NUI Galway
OÉ Gaillimh

Semester I Examinations 2013 / 2014

Exam Code(s) 4BCT
Exam(s) 4th B.Sc. in Computer Science and Information Technology

Module Code(s) CT422
Module(s) Modern Information Management

Paper No.

External Examiner(s) Prof. Liam Maguire
Internal Examiner(s) Prof. Gerard Lyons
Dr. Michael Madden
*Dr. Colm O'Riordan

Instructions: Answer any 3 questions

Duration 2 hours
No. of Pages 3 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. (9)

   (b) Using the very small document collection presented below to illustrate your answer, outline, with reference to a weighting scheme of your choice, the factors that a modern weighting scheme takes into account when assigning weights to the terms in the documents.
   - *Doc 1*: *finally delivered down*
   - *Doc 2*: *nothing was delivered*
   - *Doc 3*: *too much information about nothing. too much.*

   (12)

   (c) In the vector space model, terms are considered to be independent of each other. Outline how you might calculate or estimate the degree of relatedness between terms occurring in a collection. (12)

2. (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 briefly the Page Rank algorithm that uses information embedded in the web link structure to return relevant documents to a user. (12)

   (b) Given a simple graph of three nodes a, b, c, representing web pages with the following directed edges $(a, b), (a, c), (b, c), (c, b)$ representing web links, illustrate an approach to calculate the hub and authority score for each page. Rank the pages in terms of hub and authority scores. (12)

   (c) *Collaborative filtering* can suffer from the problem of *sparsity* in the ratings set. Explain the terms in italics. Outline an approach that may be used to help overcome this *sparsity* problem. (9)

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.* (11)

   (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. Discuss any limitations of these metrics. (11)

   (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. (11)

4. (a) 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. (11)

   (b) Query modification is often used by systems to attempt to improve precision and recall for a given information need. Discuss an approach, given user feedback, on the returned answer to improve the performance of the query. Illustrate the approach for a query "`too much data`" when a user indicates that *Doc3* in the collection is relevant. (11)

   (c) Traditionally, information retrieval systems and web search engines have presented results in a list ordered by estimated relevance. Identify any limitations with this approach and suggest an alternative presentation of results. (11)