Butlletí de la Societat Catalana de Matemàtiques
Vol. 26, núm. 1, 2011. Pàg. 29–55. DOI: 10.2436/20.2002.01.33

Les matemàtiques de Google: l'algorisme PageRank

Joan Gimbert

Resum: En aquest article presentem i analitzem l'algorisme PageRank, emprat per Google en l'ordenació dels seus resultats de cerca. La seva fonamentació teòrica ens duu a interrelacionar diferents parts de la matemàtica, com la teoria de matrius no negatives, la teoria de grafs i les cadenes de Markov. Cal dir que hi ha altres algorismes de valoració de pàgines web, basats en el còmput de vectors propis, com l'algorisme HITS, el qual exposem breument al final del treball.

Paraules clau: algorisme PageRank, matrius no negatives, grafs, cadenes de Markov, algorisme HITS.

Classificació MSC2010: 05C50, 15A51, 60J10.

1 Presentació

La primera vegada que vaig sentir parlar de la sorprenent relació entre els cercadors i els vectors propis va ser a la conferència «Search engines, eigenvectors, and chromatic numbers», del professor Herbert Wilf [35], que va impartir l'any 2001 en un congrés sobre combinatòria i teoria de grafs. Llavors feia tres anys de l'aparició de Google, desenvolupat per Sergey Brin i Larry Page [8], que ràpidament va anar guanyant «adeptes» per la millor qualitat dels seus resultats de cerca. Quin era el seu «secret»? Interpretar els enllaços com una concessió de rellevància, a partir de la qual varen definir una mesura de la valoració d'una pàgina, anomenada PageRank, atorgada únicament per la mateixa topologia de la Web.

Des d'aleshores s'ha investigat i publicat molt sobre l'algorisme PageRank, ja que ha esdevingut, juntament amb l'algorisme de clau pública RSA, un dels exemples paradigmàtics d'idees matemàtiques que han donat lloc a aplicacions i empreses exitoses. Destaquem el reconegut article divulgatiu de Pablo Fernández, «El secreto de Google y el álgebra lineal», publicat en el *Boletín de la Sociedad Española de Matemática Aplicada* (vegeu [14]), l'excel·lent llibre

d'Amy N. Langville i Carl D. Meyer, *Google's PageRank and beyond. The science of search engine rankings* (vegeu [23]), i els diferents volums de les actes dels congressos sobre *Algorithms and models for the web-graph*, dins la sèrie Lecture Notes in Computer Science (vegeu, per exemple, [2]).

Quan ja s'ha escrit molt i bé sobre un tema és difícil trobar quelcom de nou a dir. En aquest article hem procurat d'una banda exposar les idees subjacents més rellevants, i de l'altra, explicitar la interrelació i l'enriquiment mutu entre les diferents parts de la matemàtica que intervenen en la fonamentació i l'aplicació de l'algorisme PageRank, com són la teoria de matrius no negatives, la teoria de grafs i les cadenes de Markov, entre altres.

2 Breu introducció al món dels cercadors

La cerca d'informació en la Web ha suposat nous i difícils reptes, comparant-ho amb l'accés a una col·lecció estructurada de documents, com seria el fons d'una biblioteca. D'entre els trets diferencials de la Web, que condicionen el disseny dels corresponents enginys de cerca, destaca la possibilitat de navegació, proporcionada pel mateix llenguatge d'hipertext, el seu immens (i exponencialment creixent) volum de dades, que s'han d'aconseguir i tractar, així com el seu caràcter dinàmic i distribuït (vegeu l'anàlisi que fan Baeza i Ribeiro [4] d'aquests i d'altres trets característics). A més, cal tenir presents els problemes que sorgeixen en la interacció, desitjadament satisfactòria, entre l'usuari i el cercador. En aquest sentit cal plantejar-se, per exemple, com especificar una cerca i com ordenar-ne els resultats per tal que els primers siguin els més «rellevants». I a tot plegat hi hem d'afegir la immediatesa en la resposta als milers de consultes!

Certament, el disseny i la implementació d'un cercador a gran escala ha esdevingut un dels grans projectes de l'enginyeria informàtica. A continuació, descriurem breument els components d'un cercador amb la intenció de facilitar-ne una visió de conjunt.

2.1 Arquitectura d'un cercador

Com una primera aproximació, podríem dir que un cercador consta de tres grans mòduls (esquematitzats en la figura 1):

- *Robots rastrejadors de la Web*: es tracta de petits programes que, a través dels enllaços, visiten i recullen constantment noves pàgines, per ser analitzades i indexades, i actualitzen el contingut de les ja visitades. Aquestes pàgines són guardades en un *repositori* (mitjançant un sistema d'emmagatzematge escalable i distribuït).
- *Anàlisi i indexació de les pàgines*: primer es fa un buidatge del contingut de cada pàgina, en què s'extreuen les paraules amb informació relativa a la

seva localització i freqüència d'aparició,[1] així com d'enllaços,[2] indicant la pàgina d'origen i de destí de cadascun. A partir de tot aquest «escaneig» es creen diversos *índexs invertits*, semblants als que apareixen al final dels llibres però amb moltíssimes més entrades, per tal de facilitar la consulta posterior.

- *Consulta i ordenació dels resultats*: s'encarrega de construir la consulta a la base de dades, a partir de les paraules introduïdes per l'usuari, i d'ordenar les pàgines web pertinents a dites paraules d'acord amb una certa mètrica d'idoneïtat i rellevància.

Cadascun dels components anteriors està estructurat en diferents submòduls, segons la seva funcionalitat (vegeu la figura 2), els quals trobareu explicats amb detall en l'article «Searching the web» d'Arasu *et al.* [3].

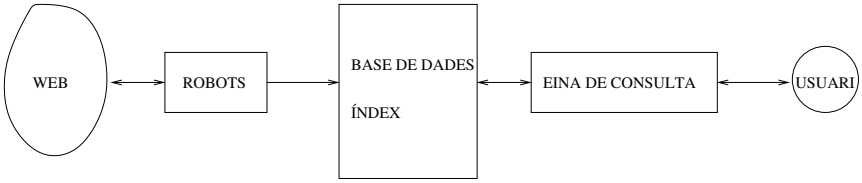

FIGURA 1: Mòduls bàsics d'un cercador.

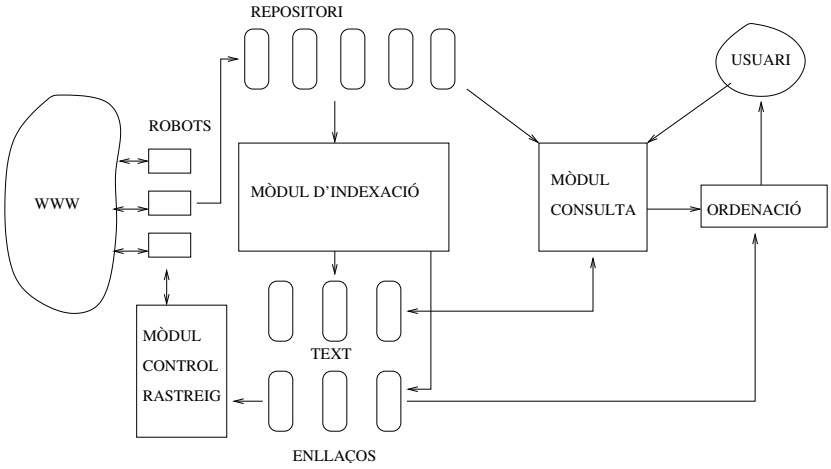

FIGURA 2: Submòduls d'un cercador, amb indicació de la seva funcionalitat.

1 A partir d'aquestes dades i d'altres com la mida de la font, s'introdueixen diferents mètriques per mesurar el grau de rellevància de cada paraula dins la pàgina en qüestió.

2 L'anàlisi dels enllaços va ser una novetat i una clau de l'èxit de GOOGLE, tal com comentarem posteriorment.

2.2 Naixement de Google

Durant els anys 1995-1998, Sergey Brin i Larry Page, llavors doctorands en ciències de la computació de la Universitat de Stanford, van estar desenvolupant un cercador que, d'una banda, pogués ser emprat a gran escala, i de l'altra, proporcionés a l'usuari resultats més satisfactoris. Varen batejar el seu cercador amb el nom de GOOGLE, tot jugant amb les lletres del terme *googol* que designa el nombre gegantí 10^{100} i amb el qual volien reflectir el volum extraordinari de dades on s'ha de trobar la informació desitjada (vegeu [8]).[3]

Quines varen ser les seves claus? Els cercadors existents fins a l'aparició de GOOGLE, com ara ALTAVISTA, basaven principalment els seus resultats de cerca en les coincidències entre les paraules buscades i les indexades dins de cada pàgina. No havien tingut en compte els enllaços entre les pàgines, tret característic dels documents d'hipertext. D'altra banda, els creadors de GOOGLE varen fer recaure l'escalabilitat del seu enginy en l'ús d'un gran nombre d'ordinadors connectats en xarxa, en lloc d'emprar unes quantes estacions de treball més potents, decisió que ha resultat ser del tot encertada. Aquestes dues característiques, en consonància amb la pròpia naturalesa de la Web, juntament amb l'optimització de les estructures de dades utilitzades (vegeu l'anatomia que Brin i Page [8] varen fer del seu prototip[4]), han esdevingut clau en l'èxit de GOOGLE.

D'ara endavant, ens centrarem en l'ús de l'estructura d'enllaços per millorar la qualitat dels resultats de les cerques. Els enllaços, entesos com a referències, concessions d'autoritat o simplement vots, permeten mesurar la rellevància de cada pàgina, tal com veurem en les seccions posteriors. Cal també mencionar el tractament diferenciat que fa GOOGLE de l'anomenat *text d'ancoratge*, aquell que es fa servir com a títol d'un enllaç. Aquest text, vist com el rètol d'una porta d'entrada, el considera com un descriptor de la pàgina destí[5] (a diferència del tractament rebut per la majoria dels altres cercadors que només l'atribuïen a la pàgina origen de l'enllaç). D'aquesta manera es pot oferir una nova descripció d'una pàgina. Ara bé, com que la publicació en la Web no està sotmesa, en general, a un procés de revisió, no es pot garantir que els enllaços ni el seu text siguin pertinents, és a dir, podrien haver estat creats amb la intenció d'enganyar el cercador, si se sospita que aquest fa un ús preferent d'aquesta característica. Així, la creació de molts enllaços cap a una mateixa pàgina, amb un text similar d'ancoratge, al llarg del temps podria fer que aquesta pàgina ocupés els primers llocs en els resultats de les consultes corresponents al text en qüestió. Mitjançant aquesta estratègia un grup d'internautes va aconseguir

3 També podríem pensar l'1 seguit de 100 zeros com una visualització de la idea de trobar una «agulla» en un immens «paller».

