Introducción

Índice

Introducción

- En la actualidad, existe una gran cantidad de información
- Cada día se generan millones de nuevos documentos
- El ser humano es incapaz de procesar toda la información y extraer conocimiento de ella
- Solución: técnicas inteligentes y automáticas para el tratamiento de la información
  - Captación
  - Indexado
  - Búsqueda
  - Ranking

Sistemas de Recuperación de la información

Introducción

Introducción

- La recuperación de la información es un área dentro de la informática (Ciencias de la Computación / *Computer Science*)
- Es un área en constante crecimiento
  - **Inicios**: indexar texto y realizar búsquedas de documentos en la colección
  - **Actualidad**: modelado, búsqueda web, clasificación de textos, arquitecturas de sistemas, interfaces de usuario, visualización de información, filtrado, idiomas (traducción, etc.)

Introducción

- Como área de investigación se divide en dos puntos de vista distintos y complementarios:
  - Informática: construcción de índices eficientes, procesamiento de las consultas de los usuarios, desarrollo de algoritmos de ranking, etc.
  - Humanidades: estudiar el comportamiento del usuario, sus necesidades, así como determinar como puede afectar al sistema de recuperación.

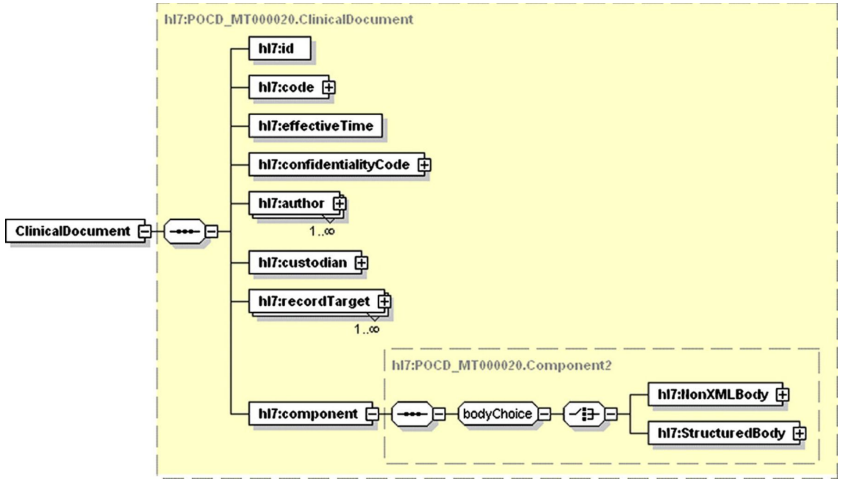
Introducción

- Su objetivo principal es facilitar a los usuarios el acceso a la información

- En términos generales se encarga de:
  - Representación
  - Almacenamiento
  - Organización
  - Acceso

