Semester II Examinations 2009/ 2010

Exam Code(s) 4IF

Exam(s) 4th Year B.Sc. Examination

Module Code(s) CT422

Module(s) Modern Information Management

Paper No.
Repeat Paper

External Examiner(s) Professor M. O'Boyle
Internal Examiner(s) Professor G. Lyons
Dr. J. Duggan
*Mr. C. O'Riordan

Instructions: Answer any **FOUR** questions.
All questions carry equal marks.

Duration 3 hours
No. of Pages 4
Department(s) Information Technology
Course Co-ordinator(s)

Q.1.

i) Describe the vector space model approach to information retrieval. Your answer should include a description of the query and document representations and also the comparison approach used. *(8)*

ii) Explain the Extended Boolean model and discuss the advantages and limitations of adopting such a model. *(8)*

iii) Assuming the following document vector has been calculated using some tf-idf weighting scheme for some document dj:

< (*galway*, 0.5), (of, 0.01), (*national*, 0.3), (*university*, 0.2), (*ireland*, 0.4)>

Show how the relevance of the document dj may be calculated with query q in the following scenarios:

a) q = (*university*, *of*, *ireland*) under the vector space model
b) q = (*university* AND *ireland*) under the extended Boolean model
c) q = (*university* OR *ireland*) under the extended Boolean model

(9)

Q.2.

i) What is meant by *relevance feedback* in information retrieval systems and what are the potential benefits and limitations of adopting relevance feedback approaches. *(6)*

ii) Describe with a suitable example the Rocchio approach to relevance feedback in the vector space model. *(9)*

iii) Describe with suitable examples, the differences between *association clusters*, *metric clusters* and *scalar clusters*. Comment on the relative efficiency of the approaches. *(10)*

Q.3.

i) Empirical evaluation of information retrieval systems plays an important role in information retrieval research. With examples discuss:

a) The components of a test collection
b) Metrics that can be used to measure the performance of an IR system

(9)

ii) Pre-processing of a test collection usually involves stop-word removal and stemming. Explain suitable approaches to both. Use the text of this question to illustrate the algorithms you describe. *(8)*

iii) Discuss with a suitable example, an appropriate approach to building an index of terms for a system employing the vector space model. *(8)*

Q.4.

i) Many modern web-based search engines attempt to take into account the web link structure in addition to the content of the pages. Describe, with the aid of an example, the *Page Rank* algorithm that uses information embedded in the web link structure to return relevant documents to a user. Discuss any limitations associated with this approach. *(11)*

ii) Explain briefly how this algorithm could be extended to take into account user-provided preferences. *(5)*

iii) In the context of distributed information retrieval, discuss suitable approaches that could be adopted to tackle the problem of *source selection*. *(9)*

Q.5.

i) Discuss, with reference to existing approaches, suitable approaches to visualising information for users using an information retrieval system. Your answer should include approaches to visualise the a) answer set and b) the relationship between the user query and the answer set. *(14)*

ii) Evolutionary computation has been used successfully to search for suitable means to combine sources of evidence in information retrieval. Discuss such an approach applied to a problem of your choice in information retrieval. Your answer should also identify the strengths and weaknesses of this approach. *(11)*

Q.6.

i) Define what is meant by *collaborative filtering.* With respect to collaborative filtering describe the problems that arise with sparse data sets. *(5)*

ii) Given the following matrix of ratings by four people for six CDS:

	"White Lies"	"Together through life"	"Yonder is the Clock"	"At the Cut"	"Townes"	"Have one on me"
Bob	4	5		4	2	
Frank	2	2		2	3	2
Lee	2	4	3			5
Charlie	5	4		4	1	

(Note: Ratings are in the range 1-5, where 1 indicates "dislike" and 5 indicates "like").

Describe a neighbourhood based approach using Pearson Correlation and weighted averages to generate a prediction for whether Bob will like the CD "Have one on me". Your answer should outline the main steps in the neighbourhood based approach and show the steps required to generate the prediction. *(10)*

iii) Decision trees represent a powerful means to mine useful information from existing datasets. Explain an approach to developing a decision tree given a set of tuples of the following format: <a1, a2, …, ai, classification>. *(10)*