UNIVERSITY OF LONDON
IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2004

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
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

Monday 10 May 2004, 14:30
Duration: 120 minutes

Answer THREE questions

Paper contains 4 questions
Calculators required

1 a Discuss two commonly-used measures for determining the overall effectiveness of a given information retrieval operation.

Discuss two further measures for determining the relative effectiveness of the retrieval operation for a given user.

b Explain what is meant by an *inverted file* and discuss the advantages of using inverted files in an information retrieval operation.

A document collection consists of four documents (D_1, D_2, D_3, D_4) containing the following terms:

D_1 = agent James Bond

D_2 = agent cellular phone

D_3 = James uses mobile uses movie

D_4 = mobile Bond uses phone

Construct an inverted file for this document collection.

c Considering the eight distinct terms that occur in the four documents given in part (b), construct a vector for each of the four documents where the individual entries in a vector indicate the relative importance of the respective term within the document in question and within the context of the document collection.

d Discuss how the vectors obtained in part (c) can be incorporated into the inverted file structure and, using an appropriate query, show how this information can be used to retrieve documents in ranked order of relevance.

The four parts of the question are equally weighted.

2a You run a reading club and intend to divide the six members $m^1, \ldots, m^6$ into two discussion groups using a clustering method. Each member has given you their choice of their favourite writer, travel destination, colour, newspaper and genre:

m^1	L Carroll	Canada	green	Daily Mail	poetry
m^2	L Carroll	Italy	green	Daily Mail	science
m^3	V Woolf	UK	red	Guardian	crime
m^4	V Woolf	Spain	blue	Mirror	fiction
m^5	V Woolf	Canada	green	Observer	science
m^6	M Proust	Spain	green	Mirror	crime

Using the simple matching coefficient method compute the corresponding difference between m^1 and m^2, showing all your workings.

Which of the following alternative clustering methods could have been used in this example — k-means; PAM k-medoids; AGNES with single, complete or group average linking; the simple one-pass algorithm; the hierarchical one-pass algorithm. State your reason for each method. You are *not* required to carry out the actual clustering.

b Consider the following triangular distance matrix of four objects A, B, C and D:

	A	B	C
B	1		
C	4	3	
D	6	5	2

Apply the hierarchical AGNES clustering algorithm and draw the corresponding dendrogram. Use the complete linkage method whereby the distance of an element X to a cluster (Y, Z) is the maximum of the distances of X to Y and X to Z, respectively. Show all your workings. Using the dendrogram, demonstrate graphically how you would split the four objects into two clusters.

Contrast hierarchical clustering and k-means clustering with respect to the two criteria of scalability wrt memory usage and run-time usage.

The two parts carry, respectively, 50%, and 50% of the marks.

3 a Sketch a block diagram of an image retrieval system that is queried with example pictures to obtain similar pictures. Explain the general workings of a content-based image retrieval system.

b Statistical properties such as centralised moments can be used for retrieval by colour and retrieval by shape. Define colour moments in general, and explain their use in image retrieval in particular. Relate your explanation to the block diagram from question 3a.

c Contrast the workings of a content-based image retrieval engine with those of a metadata-based image retrieval engine in terms of indexing and retrieval technology.

The three parts carry, respectively, 50%, 30%, and 20% of the marks.

4a Describe the purpose and characteristics of an OLAP system.

b Why is it not considered a good idea to send OLAP queries directly to the transaction processing databases?

c Draw and label a suitable architecture for a comprehensive enterprise-wide decision support system capable of handling OLAP queries efficiently as well as supporting data mining operations.

d A data table T gives the region, age, salary and status for 165 managers in a major oil company. This information has been summarised in table S below, where *Count* in each row of S gives the number of records in T having that row's given region and status, and with age and salary falling within that row's given ranges.

S

Region	Age Range	Salary Range	Status	*Count*
UK	31...35	46K...50K	Snr	*30*
UK	26...30	26K...30K	Jnr	*40*
UK	31...35	31K...35K	Jnr	*40*
US	21...25	46K...50K	Jnr	*20*
US	31...35	66K...70K	Snr	*5*
US	26...30	46K...50K	Jnr	*3*
US	41...45	66K...70K	Snr	*3*
Africa	36...40	46K...50K	Snr	*10*
Africa	31...35	41K...45K	Jnr	*4*
Asia	46...50	36K...40K	Snr	*4*
Asia	26...30	26K...30K	Jnr	*6*

Use the Naïve-Bayes method to classify by status a new US region manager, aged 32 and earning 48K.

The four parts carry, respectively, 20%, 20%, 20% and 40% of the marks.