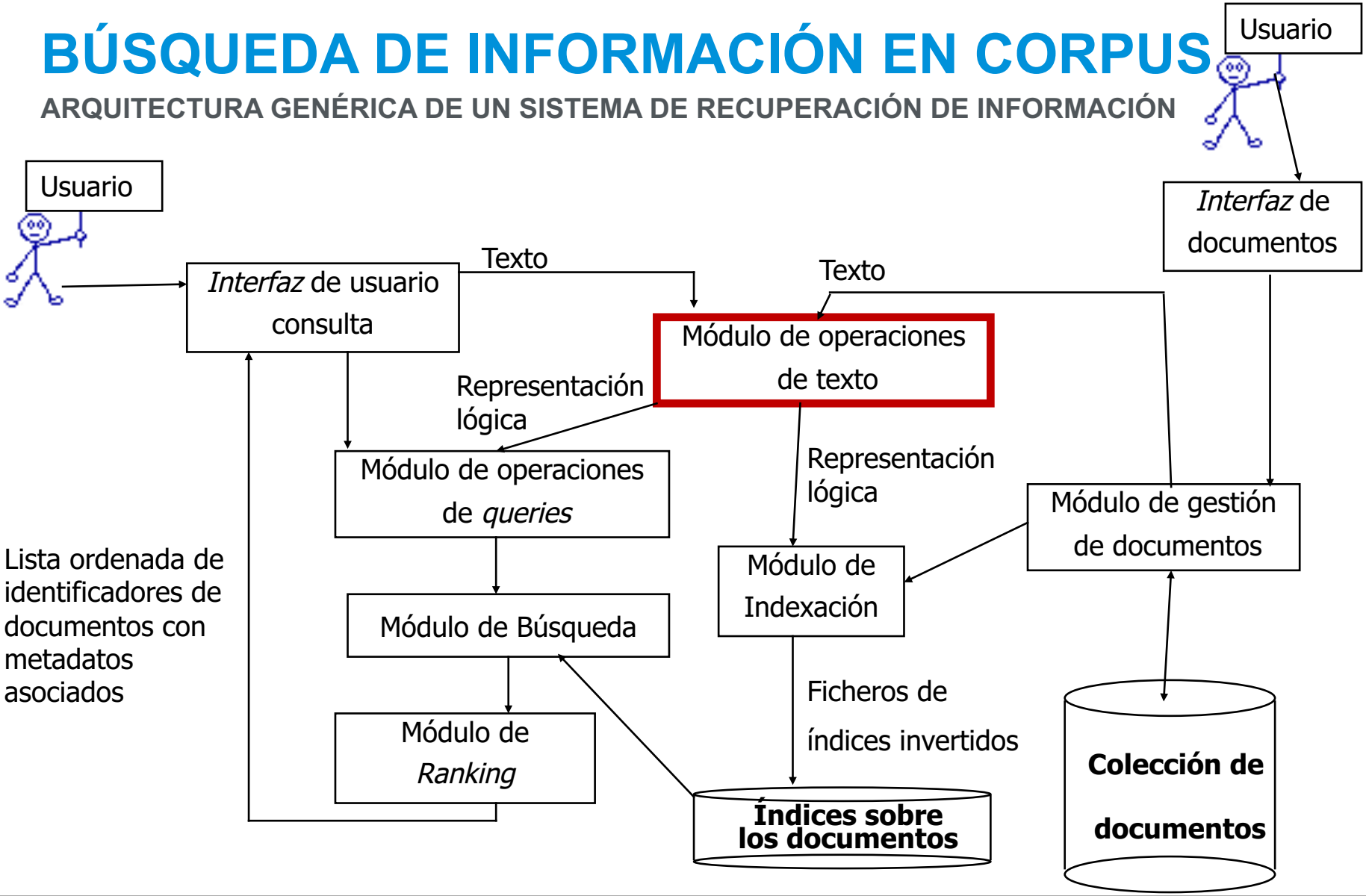
Módulo de gestión de documentos

- **Gestionar los documentos del corpus y los metadatos que tienen asociados**
  - Insertar, actualizar y/o eliminar un documento del corpus
  - *Parsear* los documentos que forman la colección para extraer información de ellos (autor, título, palabras clave, extraer metadatos, clasificar los documentos...)

  - Dado un determinado identificador de documento recuperar el documento
  - Dado un determinado identificador de documento consultar los metadatos que tiene asociados (fecha de creación, fecha de modificación, autor, etc.)

