



RECUPERACIÓ DE LA INFORMACIÓ

EXERCICIS DEL TEMA 3: Implementació, indexació i cerca

Abans de començar a resoldre els exercicis, comprova si saps respondre aquestes preguntes:

1. Inventat una petita col·lecció de documents (5–6 documents de 5–6 termes cadascun) i dibuixa l'índex invertit que produiria.
 2. Explica per què un índex invertit és adequat per recuperar documents que coincideixin amb una consulta. *per eficiència*
 3. Per què és important mantenir les llistes de *postings* ordenades per *docid*? *per fer les cerques mes llargues*
 4. L'optimització de consultes és un procés per obtenir les millors consultes per a una tasca de recuperació donada. Cert o fals? *per millorar el cost*
 5. Explica què és l'optimització de consultes. Quin és l'efecte en el *recall* i la *precision* d'una consulta?
 6. Quin és el motiu principal per comprimir l'índex? *Per les dades de disc a memoria*
-

Exercici 1

Considereu el següent fragment d'un índex invertit:

the : 3, 5, 7, 10, 12, 15, 21, 25

new : 1, 2, 3, 7, 10, 12, 21, 22

undiscovered : 2, 3, 5, 6, 7, 10, 13, 15, 21, 22

country 1, 3, 5, 7, 8, 10, 15, 21, 22, 25

que vol dir que “the” apareix en els documents 3, 5, 7, 10, etc.

1. Quins docs satisfan la consulta següent?
the AND new AND undiscovered AND country
2. Quins docs satisfan la consulta següent?
the OR new OR undiscovered OR country
3. Quins docs satisfan la consulta següent?
new AND country BUTNOT undiscovered



Exercici 2

A continuació tenim un índex més “ric” que conté informació de posició per poder respondre consultes per frase. Cada llista de *postings* conté una llista de llistes. Cada llista comença amb un identificador de document seguit d’una llista amb les posicions on el terme apareix dins el document:

the

3: 34,38,55;
5: 12,16,25,44;
7: 67,87,90,101;
10: 33,39,45,62;

undiscovered

3: 12,15,19;
5: 3,5,17,41,45,96;
6: 21,25,55,62;
7: 4,68,70,85,110;
10 :15,34,40,65,81;

country

3: 22,26;
5: 18,46,52,65;
7: 5,69,91,105;
8: 32,42,65,93;
10: 32,44,75,83;

1. Quantes vegades trobarem la frase “**the undiscovered country**” en cada document?
2. Quins documents satisfan la consulta **undiscovered AND country**?
3. L’operador NEAR és útil per consultes per frase aproximades. Un document satisfà a **NEAR b WITHIN d** si en algun lloc conté **a** i també **b** a una distància de com a molt **d-1** paraules entre ambdues. Quins documents satisfan la consulta **undiscovered NEAR country WITHIN 3**?



Exercici 3

Tornem a l'Exercici 5 del Tema 2.

1. Quin ordre de processament recomanaríeu per la consulta següent?
`computer AND client AND applications`
2. Recomaneu un pla de processament per la consulta booleana
`(computing AND programs) OR (p2p AND applications)`
`OR (computing AND networks AND applications)`

Exercici 4

Tenim indexats un conjunt de 10^7 documents. Sabent que els termes A, B, C i D apareixen, respectivament, en 2 milions, 1 milió, 800 000 i 20 000 documents, proposeu un pla d'avaluació eficient per la consulta booleana

`(A and B and C) or (A and B and D)`
`or (A and C and D) or (B and C and D)`

Expresseu el vostre pla com una seqüència d'interseccions de llistes i unions de llistes. Justifiqueu la vostra resposta. No cal que calculeu el cost estimat del vostre pla.

Exercici 5

Suposem que tenim una col·lecció de 500 000 documents amb una mitjana de 800 paraules per document. S'ha estimat que la quantitat de paraules diferents és de 700 000. Indiqueu els càlculs que feu per respondre a les qüestions següents:

1. Quina mida (en MB o GB) ocupa la col·lecció a disc (sense comprimir)?
2. Suposeu que apliquem diversos processos lingüístics (*stemming*, *stopword removal*) als documents i aconseguim una taxa de reducció del diccionari del 50%. Quina mida (nombre de termes) té el diccionari?
3. Considereu un índex en el que la longitud mitjana de la *posting list* sense posicions, és de 200. Doneu una estimació de la quantitat total de *postings* de l'índex sabent que hem aplicat 2.
4. Quants bytes necessitaríeu per codificar (sense compressió) un terme del diccionari? I un *posting* sense posicions?
5. Quina mida (en MB o GB) té el diccionari resultant i les *posting lists*?



Exercici 6

Tenim una col·lecció de 10^8 documents. La longitud mitjana d'un document és de 10000 caràcters i la longitud mitjana d'una paraula en els documents és de 7 caràcters.

1. Supposeu que la col·lecció satisfà la llei de Heaps en la forma $10 \times N^{0.5}$. Estimeu el nombre de paraules diferents que esperaríeu trobar en la col·lecció.
2. Creem un índex invertit que només conté els *docid*. Estimeu la longitud mitjana de les *posting lists*. Pista: Primer estimeu el nombre de paraules diferents per document, després el nombre total d'entrades en les *posting lists* i, a continuació, calculeu el que se us demana.