Análisis Global y Local

UCR – ECCI

CI-2414 Recuperación de Información

Prof. Kryscia Daviana Ramírez Benavides

Análisis Global

- Realiza la expansión basado en la construcción de tesauros utilizando la colección completa de documentos para añadir a la consulta los términos más cercanos a los de la consulta en el tesauro.
- La idea de este tipo de expansión de la consulta es considerar todo el conjunto de documentos en la colección.
- Se puede hacer basadas en la similaridad de conceptos o en las estadísticas de aparición de los términos que representan conceptos.
- Existen dos claros enfoques:
  - Expansión basada en un tesauro de similitud.
  - Expansión basada en un tesauro estadístico.

Análisis Global – Tesauro de Similitud

- Un tesauro de similitud está basado en relaciones de término a término, esta similitud no es establecida por la correlación entre términos.
- La similaridad es obtenida considerando que los términos son conceptos en un espacio de conceptos.
- Es este espacio, cada término es indexado por el documento en el que aparece. Así, los términos asumen el rol de documentos y los documentos como elementos de indexación.

Análisis Global – Tesauro de Similitud (cont.)

- Definición:
  - Sea t el número de términos en una colección, N el número de documentos en una colección, y $f_{i,j}$ la frecuencia de ocurrencias de un término k_i, en el documento d_j. Sea t_j el número de términos distintos en un documento d_j y itf_j la inversa de la frecuencia del término en documento d_j.
  - Entonces:

$$itf_j = \log \frac{t}{t_j}$$

$$\vec{k}_i = (w_{i,1}, w_{i,2}, \ldots, w_{i,N})$$

$$w_{i,j} = \frac{\left(0.5 + 0.5 \frac{f_{i,j}}{\max_j(f_{l,j})}\right) itf_i}{\sqrt{\sum_{m=1}^{N} \left(0.5 + 0.5 \frac{f_{i,m}}{\max_m(f_{l,m})}\right)^2 itf_i^{\,2}}}$$

$$c_{u,v} = \vec{k}_u \cdot \vec{k}_v = \sum_{\forall d_j} w_{u,j} \times w_{v,j}$$

Análisis Global – Tesauro de Similitud (cont.)

- Expansión de consulta con tesauro de similitud es dado en tres etapas:
  - Represente la consulta en el espacio de conceptos usados para representar los términos índices. Para ello:

$$\vec{q} = \sum_{k_i \in q} w_{i,q} \vec{k}_i$$

  - Basado en el tesauro de similaridad global, calcule la similaridad $sim(q,k_v)$ entre cada término k_v correlacionado con los términos en la consulta y la consulta completa. Para ello:

$$sim(q,k_v) = \vec{q} \cdot \vec{k}_v = \sum_{k_u \in q} w_{u,q} \times c_{u,v}$$

Análisis Global – Tesauro de Similitud (cont.)

- Expansión de consulta con tesauro de similitud es dado en tres etapas (cont.):
  - Expanda la consulta con los *r* mejores jerarquizados términos de acuerdo a $sim(q,k_v)$. El peso asignado al término agregado a la consulta es:

$$w_{v,q'} = \frac{sim(q,k_v)}{\sum_{k_u \in q} w_{u,q}}$$

Análisis Global – Tesauro Estático

- El tesauro global es compuesto de clases, las que agrupan términos correlacionados en el contexto de la colección completa.
- Tales términos correlacionados pueden ser usados para expandir la consulta original.
- Para ser efectivos, los términos tienen que ser altamente discriminantes (o sea, baja frecuencia). Sin embargo, es difícil agrupar términos con baja frecuencia.
- Así, el agrupamiento se hace por clases y usa los términos de baja frecuencia para definir estas clases.

Análisis Global – Tesauro Estático (cont.)

- Una estrategia es el **algoritmo *HAC* de Enlace Completo** que se describe:
  1. Calcular una matriz de similaridad M tal que $M[i, j] = sim(i, j)$ es la similaridad entre el *cluster i* y el *cluster j*.
  2. Inicialmente, colocar cada documento en un *cluster* propio.
  3. Unir los dos *clusters* más similares en uno solo.
  4. Actualizar M para reflejar el cambio producido por la unión de *clusters*.
  5. Repetir los pasos 3 y 4 hasta que haya un único *cluster*.
- La similaridad entre *clusters* es definida como la mínima de las similaridades entre pares de documentos *inter-cluster*, la función de similaridad es la fórmula del coseno del modelo vectorial.

