UNIVERSIDAD COMPLUTENSE MADRID

Procesos de recuperación de información

Sistemas de Gestión de Datos y de la Información
Enrique Martín - emartinm@ucm.es
Máster en Ingeniería Informática
Fac. Informática

Procesos principales

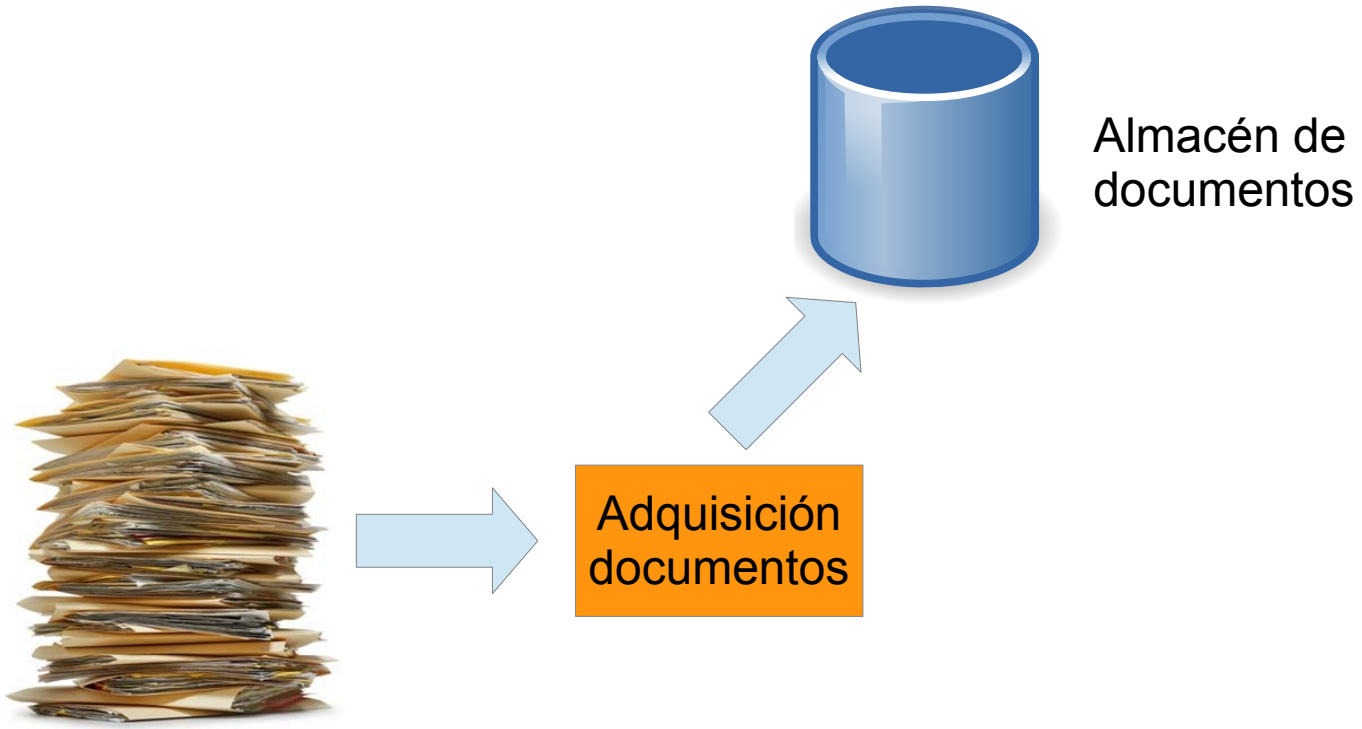
- En un sistema de recuperación de información, los procesos principales son dos:
  - **Indexado**: tratamiento de la colección de documentos hasta generar el índice sobre el que se realizarán las consultas.
  - **Consulta**: tratamiento de la consulta del usuario y encaje con los documentos, utilizando el índice.

Indexado

Indexado

- Es la fase más que más recursos consumirá, ya que tiene que procesar todos los documentos, analizándolos completamente para extraer los términos clave.
- Afortunadamente, será una fase que se realizará una sola vez por lo que su coste se amortizará con respecto al volumen de consultas contestadas.

Indexado



Indexado: adquisición

- El primer paso del indexado es es **obtener los documentos** sobre los que vamos a realizar las búsquedas.
- Según el caso, puede ser sencillo o más complicado.

