Trường Đại Học Bách Khoa Hà Nội
Viện CNTT & TT

Đề thi cuối kỳ môn học IT4853 – Tìm kiếm và trình diễn thông tin

(*Trọng số 0.7, đề thi gồm 2 trang, thời gian làm bài 120 phút, không được sử dụng tài liệu*)

Bài 1 – Cấu trúc dữ liệu chỉ mục ngược (1.0 điểm)

Cho bộ dữ liệu văn bản sau:

Doc 1: Đại học bách khoa Hà Nội
Doc 2: Bách khoa toàn thư khoa học và công nghệ
Doc 3: Đại từ điển bách khoa tòan thư

Hãy minh họa bằng hình vẽ cấu trúc chỉ mục ngược đơn giản gồm từ điển và bộ thẻ định vị. Các từ trong từ điển phải được sắp xếp tăng dần theo thứ tự bảng chữ cái, danh sách thẻ định vị cũng phải được sắp xếp theo thứ tự tăng dần mã văn bản.

Điều kiện: tách từ theo khoảng trắng; đổi tất cả chữ hoa thành chữ thường; không cần lưu bất kỳ dữ liệu nào khác ngoài từ tách được và mã văn bản; các ký tự là những ký tự Unicode dựng sẵn theo chuẩn TCVN 6909:2001; thứ tự bảng chữ cái của các ký tự đã được sử dụng như sau:

dấu_cách, a, b, c, g, h, i, k, n, o, t, v, à, á, ò, ô, đ, ư, ạ, ể, ệ, ọ, ộ, ừ

Bài 2 – Ước lượng thời gian thực hiện giải thuật sắp xếp (1.0 điểm)

Giả sử chúng ta cần $T\log_2 T$ so sánh (ví dụ, QuickSort) để sắp xếp T cặp *mã từ-mã văn bản*. Hãy ước lượng thời gian thực hiện giải thuật sắp xếp (công thức tổng quát và kết quả cụ thể tính bằng giây) trong hai trường hợp: lưu toàn bộ dữ liệu trên ổ đĩa và trong bộ nhớ. Một cách đơn giản, chúng ta giả sử rằng nếu lưu dữ liệu trên đĩa thì cần hai thao tác định vị đầu đọc và một thao tác ALU để thực hiện một phép so sánh, còn nếu sử dụng bộ nhớ thì cần hai thao tác đọc cặp *mã từ-mã văn bản* trong bộ nhớ và một thao tác ALU. Các tham số hệ thống được cho trong bảng sau, ($T = 10^6$):

Ký hiệu	*Tham số*	*Giá trị*
m	Thời gian đọc cặp *mã từ-mã văn bản* trong bộ nhớ	5E-9 s
s	Thời gian định vị đầu đọc của ổ đĩa	5E-3 s
p	Thời gian thực hiện thao tác ALU	1E-9 s

Bài 3 – Chỉ mục ngược có vị trí, truy vấn với tham số khoảng cách (1.5 điểm)

Cho chỉ mục ngược có vị trí theo định dạng sau:

từ: mã-văn-bản: <vị trí, vị trí, …>; mã-văn-bản: <vị trí, …>.

Chỉ mục ngược:

Tìm-kiếm: 1: <1>; 2: <6>; 3: <2, 15>; 4: <1>.
Dữ-liệu: 1: <3>; 3:<4, 16>; 4: <3>; 7: <14>;
Thông-tin: 1: <2>; 2: <12, 16, 21>; 3: <18>; 5: <21, 25>.

Tham số /k trong truy vấn **từ1 /k từ2** được hiểu là tìm **từ2** trong phạm vi **k** từ so với **từ1** (có tính đến thứ tự), trong đó k là số nguyên dương. Như vậy nếu k = 1 thì **từ2** là từ liền sau **từ1**.