BÚSQUEDA DE INFORMACIÓN EN CORPUS

ARQUITECTURA GENÉRICA DE UN SISTEMA DE RECUPERACIÓN DE INFORMACIÓN



BÚSQUEDA DE INFORMACIÓN EN CORPUS

ARQUITECTURA GENÉRICA DE UN SISTEMA DE RECUPERACIÓN DE INFORMACIÓN

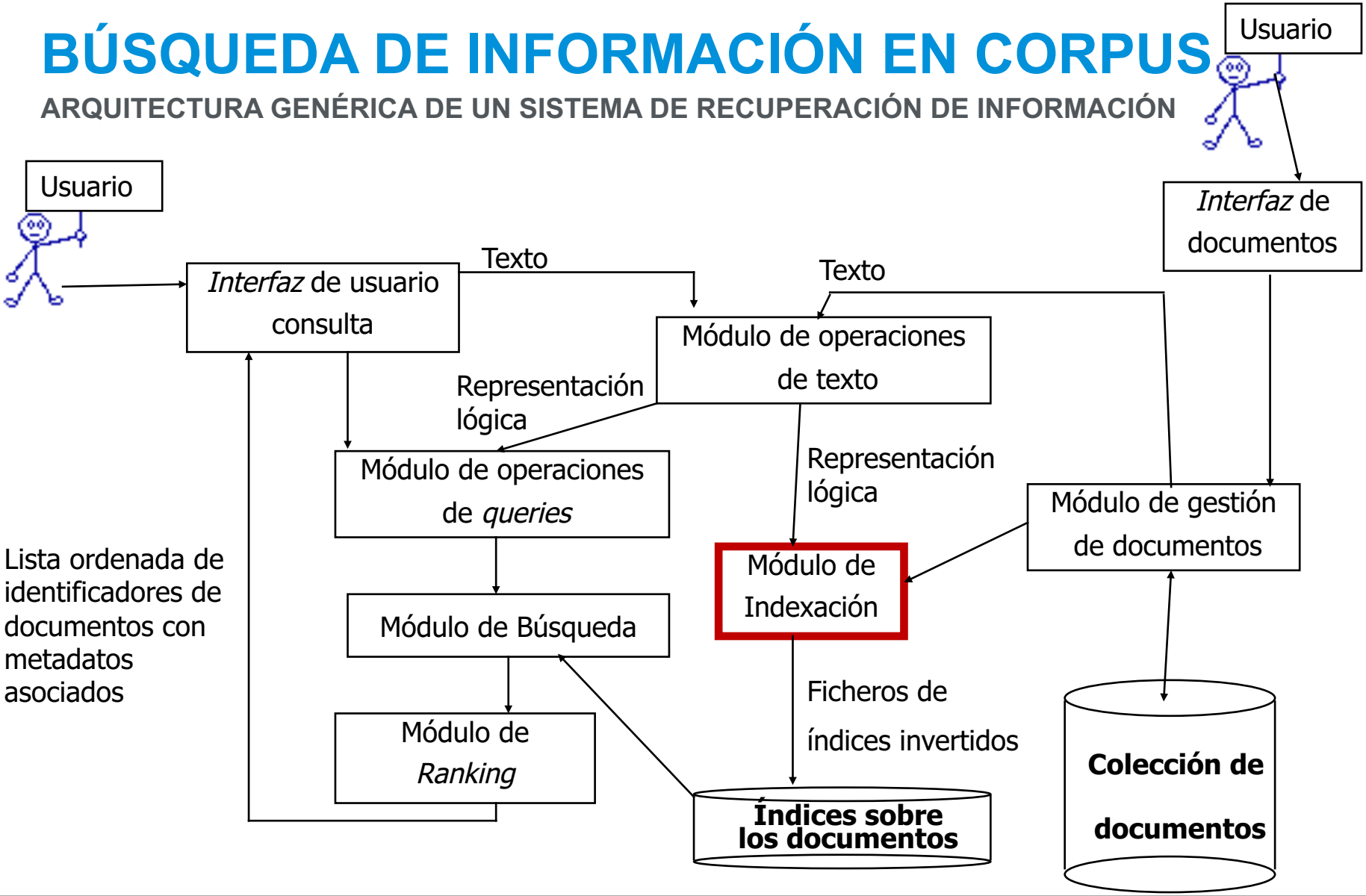
Módulo de operaciones de texto

- **Transformar el documento/consulta original en una representación del mismo/de la misma (vista lógica):**

  - En general la vista lógica consiste en una secuencia de términos.
  - Técnicas tradicionalmente empleadas:
    - Lista de palabras sin contenido semático o listas de *stopwords* (*stoplist*)
    - Lematización (*stemming*). Considerar las raíces semánticas (lemas o *stem)* de los términos.
  - Otras técnicas empleadas:
    - Procesamiento de lenguaje natural (NLP)
      - » Nombres compuestos: “bases de datos”
      - » Nombres de entidades: George Bush
      - » Determinar el significado con el que actúa una palabra polisémicas (desambiguación semántica).

BÚSQUEDA DE INFORMACIÓN EN CORPUS

ARQUITECTURA GENÉRICA DE UN SISTEMA DE RECUPERACIÓN DE INFORMACIÓN



BÚSQUEDA DE INFORMACIÓN EN CORPUS

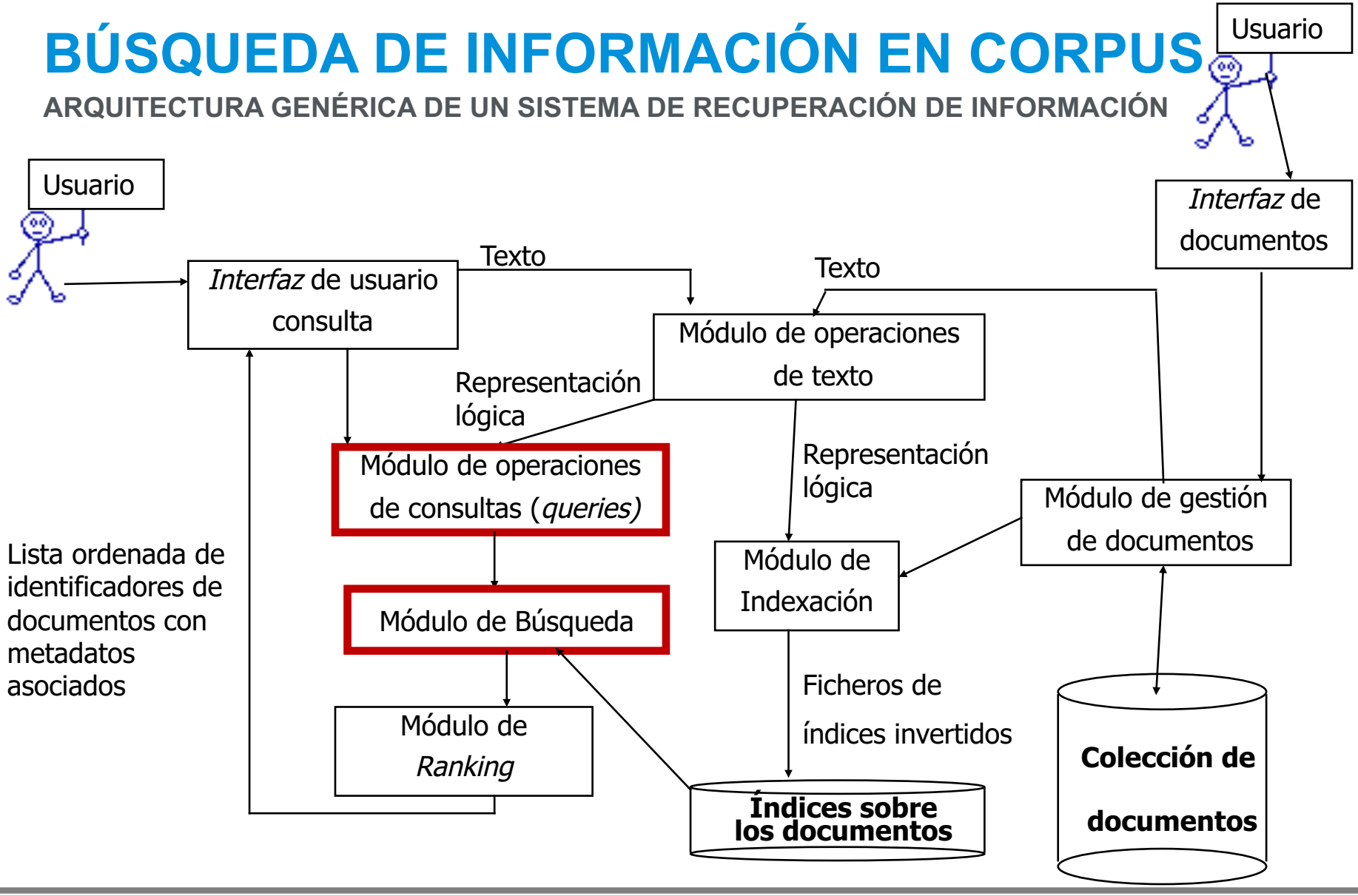
ARQUITECTURA GENÉRICA DE UN SISTEMA DE RECUPERACIÓN DE INFORMACIÓN

Módulo de indexación

- **Gestionar los índices sobre los documentos:**
  - Los más conocidos: Índices invertidos
  - Determinar la unidad de indexación (el término): Palabra, frase, oración, conjunto de 2, 3, 4, … palabras o términos (*n-grams*), raíz léxica de una palabra (lema o *stem*)
  - Determinar qué información va a almacenar el índice:
    - lista de identificadores de documentos,
    - posiciones donde aparece el término dentro del documento
    - frecuencia del término

BÚSQUEDA DE INFORMACIÓN EN CORPUS

ARQUITECTURA GENÉRICA DE UN SISTEMA DE RECUPERACIÓN DE INFORMACIÓN



BÚSQUEDA DE INFORMACIÓN EN CORPUS

ARQUITECTURA GENÉRICA DE UN SISTEMA DE RECUPERACIÓN DE INFORMACIÓN

Módulo de operaciones de consulta

- **Uso de recursos lingüísticos como Thesaurus Léxicos (WordNet) y ontologías para enriquecer la representación lógica de la consulta**
  - Expansión (o eliminación) de términos de la representación (vista lógica) de la consulta.
    - Ejemplo: 'coche' pasa a ser 'coche o auto o carro' (sinónimos)