4 Val a dir que Brin i Page foren pioners en descriure en profunditat, i d'una manera pública, l'arquitectura d'un prototip de cercador, l'accés al qual varen deixar lliure. També, en aquest sentit, varen tenir una bona visió de futur.

5 Aquesta idea ja havia estat implementada en un dels primers cercadors, anomenat *World Wide Web Worm*, desenvolupat l'any 1993 per Oliver McBryan (vegeu [26]).

que, just després de la invasió de l'Iraq, la cerca al GOOGLE de la frase «Weapons of mass destruction», emprant l'opció «I'm Feeling Lucky», donés com a resultat una pàgina amb el missatge «Not Found» (vegeu [34]). Com veiem, doncs, de la mateixa manera que els criptògrafs tenen en ment els criptoanalistes quan creen els seus codis secrets, els dissenyadors d'un cercador també han de pensar a posar-ho ben difícil a aquells que intentaran enganyar-lo.

3 L'algorisme PageRank: versió simplificada

3.1 Introducció

En l'àmbit de les publicacions científiques, hi ha diferents maneres de mesurar l'impacte o rellevància d'una revista (vegeu [1]). L'indicador més simple es basa en comptar el nombre de cites rebudes pels articles publicats en aquesta revista. Aquest factor, en el cas de les pàgines web, correspondria al nombre d'enllaços rebuts per cada pàgina. Ara bé, el fet de no poder garantir la pertinença d'aquests enllaços, per les característiques pròpies de la publicació en la Web, obliga a cercar una mesura que tingui en compte la topologia de tota la Web.

Podríem pensar que una pàgina web és *important* en la *mesura* que ho siguin les pàgines que hi apuntin. La formalització més simple d'aquesta idea suposa que la valoració d'una pàgina és proporcional a la suma de les valoracions donades per les pàgines que l'enllacen. En el cas de l'algorisme PageRank, proposat per Brin i Page [8, 28], cada pàgina reparteix equitativament la seva valoració entre totes les pàgines a les quals apunta. D'aquesta manera, els enllaços vénen a ser com *vots ponderats*, on la ponderació és directament proporcional a la rellevància de la pàgina origen i inversament proporcional al nombre d'enllaços d'aquesta. Així, una pàgina pot ser important tot i rebre pocs vots, això sí provinents de pàgines importants. Notem el caire recurrent d'aquesta mesura de rellevància, la qual es va propagant per tota la Web. Més endavant, veurem com les valoracions del PageRank descriuen un cert *estat d'equilibri* en aquesta dinàmica de la Web i com, al mateix temps, capten el comportament d'un usuari que vagi navegant per la xarxa de manera aleatòria.

EXEMPLE 1. En la figura 3 mostrem dues hipotètiques estructures d'enllaços entre diferents pàgines web. Com ordenaríeu, en cada cas, les pàgines d'acord amb la seva importància (donada únicament pels enllaços)?

En el gràfic de l'esquerra (Γ_1) veiem com la pàgina 4 és enllaçada (votada) per totes les pàgines i, en conseqüència, hauria de ser la pàgina més important. Ara bé, com que tota la seva valoració la dóna a la pàgina 1, totes dues pàgines compartirien el primer lloc. De les dues pàgines restants, que tindrien una valoració menor ja que només reben una part de la valoració de la pàgina 1, la 3 hauria d'anar per davant de la 2 ja que rep un segon *suport*. Així, doncs, sense haver escrit cap equació hem pogut deduir l'ordenació de les pàgines. Difícilment, però, aquesta drecera la trobarem en altres situacions, com podem copsar en el gràfic de la dreta (Γ_2), on ni tan sols s'intueix, a cop d'ull, la pàgina

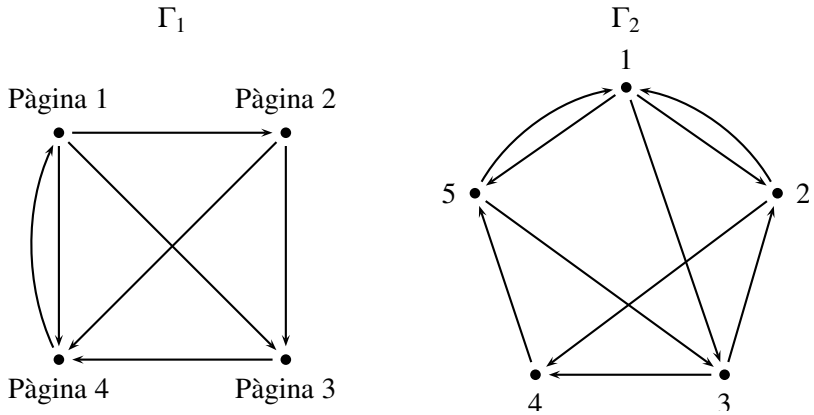

FIGURA 3: Dues hipotètiques topologies de pàgines webs, representades mitjançant grafs dirigits.

més important. Hauria de ser aquella que més cops es visités si simuléssim un bon grapat de recorreguts aleatoris per aquest diagrama de punts i arcs, anomenat *graf dirigit*.

3.2 Formulació del problema

De manera natural, la topologia de la Web pot modelitzar-se com un immens graf dirigit o digraf[6] $\Gamma = (V, A)$, on els seus vèrtexs representen les pàgines web i els seus arcs assenyalen els enllaços entre elles. Així, si $V = \{1, \ldots, n\}$ és el conjunt de pàgines web, un arc $(i, j) \in A$ indica que la pàgina i conté un enllaç cap a la pàgina j (direm que aquest enllaç té per origen la pàgina i i per destí la pàgina j).

Es vol assignar a cada pàgina i una valoració[7] $v_i > 0$ de manera que sigui proporcional a la suma ponderada de les valoracions de les pàgines que hi apunten, on la ponderació de cada sumand depèn del nombre total d'enllaços que conté la pàgina origen corresponent de l'enllaç. Així, les valoracions $v_1, \ldots, v_n$ han de satisfer les equacions lineals següents:

$$v_i = \lambda \sum_{(j,i)\in A} \frac{v_j}{n_j}, \qquad i = 1, \ldots, n, \tag{1}$$

on $\lambda > 0$ és el factor de proporcionalitat, inicialment desconegut, i n_j és el nombre total d'enllaços des de la pàgina j (en altres termes, el grau de sortida del vèrtex j).

6 Un *graf dirigit* o *digraf* $\Gamma = (V, A)$ està format per un conjunt finit $V := V(\Gamma)$, els elements del qual se'ls anomena *vèrtexs*, i per un conjunt $A := A(\Gamma)$ de parells ordenats d'elements de V, anomenats *arcs*.

7 El fet que representi una mesura de la importància justifica que es prengui com un nombre real no negatiu. D'altra banda, per tal de poder fer comparacions relatives convé que aquestes valoracions no siguin nul·les.

Per tal de formular el sistema (1) en termes matricials, definim $\boldsymbol{v} = (v_1, \ldots, v_n)^\top$ com el vector (columna) de valoracions i $\boldsymbol{P} = (p_{ij})$ com la matriu quadrada d'ordre n que té per coeficients:

$$p_{ij} = \begin{cases} \frac{1}{n_j}, & \text{si } (j,i) \in A; \\ 0, & \text{altrament.} \end{cases} \tag{2}$$

Podem pensar $\boldsymbol{P}$ com la «matriu normalitzada d'enllaços» de Γ, en el sentit que si una pàgina j té almenys un enllaç de sortida, és a dir, no és una *pàgina pou*, aleshores els coeficients no nuls de la columna j de $\boldsymbol{P}$, situats a les files de les pàgines on apunten aquests enllaços, sumen 1 (fet que es podrà interpretar en termes de probabilitats de transició).

D'aquesta manera,

$$\boldsymbol{v} = \lambda \boldsymbol{P}\boldsymbol{v}, \tag{3}$$

d'on resulta que $\boldsymbol{v}$ és un vector propi de $\boldsymbol{P}$ de valor propi $1/\lambda$. Però $\boldsymbol{v}$ no és un vector propi qualsevol, sinó un de «distingit» perquè té totes les seves components positives (direm que $\boldsymbol{v}$ és *vector positiu*, fet que denotarem per $\boldsymbol{v} > \boldsymbol{0}$). Tampoc hem de perdre de vista que la matriu $\boldsymbol{P}$ és una *matriu no negativa*, és a dir, amb tots els seus coeficients no negatius (si fossin tots positius diríem que la matriu és *positiva*).

EXEMPLE 2. Les matrius normalitzades d'enllaços corresponents als digrafs Γ_1 i Γ_2, mostrats en la figura 3, són

$$\boldsymbol{P}_1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 1 & 0 \end{pmatrix} \quad \text{i} \quad \boldsymbol{P}_2 = \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{3} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 1 & 0 \end{pmatrix},$$

respectivament.

3.3 Resolució del problema

Arribats en aquest punt, són diverses les qüestions que ens intriguen: *Podem assegurar que la matriu $\boldsymbol{P}$ té un vector propi positiu?* En cas d'existir, *és únic* (llevat d'un factor de proporcionalitat, és a dir, fixada la seva norma)? Més encara, *hi ha algun procediment eficient per calcular-lo?*

Totes les respostes a les preguntes anteriors les trobem en uns resultats clàssics, i molt rellevants, sobre les propietats espectrals d'una certa classe de matrius no negatives, anomenades *irreductibles*, teoremes que avui es coneixen sota el nom de *teoria de Perron-Frobenius*.[8] Aquesta doble atribució es deu al

8 En paraules de Carl D. Meyer, «The Perron-Frobenius theory is elegant. It is a testament to the fact that beautiful mathematics eventually tends to be useful, and useful mathematics eventually tends to be beautiful» (vegeu [27]).

fet que Oskar Perron, l'any 1907, va descobrir com la característica que una matriu fos positiva quedava reflectida en el seu espectre[9] i, posteriorment, Georg Frobenius, l'any 1912, va esbrinar quina era la condició, en termes de les posicions dels elements nuls d'una matriu no negativa, que permetia heretar la mateixa propietat.

Començarem enunciant el teorema de Perron, la demostració del qual pot consultar-se a [27].

TEOREMA 1 (PERRON). *Sigui $\boldsymbol{A} \in \mathcal{M}_n(\mathbb{R})$ una matriu positiva. Aleshores:*

1. *$\boldsymbol{A}$ té un valor propi positiu r per al qual hi ha un vector propi positiu.*
2. *r és l'únic valor propi de $\boldsymbol{A}$ que satisfà les condicions anteriors. A més,*

   2a. r té multiplicitat algebraica igual a 1.

   2b. r és un valor propi dominant de $\boldsymbol{A}$ i és l'únic de mòdul màxim; és a dir, $|\lambda| < r$, per a tot altre valor propi λ de $\boldsymbol{A}$.