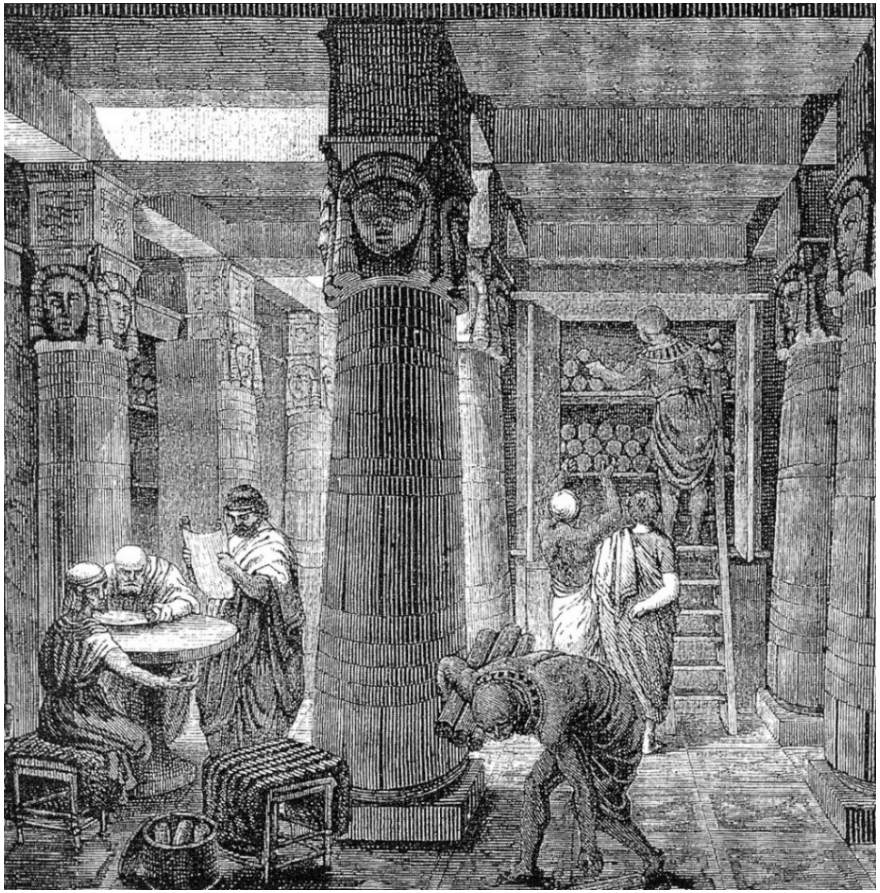
Introducción

- La información puede estar contenida en:
  - Documentos
  - Páginas web
  - Catálogos on-line
  - Registros estructurados
  - Registros semiestructurados
  - Registros no-estructurados
  - Multimedia



Historia

Historia



Historia

- La humanidad siempre ha tenido la **necesidad de organizar la información** para su posterior recuperación y acceso.
- Tradicionalmente, las bibliotecas y/o librerías se han encargado del almacenamiento.
  - La biblioteca más antigua fue creada en Elba, el actual Norte de Siria (3000 y 2500 A.C.)
  - La biblioteca de Alejandría fue la capital intelectual del mundo durante 7 siglos
  - Actualmente hay librerías en cualquier ciudad o pueblo del mundo.
- El volumen de los libros ha crecido de forma exponencial
  - Se necesitan estructuras especiales para realizar búsquedas: indexación

Historia

- Los **índices** son el **núcleo** de los sistemas de recuperación de información
  - Acceso rápido a la información
  - Procesado rápido de la consulta
- Tradicionalmente, los índices se han creado a mano: conjunto de categorías
- Cada categoría se compone de:
  - Etiquetas que identifican un tema concreto
  - Referencias a documentos relacionados con ese tema

Historia

- Los índices solían crearse por los investigadores de *Library and Information Systems*, así como por el personal de la biblioteca (Grado en Documentación -> Documentalistas)

- El surgimiento de los ordenadores y la popularización de Internet han permitido la creación de índices de gran tamaño de forma automática

  - Gran desarrollo del área de la Recuperación de la Información

Historia

- Los primeros desarrollos en Recuperación de información datan de los años 50:
  - Hans Peter Luhn
  - Eugene Garfield
  - Philip Bagley
  - Calvin Moores (acuño el término *information retrieval*)
- En 1955, Allen Kent publicó un artículo describiendo las medidas de precisión y recall
- En 1963, Joseph Becker y Robert Hayes publicaron el primer libro sobre recuperación de la información

Historia

- En los años 60, Gerard, Salton y Karen Sparck Jones, entre otros, definieron el campo de investigación, introduciendo los conceptos claves en los que se soportan las nuevas tecnologías de rankings.
- En 1968, Salton publicó su primer libro
- La primera conferencia ACM SIGIR tuvo lugar en 1978
- En 1979, C. J. Van Rijsbergen publicó "*Information Retrieval*" centrado en los modelos probabilísticos
- En 1983, Salton y McGill publicaron "*Introduction to modern information retrieval*", libro clásico en recuperación de información centrado en el modelo vectorial

Historia

- Desde entonces, la comunidad investigadora ha crecido incluyendo a miles de profesores, investigadores y profesionales.

- La *ACM International Conference on Information Retrieval* (ACM SIGIR) atrae a cientos de participantes y se someten cientos de trabajos cada año.