      'coche automóvil' pasa a ser 'coche o auto o carro' (hiperónimo)

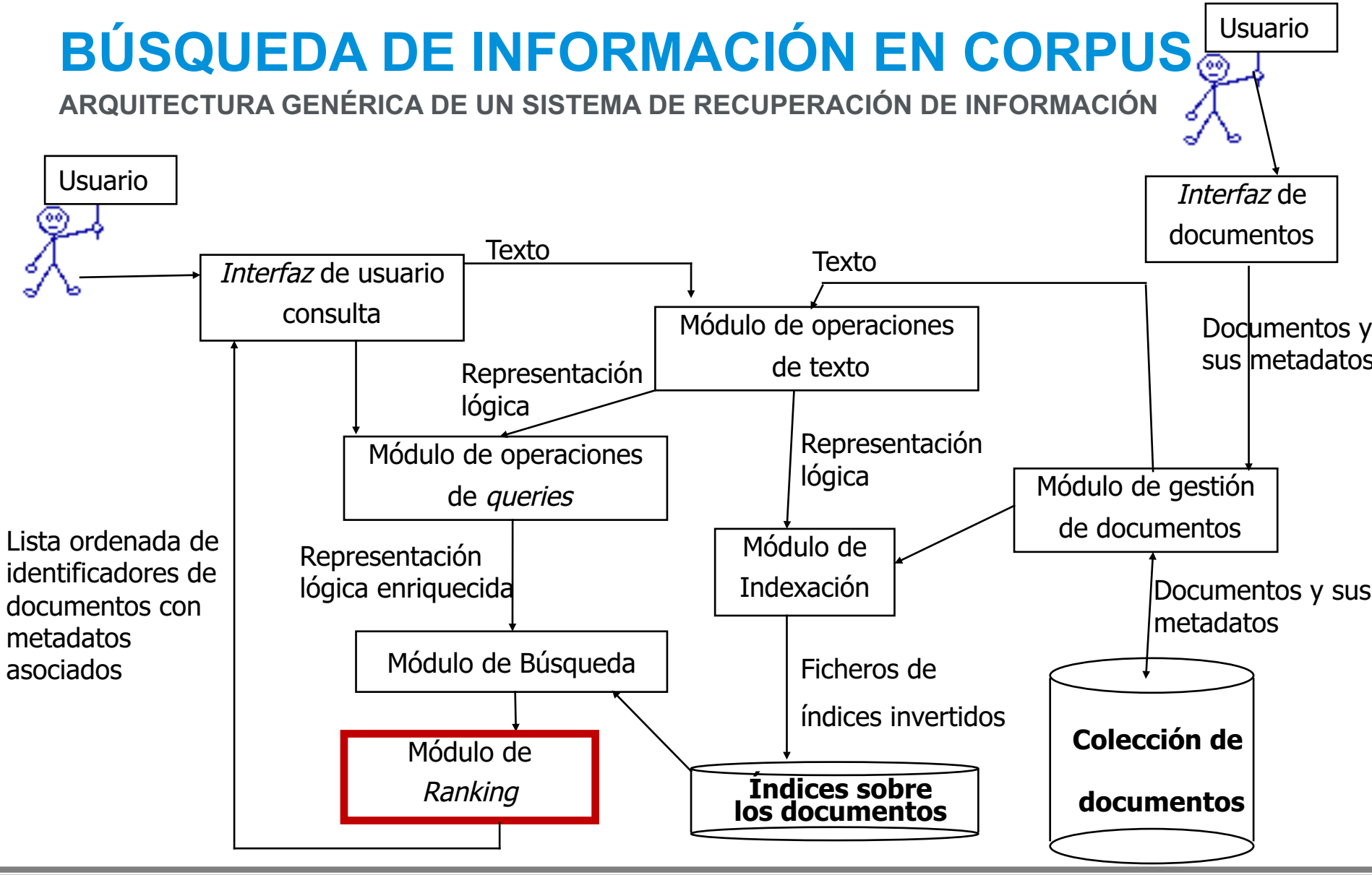
  - Solicitud de retroalimentación al usuario (*relevance-feedback*) para especificar en mayor detalle su consulta
    - Ejemplo: con 'manzana' te refieres a la fruta o al logotipo de una empresa informática

Módulo de operaciones de búsqueda

- **Dada la representación lógica de una consulta (query) realiza consultas en los índices para determinar cuáles son los documentos relevantes para dicha consulta**
    - Opción más sencilla: *matching sintáctico*

BÚSQUEDA DE INFORMACIÓN EN CORPUS

ARQUITECTURA GENÉRICA DE UN SISTEMA DE RECUPERACIÓN DE INFORMACIÓN



BÚSQUEDA DE INFORMACIÓN EN CORPUS

ARQUITECTURA GENÉRICA DE UN SISTEMA DE RECUPERACIÓN DE INFORMACIÓN

Módulo de ranking

- **Ordena los documentos recuperados de acuerdo con la relevancia respecto a la consulta que se está resolviendo**
  - Medida de similitud empleada entre las representaciones lógicas de las consultas y las representaciones lógicas de los documentos
    - » Modelo booleano
    - » Modelo vectorial
    - » Modelo probabilístico
  - Puede tener en cuenta otros aspectos:
    - – Contexto en el que se encuentra el usuario (hora del día, dispositivo, etc.)
    - – Preguntas anteriores formuladas por el usuario

BÚSQUEDA DE INFORMACIÓN EN CORPUS

MODULO DE BÚSQUEDA Y RANKING

Modelo booleano

- **1º modelo de recuperación de información**
  - Bibliotecas: Buscar información en los resúmenes de los libros.
  - Documentos legales en E.E.U.U., el sistema Lexis Nexis.
  - Historiales médicos

- **Cada documento se representa por una lista de bits (0 o 1) que indican si en ese documento aparece determinado término del vocabulario del corpus no.**

- **Ejemplos:**
  - Vocabulario: CASA, SER, VERDE, AZUL, AMARILLA, CARA, BARATA.
  - Documento: “Mi casa es amarilla y barata” -> (1,1,0,0,1,0,1)
  - Consulta 1 (query 1): Casa barata -> (1,0,0,0,0,0,1)
  - Consulta 2 (query 2): Casa and (amarilla or azul) -> (1,0,0,1,0,0,0) or (1,0,0,0,1,0,0)

BÚSQUEDA DE INFORMACIÓN EN CORPUS

MÓDULO DE BÚSQUEDA Y RANKING

Modelo booleano

- **Ejemplos:**
  - Consulta (query): (coste OR precio) AND papel
  - Documento 1: “El coste del papel aumentó un 5%” (Documento relevante)
  - Documento 2: “El precio de los alimentos aumentó” (Documento no relevante)

- **Ventajas:**
  - Sencillo de implementar y rápido debido a que usa operaciones a nivel de bit.

