Assignment 7 - Solutions

Exercise 1 (IIR 21) [1 P.]

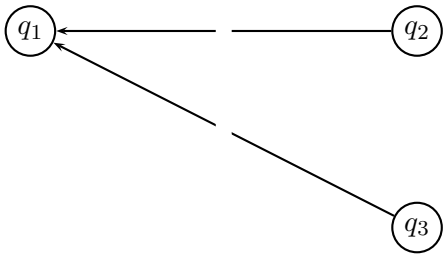
What is ergodicity and why is it important for PageRank?

Solution

ergodic = aperiodic (no periodic behavior) and irreducible (roughly: there is a path from every page to every other page)

PageRank is well-defined if surfing the web graph is ergodic

Exercise 2 (IIR 21) [3 P.]



For the web graph in the figure, compute PageRank, hub and authority scores for each of the three pages. Also give the relative ordering of the 3 nodes indicating any ties.

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. Normalize the hub and authority scores so that the maximum hub/authority score is 1.

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

Solution

Since the in-degree of A is 0, the steady-visit rate (or rank) of A is $0.1 \cdot 1/3 = 1/30$ (from teleport). By symmetry, rank(B) = rank(C). Thus, rank(B)=rank(C) = 29/60.

PageRank, (1) Solution using power method

Transition matrix P' without teleport:

	q1	q2	q3
q1	0	0	0
q2	1	0	0
q3	1	0	0

Transition matrix P with teleport:

	q1	q2	q3
q1	1/3	1/3	1/3
q2	14/15	1/30	1/30
q3	14/15	1/30	1/30

For initialization: $(1/3, 1/3, 1/3)$:

$\vec{x}P^1$ 0.733333 0.133333 0.133333

$\vec{x}P^2$ 0.493333 0.253333 0.253333
$\vec{x}P^3$ 0.637333 0.181333 0.181333
$\vec{x}P^4$ 0.550933 0.224533 0.224533
$\vec{x}P^5$ 0.602773 0.198613 0.198613
$\vec{x}P^6$ 0.571669 0.214165 0.214165
$\vec{x}P^7$ 0.590332 0.204834 0.204834
$\vec{x}P^8$ 0.579134 0.210433 0.210433
$\vec{x}P^9$ 0.585853 0.207074 0.207074
$\vec{x}P^{10}$ 0.581822 0.209089 0.209089
$\vec{x}P^{11}$ 0.584240 0.207880 0.207880
$\vec{x}P^{12}$ 0.582789 0.208605 0.208605
$\vec{x}P^{13}$ 0.583660 0.208170 0.208170
$\vec{x}P^{14}$ 0.583137 0.208431 0.208431
$\vec{x}P^{15}$ 0.583451 0.208275 0.208275
$\vec{x}P^{16}$ 0.583263 0.208369 0.208369
$\vec{x}P^{17}$ 0.583376 0.208312 0.208312
$\vec{x}P^{18}$ 0.583308 0.208346 0.208346
$\vec{x}P^{19}$ 0.583349 0.208326 0.208326
$\vec{x}P^{20}$ 0.583324 0.208338 0.208338
$\vec{x}P^{21}$ 0.583339 0.208331 0.208331
$\vec{x}P^{22}$ 0.583330 0.208335 0.208335
$\vec{x}P^{23}$ 0.583335 0.208332 0.208332
$\vec{x}P^{24}$ 0.583332 0.208334 0.208334
$\vec{x}P^{25}$ 0.583334 0.208333 0.208333
$\vec{x}P^{26}$ 0.583333 0.208334 0.208334
$\vec{x}P^{27}$ 0.583334 0.208333 0.208333
$\vec{x}P^{28}$ 0.583333 0.208333 0.208333
$\vec{x}P^{29}$ 0.583333 0.208333 0.208333
$\vec{x}P^{30}$ 0.583333 0.208333 0.208333

$\Longrightarrow$ Ranking: $d_1 > d_2 = d_3$

PageRank, (2) Solution 2

d2 and d3 have the same PageRank x. Let y be the PageRank of d1 . We have:

$$\begin{aligned} \frac{28}{30}(2x) + \frac{1}{3}y &= y \\ \frac{28}{15}x - \frac{2}{3}y &= 0 \\ \frac{28}{15}x - \frac{2}{3}(1-2x) &= 0 \\ \frac{48}{15}x &= \frac{2}{3} \end{aligned}$$

$\Longrightarrow$

$$x = \frac{2 \cdot 15}{3 \cdot 48} = \frac{5}{24} = 0.20833333333333$$
$$y = 1 - 2 \cdot x = \frac{14}{24} = \frac{7}{12} = 0.5833333333333$$

HITS, Solution 1

matrix A

$$\begin{matrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{matrix}$$

matrix A^T

$$\begin{matrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{matrix}$$

matrix AA^T

$$\begin{matrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{matrix}$$

matrix A^TA

$$\begin{matrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{matrix}$$

$\vec{a} = (1\ 1\ 1)^T$
$(A^TA)\vec{a} = (2\ 0\ 0)^T$
$(A^TA)^2\vec{a} = (4\ 0\ 0)^T$
$(A^TA)^3\vec{a} = (8\ 0\ 0)^T$