- TREC (*Text REtrieval Conference*)

Bibliotecas Digitales

Bibliotecas digitales

- Las bibliotecas fueron las primeras en aplicar un sistema de recuperación de la información
  - Desarrollados por instituciones académicas y posteriormente por empresas
- Primera generación:
  - Automatización de procesos existentes restringidos a nombres de autores y títulos
- Segunda generación:
  - Mejora de la capacidad de búsqueda: categorías, palabras clave y operadores (AND, OR, NOT)
- Tercera generación (actualidad):
  - Mejora de las interfaces gráficas, formularios electrónicos, www, sistemas abiertos, etc.

Bibliotecas digitales

- Koha
  - https://koha-community.org

El problema de Recuperación de la Información

Definición del problema

- Los usuarios de los sistemas de recuperación de información tienen necesidades de distintas complejidades:
  - Página de una empresa, gobierno o institución
  - Búsqueda de información
  - Búsqueda de una solución a un problema

"Busca todos los artículos científicos publicados por Manuel Jesús Cobo Martín"

Definición del problema

- La descripción completa de la necesidad del usuario no es la mejor entrada para el sistema
- El usuario tiene que traducir su necesidad de información en una consulta (o secuencia de consultas) que será lanzada contra el sistema
  - Conjunto de palabras clave (términos de indexación)
- El sistema de recuperación debe de recuperar **información** que sea **relevante** **y** **útil** para el usuario

información != datos

Definición del problema

"El objetivo principal de un sistema de recuperación de información es recuperar todos los documentos que sean relevantes para la consulta dada por el usuario, minimizando a sus vez el número de documentos no-relevantes recuperados"

Definición del problema

- La dificultad no está en obtener información de los documentos, sino en saber qué es relevante y qué no lo es.
- El concepto de relevancia es central en los sistemas de recuperación de información
- La relevancia es un criterio personal y depende del contexto:
  - Puede cambiar con el tiempo (aparece nueva información)
  - Localización (la mejor respuesta es el lugar más cercano)
  - Dispositivo (la mejor respuesta es la que se vea mejor en un móvil)
- Ningún sistema puede dar respuestas perfectas para todos los usuarios al mismo tiempo

Información vs Datos

Información vs Datos

- La **recuperación de datos** consiste en determinar los documentos que contienen una serie de palabras clave de la consulta del usuario
  - No suele satisfacer las necesidades del usuario
- El usuario quiere recuperar información sobre una materia
- El usuario no quiere recuperar los documentos que casan con una determinada consulta
- Los usuarios aceptarían documentos que contuvieran sinónimos de las palabras clave introducidas
- En un sistema de **recuperación de información** los objetos recuperados pueden ser erróneos según la consulta
  - Pequeños errores se obvian
- En un sistema de recuperación de datos, un objeto recuperado erróneamente implica un fallo en el sistema