Si s'admet que la matriu $\boldsymbol{A}$ pugui tenir coeficients nuls, llavors el seu radi espectral[10], $\rho(\boldsymbol{A})$, continua essent un valor propi per al qual hi ha un vector propi no negatiu (vegeu [17]). Ara bé, si es vol assegurar que aquest vector sigui únic (llevat d'un factor constant), i sigui positiu, cal afegir una certa condició sobre aquests coeficient nuls, tal com podem veure en els exemples senzills següents:

$\boldsymbol{A}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$
$\det(x\boldsymbol{I} - \boldsymbol{A})$	$(x-1)^2$	$(x-1)x$	$(x-1)(x+1)$
$\mathrm{Nuc}(\boldsymbol{A} - \boldsymbol{I})$	$\mathbb{R}^2$	$\langle(1,0)\rangle$	$\langle(1,1)\rangle$

Frobenius va adonar-se que la clau estava en com es disposaven els coeficients nuls, fet que ens duu a parlar de les matrius irreductibles.

Matrius no negatives i irreductibles Una matriu (no negativa) $\boldsymbol{A} \in \mathcal{M}_n(\mathbb{R})$ és *reductible* si existeix una matriu de permutació[11] $\boldsymbol{\Pi}$ tal que

$$\boldsymbol{\Pi}^\top \boldsymbol{A} \boldsymbol{\Pi} = \begin{pmatrix} \boldsymbol{X} & \boldsymbol{Y} \\ \boldsymbol{0} & \boldsymbol{Z} \end{pmatrix},$$

on $\boldsymbol{X}$ i $\boldsymbol{Z}$ són matrius quadrades. Altrament, direm que $\boldsymbol{A}$ és una *matriu irreductible*.

9 L'*espectre* d'una matriu quadrada és el conjunt format per tots els seus valors propis.
10 El *radi espectral* d'una matriu quadrada és el valor màxim, en mòdul, del seu espectre.
11 Una *matriu de permutació* és una matriu binària que té un únic 1 en cada fila i columna.

Notem que la matriu $\boldsymbol{\Pi}^\top \boldsymbol{A}\boldsymbol{\Pi}$ resulta de reordenar les files i les columnes de $\boldsymbol{A}$ d'acord amb la permutació π associada a $\boldsymbol{\Pi} = (\pi_{ij})$, on $\pi_{ij} = 1$ indica que $\pi(j) = i$. Així,

$$(\boldsymbol{\Pi}^\top \boldsymbol{A}\boldsymbol{\Pi})_{ij} = (\boldsymbol{A})_{\pi(i)\pi(j)}\,.$$

Determinar, doncs, el caràcter irreductible d'una matriu, emprant directament la definició anterior, comportaria de l'ordre de $O(n!)$ productes matricials. Veurem com podem «visualitzar» la noció d'irreductibilitat d'una matriu, associant-li un graf, i com, de retruc, obtindrem un procediment de cost polinòmic per verificar aquesta propietat. Per fer-ho introduirem algunes definicions que serveixen de pont entre la teoria de matrius no negatives i la teoria de grafs.

A cada matriu (no negativa) $\boldsymbol{A} \in \mathcal{M}_n(\mathbb{R})$, $\boldsymbol{A} = (a_{ij})$, li podem fer correspondre un digraf $\Gamma = \Gamma(\boldsymbol{A})$, amb conjunt de vèrtexs $V = \{1, \ldots, n\}$, on els seus arcs assenyalin, per a cada columna de $\boldsymbol{A}$, les files ocupades pels respectius coeficients no nuls:

$$(j, i) \in A(\Gamma) \iff a_{ij} \neq 0\,.$$

Així, els coeficients no nuls de $\boldsymbol{A}$ es podrien veure com els pesos dels arcs de $\Gamma(\boldsymbol{A})$.

Observem que els digrafs associats a les matrius $\boldsymbol{A}$ i $\boldsymbol{\Pi}^\top \boldsymbol{A}\boldsymbol{\Pi}$ són *isomorfs*[12] entre si, és a dir, tenen la mateixa estructura, ja que un s'obté a partir de l'altre mitjançant una reetiquetació dels seus vèrtexs (donada per la permutació π o π^{-1}, segons correspongui). D'altra banda, el fet que una matriu tingui l'estructura de blocs $\left(\begin{smallmatrix} X & Y \\ 0 & Z \end{smallmatrix}\right)$ significa que en el seu digraf associat hi ha una part dels seus vèrtexs (corresponents a les columnes de $\boldsymbol{X}$) des dels quals no es pot assolir cap dels vèrtexs de l'altra part complementària (corresponents a les files de $\boldsymbol{Z}$). Altrament, una matriu és irreductible si, i només si, en el seu digraf associat des de qualsevol vèrtex es poden assolir tots els altres, propietat que es coneix amb el nom de *connexió forta*.[13]

Si tots els coeficients no nuls d'una matriu no negativa $\boldsymbol{A} \in \mathcal{M}_n(\mathbb{R})$ fossin uns, aleshores aquesta es podria veure directament com la *matriu d'adjacència*[14] del seu digraf associat Γ. En tal situació, els coeficients de la potència

12 Dos digrafs Γ_1 i Γ_2 són *isomorfs* si existeix una aplicació bijectiva π entre els seus respectius conjunts de vèrtexs que preserva totes les adjacències, és a dir, $(i, j) \in A(\Gamma_1)$ si, i només si, $(\pi(i), \pi(j)) \in A(\Gamma_2)$.

13 Un digraf Γ és *fortament connex* si entre cada parell de vèrtexs u i v de Γ hi ha un recorregut de u a v, entès com una seqüència finita de vèrtexs, $u = u_0, u_1, \ldots, u_l = v$, on $(u_{i-1}, u_i) \in A(\Gamma)$ per a cada $i = 1, \ldots, l$, on l és el nombre de passos o longitud del recorregut.

14 Com que un digraf $\Gamma = (V, A)$ representa una relació binària definida sobre un conjunt finit $V = \{1, \ldots, n\}$, una de les seves representacions més natural, i emprada, és l'anomenada *matriu d'adjacència*, $\boldsymbol{A} = (a_{ij})$, on els seus coeficients vénen a indicar si els corresponents vèrtexs estan o no relacionats. Nosaltres prendrem

$$a_{ij} = \begin{cases} 1 & \text{si } (j, i) \in A, \\ 0 & \text{altrament,} \end{cases} \tag{4}$$

és a dir, identifiquem els índexs de les columnes i de les files com els vèrtexs origen i destí dels arcs, respectivament. Cal dir, però, que en la literatura se sol considerar a l'inrevés, fet que comporta que la matriu d'adjacència sigui la transposada de la definida anteriorment.

k-èsima de $\boldsymbol{A}$ ens comptarien el nombre total de recorreguts de longitud k entre els vèrtexs corresponents de Γ (vegeu, per exemple, [10]). Llavors, tenint en compte que la distància[15] (o grau de separació) màxima entre dos vèrtexs d'un digraf fortament connex d'ordre n és com a màxim $n-1$, la condició que Γ sigui fortament connex equival a dir que la matriu $\boldsymbol{I}+\boldsymbol{A}+\cdots+\boldsymbol{A}^{n-1}$ sigui positiva. Com a resultat obtenim la caracterització pràctica següent de les matrius no negatives i irreductibles:

PROPOSICIÓ 2. *Sigui $\boldsymbol{A}\in\mathcal{M}_n(\mathbb{R})$ una matriu no negativa. Aleshores, els enunciats següents són equivalents:*

1. *$\boldsymbol{A}$ és irreductible;*
2. *$(\boldsymbol{I}+\boldsymbol{A})^{n-1}>\boldsymbol{0}$;*
3. *$\Gamma(\boldsymbol{A})$ és un digraf fortament connex.*

En l'enunciat 2 podem substituir els coeficients positius de $\boldsymbol{A}$ per uns i fer les operacions amb l'aritmètica booleana. Tot i així, hi ha algorismes més eficients per determinar si un digraf és fortament connex com, per exemple, l'algorisme de Warshall [25] i l'algorisme de cerca per fondària prioritària [20], entre altres.

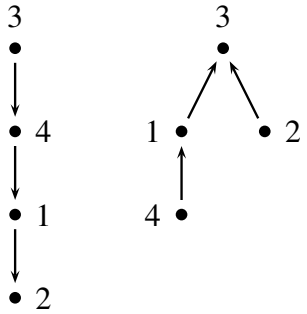

FIGURA 4: Dos subarbres dirigits del digraf Γ_1, els quals mostren que des del vèrtex 3 es poden assolir tots els altres vèrtexs i a l'inrevés.

EXEMPLE 3. La matriu no negativa $\boldsymbol{P}_1$, definida en l'exemple 2, és irreductible ja que el seu digraf associat Γ_1, mostrat en la figura 3, és fortament connex. N'hi ha prou amb comprovar que des d'un vèrtex donat v es poden assolir tots els altres vèrtexs i que des de tots aquests s'arriba a v (vegeu la figura 4, on hem pres $v=3$). D'aquesta manera tots els vèrtexs estan connectats, si més no a través del vèrtex v. Aquesta propietat també es pot verificar emprant la matriu d'adjacència $\boldsymbol{A}_1$ de Γ_1, en què els coeficients no nuls de $\boldsymbol{P}_1$ s'han reemplaçat

15 Si Γ és un digraf fortament connex, aleshores la distància d'un vèrtex i a un vèrtex j de Γ és la mínima de les longituds dels recorreguts que van de i a j.

per uns, la qual satisfà:

$$(\boldsymbol{I}+\boldsymbol{A}_1)^3 = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix}^3 = \begin{pmatrix} 6 & 4 & 3 & 4 \\ 4 & 2 & 1 & 3 \\ 7 & 4 & 2 & 4 \\ 11 & 7 & 4 & 6 \end{pmatrix} > \mathbf{0}\,.$$

En aquest cas, la potència tercera és la primera que dóna lloc a una matriu positiva ja que $(\boldsymbol{I}+\boldsymbol{A}_1)^2$ té el coeficient $(2,3)$ nul, fet que significa que la distància del vèrtex 3 al vèrtex 2 en Γ_1 és 3.

Ara estem en condicions d'enunciar el teorema de Frobenius que, com ja hem indicat prèviament, se sol citar amb el nom de *teorema de Perron-Frobenius*.

