

**EXAMINATIONS — 2009**

MID-YEAR

**COMP 423**  
**INTELLIGENT AGENTS**

**Time Allowed:** 3 Hours

**Instructions:** Attempt all questions.

The exam will be marked out of 100.

Calculators and non-electronic foreign language dictionaries are permitted.

Clean copies of the papers will be distributed for the exam.

## Questions

1. Web Search [60]
2. Web Page Clustering [20]
3. Wrapper Induction [20]

**SPARE PAGE FOR EXTRA ANSWERS**

Cross out rough working that you do not want marked.  
Specify the question number for work that you do want marked.

## Question 1. Web Search

[60 marks]

- (a) [10 marks] Discuss the “next generation” of search engines. Your answer should include the new features you want to add to the current search engines, or the improvements you want to make to existing features.
- (b) [10 marks] A small company has recently set up a web site to allow users to post advertisements such as job advertisements on the site. Currently, the web site cannot compete with other popular web sites such as Trademe and Seek; one reason is that this site has a very low Google page ranking. For example, if you do a Google search using the query “job advertisements”, this site is not listed in the first page of the search results. List your ideas for helping this company to increase the page ranking of this web site.
- (c) [10 marks] What is the main idea and the main contribution of your COMP423 project? If you had a further six weeks to work on your project full time, how would you improve your system and design/redesign your experiments to fully evaluate your system?
- (d) [10 marks] State the main limitations of the query expansion algorithm Abraq which was introduced in the lectures. Include any assumptions it makes and list your ideas for improvements.
- (e) [10 marks] The query expansion algorithm Abraq usually works well for expanding relatively short queries by adding one more word to the query. Now consider a different problem when we have very long queries (e.g. a short description of what is relevant) which do not return relevant results. How would you adapt this algorithm or design a new algorithm to address this problem?
- (f) [10 marks] “Google similarity distance” and “ontology learning” are two very different approaches to “semantics” (meaning of words, relationship of words). State the strengths and weaknesses of each approach.

## Question 2. Web Page Clustering

[20 marks]

The course introduced five clustering algorithms: AHC (Agglomerative Hierarchical Clustering), K-means, STC (Suffix Tree Clustering), QDC (Query Directed Clustering) and MAXCCLUS (MAXCCLUS is not really used for clustering though).

- (a) [10 marks] Choose one algorithm and use an example to show how it works.
- (b) [10 marks] Discuss the main strengths and weaknesses of the QDC algorithm.

## Question 3. Wrapper induction

[20 marks]

- (a) [10 marks] The course introduced many wrapper learning systems, including RoadRunner, STAVIES, WIEN, STALKER, DEPTA, ClusterWrapper and NET. In your opinion, which one is the best and why?
- (b) [10 marks] In your opinion, what should be the main directions (main focuses) of future research in the area of data extraction from semi-structured web pages and wrapper induction? Your answer should include a discussion of the main limitations of the current research and how they might be addressed.

\*\*\*\*\*