Información vs Datos

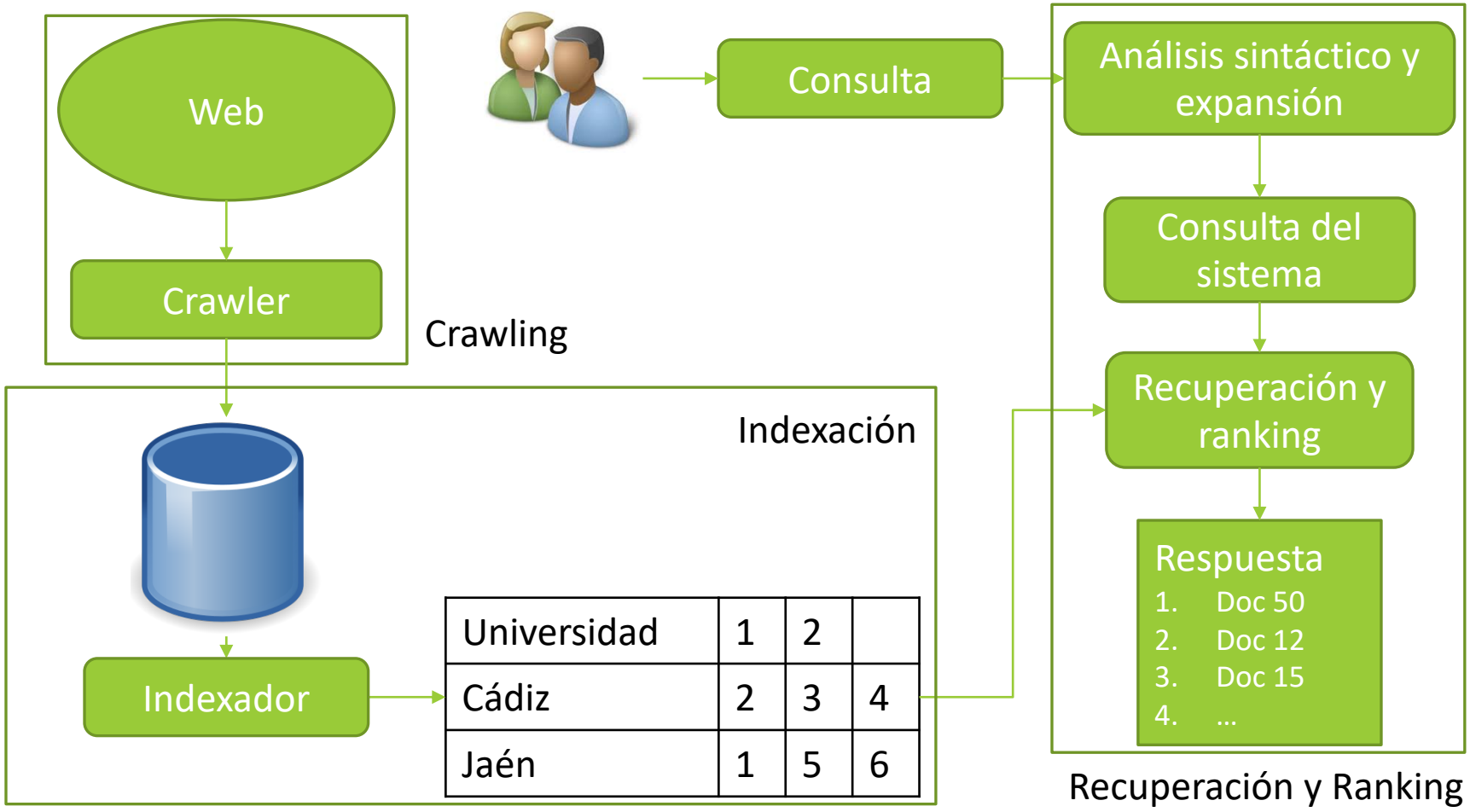
- Sistema de recuperación de datos (base de datos):
  - Datos
  - Estructura bien definida

- Sistema de recuperación de información:
  - Texto en lenguaje natural
  - Estructura no definida

