Pràctica 5
Algorisme de PageRank

Índex

1 Introducció **1**

2 Modelitzant la WWW **1**

3 Matrius estocàstiques (altra vegada) **3**

4 Exemple treballat amb Scilab **6**

1 Introducció

Els algorismes de cerca a la web, com l'algorisme PageRank de Google, constitueixen excel·lents aplicacions de l'àlgebra matricial.

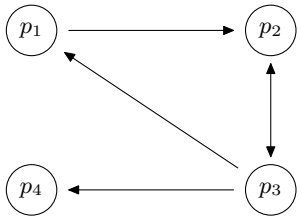
PageRank és el mètode de càlcul utilitzat pels fundadors de Google (Sergey Brin i Lawrence Page) per classificar les pàgines web d'acord amb el seu grau d'importància. L'objectiu del mètode és obtenir un vector, el vector PageRank, que proporciona la importància de les pàgines (és a dir, assigna, a cada pàgina web, un grau "d'importància", ordenant totes les pàgines d'acord amb aquest grau). Aquest vector es calcula a partir de la estructura de les connexions (enllaços) entre les pàgines web. Anem a descriure ací només els aspectes bàsics del mètode de PageRank, sense entrar en les modificacions i millores que s'han realitzat (i es continuen fent actualment).

2 Modelitzant la WWW

Suposem que voleu cercar el millor llibre d'Àlgebra Lineal. Probablement tractarieu de realitzar una cerca a Google. La llista de pàgines web que s'obtenen com a resultat de la cerca apareixen ordenades pel seu "grau d'importància". Aquest "grau d'importància" es defineix de manera que una pàgina és important si altres pàgines importants enllacen a ella. De manera

un poc més precisa: "una pàgina pot tenir més PageRank si hi ha moltes pàgines que enllacen a ella, o bé si hi ha un nombre suficient de pàgines importants que enllacen a ella". Però, com podem dir si una pàgina és important a partir de la pròpia importància de les pàgines? Aquesta és la qüestió que tractarem de resoldre.

Modelitzarem la World Wide Web (que és un recull de pàgines web connectades mitjançant enllaços) per mitjà d'un graf dirigit els vèrtexs del qual es corresponen amb les pàgines i els arcs representen els enllaços entre elles. Per a una millor comprensió del procés que explicarem a continuació, en lloc de treballar amb la totalitat de la WWW considerarem, com a model simplificat, una petita xarxa de pàgines els enllaços de la qual estan representats per mitjà del següent graf dirigit:



(La fletxa doble ha de ser interpretada com dos arcs, un en cada sentit).

La idea clau és que les pàgines haurien de ser més importants si, o bé són enllaçades molt sovint des d'altres pàgines, o bé són enllaçades des de la quantitat suficient de pàgines importants. És a dir, la importància d'una certa pàgina p_i (denotada per $I(p_i)$) ha d'augmentar cada vegada que és enllaçada des d'una altra pàgina p_j; a més a més l'increment ha de ser directament proporcional al "grau d'importància" de la pàgina p_j que l'enllaça i inversament proporcional al nombre d'enllaços a_j presents a la pàgina p_j (la importància total de la pàgina p_j es reparteix entre tots els seus enllaços). Tenint en compte això, té sentit la fórmula següent:

$$I(p_i) = \sum_{\text{pàgines } p_j \text{ que enllacen a } P_i} \frac{I(p_j)}{a_j}$$

En el nostre exemple:

- La pàgina p_1 té només un enllaç (a p_2) i, per tant, $a_1 = 1$.
- La pàgina p_2 té també només un enllaç i, per tant, $a_2 = 1$.
- La pàgina p_3 té 3 enllaços. Així $a_3 = 3$.
- La pàgina p_4 no té cap enllaç i, per tant, $a_4 = 0$.