- **Inconvenientes:**
  - Recupera o muy pocos o demasiados documentos (no es sencillo realizar el ranking)
  - Dificultad de los usuarios para expresar consultas booleanas

BÚSQUEDA DE INFORMACIÓN EN CORPUS

MÓDULO DE BÚSQUEDA Y RANKING

Modelo booleano extendido

- **En lugar de considerar sólo 0 y 1 en los vectores que constituyen las representaciones lógicas de los documentos, se considera el número de veces que aparece un término en un documento**

- **Ventajas:**
  - Se puede hacer un ranking de los documentos recuperados: los que tienen mayor número de apariciones en los términos indicados por el usuario van antes que los documentos con menos apariciones en dichos términos.

- **Dio lugar al Modelo Vectorial (Vector Space Model)**

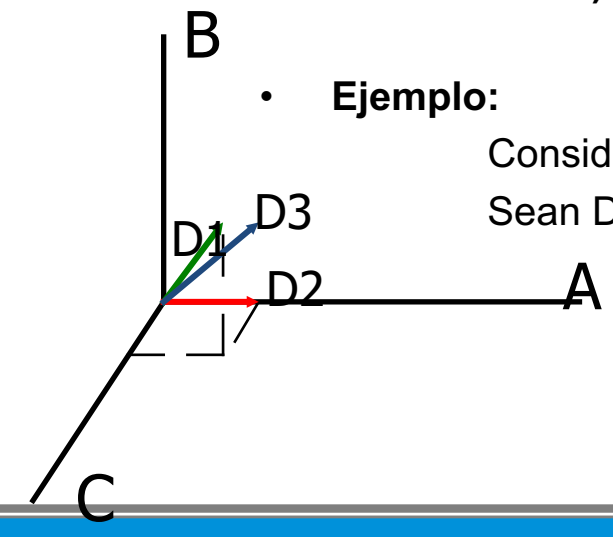
BÚSQUEDA DE INFORMACIÓN EN CORPUS

MÓDULO DE BÚSQUEDA Y RANKING

Modelo vectorial

- Marcó el inicio de la investigación en el campo de la RI por ofrecer altas prestaciones y permitir *ranking*.

- **Las *queries (consultas)* y los documentos se representan mediante vectores cuya dimensión es la cardinalidad del vocabulario (conjunto de términos considerados)**



- **Ejemplo:**

  Considerar un vocabulario de 3 términos A, B y C

  Sean Di documentos : D1 contiene los términos A, B, y C una vez

  D2 contiene el término A una vez

  D3 contiene los términos A y B una vez

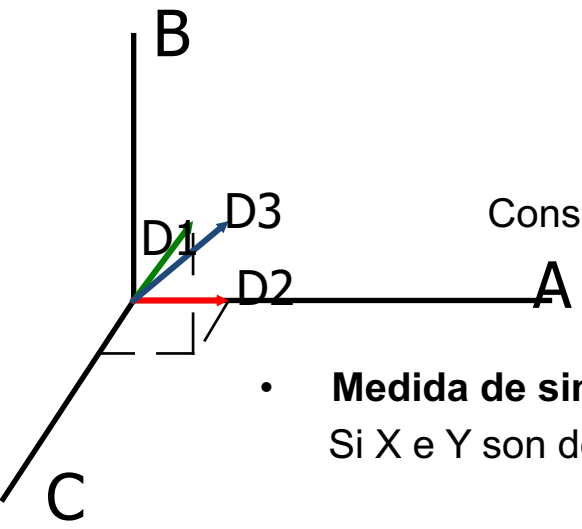
BÚSQUEDA DE INFORMACIÓN EN CORPUS

MÓDULO DE BÚSQUEDA Y RANKING

Modelo vectorial

- **Ejemplo:**

Sean Di documentos : D1 contiene los términos A, B, y C una vez <1,1,1>
D2 contiene el término A una vez <1,0,0>
D3 contiene los términos A y B una vez <1,1,0>



Consulta (query): documentos con A y B
q = <1,1,0>

- **Medida de similitud entre dos vectores: coseno del ángulo que forman**

Si X e Y son dos vectores y alfa el ángulo que forman:

X Y = |X| |Y| cos (alfa)

Cos (alfa) = X Y / |X| |Y| y Cos (alfa) € [0-1]

Mayor similitud entre X e Y, menor es el ángulo que forman → cos(alfa) tiende a 1

Cuanto más diferentes sean X e Y, mayor el ángulo que forman -> cos(alfa) tiende a 0

BÚSQUEDA DE INFORMACIÓN EN CORPUS

MODULO DE BÚSQUEDA Y RANKING

Modelo vectorial

- **Ejemplo:**

q D1 = 1 1 + 1 1 + 0 1 = 2
|q| = sqrt(1^2 + 1^2 + 0^2) = sqrt(2)
|D1| = sqrt(1^2 + 1^2 + 1^2) = sqrt(3)
Cos(q y D1) = 2/(sqrt(2) sqrt(3)) = 2 / (1.414)(1.732) = 2 / 2.449 = 0,81

q D2 = 1 1 + 1 0 + 0 0 = 1
|D2| = sqrt(1^2 + 0^2 + 0^2) = sqrt(1)=1
Cos(q y D2) = 1/(sqrt(2) 1)= 1/sqrt(2)= 1 / 1.414 = 0.707

q D3 = 1 1 + 1 1 + 0 0 = 2
|D3| = sqrt(1^2 + 1^2 + 0^2) = sqrt(2)
Cos(q y D3) = 2/(sqrt(2) sqrt(2))= 2 / 2 = 1

BÚSQUEDA DE INFORMACIÓN EN CORPUS

MÓDULO DE BÚSQUEDA Y RANKING

Modelo vectorial

- Existen muchas variantes considerando otras medidas de similitud:
  - Dice: similaridad (X, Y) = (2 X Y)/ (X^2Y^2)
  - Jaccard: similaridad (X, Y) = (X Y) / (X^2 + Y^2 - |X||Y|)

- Otras variantes:
  - Dividir la frecuencia de los términos por la longitud del documento.
  - En lugar de considerar la frecuencia de los términos (tf: term frequency), considerar también la frecuencia inversa de los términos en la colección de documentos (idf= Número total de documentos/Nº de documentos con el término t).
    - Las palabras de menor frecuencia en un corpus documental son más informativas
  - Utilizar log(idf) en lugar de valores absolutos
    - No es lo mismo cambiar de 1 a 2 que de 325 a 326

BÚSQUEDA DE INFORMACIÓN EN CORPUS

MÓDULO DE BÚSQUEDA Y RANKING

Modelo vectorial

- **Ventajas:**
  - Formalismo matemático sencillo de implementar
  - Poco coste computacional
  - Librerías que ya nos los proporcionan (SMART, Lucene)
  - Alto rendimiento
- **Inconvenientes:**
  - Asume independencia de términos (cada dimensión se trata de forma independiente de las otras)
  - No tiene en cuenta el tamaño de los documentos (en documentos más grandes es más probable encontrar más términos)

Modelo probabilístico

- Estrategia adaptativa basada en probabilidades condicionadas y el teorema de Bayes

BÚSQUEDA DE INFORMACIÓN EN LA WEB

INTRODUCCIÓN

Diferencias entre Corpus y Web

- **Necesidad de localizar los documentos con los que vamos a trabajar**
  - Localización manual (directorios de búsqueda)
  - Localización automática (*crawlers* o arañas)
  - Híbrida (las arañas localizan y los usuarios clasifican)

