Ollscoil na hÉireann, Gaillimh *GX____*
National University of Ireland, Galway

Summer Examinations 2008

Exam Code(s) 4IF, IEM
Exam(s) 4th BSc in Information Technology

Module Code(s) CT422
Module(s) Modern Information Management

Paper No.
Repeat Paper

External Examiner(s) Prof. J. A. Keane
Internal Examiner(s) Prof. G. Lyons
Mr. C. O'Riordan

Instructions: Answer ANY FOUR questions

Duration 3 hours
No. of Pages 4
Department(s) Information Technology

Requirements:
MCQ
Handout
Statistical Tables
Graph Paper
Log Graph Paper
Other Material

SECTION A

Q.1

(a) The vector space model for Information Retrieval is one of the most commonly adopted models. Outline the model explaining both the representation of queries and documents and a means to calculate similarity. Discuss the advantages and disadvantages of such an approach. *(9)*

(b) The accuracy of the vector space model depends on the quality of the weighting of the terms in both the query and documents. Discuss, with reference to Zipf's law, any suitably good weighting scheme. *(8)*

(c) Describe suitable data structures you would use to implement an information retrieval system adopting the vector space model. Discuss the efficiency of your proposed approach. *(8)*

Q.2

(a) What is meant by *Relevance Feedback* in Information Retrieval Systems and what are the potential benefits of adopting relevance feedback approaches. Describe the Rocchio approach to relevance feedback in the Vector Space Model. *(6)*

(b) Discuss, with the aid of examples, how an analysis of the document collection and the returned set, may be used to modify a user's query. *(4)*

(c) The extended Boolean model has been often been used to overcome some of the limitations of the classical Boolean model. Discuss the extended Boolean model. Your answer should explain, with examples, how queries are represented and how comparison to documents is achieved. *(8)*

(d) The assumption of term independence is made in the classical Vector space model. What is meant by the *term independence assumption.* Outline a retrieval model which attempts to overcome the term independence assumption. *(7)*

Q.3

(a) Define what is meant by *collaborative filtering*. Describe, with a suitable example, the main stages involved in generating a recommendation for a user via collaborative filtering. *(9)*

(b) Outline some of the difficulties or limitations associated with collaborative filtering. *(4)*

(c) Explain the structure of a decision tree. Explain how a decision tree could be developed from a set of tuples of the form <attribute$_1$, attribute$_{2,}$ … attribute$_n$, category> such that future tuples of the form <attribute$_1$, attribute$_{2,}$ … attribute$_n$> can be accurately placed in a correct category. *(8)*

(d) Suggest how traditional collaborative filtering or classification (using decision trees or alternative approach) could be applied in the domain of web search. *(4)*

Q.4

(a) 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)*

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

(c) In many modern search environments, many sources of evidence are available which can be used to return relevant documents in response to user's queries. Chose one such domain, outline the sources of evidence available and suggest a suitable mechanism to combines such sources of evidence. *(9)*

Q.5

(a) Describe, with the aid of an example, what is meant by *stemming*. Explain the motivations for adopting stemming in Information Retrieval and discuss any potential problems associated with the approach. *(6)*

(b) 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. *(8)*

(c) Information visualisation has come to fore in many modern systems in an attempt to easy the querying and navigation processes for users. With reference to existing systems, discuss approaches to visualising the document collection, user queries, and the relationship between the query and the returned documents. *(11)*

Q.6

(a) 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. (*13)*

(b) In recent years, there has been a move towards distributed information retrieval systems. Two of the main issues in this domain are that of *source selection* and *results merging (or fusion).* Describe these issues and outline solutions for both *selection* and *results merging.* *(12)*