Chapter 1, 7 Recuperación de Información

Recuperación da Información (Universidade da Coruña)

CHAPTER 1

- ¿Qué es un documento? Todo texto tiene propiedades comunes:
  - Contenido significativo.
  - Algún tipo de estructura (título, autor, fecha, remitente, etc).
- Documentos vs. Registros de BBDD
  - Los registros de las BBDD (o tuplas en las BD relacionales) están compuestas típicamente por campos (o atributos) bien definidos y, por ello, los datos son más fáciles de comparar ya que su semántica está bien definida.
- Comparar texto.
  - Comparar el texto de la consulta con el texto del documento y determinar si es una buena coincidencia es el eje central de la recuperación de información.
  - Muchas veces la coincidencia exacta de palabras no es suficiente ya que, en un "lenguaje natural", existen muchas formas diferentes de escribir lo mismo. Además, algunos documentos satisfarán mejor nuestras necesidades de información que otros.

- Dimensiones de RI
  - La recuperación de información no solo se centra en la búsqueda de texto en documentos o en Web (aunque estos dos puntos son los centrales), también existen otras áreas donde necesitamos recuperar información.
  - En estas áreas, al igual que en los documentos, el contenido es difícil comparar y además, también de representar (aunque se puede usar texto para representarlo a través, por ejemplo, de etiquetas).
- Tareas RI.
  - Búsqueda ad-hoc: encontrar documentos relevantes para una consulta de texto arbitrario.
  - Filtrado: identificar perfiles de usuario relevantes para un documento.
  - Clasificación: identificar etiquetas relevantes para documentos.
  - Contestar preguntas: dar una respuesta específica a una pregunta.

- Problemas en RI.
  - Relevancia.
    - Un documento es relevante si contiene la información que la persona estaba buscando en ese momento con dicha consulta. Esto hace que la relevancia sea subjetiva ya que depende de factores como: contexto, localidad, etc.
    - Por ello podemos discernir dos tipos de relevancia:
      - Relevancia temática: el tema buscado es el mismo que el encontrado.
      - Relevancia del usuario: lo demás.

    - Los algoritmos de ranking que utilizan los motores de búsqueda se basan en modelos de recuperación que se basan en propiedades estadísticas en vez de lingüísticas, es decir, tienen en cuenta características de texto simples como palabras y no analizan oraciones.

- Evaluación.

- La evaluación son procedimientos y medidas experimentales para comparar la salida del sistema con las expectativas del usuario, estos procedimientos y medidas se refieren a la eficacia de las búsquedas. Recall y precisión son medidas de eficacia.
- Necesidades del usuario de información.
  - Las consultas de palabras clave a menudo son descripciones deficientes de las necesidades reales de información. Por ello, la interacción y el contexto son importantes para comprender la intención del usuario.
  - Existen técnicas de refinamiento de consultas como: la expansión de consultas, las sugerencias de consultas, los comentarios de relevancia que mejoran la clasificación, etc.

- Motores de búsqueda.
  - Un motor de búsqueda es la aplicación práctica de técnicas de recuperación de información a colecciones de texto a gran escala.
  - Problemas de los motores de búsqueda.
    - Rendimiento.
      - Para tener un buen rendimiento debemos medir y mejorar la eficiencia de búsqueda.
        - Reduciendo el tiempo de respuesta.
        - Aumentando el rendimiento de las consultas.
        - Aumentando la velocidad de indexación.
        - […]
      - Los índices son estructuras de datos diseñadas para mejorar la eficiencia de la búsqueda. Diseñarlos e implementarlos son temas importantes para los motores de búsqueda.
    - Incorporar nuevos datos (datos dinámicos).
      - Las colecciones de documentos para la mayoría de las aplicaciones reales están cambiando constantemente.
      - Las medidas típicas son la cobertura y frescura.
      - Actualizar los índices mientras se procesan las consultas también es un problema de los motores de búsqueda.
    - Escalabilidad.
      - Se centra en hacer que todo funcione a pesar de tener millones de usuarios todos los días y manejar muchos TB de documentos. Por ello, el procesamiento distribuido es esencial.
      - Desde la perspectiva práctica, esto se consigue replicando el HW: con más memoria, el algoritmo mejor funcionará.
    - Adaptabilidad.
      - Se centra en permitir cambios y ajustes de los componentes del motor de búsqueda.
      - Cambios como cambiar el algoritmo de clasificación, la estrategia de indexación, tener interfaz para diferentes aplicaciones, etc.
    - Problemas específicos.
      - SPAM.
        - Para la búsqueda web, el spam en todas sus formas es uno de los principales problemas. Afecta a la eficiencia de los motores de búsqueda y a la eficacia de los resultados.