Així, tenim el següent:

$I(p_1) = \frac{1}{3} I(p_3)$
$I(p_2) = I(p_1) + \frac{1}{3} I(p_3)$
$I(p_3) = I(p_2)$

$I(p_4) = \frac{1}{3} I(p_3)$

Observeu que açò és equivalent a la següent igualtat matricial:

$$\underbrace{\begin{bmatrix} I(p_1) \\ I(p_2) \\ I(p_3) \\ I(p_4) \end{bmatrix}}_{\vec{I}} = \underbrace{\begin{bmatrix} 0 & 0 & 1/3 & 0 \\ 1 & 0 & 1/3 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1/3 & 0 \end{bmatrix}}_{G} \underbrace{\begin{bmatrix} I(p_1) \\ I(p_2) \\ I(p_3) \\ I(p_4) \end{bmatrix}}_{\vec{I}} \tag{1}$$

Així doncs, el "vector d'importàncies" $\vec{I}$ (que és el que volem calcular) ha de satisfer la igualtat $G\vec{I} = \vec{I}$, on $G = [g_{ij}]$ és la matriu 4×4 tal que

$$g_{ij} = \frac{1}{\text{nombre d'enllaços de } p_j}$$

si p_j enllaça a p_i, i $g_{ij} = 0$ si p_j no enllaça a p_i. El vídeo que ve en el següent enllaç (en anglès) t'ajudarà a entendre tot això:

http://www.youtube.com/watch?v=ZstQKxUW7oM

La igualtat (1) vol dir que volem un *vector estacionari* $\vec{I}$ de la matriu G (és a dir, un vector que no canvia si el multipliquem per G). Volem, a més, que $\vec{I}$ siga un *vector de probabilitat*, que significa que totes les components de $\vec{I}$ siguen no negatives i sumen 1. Un vector $\vec{I}$ satisfent totes aquestes propietats *ordenaria* les pàgines assignant, a cadascuna d'elles, un nombre entre 0 i 1 que representa el seu "grau d'importància". Però, a més a més, hauria d'haver un **únic** vector amb aquestes característiques (atès que l'existència de dos vectors diferents significaria l'existència de dues ordenacions diferents de les pàgines, i no és això el que volem). Resumint, el que volem és que hi haja un vector $\vec{I}$ satisfent les següents propietats:

(1) $\vec{I}$ és un *vector estacionari* per a G, és a dir, $G\vec{I} = \vec{I}$,

(2) $\vec{I}$ és un *vector de probabilitat*, és a dir, els seus components són no negatius i sumen 1.

(3) $\vec{I}$ és l'únic satisfent (1) i (2).

Satisfà la matriu anterior G aquests requisits? En aquest cas la resposta és NO (es pot comprovar fàcilment que l'únic vector estacionari és $\vec{0}$). No obstant això, com veurem, podem modificar lleugerament (i de manera intel·ligent) la matriu G per forçar-la a satisfer les 3 condicions anteriors.

3 Matrius estocàstiques (altra vegada)

Recordem que una matriu $n \times n$ és *estocàstica* si totes les seves entrades són no negatives i la suma de les entrades en cada columna és 1. Les matrius estocàstiques són especialment interessants a causa de la següent propietat (que ja vam veure en la Pràctica 3):

Teorema 1. Qualsevol matriu estocàstica té, almenys, un vector de probabilitat estacionari.

Es segueix de la definició de G que la suma de les entrades en cadascuna de les seves columnes no nul·les és 1. No obstant això, l'última columna és nul·la i, per tant, G no és estocàstica; el problema és *l'existència de columnes nul·les*, i això passa perquè hi ha una pàgina (P_4) que no té cap enllaç (es correspon amb un vèrtex "pou" del graf dirigit). La manera més simple d'evitar aquest problema és imaginar que, quan algú està "navegant a través de la xarxa" i arriba a una pàgina com aquesta (pàgina pou), el que fa per continuar és triar una nova pàgina "totalment a l'atzar". Així doncs, podem pensar que una pàgina "pou" té, realment, un enllaç a cadascuna de les altres pàgines (i a ella mateixa). Per tant, podem redefinir la matriu G substituint les columnes nul·les per vectors $(1/n, 1/n, \ldots, 1/n)$, on n és el nombre total de pàgines (4, en el nostre cas):

$$G = \begin{bmatrix} 0 & 0 & 1/3 & 1/4 \\ 1 & 0 & 1/3 & 1/4 \\ 0 & 1 & 0 & 1/4 \\ 0 & 0 & 1/3 & 1/4 \end{bmatrix} \tag{2}$$

A causa del teorema anterior, aquesta matriu (i qualsevol matriu definida d'aquesta manera a partir de qualsevol xarxa de pàgines) satisfà les condicions (1) i (2). En aquest exemple som afortunats i, com vosaltres mateixos podeu comprovar, la condició (3) es satisfà també, però, en general, això no és veritat. Per tant, necessitem modificar una altra vegada la definició de G per assegurar-nos que el vector de probabilitat estacionari és sempre únic. El següent resultat (que ja vam veure en la Pràctica 3 amb més generalitat i que és una conseqüència del teorema de Perron-Fröbenius) ens proporciona la propietat que necessitem [1]:

Teorema 2. Si G és una matriu estocàstica que té totes les seves entrades estrictament positives llavors G té un **únic** vector de probabilitat estacionari.

