UDLA PUEBLA

Departamento de Ingeniería en Sistemas Computacionales

Administración de la Información
(IS 346)

Colecciones no estructuradas:
Modelo Vectorial de RI, Parte II

© 2004, Alfredo Sánchez

Colecciones no estructuradas: Modelo Vectorial de RI, Parte II

- Asignación de pesos
- Cálculo del índice de similitud
- Uso de DBMSs
- Implementación del modelo vectorial
- SQL y el modelo vectorial

© 2004, Alfredo Sánchez

Modelo vectorial: pesos y similitud

- Asigna pesos no binarios a cada término en los documentos de una colección
  - Los esquemas de asignación de pesos varían: TFIDF es frecuente
- Jerarquiza documentos de acuerdo a su grado de similitud con una consulta dada
  - El cálculo de similitud varía: el coseno del ángulo entre vectores es típico

© 2004, Alfredo Sánchez

Asignación de pesos

- A partir de experimentos:

$$w_{ij} = \frac{(\log tf_{ij} + 1) * idf_j}{\Sigma_{j=1,t} [(\log tf_{ij} + 1) * idfj]^2}$$

- otros esquemas exploran como contrarrestar diversos problemas, como el impacto de términos con frecuencias altas

© 2004, Alfredo Sánchez

Medidas de similitud

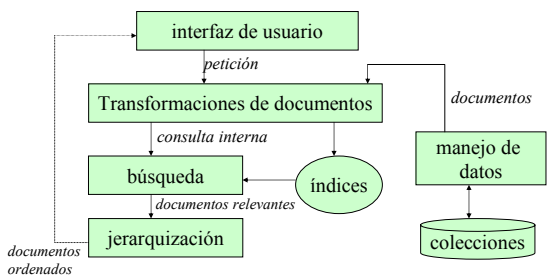
- Coseno del ángulo entre *Q* y D_i:

$$sim(Q, D_i) = \frac{\Sigma_{j=1,t} w_{qj} \times d_{ij}}{\sqrt{\Sigma_{j=1,t} (d_{ij})^2 \Sigma_{j=1,t} (w_{qj})^2}}$$

- Otros esquemas buscan reducir problemas como el impacto del tamaño de documentos o la eficiencia en cálculos

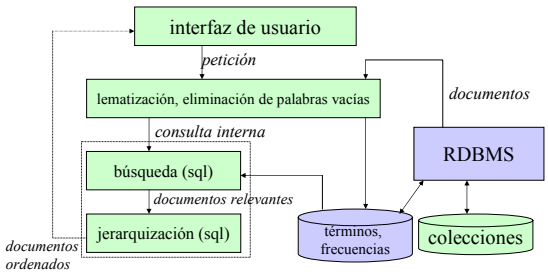
© 2004, Alfredo Sánchez

Implementación de modelos de IR



© 2004, Alfredo Sánchez

Modelo vectorial usando RDBMSs



© 2004, Alfredo Sánchez

Tablas posibles para el modelo vectorial

- *Docs* (IdDoc, NombreDoc, FechaPub, …, texto)
- *IndiceInv* (IdDoc, Term, tf)
- *Terms* (term, idf)
- *Consulta* (term, tf)
- Un pre-proceso puede producir *Docs* (extrayendo datos de archivos) e *IndiceInv* (vía parser, lematización, eliminación de palabras vacías)
- *Terms* puede producirse mediante:
  - insert into terms
    select term, log(N/count(*))
    from IndiceInv
    group by term
- donde N es el número total de documentos y se conoce o se puede obtener previamente (select count(*) from Docs)
- *Consulta* es solamente una representación tabular de la petición del usuario para facilitar el cálculo del índice de similitud

© 2004, Alfredo Sánchez

Uso de SQL para el modelo booleano

- Encontrar los ids de documentos que contengan los término "casa" o "blanca"
  - select distinct IdDoc from IndiceInv
    where term = 'casa' or term = 'blanca'
- Recuperar el texto de los documentos…
  - select texto from Docs
    where IdDoc in
    (select distinct IdDoc from IndiceInv
    where term = 'casa' or term = 'blanca')
- ¿Cómo se obtendrían los documentos que contienen "casa" Y "blanca"?

© 2004, Alfredo Sánchez

SQL para el modelo booleano

- "AND" usando agregación y agrupación:
  - select IdDoc from IndiceInv i, consulta q
    where i.term = q.term
    group by i.IdDoc
    having count(i.term) = (select count(*) from consulta)
- en este caso, "where" asegura que se incluyen los términos pedidos, "group by" agrupa los términos para cada documento, y "having" que haya tantos términos en el documento como en la consulta

© 2004, Alfredo Sánchez

SQL para espacios vectoriales usando TFIDF

- select i.IdDoc, sum(q.tf * t.idf * i.tf * t.idf)
  from consulta q, IndiceInv i, Terms t
  where q.term = t.term
  and i.term = t.term
  group by i.IdDoc
  order by 2 desc
- "where" garantiza que se incluyen sólo terminos de la consulta, "order by", que se ordenan según el coeficiente de similitud

© 2004, Alfredo Sánchez

Cálculo del coseno del ángulo

- Tablas intermedias:
  - pesos_docs(IdDoc, peso)
  - peso_q(peso)
- Generación de tablas intermedias:
  - insert into pesos_docs
    select IdDoc, sqrt(sum(i.tf * t.idf * i.tf * t.idf))
    from IndiceInv i, Terms t
    where i.term = t.term
    group by IdDoc
  - insert into peso_q
    select sqrt(sum(q.tf * t.idf * q.tf * t.idf))
    from consulta q, Terms t
    where q.term = t.term

© 2004, Alfredo Sánchez

… Cálculo del coseno del ángulo

- select i.IdDoc, sum(q.tf * t.idf * i.tf * t.idf) / (dw.peso * qw.peso)
  from consulta q, IndiceInv i, terms t, pesos_docs dw, peso_q qw
  where q.term = t.term AND
  i.term = t.term AND
  i.IdDoc = dw.IdDoc
  group by i.IdDoc, dw.peso, qw.peso
  order by 2 desc

© 2004, Alfredo Sánchez

Resumen

- Pesos e índices de similitud son las características más importantes del modelo vectorial
- Gran parte de la investigación en RI se enfoca a la determinación de mejores métodos para asignar pesos y calcular similitudes
- El uso de RDBMSs y SQL puede facilitar la implementación del modelo vectorial

© 2004, Alfredo Sánchez