CAIM, examen parcial

4 de novembre de 2021. Temps: 1 hora 30 minuts

Exercici 1 (1 punt)

Les llengües satisfan una sèrie de lleis estadístiques (llei de Zipf, llei de Herdan/Heaps,...). Esmenta aplicacions en el context de l'assignatura de la llei de Zipf més enllà de purament descriure com són les llengües. Aquestes aplicacions han de demostrar-ne la seva utilitat i basar-se en el temari de l'assignatura fins a l'examen.

Resposta: Segons Luhn (1958) les paraules de freqüència intermitja (un cop ordenades per decreixentment per freqüència) són les més representatives de la temàtica d'un text. La llei de Zipf també és a la base del pesos tf-idf concebuts per Spärk-Jones (1972).

En el context de la compressió de *posting lists*, la llei de Zipf també el motiu per escollir una codificació unària autodelimitant per a les freqüències dels termes. Aplicant la llei de Zipf es pot predir que les freqüències de les paraules en les *posting lists* ocuparan al volant de 2 bits en mitjana aproximadament.

(Algunes referències per als interessats en saber-ne més

1. Luhn, H. (1958). The automatic creation of literature abstracts. IBM Journal of Research and Development 2, 159–165
2. Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 28, 11–21.

)

Exercici 2 (3.5 punts)

En el marc del model booleà, volem analitzar els plans d'execució (altrament coneguts amb el nom de plans d'avaluació) per a la consulta

$$a\ AND\ b\ AND\ c\ AND\ d$$

on a, b, c i d són termes diferents. La següent taula dóna la mida de les *posting lists* de cadascun dels termes:

terme	mida
a	100000
b	30000
c	5000
d	60000

Us demanem de respondre una sèrie de qüestions assumint que les *posting list* estan implementades amb llistes i que les interseccions es calculen de forma simple, és a dir, mitjançant un algorisme de fusió seqüencial de llistes (*merge*) i sense *skip pointers* ni paral·lelització. En concret, volem saber

1. Quants plans d'execució diferents hi ha?
2. Mostra un pla d'execució òptim en temps (cas pitjor) parentitzant adientment l'expressió original.
3. Dóna una estimació del cost temporal (cas pitjor) del pla proposat en l'apartat anterior.
4. Explica perquè el pla d'execució mostrat en l'apartat anterior seria òptim.
5. Dedueix una fórmula per al cost total d'execució temporal (cas pitjor) per al pla d'avaluació següent

$$((t_1\ AND\ t_2)\ AND\ t_3)\ AND\ t_4$$

on t_1,...,t_4 són termes diferents amb *posting lists* $L_1, ..., L_4$ respectivament. En la teva resposta usa $|L_i|$ per indicar la mida de la llista L_i. **Pots fer servir la resposta a aquesta pregunta per a respondre les anteriors.**

Raona la teva resposta a les preguntes anteriors.

Resposta:

1. Hi ha dos tipus de plans d'execució possibles. El primer és de la forma

$$((t_1\ AND\ t_2)\ AND\ t_3)\ AND\ t_4$$

Cal executar tres ANDs. Per a la primera AND tenim $\binom{4}{2}$ aparellaments possibles de termes. Per a la segona AND ens queden 2 termes possibles. Per la tercera AND només queda un candidat (no cal escollir). Per tant, hi ha

$$\binom{4}{2} \cdot 2 \cdot 1 = 12$$

plans d'execució possibles. El segon tipus és de la forma

$$((t_1\ AND\ t_2)\ AND\ (t_3\ AND\ t_4)$$

Per a la primera AND tenim $\binom{4}{2}$ possibilitats. Un cops triats els dos termes per a la primera AND, els termes per a la tercera AND ja estan determinats (no cal escollir). Per tant, hi hauria

$$\binom{4}{2} \cdot 1 = 6$$

plans d'execució possibles però adonant-nos que que hi ha plans d'execució simètrics respecte la darrera AND que s'executa, resulten $6/2 = 3$. En total, $12 + 3 = 18$ plans d'execució ($12 + 6 = 18$ sense tenir en compte simetries).

2. Per exemple, ((b AND c) AND d) AND a (qualsevol pla d'execució del primer tipus que processi la llista més curta en la primera AND serà òptim).

