GX1341

OLLSCOIL na hÉIREANN
THE NATIONAL UNIVERSITY of IRELAND

NATIONAL UNIVERSITY OF IRELAND, GALWAY

SUMMER EXAMINATIONS 2000

Fourth University Examination in Information Technology

CT422 Modern Information Management

Prof. D. Bell
Dr. G. Lyons
Mr. C. O' Riordan

Time allowed: **Three hours**

Answer any 4 questions
All questions carry equal marks

Q.1. i) Compare and contrast the vector-space model and probabilistic models in information retrieval systems with respect to representation and comparison of information.

ii) Explain and discuss weighting schemes that have been used in systems adopting the vector space approach.

iii) Latent semantic indexing has been used to attempt to improve upon the performance of the vector-space model. Explain the latent semantic indexing approach. Discuss any short-comings associated with this approach.

Q.2. i) The usefulness of an IR system is often measured using the metrics of precision and recall. Describe, with an example, these metrics. Discuss the suitability of these metrics for different types of information retrieval and filtering systems.

ii) Relevance feedback is an integral component of information retrieval systems. Discuss relevance feedback mechanisms for systems adopting the following models:

a) vector space
b) probabilistic

iii) Outline problems associated with attaining relevance feedback from users and suggest mechanisms to attempt to overcome these problems.

iv) Outline, briefly, how user-specified queries can be automatically expanded using automatic analysis of returned documents.

Q.3. i) Discuss, with references to existing systems, the advantages of collaborative filtering over traditional content based filtering.

ii) Describe, with examples, neighbourhood-based approaches to collaborative filtering including in your answer an explanation of different correlation metrics that could be used.

iii) Suggest approaches to overcome the difficulties that arise in neighbourhood based *recommender systems* (e.g., scarcity of available ratings, ascertaining optimal size of neighbourhood).

Q.4. i) Many different activities may be involved in an information retrieval/filtering systems: browsing, querying, viewing results, reformulating query etc. Explain, with reference to existing systems, the issues involved in designing an interface to aid the user with the involved activities.

ii) Outline possible approaches to aid the user in visualising both the relationship between the given query and the document set, and the relationship between returned documents and the remainder of the document set. Include in your answer a discussion on the interface design and on the underlying algorithms.

Q.5. i) With respect to compression, outline, with an example, Huffman based compression. Discuss the relative merits and shortcomings of word-based Huffman encoding over character-based Huffman encoding.

ii) Suggest, giving reasons, suitable indexing strategies to allow efficient evaluation of the following types of queries: single term, prefixes, suffixes and vector space representation.

Q.6. i) With respect to parallel IR systems, explain techniques that can be used to improve the speed of any one query

ii) Newer approaches to indexing web-based material have exploited data mining techniques. Explain, with reference to any system, any such technique.

iii) Discuss the operation of web-based meta-searchers (i.e. those that piggy-back on traditional search engines). Explain the difficulties that may arise.

Q.7. i) Explain how a decision tree could be developed from a set of tuples of the form <attribute_1, ..., attribute_n, category> such that future tuples of the form <attribute_1, ..., attribute_n> can be accurately placed in a correct category.

Outline any problems or shortcomings associated with this approach.

ii) A common problem in e-commerce systems is the identification of *item-sets* (sets of items that are purchased together). Outline a suitable, efficient algorithm to identify all *item-sets* which occur with a frequency above a given threshold.