UNIVERSIDAD DE CONCEPCION
FACULTAD DE INGENIERIA
DEPARTAMENTO DE INGENIERIA INFORMÁTICA Y CIENCIAS DE LA COMPUTACION

Tarea N° 1
Consideraciones Generales en Reformulación y Expansión de Consulta

Cecilia Labraña Cabrera
Manuel Novoa Olivares

En la expansión automática el sistema responsable por la expansión inicial o subsecuente de la consulta utilizando algún método definido.

Se debe considerar:

- La retroalimentación de relevancia.
- Los términos de la consulta inicial
- Juicios de relevancia con respecto a términos.

Consideremos cada uno de estos puntos.

a) Retroalimentación de relevancia

Después de la generación inicial de la lista de documentos, el sistema consulta por la relevancia o no relevancia del documento como condición binaria. Con esta respuesta por parte del usuario el sistema se retroalimenta para generar un nuevo listado de documentos.

Los documentos seleccionados sirven para modificar la consulta recalculando o reasignando los pesos de los términos de la consulta y/o agregando términos que orientan de mejor forma la consulta y eliminando aquellos que no lo hacen.

Existen dos forma de retroalimentación de relevancia:

- Basada en un documento único el sistema busca documentos similares al ofrecido.
- El sistema busca los términos del documento relevante agregando en la búsqueda los sinónimos de esos términos.

Se definen cuatro métodos de selección de términos para reformulación y expansión de consulta:

- Utiliza los términos de la consulta original para usarlos en una nueva consulta.
- Utiliza los términos de la consulta original y agrega nuevos términos de otra fuente; por ejemplo, los términos adyacentes en el árbol de expansión máxima o los vecinos más cercanos.
- Es una mezcla de métodos, ya que usa términos derivados de la consulta original de los documentos entradas y de los juicios de relevancia.
- Se abandonan los términos de la consulta original y se usan sólo los términos encontrados en los documentos que arrojó la primera consulta.

En todos los casos, después de la formulación de la consulta inicial, la única forma de retroalimentación para el usuario son documentos y la elección de esos documentos.

La mayoría de la investigación en retroalimentación de relevancia y expansión de consulta se ha hecho usando expansión de consulta y reasignación de pesos a los términos.

b) Términos de consulta:

En la mayoría de los casos se utilizan los términos iniciales de la consulta y los términos generados por expansión en muchas formas posibles variando la cantidad de términos que se agregan.

c) Juicios de relevancia en línea

Una vez que los términos de la consulta inicial son seleccionados, se realiza la búsqueda; se despliegan los resultados al usuario; y se obtienen juicios de relevancia.

Se asocian 3 preguntas con los juicios de relevancia.

- tamaño de la muestra de documentos relevantes para la retroalimentación y expansión de consulta.

  El sistema, con la ayuda del usuario, trata de definir un modelo de distribución probabilística de documento relevantes o no relevantes. Con esto la probabilidad de relevancia del documento puede ser estimada. No está claro cual es el mejor tamaño de muestra para realizar juicios de relevancia, sólo que se requiere un documento como mínimo. Se piensa también que mientras mayor sea el tamaño de la muestra de documentos, mejor será la estimación.

  En la ausencia de asignación de juicios de relevancia por el usuario, el sistema utilizará los primeros documentos arrojados por la consulta inicial y se considerarán como relevantes. Estos serán utilizados como muestra para recalcular los pesos de los términos de la consulta y para agregar términos a ella.

- En cuales representaciones de documentos debieran basarse los juicios de relevancia

  Es importante considerar que los usuarios asignan relevancia de acuerdo a la forma o tipo del documento.

  Por ejemplo: El juicio de relevancia es distinto si el término aparece en un título, como texto , en el abstract, etc.

  Se propone que los usuarios realicen sus juicios de relevancia observando la representación más completa disponible.

- Afirmación de relevancia

  Al usuario se le pregunta sobre la relevancia del documento, este juicio es binario (Si / No). El usuario debe contestar acerca de la relevancia del documento y no de su utilidad. Esto es necesario para realizar una estimación de la probabilidad de la relevancia.

Es importante hacer la distinción entre relevancia y utilidad como un documento que el usuario puede utilizar.

Algoritmos de rankeo y selección de términos para expansión de consulta.