TEOREMA 3 (FROBENIUS). *Sigui $\boldsymbol{A} \in \mathcal{M}_n(\mathbb{R})$ una matriu no negativa i irreductible. Llavors:*

1. *$\boldsymbol{A}$ té un valor propi positiu*[16] *r per al qual hi ha un vector propi positiu.*
2. *r és l'únic valor propi que satisfà les condicions anteriors.*[17] *A més,*
   - 2a. *r té multiplicitat algebraica igual a 1.*
   - 2b. *r és un valor propi dominant de $\boldsymbol{A}$, $r = \rho(\boldsymbol{A})$.*
   - 2c. *Si $\boldsymbol{A}$ té h valors propis de mòdul màxim, aleshores tots són simples i es corresponen amb les arrels de l'equació $\lambda^h = r^h$. Més encara, l'espectre de $\boldsymbol{A}$, vist com un conjunt de punts del pla complex, es manté invariant per l'acció d'un gir d'angle $2\pi/h$. Si $h > 1$, aleshores la matriu $\boldsymbol{A}$, via una reordenació de les seves files i columnes, pren l'«estructura cíclica» següent:*

$$\begin{pmatrix} 0 & \boldsymbol{A}_{12} & 0 & \dots & 0 \\ 0 & 0 & \boldsymbol{A}_{23} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \boldsymbol{A}_{h-1\,h} \\ \boldsymbol{A}_{h1} & 0 & 0 & \dots & 0 \end{pmatrix}, \tag{5}$$

   en què els blocs situats sobre la diagonal principal són matrius quadrades.

La literatura sobre matrius no negatives aplega un bon grapat de demostracions diferents del teorema de Perron-Frobenius (vegeu [24]). Una de les més senzilles, tal com es recull en la tesi d'Enric Ventura [32, cap. 1], aplica el teorema del punt fix de Brouwer a la funció contínua

$$\begin{aligned} f_{\boldsymbol{A}} : \Delta_{n-1} &\longrightarrow \Delta_{n-1} \\ \boldsymbol{x} &\longrightarrow \tfrac{\boldsymbol{A}\boldsymbol{x}}{\|\boldsymbol{A}\boldsymbol{x}\|_1}, \end{aligned}$$

16 Llevat que $\boldsymbol{A}$ sigui la matriu nul·la 1×1, que anomenarem *matriu trivial*, per la qual $r = 0$.
17 Aquest valor propi distingit r se l'anomena *valor propi de Perron-Frobenius.*

definida sobre el $n-1$ símplex

$$\Delta_{n-1} = \{\boldsymbol{x} \in \mathbb{R}^n \,:\, \boldsymbol{x} \geq \mathbf{0} \text{ i } \|\boldsymbol{x}\|_1 = 1\}\,.$$

Notem que f_A està ben definida ja que $\boldsymbol{A}\,\boldsymbol{x} = \mathbf{0}$ implicaria que $\left(\sum_{i=1}^n \boldsymbol{A}^i\right)\,\boldsymbol{x} = \mathbf{0}$, igualtat que no es pot donar quan la matriu $\boldsymbol{A}$ és irreductible, on

$$\sum_{i=1}^{n} \boldsymbol{A}^i = \boldsymbol{A}\left(\sum_{i=0}^{n-1} \boldsymbol{A}^i\right) > \mathbf{0}\,.$$

Així mateix, si $f_{\boldsymbol{A}}(\boldsymbol{v}) = \boldsymbol{v}$ aleshores $\boldsymbol{A}\,\boldsymbol{v} = r\,\boldsymbol{v}$, en què $r = \|\boldsymbol{A}\,\boldsymbol{v}\|_1$, d'on resulta

$$\mathbf{0} < \left(\sum_{i=1}^{n} \boldsymbol{A}^i\right)\,\boldsymbol{v} = \left(\sum_{i=1}^{n} r^i\right)\,\boldsymbol{v}$$

i, consegüentment, $\boldsymbol{v} > \mathbf{0}$.

El fet d'haver utilitzat el teorema del punt fix per deduir l'existència d'aquest vector propi positiu $\boldsymbol{v}$ ens duu, de manera natural, a plantejar el mètode numèric següent per calcular-lo:

$$\begin{aligned} &\boldsymbol{v}^{(0)} \geq \mathbf{0}, && \|\boldsymbol{v}^{(0)}\|_1 = 1,\\ &\boldsymbol{u}^{(k)} := \boldsymbol{A}\boldsymbol{v}^{(\boldsymbol{k-1})}, && \boldsymbol{v}^{(k)} := \tfrac{\boldsymbol{u}^{(k)}}{\|\boldsymbol{u}^{(k)}\|_1},\ k = 1, 2, \ldots \end{aligned}$$

Aquest procediment iteratiu és el conegut *mètode de la potència*,

$$\boldsymbol{v}^{(k)} = \frac{\boldsymbol{A}^k\boldsymbol{v}^{(0)}}{\|\boldsymbol{A}^k\boldsymbol{v}^{(0)}\|_1}\,,$$

del qual sabem que, sota la hipòtesi addicional que r és l'únic valor propi de mòdul màxim, tenim garantida la seva convergència ($\boldsymbol{v}^{(k)} \to \boldsymbol{v}$), per a qualsevol vector inicial $\boldsymbol{v}^{(0)}$ tal que la seva component respecte a la direcció buscada $\boldsymbol{v}$ no sigui nul·la[18] (vegeu [12]). Recordem com es dedueix aquesta convergència en el cas que la matriu $\boldsymbol{A}$ sigui diagonalitzable. Així, si $\boldsymbol{v}^{(0)} = c_1\boldsymbol{v}_1 + c_2\boldsymbol{v}_2 + \cdots + c_n\boldsymbol{v}_n$, on $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\}$ és una base formada per vectors propis de $\boldsymbol{A}$, essent $\boldsymbol{v}_i$ un vector propi de valor propi λ_i i $r = \lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_n|$, aleshores

$$\boldsymbol{A}^k\boldsymbol{v}^{(0)} = c_1\lambda_1^k\boldsymbol{v}_1 + c_2\lambda_2^k\boldsymbol{v}_2 + \cdots + c_n\lambda_n^k\boldsymbol{v}_n\,.$$

D'aquesta manera, si $c_1 \neq 0$ tenim que

$$\frac{\boldsymbol{A}^k\boldsymbol{v}^{(0)}}{r^k} = c_1\boldsymbol{v}_1 + c_2\left(\frac{\lambda_2}{r}\right)^k\boldsymbol{v}_2 + \cdots + c_n\left(\frac{\lambda_n}{r}\right)^k\boldsymbol{v}_n \longrightarrow c_1\boldsymbol{v}_1\,,$$

18 Des del punt de vista pràctic, aquesta condició no ens hauria de preocupar ja que els errors d'arrodoniment en els còmputs ens ajudarien, aquest cop, a redreçar la situació, en el sentit que tot i haver començat per una «direcció problemàtica» ho deixaria de ser al cap d'uns passos. A més, des del punt de vista teòric, sabem que el conjunt de vectors inicials amb component nul·la, respecte al vector buscat, té mesura nul·la.

d'on resulta que $\boldsymbol{v}^{(k)} \longrightarrow \boldsymbol{v}_1/\|\boldsymbol{v}_1\|_1$. La velocitat de convergència del mètode de la potència depèn del quocient $\frac{\lambda_2}{r}$, és a dir, com més gran sigui la diferència relativa entre els dos valors propis amb mòdul més gran, més ràpidament aconseguirem apropar-nos a la direcció privilegiada.

En el cas que ens ocupa, el fet de prendre com a punt de partida un vector $\boldsymbol{v}^{(0)}$ no negatiu, com és natural, ens assegura que la direcció cercada hi estarà representada, fet que permetrà atreure els vectors successius cap a aquesta direcció. Aquesta remarca es basa en el fet que si prenem els vectors propis positius, $\boldsymbol{v}_1$ i $\boldsymbol{w}_1$, de les matrius $\boldsymbol{A}$ i $\boldsymbol{A}^\top$, respectivament, i considerem una base de Jordan de la matriu $\boldsymbol{A}$, $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\}$, aleshores el vector $\boldsymbol{w}_1$ resulta ser ortogonal a cadascun dels vectors $\boldsymbol{v}_i \neq \boldsymbol{v}_1$. Vegem-ho per al cas que $\boldsymbol{v}_i$ sigui un vector propi de $\boldsymbol{A}$ associat al valor propi $\lambda_i \neq r$. Així,

$$r\langle \boldsymbol{w}_1, \boldsymbol{v}_i\rangle = \langle \boldsymbol{A}^\top \boldsymbol{w}_1, \boldsymbol{v}_i\rangle = \langle \boldsymbol{w}_1, \boldsymbol{A}\boldsymbol{v}_i\rangle = \lambda_i \langle \boldsymbol{w}_1, \boldsymbol{v}_i\rangle,$$

d'on resulta $\langle \boldsymbol{w}_1, \boldsymbol{v}_i\rangle = 0$, per a cada $i = 2, \ldots, n$. Com a conseqüència,

$$0 < \langle \boldsymbol{v}^{(0)}, \boldsymbol{w}_1\rangle = c_1 \langle \boldsymbol{v}_1, \boldsymbol{w}_1\rangle \;\Rightarrow\; c_1 > 0\,.$$

Reprenem el nostre problema inicial Si la matriu normalitzada d'enllaços $\boldsymbol{P}$, definida a (2), fos irreductible i tingués un únic valor propi de mòdul màxim, llavors tindríem resoltes totes les qüestions plantejades sobre l'existència, unicitat i còmput del vector de valoracions $\boldsymbol{v}$. I aquestes són precisament les dues característiques que defineixen les anomenades *matrius primitives*, que exposarem a continuació. Abans, però, notem que si la matriu $\boldsymbol{P}$ és irreductible, en particular, totes les seves columnes sumen 1 (el seu digraf associat, que és fortament connex, no pot contenir vèrtexs amb grau de sortida 0). Llavors, $\boldsymbol{P}^\top \boldsymbol{j} = \boldsymbol{j}$, on $\boldsymbol{j} = (1, \ldots, 1)^\top$. En conseqüència, $\rho(\boldsymbol{P}) = \rho(\boldsymbol{P}^\top) = 1$, és a dir, el valor propi de Perron-Frobenius de $\boldsymbol{P}$ és $r = 1$. Per tant, l'equació (3) esdevé

$$\boldsymbol{v} = \boldsymbol{P}\boldsymbol{v}\,. \tag{6}$$

Notem també que en el còmput de les diferents aproximacions $\boldsymbol{v}^{(k)} = \boldsymbol{P}\boldsymbol{v}^{(k-1)}$ no ens haurem de preocupar de la normalització, sempre que partim d'un vector inicial $\boldsymbol{v}^{(0)}$ amb $\|\boldsymbol{v}^{(0)}\|_1 = 1$, ja que

$$\|\boldsymbol{P}\boldsymbol{v}\|_1 = \sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij}v_j = \sum_{j=1}^{n} v_j \sum_{i=1}^{n} p_{ij} = \|\boldsymbol{v}\|_1\,.$$