Hãy xác định: a) Tập văn bản thỏa mãn truy vấn: **Tìm-kiếm /2 Dữ-liệu**

b) Tập giá trị k sao cho truy vấn: **Tìm-kiếm /k Thông-tin** trả về tập kết quả {1, 3}.

c) Tập giá trị k sao cho truy vấn **Thông-tin /k Thông-tin** trả về tập kết quả khác rỗng.

Bài 4 – Mô hình tìm kiếm thông tin, mô hình không gian vec-tơ (1.0 điểm)

Sử dụng chỉ mục ngược đã cho ở bài 3. Hãy tính độ tương đồng cosine với truy vấn *Tìm-kiếm Thông-tin* (gồm hai từ *Tìm-kiếm* và *Thông-tin*) cho ba văn bản với mã số 1, 2, 3 và sắp xếp các văn bản này theo thứ tự giảm dần độ tương đồng cosine. Sử dụng phương pháp xác định trọng số từ tf.idf theo cấu trúc lnc.ltc.

** Gợi ý giải mã ký hiệu SMART:* bộ ba ký tự đầu tiên áp dụng cho văn bản, bộ ba ký tự tiếp theo áp dụng cho câu truy vấn, các ký tự theo thứ tự áp dụng cho tf, df, và chuẩn hóa. Ý nghĩa các ký hiệu như sau:

l (logarithm): $1 + \log(tf)$ n (no): df = 1 t (idf): $\log(N/df)$

c (cosine): $1/\sqrt{w_1^2 + w_2^2 + ... + w_M^2}$ trong đó M là số từ trong từ điển.

Bài 5 – Đánh giá kết quả tìm kiếm **(1.5 điểm)**

Giả sử có 3 văn bản phù hợp với nhu cầu thông tin thứ nhất và 5 văn bản phù hợp với nhu cầu thông tin thứ 2 trong bộ dữ liệu. Kết quả đánh giá tính phù hợp cho 10 văn bản đầu tiên được trả về như sau (ký tự bên trái nhất đại diện cho kết quả đầu tiên được trả về, R – phù hợp, N – không phù hợp):

Nhu cầu thông tin 1:

Hệ thống 1: RNRNNNNRNN

Hệ thống 2: NRNNNRNRNN

Nhu cầu thông tin 2:

Hệ thống 1: NRNRNNRRNR

Hệ thống 2: RRNRNNRNNR

Hãy so sánh hai hệ thống dựa trên những dữ liệu đã cho:

a) Hãy tính MAP của hai hệ thống? Các giá trị MAP thu được cho thấy hệ thống nào ưu việt hơn?

b) Hãy tính các giá trị F1 cho từng tập kết quả trả về? Xác định giá trị trung bình của độ đo F1 cho mỗi hệ thống? Dựa trên các giá trị thu được hãy đưa ra kết luận hệ thống nào ưu việt hơn trong từng trường hợp (truy vấn 1, truy vấn 2, trường hợp tổng quát)?

c) Trong danh sách có thứ tự của kết quả trả về, chúng ta định nghĩa điểm cân bằng là điểm có độ chính xác bằng độ đầy đủ. Hãy đưa ra điều kiện để tồn tại một điểm như vậy và giải thích?

* *Gợi ý:* MAP được định nghĩa như sau: Nếu tập văn bản phù hợp cho nhu cầu thông tin $q_i \in Q$ có dạng $\{d_1, d_2, ..., d_{m_i}\}$ và R_{ik} là tập kết quả có xếp hạng từ kết quả đầu tiên theo thứ tự tới văn bản d_k, thì

$$MAP(Q) = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{m_i}\sum_{k=1}^{m_i}P(R_{ik})$$, trong đó $P(R_{ik})$ là độ chính xác trên tập R_{ik}.

F1 (trung bình điều hòa của P và R) được xác định theo công thức sau F1 = 2PR/(P + R).

