NUI Galway
OÉ Gaillimh

Autumn Examinations 2011

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
	* 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 *d*:

< (*computer*, 0.5), (*science*, 0.3), (*galway*, 0.3), (*university*, 0.3), (*ireland*, 0.2),>

Show how the relevance of the document *d* to query *q* may be calculated in the following scenarios:

a) q = (*computer science ireland*) under the vector space model
b) q = (*computer* AND *science*) under the extended Boolean model
c) q = (*computer* OR *science*) under the extended Boolean model

(9)

Q.2.

i) Discuss with suitable examples, an appropriate approach to developing a system to perform information retrieval adopting a vector space model. Your answer should include a discussion of the indexing structure adopted. *(11)*

ii) Write short notes, with examples, on any two of the following topics:

a) Discuss how you might pre-process the document collection prior to building your index for a vector space model.
b) Discuss how you might augment, and/or use, your indexing structure to incorporate information regarding proximity of query terms in the documents.
c) Discuss how you might augment your retrieval model to incorporate aspects of data retrieval for documents that have both structured data fields (e.g. date, title and author) and unstructured content (the content of the article).

(14)

Q.3.

i) Retrieval of relevant documents (and sub-documents) from a collection of XML documents is a well-known example of structured retrieval. Explain the main differences between structured retrieval and classical information retrieval. Outline some of the difficulties encountered in structured retrieval and suggest means to deal with these difficulties. *(10)*

ii) 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 IR system: *precision, recall, novelty and coverage.* *(10)*

iii) The concepts of *topical relevance* and *component coverage* have been used to evaluate approaches to retrieval from XML collections. Describe these approaches. *(5)*

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 and suggest how one might overcome these limitations. *(12)*

ii) Learning mechanisms 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. *(13)*

Q.5.

i) Clustering approaches have been used in a number of domains and applications of information retrieval. Given a collection of document vectors, outline an algorithm to cluster the documents together. Outline any limitations of this approach. *(12)*

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

	The Aviator	Crash	Vera Drake	Ray	The Incredibles	Don't Look Back
Bob	2	5		2	4	
Eoin	4	2		5	1	2
Lisa	2	4	3			5
Maura	2	4		5	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 Bob for the film "Don't Look Back". Your answer should: *(8)*

- Describe the data given.
- Outline the main steps in the neighbourhood based approach.
- Show the steps required to predict a rating for "Don't Look Back" for Bob.

(13)

Q.6.

i) 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>. *(9)*

ii) Outline an algorithm for efficiently identifying frequently occurring itemsets. *(8)*

iii) In distributed information retrieval, two of the main problems are *source selection* (identifying a good source to which send queries) and *result fusion* (merging results from different collections. The ability to perform these tasks well is often dependent on the quality of the site descriptions. Propose a suitable approach to generating useful descriptions. *(8)*