La nostra matriu G (i en la pràctica, qualsevol matriu G obtinguda d'aquesta manera) té entrades nul·les i, per tant, no se li pot aplicar aquest teorema. No obstant això hi ha una manera coherent de modificar G per aconseguir que totes les seves entrades siguen estrictament positives:

Imaginem una persona que està "navegant" per la xarxa seguint els enllaços de les pàgines que es troba: cada vegada que ell/ella visita una pàgina p_i, va a una altra pàgina usant un dels enllaços de p_i; si p_i és una "pàgina pou" llavors ell/ella va a una pàgina aleatòria (escrivint una adreça aleatòria a la barra d'adreces). Potser, de tant en tant (i independentment de si la pàgina visitada en aquest moment és una "pàgina pou" o no ho és) ell/ella podria voler escriure, a la barra d'adreces, una adreça aleatòria en lloc de "seguir" els enllaços de la pàgina web actual. És a dir, podria "voler seguir", en lloc de la matriu G, la següent matriu (que

[1] L'enunciat i la prova del Teorema de Perron-Fröbenius fan servir, com a ingredients clau, els conceptes de valor propi i vector propi, que veurem en el Tema 6.

anomenarem **matriu d'aleatorietat**:

$$E := \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix};$$

això vol dir que, quan "es segueix" aquesta matriu, totes les pàgines web tenen la mateixa probabilitat de ser escollides.

Ara assumim el següent: cada vegada que un "navegant" visita una pàgina té dues possibilitats per actuar: "seguir" la matriu G (és a dir, continuar usant un dels enllaços de la pàgina web en la qual es troba) o bé "seguir" la **matriu d'aleatorietat** E (és a dir, triar una pàgina web a l'atzar).

Fixem un nombre real $\alpha \in]0,1[$ que denota la probabilitat de navegar "seguint" la matriu G. Llavors, la probabilitat de navegar "seguint" la matriu E és $1-\alpha$. α mesura, en certa manera, el "grau de llibertat" que tenim per navegar "seguint" cadascuna de les dues matrius. (Observa que, com estem interessats en donar molta més importància a la matriu G que a la matriu E, α hauria de ser un valor proper a 1). Aquesta nova situació equival a considerar, en lloc de la matriu G, aquesta altra:

$$\mathbf{G} = \alpha G + (1-\alpha) I.$$

És fàcil comprovar que aquesta matriu és estocàstica i que, a més a més, totes les seves entrades són estrictament positives. Llavors podem aplicar els teoremes que hem vist abans, deduint que existeix un vector $\vec{I}$ satisfent les condicions anteriors (1), (2) i (3). Aquest és el vector PageRank que estàvem buscant.

Una manera de calcular este vector $\vec{I}$ es resoldre el següent sistema d'equacions lineals[2]:

$$(\mathbf{G} - I_{4\times 4})\vec{x} = \vec{0}. \tag{3}$$

El paper del paràmetre α és molt important. Observa que si $\alpha = 1$ llavors $\mathbf{G} = G$. Això vol dir que estem treballant amb l'estructura d'enllaços original del web. No obstant això, si $\alpha = 0$ llavors $\mathbf{G} = E$; en altres paraules, estem considerant que qualsevol pàgina té un enllaç a qualsevol altra perdent-se, així, l'estructura original de la xarxa. Clarament ens agradaria que α fos un valor proper a 1 per donar molta importància a l'estructura original.

