CSE 494: Information Retrieval, Mining and Integration on the Internet

Midterm. October 23, 2002 (Instructor: Subbarao Kambhampati)
In-class Duration: Duration of the class 1hr 15min (75min)
For At home version: Due date: 10/28; 12:15pm in class. No exceptions.

Total points: 75 (a minute a point—what a deal!)

Name:______________________________ Student ID:__________________________

There are 11 pages, including the front page, in this exam. Open book; open notes; No web-access; no discussions with others.

Must be answered on this document, in the space provided (*answers on separate ruled sheets etc won't be accepted)*. If you need more space, you may use the backs of the sheets (but then put a note so I won't miss them).

Calculators are not really needed. However, if you forgot your number-work, you may use *basic* calculator (with normal arithmetic plus sqrt and log functions).

Qn I (Vector Space Ranking) 13pt	
Qn II(A/H; PageRank)13pt	
Qn III (LSI) 18pt	
Qn IV Clustering 8pt	
Qn V Short-answer 23pt	

For the At-home version, please read and sign the following declaration:
"*I declare, under the penalty of academic dishonesty, that I have not (a) discussed this exam with* <u>anyone</u> *other than the instructor since 1:30pm on October 23*rd *and (b) I have not used any source other than the notes, papers and homeworks distributed in the class".*

*Signed:*________________________ *date:*______________________________

Qn I. [In the following, you must SHOW YOUR WORK to get partial credit] Assume that the total number of documents in a corpus is 1024 and that the following words occur in the following number of documents:

“Computer” occurs in 32 documents
“software” occurs in 8 documents
“intelligent” occurs in 16 documents
“robust” occurs in 1024 documents

1. [6pt] Calculate the TF-IDF weighted term vector for the following document D. Assume that the log in the idf weight is taken to the base 2. (Hint: all the numbers above are powers of 2).

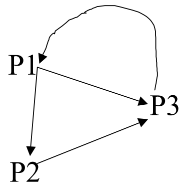
“Computer intelligent software robust computer software”

2.[4pt]Suppose I have a query Q which is specified as
“Intelligent Software”

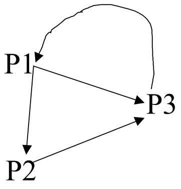
Assuming that query vector is computed just in terms of TF weights (no IDF weights), and similarity is measured by the cosine metric, what is the similarity between Q and D?

3.[3pt] Suppose the user is shown D in response to the query Q, and the user says that D is relevant to his query. If we now use relevance feedback to modify Q, what will the query vector become? Assume that alpha, beta and gamma are all 1.

Qn II. Suppose we have 3 web pages p1, p2 and p3, such that p1 has links to p2 and p3; p2 has link to p3 and p3 has link to p1 (see the picture)



(a) [6pt] Show one iteration of authorities and hubs algorithm. Assume you set all the authorities and hub values to 1 in the beginning. Show all the steps.



(b) [5pt] Show the augmented transition matrix, that will be used by the PageRank algorithm , assuming that with **c** probability a random surfer will follow the links on the current page, and that with **(1-c)** probability she will transition to any of the (three) pages with uniform probability; where **c** is set to 0.8

(c) [2pt] Suppose we set **c** to 0, then what will be the page ranks associated with the three pages?

Qn III. Consider the following T-D matrix defining 6 documents defined in terms of 4 keywords.

	D1	D2	D3	D4	D5	D6
Bush	5	15	7	9	7	0
Kalahari	5	7	1	0	1	0
Iraq	1	0	7	4	6	0
Saddam	0	1	6	4	0	4

