Semester 2 Examinations 2011 / 2012

Exam Code(s)	4 IF
Exam(s)	4th B.Sc. in Information Technology
Module Code(s)	CT422
Module(s)	Modern Information Management
External Examiner(s)	Professor M. O'Boyle
Internal Examiner(s)	Professor G. Lyons
	Dr. M. Madden
	*Dr. C. O'Riordan

Instructions: Answer 4 questions. All questions will be marked equally.

Duration	3hrs
No. of Pages	4
Discipline	Information Technology
Requirements	None

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:

q = (*university*, *of*, *ireland*) under the vector space model
q = (*university* AND *ireland*) under the extended Boolean model
q = (*university* OR *ireland*) under the extended Boolean model

(9)

Q.2.

i) 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)*

ii) 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. *(3)*
Outline some of the difficulties encountered in structured retrieval and suggest means to deal with these difficulties. *(7)*

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.3. Answer one of the three following questions ((i), (ii) or (iii)). *(25)*

i) A system is required to allow efficient retrieval of relevant documents from a document collection given user queries. Describe the approaches and data structures you would use in building such a system. Your answer should include a description of:
    a) Pre-processing approaches
    b) Suitable indexing structures
    c) Data needed to be stored to rank documents using a weighting scheme such as BM25.

ii) Traditional information retrieval systems typically rank documents based on the similarity of the content in the documents to the content in the user queries. In addition to ranking based on content matching, it is decided to build a system that also ranks based on the style of the documents. A system is required to rank documents depending on how well-written they are. Describe how one could achieve this requirement. Your answer should describe:

    a) Heuristics that could be used to measure style
    b) Suitable approaches to implementing these heuristics. Include a description of the indexing structures or other data structures you would use.

iii) Social search relates to using social relationships between people to recommend information or people to users. Given twitter data (content of tweets, sender, re-tweets, follower data), various graphs can be created to represent the social relationships between users. This graph can then be used to extract features regarding groups of users which may be useful in making recommendations.
    a) Explain what heuristics you could use to create such a graph from the underlying twitter data.
    b) Outline a suitable approach and data structures to represent this graph.
    c) Explain how you could use this graph to identify small cliques or sub-groups of users who are related.

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

iii) Outline a suitable compression algorithm to deal with large document collections 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 the page rank algorithm could be extended to take into account other sources of evidence available (e.g. query logs, user-provided preferences, etc.). *(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) Supervised 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. Discuss any 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)*

ii) 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)*