

CS 276B: Web Search and Mining
Open Book Midterm Examination
Tuesday, February 15, 2005

This midterm examination consists of 6 pages, 5 questions and 100 points and counts for 20 percent of your final grade. Please write your answers on the exam paper in the spaces provided. You may use the back of a page if necessary. You have 75 minutes to complete the exam.

Stanford University Honor Code:

I attest that I have not given or received aid in this examination, and that I have done my share and taken an active part in seeing to it that others as well as myself uphold the spirit and letter of the Honor Code.

Name (printed): _____

Signature: _____ SUID: _____

Question	Score	Possible
1. Pagerank		10
2. Pattern-based extraction		10
3. Link analysis		20
4. Parallel crawling		30
5. Named entity recognition		30
Total		100

The standard of academic conduct for Stanford students is as follows:

- A. The Honor Code is an undertaking of the students, individually and collectively:
 - 1. that they will not give or receive aid in examinations; that they will not give or receive unpermitted aid in class work, in the preparation of reports, or in any other work that is to be used by the instructor as the basis of grading;
 - 2. that they will do their share and take an active part in seeing to it that they as well as others uphold the spirit and letter of the Honor Code.
- B. The faculty on its part manifests its confidence in the honor of its students by refraining from proctoring examinations and from taking unusual and unreasonable precautions to prevent the forms of dishonesty mentioned above. The faculty will also avoid, as far as practicable, academic procedures that create temptations to violate the Honor Code.
- C. While the faculty alone has the right and obligation to set academic requirements, the students and faculty will work together to establish optimal conditions for honorable academic work.

1. Pagerank (10 points)

We focus on two web pages A and B, with no hyperlink from A to B. (There are many other web pages of course.) Let $Old(B)$ be the pagerank of B. We now add a hyperlink from A to B. Intuitively, the new pagerank of B, $New(B)$, should be higher than $Old(B)$. Briefly describe a scenario in which this is not obvious by inspection. In other words, why is this not trivial to prove?

2. Pattern-based Extraction (10 points)

Suppose you have a named entity recognizer that can recognize company and place names. You want to build on top of that an information extraction system that extracts the relation **headquarters-located-in(company, location)**. Using the extraction pattern notation of RAPIER, hand-write a pattern that will extract instances of this relation. You may assume that company and place names have already been chunked together as a single unit. The rule should cover the three positive instances below:

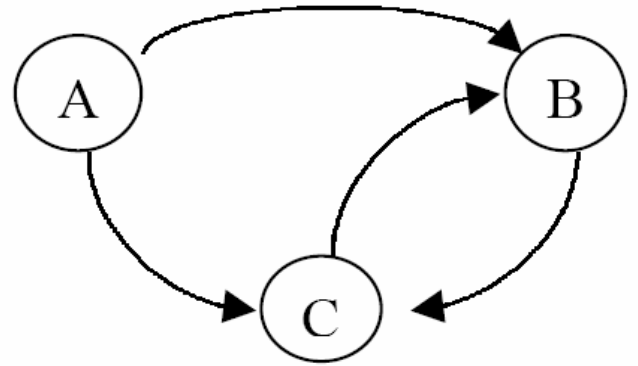
Covance, with headquarters in Princeton, New Jersey, is one of the world's largest and most comprehensive drug development services companies.

AmerisourceBergen is headquartered in Valley Forge, PA, and employs more than 14,000 people.

Camden National Corporation, headquartered in Camden, Maine, is the holding company for a family of three financial services companies.

3. Link analysis (20 points)

Consider a small web with 3 pages A, B and C. A links to B and C, while B links to C and C links to B. Compute pagerank, hub and authority scores for each of the three pages. Also give the relative ordering of the 3 nodes for each of these scores, indicating any ties.



Pagerank:

Assume that at each step of the pagerank random walk, we teleport to a random page with probability 0.1, with a uniform distribution over which particular page we teleport to.

Hubs/Authorities:

Normalize the hub scores so that the maximum hub score is 1.

Normalize the authority scores so that the maximum authority score is 1.

Hint 1: using symmetries to simplify and solving with linear equations might be easier than using iterative methods.

Hint 2: for partial credit, provide at least the relative ordering (indicating any ties) of the three nodes for each of the three scoring measures.

4. Parallel crawling (30 points)

Recall the treatment of Cho/Garcia-Molina covered in class (see lecture 7). Provide brief (one or two sentences) responses to the following questions.

(a) In firewall mode, why does it hurt to have more c-procs? What explains the relatively modest improvement in coverage as the number of seeds grows, following the initial jump?

(b) In firewall mode, what would be the effect on coverage of demanding more freshness in firewall mode?

(c) In crossover mode, for 4 c-procs to attain coverage close to 1, the overlap approaches 2.5. At this data point, what is the average number of times each page is downloaded?

(d) In exchange mode, why does URL hash result in a much higher communication overhead than site hash?

(e) In exchange mode, where is the impact of URL exchange high – on network bandwidth or on system resources at the sender/receiver?

(f) In exchange mode, why doesn't communication overhead increase linearly as the number of exchanged URLs grows?

5. Named entity recognition (30 points)

Consider the following HMM named entity recognition problem. The model has 4 states for P(refix), S(uffix), O(ther) and E(ntity). We have trained/built it to recognize locations. Well, really precisely one location, *Massachusetts*.

Here are the A and B matrices and the Π vector for the HMM:

A	P	S	O	E		Π	
P	0	0	0	1.0		P	0.05
S	0	0	1.0	0		S	0
O	x	0	$(1-x)$	0		O	y
E	0	1.0	0	0		E	$(0.95-y)$

B	Massachusetts	lived	in	Lovely Bay	State	...
P	0.01		0.01	0.2	0.01	0.01
S	0.01		0.01	0.1	0.01	0.1
O	0.1		0.1	0.1	0.01	0.01
E	1.0		0	0	0	0

Two parameters of the model have been left unspecified above. Consider the text:

Massachusetts lived in Massachusetts

where the intended interpretation is that the first “Massachusetts” refers to a Native American tribe, and the second to the location Massachusetts.

a) What are the conditions on the values of the parameters x and y that will lead each of the first and second occurrence of “Massachusetts” to be tagged as a location versus something else (when tagging using the Viterbi algorithm)?

b) The Massachusetts state poem begins thus:

Massachusetts Massachusetts
Lovely Bay State by the sea
Chosen by the Pilgrim Fathers
In their search for liberty

The above HMM will not correctly tag the two instances of the location Massachusetts above (regardless of how x and y are set). Why not? Modify the parameters of the HMM so that the correct tagging is achieved (just ignore things past the first 5 words of the poem).