$\vec{h} = (1\ 1\ 1)^T$
$(AA^T)\vec{h} = (0\ 2\ 2)$
$(AA^T)^2\vec{h} = (0\ 4\ 4)$
$(AA^T)^3\vec{h} = (0\ 8\ 8)$

After normalization: $\vec{a} = (1\ 0\ 0)$, $\vec{h} = (0\ 1\ 1)$
Authority ranking: $d1 > d2 = d3$
Hub ranking: $d2 = d3 > d1$

HITS, Solution 2

Authorities: $authority(d_2) = authority(d_3) = 0$ since nobody is pointing to these two pages. $authority(d_1) > 0$ since somebody is pointing to d_1, thus value greater zero. After normalization (there is no page with a greater authority) this value is 1.0.
Hubs: By similar reasoning: $hub(d_1) = 0$, $hub(d_2) = hub(d_3) > 0$.
There is no page with a hub score higher than d_2 and d_3, thus $hub(d_2) = hub(d_3) = 1$.

Exercise 3 (IIR 6) [3 P.]

One measure of the similarity of two vectors is the Euclidean distance between them: $|\vec{x} - \vec{y}| = \sqrt{\sum_{i=1}^{M}(x_i - y_i)^2}$. Given a query q and documents $d_1, d_2, \ldots$, we may rank the documents d_i in order of increasing Euclidean distance from q. Show (by a mathematical proof) that if q and the d_i are all normalized to unit vectors, then the rank ordering produced by Euclidean distance is identical to that produced by cosine similarities.

Solution

$\sum(q_i - w_i)^2 = \sum q_i^2 - 2\sum q_i w_i + \sum w_i^2 = 1 - 2\sum q_i w_i + 1 = 2(1 - \sum q_i w_i)$

(Note that for a normalized vector $\vec{x}$, we have: $\sum x_i^2 = 1$.)

Thus: $|\vec{q} - \vec{v}| < |\vec{q} - \vec{w}| \Leftrightarrow |\vec{q} - \vec{v}|^2 < |\vec{q} - \vec{w}|^2 \Leftrightarrow \sum(q_i - v_i)^2 < \sum(q_i - w_i)^2 \Leftrightarrow 2(1 - \sum q_i v_i) < 2(1 - \sum q_i w_i) \Leftrightarrow \sum q_i v_i > \sum q_i w_i \Leftrightarrow \cos(\vec{q}, \vec{v}) > \cos(\vec{q}, \vec{w})$

This proves that ordering normalized vectors according to increasing distance is the same as ordering them according to decreasing cosine similarity.

Exercise 4 (IIR 8) [3 P.]

An unranked document retrieval approach is tested on a test set that consists of 300 documents. In response to a query 200 documents are retrieved of which 170 docs are relevant to the query and 30 not relevant. From the entire test corpus 190 documents are considered to be relevant for the mentioned query.

(a) Calculate precision, recall, accuracy and (balanced) f-measure of the presented classifier.

(b) Why do we usually have to face a tradeoff between precision and recall?

Solution

(a)

	relevant	nonrelevant
retrieved	170	30
not retrieved	20	80

- $Precision = tp/(tp + fp) = 170/(170 + 30) = 0.85$
- $Recall = tp/(tp + fn) = 170/(170 + 20) \approx 0.895$
- $Accuracy = (tp + tn)/(tp + fp + fn + tn) = (170 + 80)/(170 + 30 + 20 + 80) \approx 0.83$
- F-measure $= 2PR/(P + R) \approx 0.87$

(b) Because different users have different needs. Some users want to get documents that match their query as exact as possible (precision). They do not want to read all the documents that are available for a certain topic (recall). On the other side there are people who want to obtain all documents related to a topic, e.g. lawyers who want to get all law documents related to drug possession. They need high recall.
With respect to an information retrieval system we can always achieve a recall of 1 when we retrieve the whole collection, but then precision will be very low. When we want high precision, we can do this by only returning the documents where we are very sure that they are relevant, in the extreme only 1 document. This will of course create a very low recall. In practice, we will never have these extreme behaviours, but we nearly always face a decision if we want to increase precision or recall.

Exercise 5 (IIR 13-16) [5 P.]

As we have seen in chapter 14 there exist several types of classification algorithms.

(a) List the classification algorithms we have seen in chapters 13, 14 and 15 and give their key properties.

(b) Usually, we have dealt with only 2 classes in our examples. What changes with respect to the classification algorithms in (a) do we need to make if we want to classify more than 2 classes?

(c) Explain the difference between classification and clustering.

Solution

(a) Linear: Naive Bayes [Probabilistic; Independence assumption: One feature is independent from other features], Rocchio [Calculates Centroids and assigns new documents the class of the nearest centroid], SVM [Large margin, uses support vectors to calculate a decision hyperplane between classes]
Non-linear: kNN [decision boundary consists of locally linear segments, no training needed]

(b) If we have e.g. 4 classes: Perform a first binary classification for c_1 and $\{c_2, c_3, c_4\}$. In the next step classify c_2 and $\{c_3, c_4\}$ etc.

(c) Classification is supervised, i.e. a classifier is trained on a labeled dataset. Clustering on the other side is unsupervised: It is carried out on unlabeled data.