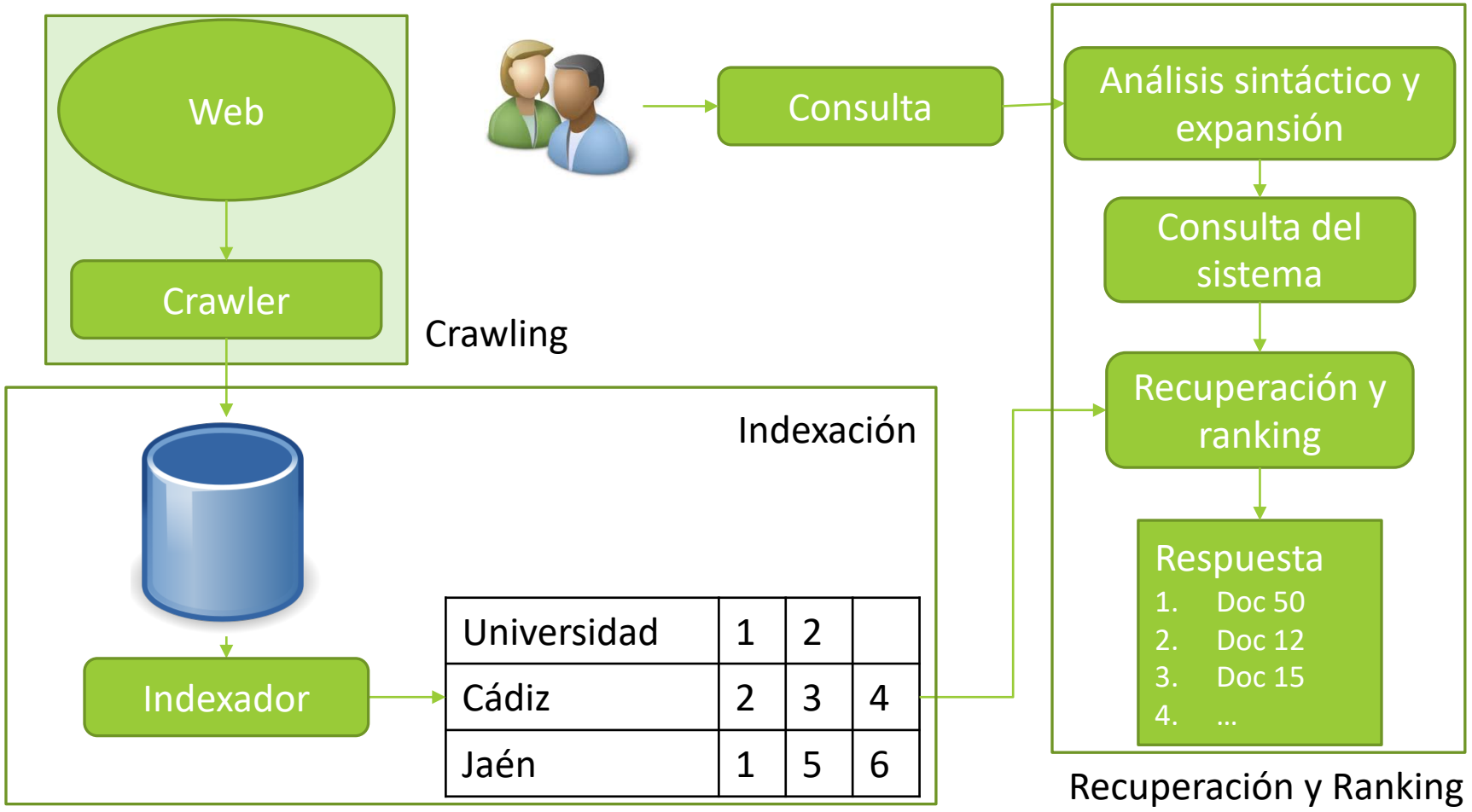
Sistema de Recuperación de la Información

MODELO GENÉRICO

Sistema RI

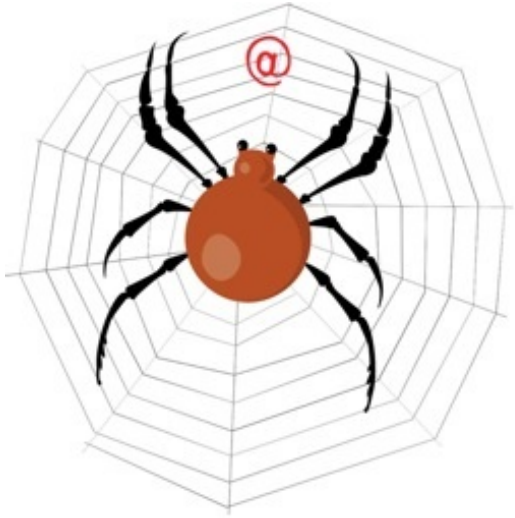

Web
Crawler
Crawling
Consulta
Análisis sintáctico y expansión
Consulta del sistema
Recuperación y ranking
Respuesta
1. Doc 50
2. Doc 12
3. Doc 15
4. …
Recuperación y Ranking
Indexación
Indexador
Universidad 1 2
Cádiz 2 3 4
Jaén 1 5 6

