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

EXAMINATIONS — 2011

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 [40]
2. Query expansion [40]
3. Clustering [80]
4. Project [20]

Question 1. Information retrieval and Web search [40 marks]

(a) [20 marks] The "Vector Space Model" is one of the most commonly used information retrieval models. Explain how it works and your answers should include how the documents and queries are represented, how the weights are calculated, and how the similarity between a query and a document is calculated.

(b) [20 marks] Text is typically represented as "bag of words" in information retrieval, document clustering and classification, and many other text mining systems. Explain the main problems of this representation and state any solutions you may have for solving or reducing the problems.

Question 2. Query expansion [40 marks]

(a) [20 marks] Briefly explain the two main approaches in query expansion. State the main problems or limitations of each approach and any solutions you may have for solving the problems.

(b) [20 marks] Recent research has suggested that Wikipedia is useful for query expansion and there are many different ways of integrating Wikipedia into query expansion systems. In your opinion, how would you use Wikipedia to improve the current query expansion systems?

Question 3. Clustering [80 marks]

(a) [20 marks] The lectures introduced two well known clustering algorithms HAC and K-means. Choose ONE of them and very briefly explain how it works. Your explanation should use the following example, showing the steps in the first iteration, and showing how the data set is updated after the first iteration.

The following example describes seven food items (a to g) using two numeric attributes. You may assume that these food items can be clustered into four groups.

Food item	Protein Content (P)	Fat content (F)
a	1	60
b	8	20
c	4	30
d	1	20
e	7	10
f	2	50
g	4	40

(b) [10 marks] Describe the main steps of a web page clustering process.

(c) [10 marks] TF-IDF is a typical way of weighting terms for text representation. Discuss its advantages and disadvantages.

(d) [20 marks] It is generally agreed that "phrases" carry more semantics than single words. A simple idea for designing a new clustering algorithm would be "to represent documents using key phrases and to compare the similarity of documents by counting the number of key phrases shared between documents". Identify the main challenge problems in this approach and explain how you would address these challenge problems so that you can improve this simple idea and use it in web page clustering.

(e) [20 marks] Design your own web page clustering algorithm, pointing out the new ideas and the main contributions. You may design a new algorithm from scratch, or improve one of the existing algorithms, or combining different algorithms. You may also consider to use external resources such as Wikipedia.

Question 4. Project [20 marks]

What are the main ideas and the main contributions of your COMP423 project? If it were a Masters project and you had one year to work on your project full time, how would you improve your system and design/redesign your experiments to fully implement and evaluate your system?

****