VICTORIA UNIVERSITY OF WELLINGTON
Te Whare Wānanga o te Ūpoko o te Ika a Māui

EXAMINATIONS — 2004

MID-YEAR

COMP 423

INTELLIGENT AGENTS

Time Allowed: 3 Hours

Instructions: Attempt all questions in part I
Attempt any **three** of the four questions in part II

The exam will be marked out of 180.

Calculators and non-electronic foreign language dictionaries are permitted.

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

Questions

Part I (Attempt **all** questions) [90 marks]

1. Intelligent Agents [10]
2. Information Retrieval *vs* Information Extraction [10]
3. Information Agents [10]
4. Mobile Robots Agents [10]
5. Information Retrieval [20]
6. Wrapper Induction for Information Extraction [30]

Part II (Attempt **three** out of four questions) [90 marks]

7. Spam Filtering [30]
8. Lumière and the Microsoft Office Assistant [30]
9. Version Spaces and Programming by Demonstration [30]
10. Qualitative Spatial Reasoning [30]

Part I

Attempt all of questions 1 to 6

Question 1. Intelligent Agents [10 marks]

The term "intelligent agent" is not precisely defined, but there are a number of properties that are significant for intelligent agents.

(i) List a set of such properties.

(ii) Briefly discuss two of the systems that we looked at during the course in terms of how they did or did not satisfy these properties.

Question 2. Information Retrieval *vs* Information Extraction [10 marks]

Information retrieval and information extraction are two different research areas. Give one example application for each area and describe the differences between the two areas.

Question 3. Information Agents [10 marks]

Consider the information agent you developed in assignment 2. Describe its behaviour in terms of the PAGE framework: Percepts, Actions, Goals, Environment.

Question 4. Mobile Robots [10 marks]

The design of an intelligent mobile robot to operate in places with lots of people (eg, a museum guide, footpath cleaning robot, cocktail party waiter, etc) must deal with a wide range of problems at different levels of abstraction. Briefly (in one or two sentences each) describe at least five of these problems.

Question 5. Information Retrieval [20 marks]

(a) [8 marks] Suppose we use the Vector Space Retrieval Model for retrieving relevant documents. Each document is represented as a term vector defined by a set of term weights. Each term weight reflects the estimated importance of one particular term in the document and is often calculated as follows:

$$T_i = TF * IDF \qquad \text{where} \quad TF = \frac{tf}{\textit{doc-length}} \quad \text{and} \quad IDF = \log \frac{N}{\textit{df}} + 1$$

Suppose our information source contains the following three documents:

Document1: `cat eat mouse, mouse eat chocolate`
Document2: `cat eat mouse`
Document3: `mouse eat chocolate mouse`

Suppose we represent the three documents as term vectors using all the words in the documents. Give the term vector for Document3. Show your working. You may use the log table below.

(b) [12 marks] Google uses a retrieval model based on an index from words to pages and a page ranking algorithm. Briefly describe the main differences between the model used in Google and other classic retrieval models such as the Vector Space Retrieval Model and discuss the relative advantages and disadvantages.

Table of $\log_2(\frac{n}{m})$

$m \setminus n$	**1**	**2**	**3**	**4**	**5**	**6**
1	0	1	1.58	2	2.32	2.58
2	-1	0	0.58	1	1.32	1.58
3	-1.58	-0.58	0	0.42	0.74	1
4	-2	-1	-0.42	0	0.32	0.58
5	-2.32	-1.32	-0.74	-0.32	0	0.26
6	-2.58	-1.58	-1	-0.58	-0.26	0

Question 6. Wrapper Induction for Information Extraction [30 marks]

Consider the following two example pages:

```
<HTML><HEAD><TITLE>Used Cars for Sale</TITLE></HEAD>
<BODY><B>Used Cars for Sale</B><p>
<B>Toyota</B> <I>1992</I> <I>$3242</I> <BR>
<B>Toyota</B> <I>1989</I> <I>$2000</I> <BR>
<B>Ford</B> <I>1995</I> <I>$4000</I> <BR>
<HR><B>End</B></BODY></HTML>
```

and

```
<HTML><HEAD><TITLE>Used Cars for Sale</TITLE></HEAD>
<BODY><B>Used Cars for Sale</B><p>
<B>Toyota</B> <I>1999</I> <I>$6242</I> <BR>
<B>Ford</B> <I>1995</I> <I>$4000</I> <BR>
<HR><B>End</B></BODY></HTML>
```

Suppose we want to extract the Make, Year, and Price for each car in the pages. If the learning algorithm requires labelled training data, you may assume that the two pages are manually labelled properly in the format that is required.

(a) [7 marks] Give one wrapper that could be learned for the examples above by the BuildHLRT algorithm described in the paper "Wrapper Induction for Information Extraction" by Kushmerick, Weld, and Doorenbos.

(b) [7 marks] Give one wrapper that could be learned for the examples above by the STALKER algorithm described in the paper "A Hierarchical Approach to wrapper Induction" by Muslea, Minton, and Knoblock.

(c) [7 marks] Give one wrapper that could be learned for the examples above by the RoadRunner algorithm described in the paper "RoadRunner: Towards Automatic Data Extraction from Large Web Sites" by Crescenzi, Mecca, and Merialdo.

(d) [9 marks] Compare the three algorithms above and list the advantages and disadvantages of each algorithm.

Part II

Attempt **three** of questions 7 to 10

Question 7. Spam Filtering Agents [30 marks]

(a) [12 marks] Briefly outline how the Naïve Bayes spam filter classifies a message as mail or spam. In your explanation, include at least the following points:

- the probability formulas that the filter is based on,
- the assumptions underlying Naïve Bayes,
- the effect of using a threshold,
- the method for training the spam filter.

(b) [5 marks] Many spam messages now contain lists of random words chosen from a dictionary. Explain how this would probably trick a Naïve Bayes spam filter into classifing such spam as real mail.

Example text from such a spam message:

```
closeup architects peered beseech harvester grouse facade
apparently loadings enchant knighted stamens foresight mahogany
feathered shared temperate tolerably leans returnable ...
```

(c) [6 marks] Suggest how a Naïve Bayes filter could be modified to detect spam messages with lists of random words. What difference would this modification make to the training phase?

(d) [7 marks] Some spam messages contain chunks of English text from books, rather than lists of unrelated words. This text may be 10 times larger than the "real" content of the spam message. Explain why this is even harder for a Naïve Bayes filter to detect, and suggest some approaches for a spam filtering agent that could deal with these messages.

Example text from such a spam message:

```
But now the shouts of a vast concourse of amazed spectators
reached the boy's ears He remembered that he was suspended in
mid-air over the crowded street of a great city, while thousands
of wondering eyes were fixed upon him ...
```

[Attempt **three** of questions 7 to 10]

Question 8. Lumière and the Microsoft Office Assistant [30 marks]

(a) [10 marks] Bayesian networks, such as the one in the diagram below, are used widely for intelligent agents that must reason about uncertain knowledge.

- **(i)** Explain what the nodes and edges in a Bayesian network represent.
- **(ii)** Explain what probability data is needed for a Bayesian Network.
- **(iii)** Explain why Baysian Networks are a good way of representing probability information in a complex domain?

(b) [10 marks] A major component of the the Lumière system was concerned with inferring the user's goal from the observations of the user's behaviour.

- **(i)** Explain what kinds of observations Lumière used.
- **(ii)** Outline how the inference was done. Use the diagram above (or a similar one) in your explanation.
- **(iii)** What complications does time introduce?

(c) [10 marks] As an intelligent agent for assisting users, the Microsoft Office Assistant is generally viewed as a failure. Explain at least three causes of its failure. For each one, suggest an approach that would have improved users' perceptions of the assistant.

[Attempt **three** of questions 7 to 10]

Question 9. Version Spaces and Programming by Demonstration [30 marks]

(a) [3 marks] Version Spaces are a way of representing hypotheses while learning from examples. Explain the key idea of version spaces.

(b) [3 marks] If a version space is represented by two boundaries (maximally general generalisations and maximally specific generalisations), explain how a learning algorithm would modify the version space in response to (i) a new positive example and (ii) a new negative example.

(c) [10 marks] Suppose that you are using version spaces to infer the description of a class of unix commands that causes a certain kind of error, and that you are given the following three commands – two positive examples of the class and one negative example not in the class:

Example 1 (positive): `copy mydata1 mydata2 temp/ ; lpr temp/my*`
Example 2 (positive): `copy mydata1 old1 old2 temp/ ; ls temp/*`
Example 3 (negative): `copy old1 mydata2 temp/; move mydata2 /dev/null/`

(i) Show the Maximally Specific Generalisation(s) after the first two examples.

(ii) Show the Maximally General Generalisation(s) after the third example.

Assume there is a grammar for unix commands and that the parse-trees of the examples are given below:

Example 1 (+ve):

Example 2 (+ve):

Example 3 (–ve):

(d) [10 marks] Explain the task that the SmartEdit system was intended to solve, and discuss the strengths and limitations of SmartEdit.

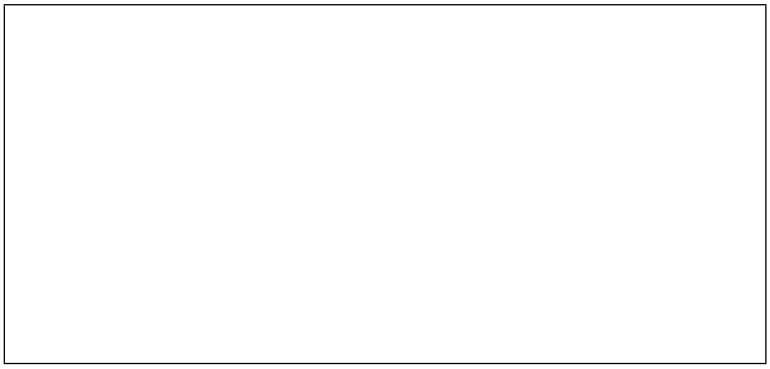
(e) [4 marks] In assignment 4, you addressed the task of inferring patterns to specify locations in text. Explain why this task is difficult to do in a general way.

[Attempt **three** of questions 7 to 10]

Question 10. Qualitative Spatial Reasoning [30 marks]

(a) [12 marks] Forbus's qualitative spatial reasoning system enabled an agent to efficiently simulate possible behaviour of an object moving in a complex space.

The diagram below shows a ball that is currently rolling on a surface towards a hole. The space can be divided into 12 qualitative regions: **A**, **B**, ... **F** are surfaces, and **G**, **H**, ... **L** are open regions.

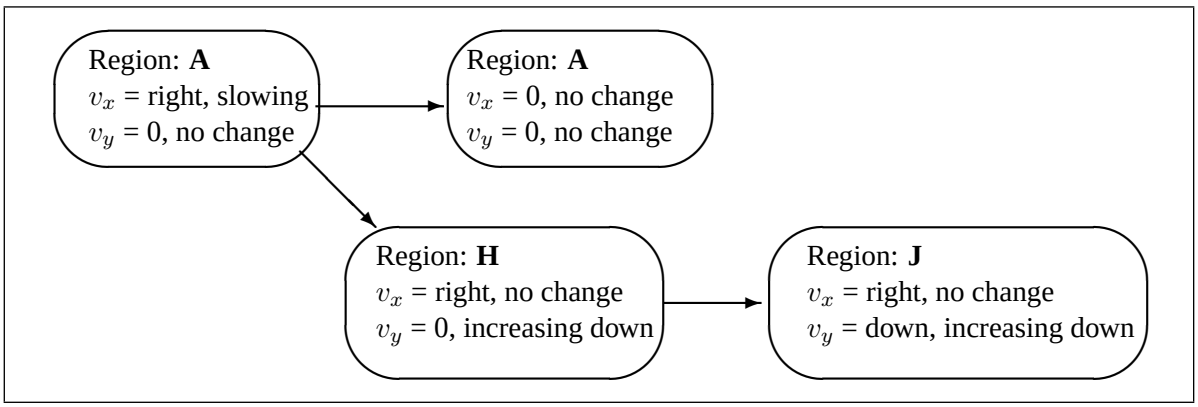


The behaviour of the ball can be described by a directed graph in which each node describes a qualitatively different part of the behaviour. Each node should specify

- The region the ball is in,
- The horizontal velocity $v_x \in \{$*left*, 0, *right*$\}$, and
  how v_x is changing: *no change*, *slowing*, *increasing left*, *increasing right*,
- The vertical velocity $v_y \in \{$ *up*, 0, *down*$\}$, and
  how v_y is changing: *no change*, *slowing*, *increasing up*, *increasing down*.

The first few states of the behaviour graph are shown below: the ball may either slow down to a stop on surface **A**, or may move into region **H** then fall into **J**. (We assume friction, but no wind resistance.)

Draw *four* further states of this graph, including at least two ways that the ball could exit from region **J**, and at least a bounce or a slide. (You do *not* need to copy the graph below.)



(b) [18 marks] In their paper, Forbus, Mahoney, and Dill suggest that qualitative spatial representation could be used in strategy games. Explain

(i) how such a representation could be constructed and used for path planning.

(ii) the advantages of this representation over just a low level quantitative representation.

(iii) the limitations and disadvantages of this representation.

****