Autumn Examinations 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) Describe and discuss, with the aid of examples, suitable indexing strategies and algorithms to deal with *single term queries, Boolean queries* and *prefix queries*. (*12)*

ii) Outline a compression algorithm to deal with large document collections suitable in the domain of Information retrieval. *(7)*

iii) With respect to compression, outline techniques that may be adopted to compress an inverted index. *(6)*

Q.5.

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 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.6.

i) Natural Language Processing techniques have been used to attempt to resolve ambiguity in user queries. Provide examples of how ambiguity might arise in natural language and suggest an approach to resolving such ambiguities. *(11)*

ii) Learning approaches have been adopted in information retrieval systems to either adapt to changes in user behaviours or to learn an optimal manner in which to combine information or process information to give good performance. Neural networks and evolutionary computation are two such learning approaches. Discuss either learning approach in relation to a problem of your choice in information retrieval. Your answer should also identify the strengths and weaknesses of this approach. *(14)*