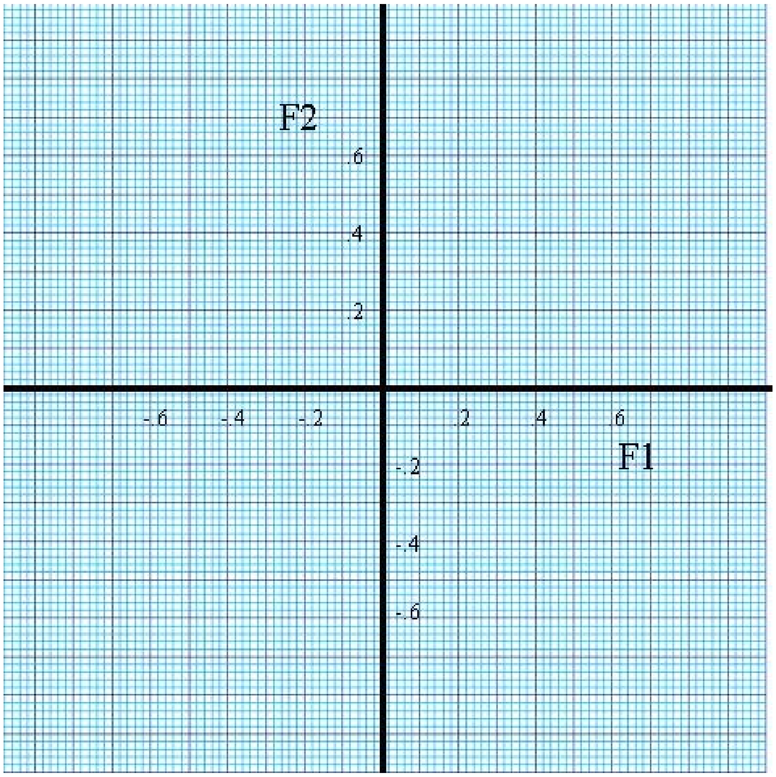
We decide to reduce the noise and dimensionality of this data through SVD analysis The SVD of this T-D matrix, according to MATLAB is: USV^t where U,S,V are given by:

```
 0.8817    0.1969   -0.0444   -0.4264
  0.2887    0.4928    0.1190    0.8122
  0.3033   -0.6652   -0.5674    0.3790
  0.2173   -0.5253    0.8136    0.1222

23.33        0         0         0         0         0
     0    9.76         0         0         0         0
     0         0    5.03         0         0         0
     0         0         0    3.27         0         0


0.2638    0.6627    0.4237    0.4293    0.3549    0.0373
 0.2850    0.6018   -0.6079   -0.3061   -0.2171   -0.2151
-0.0385    0.1948    0.1425    0.1162   -0.7138    0.6460
 0.7038   -0.1795    0.3700   -0.5590    0.0308    0.1491
 0.5557   -0.3294   -0.1526    0.6077   -0.3198   -0.2965
-0.2090    0.1411    0.5201   -0.1635   -0.4629   -0.6519
```

(1) [3pt] Suppose we are willing to sacrifice upto a maximum of 10% of the total variance in the data, then what is the least number of dimensions we need to keep? Explain how you arrived at your answer.

(2) [4pt] Suppose we decided to just keep top two most important dimensions after the LSI analysis. Draw a bounding box around the parts of U,S,V matrices above that will be retained after this decision. [You answer this question by directly marking the matrices above]

(3) [6pt] Suppose the two most important dimensions after LSI are called *f1* and *f2* respectively. Plot the six documents as points in the factor space (use the plot below). (It is okay if you put the points in the rough place they will come; no need to spoil your eyesight counting all the small grid lines). *Comment on the way the documents appear in the plot—is their placement related in any rational way to their similarity you would intuitively attach to them?*



(4) [5] What is the vector space similarity between D5 and D6 *before* and *after* the LSI transformation (assume, in the latter case, that we are using the top two dimensions). Is the change intuitively justified?

Qn IV Suppose you have a set of documents that basically contain only one key word, repeated multiple times. Suppose the dissimilarity between the documents is judged in terms of the difference in frequency of occurrence of that single keyword. The documents are given by:

D1: 5
D2: 7
D3: 9
D4: 14
D5: 15

(a)[6pt] Suppose we want to use K-Means algorithm to cluster this data into 2 clusters. Show how the clustering progresses, if you start K-means off with D3 and D4 as the seeds. What is the cumulative intra-cluster dissimilarity measure for the final clustering?

(b)[2pt] Suppose we are allowed to vary K, the number of clusters that K-means looks for. What is the lowest intra-cluster dissimilarity measure that can be achieved this way? When will it happen?

Qn V. Short answer questions. Except for the first question, all other questions carry 3 points.

1. [5pt] Suppose the number of keywords (size of vocabulary) is V, the average length of a document (in terms of words) is N, the number of documents in the corpus is M, the average length of a query is Q, and the average number of documents in which a query word appears is B. What is the time complexity, in vector-space retrieval, of: (a) Naïve query processing (without inverted index) and (b) query processing with inverted index. Why is b better?

2. [3] In the class, I mentioned that one way of making A/H computation more stable is to define the page importance in terms of subspaces rather than eigen vectors of the adjacency matrix. Explain how/why this is supposed to help. (Short answer in terms of examples is enough)

3. [3] In the class, someone asked why Google doesn’t remove all sink nodes (i.e., nodes that do not have any outlinks) from the page graph altogether before computing the page rank. What useful capability of Google will be lost if this were to be done?

4. [3] We talked about “stemming” as a technique that many text retrieval systems use. Comment on how stemming affects the precision and recall (i.e. improve/worsen)

5. [3] The HITS analysis assumes that all outlinks on a page are relevant to the given query. In many cases, however, even pages that are among the top K in terms of their similarity to a query Q, may have links to pages that have nothing to do with that particular query. Give a technique that can offset this problem

6. [3] Give one good reason why we shouldn’t replace cosine-metric with the inverse of eucledian distance (between query and document vectors) for deciding query-document similarity.

7. [3] List four (4) magic parameters that Google uses (a magic parameter is a number that needs to be set by Brin & Page—or their underlings).