UNIVERSITY OF LONDON
IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE

EXAMINATIONS 2001

BEng Honours Degree in Computing Part III
MSc in Computing Science
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
MSc in Advanced Computing
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 C440=I3.34

ADVANCED KNOWLEDGE MANAGEMENT TECHNIQUES

Wednesday 2 May 2001, 10:00
Duration: 120 minutes

Answer THREE questions

Paper contains 4 questions
Calculators not required

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

Explain what is meant by

i) a *stoplist*

ii) the process of *stemming*

Discuss the advantages of incorporating both of these features in an Information Retrieval system.

b 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.

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

D_1 = shipment of gold damaged in a fire

D_2 = delivery of silver arrived in a silver truck

D_3 = shipment of gold arrived in a truck

The documents are *not* to be subjected to any processes that will reduce or alter them in any way.

A query, Q, is to be run against the documents, where

Q = gold silver truck

Considering the eleven distinct terms that occur in these three documents, construct a vector for each of the three documents D_1, D_2, D_3 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.

A vector is to be constructed for the given query where the entries in the vector for the terms appearing in the query may be assumed to be:

gold = 0.176
silver = 0.477
truck = 0.176

Using the document vectors and the query vector obtain a ranking of the three documents based on the order of their relative importance.

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

2a Sketch a block–diagram of a MIDI music retrieval system that is to be queried using a MIDI sample to obtain similar music pieces. Explain the general workings of this kind of content–based music retrieval system.

b Histograms are in common use in many content–based multimedia retrieval systems. Explain how a colour histogram is computed and how it is used in image retrieval systems. Motivate why a perceptually uniform colour space should be used for the histogram computation, but do not go into the details of creating such a colour space representation. How can these colour histograms be used for the task of simple video segmentation?

c Semitone histograms, which capture the semitone distribution, are global features. In which way could they be changed to capture some of the temporal aspects of a music piece? Draw a figure to illustrate your idea.

The three parts carry, respectively, 50%, 35%, 15% of the marks.

3a Knowledge Management is becoming more and more important for enterprises. Historically, what are the main interests of Knowledge Management and what are the main steps of the Knowledge Capitalising Cycle of Barthes? What are the links between Knowledge Management and innovation?

b In order to reduce the time to market of the product, the Integrated Team organisation is to be applied. In this context, could you explain the drawbacks of this kind of organisation? In this context, what are the difficulties to capitalise knowledge and know–how for verbal communication and Groupware tools? Describe one approach that attempts to overcome these difficulties. Is this approach able to treat sketches? Give a reason for your answer.

c What are the two main advantages of Integrated Teams? In the context of an innovative project, what are the two main approaches to favour innovation? What are their purposes? What are the three main steps of the TRIZ methodology?

The three parts carry, respectively, 25%, 40%, 35% of the marks.

4a Describe, with the aid of a diagram, the components of a typical modern large-scale decision support system.

b What are the main types of OLAP activities involved when navigating a multidimensional database

c Carefully explain the operation of the ID3 data mining algorithm, by showing how it produces the first level (i.e. the root node and branches emanating from it) of a decision tree to advise whether it is suitable weather for playing cricket, on the basis of the following training set:

Day	**Outlook**	**Temperature**	**Humidity**	**Wind**	**Play cricket?**
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

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