Però hi ha una altra consideració. A la pràctica, la matriu $\mathbf{G}$ corresponent amb la WWW és una matriu extremadament gran. Com a conseqüència d'això, resoldre un sistema com (3) utilitzant els mètodes habituals no és una bona idea [3]. En lloc d'això, en la pràctica, el vector $\vec{I}$ es calcula utilitzant un mètode iteratiu conegut com *m*ètode de la potència. *Essencialment*

[2] Observa que $\mathbf{G}\vec{x} = \vec{x} \;\Leftrightarrow\; \mathbf{G}\vec{x} = I_{4\times 4}\vec{x} \;\Leftrightarrow\; \mathbf{G}\vec{x} - I_{4\times 4}\vec{x} = \vec{0} \;\Leftrightarrow\; (\mathbf{G} - I_{4\times 4})\vec{x} = \vec{0}$

[3] L'acumulació d'errors d'arrodoniment és una raó. Una altra és que, com el nombre de pàgines web és tan elevat, la matriu E és "gairebé" la matriu nul·la

consisteix a calcular, per iteració, una bona aproximació al límit de la següent cadena de Markov [4] :

$$\vec{x}_0, \quad \vec{x}_1 = \mathbf{G}x_0, \quad vecx_2 = \mathbf{G}x_1, \ldots$$

$on\ _0$ és *qualsevol* vector de probabilitat inicial.

Ometrem aquí més explicació sobre aquest mètode excepte el fet que, quan el paràmetre α està " massa proper" a 1 la convergència d'aquest mètode és molt lenta. Sergey Brin i Larry Page, els creadors del PageRank, van triar un valor de α pròxim a $0,85$.

4 Exemple treballat amb Scilab

Com a exemple d'aplicació del mètode, anem a calcular el vector PageRank corresponent a l'exemple anterior amb l'ajuda d'Scilab. Primer introduirem la matriu G donada en (**??**) a partir del graf que descriu els enllaços, però substituint els zeros de les columnes nul·les per $1/n$, on n és el nombre de pàgines ($n = 4$ en el nostre cas):

```
-->G=[0 0 1/3 1/4; 1 0 1/3 1/4; 0 1 0 1/4; 0 0 1/3 1/4]
 G  =

    0.    0.    0.3333333    0.25
    1.    0.    0.3333333    0.25
    0.    1.    0.           0.25
    0.    0.    0.3333333    0.25
```

Ara definim la matriu E:

```
-->E=1/4*ones(4,4)
 E  =

    0.25    0.25    0.25    0.25
    0.25    0.25    0.25    0.25
    0.25    0.25    0.25    0.25
    0.25    0.25    0.25    0.25
```

Definim ara la *matriu Google* $\mathbf{G}$ agafant $\alpha = 0.85$:

```
-->G=0.85*G+(1-0.85)*E
 G  =

    0.0375    0.0375    0.3208333    0.25
    0.8875    0.0375    0.3208333    0.25
    0.0375    0.8875    0.0375       0.25
    0.0375    0.0375    0.3208333    0.25
```

[4] En la pràctica s'usa una versió modificada del *mètode de la potència* que permet usar només la matriu G en lloc de $\mathbf{G}$, evitant l'ús de E en els càlculs.

Resolem el sistema (3), és a dir, calculem el nucli de la matriu $\mathbf{G} - I_{4\times 4}$:

```
-->x=kernel(G-eye(4,4))
 x  =

    0.3254602
    0.6021013
    0.6523997
    0.3254602
```

L'únic vector de probabilitat que és solució del sistema (3) pot calcular-se fàcilment dividint el generador del nucli que hem obtingut per la suma dels seus components:

```
-->x/sum(x)
 ans  =

    0.1708075
    0.3159938
    0.3423913
    0.1708075
```

Este és el vector PageRank. Açò significa que les pàgines s'ordenen, en ordre d'importancia decreixent, de la següent manera:

$$Page\ 3$$

$$Page\ 2$$

$$Page\ 1$$

$$Page\ 4$$

En aquest cas, com només tenim 4 pàgines, és molt fàcil ordenar les pàgines "a mà". No obstant això, en casos amb un major nombre de pàgines, pot ser útil escriure la següent comanda de Scilab:

```
-->[w,k]=gsort(x/sum(x))
 k  =

    3.
    2.
    1.
    4.
 w  =

    0.3423913
```

```
0.3159938
0.1708075
0.1708075
```

El vector w que retorna escriu, en ordre decreixent, el PageRank de les pàgines. El vector k proporciona directament la llista ordenada de les pàgines.