Ollscoil na hÉireann, Gaillimh ***GX_____***

National University of Ireland, Galway

Summer Examinations 2009

Exam Code(s)	4IF
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

Q.1.

i) 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)*

ii) 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)*

iii) Many modern weighting schemes adopt the *term-independence* assumption. Explain this term. Outline approaches that may be used to overcome this assumption. *(8)*

Q.2.

For either of the two suggested approaches specified below, provide a high level design of a system that would satisfy the requirements provided. For all components of your system, outline the algorithm(s) and technique(s) used. Outline any limitations with your designed system. *(25)*

a) **Incorporation of proximity information**
Often, extra evidence can be gleaned from taking into account the actual position of occurrences of query terms in returned documents. The proximity of these occurrences can be used to increase or decrease the relevance score assigned to a document for a given query in an effort to improve precision.

b) **Fusion of features for information retrieval.**
In many IR domains, many sources of evidence are available to improve retrieval performance (content, structured fields, links, citations etc.). These sources of evidence should be exploited to return suitable results to the user.

Q.3.

i) Many models for Information retrieval have been proposed. Discuss any of two of the following models. Your answer should explain document and query representation, comparison methods and the advantages and disadvantages associated with the model. *(8)*

a) Boolean Model
b) Extended Boolean Model
c) Latent Semantic Indexing
d) Fuzzy Set Approaches

ii) Define what is meant by *collaborative filtering*. Describe, with a suitable example, an approach to generate recommendations for a user. *(8)*

iii) The following database contains tuples on users (*cust_id*), transactions (*trans_id*), and the items purchased in each transaction (*item_id*).

trans_id	***cust_id***	***item_id***
1	201	c76
1	201	mo9
1	201	dm1
2	201	hd00
2	201	dm1
3	202	c76
3	202	hd00
3	202	dm1

Assuming a support level of 60%, describe the Apriori approach to itemset recognition, finding the frequent itemsets in the above database, and also the association rules and their confidence levels. *(5)*

iv) Suggest techniques to implement the Apriori approach in an efficient manner. *(4)*

Q.4.

i) Describe and discuss, with the aid of examples, suitable indexing strategies and algorithms to deal with *single term queries* and *prefix queries*. *(8)*

ii) Explain the term *stemming*. Suggest algorithms to provide effective and efficient stemming. Discuss their efficiency. *(6)*

iii) Outline a compression algorithm to deal with large document collections suitable in the domain of Information retrieval. *(6)*

iv) With respect to compression, outline techniques that may be adopted to compress an inverted index. *(5)*

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) Feedback mechanisms have been adopted to 'learn' more accurately the user's information need. Discuss approaches to improve performance by extending the user's query. *(8)*

ii) Self organising maps have been adopted in a range of domains to cluster information. Discuss self organising maps and their application to information visualisation. *(8)*

iii) 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 an problem of your choice in information retrieval. Your answer should also identify the strengths and weaknesses of this approach. *(9)*