Information Organization Midterm Examination (November 28th, 2012)

Notice

- Write your name and student number to all the answer sheets.
- You can open printed materials (slides, notes, memos, etc).
- You can use a calculator. But other electronic devices such as laptop PC, tablet, mobile phone, electronic dictionary, are not allowed.
- You can write answers either in English or Japanese.

$\frac{9+3+5+5+11+9}{12+10+20}$ 32

20'

Question 1.

Suppose that the following document IDs are in a postings list.

000 11111 63=

22, 24, 31, 38, 101, 132

40+

16+4+2

Show a compact representation of the above postings list, using bits as few as possible. How many bits are necessary?

$2^4+2^2+2^1$ 10110 4+2+1 64

0111111 64+32+16+8+4

111111

40+22

20'

Question 2.

In information retrieval, it is common to retrieve top-k relevant documents.

1. Explain an algorithm that uses min-heap and finds top-k documents from a collection of N documents.
2. Illustrate how min-heap changes when the following relevance scores of $N = 5$ documents are processed in this order. Here, assume that $k = 3$, and a higher score should be ranked higher.

0.21, 0.33, 0.15, 0.82, 0.64

$\frac{1}{F} = \frac{1}{2}(\frac{4}{3} + \frac{4}{5}$

$\frac{8}{3}$

35'

Question 3.

Suppose that there is a collection of $N = 100000$ documents, with document frequencies (df) of terms

orange, apple, melon, mango

are 10000, 50000, 1000, 100, respectively. There are two documents d_1 and d_2:

d_1: orange melon mango
d_2: apple melon melon

$\frac{1}{F} = \frac{1}{2}(4+1)$

$\frac{1}{F} = \frac{5}{2}$

Now consider the following query q:

q: melon mango

10'

1. Compute the tf-idf score between query q and document d_1. For tf-idf score, use ltn.lnc, namely logarithmic term frequency $(1 + \log(tf_{t,d}))$, idf ($\log N/df_t$) document frequency, no normalization for the query, and logarithmic term frequency, no document frequency, and cosine normalization ~~for the document~~. You can use $\log 2 = 0.30$ for approximation.

10'

2. Compute the tf-idf score between query q and document d_2, in the same manner as (1). Then answer which document is more relevant to q.

$F = $ 10

3. Compute the Jaccard coefficient (a) between q and d_1, and (b) between q and d_2. Which one this result is indicating as more relevant, (a) or (b)?

5'

4. As a relevance measure, what are advantages of tf-idf score over Jaccard coefficient?

4+1

$\frac{4}{3} + \frac{4}{3}$ $\frac{4}{3}$ $\frac{11}{20}$ $\frac{10^5}{20}$ 5×10^6 (log 2) $\frac{8}{20}$ $\frac{15}{20}$

25'

Question 4.

15'

1. Suppose that System A and System B for information retrieval were evaluated on the same document collection, over query to find "ramen recipe". The results were as follows:

```
System A:  Retrieved  80 documents, in which  20 documents were about
ramen recipe.  All the documents about ramen recipe in the collection were retrieved.
```

$p = \frac{20}{80}$ $r = 60$

```
System B:  Retrieved 20 documents, in which 15 documents were about
ramen recipe.
```

$\frac{1}{\sqrt{1+1.3^2}}$

$\frac{4+1}{2}$

$\frac{1}{F}$

10'

Now compare qualities of System A and System B using appropriate measures.

2. In web search, potentially all the web pages in the Internet can be the document collection. Why recall is difficult to measure in this situation? What is a good measure to evaluate quality of retrieved results in web search?