Indexado: adquisición

- Adquisición sencilla sería:
  - Catálogo de una biblioteca
  - E-mails en un cliente de correo
- Ejemplos más complicados:
  - Páginas web en Internet
  - Documentos en el ordenador
  - Ficheros en los ordenadores conectados a la red local de una empresa.

Indexado: adquisición

- En muchos casos la adquisición la realiza un ***crawler***.
- Este sistema recorre los documentos y las distintas localizaciones buscando nuevos documentos a incluir.
- Puede buscar únicamente documentos actualizados después de una fecha, sobre una temática concreta, en un determinado dominio, etc.

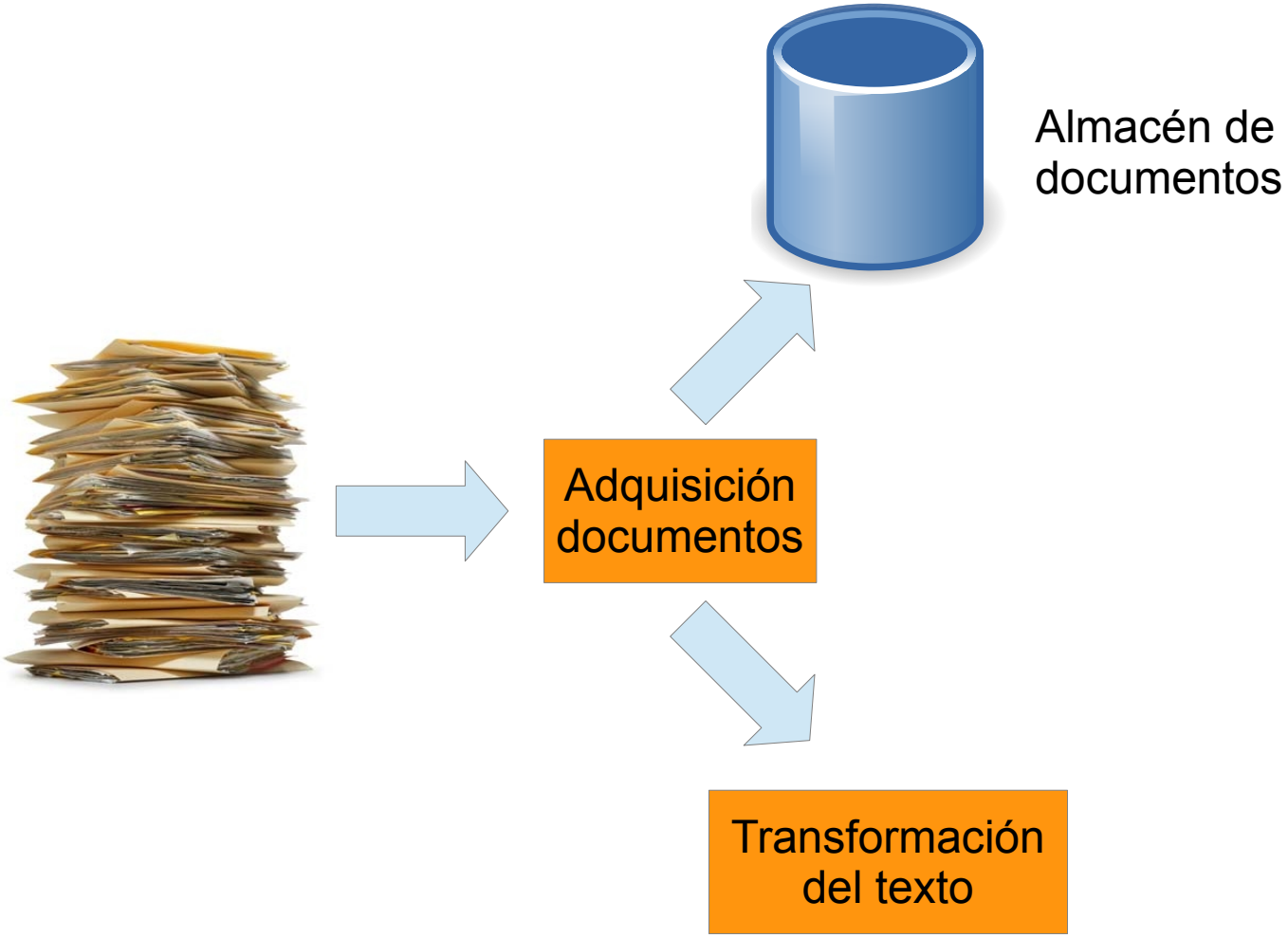
Indexado: adquisición

- Una vez hemos encontrado los documentos a incluir, el siguiente paso es la **conversión**.
- Los distintos documentos pueden tener formatos variopintos: HTML, XML, PDF, DOC, PPTX, etc.
- Es muy recomendable tenerlos representados en un formato de texto común.
- En algunos casos, el fichero puede necesitar ser interpretado (p.ej. PDF con imágenes)