El objetivo es entregar una lista de documentos en la cual los documentos más útiles estén en el tope de esta. Para hacer esta lista se pueden agregar decisiones heurísticas como por ejemplo que los términos malos sean eliminados de la consulta en vez de darles un bajo peso.

Existe una relación entre la frecuencia de un término en un documento y la relevancia del documento.

- Términos frecuentes, no muy útil
- Términos de frecuencia media , útiles
- No frecuentes, útiles, pero no tanto como los de frecuencia media.
- Términos muy infrecuentes, son útiles en el sentido que cuando se presentan son buenos indicadores de relevancia.

Entonces, un buen algoritmo de rankeo de términos va a ubicar los términos de media frecuencia en el tope de la lista.

Existen criterios cuantitativos y cualitativos para estimar el valor de un término.

Los algoritmos cualitativos se preocupan del valor de un término.

Los algoritmos cuantitativos se preocupan de algunos criterios específicos tales como pruebas de desempeño.

Se puede utilizar la siguiente fórmula para realizar la expansión de consulta, para la asignación de pesos a cada término t del documento relevante.

$$w_t = \log \frac{(r+c)(N-n-R+r+1-c)}{(n-r+c)(R-r+1-c)}$$

donde:
c = n / N
N: número total de documentos en la colección
R: Muestra de documentos relevantes definidos por la retroalimentación del usuario
n: Número de documentos indexados por el término t
r: Número de documentos relevantes (desde la muestra R) asignados al término t.

Con esto se asigna un grado de importancia mayor a los términos de baja frecuencia y los ubica al tope de la lista.

a) Algoritmo WPQ

Se asume que los distintos términos se distribuyen independientemente unos de otros en los documentos no relevantes y lo mismo sucede en los documentos relevantes. En un estado de expansión de consulta, se considera también una independencia estadística entre el término de expansión de consulta y los términos en la formulación de búsqueda previa.

La presencia o ausencia del término de expansión de consulta no afecta la distribución inicial. Luego:

$$WPQ_t = w_t(p_t - q_t)$$

donde:
w_t: es una función de peso
p_t: es la probabilidad de que el término t aparezca en un documento relevante
q_t: es la probabilidad de que el término t aparezca en un documento no relevante

Le quita peso al término que aparece en los no relevantes.

Asigna un peso a aquellos términos que están en los documentos relevantes en vez de la fórmula original. Para decidir la inclusión de un término en una expansión de consulta considera el ranking WPQ_t en vez de W_t. Así la primera fórmula queda como:

$$WPQ_t = \log\frac{(r+.5)(N-n-R+r+.5)}{(n-r+.5)(R-r+.5)} * \left(\frac{r}{R} - \frac{n-r}{N-R}\right)$$

b) Algoritmo Porter

$$Porter = \frac{r}{R} - \frac{n}{N}$$

La fórmula Porter pone más énfasis en los términos que ocurren frecuentemente en el conjunto de documentos relevantes.

c) Algoritmo ENIM

ENIM es un modelo de asignación de pesos que incorpora información de relevancia donde se asume que los términos índices no de distribuyen de manera independientes unos de otros.

$$EMIM = E_{iq} = \sum_{t_i, w_q} \Delta_{iq} P(t_i, w_q) \log\frac{p(t_i, w_q)}{P(t_i)P(w_q)}$$

o más generalizadamente:

$$G_{iq} = \sum_{t_i, w_q} \Delta_{iq} D_{iq} P_{iq}$$

donde:
t_i: indica la presencia (1) o ausencia (0) de un término
w_q: indica que un documento es relevante (1) o no relevante (0)
Δ_{iq}: indica el valor de un término como un discriminador de relevancia; es 1 si $t_i = w_q$ o –1 si $t_i \neq w_q$
D_{iq}: es el grado de involucramiento
P_{iq}: es la contribución probabilística dada por la expresión logarítmica.

d) Rankeo ZOOM

Provee frecuencia automática de análisis de frases, palabras, códigos o combinación de estos en referencias seleccionadas.

e) Algoritmo R-LOHI

Rankea términos de acuerdo a la frecuencia de ocurrencia en un documento relevante en orden descendente y resuelve empates de acuerdo a la frecuencia del término de baja a alta frecuencia.