CHAPTER 7

- Retrieval Models.
  - Proveen de un modelo matemático para definir el proceso de búsqueda.
  - El progreso en los modelos de búsqueda está relacionado con mejoras de la efectividad.

- Boolean Retrieval (modelo antiguo).
  - Dos posibles salidas para el procesamiento de la query.
    - TRUE / FALSE.
    - "exact-match"
    - La forma más simple de ranking.
  - Funciona mejor o peor en función de si el usuario sabe formular la query.
- Vector Space Model (modelo antiguo).
  - Los documentos y queries están representadas por un vector de términos.
  - Cada término tiene asociado un peso.

    Sparse matrix: los valores 0 no se codifican.
    D1 Tropical Freshwater Fish
    D2 Tropical Fish Aquarium Care Fish
    Se está guardando el term frequency.

	D1	D2
tropical	1	1
freshwater	1	0
aquarium	0	1
fish	1	2
care	0	1

  - Un índice invertido en el fondo es una implementación sparse de la matriz de términos por documento.
  - El modelo de espacio vectorial tiene tantas dimensiones como términos de indexación (index terms).
  - Se puede mide la similaridad entre documentos mediante la medida del coseno (cosine similarity).

- Esquemas de pesado.
  - Term frequency: número de ocurrencias del término k en el documento i. Mide la importancia en el documento.
  - Inverse document frequency: mide la importancia en la colección. Se suele representar de forma logarítmica.
  - Term Frequency logarítmico: si el término no aparece en el documento, 0, si el término sí aparece en el documento, $1 + \log f_{ik}$
    - Si pasa de aparecer 0 veces a 1 vez es más importante que si pasa de aparecer 1 vez a 2 veces.
    - Comprime el raw tf.

- Relevance feedback. Proceso por el que se formula una query y el sistema proporciona un ranking de documentos.
    - Rocchio algorithm.
    - El usuario marca qué es relevante.
    - El propósito es que dada la query original *q* y con la información del usuario el sistema es capaz de generar una query nueva *q'* que debería ser mejor.

- Pseudo-relevance feedback. Proceso por el cual un search engine obtiene un ranking de documentos que cree que están ordenados por relevancia y supone que el primero (o los 5 primeros, o los 10 primeros) son relevantes. Y a partir de esa suposición construye una nueva query.

- Implicit Relevance. Aquella que aparece normalmente cuando alguien le da al play en una lista de canciones.

- Language Model.
    - Distribución de probabilidad sobre las palabras de un lenguaje.
    - Multinomial distribution over words.
        - El texto se va a modelar como una secuencia de palabras donde hay, en cada posición de esa secuencia, una palabra con una probabilidad.

- Query-Likelihood Model.
    - Para todos los query terms hay que calcular la probabilidad de q_i dado D.

- Maximum Likelihood Estimate (MLE).
    - Asigna las probabilidades en función de lo que ve.
    - Utiliza el número de ocurrencias de la palabra en el documento y lo divide entre el número de veces que aparece el documento.

- Smoothing. Técnica para estimar la probabilidad de las *missing words* (palabras no vistas).
- Jelinek-Mercer Smoothing.
    - Mezcla la probabilidad en el documento con la frecuencia de la palabra en la colección general.
    - Problemas de precisión al multiplicar números pequeños.
    - λ da el peso de suavizado.
    - En exámenes él pide el ranking score, no el logaritmo.

$$P(Q|D) = \prod_{i=1}^{n}((1-\lambda)\frac{f_{q_i,D}}{|D|} + \lambda\frac{c_{q_i}}{|C|})$$

- Dirichlet Smoothing.
    - Un término que no ocurre en un documento largo debería tener asignado un valor de smoothed probability bajo (cuanto más largo el documento, más baja la probabilidad).
- En el IndexWriter hay que indicarle que estamos trabajando con Language Models (a la hora de indexar y a la hora de hacer la búsqueda).

- Modelos de Relevancia.
    - Distribución de probabilidad ideal de las palabras en ese modelo de relevancia.
    - Modelo de lenguaje que representa la información necesaria.
    - Asume que hay un modelo de relevancia y lo va a estimar.
    - P(D/R) probabilidad de generación de texto en un documento dado un modelo de relevancia.
        - document likelihood model
            - Menos eficaz que query likelihood.
    - Se puede estimar sólo con la query. Pero también con la query y los top rated documents - Pseudo-Relevance Feedback.
    - Rank los documentos dependiendo de la similaridad del modelo de documento con el relevance model.
    - Kullback-Leibler divergence (KL-divergence) - medida de diferencia entre dos distribuciones de probabilidad.
        - Siempre no negativa.
    - Estimar el Relevance Model.
        - Probabilidad de la palabra dados los términos de la query.