Matrius primitives Les matrius no negatives i irreductibles es poden classificar en dos grups, segons si tenen un únic valor propi dominant o si en tenen més d'un (distribuïts uniformement sobre el seu cercle espectral). Les primeres se les anomena *matrius primitives*; altrament, se'n diuen *matrius imprimitives*. D'acord amb el teorema de Perron-Frobenius, si una matriu $\boldsymbol{A}$ és imprimitiva aleshores hi ha una reordenació de les seves files i columnes

que la transforma en una matriu per blocs, que té la forma d'una matriu de permutació cíclica; vegeu (5). De nou ens trobem amb una condició sobre la matriu que depèn únicament de les posicions dels seus elements no nuls i, en conseqüència, pot ser formulada en termes de grafs. Així, si una matriu $\boldsymbol{A}$ és imprimitiva (amb índex d'imprimitivitat h), el seu digraf associat $\Gamma(\boldsymbol{A})$ és un *h-cicle generalitzat*,[19] en el sentit que el seu conjunt de vèrtexs V admet una partició en $h > 1$ parts, $V = V_1 \cup \cdots \cup V_h$, tal que els vèrtexs de cada part V_i només poden ser adjacents cap als vèrtexs de la part (següent) V_{i+1}, on els subíndexs es prenen mòdul h (notem que el cas $h = 2$ correspon als anomenats *digrafs bipartits*). Fiol *et al.* [16] varen provar que, dins la classe dels digrafs fortament connexos, els digrafs cicle generalitzats són els únics per als quals no és possible trobar recorreguts d'una mateixa longitud entre cada parell de vèrtexs, és a dir, no hi ha cap potència de la seva matriu d'adjacència que doni lloc a una matriu positiva. Notem que un digraf amb estructura de h-cicle generalitzat només pot contenir cicles de longitud múltiple de h. Es demostra que el recíproc també és cert, suposant que el digraf és fortament connex. Recollim aquestes caracteritzacions de les matrius primitives en el resultat següent:

TEOREMA 4. *Sigui $\boldsymbol{A} \in \mathcal{M}_n(\mathbb{R})$ una matriu no negativa (i no trivial). Aleshores, els enunciats següents són equivalents:*

1. *$\boldsymbol{A}$ és una matriu primitiva.*
2. *Existeix un enter positiu m tal que $\boldsymbol{A}^m > \boldsymbol{0}$.*
3. *El digraf associat $\Gamma(\boldsymbol{A})$ és fortament connex i el màxim comú divisor de les longituds dels seus cicles és 1.*

A [5] es recullen diferents fites per al valor mínim de l'exponent m tal que $\boldsymbol{A}^m > 0$, les quals mostren que el caràcter primitiu d'una matriu no negativa també pot verificar-se en temps polinòmic.

EXEMPLE 4. La matriu no negativa $\boldsymbol{P}_1$, definida en l'exemple 2, és primitiva ja que el seu digraf associat Γ_1, mostrat en la figura 3, és fortament connex i no és un cicle generalitzat, ja que conté cicles de longituds 2 i 3, valors primers entre si. Una justificació alternativa resulta de comprovar que $\boldsymbol{P}_1^5 > \boldsymbol{0}$.

3.4 El PageRank i el comportament d'un «surfista de la Web»

Sota la hipòtesi, poc realista, que el digraf que modelitza la Web fos fortament connex (i no fos un cicle generalitzat, supòsit que sí que és plausible), la teoria de Perron-Frobenius ens diu que el vector de valoracions $\boldsymbol{v}$ de les pàgines web pot obtenir-se com el límit de la successió de vectors $\{\boldsymbol{P}^k\boldsymbol{v}^{(0)}\}_k$, independentment del vector de valoracions inicials $\boldsymbol{v}^{(0)} \geq \boldsymbol{0}$, on

$$\|\boldsymbol{P}^k\boldsymbol{v}^{(0)}\|_1 = \|\boldsymbol{P}^{k-1}\boldsymbol{v}^{(0)}\|_1 = \cdots = \|\boldsymbol{v}^{(0)}\|_1 = 1\,.$$

19 Se'n diu així ja que en el cas que totes les parts es redueixin a un sol vèrtex obtenim el digraf cicle.

Vegem com podem interpretar les components del vector $\boldsymbol{v}$. Per començar, com que totes les columnes de la matriu no negativa $\boldsymbol{P} = (p_{ij})$ sumen 1 ($\boldsymbol{P}$ és una *matriu estocàstica per columnes*), els seus coeficients poden pensar-se com la probabilitat que un «surfista de la Web»[20] salti d'una pàgina a una altra. Així, el coeficient $p_{ij} = \frac{1}{n_j}$, on n_j és el total d'enllaços des de la pàgina j, representaria la probabilitat que el surfista esculli com a nova destinació la pàgina i, en el supòsit que la seva navegació es faci de manera aleatòria tenint tots els enllaços de la pàgina j les mateixes possibilitats de ser triats. Llavors, el coeficient $p_{ij}^{(k)}$ de la matriu $\boldsymbol{P}^k$ indicaria la probabilitat que, partint de la pàgina j, després de k passos arribem a la pàgina i. Com que estem suposant que la matriu $\boldsymbol{P}$ és primitiva, aquestes probabilitats tendeixen a estabilitzar-se, a mesura que el nombre de passos k augmenta, i només depenen de la pàgina destí del recorregut; és a dir, tots els vectors columna de $\boldsymbol{P}^k$ convergeixen cap a un mateix vector, que és precisament la solució buscada $\boldsymbol{v}$. Així, doncs, la valoració v_i ($\approx p_{ij}^{(k)}$, k prou gran) d'una pàgina i pot pensar-se com la probabilitat que un surfista vagi cap a aquesta pàgina. D'aquesta manera, les pàgines més importants, d'acord amb aquesta valoració, són aquelles que la mateixa topologia de la Web fa que un surfista les visiti més sovint. Aquest és el significat que Brin i Page donen al seu PageRank [28] i que reforçaria la idoneïtat d'aquesta mesura.

Les argumentacions anteriors poden formular-se en un context més ampli, com és el de les anomenades *cadenes de Markov*, com ja us n'haureu adonat (i potser haureu trobat a faltar) tots aquells que estigueu avesats a treballar amb processos estocàstics. Així, podríem considerar la successió de variables aleatòries $\{X_t\}_{t=0}^{\infty}$, on X_t indicaria en quina pàgina (estat), $\{S_1, S_2, \ldots, S_n\}$, es troba un surfista en el pas (instant) t. La suposició que aquest surfista triï, en cada pas, el seu nou destí tenint en compte únicament el lloc on es troba (com si no tingués memòria), es tradueix en l'anomenada *propietat de Markov* sobre les anomenades *probabilitats de transició*:

$$\text{Prob}(X_{t+1} = S_i \mid X_t = S_{j_t}, \ldots, X_0 = S_{j_0}) = \text{Prob}(X_{t+1} = S_i \mid X_t = S_{j_t})\,.$$

Si, a més, suposem que cada cop que el surfista arriba a una pàgina S_j les probabilitats de moure's cap a cadascuna de les pàgines restants no han canviat, és a dir,

$$\text{Prob}(X_{t+1} = S_i \mid X_t = S_j) = p_{ij}\,,$$

llavors $\{X_t\}_{t=0}^{\infty}$ és una *cadena de Markov homogènia*, amb un nombre finit d'estats, i la matriu $\boldsymbol{P} = (p_{ij})$ esdevé la seva *matriu de transició*. En aquest context, la condició que la matriu $\boldsymbol{P}$ sigui primitiva es tradueix amb dir que la cadena de Markov sigui *irreductible* i *aperiòdica*. I per a dites cadenes existeix la *distribució de probabilitats estacionària* $\boldsymbol{\pi}$, la qual satisfà $\lim_{k\to\infty} \boldsymbol{P}^k \boldsymbol{\pi}^{(0)} = \boldsymbol{\pi}$, per a qualsevol distribució inicial de probabilitats $\boldsymbol{\pi}^{(0)}$.

20 Navega per la Web seguint únicament els enllaços («onades»).

EXEMPLE 5. En la taula següent mostrem les aproximacions successives a les valoracions de les pàgines enllaçades segons indica el digraf Γ_2 de la figura 3.

k	$v_1^{(k)}$	$v_2^{(k)}$	$v_3^{(k)}$	$v_4^{(k)}$	$v_5^{(k)}$
0	0,2	0,2	0,2	0,2	0,2
1	0,2	0,167	0,167	0,2	0,267
⋮	⋮	⋮	⋮	⋮	⋮
11	0,208	0,167	0,194	0,181	0,25
12	0,208	0,167	0,194	0,181	0,25

L'ordenació de les pàgines, d'acord amb la valoració calculada, és $v_5 > v_1 > v_3 > v_4 > v_2$. Això significa que la mateixa topologia dels enllaços ens duria més sovint a visitar la pàgina 5, fet que podríem experimentar generant un nombre prou gran de recorreguts aleatoris suficientment llargs i comptabilitzant la freqüència de cada pàgina com a destí d'aquests recorreguts.

Pel que fa a la valoració de les pàgines representades en el digraf Γ_1 de la figura 3, aquesta és $v_1 = v_4 = 0,353$, $v_2 = 0,118$ i $v_3 = 0,176$. D'on resulta l'ordenació $v_1 = v_4 > v_3 > v_2$, tal com havíem intuït.

4 Algorisme PageRank: formulació completa

En la figura 5 mostrem dos petits exemples de situacions problemàtiques pel que fa al còmput del *PageRank*, tal com s'ha formulat en la secció anterior.

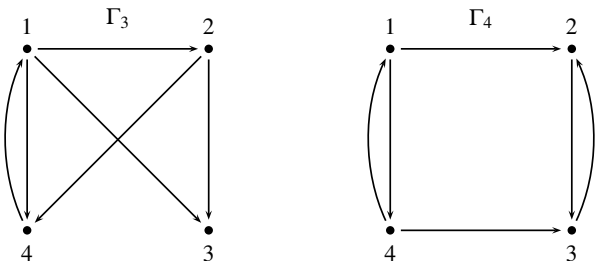

FIGURA 5: Dígrafs amb més d'un component fortament connex.

Així, si partim del vector de valoracions inicials $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})^\top$, mentre que en el digraf Γ_3 anirem a parar al vector idènticament nul (si no normalitzem el vector resultant després de cada iteració), en el digraf Γ_4 tendirem cap al vector $(0, \frac{1}{2}, \frac{1}{2}, 0)^\top$. En qualsevol cas ens apareixen valoracions finals nul·les, fet que no ens interessa. Tot seguint analitzarem més a fons aquestes dues situacions i veurem com podem resoldre-les.