Bài 6 – Nén danh sách mã số văn bản **(1.0 điểm)**

a) Giả sử có một danh sách mã số văn bản đã được chuyển thành danh sách khoảng cách và được mã hóa với số bytes thay đổi (VB), kết quả mã hóa như sau (dữ liệu được cho dưới dạng mã nhị phân):

10000101 00000011 10000001 10001001

Hãy giải mã VB đã cho để lấy danh sách mã số văn bản ban đầu.

b) Hãy mã hóa danh sách khoảng cách của danh sách mã số văn bản đó sử dụng mã gamma (γ-code).

Bài 7 – Lưu từ điển **(1.5 điểm)**

Xét một bộ từ vựng (từ điển) trong đó không có từ độ dài 1 hoặc 2 ký tự. Giả sử rằng số từ có độ dài *i* tỉ lệ thuận với $1/i^2$, với $i > 2$ và độ dài cực đại của từ là 30. Ngoài ra bộ từ vựng có M = 100000 từ và sử dụng 1 byte để biểu diễn 1 ký tự.

a) Hãy viết công thức tổng quát xác định số lượng ký tự cần sử dụng để viết tất cả từ có độ dài *i*.

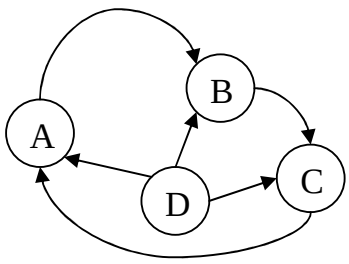
b) Nếu chúng ta lưu tất cả từ như một chuỗi ký tự dài và con trỏ tới ký tự bắt đầu của mỗi từ, thì sẽ cần bao nhiêu bytes? (viết công thức tổng quát và kết quả cuối cùng, giả sử mỗi con trỏ chiếm 4 bytes).

c) Nếu sử dụng phân đoạn: lưu con trỏ tới ký tự đầu tiên của mỗi khối mười từ liên tiếp, sử dụng một byte đặt trước ký tự đầu tiên của mỗi từ để lưu độ dài của từ đó. Hãy tính số bytes cần sử dụng trong trường hợp này? (công thức tổng quát và kết quả cuối cùng).

* Gợi ý: Có thể làm tròn tổng $1 + 1/2^2 + 1/3^2 + \ldots + 1/30^2 = \pi^2/6$; $\pi^2/6 \approx 1{,}65$.

Bài 8 – Phân tích liên kết **(1.5 điểm)**

Cho một đồ thị web nhỏ với bốn trang A, B, C, D như trên hình vẽ. Hãy tính PageRank, điểm giới thiệu (hub) và điểm uy tín (authority) cho mỗi trang. Đồng thời hãy xác định thứ tự xếp hạng các trang theo những tiêu chí này.



PageRank:

Giả sử tại mỗi bước di chuyển ngẫu nhiên, với xác suất 0.2 chúng ta sẽ nhảy tới một trang bất kỳ, với xác suất còn lại chúng ta sẽ di chuyển theo liên kết. Xác suất bước nhảy tới mỗi trang là bằng nhau.

Điểm giới thiệu/điểm uy tín:

Hãy chuẩn hóa các giá trị điểm giới thiệu và điểm uy tín sao cho giá trị cực đại bằng 1.

Đáp án

Bài 1 – Cấu trúc dữ liệu chỉ mục ngược **(1.0 điểm)**

bách	→	1	→	2	→	3
công	→	2				
hà	→	1				
học	→	1	→	2		
khoa	→	1	→	2	→	3
nghệ	→	2				
nội	→	1				
thư	→	2	→	3		
toàn	→	2				
tòan	→	3				
từ	→	3				
và	→	2				
điển	→	3				
đại	→	1	→	3		

Các mã văn bản có thể đặt tùy ý đủ để hiểu là mã của các văn bản đã cho. Kết quả đúng hoàn toàn đạt 1 điểm. Đúng cấu trúc chỉ mục ngược nhưng sắp xếp sai đạt 0,5 điểm. Nếu không phân biệt *toàn* và *tòan* thì chỉ đạt một nửa số điểm của hai trường hợp vừa nêu.