20	60	80
50	a	

0	15	10^5	20 10
0	5	10^3	log

$\frac{1}{\sqrt{3}} = \frac{1}{\sqrt{3}}$

$\frac{\sqrt{3}}{3}$

解　　答　　用　　紙

2012年 11月 28日提出

科目/Subject Introduction to Information Retrieval	4 4 1 2 1 5 4 0 - 7 氏名/Name XU, Shan (ジョ サン)	Score 91
担当教員/Lecturer		

Question 1: There are maybe two ways possible to represente using gap: 22, 2, 7, 7, 63, 31

Variable byte code: 00010110, 00000010, 00000111, 00000111, 00111111, 00011111　48 bits　47

Gamma codes: 111100110, 100, 11011, 11011, 1111011111, 111101111　42 bits.　44

It's obvious that using Gamma codes is better, the representation is shown below:

111100110, 100, 11011, 11011, 1111011111, 111101111

42 bits are necessary.　20

Question 2: 1. Use a binary min heap. Takes O(N logk) operations to construct then read off k winners in O(k logk) steps. A binary min heap is a binary tree in which each node's value is less than the values of its children. To process a new document d' with score s' following the steps:

① Get current minimum hm of heap (O(1))

② If s' ≤ hm skip to next documone　10　34

③ If s' > hm heap-delete-root (O(log k))

④ Heap-add d'/s' (O(log k))

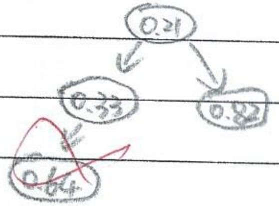
2. As descripted in 1, firstly, find current minimum $h_m = 0.21$.

then read $s' = 0.33 > 0.21$, So add to this tree $(0.21 \rightarrow 0.33)$

then read $s' = 0.15$, minmum $h_m = 0.21$, $s' < h_m$, So skip to next one.

then read $s' = 0.82$ minmum $h_m = 0.21$ $s' > h_m$. So add to this tree ($0.21 \begin{smallmatrix} \nearrow 0.33 \\ \searrow 0.82 \end{smallmatrix}$)

then read $s' = 0.64$ minmum $h_m = 0.21$, $s' > h_m$, and turn to the children note of 0.33, as in the children node, $h_m = 0.33$, $s' > h_m$, So add to the children node of 0.33, it is shown as



−2 ×3

4

Question 3: 1. As the figure show:

10

10

term	query					document				product
	tf	wf	df	idf	$W_{t,q}$	tf	wf	$W_{t,d}$	n'lized	
orange	0	0	10000	1	0	1	1	1	0.577	0
melon	1	1	1000	2	2	1	1	1	0.577	1.15
mango	1	1	100	3	3	1	1	1	0.577	1.73

So the tf-idf score is $0 + 1.15 + 1.73 = 2.88$

次ページへつづく

解　答　用　紙

2012年 11月 28日提出

科目/Subject

前ページのつづき

2 term	query					document				Product
	tf	wf	df	idf	weight	tf	wf	Wt'd	n'lized	
apple	0	0	10000	0.3	0	1	1	1	0.61	0
melon	1	1	1000	2	2	2	1.3	1.3	0.79	1.58
mango	1	1	100	3	3	0	0	0	0	0

10

So tf-idf score is 1.58

As 2.88 > 1.88, So document 1 is more relevan to q.

3. (a) Jaccard (q, d₁) = $\frac{2}{3}$

(b) Jaccard (q, d₂) = $\frac{1}{3}$

10

As Jaccard is an measure of overlap of two sets, So, (a) is more relavant.

25

5

4. Jaccard coefficient doesn't consider term frequency. And Rare terms are more informative than frequent terms, Jaccard does not consider this information. However, tf-idf score both take the term frequency and rare terms more informative into consideration. At the same time, it normalizes for the length of a document.

Question 4 : 1. For the document collection, we can calculate precision, recall, and F_1 to compare.