4.1 Existència de pàgines pou

Observem que el digraf Γ_3 conté un vèrtex, el 3, sense cap enllaç de sortida (vèrtex pou). Aquest podria representar una pàgina web que sigui tan sols una imatge o un document PDF; també podria tractar-se d'una pàgina que, tot i haver estat descarregada en el repositori del cercador, no hagi estat analitzada i, per tant, no s'hagin descobert els seus enllaços. S'estima que hi ha una proporció significativa d'aquestes pàgines pou en el conjunt de la Web i, en conseqüència, cal decidir què fer-ne.

Si no es fes cap tractament especial podria succeir, tal com passa en el digraf Γ_3, que totes les valoracions s'acabessin «perdent», en el sentit que tot recorregut prou llarg quedaria atrapat en un vèrtex pou i, en conseqüència, no es podria continuar transmetent la valoració rebuda (fet que es produiria quan des de qualsevol vèrtex pogués assolir-se almenys un vèrtex pou). La manera més senzilla d'evitar-ho seria considerar que si un surfista cau en una pàgina pou aleshores en el següent pas pot saltar (virtualment) a qualsevol pàgina, escollida a l'atzar, i amb igual probabilitat, entre totes les pàgines. Això equival a reemplaçar cada columna idènticament nul·la de la corresponent matriu de pesos $\boldsymbol{P}$ pel vector $(\frac{1}{n},\ldots,\frac{1}{n})^\top$, on n representa el total de pàgines.[21] D'aquesta manera, la matriu resultant $\boldsymbol{P}_e$ és una matriu estocàstica.

EXEMPLE 6. En el cas del digraf Γ_3, les matrius $\boldsymbol{P}$ i $\boldsymbol{P}_e$ són:

$$\boldsymbol{P}=\begin{pmatrix} 0 & 0 & 0 & 1\\ \frac{1}{3} & 0 & 0 & 0\\ \frac{1}{3} & \frac{1}{2} & 0 & 0\\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{pmatrix} \quad \text{i} \quad \boldsymbol{P}_e=\begin{pmatrix} 0 & 0 & \frac{1}{4} & 1\\ \frac{1}{3} & 0 & \frac{1}{4} & 0\\ \frac{1}{3} & \frac{1}{2} & \frac{1}{4} & 0\\ \frac{1}{3} & \frac{1}{2} & \frac{1}{4} & 0 \end{pmatrix}.$$

Com que la matriu $\boldsymbol{P}_e$ resulta ser irreductible, prenem el seu únic vector propi positiu i normalitzat, $(0{,}320,\ 0{,}170,\ 0{,}255,\ 0{,}255)^\top$, com el vector de valoracions. En canvi, la matriu $\boldsymbol{P}$, que òbviament no és irreductible, té dos vectors propis no negatius i normalitzats, $(0{,}340,\ 0{,}152,\ 0{,}254,\ 0{,}254)^\top$ i $(0,0,1,0)^\top$, els quals corresponen als valors propis 0,746 i 0, respectivament. Notem que mentre el radi espectral de $\boldsymbol{P}_e$ és 1, com que es tracta d'una matriu estocàstica, el corresponent a $\boldsymbol{P}$ és inferior a 1.

Tenint en compte que els navegadors web ens permeten «tornar enrere», per què no incorporem aquesta acció, si més no per escapar-nos d'una pàgina pou, en lloc d'inventar-nos els «salts virtuals»? Hi ha hagut diferents propostes al respecte. En aquest sentit, Langville i Meyer [23] suggereixen reemplaçar cada pàgina pou, z, per tantes «pàgines fictícies», $y_1,\ldots,y_k$, com enllaços d'entrada tingui z, i substituir cada arc (x_i,z) pel camí d'anada i tornada

21 Cal dir, però, que Brin i Page, en la seva formulació inicial de l'algorisme PageRank [7], havien proposat de suprimir (recurrentment) les pàgines pou, per després efectuar el còmput de les valoracions amb la resta de les pàgines i, finalment, tornar-les a introduir per rebre la corresponent valoració.

x_i, y_i, x_i. D'aquesta manera, el surfista no hauria de recordar de quina pàgina prové, quan cau a una pàgina pou, i el seu comportament podria continuar modelitzant-se com una cadena de Markov. En aquest nou marc, la valoració que s'atorga a una pàgina pou és igual a la suma de les valoracions de les seves còpies. L'inconvenient principal d'aquesta alternativa és l'augment considerable en la grandària del digraf que modelitza la nova situació. Una manera d'evitar-ho seria afegir «memòria» en el model, així esdevindria més complex però, possiblement, reflectint millor la navegació per la Web (vegeu, per exemple, [13]).

4.2 Existència de més d'un component fortament connex

D'ara endavant, suposarem que hem solucionat el problema de les pàgines pou permetent «saltar» des d'aquestes a qualsevol pàgina. Dit d'una altra manera, cada cop que fem cap a una pàgina pou triem una pàgina a l'atzar i hi reiniciem el recorregut. Aquesta idea de donar l'oportunitat de començar de nou la navegació, com si el navegant s'avorrís de seguir els enllaços, serveix per resoldre el problema de la desconnexió de la Web. Així, si admetem que, en cada pas i independentment de la pàgina on ens trobem, sempre hi ha una certa probabilitat d'eixida, cap a qualsevol pàgina, llavors el navegant veuria la Web formada d'una sola «peça». Dita probabilitat hauria de ser petita per tal de donar més pes a la topologia real de la Web.

Si anomenem $1 - \alpha > 0$ la probabilitat d'eixida, llavors es tractaria de substituir la matriu normalitzada d'enllaços $\boldsymbol{P}$ per la matriu

$$\boldsymbol{P}(\alpha) = \alpha \boldsymbol{P} + (1 - \alpha)\frac{1}{n}\boldsymbol{J}\,, \tag{7}$$

on $\boldsymbol{J}$ denota la matriu $n \times n$ tota d'uns. Notem que $\boldsymbol{P}(\alpha)$ és una matriu estocàstica per columnes, ja que és una combinació convexa de matrius que també ho són. D'aquesta manera, podem continuar pensant els coeficients de $\boldsymbol{P}(\alpha)$ com a probabilitats de saltar d'una pàgina a una altra, però ara no necessàriament a través d'enllaços reals entre si. Així, si estem en la pàgina j, llavors amb probabilitat α optem per seguir per un dels n_j enllaços de la pàgina, tal com es feia en la formulació simple del PageRank, i amb probabilitat $1 - \alpha$ decidim saltar a una de les n pàgines de la Web, escollida a l'atzar i de manera uniforme entre totes (hi hagi o no enllaç cap a aquesta). Com que en aquesta nova formulació la matriu de transició $\boldsymbol{P}(\alpha)$ és positiva, tenim garantida l'existència de la distribució estacionària de probabilitats (totes no nul·les), les quals representaran les valoracions de les corresponents pàgines web.

Si fos necessari, podríem «personalitzar» les probabilitats de «saltar virtualment» cap a cadascuna de les pàgines. És a dir, podríem reemplaçar la matriu $\frac{1}{n}\boldsymbol{J}$, que té tots els coeficients iguals a $\frac{1}{n}$, per una matriu que tingués tots els coeficients d'una mateixa fila i iguals a un cert valor $q_i > 0$, on la suma d'aquests valors, $q_1, \ldots, q_n$, hauria de continuar essent 1 (per ser una distribució de probabilitats).

EXEMPLE 7. En la taula següent mostrem les valoracions resultants per a les pàgines enllaçades segons el digraf Γ_4 (vegeu la figura 5), prenent diferents valors per al paràmetre α, el primer dels quals, $\alpha = 0,85$, correspon al valor suggerit per Brin i Page [8, 28]. Com és d'esperar, a mesura que anem disminuint el valor de α, les valoracions de les pàgines es van equilibrant, ja que anem rebaixant el pes dels enllaços reals fins que esdevenen totes iguals quan $\alpha = 0$.

α	v_1	v_2	v_3	v_4
0,85	0,065	0,435	0,435	0,065
0,5	0,167	0,333	0,333	0,167
0	0,25	0,25	0,25	0,25

5 L'algorisme PageRank: aspectes computacionals

El pas de la formulació simple a la formulació completa de l'algorisme PageRank, resumida en la substitució de la matriu normalitzada d'enllaços $\boldsymbol{P}$ per la matriu $\boldsymbol{P}(\alpha)$, ens ha resolt els problemes plantejats inicialment, però ens en suggereix de nous: *Com afecta el paràmetre α al procés de còmput i a l'ordenació dels resultats*? i *Quines dificultats afegides comporta passar de tenir una matriu immensa, però escassa (amb molts zeros), a haver-ho de fer amb una matriu on tots els seus coeficients són no nuls?*

5.1 El paràmetre α i el mètode de la potència

Veurem com el mètode iteratiu de la potència, aplicat a la matriu $\boldsymbol{P}(\alpha)$, convergeix més ràpidament a mesura que disminueix el factor d'amortiment α i, en conseqüència, menystenim la topologia real de la Web. Llavors, podrem interpretar el valor triat per Brin i Page, $\alpha = 0,85$, com una solució de compromís entre l'eficiència i l'efectivitat, tal com assenyalen Langville i Meyer [23].

Per determinar la rapidesa del mètode de la potència hem de conèixer com de diferents són els dos primers valors propis, amb mòdul més gran, de la matriu a la qual s'aplica. El resultat següent ens relaciona precisament l'espectre de les matrius que apareixen en les dues formulacions de l'algorisme PageRank (vegeu [23]).

TEOREMA 5. *Si l'espectre d'una matriu estocàstica per columnes $\boldsymbol{P}$ és $\{1, \lambda_2, \ldots, \lambda_n\}$, aleshores l'espectre de la matriu $\boldsymbol{P}(\alpha) = \alpha\boldsymbol{P} + (1-\alpha)\boldsymbol{q}\,\boldsymbol{j}^\top$, on $\boldsymbol{j}^\top$ és el vector tot d'uns i $\boldsymbol{q}$ és un vector de probabilitats, és*

$$\{1, \alpha\lambda_2, \ldots, \alpha\lambda_n\}.$$

En conseqüència, com més petit sigui el valor de α més allunyat es trobarà el segon valor propi ($\alpha\lambda_2$) del primer (1) i, per tant, més aviat convergirà el mètode de la potència.