Directorios de búsqueda

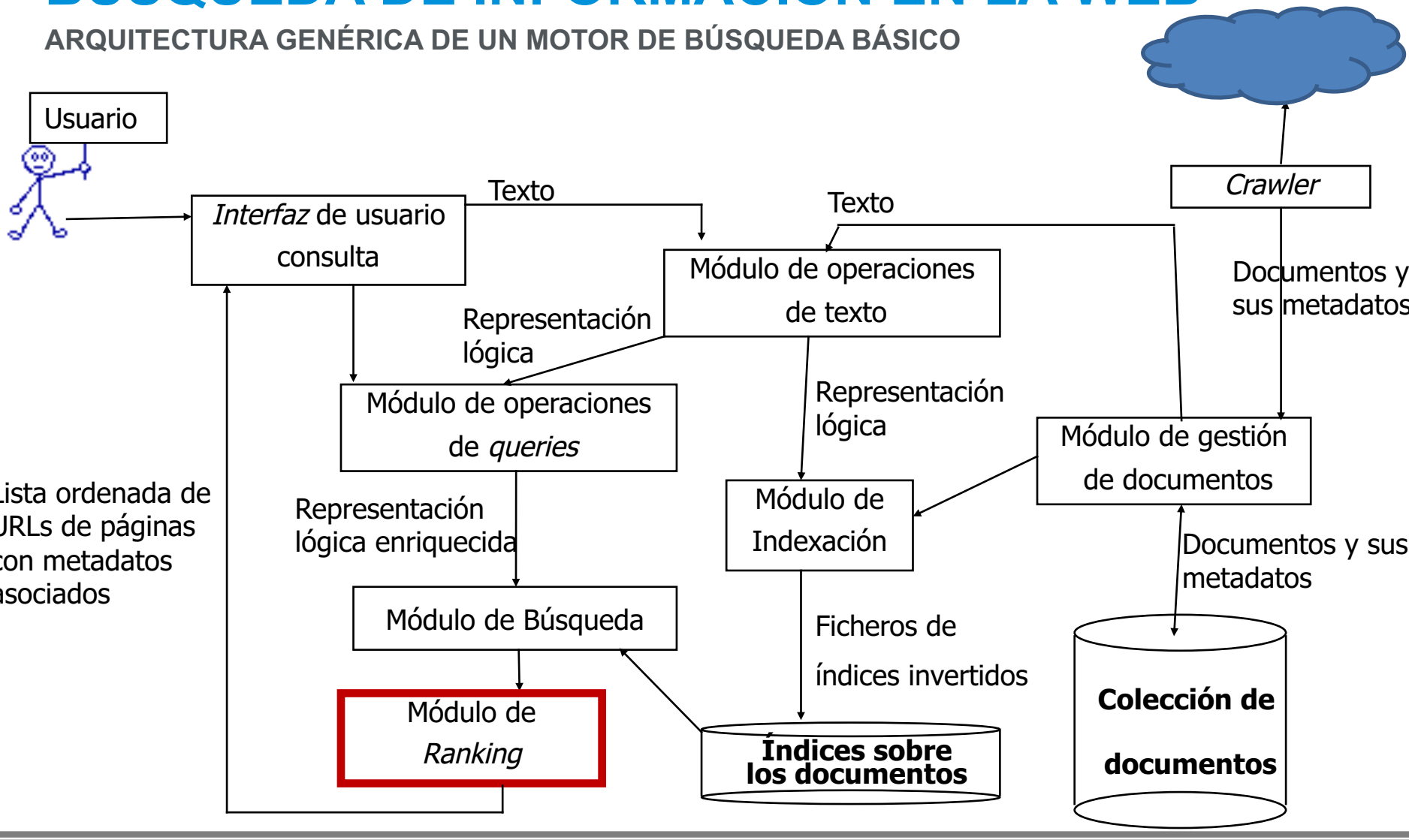
- **Organización manual** de las páginas en categorías (análogo a las carpetas)
- Ejemplos: Yahoo en sus inicios, DMOZ
- **Inconvenientes**: Escalabilidad, definición de la estructura

Motores de búsqueda o buscadores

- Adaptación de las técnicas de RI en grandes corpus a la Web mediante el uso de arañas (*crawlers)* generalmente con interfaces basadas en palabras clave
- Ejemplos: Altavista, Google, AlltheWeb, Bing

BÚSQUEDA DE INFORMACIÓN EN LA WEB

ARQUITECTURA GENÉRICA DE UN MOTOR DE BÚSQUEDA BÁSICO



BÚSQUEDA DE INFORMACIÓN EN LA WEB

ARQUITECTURA GENÉRICA DE UN MOTOR DE BÚSQUEDA BÁSICO

Motores de búsqueda o buscadores

- Construcción de un gran índice de palabras sobre todos los documentos del web estático.
- Búsquedas por palabras clave sobre el índice, obteniendo granularidad de documento.

Aspectos a tener en cuenta en los buscadores

- Construcción del índice (arañas o crawling)
- Distribución del índice (gran volumen de información)
- Algoritmo de ejecución distribuida de consultas.
- Algoritmos de relevancia mucho más críticos. Hay que sacar partido de la estructura proporcionada por los hiperenlaces
- Problema de la web oculta

BÚSQUEDA DE INFORMACIÓN EN LA WEB

ARQUITECTURA GENÉRICA DE UN MOTOR DE BÚSQUEDA BÁSICO

Algunos ejemplos de motores de búsqueda o buscadores

- **Altavista**
  - Arquitectura basada en grandes servidores -> No escala
  - Algoritmos de relevancia de las páginas basados en los tradicionales de bases de datos documentales:
    - » Se inundaron por la cantidad de información.
    - » Apenas uso de los hiperenlaces.
  - Sólo tratan la web estática
- **Google**
  - Arquitectura distribuida basada en miles de estaciones de bajo precio
  - Algoritmos de relevancia que sacan partido a la estructura basada en hiperenlaces.
  - Siguen limitándose a la web estática aunque investigan en la línea de la web oculta.

BÚSQUEDA DE INFORMACIÓN EN LA WEB

META BUSCADORES

- Definición: Buscadores que buscan en buscadores y luego integran sus resultados en tiempo real (p.e.,: www.dogpile.com).
- Se han utilizado para mejorar la relevancia en buscadores en Internet. Mediante algoritmos de ponderación: MetaCrawler, SavvySearch, Multibuscador,...
- En entorno corporativo surgen como soluciones de 'búsqueda federada'

- Dificultades:
  - Traducción de consultas del formato general al de la fuente: Traducción de sintaxis y post-procesados
  - Construcción de 'envoltorios' sobre los buscadores origen.
  - Relevancia ponderada de resultados: relevancia del origen, de la fuente...
  - Eficiencia en las consultas

BÚSQUEDA DE INFORMACIÓN EN LA WEB

EL ALGORITMO PAGE-RANK

- Los enlaces son considerados como ‘citas’ de otros documentos.
- Se asumen como más relevantes lo documentos más citados pero también importa quién es el que te cita.
- Una página tiene un pagerank alto si tiene muchas páginas que la apuntan o la apuntan páginas con un PageRank alto (“Hubs”).
- El texto en los enlaces se asocia también a la página destino.

