Gestión de Información en la Web
Máster en Ingeniería Informática
Parte II: Recuperación de información y Sistemas de recomendación

Tema 4: Recuperación de información
4.7. Técnicas avanzadas de RI

Juan Manuel Fernández Luna
Dpto. Ciencias de la Computación e Inteligencia Artificial
jmfluna@decsai.ugr.es

Índice

- Técnicas de modificación de consulta.
- Clasificación y agrupamiento documental.
- Sistemas de búsqueda de respuesta.
- RI Patrocinada.
- Detección y seguimiento de temas de actualidad.
- Construcción automática de resúmenes.

Técnicas de modificación de la consulta

Cuando buscamos por “coche” no recuperamos documentos que tratan de “autos” o “automóviles” ¿Estamos perdiendo documentos relevantes?

Necesidad de mejorar los resultados

¿Cómo?

Técnicas de modificación de la consulta

- Métodos globales: expansión de consulta.
- Métodos locales: realimentación por relevancia y pseudo realimentación.

Realimentación de relevancia

- Realimentación por parte del usuario sobre el conjunto de documentos inicialmente recuperados:
  - El usuario formula una consulta simple y corta.
  - El usuario marca algunos resultados como relevantes o no relevantes.
  - El sistema calcula una mejor representación de la necesidad de información basada en la realimentación.
  - Varias iteraciones.

Algoritmo de Rocchio

- En la práctica:

$$\vec{q}_m = \alpha \vec{q}_0 + \beta \frac{1}{|D_r|} \sum_{d_j \in D_r} \vec{d}_j - \gamma \frac{1}{|D_{nr}|} \sum_{d_j \in D_{nr}} \vec{d}_j$$

- *D_r = Vectores del conjunto de documentos relevantes conocidos.*
- *D_{nr}* = Vectores del conjunto de documentos no relevantes conocidos.
  - Diferentes de los conjuntos C_r y C_{nr}
- *q_m* = consulta modificada; *q_0* = vector de la consulta original; *α,β,γ*: pesos (elegidos a mano o empíricamente)
- La consulta nueva se mueve hacia los documentos relevantes y se aleja de los irrelevantes.

Expansión de consulta

- Para cada término t de la consulta, expandir la consulta con sinónimos y palabras relacionadas con t en el **tesauro**.
  - Gato → felino.
- Se suelen ponderar más bajo a los términos añadidos que a los originales.
- Incrementa el recall.
- Ampliamente usados en entornos científicos y de ingeniería.
- Puede decrementar la precisión, sobre todo con términos ambíguos:
  - “interest rate” → “interest rate fascinate evaluate”
- La construcción manual del tesauro es costosa y su actualización.

Clasificación y agrupamiento

- En recuperación de información:
- Clasificación o categorización documental:
  - Cada documento pertenece a una clase conocida.
  - Nuevos documentos se tienen que situar en uno de los grupos existentes de manera correcta.
- Agrupamiento documental:
  - Los documentos no están etiquetados y los grupos deben ser descubiertos.

Aplicaciones del agrupamiento

- Navegación y visualización de conjuntos de documentos recuperados.
- Análisis de colecciones completas.
- Mejora de la efectividad de recuperación de los motores de búsqueda (modelo de recuperación basado en agrupamiento).

Aplicaciones de la clasificación

- Organización documental:
  - Por ejemplo, en un periódico, previamente a la publicación de anuncios, éstos pueden ser clasificados en “venta de coches”, “venta de casas”,...

Aplicaciones de la clasificación

- Filtrado de texto:
  - Clasificación de un flujo de documentos entrantes, dependiendo de su relevancia para el usuario. (Por ejemplo, una agencia de noticias, que manda unas noticias a unos clientes y otras, a otros).
  - Normalmente binario (relevante – no relevante).
  - Es común tener un perfil del usuario, donde éste establezca sus gustos, y que pueda ser actualizado, explícitamente por el propio usuario o implícitamente por el sistema (Filtrado adaptativo).