5.2 Requeriments computacionals i de memòria

En un principi, pot sorprendre que s'hagi escollit el mètode de la potència per trobar el vector (propi) de valoracions, havent-hi altres mètodes iteratius que, en general, són més ràpids. La raó, a banda de la senzillesa, es pot trobar en l'estructura de la matriu d'iteració $\boldsymbol{P}(\alpha) = \alpha\boldsymbol{P} + \frac{(1-\alpha)}{n}\boldsymbol{J}$, la qual permet reformular els còmputs de cada iteració, si s'empra el mètode de la potència, en termes de la matriu $\boldsymbol{P}$, on la major part de les seves entrades són nul·les, fet que comporta un estalvi de còmput i de memòria. Així, si k és el nombre esperat d'enllaços d'una pàgina web, que alguns autors estimen de l'ordre de 10 (vegeu, per exemple, [9]), i n és el total de pàgines web, aleshores el nombre estimat de coeficients no nuls de $\boldsymbol{P}$ és d'ordre lineal en n ($k \cdot n$), en comparació dels n^2 coeficients no nuls de la matriu $\boldsymbol{P}(\alpha)$. Així mateix, el nombre estimat d'operacions també seria d'ordre lineal respecte a n, quantitat que seria multiplicada pel nombre d'iteracions requerides.

A continuació, presentem el pseudocodi de l'algorisme PageRank, corresponent a la seva formulació completa, on en lloc de tenir guardat el digraf de la Web mitjançant la seva corresponent matriu d'adjacència emprem l'anomenada *llista d'adjacències*, la qual ens indicarà, per a cada pàgina, quines són les pàgines a les quals apunta.

Precondició: Γ és un digraf d'ordre n amb *llista d'adjacències* $(L_1, \ldots, L_n)$, on L_i és la llista de vèrtexs adjacents des del vèrtex i i $n_i = |L_i|$. El paràmetre α és un nombre real tal que $0 \leq \alpha < 1$.

```
algorisme PageRank (Γ,α) es
        const
              precisio: real = 1.e − 6;
        fconst
        var
            L: llista de llista adjacencies;
            v1, v2: taula[1..n] de real;
            error: real;
        fvar
        per i := 1 fins n fer
            v1[i] := 1/n;
        fper
        error:= 1;
        mentre error > precisio fer
                per i := 1 fins n fer
                    v2[i] := 0;
                fper
                per i := 1 fins n fer
                    si n_i > 0 aleshores
                       per a cada j ∈ L_i fer
                                v2[j] := v2[j] + v1[i]/n_i;
```

```
                fper
            sino
                per j := 1 fins n fer
                    v2[j] := v2[j] + v1[i]/n;
                fper
            fsi
        fper
        per i := 1 fins n fer
            v2[i] := α * v2[i] + (1 − α)/n;
        fper
        error := ‖v1 − v2‖_1;
        v1 := v2;
    fmentre
    retornar(v1);
falgorisme
```

Val a dir que s'han desenvolupat diferents tècniques per comprimir la informació continguda en la llista d'adjacències del digraf que modelitza la Web, a fi i efecte que pugui encabir-se en la memòria principal (vegeu [23]). Un d'aquests mètodes, ideat per Bharat *et al.* [6], es basa en l'observació que bona part dels enllaços d'una mateixa pàgina tenen com a destí pàgines «properes» quant a la seva localització, fet que comporta que els seus corresponents identificadors tinguin valors pròxims. Això suggereix recuperar aquests identificadors a partir d'un sol d'aquests i dels valors (petits) dels salts corresponents respecte d'aquest. Altres propostes treuen profit del fet que pàgines amb adreces d'un mateix domini possiblement comparteixen molts dels seus enllaços (vegeu [30]). Així mateix, s'han desenvolupat implementacions eficients del PageRank pel que fa a les operacions de lectura/escriptura de dades (vegeu [19]).

6 Altres mètodes de valoració de pàgines web: l'algorisme HITS

6.1 Introducció

Cal dir que, a banda de l'algorisme PageRank, hi ha hagut altres propostes per mesurar el grau de rellevància d'una pàgina web a partir de l'anàlisi dels enllaços. D'entre aquestes destaca l'algorisme HITS, acrònim de *Hypertext Induced Topic Search*, ideat per en Jon Kleinberg, que el va presentar el 1998[22] a l'ACM-SIAM Symposium on Discrete Mathematics i fou publicat l'any següent al *Journal of the ACM* (vegeu [22]). En aquell temps, Kleinberg feia un postdoctorat en un centre de recerca de la IBM, i treballava en un projecte, dirigit per Prabhakar Raghavan, sobre com millorar la qualitat de les cerques a la Web (vegeu [29]). A diferència, però, de Brin i Page, Kleinberg no va convertir la seva idea en una empresa sinó que va decidir continuar en el món acadèmic,

22 El mateix any que fou presentat l'algorisme PageRank.

concretament a la Universitat de Cornell, on ha aconseguit un gran renom com a investigador i com a docent. En aquest sentit, cal recordar que Kleinberg va rebre el prestigiós Premi Nevanlinna, atorgat per la International Mathematical Union durant el Congrés Internacional de Matemàtiques (ICM), celebrat a Madrid l'any 2006, per les seves importants contribucions algorísmiques en l'estudi de la Web i, més en general, de les xarxes complexes anomenades *petit-món* (vegeu [11]).

6.2 Principals diferències entre l'algorisme HITS i l'algorisme PageRank

Mentre que l'algorisme PageRank puntua cadascuna de les pàgines, independentment de la cerca, l'algorisme HITS valora, per a cada consulta, un conjunt (reduït) de pàgines relacionades amb el terme buscat. Ara, doncs, els còmputs s'aplicaran sobre un cert subdigraf, centrat en el tema de la cerca, en lloc de fer-ho sobre l'immens digraf que representa tota la Web. Però, per contra, s'hauran de realitzar en temps de consulta (no es podran tenir precalculats), fet que condicionarà la grandària de l'esmentat subdigraf.

D'altra banda, mentre l'algorisme PageRank resumeix la valoració d'una pàgina en un sol nombre, l'algorisme HITS li assigna dues valoracions, una d'acord amb el seu *grau d'autoritat* (*authority value*) i l'altra segons el seu paper com a *pàgina recurs* o *guia* (*hub value*). Segons diu Kleinberg (vegeu [29]), la idea de l'existència d'aquests dos tipus importants de pàgines responia a l'experiència que es tenia quan es movia per la Web. Així, un troba certes pàgines que, pel fet de recopilar molta informació i enllaços, farien el paper dels *hubs* en la navegació per la Web, i d'altres que, pel seu propi contingut, esdevenen veritables pàgines de referència, avalades precisament pels enllaços (concessions d'autoritat) donats per les pàgines guia. Per exemple, si pensem en el tema dels automòbils, possiblement inclourem les webs dels principals fabricants com a pàgines de referència. Ara bé, segurament aquestes pàgines no contindran enllaços entre si (per no fer publicitat a la pròpia competència) però, en canvi, seran apuntades per aquelles pàgines considerades com a recurs o directori sobre aquest tema.

6.3 Construcció d'un subdigraf centrat en un tema

Donat un tema de cerca T, idealment s'hauria d'escollir una col·lecció relativament petita i prou bona de pàgines, en el sentit que abundessin les «pàgines rellevants» i que, entre aquestes, hi figuressin les de més «autoritat». A la pràctica, Kleinberg proposava seguir el procediment següent:

- Prendre, com a punt de partida, les t primeres pàgines (per exemple $t = 200$) que un cercador, que ordeni els resultats d'una cerca basant-se només en la concordància amb el text buscat, retorni davant la consulta del terme T. Aquestes pàgines formarien l'anomenat *conjunt arrel* R_T.
- Afegir-hi totes aquelles pàgines que siguin apuntades per alguna de les pàgines de R_T. A més, per a cadascuna de les pàgines de R_T incloure-hi

totes les pàgines, fins a un valor llindar d (p. e. $d = 50$), que hi apuntin. Totes aquestes pàgines, amb els seus corresponents enllaços, enriqueixen la «col·lecció».

Si anomenem S_T al conjunt de pàgines resultant i considerem tots els enllaços existents entre elles, obtenim l'anomenat *subdigraf centrat en el tema T*, és a dir, el subdigraf induït $\Gamma_T = \Gamma[S_T]$.

6.4 Formulació de l'algorisme *HITS*

Kleinberg considera que una «bona pàgina guia» és aquella que apunta a «bones pàgines de referència» i que una «bona pàgina de referència» es aquella que és apuntada per «bones pàgines guia». Per fer un símil, podríem pensar com el renom d'un restaurant s'enforteix quan les guies qualificades de restauració l'inclouen i com la reputació de les guies augmenta per l'encert de les seves recomanacions.

Ara veurem com la premissa anterior es tradueix en termes algorísmics. Per a cada pàgina i del conjunt S_T, $i = 1, \ldots, n$, denotem per $x_i^{(k)}$ i $y_i^{(k)}$ l'estimació de la seva valoració com a autoritat i com a recurs, respectivament, obtinguda en el pas k. Si denotem per $\Gamma^-(i)$ el conjunt de pàgines que apunten cap a la pàgina i i per $\Gamma^+(i)$ el conjunt de pàgines apuntades des de la pàgina i, llavors

$$\left.\begin{array}{rcl} x_i^{(k)} & = & \displaystyle\sum_{j\in\Gamma^-(i)} y_j^{(k-1)}, \\ y_i^{(k)} & = & \displaystyle\sum_{j\in\Gamma^+(i)} x_j^{(k)}, \\ x_i^{(k)} & := & \dfrac{x_i^{(k)}}{\|\boldsymbol{x}^{(k)}\|_\infty} \quad \text{i} \quad y_i^{(k)} := \dfrac{y_i^{(k)}}{\|\boldsymbol{y}^{(k)}\|_\infty}, \end{array}\right\} \; k = 1, 2, \ldots$$

$$y_i^{(0)} = 1, \; i = 1, \ldots, n\,.$$

Formulació matricial Sigui $\boldsymbol{A}$ la matriu d'adjacència del digraf $\Gamma[S_T]$, tal com l'hem definida a (4), i siguin $\boldsymbol{x}^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})^\top$ i $\boldsymbol{y}^{(k)} = (y_1^{(k)}, \ldots, y_n^{(k)})^\top$ els corresponents vectors de valoracions dels seus vèrtexs, d'acord amb la seva rellevància com a autoritats i com a pàgines recurs, respectivament. Aleshores,

$$\boldsymbol{x}^{(k)} = \boldsymbol{A}\boldsymbol{y}^{(k-1)} \quad \text{i} \quad \boldsymbol{y}^{(k)} = \boldsymbol{A}^\top \boldsymbol{x}^{(k)},$$

de manera que

$$\boldsymbol{y}^{(k)} = \frac{\boldsymbol{A}^\top \boldsymbol{x}^{(k)}}{\|\boldsymbol{A}^\top \boldsymbol{x}^{(k)}\|_\infty} = \frac{\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{y}^{(k-1)}}{\|\boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{y}^{(k-1)}\|_\infty} = \cdots = \frac{(\boldsymbol{A}^\top \boldsymbol{A})^k \boldsymbol{y}^{(0)}}{\|(\boldsymbol{A}^\top \boldsymbol{A})^k \boldsymbol{y}^{(0)}\|_\infty}\,.$$