BÚSQUEDA DE INFORMACIÓN EN LA WEB

EL ALGORITMO HITS

Relevancia basada en hiperenlaces fue: HITS

- Búsqueda previa sobre un índice pre-construido.
- Algorimo iterativo sobre los enlaces entre documentos.
- Hubs son páginas que enlazan muchas ‘páginas buenas’ (autoridades)
- Autoridades son páginas enlazadas desde muchos ‘referentes buenos’ (hubs).
- Si buscamos autoridades relacionadas con un cierto tema, no llega con que sean páginas apuntadas desde muchas otras: debe existir cierto solape entre las páginas que las apuntan (hubs).

Procesamiento básico en HITS

- Se realiza la búsqueda en un motor tradicional.
- Se expanden sus resultados con páginas que ‘son apuntadas’ por páginas de los resultados y con páginas que ‘apuntan’ a páginas resultados, hasta un cierto nivel de profundidad.
- Cada página (nodo del grafo) comienza con un ‘peso de hub’ y un ‘peso de autoridad’
- En cada iteración, el peso de’autoridad’ de un nodo se calcula como la suma del ‘peso de hub’ de la iteración anterior de los nodos que lo apuntan.
- El ‘peso de hub’ se calcula como la suma del ‘peso de autoridad’ de los nodos a los que apunta.
- Se demuestra que el algoritmo converge.

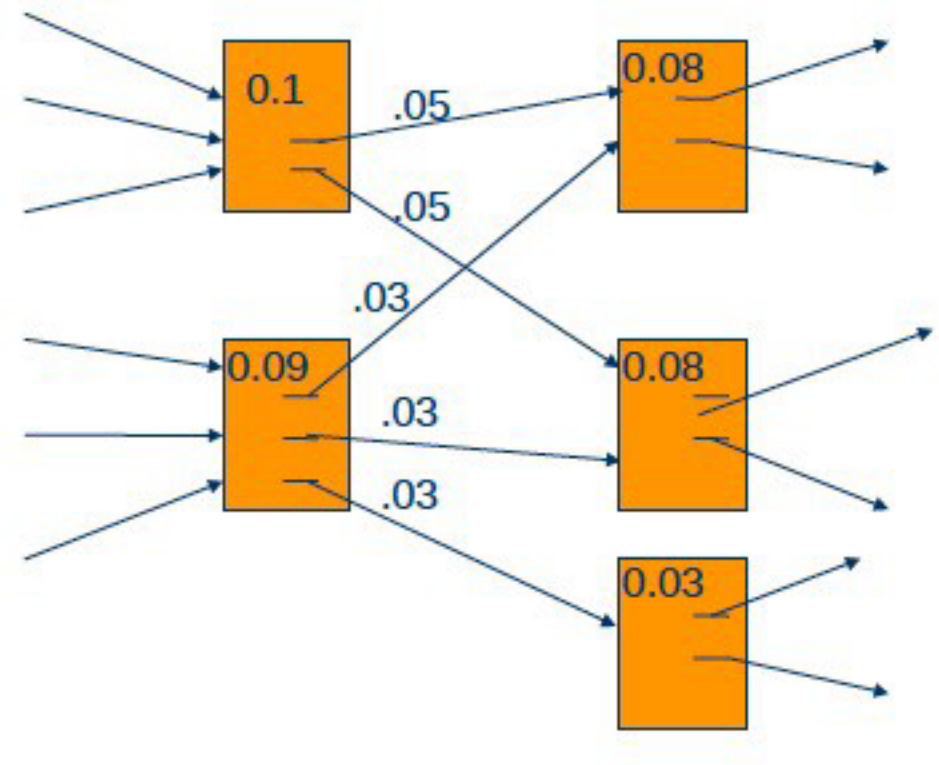
BÚSQUEDA DE INFORMACIÓN EN LA WEB

EL ALGORITMO PAGE-RANK

- Idea similar a HITS
- PR(A) = (1-d) + d(PR(t1)/C(t1)+..... (PR(tn)/C(tn))
- T1,Tn son las páginas que apuntan aA
- C(Ti) número de enlaces salientes de Ti.
- D. ‘damping factor’ es 0.85.
- Una página tiene un PageRank alto si la apuntan muchas páginas, o la apuntan menos páginas pero con un PageRank muy alto.
- Eficiente y no dependiente de búsqueda inicial como HITS.

