

Name	Surname	Matricola	Signature

Web Mining course, second semester, Academic year 2007-2008

Written exam

Mauro Brunato

Elisa Cilia

June 25, 2008

- Please write name, surname, matricola and signature in the appropriate box above. Teachers and assistants can request a document for identification.
- The solution and procedure must be written in the same sheet containing the exercise text; if needed, use the back. All papers must be handed back at the end.
- No exercise requires complex calculations, only additions, multiplications by one- or two-digit numbers and simple fractions.
- It is important that the procedure is reported as well as the result. Unjustified results will not be taken into account.
- Every exercise is worth a maximum of 10 points.
- Use of notes, books, computers, calculators and PDAs is forbidden.

Exercise 1

Let V_1 and V_2 be two finite sets. Then the set of edges in a complete directed bipartite graph having V_1 as source nodes and V_2 as destination nodes is the Cartesian product of the two sets:

$$V_1 \times V_2 = \{(i, j) : i \in V_1 \wedge j \in V_2\}$$

Let us define graph $G = (V, E)$ where:

$$\begin{aligned} V &= \{1, \dots, 12\} \\ E &= (\{1, 2, 3\} \times \{4, 5, 6\}) \cup (\{5, 6\} \times \{7, 8\}) \cup (\{9, 10\} \times \{11, 12\}). \end{aligned}$$

The three subsets of E identify three bipartite components of G :

$$\begin{aligned} G_1 &= (\{1, \dots, 6\}, \{1, 2, 3\} \times \{4, 5, 6\}) \\ G_2 &= (\{5, \dots, 8\}, \{5, 6\} \times \{7, 8\}) \\ G_3 &= (\{9, \dots, 12\}, \{9, 10\} \times \{11, 12\}) \end{aligned}$$

Note that the three components are not disjoint, but the graph is not connected.

For every node n , according to the HITS scoring system, let $h(n)$ be its *hub score* and let $a(n)$ be its *authority score*. Moreover, if $B = (V_B, E_B)$ is a bipartite graph, let us define its *importance* $I(B)$ as the sum of hub scores of its source nodes plus the sum of the authority scores of its destination nodes:

$$I(B) = \sum_{i: \exists j (i, j) \in E_B} h(i) + \sum_{i: \exists j (j, i) \in E_B} a(i).$$

1.1) Which bipartite component (among G_1 , G_2 and G_3) will asymptotically achieve the maximum importance, and why?

1.2) Simulate three iterations of the HITS system starting with a uniform value of 1 to all hub and authority scores. What is the importance of each bipartite component, at the end?

1.3) If the edge $(3, 9)$ is added to G , how do you expect the importance scores of the three components to change, and why?

1.4) With the further addition of edge $(10, 3)$ to the graph, how do you expect the importance scores of the three components to change, and why?

Exercise 2

The singular value decomposition of a term-document matrix $A = U\Sigma V^T$ is

$$U = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & \sqrt{2} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad V = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 & 0 & -1 \\ 1 & 1 & 1 \\ 0 & 1 & -1 \\ 1 & -1 & 0 \end{pmatrix}$$

- 2.1) What are the size and the rank of the matrix A ?
- 2.2) Perform a reduction of the LSI space of one dimension.
- 2.3) Given the new representation of matrix \hat{A} , apply an agglomerative clustering procedure to the collection. Merge the clusters at each step according to the *self-similarity* measure by using as a measure of inter-document similarity simply the dot-product $\langle d_1, d_2 \rangle$.
- 2.4) Draw the resulting dendrogram. How many clusters can you find at a level of similarity of 2?
- 2.5) Check the clustering results you get by cutting across the dendrogram, by plotting them.

Exercise 3

As a part of a clustering method, we decide to compute for each document d the number of occurrences of the most frequent term in that document:

$$N(d) = \max_{t \in T} n(t, d).$$

We want to model $N(d)$ as a Poisson random variable with parameter λ , so that given a random document d in our corpus we have

$$\Pr(N(d) = n; \lambda) = \frac{\lambda^n e^{-\lambda}}{n!}$$

3.1) Given a sampled document set d_1, \dots, d_k , show that the maximum likelihood estimate of λ in the Poisson model is the average of $N(d_i)$.

3.2) Determine λ based on the following sample:

- $d_1 = (1, 2, 3, 4, 3, 2, 3, 3, 2, 1, 5, 6, 3)$
- $d_2 = (3, 2, 4, 4, 2, 3, 2, 4, 5, 1, 6)$
- $d_3 = (6, 4, 3, 5, 2, 6, 1, 7)$
- $d_4 = (6, 5, 4, 3, 2, 1, 2, 6, 2, 2)$
- $d_5 = (4, 3, 4, 2, 5, 4, 1, 6, 3)$