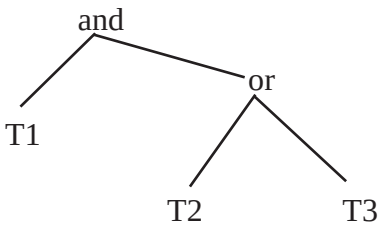
ESPECIFIACION DE CONSULTAS

Tipos de Consulta

1. Texto

- Consulta por palabra única
- Consulta in contexto
  - frase
  - proximidad
- Consulta booleana



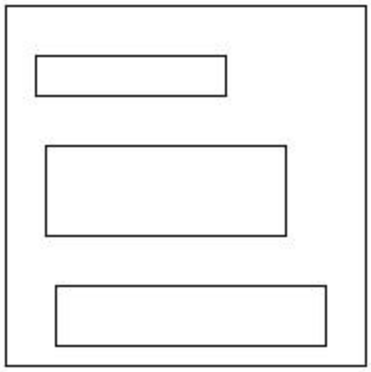
- Lenguaje natural

2. *Correspondencia de patrones*

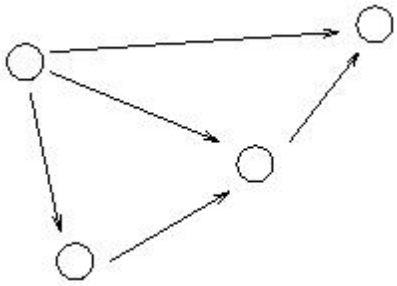
- Lenguaje natural
- Palabra
- Prefijos
- Sufijos
- Substring
- Rango
- Permitiendo errores
- Expresiones regulares
  - Union
  - Concatenación
  - Repetición
- Patrones extendidos
  - Clase de caracteres
  - Expresiones condicionales
  - Combinations

3. *Consulta estructural*

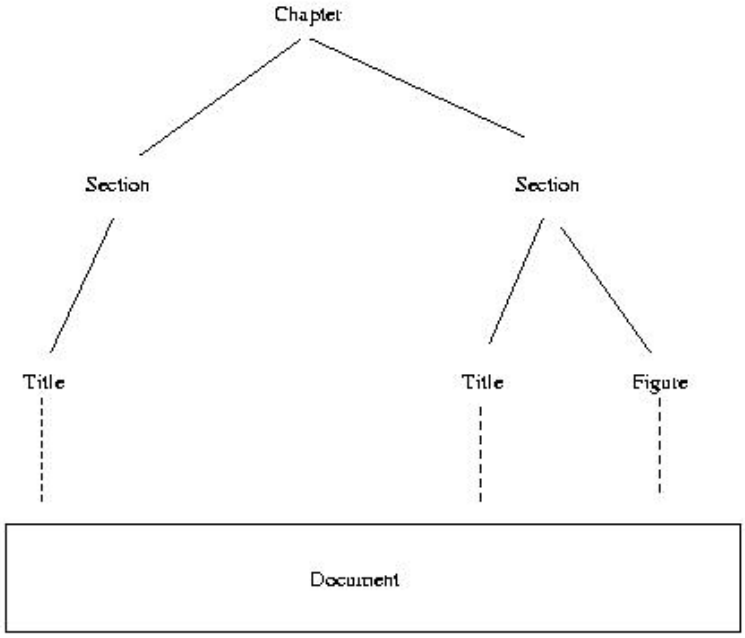
Estructura fija



Hypertexto



Estructura jerárquica



4. *Consulta de información espacial*

- Ventana (qué)
- Objectos (dónde)
- Consulta combinada o estructural:
  - Objects
  - Relaciones
  - Problema de satisfacción de restricciones.

Operaciones sobre Consultas

La idea es mejorar la formulación de la consulta inicial a través de la expansión de la consulta y re-peso de palabras claves. Estos enfoques se agupan en tres categorias:
(1) retroalimentación de parte de usuario, (2) derivación del grupo de documentos incialmente derivados (análisis local) y (3) derivados de la información global de los documentos en la colección (análisis global).

1. *Retroalimentación del usuario.* El usuario es presentado con una lista de documentos recuperadosd y después de examinarlos, marca los que son relevantes. A los términos o plabras claves de los documentos seleccionados como relevantes se les da más importancoa en la reformulación de la consulta.

En el caso del modelo vectorial, la retroalimenatción considera que los vectores de peso de los documentos identificados como relevantes son similares entre ellos. Más aún, documentos no relevantes tienen vectores que son distintos a los de los relevantes. Entonces, la consulta es reformulada de manera que sea más cercana al espacio de vectores de pesos de los documentos relevantes.

Considere la siguiente terminología:

D_r: conjunto de documentos relevantes identificados por el usuario entre los documentos recuperados inicialmente.
D_n: conjunto de documentos no relevantes entre los documentos recuperados inicialmente.
C_r: conjunto de todos los documentos relevantes en la colección.
$|D_r|,|D_n|,|C_r|$: número de docuementos en los respectivos conjuntos.