Sistema RI: crawling

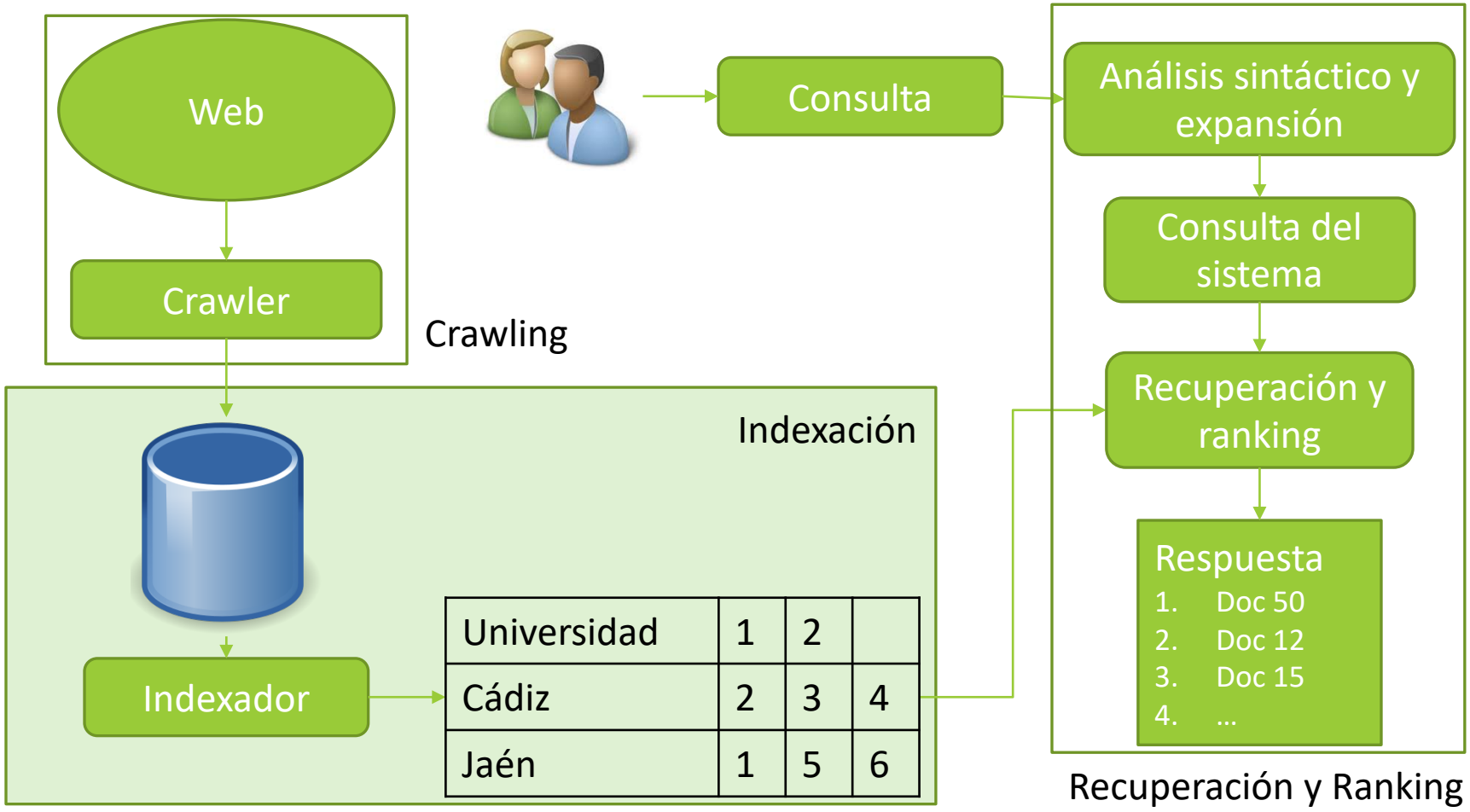


Sistema RI: crawling

- El sistema de recuperación de la información necesita un conjunto de documentos para comenzar a trabajar
  - Colección de documentos privada
  - Colección de documentos recuperada de la web mediante un robot (crawler)
- La colección de documentos se almacena en disco:
  - Repositorio central