Bài 2 – Ước lượng thời gian thực hiện giải thuật sắp xếp **(1.0 điểm)**

Trường hợp lưu toàn bộ dữ liệu trên ổ đĩa: [0,5]

$T * \log_2 T * (2 * s + p) = 10^6 * 6 * \log_2 10 * (2 * 5 * 10^{-3} + 10^{-9}) = 199315.7$ (s)

Thời gian thực hiện giải thuật sắp xếp là 199315.7 (s) (khoảng 2 ngày 7 giờ)

Trường hợp lưu toàn bộ dữ liệu trong bộ nhớ [0,5]

$T * \log_2 T * (2 * m + p) = 10^6 * 6 * \log_2 10 * (2 * 5 * 10^{-9} + 10^{-9}) = 0.22$ (s).

Thời gian thực hiện giải thuật sắp xếp là 0.22(s)

Bài 3 – Chỉ mục ngược có vị trí, truy vấn với tham số khoảng cách **(1.5 điểm)**

Điều kiện để văn bản d thỏa mãn truy vấn dạng T1 /k T2 là tồn tại ít nhất 1 cặp giá trị post1 và post2 (vị trí xuất hiện của T1 và T2) sao cho post2 – post1 >= k.

a) Tập văn bản thỏa mãn truy vấn **Tìm-kiếm /2 Dữ-liệu** là: {1, 3, 4} [0,5]

b) Tập giá trị k sao cho truy vấn **Tìm-kiếm /k Thông-tin** trả về tập kết quả {1, 3} là: {3, 4, 5}. Nếu k > 5 thì tập kết quả trả về là {1, 2, 3} [0,5]

c) Để truy vấn **Thông-tin /k Thông-tin** trả về tập kết quả khác rỗng thì k>=4, hay tập giá trị của k là: {4, 5, …, +Inf} [0,5]

Bài 4 – Mô hình tìm kiếm thông tin, mô hình không gian vec-tơ **(1.0 điểm)**

Từ chỉ mục ngược suy ra N = 6, các mã số văn bản là 1, 2, 3, 4, 5, 7.

Thống kê:

	Truy vấn		D1		D2	D3	
	tf	df	tf		tf	tf	
Tìm-kiếm	1	4	1		1	2	
Thông-tin	1	4	1		3	1	
Dữ liệu	0	4	1		0	2	

lnc.ltc:

	Truy vấn	D1	D2	D3
	(1+log tf) * log(N/df)	1 + log tf	1 + log tf	1 + log tf
Tìm-kiếm	0,18	1,00	1,00	1,30
Thông-tin	0,18	1,00	1,48	1,00
Dữ liệu	0,00	1,00	0,00	1,30
c-norm	4,02	0,58	0,56	0,48

Biểu diễn vec-tơ và độ tương đồng cosine

	Truy vấn	D1	D2	D3
Tìm-kiếm	0,71	0,58	0,56	0,62
Thông-tin	0,71	0,58	0,83	0,48
Dữ liệu	0,00	0,58	0,00	0,62
cosine		0,82	0,98	0,78

Thứ tự xếp hạng của ba văn bản D1, D2, D3 theo truy vấn là: D2, D1, D3

* Đúng biểu diễn vec-tơ đạt 0,5. Đúng độ tương đồng cosine và sắp xếp đúng các văn bản đạt 0,5.

Bài 5 – Đánh giá kết quả tìm kiếm **(1.5 điểm)**

a) Tính MAP và so sánh các hệ thống [0,5]

$AP_{s1q1} = (1/1 + 2/3 + 3/8) / 3 = 0,68056$ $AP_{s2q1} = (1/2 + 2/6 + 3/8) / 3 = 0,40278$

$AP_{s1q2} = (1/2 + 2/4 + 3/7 + 4/8 + 5/10) / 5 = 0,48571$ $AP_{s2q2} = (1/1 + 2/2 + 3/4 + 4/7 + 5/10) / 5 = 0,76429$