For SystemA	relevant	not relevant		For systemB	relevant	not relevant	
retrived	20.	60	80	retrived	15	5	20
not retrived.	0			not retrived	5		
	20				20		

15

22

For system A precision $P_1 = \frac{20}{80} = \frac{1}{4}$, recall $R_1 = \frac{20}{20} = 1$, $F_1 = \frac{2}{5}$

system B precision $P_2 = \frac{15}{20} = \frac{3}{4}$, recall $R_2 = \frac{15}{20} = \frac{3}{4}$, $F_2 = \frac{3}{4}$

We can know that B has a higher precision, and lower. recall than A, because it returns less results. and B has a better harmonic mean than A.

2. Recall (R) is the fraction of relevant documents that are retrieved.

7

$$\text{Recall} = \frac{\#(\text{relevant items retrieved})}{\#(\text{relevant items})}$$

−1 ×3

Recall is a non-decreasing function of the number docs retrieved. And users care about the high-ranking documents, like the top 4. So web search is better to use precision at top k or use measures that reward users more for getting rank 1 right than for getting rank 10 right.

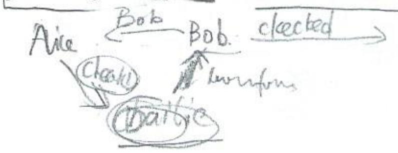
Information Organization Final Examination (January 23th, 2013)

Notice

- Write your name and student number to all the answer sheets.
- You can bring printed materials (slides, notes, memos, etc). But NO electronic device, such as cell phone, laptop, electronic dictionary, or calculator is allowed.
- You can write answers either in English or Japanese.

Question 1. Alice wants to send a document D to Bob securely over the Internet. But Bob wants that D must be checked by Charlie. So Charlie receives D from Alice first. After he checks D, he sends D and his confirmation message to Bob. Bob wants that copies of D received from Alice and Charlie are authentic (truly from them) and identical. Alice wants a message that Bob correctly received D.

Describe a sequence of communications over the Internet which uses public-key encryption and satisfies these requirements.



Question 2. Consider the following transactions T1, T2 and T3:

T1: R(A), R(B), W(A), W(C)
T2: R(B), R(A), W(B)
T3: R(C), W(C)

Show schedules that demonstrate the following notions (1)-(4), by interlacing one or more transactions from T1, T2, and T3. You can add S(), X(), U(), commit, and abort, if necessary. Here, S(A) means a shared-lock on A, X(A) means an exclusive-lock on A, and U(A) means unlocking A. A schedule should be written like:

T3: S(A), T3:X(C), T3: R(C), T1: S(A), T1:R(A), T3:X(C), T3: W(C), T3:U(C), T3: U(A), T3: commit

Also you need to explain why each schedule satisfies its notion.

1. Read-write conflict.
2. Deadlock.
3. Schedule that is conflict-serializable, but causing cascading aborts.
4. Schedule that is view-serializable, but not conflict-serializable.

T1	T2	T3
R(A)		
	R(A)	
	W(A)	
	C	
R(A)		
W(A)		
C		

T1	T2	T3
R(A)	R	
	R(B)	
	W(B)	
		R(C)
~~R(B)~~		
R(A)		

Question 3.
Answer following questions.

1. In what situations SIX lock mode is used? Is IX lock compatible with SIX lock? Justify your answers.
2. Suppose that table emp(EID, salary) holds records of employee IDs (eid) and his/her salary (Salary). Now consider the following transaction T:

   Compute the average salary (AvgSalary) of all the employees. Then for each employee, if he/she has salary below AvgSalary, then increase his/her salary by (AvgSalary - Salary)*0.1.

   In the questions below, justify your answers.

   (a) Describe a lock schedule for T that guarantees conflict serializability.

   (b) Describe a lock schedule for T that may not be conflict-serializable, may produce an inaccurate result, but has more concurrency than (a).

T1

T1	T2	T3
	~~R(B)~~	
S(B)	~~R(A)~~	
R(B)	[illegible]	
~~U(B)~~	X(B)	
	R(A)	
	W(B)	
~~R(A)~~		
R(B)		

Questions continue to the back.