Indexado: adquisición

- Cuando ya tenemos los ficheros localizados y convertidos a un formato textual uniforme, el siguiente paso es almacenarlos en un almacén de documentos.

Indexado


Almacén de documentos
Adquisición documentos
Transformación del texto

Indexado: transformación

- La fase posterior a la adquisición es la **transformación del texto**.
- Tenemos los documentos necesarios almacenados en un formato textual, pero ahora tenemos que procesarlos para conocer su contenido.

Indexado: transformación

- **Análisis léxico (*tokenizing*)**
  Se recorre el documento separando las unidades mínimas (tokens) que se van a considerar. Normalmente se tratará de palabras.
- Hay que tener cuidado con situaciones especiales como: guiones (*text-based*), comillas (“Strawberry fields”), apóstrofes (O'Briens), etc.

Indexado: transformación

- **Eliminación de palabras vacías *(stop words).***
  Un **paso opcional** es eliminar aquellas palabras que no contribuyen mucho al contenido del texto.
- Estas palabras suelen ser:
  - **Artículos**: the, a, an.
  - **Conectivas**: thus, therefore, and, or.
  - **Preposiciones**: of, in, to.

Indexado: transformación

- La eliminación de palabras vacías se suele llevar a cabo utilizando una lista de *stop words*.
- La eliminación de estas palabras suele tener poco impacto en la calidad de los documentos obtenidos. Pensad que **muchas de ellas aparecerán en todos los documentos**.
  - Sin embargo son **imprescindibles** para búsquedas de **frase**.

Indexado: transformación

- **Obtención de raíces *(stemming)***.
- Transformación a nivel de palabra que trata de agrupar palabras de significado muy similar basándose en su raíz (*stem*).
- **Ejemplo**: *comer*, comí y comimos son palabras diferentes pero tienen la misma raíz: **com**.

