UNIVERSITY OF LONDON
IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2003

BEng Honours Degree in Computing Part III
MSc in Computing for Industry
BEng Honours Degree in Information Systems Engineering Part III
MEng Honours Degree in Information Systems Engineering Part III
BSc Honours Degree in Mathematics and Computer Science Part III
MSci Honours Degree in Mathematics and Computer Science Part III
for Internal Students of the Imperial College of Science, Technology and Medicine

This paper is also taken for the relevant examinations for the Associateship of the City and Guilds of London Institute
This paper is also taken for the relevant examinations for the Associateship of the Royal College of Science

PAPER C340=I3.34

KNOWLEDGE MANAGEMENT TECHNIQUES

Tuesday 6 May 2003, 14:00
Duration: 120 minutes

Answer THREE questions

Paper contains 4 questions
Calculators required

1 a Define what is meant by *precision* and *recall* in the context of retrieval effectiveness.

Explain why it is not normally possible to achieve high values for both precision and recall.

b Explain what is meant by a *text surrogate*.

A database of documents contains a maximum of t different terms.

A document D_i is represented by the vector

$$D_i = (a_{i1}, a_{i2}, \ldots . a_{it})$$

where a_{in} (n = 1…t) represents the significance of term a_n within the context of the document and within the context of the document collection.

Explain how the values a_{in} (n = 1…t) may be determined and discuss how a *thesaurus* may be used in this process.

c Describe the structure of a typical *clustered file* organisation and explain the purpose of the *cluster centroid.*

A clustered file is to be constructed for four documents.

The same four terms are to be used to represent each of the documents; the importance of a given term in a given document is given by its respective weight.

The four documents (D_1, D_2, D_3, D_4) are represented by the following vectors:

D_1 = (0.1, 0.5, 0.8, 0.6)

D_2 = (0.3, 0.7, 0.9, 0.2)

D_3 = (0.4, 0.8, 0.3, 0.7)

D_4 = (0.9, 0.9, 0.01, 0.01)

Using the *single-pass* method, construct a clustered file organisation for the four documents using the *Cosine Coefficient* similarity measure with threshold value = 0.85.

Explain each step of the process clearly.

The maximum cluster size may be assumed to equal three; overlap is permitted between clusters.

The three parts carry, respectively, 20%, 25%, and 55% of the marks.

2a Which two characteristics signify a good set of clusters in cluster analysis? Which three main factors influence the quality of the clustering?

b Consider the following six one-dimensional data points:

$$x_1 = 1,\ x_2 = 1.8,\ x_3 = 3,\ x_4 = 4,\ x_5 = 5,\ x_6 = 6$$

Work out the resulting clusters using the k-means algorithm assuming that the $k = 3$ initial cluster centres were randomly chosen to be x_1, x_2 and x_4. Show all your workings.

c Show that the resulting clustering is suboptimal with respect to the squared error criterion. What has caused this sub-optimality? How could the k-means algorithm be changed to alleviate this problem?

d Comment on whether or not, in general, the k-means algorithm i) is scalable; ii) is able to deal with nominal data types; iii) needs prior knowledge about the cluster structure and iv) is sensitive to the order in which the data points are presented. Include the reasons in your answer.

(The four parts carry, respectively, 25%, 20%, 35%, 20% of the marks.)

3a Colour histograms are often used in image search engines. Explain how a normalised 3-d colour histogram of an RGB image is computed using, say, eight equally sized histogram bins to cover the 3-d RGB space.

b How are these colour histograms used to retrieve similar images from an image database when the query consists of an example image? Describe the whole process, both the offline processing of the database and the online processing when a query image is actually presented. How does the runtime of a query depend on the number n of images in the database, the number p of histogram bins and the number s of pixels in the query image?

c Explain how R-Trees can reduce the runtime of a query. What is the saving in the best case? Write down the pseudo-code for identifying the vector in an R-tree which is closest to a given query vector.

(The three parts carry, respectively, 30%, 40%, 30% of the marks.)

4 A small travel company offered accommodation at six hotels (in various resorts) last summer, with mixed results. Details are given in the following table:

Hotel	**Climate**	**Pool**	**Stars**	**Popular**
Buonavista	hot	yes	3	yes
Excelsior	hot	yes	4	no
Bristol	warm	yes	5	yes
Solara	hot	no	3	no
Ritz	cool	no	4	no
Mayfair	cool	no	5	yes

They are considering offering a new hotel this summer, but are concerned whether it will be popular or not with their clients. The three-star hotel in question is situated in a cool location and has no pool.

a Use the ID3 algorithm to derive **the first branching** of a decision tree from the data on the six current hotels.

NOTE THAT: $\log_2 x = \dfrac{\log_{10} x}{\log_{10} 2}$

b Employ the Naïve-Bayes method to predict the popularity of the new hotel.

The two parts carry, respectively, 55%, 45% of the marks.