Birkbeck
(University of London)

MSc Examination for Internal Students

Department of Computer Science and Information Systems

Information Retrieval and Organisation (COIY064H7)
Credit Value: 15

Date of Examination: xxxday xx xxx 2010
Duration of Paper: xx:00 - xx:00

RUBRIC

1. *This paper contains* 15 *questions for a total of* 100 *marks.*
2. *Students should attempt to answer* **all** *of them.*
3. *This paper is not prior-disclosed.*
4. *The use of non-programmable electronic calculators is permitted.*

1. (3 marks)

Briefly describe how pseudo-relevance feedback works. What is the meaning of query drift in this context?

2. (6 marks)

While building an inverted file index 32 GByte of data need to be sorted using external sorting. You have 4 GByte of main memory at your disposal. Draw a schematic diagram showing how the data will be sorted (i.e. how many blocks of which size you have on each level). Assume that during each merge step you can merge (up to) three different blocks at the same time.

3. (6 marks)

In a Boolean IR system using positional information you find the following inverted file index:

cold, 2:	$< 1, 1 :< 6 >;\ 4, 1 :< 4 >>$
days, 1:	$< 3, 1 :< 2 >>$
eat, 1:	$< 6, 1 :< 1 >>$
hot, 2:	$< 1, 1 :< 3 >;\ 4, 1 :< 8 >>$
in, 3:	$< 2, 1 :< 3 >;\ 4, 2 :< 1, 5 >$
lot, 1:	$< 6, 1 :< 3 >>$
nine, 1:	$< 3, 1 :< 1 >>$
old, 1:	$< 3, 1 :< 3 >>$
pease, 5:	$< 1, 2 :< 1, 4 >;\ 2, 1 :< 1 >;\ 5, 2 :< 1, 3 >>$
porridge, 5:	$< 1, 2 :< 2, 5 >;\ 2, 1 :< 2 >;\ 5, 2 :< 2, 4 >>$
pot, 3:	$< 2, 1 :< 5 >;\ 4, 2 :< 3, 7 >>$
the, 4:	$< 2, 1 :< 4 >;\ 4, 2 :< 2, 6 >;\ 6, 1 :< 2 >>$

(a) List all the IDs of all the documents in which the term "pease" is immediately followed by "porridge". (3 marks)

(b) Which document IDs would the following query return? "in /3 pot" (3 marks)

4. (6 marks)

The evaluation of an IR system yields the following results:

Ranking	Recall	Precision
1. d_{34}	20%	100%
2. d_{12}	40%	100%
3. d_{89}	40%	67%
4. d_{39}	40%	50%
5. d_{17}	60%	60%
6. d_{48}	60%	50%
7. d_{23}	60%	43%
8. d_{61}	60%	38%
9. d_{72}	80%	44%
10. d_{25}	80%	40%

Draw the interpolated Recall-Precision curve for the recall values 0%, 20%, 40%, 60%, 80%, and 100%.

5. (8 marks)

Assume you are building indexes for an IR system supporting tolerant retrieval and would like to index the search terms "letters" and "string".

(a) List the entries for the two search terms when using a 2-gram index. (2 marks)

(b) List the entries for the two search terms when using a permuterm index. (2 marks)

(c) Calculate the total size of the entries for the two search terms in bytes for the 2-gram index and the permuterm index. Assume that each character uses one byte. (2 marks)

(d) You will notice that the total size for the entries in the 2-gram index take up less space. However, this involves a trade-off. Briefly describe which drawback the 2-gram index has. (2 marks)

6. (7 marks)

Compute the minimum-editing distance (also called Levenshtein distance) between the terms "system" and "item".