Sistema RI: indexación



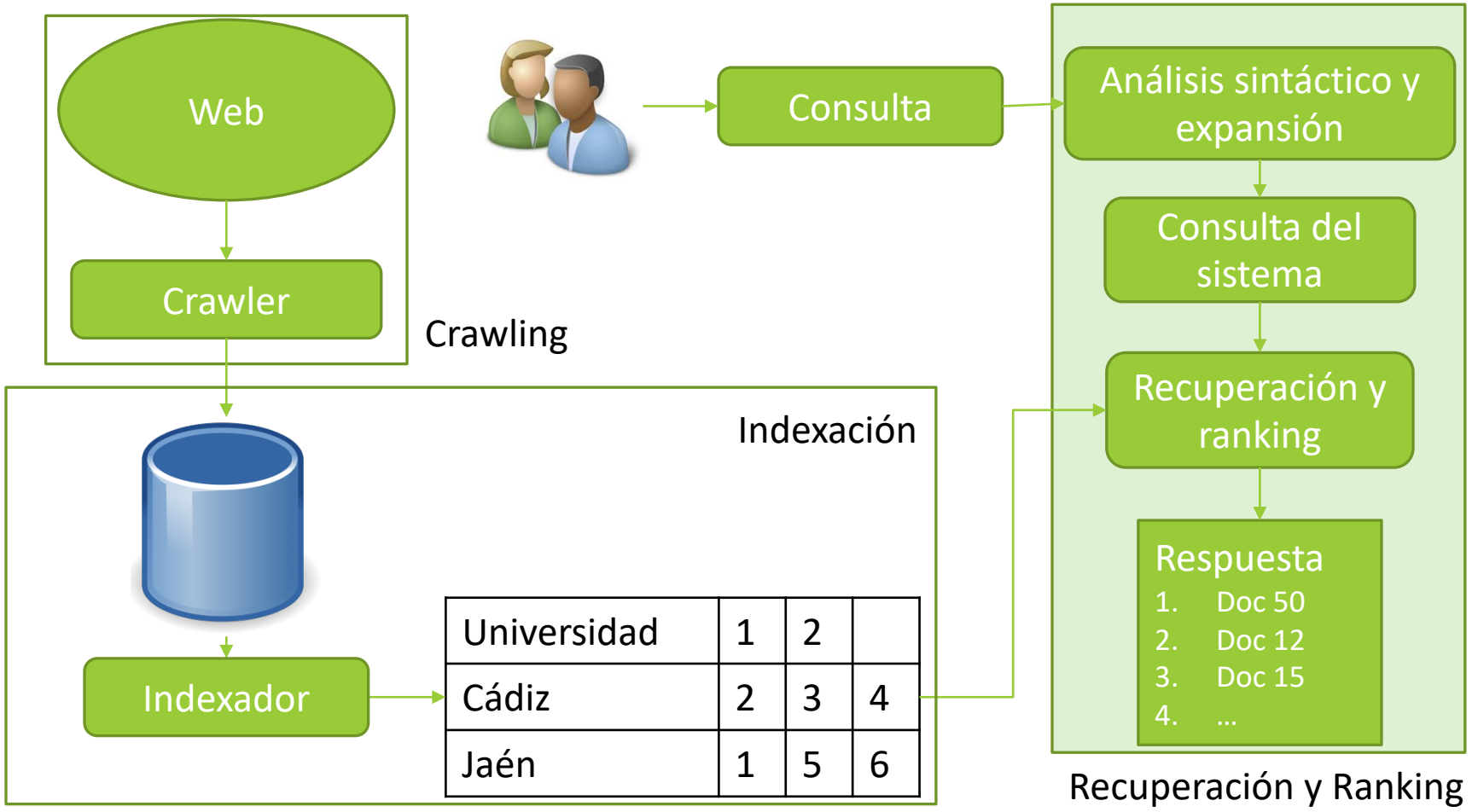
Sistema de RI: indexación

- La colección principal tiene que pre-procesarse:
  - Eliminación de palabras vacías (stopwords)
  - Lematización (stemming)
  - Selección de los términos de indexación
    - Utilizados como representación de un documento
    - Deben ser más pequeños que los documentos en si.
- La colección de documentos del repositorio central se tiene que indexar para permitir:
  - Búsquedas rápidas
  - Ordenación y ranking

Sistema de RI: indexación

- La estructura más utilizada es el Índice Invertido:
  - Lista de todas las palabras distintas de la colección
  - Para cada palabra, lista de los documentos en los que aparece
- La creación del índice invertido tiene que realizarse mediante un proceso offiline:
  - Proceso costoso
  - Se tiene que realizar al principio, antes de que el sistema pueda recibir alguna consulta
- Los recursos de tiempo y almacenamiento invertidos se amortizarán lanzando consultas sobre el sistema muchas veces