$MAP_{s1} = (0,68056 + 0,48571) / 2 = 0,58313$ $MAP_{s2} = (0,40278 + 0,76429)/ 2 = 0,58353$

$MAP_{s2} > MAP_{s1}$ cho thấy hệ thống 2 ưu việt hơn hệ thống 1 dù chênh lệch điểm là không lớn.

b) Tính F1 và so sánh các hệ thống [0,5]

$P_{s1q1} = P_{s2q1} = 3/10 = 0,3;$ $P_{s1q2} = P_{s2q2} = 5/10 = 0,5;$

$R_{s1q1} = R_{s2q1} = R_{s1q2} = R_{s2q2} = 1;$

$F_{s1q1} = F_{s2q1} = 2 * 0,3 * 1 / (1 + 0,3) = 0,46$ $F_{s1q2} = F_{s2q2} = 2 * 0,5 * 1 / (1 + 0,5) = 0,67$

Các giá trị trung bình:

$Avg(F1_{s1}) = Avg(F1_{s2}) = (0,46 + 0,67) / 2 = 0,57$

Vì $F1_{s1} = F1_{s2}$ trong mọi trường hợp, nên có thể kết luận là độ đo F1 xếp hạng cả hai hệ thống như nhau trong trường hợp này. (*Lưu ý: P, R và F1 là các độ đo áp dụng cho tập kết quả không xếp hạng. Cả hai hệ thống đều trả về tất cả kết quả phù hợp trong 10 văn bản đầu tiên*)

c) Điểm cân bằng [0,5]

Theo như định nghĩa, K là điểm cân bằng khi và chỉ khi P@K = R@K. Giả sử trong bộ dữ liệu có R văn bản phù hợp, và trong K kết quả đầu tiên trả về có r kết quả phù hợp. Chúng ta có:

$$P@K = R@K \Leftrightarrow r/K = r/R$$

Nếu R = 0 thì không tồn tại điểm cân bằng.

Nếu R > 0, điểm cân bằng như trong định nghĩa tồn tại trong các trường hợp sau:

K = R ⇔ số văn bản trả về lớn hơn hoặc bằng số văn bản phù hợp trong bộ dữ liệu (trường hợp cơ bản).

r = 0 ⇔ văn bản được trả về đầu tiên không phù hợp.

*5a, 5b: Tính đúng các giá trị đạt 0,25; so sánh đúng đạt 0,25.

*5c: Đạt 0,5 nếu chỉ ra được trường hợp cơ bản.

Bài 6 – Nén danh sách mã số văn bản **(1.0 điểm)**

a) Danh sách khoảng cách: [0,5]

mã nhị phân:	101	110000001	1001
hệ thập phân:	5	385	9
Danh sách mã số văn bản:	5	390	399

b) Mã gamma của danh sách khoảng cách

giá trị	**Xóa 1 bít trái**	**Độ dài**
101	01	110
110000001	10000001	111111110
1001	001	1110001

Kết qủa: 11001 11111111010000001 1110001

* 6a, xác định đúng danh sách khoảng cách đạt 0,25; xác định đúng danh sách mã số văn bản đạt 0,25.

Bài 7 – Lưu từ điển **(1.5 điểm)**

a) Số lượng từ có độ dài *i* là C/i^2, trong đó C là hằng số. [0,5]

$$C\sum_{i=3}^{30}\frac{1}{i^2} = M \rightarrow C = M\ /(1{,}65 - 1{,}25) \rightarrow \text{C} = 2{,}5 * \text{M}$$

Số ký tự cần thiết để viết tất cả từ có độ dài i là: 2,5 * M/i

b) Số bytes cần sử dụng để lưu tất cả con trỏ là 4 * M [0,5]

Số bytes cần sử dụng để lưu các ký tự là 2,5 * M * (1/3 + 1/4 + … + 1/30) ≈ 6,23747*M
Tổng số bytes cần sử dụng là (4 + 6,23747) * M ≈ 1023747 bytes.
c) Số phân đoạn là M/10 [0,5]
Số bytes cần sử dụng để lưu con trỏ là 4 * M / 10 = 0,4 * M
Số bytes cần sử dụng để lưu độ dài là M
Số bytes để lưu các ký tự không thay đổi và bằng 6,23747 * M
Tổng số bytes cần sử dụng trong trường hợp này là : (6,23747 + 1 + 0,4) * M ≈ 763747 bytes.
* Để tính tổng 1/3 + 1/4 + ... + 1/30 có thể sử dụng công thức làm tròn: ln(n) + 0.5772156649 – 1,5 hoặc tính theo cách thông thường.

