Tópicos em Recuperação de Informação[1]

Nivio Ziviani

[1]Conjunto de transparências elaborado por Nivio Ziviani e Patrícia Correia

Tipos de Consultas

Existem dois tipos de consultas:

- Lógica (*boolean*)
  - lista de termos que são combinada através dos operadores lógicos AND, OR e NOT.

- Ordenada (*ranked*)
  - envolve uma heurística que é aplicada para medir a similaridade de cada documento com a consulta.

Consulta Lógica

- A resposta à consulta é formada pelos documentos que satisfazem a condição estipulada.

  Ex.: *text* AND *compression* AND *retrieval*

- Busca ampla ⇒ aumenta revocação (*recall*).

- Maioria do documentos são relevantes ⇒ aumenta precisão (*precision*).

  Ex.: (*text* OR *data* OR *image*) AND
  (*compression* OR *compaction* OR *decompression*) AND
  (*archiving* OR *retrieval* OR *storage* OR *indexing*)

  ( ) indicam a ordem das operações.

Consulta Ordenada

- Esforço para obter:

  - medidas de similaridade.
  - estratégias de ordenação.

- Com o objetivo de manter:

  - revocação alta.
  - precisão alta.

1. Técnica simples:

   - contagem do número de termos da consulta que aparecem no documento.

2. Técnica sofisticada (tamanho do documento):

   - atribui um peso numérico a cada termo.
   - medida do cosseno.
   - modelo vetorial.

Consulta Ordenada

3. Muitas outras técnicas...

4. Coleção TREC:

   - 50 consultas.
   - medida do cosseno consegue precisão acima de 40% na média (difícil de conseguir com modelo booleano).

Processamento de Consultas Lógicas

1. Vocabulário é pesquisado para cada termo.
2. Cada entrada do arquivo invertido é recuperada (e decodificada).
3. As entradas são intercaladas (interseção, união ou complemento).
4. Documentos são recuperados de acordo com a lista final.

Tipicamente, consultas com 5 a 10 termos levam cerca de 1 segundo para o processamento das listas.

- Visualização depende do tamanho da resposta.

Tipos de Consultas Lógicas

Conjuntivas:

text AND *compression* AND *retrieval*

- Termos são ordenadas pela freqüência do termo na coleção f_t.

- Lista de documentos candidatos é inicializada a partir do termo t de menor f_t: $C \leftarrow I_t$

Algoritmo para Consultas Conjuntivas

1. **Para cada** termo t da consulta **faça**
   (a) Obtenha t.
   (b) Pesquise no vocabulário.
   (c) Armazene f_t e o endereço de I_t, a entrada do arquivo invertido para t.
2. Identifique o termo t da consulta com o menor f_t.
3. Leia a entrada I_t correspondente do arquivo invertido. Faça $C \leftarrow I_t$, sendo C a lista de $candidatos$.
4. **Para cada** termo t restante **faça**
   (a) Leia a entrada I_t do arquivo invertido.
   (b) **Para cada** $d \in C$ **faça**
   Se $d \notin I_t$ **então**
   $C \leftarrow C - \{d\}$.
   (c) **Se** $|C| = 0$ **então**
   Retorne *não existem respostas*.
5. **Para cada** $d \in C$ **faça**
   (a) Procure o endereço do documento d.
   (b) Recupere o documento d e o apresente ao usuário.

Ordem de Processamento dos Termos

Razões para iniciar C com termo de menor f_t:

1. Minimizar espaço temporário de memória

    - Tamanho de C é não crescente.

    - Maior tamanho é na inicialização.

    - C pode ser reduzido rapidamente, algumas vezes a 0.

Ordem de Processamento dos Termos

Razões para iniciar C com termo de menor f_t:

2. Minimizar tempo de resposta

- Processamento envolve consulta a membros da lista e não intercalação.
  - Intercalação envolvendo $\mid C \mid$ candidatos e f_t apontadores: $\mid C \mid + f_t$ passos.
  - Se $\mid C \mid = 60$ e $f_t = 60000 \Rightarrow 60060$ passos.
  - Como as listas estão ordenadas: busca binária $\Rightarrow$ $\mid C \mid \log 60060$.
  - Para j>2 tamanho da menor lista é O(1).
  - Para j=2 tamanho da menor lista é O($n^{\beta(1-2\beta)} \log n$).

Tipos de Consultas Lógicas

Não-Conjuntivas:

(*text* OR *data* OR *image*) AND
(*compression* OR *compaction*) AND
(*retrieval* OR *indexing* OR *archiving*)

- Cada termo entre parêntesis pode ser processado simultaneamente:
  - candidatos ficam no conjunto se aparecem em qualquer membro do grupo OR.

- C deve ser inicializado com o menor conjunto resultante da união dos membros.

- Aproximação para tamanho de cada conjunto:
  - soma dos f_t de cada termo constituintes ignorando possibilidade de repetição no OR (aproximação pessimista).