① Alice $\xrightarrow{D}$ Charlie $\xrightarrow{\text{confirmation}}$ B
authentic

① Alice ~~[illegible]~~get chailie public key for Chalie
② ~~[illegible]~~use D+ encryption key → send to Charlie

T1	T2	T2
S(B)		
R(B)		
	X(B)	
	R(B)	
	W(B)	
	commit	

T2. T1

con

Question 4. Consider the following XML document d_1:

```
<cars>
   <car> <maker>Toyota</maker>
      <year>2012</year>
      <model>Voxy <special>Tourist</special></model>
      <option>Air bag</option>
      <color>blue</color>
   </car>
   <car> <maker>Honda</maker>
     <year>2008</year>
     <model grade=''G''>Insight</model>
     <color>red</color>
     <special>navi
          <option>radar</option>
     </special>
   </car>
</cars>
```

car[//model][

1. Write the answers to the following XPath queries applied on d_1.

   (1) /*/car/[color][option]

   (2) //[color[text() =”blue”]/ancestor:://*[text()=”radar”]]

   (3) //option/../../[//model][following:*]

2. Show a DTD e such that the above document d_1 is valid against e.

3. Consider the following XPath queries (4) and (5).

   (4) //a/following::b//a

   (5) //a/ancestor::*/following::b/*/a

   (a) Show a document d_2 that returns an empty result to query (4) and returns a non-empty result to query (5).

   (b) Show a document d_3 that returns a non-empty result to query (4) and returns an empty result to query (5).

   (c) The product (6) of queries (4) and (5) is an XPath query such that i) if query (4) and query (5) return the same result r on a document, then (6) also returns r, and ii) otherwise (6) returns an empty result. Show such an XPath query (6).

4. Concisely describe advantages and disadvantages of using XML.

解　　答　　用　　紙

2013年 1月23日提出

科目/Subject Information Organization	4 4 1 2 1 5 4 0 - 7	Score
担当教員/Lecturer	氏名/Name XU, Shan	64

24

Question1: Public-key Encryption: User's public encryption key is known to all, decryption key is only known by user.

40

① Bob generates a public encryption key and sends it to Charlie. Charlie issues a certificate to Bob.

② This certificate is stored in encrypted form, encrypted with Charlie's private key, known only to Charlie.

③ Charlie's public key is know to all users, including A, which can decrypt the certificate and obtain Bob's public key.

⑤ Alice encrypted the document first with session key, then with Bob's public key to Charlie. Charlie check it, and send the confirmation message to Bob.

④ Bob ask Alice to send copies of D.

20

⑥ Bob received D, and decrypted it with the decryption key.

⑦ Bob send a message to Alice to inform that he has received D.

Question2:

1. T₁: S(B)., T₁: R(B), T₂: S(B)., T₂: R(B). T₂: X(B), T₂: W(B). T₂: U(B), T₁: R(B). T₁: Commit.

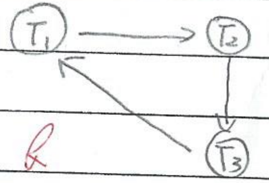
It is obvious that T₂ can not repeatable read −1

−1

2: T1 : S(A), R(A), S(B) (c) 4

T2 : X(B) W(B) X(C). -1

T3 : S(C), R(C), X(A). 4 12



There is a cycle in the dependency graph. of the schedule.

3. T1 : X(A), W(A). U(A) X(A). W(A) abort. 3

T2 : S(A), R(A), U(A), commit

Because T2 read ~~A~~ written by T1, so T1 abort, T2 will abort. +1

The result of the schedule is serializable to conflict equivalence of the commited transactions

4. T1 :

T2 :

Question3: 1. If T1 like S and IX at the same time, it will use SIX lock mode.

For example. T1 scan R, and updates a few turples:

T1 get an SIX lock on R, then repeatedly get an S lock on turples of R, and occasionally upgrade To X on turples. 8

IX lock is compatible with SIX lock. As mentioned SIX is S and IX, S and IX is compatible, so SIX and IX lock is compatible. -2