Bài 8 – Phân tích liên kết (1.5 điểm)

PageRank: [0,5]

Phương pháp: Tính ma trận xác suất chuyển rồi sử dụng phương pháp lũy thừa hoặc giải hệ phương trình.

Tuy nhiên có thể sử dụng tính chất đối xứng như sau (cách thông thường trình bày ở bên dưới):

Trang D không có liên kết đi vào, do đó

D = 0.2 * 1/4 = 0.05

Các trang A, B, C có cấu trúc liên kết như nhau, do đó

A = B = C = (1 – 0.05)/3 = 0.316

Thứ tự các trang theo PageRank là: A B C D

Điểm uy tín: [0,5]

Trang D không có liên kết đi vào, do đó điểm uy tín D = 0

Các giá trị chuẩn hóa của A, B, C là A = B = C = 1

Thứ tự các trang theo điểm uy tín là: A B C D

Điểm giới thiệu: [0,5]

Sử dụng các điểm uy tín như trên chúng ta có các điểm giới thiệu như sau:

D = A + B + C → D = 3 A = B = C = 1

Các giá trị điểm giới thiệu sau chuẩn hóa:

A = B = C = 0,333 D = 1

Thứ tự các trang theo điểm giới thiệu là: D A B C

* Các trang A, B, C có giá trị tham số như nhau, vì vậy thứ tự tương đối bất kỳ giữa ba trang này đều là những thứ tự đúng. Tính đúng các tham số đạt 0,25. Xác định đúng thứ tự đạt 0,25.

Tính PageRank thông qua hệ phương trình:

Ma trận kề A[4 x 4]:

	A	B	C	D
A	0	1	0	0
B	0	0	1	0
C	1	0	0	0
D	1	1	1	0

Ma trận xác suất chuyển T[4 x 4]:

0	1	0	0		1/4	1/4	1/4	1/4			0,05	0,85	0,05	0,05
0	0	1	0	*0,8 +	1/4	1/4	1/4	1/4	* 0,2	=	0,05	0,05	0,85	0,05
1	0	0	0		1/4	1/4	1/4	1/4			0,85	0,05	0,05	0,05
1/3	1/3	1/3	0		1/4	1/4	1/4	1/4			0,95/3	0,95/3	0,95/3	0,05

Gọi A, B, C, D là các giá trị PageRank của các trang tương ứng. Ở trạng thái ổn định chúng ta có:

[A B C D] * T = [A B C D] (vec-tơ riêng chính trái của T).

Chúng ta có hệ phương trình sau:

{ (1) -0,95A + 0,05B + 0,85C + 0,95D/3 = 0
(2) 0,85A - 0,95B + 0,05C + 0,95D/3 = 0
(3) 0,05A + 0,85B - 0,95C + 0,95D/3 = 0
(4) 0,05A + 0,05B + 0,05C - 0,95D = 0

Thay A + B + C + D = 1 vào (4) ta có D = 0,05

Thay D = 0,05 và nhân mỗi phương trình (1), (2), (3) với 3/0,05 (60) ta được hệ phương trình sau:

{ (1) -57A + 3B + 51C = -0,95
(2) 51A - 57B + 3C = -0,95
(3) 3A + 51B - 57C = -0,95

Giải hệ chúng ta thu được A = B = C = 0,95/3 ≈ 0,32. Thứ tự xếp hạng các trang theo PageRank là A B C D