Análisis Global – Tesauro Estático (cont.)

- Dado la jerarquía de *cluster* para una colección completa, la selección de términos se hace por lo siguiente:
  - Obtenga los parámetros: TC, NCD y MINDF.
  - Use TC para determinar los *clusters* de documentos a ser usados.
  - Use el NDC como límite del tamaño del *cluster*.
  - Seleccione los documentos con baja frecuencia como origen de términos.
  - El parámetro MINDF define el valor mínimo de la frecuencia del documento inversa para cualquier término seleccionado para participar el tesauro de clases.

Análisis Global – Tesauro Estático (cont.)

- Una vez que se ha definido la jerarquía de *clusters* o tesauro de clases, el promedio del peso de un término para cada clase del tesauro es:

$$wt_C = \frac{\sum_{i=1}^{|C|} w_{i,C}}{|C|}$$

- Donde:
  - $|C|$ es el número de términos en la clase tesauro C.
  - $w_{i,C}$ es el peso pre-calculado asociado con el par término-clase $[k_i, C]$.
- El promedio del peso de un término puede ser usado para calcular el peso de una clase en el tesauro:

$$w_C = \frac{wt_C}{|C|} \times 0.5$$

Análisis Local

- Realiza la expansión con los documentos recuperados por una consulta donde se realizan clusters basados en la correlación de términos
- Utiliza esta información para expandir las consultas añadiendo a las mismas los términos correlacionados con los de la consulta.
- Este enfoque trata de obtener un conjunto más grande de objetos relevantes automáticamente.
- Esto usualmente consiste en identificar sinónimos, variaciones terminales, o términos que están cercanos a los términos de la consulta en el texto.

Análisis Local (cont.)

- En el análisis local, los documentos recuperados para una consulta son examinados para determinar términos de expansión, lo cual es hecho sin el apoyo del usuario.
- Existen dos claros enfoques:
  - Agrupamiento local.
    - Agrupamiento de Asociación.
    - Agrupamiento Métrico.
    - Agrupamiento Escalar.
  - Análisis de contexto local.

Análisis Local – Agrupamiento

- Sea *V(s)* el conjunto no vacío de palabras que son variaciones gramaticales entre ellas.
- Por ejemplo:
  - *V(s)* = {computador, computadores, computacional, computación}
  - *s* = computa (prefijo común)
- Para cada consulta *q* dada:
  - El conjunto D_l de documentos recuperados es llamado conjunto de documentos locales.
  - El conjunto V_l de todas las palabras distintas en los documentos locales es llamado vocabulario.
  - El conjunto de todos los prefijos comunes es llamado S_l.

Análisis Local – Agrupamiento (cont.)

- *Agrupamiento de Asociación*
  - El agrupamiento de asociación está basado en la concurrencia de raíces o términos dentro de los documentos.
  - Dos raíces que concurren frecuentemente en los documentos recuperados tienen una asociación de sinonimia.
  - La idea es que prefijos comunes que frecuentemente concurren en los documentos tienen asociación de sinonimia.

Análisis Local – Agrupamiento (cont.)

- *Agrupamiento de Asociación*
  - Definición:
    - La frecuencia de un prefijo en un documento d_j, $d_j \in D_l$ es llamado $f_{si,j}$.
    - Sea $m = (m_{i,j})$, la matriz de asociación con $|S_l|$ filas y $|D_l|$ columnas, donde $m_{i,j} = f_{si,j}$.
    - Sea m^t la matriz transpuesta de m.
    - La matriz $s = mm^t$ es la matriz de asociación local de prefijo-prefijo.
    - Cada elemento $s_{u,v}$ expresa la correlación $c_{u,v}$ entre prefijos s_u y s_v:

$$c_{u,v} = \sum_{d_j \in d_l} f_{s_u,j} \times f_{s_v,j}$$

    - Una normalización del factor de correlación $c_{u,v}$ es:

$$s_{u,v} = \frac{c_{u,v}}{c_{u,u} + c_{v,v} - c_{u,v}}$$

Análisis Local – Agrupamiento (cont.)