Aplicaciones de la clasificación

- Desambiguación del significado de las palabras:
  - ¿banco = institución financiera o asiento?
  - Desambiguación: dada la ocurrencia ambigua de una palabra en un texto, asignarle su significado correcto.
  - Los contextos de las ocurrencias de las palabras se pueden ver como documentos, mientras que los significados como categorías.
  - Se dispone de un conjunto de “documentos” asignados a las “categorías” correctas, y se intenta encontrar el significado correcto de nuevas palabras en un contexto.

Sistema de búsqueda de respuestas

- Tratan de proporcionar la respuesta exacta a una pregunta, no el documento que la contenga.

  ¿Cuándo nació Charles Darwin?

  12 de febrero de 1890

  ¿Cuál es el PNB de Japón?

  Tabla con los datos históricos

- Respuesta a pregunta explícita ó lista de términos.

Sistema de búsqueda de respuestas

- Preguntas factuales:
  - Métodos superficiales.
  - Identificación del tipo de pregunta (clasificación: qué, cuándo, quién, dónde,...).
  - Extracción de contextos (lanzar la respuesta a un SRI para obtener párrafos).
  - Selección del dato concreto a partir de los diferentes contextos.

Sistema de búsqueda de respuestas

- Preguntas factuales:
  - Métodos superficiales.
  - Identificación del tipo de pregunta (clasificación: qué, cuándo, quién, dónde,...).
  - Extracción de contextos (lanzar la respuesta a un SRI para obtener párrafos).
  - Selección del dato concreto a partir de los diferentes contextos.

RI Patrocinada

- Proporcionar a los usuarios anuncios altamente relevantes a sus necesidades de información en un momento concreto.
- La mayor parte de motores de búsqueda de la web proporcionan este servicio.
- Forma de financiación.
- Resultados orgánicos vs patrocinados.
- ¿Diferenciados o integrados?

RI Patrocinada

- **Documento = información suministrada por el anunciador**:
  - Título, breve descripción, página web destino y posibles consultas para las cuales el anuncio es relevante (palabras clave), zona geográfica, idioma, ...
  - ¿Dónde se puede hacer mejor RI Patrocinada?
  - es decir, ¿dónde existe muchas más información del usuario útil para la RI Patrocinada?

Detección y seguimiento temas actualidad

- **Contexto**: noticias → documento de texto donde se informa sobre un determinado suceso, con narrativa que responde el qué, dónde, cómo y porqué del mismo.
- **Topic Detection and Tracking (TDT)**.
- Línea de investigación que permite al usuario realizar un seguimiento de los temas de actualidad a partir de varias fuentes.

Detección y seguimiento temas actualidad

- **Objetivo**:
  - Organización de sucesos y detectar aquellos nuevos.
  - Identificar noticias que conforman un suceso al que hay que darle seguimiento.

Detección de novedad

¿Novedad?

- *"Novedad o nueva información significa nuevas respuestas a las preguntas potenciales que representan una petición del usuario o necesidad de información" (Li y Croft, 2008).*
- Dos vertientes:
  - Necesidad de información → consultas.
  - Información novel → detectando documentos que contienen preguntas que no han sido respondidas en respuestas previas (documentos vistos por el usuario).

Detección de novedad

- **Diversidad**:
  - Cada una de las posibles interpretaciones de una consulta.
  - Irak → conflictos, información geográfica, histórica, etc.
  - Subtema o faceta: cada una de las interpretaciones o subtemáticas relacionadas con una consulta.
  - Es interesante introducir diversidad en una salida del SRI.

Detección de novedad

¿Novedad = diversidad?

Normalmente no intercambiables, pero...

… al buscar novedad se promueve la diversidad.

- **Detección de novedad**: encontrar información novel. Dado un documento y un conjunto de documentos vistos previamente, un documento deberá tratar de un tema o interpretación que es diferente a cualquier otra ya tratado previamente o, al menos, hacer referencia a una faceta o subtema distinto de los cubiertos anteriormente.

Construcción de resúmenes

- La **construcción automática de resúmenes de documentos** (summarization) consiste en, dados una fuente de información (uno o más documentos textuales) y un demandante (usuario o aplicación), extraer el contenido de la fuente de información y presentarlo al demandante de forma condensada, comprensible, y que satisfaga sus necesidades.
- **Resumen**= corto, preservar información y estar desprovisto de redundancia.

Construcción de resúmenes

Tipos de resúmenes: