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

EXAMINATIONS — 2005

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.

Questions

1.	Information Retrieval and Search	[30]
2.	Information Retrieval and TF-IDF	[15]
3.	Wrapper Induction for Information Extraction	[30]
4.	Computer Games: Pathfinding	[15]
5.	Robot Agents: Rhino	[18]
6.	Help agents and Bayesian Networks	[20]
7.	Language agents: Grammar Checkers	[25]
8.	Robot Agents: The Mars Rover	[27]

Question 1. Information Retrieval and Search [30 marks]

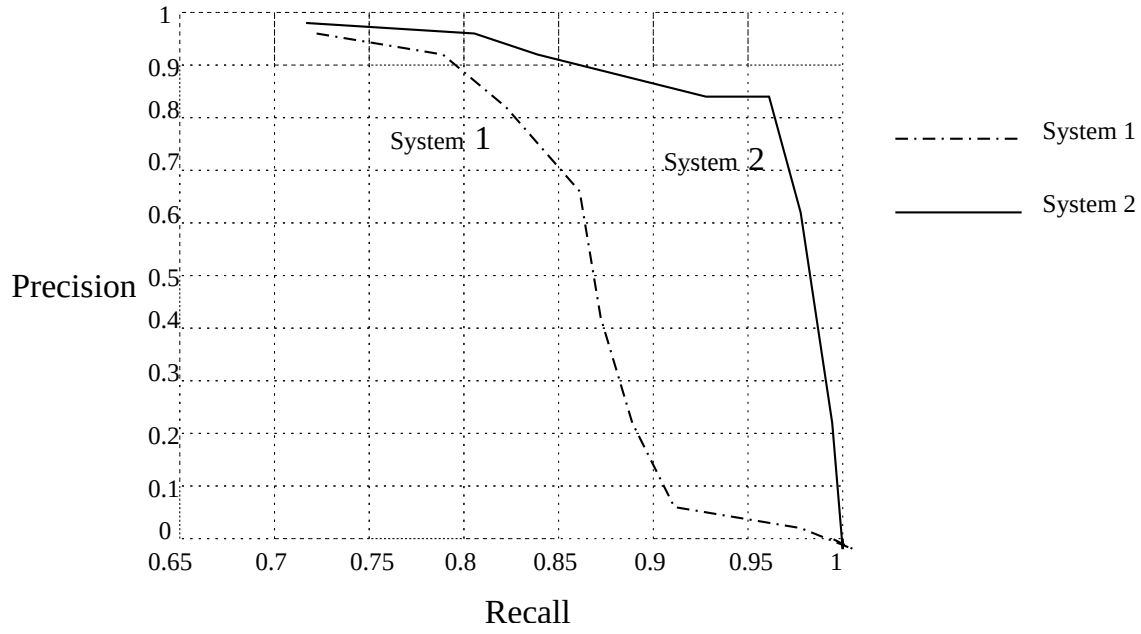
(a) [5 marks] State the main problems of current search engines.

(b) [5 marks] State the features/characteristics/technologies you think that are important for an ideal Web search engine (or the so-called "third generation search engines").

(c) [10 marks] What is the main idea of PageRank? State its limitations/problems.

(d) [5 marks] Precision and Recall are two commonly used parameters for the evaluation of Information Retrieval and Information Extraction systems. Define Precision and Recall using examples.

(e) [5 marks] The following figure shows the performance of two information retrieval systems. State which system is better and briefly explain why.



Question 2. Information Retrieval and TF-IDF [15 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:

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

Suppose our information source contains the following three documents:

Document1: `Programming Agent Java Programming`
Document2: `Software Agent`
Document3: `Software`

The query is

Query: `Agent`

Suppose we represent the three documents as term vectors using all the words in the documents.
Give the term vector for Document1 and the Query. Show your working.
You may use the log table below.

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

m \ 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 3. Wrapper Induction for Information Extraction [30 marks]

Consider the following example page (the attributes of tags have been deleted):

```
<HTML><HEAD><TITLE>BigBook</TITLE></HEAD>
<BODY><FONT>BigBook</FONT><TABLE>
<TR> <TD><A><IMG></A><A> <IMG></A></TD>
<TD><A><FONT>1 2 3 Convenience Store</FONT></A></TD>
<TD><FONT>144 Hempstead Tpke</FONT></TD>
<TD><FONT>W Hempstead, NY</FONT></TD> </TR>
<TR> <TD><A><IMG></A></TD>
<TD><A><FONT>1 hour Auto Glass Inc</FONT></A></TD>
<TD><FONT>403 West St</FONT></TD>
<TD><FONT>New York, NY </FONT></TD> </TR>
</TABLE><P>
</BODY>
</HTML>
```

Suppose a number of similar pages can be downloaded from the same Web site and we want to extract the three data fields (Name, Address, City) from the pages.

(a) [20 marks] Show the wrappers that could be learned for the example page above using **TWO** different wrapper induction systems. Choose the two systems from the systems introduced in lectures (HLRT/WIEN, STALKER, RoadRunner and AutoWrapper) and/or your project-1.

If the learning algorithm requires labelled training data, you may assume that the pages are manually labelled properly in the format that is required.

(b) [10 marks] State the limitations of **ONE** of the systems you chose in (a).

Question 4. Computer Games: Pathfinding [15 marks]

A* is a fast algorithm for finding shortest paths in a graph, as long as there is a good estimate of the remaining distance to the goal from each point. In games that have a grid-based map of the world, the straight line distance from a grid cell to the goal cell is generally a good heuristic.

(a) [5 marks] Describe a category of obstacles that will make this heuristic less effective and outline one approach to improving the A* search in these cases.

(b) [10 marks] Even with an efficient search algorithm, finding paths in a game with a large grid by searching the graph of grid cells may be too slow. Outline an approach that can speed up path finding in large maps. Specify any assumptions about the map and the paths that your approach depends on.

Question 5. Robot Agents: Rhino [18 marks]

The papers on Rhino addressed the design of mobile autonomous robot agents that have to interact with people in a geographically limited environment (*e.g.* a particular building).

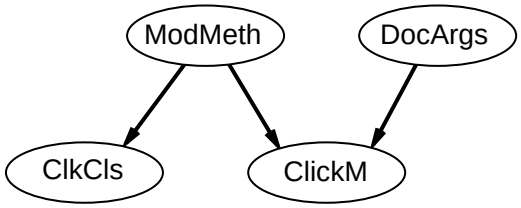
(a) [10 marks] Briefly list and explain the different computational tasks that such an agent must solve.

(b) [8 marks] What assumptions about its world did Rhino rely on? Explain how Rhino would have failed if those assumptions had not been true.

Question 6. Help agents and Bayesian Networks [20 marks]

Suppose you are trying to build a system to provide help for the user of a programming development environment. A small part of your Bayesian Network is shown below.

- All four nodes are boolean.
- ClkCls ("right click on class name in side bar") and ClickM ("right click on method call") represent user actions.
- ModMeth ("modify code of a method") and DocArgs ("view documentation on method arguments") represent goals that the user may have (possibly both goals at the same time).
- The tables give the prior probabilities of the goals, and the conditional probabilities of the actions.



$p(\text{ModMeth}) = 2/10$

$p(\text{DocArgs}) = 4/10$

ModMeth	$p(\text{ClkCls} \mid \text{ModMeth})$
true	$7/10$
false	$1/10$

ModMeth	DocArgs	$p(\text{ClickM} \mid \text{ModMeth \& DocArgs})$
true	true	9/10
true	false	8/10
false	true	2/10
false	false	0

(a) [2 marks] There is no link in the network from DocArgs to ClkCls. State what the absence of this link represents.

(b) [3 marks] The table says that $p(\text{ClickM} \mid \text{ModMeth} = false \ \& \ \text{DocArgs} = false) = 0$. Why is this unreasonable in practice ?

(c) [8 marks] When the system starts up, it has not observed any user actions.

- What is the system's estimated probability that the user has the ModMeth goal?
- What is the probability of observing the ClkCls action in the near future?
- What is the probability of observing the ClickM action in the near future?

(d) [7 marks]

At a later time, the system observes the ClickM action (but has not observed any other relevant actions).

- What is the new probability that the user has the ModMeth goal?
- What is the new probability of observing the ClkCls action in the near future?

Question 7. Language agents: Grammar Checkers [25 marks]

Many ESL writers have difficulty learning when to use "the" and when not use "the" in a noun phrase. Suppose you are trying to construct patterns for a grammar checking agent that will help ESL writers to correct their writing.

The list below shows a number of simple sentences without a "the" in the first noun phrase. The incorrect sentences are in italic and need a "the" where the $*$ is.

Assume that the grammar checking agent can tag words with their part of speech and that the lexicon contains information about categories of words, such as

- whether a noun is singular or not, and plural or not.
- whether a noun is a mass noun (something you can have an amount of).
- whether a noun is countable or not (something you can have 5 of).
- whether a noun can be labelled with a number (*e.g.*, "section 5").

Construct a set of patterns for identifying sentences that are missing a "the", such as the ones below. Make sure you explain any constraints used in your patterns.
Note, your patterns should **NOT** match any correct sentences.

```
Give apples to John
Give big apples to John
```
Give $*$ apple to John

Give $*$ red apple to John
```
Give sheep to John
He read section 5 to the teacher
```
He read $*$ section to the teacher
```
He read sections to the teacher
She checked theorem 6.8 carefully for errors
```
She checked $*$ theorem carefully for errors
```
Give apple juice to John
Give apple pies to John
```

Question 8. Robot Agents: The Mars Rover [27 marks]

Unlike classical planners, the planner for the Mars Rover had to take duration and resource usage of actions into account.

(a) [3 marks] Why were duration and resource usage important?

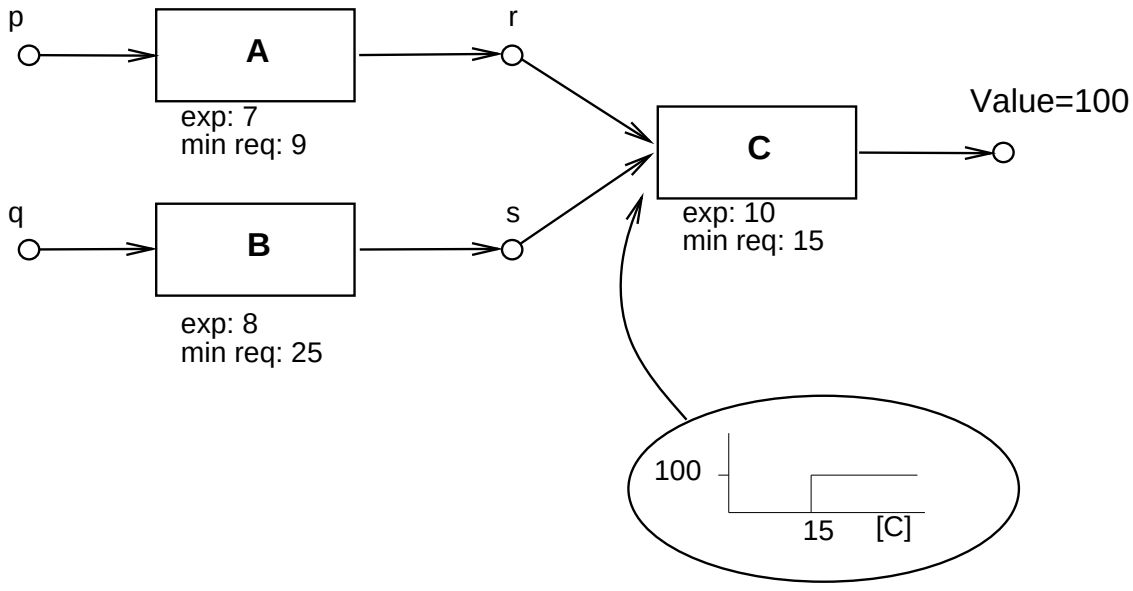
(b) [3 marks] Explain how the planner represented the duration and resource usage of an action.

(c) [6 marks] The planning algorithm for the Rover built a conditional plan that included not only the ideal plan, but also alternative branches, along with conditions for taking the alternative branches. What constraints made them take this approach to planning instead of simply replanning from scratch when it found it had insufficient resources for the next action?

(d) [15 marks] Part of the planning algorithm involved "regressing" resource constraints backwards from the goal states. For this algorithm, the constraints were represented as the expected output value of a sequence of actions as a function of resources.

Given the following plan graph and resource constraints (expected cost for each action and minimum resource required to start action), use their algorithm to regress the constraints to identify the best sequence of actions, and the minimum resource level required to achieve the outcome of 100 units. Show your working.

Note: The precondition for action "C" is "r & s". The first "regression" step is done for you.



****