Autumn Examinations 2013

Exam Code(s)	4BCT1
Exam(s)	4th B.Sc. Computer Science and Information Technology
Module Code(s)	CT422
Module(s)	Modern Information Management
Paper No.	
External Examiner(s)	Prof. Michael O'Boyle
Internal Examiner(s)	Prof. Gerard Lyons
	Dr. Michael Madden
	Dr. Colm O' Riordan

Instructions: Answer any 3 questions

Duration	3 hours
No. of Pages	4 including this one
Department(s)	Information Technology
Requirements	None

PTO

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 a suitably good weighting scheme of your choice. *(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) Outline suitable compression approach(es) to deal with large document collections typically found in the domain of Information Retrieval. *(7)*

(b) Given an inverted list representation of a term-document matrix, suggest how this may be extended to work in a parallel computer. Illustrate your answer with a small example. *(6)*

(c) Twitter and other similar social media platforms have led to the creation of vast streams of data and information. Identify the key properties of this source of data that distinguishes it from traditional information retrieval domains. With reference to a problem domain of your choice (e.g. identifying important tweets, users, capturing sentiment) identify the sources of data available, and how they could be used to tackle your chosen problem. *(12)*

PTO

Q.5.

(a) Empirical evaluation of information retrieval systems plays an important role in information retrieval research. Define and discuss the following metrics that can be used to measure the performance of an Information Retrieval system: precision, recall, novelty and coverage. *(10)*

(b) The concepts of topical relevance and component coverage have been used to evaluate approaches to structured retrieval (e.g., retrieval of XML documents). Describe, with a suitable example, these concepts. *(7)*

(c) With reference to a clustering algorithm of your choice, describe suitable approaches to measuring the quality of the clustering algorithm. Your answer should distinguish between internal and external criteria. *(8)*

Q.6.

(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. *(13)*

(b) In the context of distributed information retrieval from a heterogeneous set of collections, suggest an approach to merging answer sets from various sites. Your answer should include an overview of the factors involved that may influence how one would merge the results. With an example, show how these factors are accounted for in your approach. *(12)*