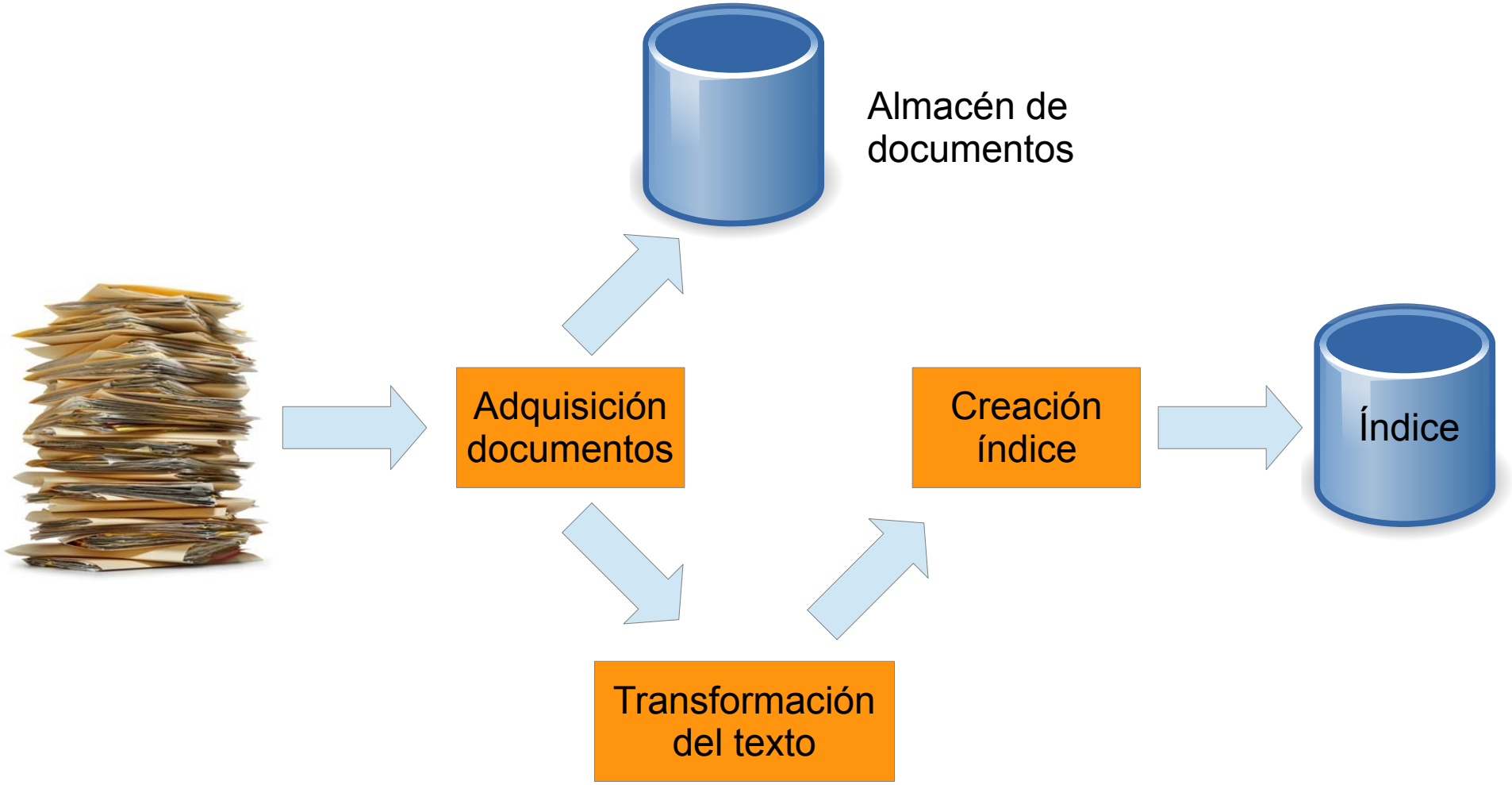
Indexado: transformación

- **Obtención de lemas *(lematización)***.
- Transformación a nivel de palabra que trata de sustituir cada palabra **por otra palabra** general que las representa a todas.
- **Ejemplo**: *comer*, comí y comimos son palabras diferentes representadas por el mismo lema ***comer***.

Indexado: transformación

- Dependiendo del sistema de recuperación de información concreto, podemos tener más fases. Por ejemplo en un buscador Web podríamos tener:
  - Extracción y análisis de enlaces.
  - Categorización por temas.

