Fundamentos de la Minería de Datos Web

Máster Online en Ciencia de Datos

UCO ONLINE

Dr. José Raúl Romero

Profesor Titular de la Universidad de Córdoba y Doctor en Ingeniería Informática por la Universidad de Málaga. Sus líneas actuales de trabajo se centran en la democratización de la ciencia de datos (*Automated ML* y *Explainable Artificial Intelligence*), aprendizaje automático evolutivo y analítica de software (aplicación de aprendizaje y optimización a la mejora del proceso de desarrollo de software).

Miembro del Consejo de Administración de la *European Association for Data Science*, e investigador senior del Instituto de Investigación Andaluz de *Data Science and Computational Intelligence*.

Director del **Máster Online en Ciencia de Datos** de la Universidad de Córdoba.

Fundamentos de la Recuperación de la Información

IR – *Information Retrieval*

UCO ONLINE

IR - *Information Retrieval*

- Es el campo de estudio que ayuda al usuario a **encontrar información** de entre una gran colección de documentos de texto
  - La IR tradicional considera al **documento** como unidad básica de información
  - En IR sobre Web, los documentos son **páginas Web**
- "**Encontrar información**" significa encontrar el conjunto de documentos que es relevante para la consulta (***query***) del usuario
  - Normalmente los documentos se devuelven en forma de *ranking*
  - La forma más habitual de *query* es un listado de palabras clave (***keywords***), llamadas términos (***terms***)
- La **búsqueda Web no es estrictamente IR**: además de los modelos IR tradicionales, utiliza algoritmos propios y también aporta técnicas a IR (bidireccionalidad)

IR - *Information Retrieval*

- **Recuperar datos** (**RD**) Vs. **Recuperar información** (**RI**)
  - Los datos se pueden estructurar en tablas, árboles, …
  - Recuperar exactamente lo que el usuario quiere
- El texto no tiene estructura clara y no es fácil crearla

```
Quiero información sobre las consecuencias de la burbuja inmobiliaria en la crisis económica española.
```

```
SELECT nombre, puesto, salario
FROM Employees
WHERE empresa="La Caixa" and salario>=3000
```

IR - *Information Retrieval*

- **Recuperar datos** (**RD**) Vs. **Recuperar información** (**RI**)
  - En **RD** se sabe exactamente lo que se quiere
  - En **RI** no existe ***la*** respuesta correcta
  - En **RD** importa la eficiencia (velocidad y espacio)
  - En **RI** importa la ***calidad*** de la respuesta

La **RI** busca una ***aproximación*** a responder lo que el usuario busca

Objetivos de IR

El objetivo es encontrar información relevante, útil y significativa

SUBOBJETIVO 1: **Recall**. Recuperar todos los documentos relevantes

SUBOBJETIVO 2: **Precisión**. Recuperar la mayoría de los documentos relevantes

SUBOBJETIVO 3:

- Recuperar tan pocos documentos no-relevantes como sea posible
- Recuperar **documentos relevantes** y posicionarlos antes que los no-relevantes

Información – documentos – conocimiento

Recuperación de la información

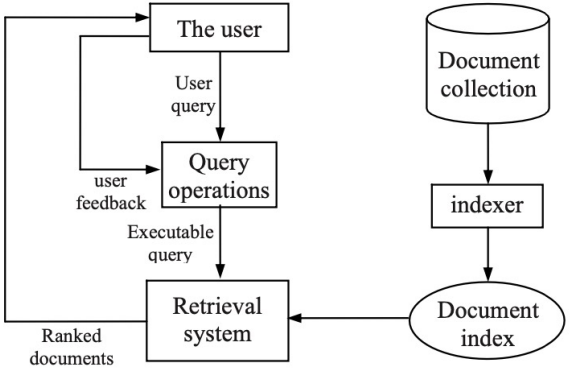
Gestión de documentos

Ingeniería del conocimiento

	Documento	Información	Conocimiento
IR	Indexado	Ranking	Razonamiento
Gestión	Escaneado	Filtrado	Aprendizaje
Ingeniería	Estructuración	Modelado	Anotación

Sistema IR

- El *usuario* con necesidad de información emite una consulta (*user query*) al *sistema de recuperación* a través del módulo de *operaciones de consulta*
- La *colección de documentos* también se denomina base de datos de texto, que es indexada por el *indexador* para una recuperación eficaz
- El *sistema de recuperación* utiliza el *índice de documentos* para recuperar aquellos documentos que contienen algunos términos de la consulta (probable que sean relevantes para la consulta), calcular sus puntuaciones de relevancia y clasificarlos de acuerdo con las puntuaciones (*ranking*)
- Los documentos clasificados y ordenados se presentan al usuario



Arquitectura general de un sistema IR
(B. Liu, "Web Data Mining", 2011)

Procesos Básicos del Sistema de IR

- **Representación de los documentos** (*indexing*)
  - Proceso off-line sin intervención del usuario
  - Dos modos de indexación:
    - **Texto completo –** p.ej. algoritmo que identifica palabras en inglés, las pone en minúscula
    - **Texto parcial –** p.ej. extrae título y resumen del documento, además de su localización
- **Representación de la necesidad de información** (*query formulation*)
  - Denota el diálogo o interacción completa entre sistema y usuario
  - Pretende la búsqueda de la consulta adecuada y una mayor comprensión de la necesidad de información del usuario
- **Comparación entre la representación del documento y de la consulta** (*matching*)
  - Devuelve ranking de documentos (**¿relevantes en la cima?**)
  - Los algoritmos más sencillos utilizan la frecuencia de distribución de términos en documentos, u otras estadísticas (p.ej. número de hiperenlaces que apuntan al documento)
  - Un ranking acertado reduce al usuario significativamente el tiempo necesario para leer los documentos