Anàlogament,

$$\boldsymbol{x}^{(k)} = \frac{(\boldsymbol{A}\boldsymbol{A}^\top)^{k-1} \boldsymbol{x}^{(1)}}{\|(\boldsymbol{A}\boldsymbol{A}^\top)^{k-1} \boldsymbol{x}^{(1)}\|_\infty}\,.$$

Així, doncs, en el cas que les successions de vectors $\{\boldsymbol{x}^{(k)}\}_k$ i $\{\boldsymbol{y}^{(k)}\}_k$ convergeixin aquestes ho fan cap a unes 'direccions privilegiades', les corresponents a vectors propis (no negatius) de les matrius $\boldsymbol{A}\boldsymbol{A}^\top$ i $\boldsymbol{A}^\top\boldsymbol{A}$, respectivament, associades al valor propi dominant (radi espectral) d'aquestes dues matrius. Observem que les dues matrius esmentades són simètriques i semidefinides no negatives, ja que

$$\boldsymbol{x}^\top(\boldsymbol{A}\boldsymbol{A}^\top)\boldsymbol{x} = (\boldsymbol{A}^\top\boldsymbol{x})^\top(\boldsymbol{A}^\top\boldsymbol{x})\,.$$

Per tant, les dues matrius són diagonalitzables sobre els reals, via sengles bases ortogonals de vectors propis, i els seus valors propis són tots no negatius. Més encara, si $\boldsymbol{x}$ és un vector propi de $\boldsymbol{A}\boldsymbol{A}^\top$ associat a un valor propi $\lambda > 0$, aleshores $\boldsymbol{y} = \boldsymbol{A}^\top\boldsymbol{x}$ és un vector propi de $\boldsymbol{A}^\top\boldsymbol{A}$, associat al mateix valor propi λ, ja que

$$(\boldsymbol{A}^\top\boldsymbol{A})\boldsymbol{A}^\top\boldsymbol{x} = \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)\boldsymbol{x} = \lambda\boldsymbol{A}^\top\boldsymbol{x}$$

i $\boldsymbol{A}^\top\boldsymbol{x} \neq \boldsymbol{0}$, perquè en cas contrari

$$\lambda\|\boldsymbol{x}\|_2^2 = \langle\lambda\boldsymbol{x}, \boldsymbol{x}\rangle = \langle\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{x}, \boldsymbol{x}\rangle = 0\,.$$

Convergència La convergència de les successions $\{\boldsymbol{x}^{(k)}\}_k$ i $\{\boldsymbol{y}^{(k)}\}_k$ queda garantida pel fet que no poden haver-hi dos valors propis diferents sobre el cercle espectral, perquè són tots ells no negatius, i per haver pres, com a vector inicial, un vector positiu, el qual no pot ser ortogonal al subespai de vectors propis associats al valor propi dominant ja que almenys un d'aquests és no negatiu.

Si volguéssim que els vectors de valoracions $\boldsymbol{x}$ i $\boldsymbol{y}$ fossin positius, aleshores, pel teorema de Perron-Frobenius, bastaria que les matrius $\boldsymbol{A}\boldsymbol{A}^\top$ i $\boldsymbol{A}^\top\boldsymbol{A}$ fossin irreductibles o, equivalentment, que els seus grafs associats fossin connexos. I què significaria, per exemple, que el graf $\Gamma(\boldsymbol{A}^\top\boldsymbol{A})$ fos connex? Voldria dir que entre qualsevol parella de pàgines podríem formar una «cadena», on cada parell de pàgines consecutives de la mateixa comparteixen almenys un enllaç de sortida, és a dir, apunten cap a una mateixa pàgina.

7 Comentaris finals

D'on sorgeix la idea d'utilitzar vectors propis per a fer «ordenacions»? Segons indica Herbert Wilf [35], aquesta idea apareix per primer cop en els treballs de Kendall i Wei [21, 33], l'any 1950, dins l'àmbit de l'estadística. També s'havien emprat en altres àrees, com la geografia, per tal de mesurar la importància estratègica de diferents nodes de comunicació (vegeu [18, 31]). Actualment, són moltes les aplicacions que tenen en el camp de la teoria espectral de grafs, on es pretén esbrinar fins a quin punt l'espectre de la matriu d'adjacència d'un graf conté informació sobre la seva estructura (per fer-ne un bon tast us recomanem el treball de Miquel Àngel Fiol, [15], corresponent a la conferència plenària que va impartir al MAT.ES 2005 a València).

Agraïments

L'autor vol agrair als professors Armengol Gasull i Oriol Serra el suggeriment de fer aquest treball i la confiança rebuda. Així mateix, dóna les gràcies al revisor per les seves acurades correccions.

Referències

[1] ADLER, R.; EWING, J.; TAYLOR, P. «Citations statistics: A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS)». *Statist. Sci.*, 24 (1) (2009), 1–14.

[2] AVRACHENKOV, K.; DONATO, D.; LITVAK, N. (ed.) *Algorithms and models for the web graph*: Proceedings of the 6th International Workshop WAW, Barcelona, 2009. Berlín: Springer, 2009. (Lecture Notes in Computer Science; 5.427)

[3] ARASU, A.; CHO, J.; GARCIA-MOLINA, H.; PAEPCKE, A.; RAGHAVAN, S. «Searching the web». *ACM Trans. Internet Technology*, 1 (1) (2001), 2–43.

[4] BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern information retrieval*. Nova York: ACM Press, 1999.

[5] BERMAN, A.; PLEMMONS, R. J. *Nonnegative matrices in the mathematical sciences*. Nova York: Academic Press, 1979.

[6] BHARAT, K.; BRODER, A.; HENZINGER, M.; KUMAR, P.; VENKATASUBRAMANIAN, S. «The connectivity server: Fast access to linkage information on the web». A: *The seventh world wide web conference*. Brisbane: Elsevier Science, 1998, 469–477.

[7] BRIN, S.; MOTWANI, R.; PAGE, L.; WINOGRAD, T. «What can you do with a web in your pocket». *Data Engineering Bulletin*, 21 (1998), 37–47.

[8] BRIN, S.; PAGE, L. «The anatomy of a large-scale hypertextual web search engine». *Computer Networks and ISDN Systems*, 30 (1998), 107–117.

[9] BRODER, A.; KUMAR, R.; MAGHOUL, M. «Graph structure in the web». A: *The ninth international world wide web conference*. Nova York: ACM Press, 2000, 309–320.

[10] CHARTRAND, G.; LESNIAK, L. *Graphs & Digraphs*. Boca Raton, FL: Chapman & Hall/CRC, 2005.

[11] COMELLAS, F. «Models deterministes de xarxes complexes». *Butll. de la SCM*, 22 (1) (2007), 23–43.

[12] DAHLQUIST, G.; BJÖRCK, Å. *Numerical methods*. Nova Jersey: Prentice Hall, 1974.

[13] FAGIN, R.; KARLIN, A. R.; KLEINBERG, J.; RAGHAVAN, P.; RAJAGOPALAN, S.; RUBINFELD, R.; SUDAN, M.; TOMKINS, A. «Random walks with 'back buttons'». A: *32nd ACM Symposium on Theory of Computing*, 2000.

[14] FERNÁNDEZ, P. «El secreto de Google y el álgebra lineal». *Boletín de la Sociedad Española de Matemática Aplicada*, 30 (2004), 115–141.

[15] FIOL, M. A. «Aplicaciones de la teoría de matrices en problemas de matemática discreta» [en línia]. Disponible a Internet: http://www.uv.es/mat.es2005/Fiol-Mora.pdf

[16] FIOL, M. A.; ALEGRE I.; YEBRA, J. L. A.; FÁBREGA, J. «Digraphs with walks of equal length between vertices». A: *Graph theory with applications to algorithms and computer science*. Nova York: Wiley, 1985. 313–322.

[17] GANTMACHER, F. R. *The theory of matrices*. Vol. II. Nova York: AMS Chelsea Publishing, 1959.

[18] GOULD, P. «The geographical interpretation of eigenvalues». *Transactions of the Institute of British Geographers*, 42 (1967), 53–85.

[19] HAVELIWALA, T. H. «Efficient computation of Page Rank». Technical Report, 1999.

[20] JUNGNICKEL, D. *Graphs, networks and algorithms*. Berlín: Springer, 2008.

[21] KENDALL, M. G. «Further contributions to the theory of paired comparisons». *Biometrics*, 11 (1955), 43–62.

[22] KLEINBERG, J. M. «Authoritative sources in a hyperlinked environment». *J. ACM*, 46 (5) (1999), 604–632.

[23] LANGVILLE, A. N.; MEYER, C. D. *Google's PageRank and beyond*: *The science of search engine rankings*. Princeton: Princeton University Press, 2006.

[24] MACCLUER, C. R. «The many proofs and applications of Perron's theorem». *SIAM Review*, 42 (3) (2000), 487–498.

[25] MAURER, S. B.; RALSTON, A. *Discrete algorithmic mathematics*. Reading, Mass: Addison-Wesley, 1991.

[26] MCBRYAN, O. A. «GENVL and WWWW: Tools for taming the web». A: *Proceedings of the First International World Wide Web Conference*. Ginebra, 1994.

[27] MEYER, C. D. *Matrix analysis and applied linear algebra*. Filadèlfia: SIAM, 2000.

[28] PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. «The PageRank citation ranking: bringing order to the web». Technical Report. Stanford InfoLab, 1999.

[29] PIATETSKY, G. «Interview with Jon Kleinberg». *SIGKDD Explorations*, 9 (2) (2007), 47–50.

[30] RAGHAVAN, S.; GARCÍA-MOLINA, H. «Representing web graphs». A: *Proceedings of the 19th IEEE Conference on Data Engineering*. Bangalore (Índia), 2003.

[31] STRAFFIN, P. D. «Linear algebra in geography: Eigenvectors of networks». *Math. Mag.*, 53 (5) (1990), 269–276.

[32] VENTURA, E. *Endomorfismes de grups lliures finitament generats*. Tesi doctoral. UAB, 1995.

[33] WEI, T. H. *The algebraic foundations of ranking theory*. Cambridge: Cambridge Univ. Press, 1952.

[34] WIKIPEDIA. «Google bomb» [en línia]. Disponible a Internet: http://en.wikipedia.org/wiki/Google_bomb [Consulta: 5 de juliol 2010].

[35] WILF, H. «Searching the web with eigenvectors» [en línia]. Disponible a Internet: http://www.math.upenn.edu/~wilf/ [Consulta: 5 juliol 2010].

DEPARTAMENT DE MATEMÀTICA
ESCOLA POLITÈCNICA SUPERIOR
UNIVERSITAT DE LLEIDA
AVINGUDA JAUME II, 69, 25001 LLEIDA
joangim@matematica.udl.cat