TE WHARE WĀNANGA O TE ŪPOKO O TE IKA A MĀUI
VICTORIA
UNIVERSITY OF WELLINGTON

EXAMINATIONS — 2010

MID-YEAR

COMP 423

INTELLIGENT AGENTS

Time Allowed: 3 Hours

Instructions: Attempt all questions.
The exam will be marked out of 180.
Calculators and non-electronic foreign language dictionaries are permitted.
Clean copies of the papers will be distributed for the exam.

Questions

1. Information retrieval and Web search [35]
2. Query expansion [45]
3. Clustering [80]
4. Web mining and others [20]

Question 1. Information retrieval and Web search [35 marks]

(a) [10 marks] Explain how you can evaluate an information retrieval system. Use an example to show how the evaluation parameters are calculated.

(b) [13 marks] Use example documents to show how the terms are weighted in information retrieval systems.

(c) [12 marks] State the main advantages and disadvantages of the PageRank algorithm.

Question 2. Query expansion [45 marks]

(a) [10 marks] Briefly explain the main approaches in query expansion.

(b) [15 marks] In your opinion, what are the main problems of current query expansion techniques? State any solutions you may have for solving the problems.

(c) [20 marks] Suppose you are required to develop an "Automatic Thesaurus Generation" system. (You may consider it as a sub-problem of query expansion.) The input is a term that can be a word or a phrase, and the output should be a ranked list of terms that are closely related to the input term. Design a good method to solve the problem. Explain the process and detail the function that you use to rank the terms.

Question 3. Clustering [80 marks]

This course covered three closely related areas: data clustering, document clustering and web search results clustering, and introduced six clustering algorithms: HAC (Hierarchical Agglomerative Clustering), K-means, BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), STC (Suffix Tree Clustering), QDC (Query Directed Clustering), and SRC (Learning to cluster web search results).

(a) [15 marks] Suppose someone has done a survey and you are required to use clustering to analyse the data. The survey consists of some fill-the-gap questions that ask for numerical or textual data, some multi-choice questions and some open questions. There might be missing data, for example, some questions are left unanswered. Briefly write a step by step process to show how you are going to apply clustering to this task.

(b) [15 marks] Suppose you have a big collection of customer data and you are required to use clustering to find the customer groups. The data set is large and noisy, including information such as customer's age, gender, annual income, number of children, etc. You know very little about the customer groups, and there might be outliers (customers that do not belong to any groups). What is a good clustering algorithm for this task? Justify your answer.

(c) [15 marks] Explain how data clustering algorithms can be applied to document clustering. Give examples to show the document representation and similarity measures.

(d) [20 marks] Compare the three web page clustering algorithms STC, QDC and SRC, and discuss the main strengths and weaknesses of each algorithm.

(e) [15 marks] In your opinion, what should be the main directions of future research in the area of web page clustering? Your answer should include a discussion of the main limitations of the current research and how they might be addressed.

Question 4. Web mining and others [20 marks]

(a) [9 marks] Very briefly outline the following research area: Web usage mining, information extraction and opinion mining.

(b) [11 marks] Discuss how "Artificial intelligence" can be used to improve Web search. In your opinion, what should be the focus of future research in this area?

****