Formulación de la consulta de usuario (1/3)

Consulta de palabras clave (*keyword query*):

- El usuario proporciona uno o más **términos** para encontrar documentos que contengan alguno de ellos
- En algunos sistemas IR, el orden de los términos es relevante para el resultado

Consulta booleana:

- El usuario puede utilizar operadores AND, OR, NOT para construir *queries* complejas
- Un documento es devuelto si la *query* es lógicamente verdad para ese documento (***exact match***)

Consulta de frase:

- La consulta es una secuencia de palabras que forman una frase
- El documento devuelto debe contener una instancia de la frase (la frase exacta entre comillas “”)

Formulación de la consulta de usuario (2/3)

Consulta de proximidad:

- Versión relajada de la consulta de frase, que puede combinarse con términos y frases
- Estas consultas buscan términos dentro de una proximidad cercana entre ellos (la **cercanía** es un factor en el cálculo del ranking de documentos)
  - Un documento que contiene todos los términos cercanos entre sí se "rankea" más alto
  - Algunos sistemas de IR permiten establecer umbrales de proximidad
  - La mayoría de sistemas de búsqueda consideran tanto la proximidad de términos como su orden en su recuperación

Consulta de documento completo:

- El usuario quiere encontrar documentos que sean similares al documento de consulta
- Algunos buscadores (p.ej. Google) permiten esta consulta cuando se pasa la URL de una página Web como *query* (a veces se puede ver el resultado como "*more like this*" / "*similar pages*")

Formulación de la consulta de usuario (3/3)

Pregunta en lenguaje natural:

- El tipo más complejo de búsqueda, pero el ideal para el usuario
- Es un **área de investigación activa**: question answering
  - Algunos sistemas ya funcionan bien para preguntan acotadas: definiciones de términos técnicos o patrones lingüísticos muy estrictamente definidos (“refers to”, “defined as”, etc.)
  - Algunas de las operaciones se hacen *off-line* (p.ej. extracción de definiciones)

Habitualmente, las consultas requieren algún tipo de preprocesado

- Forma más sencilla – identificación de ***stop-words*** (términos muy frecuentes pero poco significativos semánticamente, como preposiciones, artículos, etc.)
- Forma más compleja – transformación de lenguaje natural a consultas procesables que podría requerir aceptar **feedback relevante** del usuario para refinar o extender la query original

IR tradicional Vs. IR sobre páginas Web

1

En **IR** → documentos de texto convencional

En Web → **documentos con hiperenlaces y anclas** (= texto asociado a los hiperenlaces y que resulta una descripción más precisa sobre la página destino que el hiperenlace en sí)

- En IR: no existen los hiperenlaces (¿citas?)
- En Web: Los hiperenlaces determinan el resultado de las búsquedas y condicionan los rankings resultado de los algoritmos

2

Las páginas Web son **semi-estructuradas** → texto convencional con bloques y campos prefijados

3

El **spamming** no es considerado por IR tradicional pero crucial en Web

- Por ejemplo: resultados relevantes para la *query* pero con un *ranking* muy bajo

Base de datos documental

Con la evolución tecnológica, se tiende a **documentos multimedia** (combinan texto, video, imágenes, etc.)

La BDD **no almacena el documento entero** directamente

- Se guardan descriptores del documento (representación)
- BD más pequeña

  Más eficiencia

  Menos tiempo de búsqueda

BDD – Formulación matemática

- La BD es una matriz o tabla en la que cada **fila** es un documento y cada **columna** es un descriptor (p.ej. término)
  - Cada **celda** es valor UNO si el documento está representado por el descriptor, o valor CERO en caso contrario
  - Cada **documento** es un vector de 0 y 1, según el descriptor represente al documento

- N documentos → D= {d1, d2, ...d_N}
- T descriptores → T= {t1, t2, … t_T}

	t1	t2	t3	t4	t5	...	...	...	t_T
d1	■		■	■					
d2			■		■				■
d3	■						■		
d4		■	■	■					■
...									
...									
d_N		■		■		■			

Caso de la representación binaria

BDD – Formulación matemática

- Cada documento es un vector de **T** valores (**vector documental**)
  - La interpretación de los valores dependerá del **modelo de recuperación** que se trate:
    - **Modelo booleano** – La existencia o no del descriptor en el documento
    - **Modelo vectorial** – Grado en que el descriptor describe al documento
    - **Modelo probabilístico** – Probabilidad de que el descriptor sea relevante
- Al proceso de construcción de vectores documentales se le denomina **indexación**:
  - Mejora el acceso a los documentos
  - Define áreas de conocimiento que permiten relacionar unos documentos con otros
  - Permiten predecir la relevancia del documento frente a una necesidad de información
- La indexación puede ser manual o automática:
  - **Manual** – multiples problemas (pérdida de consistencia, diferentes niveles de exhaustividad, etc.)
  - **Automática** – require módulo indexador

BDD – Indexador

- Asocia automáticamente una representación a cada documento en función de los contenidos de información de este
- Determina los valores de cada descriptor **D** en el vector documental **T**
- **Función de ponderación**

$$F: D \times T \rightarrow [0, 1]$$

  - Los descriptores referenciados tendrán un valor $\neq 0$
  - Los descriptores no referenciados tendrán un valor 0

- En el caso de indexación binaria, toma 2 valores (0, 1), mientras que la indexación basada en pesos reales pondera según un rango [0,1] menos estricto
- Juega un papel fundamental en la calidad de la recuperación

¡Gracias!

UCO ONLINE