Indexado

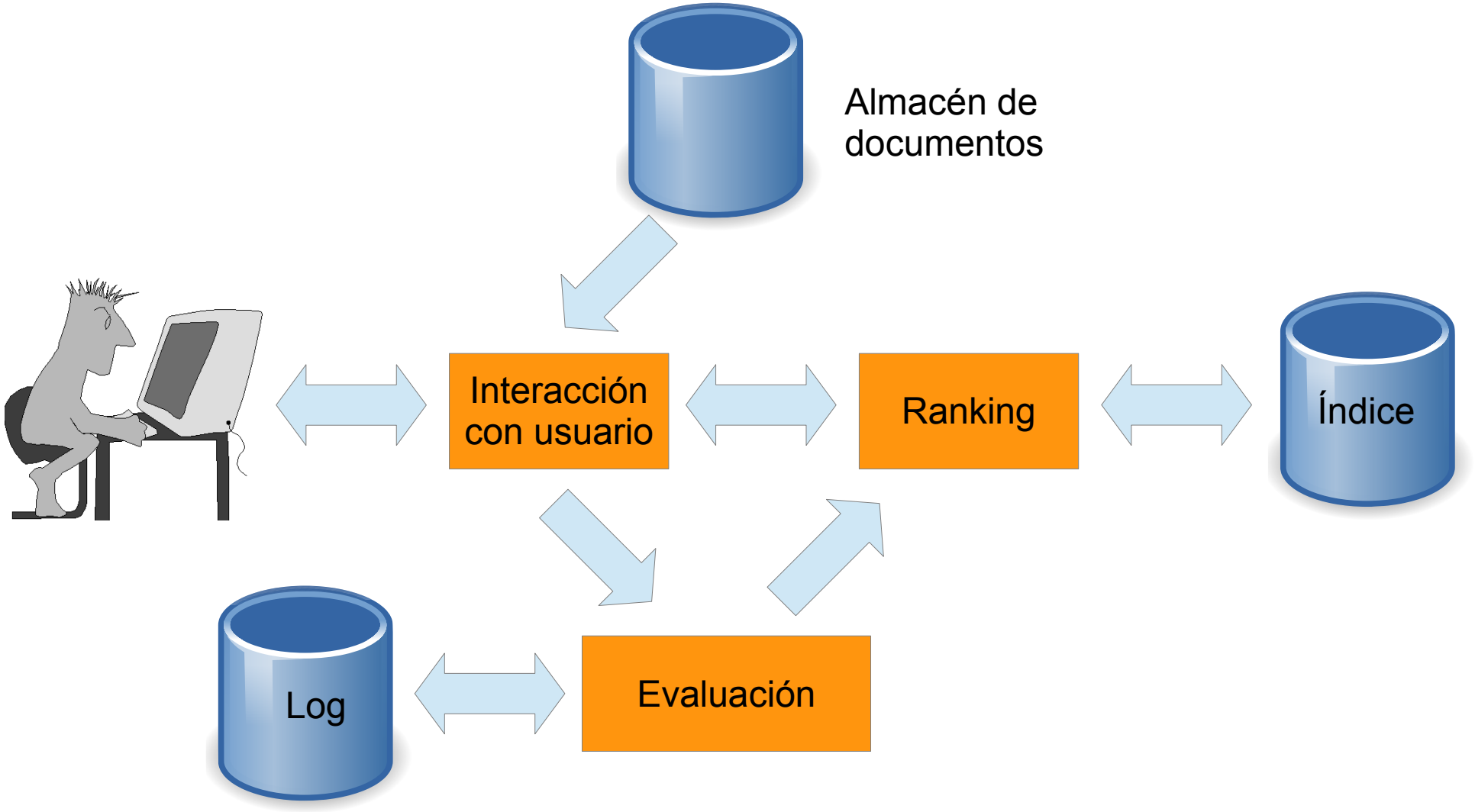

Almacén de documentos
Adquisición documentos
Creación índice
Índice
Transformación del texto

Indexado: creación

- Una vez tenemos los documentos adquiridos y transformados, el siguiente paso es la **creación del índice invertido**.
- Depende de la estructura de datos utilizada.
- En esta etapa obtendremos posiciones, repeticiones, pesos, etc. en consonancia con el modelo de recuperación que vayamos a utilizar.

Consulta

Consulta


Almacén de documentos
Interacción con usuario
Ranking
Índice
Log
Evaluación

Consulta

- El proceso de consulta es un proceso complejo en el que intervienen varios actores:
  - El **usuario**: realiza las consultas, recibe los resultados y los navega.
  - El **almacén de documentos**: útil para obtener fragmentos con los que acompañar los resultados (*snippets*).
  - El ***log* de consultas previas**: almacena consultas previas y las respuestas del usuario.
  - El **índice**

Consulta: interacción

- La componente de **interacción con el usuario** se encarga de:
  - Recibir la consulta en un lenguaje de consultas establecido.
  - Transformar la consulta (*tokenizing*, análisis, *stopping*, *stemming*) y ofrecer sugerencias de consultas relacionadas o expandir la consulta con términos similares.
  - Mostrar los resultados y monitonizar el comportamiento del usuario.

Consulta: *ranking*

- La componente de **ranking** se encarga de encontrar aquellos documentos más relevantes para una consulta y además obtiene un orden.
- Para ello hace uso de un **modelo de recuperación de información**.
- También puede utilizar información del *log*, como documentos populares por los usuarios.

Consulta: evaluación

- La componente de **evaluación** realiza las siguiente tareas:
  - Almacena las consultas y las interacciones del usuario en el *log*.
  - Analiza la correspondencia entre la relevancia del modelo de recuperación y la relevancia real para los usuarios.
- Esta información es importante para mantener el sistema de recuperación en buen estado, recomendando mejoras de modelo.

Bibliografía

Bibliografía

- **Modern Information Retrieval, the concepts and technology behind search**, second edition. *Ricardo Baeza-Yates y Berthier Ribeiro-Neto*. Pearson Education Limited (2011).
- **Introduction to Information Retrieval**. *Christopher D. Manning, Prabhakar Raghavan y Hinrich Schütze*. Cambridge University Press. 2008. `http://www-nlp.stanford.edu/IR-book/`

Bibliografía

- **Search Engines: Information Retrieval in Practice** (International Edition). *W. Bruce Croft, Donald Metzler y Trevor Strohman*. Person Education. 2010.