Si se supiera a-priori el conjunto de todos los documentos relevantes a una consulta, la consulta óptima estaría dada por:

$$\bar{q}_{optimal} = \frac{1}{|C_r|} \sum_{\forall \bar{d}_j \in C_r} \bar{d}_j - \frac{1}{1-|C_r|} \sum_{\forall \bar{d}_j \notin C_r} \bar{d}_j$$

Como no se sabe el conjunto C_r, se realiza una expansión incremental, donde hay tres clásicas y similares formas de determinarla son:

$$\bar{q}_m = \alpha\bar{q} + \frac{\beta}{|D_r|}\sum_{\forall \bar{d}_j \in D_r} \bar{d}_j - \frac{\gamma}{|D_n|}\sum_{\forall \bar{d}_j \in D_n} \bar{d}_j$$

$$\bar{q}_m = \alpha\bar{q} + \beta\sum_{\forall \bar{d}_j \in D_r} \bar{d}_j - \gamma\sum_{\forall \bar{d}_j \in D_n} \bar{d}_j$$

$$\bar{q}_m = \alpha\bar{q} + \beta\sum_{\forall \bar{d}_j \in D_r} \bar{d}_j - \gamma \max_{non-relevante(\bar{d}_j)}$$

2. *Análisis Local Automático*. Este enfoque trata de obtener un conjunto más grande de objetos relavantes automáticamente. Esto usualmente consiste en identificar sinónimos, variaciones terminales, o términos que están cercanos a los términos de la consulta en el texto. En el análisis local, los documentos recuperados para una consulta son examinados para determinar términos de expansión. Esto es hecho sin el apoyo del usuario. Existen dos claros enfoques: agrupamiento local y análisis de contexto local.

2.1 Agrupamiento.

Definición: 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).

Definición: Para cada consulta dada q, el conjunto D_l de documentos 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 predijos comunes es llamdo S_l.

Agrupamiento de Asociación. El agrupamiento de asociación está basado en la co-ocurrencia de términos dentro del documento. La idea es que prefijos comunes que frecuentemente co-ocurren en los documentos tienen asociación de sinonimia.

Defínción: La de fecuencia de un prefijo si en un documento d_j, $d_j \in D_l$ es llamado $f_{s_i,j}$. Sea $\bar{m} = (m_{i,j})$ la matriz de asociación con $|S_l|$ filas y $|D_l|$columnas, donde $m_{i,}j = fs_{i,j}$. Sea $\bar{m}^t$ la matriz transpuesta de $\bar{m}$. La matriz $\bar{s} = \bar{m}\bar{m}^t$ la matriz de asociación local de prefijo-prefijo. Cada elemento $s_{u,v}$ en 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:

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

Usando esta correlación, se construyen agrupamiento de asociación de la siguiente forma

Definición. Consider la *u*-ésima fila en la matriz de asociación $\bar{s}$. Sea $S_u(n)$ la función que toma la *u*-ésima fila y retorna el conjunto de los *n* valores más grandes $s_{u,v}$, donde *v* varía sobre el conjunto de prefijos locales y $v \neq u$. Entonces $S_u(n)$ define un agrupamiento de asociación alrededor de s_u. Si $s_{u,v}$ es dado por la ecuación normalizada, el agrupamiento de asociación se dice normalizado.

Agrupamiento Métrico. El agrupamiento no toma encuenta dónde los términos ocurren en el documento. La idea del agrupamiento métrico es considerar la distancia entre los términos para determinar su co-ocurrencia.

Definición: La distance $r(k_i,k_j)$ 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 matrix de correlación de prefijos es definida como:

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

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 $\bar{s}_u$, organizar todas las correlaciones $s_{v,i}$ en otro vector $\bar{s}_v$ y comparar estos vectores por una medidad escalar. Por ejemplo, el coseno del ángulo entre los vectores. Así,

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

2.2 Análisis de Contexto Local

Basado en el suo de grupos de sustantivos (sustantivos único, 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 cosiderar el documento, una ventana de texto es usada para determinar la co-ocurrencia (como se haría en un análisis global). Las tres 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 mejor 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 $1 - 0.9i/m$ donde u es la posición del concepto *i* en el ranking del concepto. Los términos en la consulta original pueden ser remarcados al duplicar su peso.

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

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

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

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

donde $pf_{i,j}$ es la frecuencia del término 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, y δ es un valor pequeño y distinto de zero.

3. Análisis Global

La idea de este tipo de expansión de la consulta es considerar todo el conjunto de documentos en la colección.

3.1 Expansión Basada en un Tesauro de Similitud

Un tesauro de similitud es basado en relaciones de término a término. Estas similtud 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 indezado por el documento en el que aparece. Así términos asumen el rol de documentos y los documentos como elementos de indezación.

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 documetno d_j y itf_j el inverso de la frecuencia de términos en documento d_j. Entonces:

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

$$\bar{k}_i = (w_{i,1}, w_{i,2}, ..., w_{i,N})$$

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

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

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 de índices. Para ello,

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

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

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

- Expanda la consulta con los *r* mejor jerarquizados términos deacuerdo a $sim(q,k_v)$. El peso asignado al término agregado a la consulta es:

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

3.1 Expansión Basada en un Tesauro Estadístico

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 altamento discriminantes (osea baja frecuencia). Sin embargo, es dificil agrupar términos con baja frecuencia. Así, el agrupamiento se hace por clases y usa los térmknos de baja frecuencia para definir estas clases.
Una estrategia es el *algoritmo de enlace completo* que se describe:

- Asigne los documentos a diferentes clases
- Calcule la similaridad entre pares de clases
- Determine el par de clases $[C_u,C_v]$ con la mayor similaridad inter-clusters.
- Mezcle los clusters C_u y C_v.
- Verifique un criterio de parada sino vuelva a el segundo paso.

La similaridad entre clusters es definida como la mínima de las ismilaridad entre pares de documentos inter-cluster. La similaridad entre documentos como el coseno de la formula del modelo vectorial.

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 de documento inversa para cualquier término seleccionado para participar el el tesauro de clases.

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

$$wt_c = \frac{\Sigma_{i=1}^{|c|} w_{i,C}}{|C|}$$

donde |C| es el número de términos en la clase tesauro C y $w_{i,C}$ es 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:

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