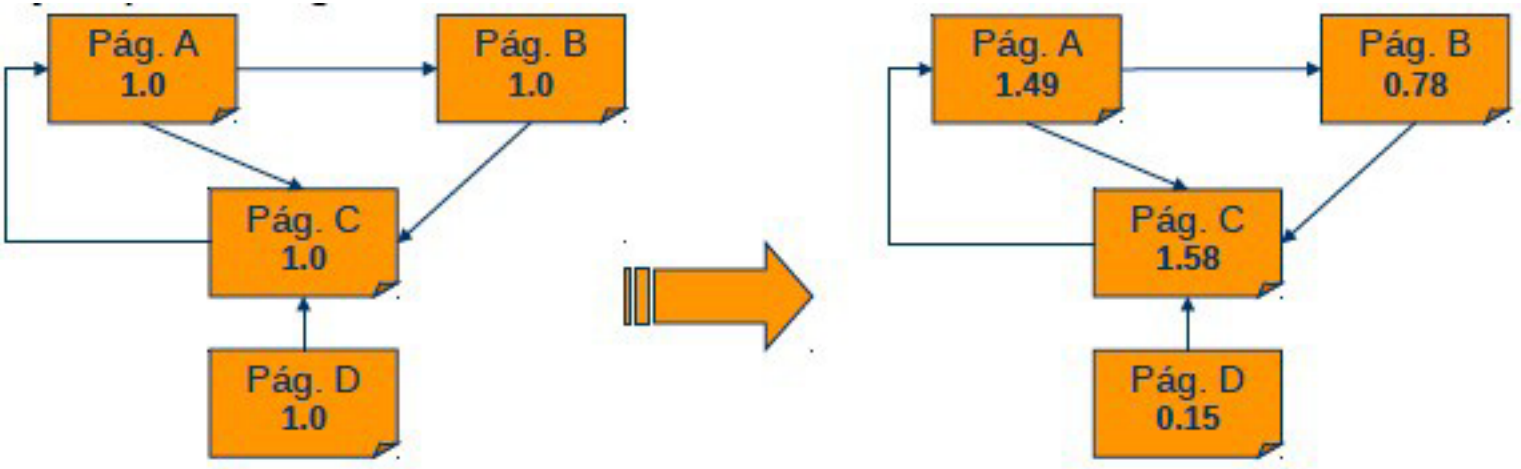
BÚSQUEDA DE INFORMACIÓN EN LA WEB

EL ALGORITMO PAGE-RANK



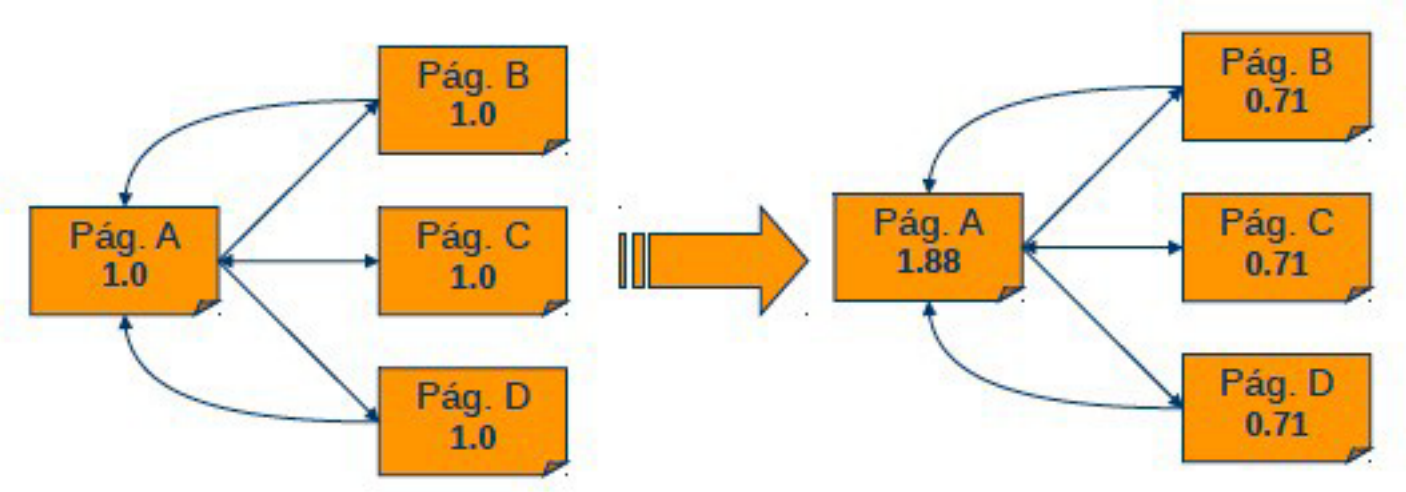
BÚSQUEDA DE INFORMACIÓN EN LA WEB

EL ALGORITMO PAGE-RANK



BÚSQUEDA DE INFORMACIÓN EN LA WEB

EL ALGORITMO PAGE-RANK

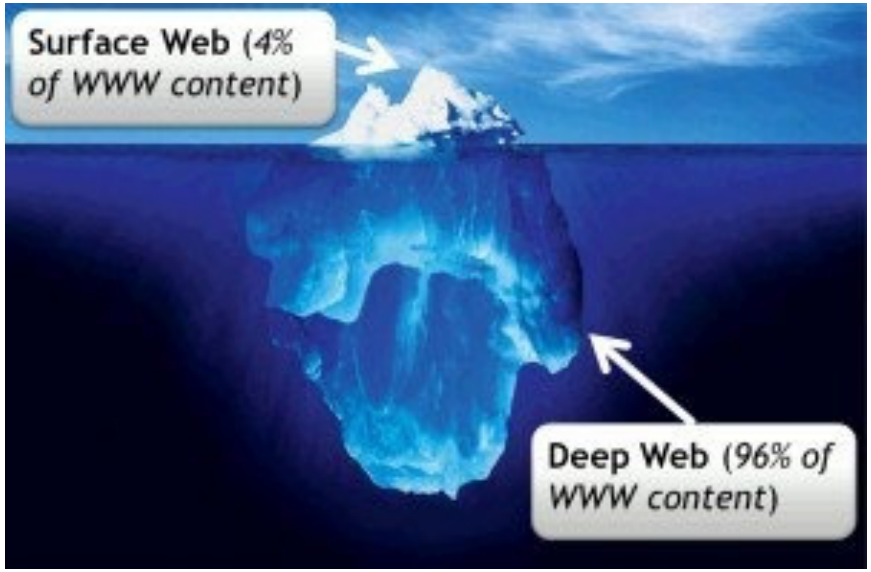


BÚSQUEDA DE INFORMACIÓN EN LA WEB

LA WEB OCULTA (HIDDEN WEB)

Definición

- **El contenido de la Web no indexado por los motores de búsqueda y por tanto difícilmente localizable a no ser que se conozca su existencia.**
- Sinónimos: Deep Web, Invisible Web



BÚSQUEDA DE INFORMACIÓN EN LA WEB

LA WEB OCULTA (HIDDEN WEB)

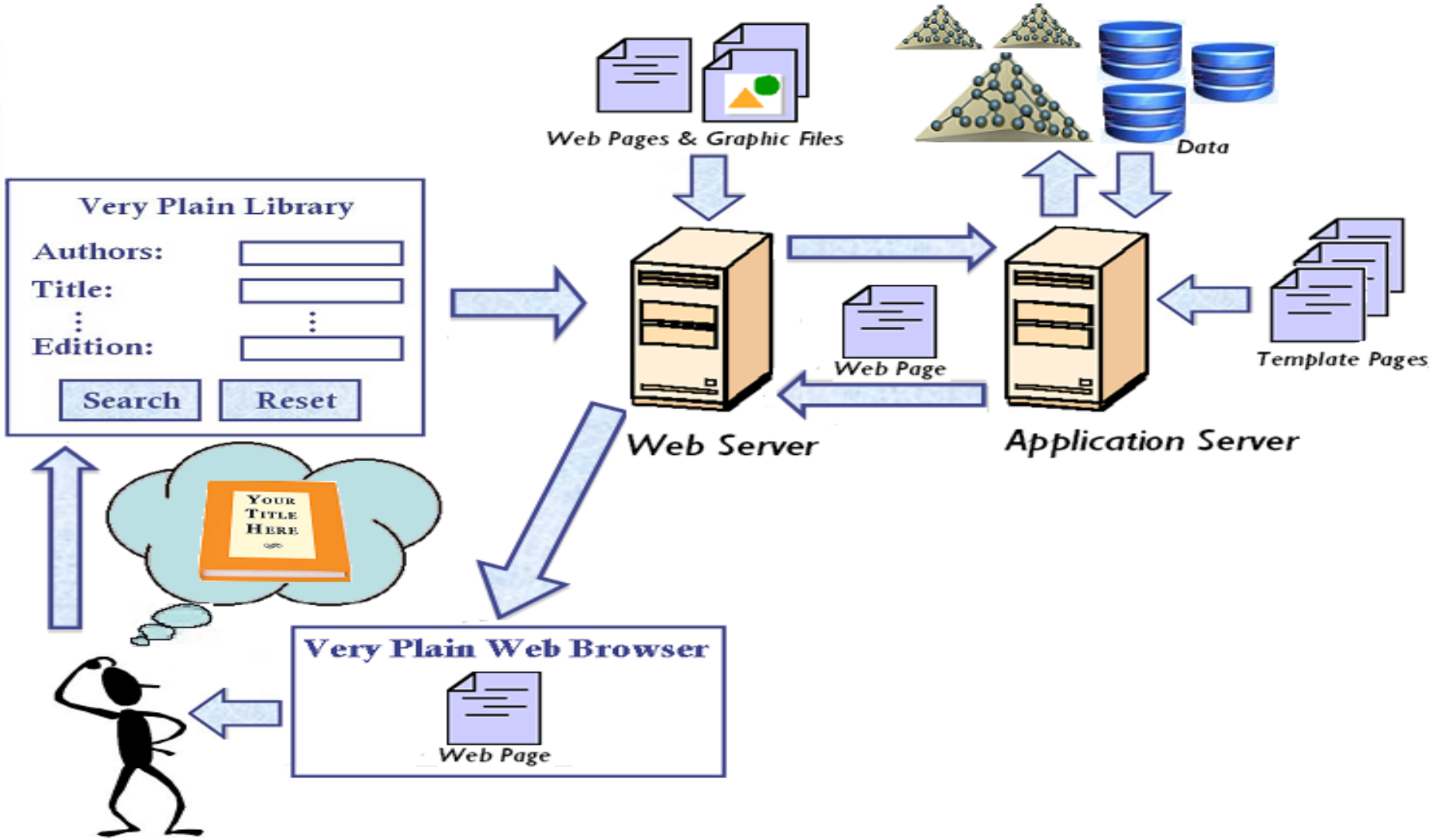
Definición

- **¿Por qué no se indexan esos contenidos?**
  - Los servidores que los alojan se encuentran aislados
    - » Solución: Proporcionarle la URL del contenido de dicho servidor a un motor de búsqueda
  - Las páginas se generan dinámicamente en función de las peticiones del usuario:
    - » Los crawlers tradicionales no pueden acceder a esta información:
    - » ¿Cómo entender formularios, aprender a consultarlos y saber consultar en ellos? ¿Problemas de acceso mediante claves?