- *Agrupamiento Métrico*
  - El agrupamiento no toma encuentra dónde los términos ocurren en el documento.
  - La idea del agrupamiento métrico es considerar la distancia entre los términos para determinar su concurrencia.
  - Dos términos que ocurren en la misma frase parecen más correlacionados que aquellos que aparecen uno al inicio y otro al final, función de distribución de aparición.

Análisis Local – Agrupamiento (cont.)

- *Agrupamiento Métrico*
  - Definición:
    - La distancia $r(k_i,k_j)$ entre dos palabras está dada por el número de palabras entre ellas en el mismo documento.
    - Si las palabras están en distintos documentos, su distancia es ∞. La matriz de correlación de prefijos es definida como:

$$c_{u,v} = \sum_{k_i \in V(s_u)} \sum_{k_j \in V(s_v)} \frac{1}{r(k_i, k_j)}$$

Análisis Local – Agrupamiento (cont.)

- *Agrupamiento Escalar*
  - Otra forma de determinar sinonimia entre dos términos locales s_u y s_v es comparando el $S_u(n)$ y $S_v(n)$.
  - La idea es que dos palabras con similar vecindad tienen una relación de sinonimia.
  - Una forma de cuantificar la vecindad es organizar todos los valores de correlación $s_{u,i}$ en un vector s_u, organizar todas las correlaciones $s_{v,i}$ en otro vector s_v, y comparar estos vectores por una medida escalar.
  - Por ejemplo, el coseno del ángulo entre los vectores, así:

$$s_{u,v} = \frac{\vec{s}_u \cdot \vec{s}_v}{\left|\vec{s}_u\right| \times \left|\vec{s}_v\right|}$$

Análisis Local – Análisis de Contexto Local

- Basado en el uso de grupos de sustantivos (únicos, dos sustantivos adyacentes, o tres sustantivos adyacentes en el texto) como conceptos de documentos.
- Para una expansión de consulta, los conceptos son seleccionados dentro de los documentos mejor jerarquizados basado en su correlación con términos (sin análisis de prefijos) de la consulta.
- Sin embargo, en vez de considerar el documento, una ventana de texto es usada para determinar la concurrencia (como se haría en un análisis global).

Análisis Local – Análisis de Contexto Local (cont.)

- Las etapas de este análisis son:
  - Recuperar los *n* mejores documentos de respuesta a una consulta. Estos documentos son divididos en pasajes o ventanas de texto.
  - Para cada concepto *c* dentro de los mejores evaluados pasajes se calcula la similaridad *sim*(*q*,*c*) entre toda la consulta *q* y el concepto *c* usando una variación del *ranking tf-idf*.
  - Los *m* mejores conceptos son entonces agregados a la consulta. Para cada concepto agregado se le asigna un peso $1 - 0.9i/m$.
  - Los términos en la consulta original pueden ser remarcados al duplicar su peso.

Análisis Local – Análisis de Contexto Local (cont.)

$$sim(q,c) = \prod_{k_i \in q} \left(\delta + \frac{\log(f(c,k_i) \times idf_c)}{\log n} \right)^{idf_i}$$

$$idf_i = \max\left(1, \frac{\log_{10} N / np_i}{5}\right)$$

$$f(c,k_i) = \sum_{j=1}^{n} pf_{i,j} \times pf_{c,j}$$

$$idf_c = \max\left(1, \frac{\log_{10} N / np_c}{5}\right)$$

- Donde:
  - $pf_{i,j}$ es la frecuencia del término *i* en el *j-ésimo* pasaje.
  - k_i en el *j-ésimo* pasaje.
  - $pf_{c,j}$ es la frecuencia del concepto *c* en el *j-ésimo* pasaje.
  - np_i es el número de pasajes que contienen el término k_i.
  - np_c es el número de pasajes que contienen el concepto *c*.
  - δ es un valor pequeño y distinto de cero.

Referencias Bibliográficas

- La información fue tomada de:
  - Libro de texto del curso.
  - http://www.inf.udec.cl/~andrea/cursos/retrieval/consultas.pdf.