Sistema RI: recuperación y ranking



Sistema RI: recuperación y ranking

- En primer lugar, el usuario tiene que especificar una consulta que refleje sus necesidades de información
- El sistema aplica un procesamiento similar al realizado sobre los documentos:
  - Eliminación de palabras vacías
  - Corrección de errores
  - Lematización
- El sistema puede expandir la consulta:
  - Sugerencias hechas por el sistema y confirmadas por el usuario
- Proceso de recuperación: la consulta del sistema se lanza contra el motor de recuperación para obtener el conjunto de documentos que contienen las palabras de la consulta

Sistema RI: recuperación y ranking

- El sistema puede recuperar miles de documentos.
- El usuario sólo estará interesado en los más relevantes.
- El sistema tiene que ordenar los documentos en función de la probabilidad de ser relevantes para el usuario.
- El proceso de ordenación es sin duda el más crítico
  - El usuario percibirá la calidad del sistema a través de los documentos devueltos y su orden.
- Finalmente, los documentos que el sistema ha detectado como más relevantes son mostrados al usuario:
  - Título
  - Resumen
  - Etc.

La RI en la era de la Web

La RI en la era de la Web

- Desde su creación, la web ha tenido un gran éxito y ha crecido de forma exponencial
  - El número de páginas web exceden los 20,000 millones
  - El número de usuarios exceden los 1,700 millones
  - Existen más de 1 trillón de URLs distintas
- ¿Existe alguna característica que haya sido decisiva para el éxito de la web?
  - ¿La simpleza del HTML?
  - ¿El bajo coste de acceso?
  - ¿La amplia cobertura de Internet en la actualidad?
  - ¿Los navegadores web?
  - ¿Los buscadores?

La RI en la era de la Web

- Sin lugar a dudas, las características anteriores han influido notablemente en el éxito de Internet, pero no son la causa.
- La verdadera causa: ***libertad para publicar***
- Las personas pueden publicar ahora sus ideas en la web y llegar a un público objetivo
  - No hay que pagar nada
  - No hay que convencer a ningún editor
- En definitiva,
  - las limitaciones impuestas por los medios de comunicación (*mass-media*) han desaparecido
  - las barreras geográficas se han ido

La RI en la era de la Web

- El surgimiento de la web ha hecho que la Recuperación de la Información haya crecido
- La búsqueda web es una de las principales tareas de la Recuperación de la Información
- Varios factores han afectado a la Recuperación de la Información:
  - La estructura de la web (documentos conectados por enlaces) ha hecho necesarios los crawlers para acceder y descargar la información de la web
  - El volumen de información. Los motores de búsqueda tienen que hacer frente a:
    - Millones de documentos
    - Millones de consultas diarias

La RI en la era de la Web

- Medir la relevancia de millones de documentos no es fácil
  - Cualquier consulta recuperará miles de documentos que contienen las palabras clave
  - Muchos de ellos son ruido
  - Afortunadamente, la web a traído nuevas formas de medir la relevancia (clics)
- Los usuarios ya no sólo buscan información textual:
  - Precios de un libro
  - Horario de vuelos
  - Teléfono de un hotel
- Nuevo modelo de negocio: anuncios on-line
  - Nuevos problemas: web spam

Conclusiones

Conclusiones

- Para procesar la gran cantidad de información que existe en la actualidad, son necesarias técnicas inteligentes
- La recuperación de la información (RI) es una técnica informática centrada en el desarrollo de modelos eficientes y eficaces de búsqueda de información dentro de una colección de documentos
- Tradicionalmente ha sido un área asociada a las bibliotecas, pero en la actualidad se aplica principalmente en la web
- La RI trata de dar respuesta a una necesidad de información particular
- Intenta recuperar todos los documentos relevantes para una consulta dada, minimizando el número de no relevantes

Conclusiones

- Un sistema RI se compone de tres grandes módulos:
  1. Crawler
  2. Indexación
  3. Recuperación y ranking
- Los sistemas RI en la web han supuesto un reto enorme
  - Gran parte del éxito de la web puede asociarse a los motores de búsqueda