

Written exam

Mauro Brunato

Elisa Cilia

June 25, 2008

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.

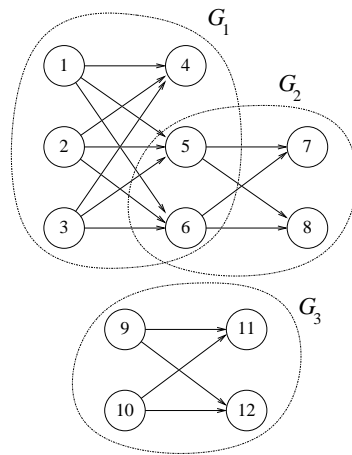
For every node n , according to the HITS scoring system, let $h(n)$ be its *hub score* and let $a(n)$ be its *authority score*. Moreover, if $B = (V_B, E_B)$ is a bipartite graph, its *importance* $I(B)$ is the sum of hub scores of its source nodes plus the sum of the authority scores of its destination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

Solution 1

The initial graph is the following (the three bipartite components are also shown):



1.1) The HITS ranking system favors the largest bipartite component, which corresponds to the principal eigenvector of $E^T E$. Therefore, component G_1 will asymptotically prevail.

1.2) Authority scores:

Node	1	2	3	4	5	6	7	8	9	10	11	12
Initial value	1	1	1	1	1	1	1	1	1	1	1	1
Step 1	0	0	0	3	3	3	2	2	0	0	2	2
Step 2	0	0	0	9	9	9	4	4	0	0	4	4
Step 3	0	0	0	27	27	27	8	8	0	0	8	8

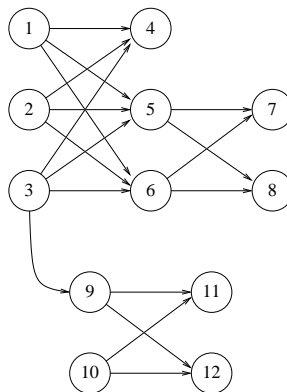
Hub scores:

Node	1	2	3	4	5	6	7	8	9	10	11	12
Initial value	1	1	1	1	1	1	1	1	1	1	1	1
Step 1	3	3	3	0	2	2	0	0	2	2	0	0
Step 2	9	9	9	0	4	4	0	0	4	4	0	0
Step 3	27	27	27	0	8	8	0	0	8	8	0	0

Hub scores:

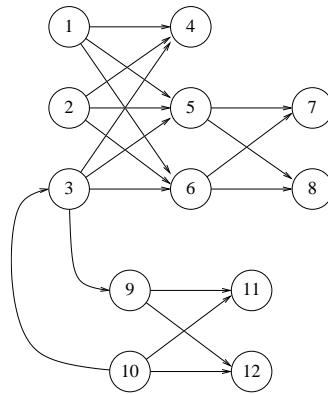
Component	G_1	G_2	G_3
Initial value	6	4	4
Step 1	18	8	8
Step 2	54	16	16
Step 3	162	32	32

1.3) After the new edge, the graph is the following:



Due to the current hub value of node 3, the authority value of node 9 increases, and in turn also the hub value of node 3 will increase, and therefore the authority values of nodes 4, 5, and 6. Therefore, the new edge causes $I(G_1)$ to increase. On the other hand, the authority value of node 9 does not impact on $I(G_3)$, where it is a source, and for the same reason the new edge has no impact on $I(G_2)$.

1.4) Finally, after the addition of the last edge:



The edge impacts on the authority score of node 3, therefore $I(G_1)$ and $I(G_2)$ do not change, and the hub score of node 10 (and hence the authorities of nodes 11 and 12) is increased. Therefore, the new edge only impacts on $I(G_3)$.

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 is 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.

Solution 2

- 2.1) Matrix A was originally a 3×4 matrix. The three elements in the diagonal matrix Σ are non-null, therefore matrix $A^T A$ (hence, matrix A) has full rank (3).
- 2.2) Let us obtain \hat{U} , \hat{V} and $\hat{\Sigma}$ by removing the third column from U and V , and the third row and column from Σ , corresponding to the smallest eigenvalue of $A^T A$:

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

2.3) After rank reduction, we can compute the similarity by

$$\hat{A}^T \hat{A} = \hat{V} \hat{\Sigma}^2 \hat{V}^T = \frac{1}{3} \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 9 & 0 \\ 0 & 4 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & -1 \end{pmatrix}$$

Therefore, we obtain the following table of unnormalized dot product similarities:

	2	3	4
1	3	0	3
2		$\frac{4}{3}$	$\frac{5}{3}$
3			$-\frac{4}{3}$

2.4) $\{1, 2\}$ and $\{1, 4\}$ are both candidates as the first cluster. Let us choose the first pair. Therefore, at level 3 the first clustering step yields

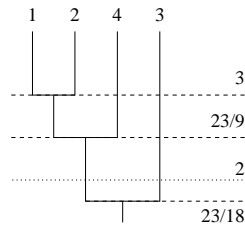
$$\begin{array}{c|cc} & 3 & 4 \\ \hline \{1, 2\} & \frac{13}{9} & \frac{23}{9} \\ 3 & & -\frac{4}{3} \end{array}$$

Now, the highest self-similarity value is achieved by cluster $\{1, 2, 4\}$ at level $\frac{23}{9}$, so that the similarity matrix becomes

$$\begin{array}{c|c} & 3 \\ \hline \{1, 2, 4\} & \frac{23}{18} \end{array}$$

Therefore, at similarity level 2 we have two clusters: $\{1, 2, 4\}$ and 3.

2.5) The corresponding dendrogram is



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:

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

Solution 3

3.1) The likelihood of λ with respect to the sample set is

$$L(\lambda; N(d_1), \dots, N(d_k)) = \Pr(N(d_1), \dots, N(d_k); \lambda) = \prod_{i=1}^k \Pr(N(d_i); \lambda) = \prod_{i=1}^k \frac{\lambda^{N(d_i)} e^{-\lambda}}{N(d_i)!}.$$

Therefore, the log-likelihood is

$$\log L(\lambda : N(d_1), \dots, N(d_k)) = \sum_{i=1}^k (N(d_i) \log \lambda - \lambda - \log N(d_i)!),$$

and its derivative with respect to λ is

$$\frac{d}{d\lambda} \log L(\lambda; N(d_1), \dots, N(d_k)) = \sum_{i=1}^k \left(\frac{N(d_i)}{\lambda} - 1 \right) = \frac{1}{\lambda} \sum_{i=1}^k N(d_i) - k.$$

Equating the derivative to zero, we get

$$\lambda = \frac{1}{k} \sum_{i=1}^k N(d_i).$$

3.2) By counting:

$$N(d_1) = 5, \quad N(d_2) = 3, \quad N(d_3) = 2, \quad N(d_4) = 4, \quad N(d_5) = 3,$$

the maximum likelihood estimate is the average of these values over the 5-document sample:

$$\hat{\lambda} = \frac{5 + 3 + 2 + 4 + 3}{5} = \frac{17}{5}.$$