7. (13 marks)

In an IR system on textiles using the vector space model with impact ordering (for inexact top-K retrieval) the following inverted file has been created (the tf and df values in this file are raw values):

pattern	$< 4 : (8, 1 : 1), (4, 2 : 5, 8), (1, 1 : 3) >$
textile	$< 8 : (8, 2 : 1, 2), (4, 2 : 3, 4), (2, 1 : 7), (1, 3 : 5, 6, 9) >$
thread	$< 2 : (4, 1 : 2), (1, 1 : 5) >$
warp	$< 4 : (16, 1 : 5), (2, 2 : 3, 7), (1, 1 : 4) >$
weft	$< 4 : (4, 1 : 3), (2, 1 : 6), (1, 2 : 5, 9) >$

A user sends the query consisting of the search terms "textile", "thread", and "warp" to the system. Counters to compute the score for **three** documents have been allocated.

(a) Determine the three documents that will have their complete score calculated by the system. Assume that the following variants are used for tf-idf and normalisation:

tf	$1 + \log_2(\text{tf})$
idf	$\log_2(\frac{N}{\text{df}})$
normalisation	none

Furthermore, assume that there are a total of $N = 64$ documents in the document collection. Briefly explain your answer. (9 marks)

(b) Compute the complete score for document 3 for the above query. (4 marks)

8. (6 marks)

The following sequence represents a postings list encoded with a γ code. Decode the sequence: 10111101001100111110100111110000

9. (10 marks)

Rank the documents in collection $\{d_1, d_2\}$ for query q using a unigram language model that mixes the distributions estimated from the specific document and the entire collection with the mixture coefficient (i.e., the weight for document distribution) $\lambda = 0.4$.
d_1: "Scottish sheep getting smaller due to climate change study says."

d_2: "The analysis has shown a dramatic shift in the natural ranges for US Bird species in response to climate change."
q: "climate change"

10. (5 marks)

What roles does smoothing play in the language modelling approach to IR?

11. (10 marks)

Consider the following collection of documents that belong to two classes:
China (C) and Japan (J).

	docID	docText	class
TRAINING	d_1	Taipei Taiwan	C
	d_2	Macao Taiwan Shanghai	C
	d_3	Tokyo Sapporo	J
	d_4	Sapporo Osaka Taiwan	J
TEST	d_5	Taiwan Taiwan Sapporo	?

Show how the Naive Bayes algorithm (with Laplace smoothing) can be used to train a classifier and predict the class of test document.

12. (4 marks)

The following table shows the performance of two individual binary classifiers that correspond to the two classes in a document collection, where "truth" is the true class and "call" is the decision of the classifier.

class 1		
	truth: yes	truth: no
call: yes	80	20
call: no	80	300

class 2		
	truth: yes	truth: no
call: yes	180	120
call: no	180	900

(a) Compute the macroaveraged precision and recall.

(b) Compute the microaveraged precision and recall.

13. (6 marks)

Consider the following 6 points in a two-dimensional vector space representing 6 documents in a document collection:
$\vec{d_1} = (1, 2)$; $\vec{d_2} = (2, 2)$; $\vec{d_3} = (3, 2)$; $\vec{d_4} = (1, 1)$; $\vec{d_5} = (2, 1)$; $\vec{d_6} = (3, 1)$.
Suppose that the dissimilarity between each pair of documents is measured by Euclidean distance (straight-line distance). How will the k-means clustering algorithm (with $k = 2$) group these documents using the following two seeds?

(a) $\vec{d_1}$ and $\vec{d_2}$.

(b) $\vec{d_1}$ and $\vec{d_4}$.

(c) $\vec{d_1}$ and $\vec{d_6}$.

14. (5 marks)

Consider the following 5 points in a two-dimensional vector space representing 5 documents in a document collection:
$\vec{d_1} = (0, 1)$; $\vec{d_2} = (0, 2)$; $\vec{d_3} = (1, 2)$; $\vec{d_4} = (3, 1)$; $\vec{d_5} = (2, 0)$.
Suppose that the similarity between each pair of documents is measured by cosine similarity. Draw the dendrogram generated by the single-link Hierarchical Agglomerative Clustering (HAC) algorithm.

15. (5 marks)

The following table shows the result of flat clustering on a document collection, where each letter "A", "B" or "C" represents a document in the true class "A", "B" or "C" respectively.

cluster 1	A B A A A A
cluster 2	A B B B C B
cluster 3	B C C C

What is the purity of the above clustering?