3. Segons la resposta al darrer apartat el cost seria $195000+2*5000 = 205000$.

4. Per respondre aquest apartat cal tenir en compte els dos tipus de plans d'execució. Per al primer tipus tenim en compte la resposta del darrer apartat, que correspon al primer tipus de plans d'execució. Segons la resposta al darrer apartat el pla es òptim entre els plans d'execució del primer tipus perquè la *posting list* més curta (la del terme c) es processa en la primera AND que s'executa. La fórmula del darrer apartat és minimitza quan es processa la llista més curta en la primera AND (en l'exemple, $t_1 = c$ o $t_2 = c$). Per determinar si és l'òptim entre tots els tipus de plans, ens basem en la fórmula per al cost del segon tipus de plans d'execució. Aquesta s'obté considerant el cost temporal (cas pitjor) de la primera AND, la tercera i la segona AND (en ordre d'esquerra a dreta) que dóna

$$\begin{aligned} &|L_1| + |L_2| \\ &|L_3| + |L_4| \\ &\min(|L_1|, |L_2|) + \min(|L_3|, |L_4|). \end{aligned}$$

   Sumant tots els costos obtenim el costa total

$$\sum_{i=1}^{4} |L_i| + \min_{1 \le i \le 2} |L_i| + \min_{3 \le i \le 4} |L_i| \tag{1}$$

   Aquesta formula dóna que el millor pla d'execució del segon tipus tindria cost $195000+5000+30000$, que és pitjor que el que obtenim amb el primer tipus.

5. Els costs temporal (cas pitjor) de cada AND, d'esquerra a dreta, són

$$\begin{aligned} &|L_1| + |L_2| \\ &\min(|L_1|, |L_2|) + |L_3| \\ &\min(|L_1|, |L_2|, |L_3|) + |L_4|. \end{aligned}$$

   Sumant tots els costos obtenim el costa total

$$\sum_{i=1}^{4} |L_i| + \min_{1 \le i \le 2} |L_i| + \min_{1 \le i \le 3} |L_i| \tag{2}$$

Exercici 3 (3 punts)

Suposem que la *posting list* d'un terme que està formada per parells $(docid, f)$ on $docid$ és un identificador de document i f és la freqüència del terme en el document. Per exemple, la *posting list*

$$(3, 1), (8, 5)$$

indica que el terme apareix 1 cop al document 3 y 5 cops al document 8.

Hem comprimit una *posting list* i hem obtingut la següent tirallonga de bits

$$00010110101000011001000011100000110001111111110$$

Descodifica la tirallonga per obtenir la *posting list* original suposant que

1. Les freqüències han estat codificades usant codis unaris autodelimitants (*unary self-delimiting codes*) acabats en zero.
2. Els identificadors de documents s'han codificat usant *gap compression* i codis Elias γ (un codi Elias γ té dues parts; la primera part està formada per una seqüència de zeros que indica la llargada de la segona part del codi).

La vostra resposta ha de mostrar la seqüència de passos que van transformant la tirallonga de bits fins a obtenir la *posting list* original.

Resposta: Segmentant la seqüència, obtenim els parells en base 2

$$(0001011, 0), (1, 0), (1, 0), (0001100, 10), (0001110, 0), (0001100, 0), (1, 11111110)$$

que corresponen a la llista de parells en base 10

$$(11, 1), (1, 1), (1, 1), (12, 2), (14, 1), (12, 1), (1, 8)$$

Desfent *gap compression*, arribem a la *posting list* original, que és

$$(11, 1), (12, 1), (13, 1), (25, 2), (39, 1), (51, 1), (52, 8).$$

Exercici 4 (2.5 punts)

A partir de la teoria del càlcul dels pesos de PageRank, deduïu l'equació

$$\vec{p} = (I - \lambda M^T)^{-1} \begin{pmatrix} \frac{1-\lambda}{n} \\ . \\ . \\ . \\ \frac{1-\lambda}{n} \end{pmatrix}$$

on n és el nombre de vèrtexs del graf, λ és el factor d'esmorteïment, M és la matriu d'adjacència normalitzada a nivell de rengles i I és la matriu identitat.

Resposta: Segons la definició del PageRank,

$$\vec{p} = G^T \vec{p}$$
$$\sum_{i=1}^{n} p_i = 1,$$

on

$$G = \lambda M + \frac{1-\lambda}{n} J.$$

Per tant,

$$\begin{aligned} \vec{p} &= \left(\lambda M + \frac{1-\lambda}{n} J\right)^T \vec{p} \\ &= \left(\lambda M^T + \frac{1-\lambda}{n} J\right) \vec{p}. \end{aligned}$$

A partir d'aquí, movent λM^T a l'esquerra de la igualtat, obtenim

$$\vec{p} - \lambda M^T \vec{p} = \frac{1-\lambda}{n} J\vec{p}$$
$$(I - \lambda M^T)\vec{p} = \frac{1-\lambda}{n} J\vec{p}.$$

Aplicant $\sum_{i=1}^{n} p_i = 1$ obtenim

$$(I - \lambda M^T)\vec{p} = \begin{pmatrix} \frac{1-\lambda}{n} \\ . \\ . \\ . \\ \frac{1-\lambda}{n} \end{pmatrix}.$$

Assumint que $I - \lambda M^T$ és una matriu invertible, obtenim finalment el resultat esperat.