NUI Galway
OÉ Gaillimh

Semester II Examinations 2012 / 2013

Exam Code(s) 4IF1
Exam(s) 4th B.Sc. (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 4 questions

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

Requirements None

PTO

1. (a) 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. (10)

   (b) Assuming the following document vector has been calculated using some weighting scheme for some document d_j:

   $$< (information, 0.4), (retrieval, 0.2), (filtering, 0.3), (cluster, 0.4), (and, 0.01) >$$

   Show how the relevance of the document d_j may be calculated with query q in the following scenarios:
   - $q =$ (information retrieval and filtering) under the vector space model
   - $q =$ (information) OR (cluster) under the extended Boolean model

   (8)

   (c) Given a weighting scheme such as BM25 or pivoted normalisation, identify the information you need to store in order to calculate the similarity between a query and a document. Suggest a suitable way of indexing the document collection so that queries can be handled efficiently. (7)

2. (a) Feedback mechanisms have been adopted to achieve better representations of the user's information need. Discuss a suitable explicit feedback mechanism that could be used with the vector space model. (10)

   (b) In many scenarios, users are reluctant to offer feedback. Suggest an approach to automatically expand and refine a user's query in the absence of explicit user feedback. Outline the advantages and disadvantages of such an approach. (8)

   (c) Given a submitted query, we can process the query in many ways e.g query expansion. These additional techniques may be more beneficial for *difficult* queries. Suggest a suitable means to identify a *difficult* query. (7)

3. (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)

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. (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)

5. (a) Given a collection of emails each with a classification as either spam or non-spam, suggest a means to learn how to classify unseen emails. Outline any limitations of your approach and discuss how you would evaluate the success of the learned solution. (11)

(b) What is meant by *collaborative filtering*. Illustrate how a collaborative filtering approach could be used to make a prediction regarding *The Incredibles* for user *Ciara* given the following data set: (8)

	Up	*The Incredibles*	*Peter Pan*	*Toy Story*
Ciara	4		3	2
Muireann	1	4	3	
Daria	3	3		
Mia	4	5	4	3

(c) Discuss what you consider to be the main limitations of a collaborative filtering approach and suggest approaches to overcome these limitations. (6)

6. (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)