BÚSQUEDA DE INFORMACIÓN EN LA WEB

LA WEB OCULTA (HIDDEN WEB)

Definición



BÚSQUEDA DE INFORMACIÓN EN LA WEB

LA WEB OCULTA (HIDDEN WEB)

Caracterización

- **Algunos datos**
  - Unos 300.000 sitios web
  - Sitios de la web oculta (*Hidden Web*) reciben en torno al 65% más de tráfico que los sitios de la web navegable (*Surface Web*).
  - La web oculta crece mucho más rápido que la web estática.
  - Más del 60% de los sitios web de la Web oculta son bases de datos de temas específicos que proporcionan información de alta calidad.
  - Más del 80% de los datos están accesibles públicamente.

BÚSQUEDA DE INFORMACIÓN EN LA WEB

LA WEB OCULTA (HIDDEN WEB)

Avances en el descubrimiento de la Web oculta

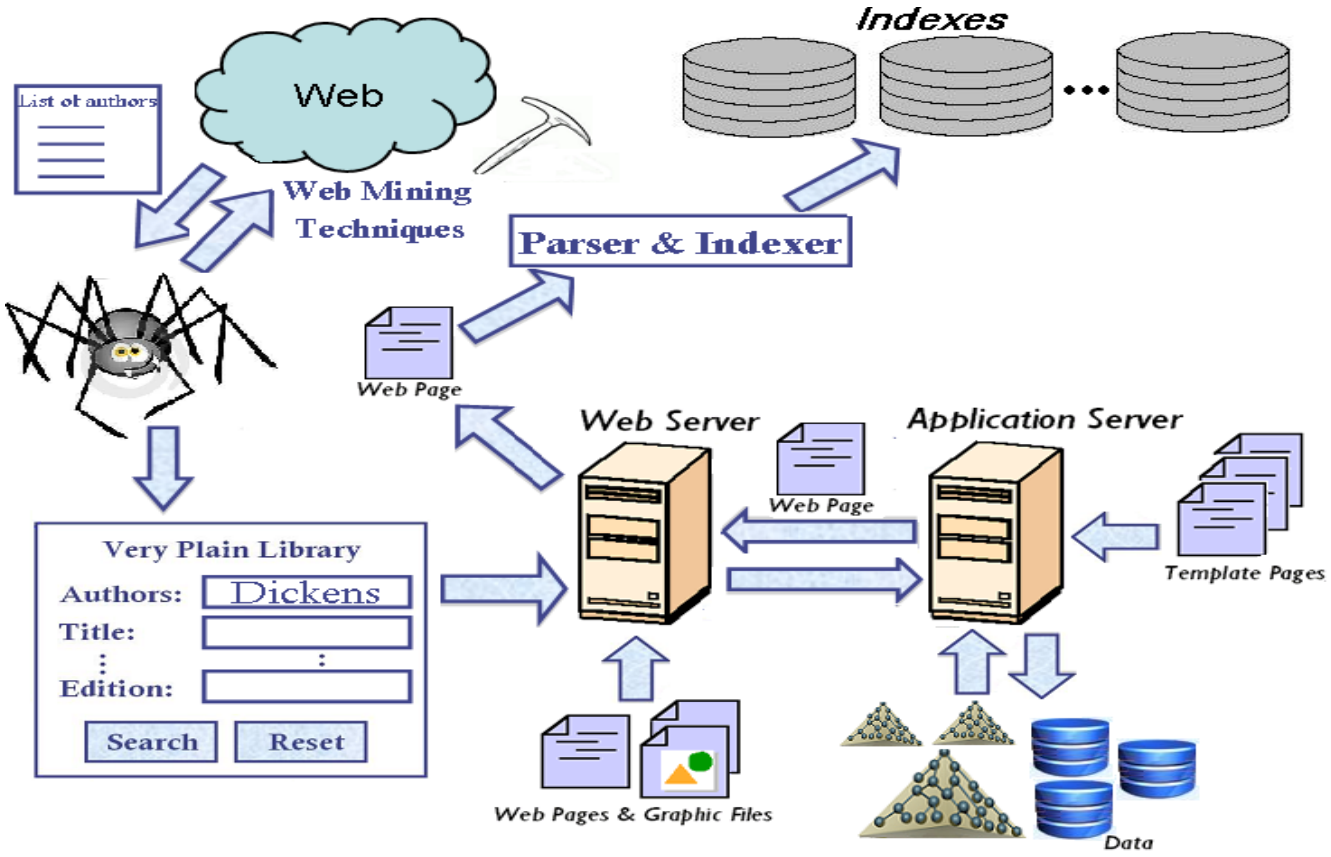
- **Se parte de una especificación de dominio** (por ejemplo: tiendas electrónicas de libros)
- **Se estudian los atributos de los formularios que aparecen en esos dominios:** tipos de datos y nombres posibles dándose ejemplos de datos reales configurando el sistema específicamente

- Identificar automáticamente nuevos formularios relevantes.
- Aprender a realizar consultas sobre dichos formularios: Generar consultas relevantes partiendo de los ejemplos del dominio y de datos obtenidos en consultas previas**.**

BÚSQUEDA DE INFORMACIÓN EN LA WEB

LA WEB OCULTA (HIDDEN WEB)

Avances en el descubrimiento de la Web oculta



BÚSQUEDA DE INFORMACIÓN EN LA WEB

LA WEB OCULTA (HIDDEN WEB)

Problemas abiertos:

- Se restringen las posibles consultas que puede resolver la fuente de datos a sólo aquellas que se pueden resolver a través del formulario

- **Ejemplo:** Select count(*)
  from fuenteDeDatos
  where publicationDate > ‘2001/01/01’ and publicationDate < ‘2010/01/01’;

BÚSQUEDA DE INFORMACIÓN EN LA WEB

REFERENCIAS Y BIBLIOGRAFÍA

- **Information Retrieval y Modern Information Retrival, Berthier Ribeiro-Netoy Ricardo Baeza-Yates**
  - 2 volúmenes. Cubre algoritmos de stemming y búsqueda de cadenas
- **Managing Gigabytes, Moffat y Zobel**
  - Cubre detalles de implementación de RI y Recuperación de imágenes
- **Information Retrieval, Gerard Salton**
  - Es un libro clásico, la última versión es de 1989
- **Information Retrieval, Jerry Kowalski**
  - Un buen resumen de las arquitecturas de los sistemas de RI


Universidad
Zaragoza
1542