

L'ÀLGEBRA LINEAL ÉS DARRERE DELS CERCADORS D'INTERNET

CARLOS D'ANDREA

*La vérité est trop compliqué.
Les mathématiques sont simples.*
Cédric Villani

1. ÀLGEBRA LINEAL PER A INFORMÀTICS

Els alumnes que es matriculin al grau d'Enginyeria en Informàtica que ofereix aquesta facultat, en el primer semestre del seu primer any d'estudis es trobaran amb l'assignatura “Àlgebra”, que entre els seus blocs temàtics ofereix el següent menú:

- Sistemes d'equacions lineals.
- Matrius i determinants.
- Espais vectorials. Subespais.
- Transformacions lineals. Nucli, imatge, isomorfismes.
- Polinomis.
- Nombres complexos.
- Diagonalització.

Qualsevol persona amb un mínim de coneixement en aquests temes s'adonarà ràpidament que la diagonalització és un procés que involucra *tots* els temes anteriors; i conclourà —amb bastant certesa— que aquest és un curs on s'aprèn a (decidir quan es pot) diagonalitzar matrius.

No hi ha res de trivial ni de sarcàstic en aquesta conclusió. És indubtable que l'àlgebra lineal en general —i el problema de càlcul de vectors i valors propis (que necessitem conèixer per decidir si una matriu és diagonalitzable) en particular— són molt importants en la informàtica, ja que estan presents en diversos processos centrals d'aquesta disciplina. Com a exemples, podem esmentar els següents:

- Agrupament i classificació de dades.
- Programació gràfica.
- Xarxes socials.
- Sistemes de recomanació (Netflix, PageRank, etc).
- Reconeixement de formes (cançons, empremtes digitals, fotografies).
- Intel·ligència artificial.

En el grau d'enginyeria en informàtica d'aquesta facultat, alguns d'aquests temes seran coberts al llarg de la carrera. Naturalment, els alumnes els veuran després d'haver acabat el curs d'àlgebra. És doncs raonable que no resulti gaire motivador per a l'alumnat aprendre a utilitzar unes eines que indubtablement

seran importants, però que ara com ara no podem explicar-los de quina manera ho seran, ni com s'utilitzaran.

És per aquest motiu que he triat presentar-lo en aquesta classe: per tal de motivar els alumnes que comencen a estudiar l'àlgebra que els oferim en aquesta Casa. I també per mostrar, als més avançats en ambdues carreres (matemàtica i informàtica), un problema de valors vectors propis (diagonalització) senzill d'enunciar que es'utilitzà recentment i amb molt èxit al món de la informàtica per resoldre un problema d'un dels tipus esmentats més amunt. Es tracta del problema de recomanació que tenen per davant els motors de cerca (o cercadors) d'internet a l'hora de suggerir a l'usuari quines pàgines li cal visitar com a resposta a unes certes paraules clau prèviament introduïdes pel mateix usuari en l'ordinador.

Per aquesta raó ens fixarem en un cercador específic, que és el més usat mundialment, i en l'algorisme que s'utilitzava inicialment i que s'ha seguit utilitzant fins a fa ben poc. Aquest algorisme va produir una veritable revolució al món del tràfic d'informació en línia.

¡I tot això gràcies a l'àlgebra lineal!

2. UN CERCADOR D'INTERNET MOLT RÀPID I EFICIENT

L'any 1996, dos joves alumnes de doctorat de la Universitat de Stanford (EUA), Sergei Brin i Lawrence Page, van començar a treballar en el disseny d'un cercador d'internet. Ambdós tenien 23 anys. Brin s'havia graduat en matemàtica i Page, en informàtica.

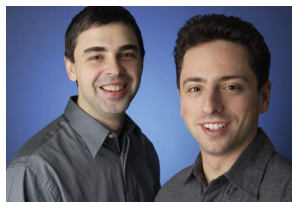


FIGURA 1. Sergei Brin (esquerra) i Larry Page (dreta)

L'algorisme que van dissenyar per a Google va ser denominat "PageRank", atès que Page havia començat inicialment el projecte al qual després s'incorporarà Brin (cf. [BP98]), i va acabar sent implementat per **Google**. En 1998 l'algorisme es va patentar, i al mateix temps va aparèixer en internet el cercador Google que també va ser dissenyat per Brin i Page. Des dels seus inicis, Google va utilitzar amb èxit aquest algorisme per posicionar-se des de molt primerenc (i fins al dia d'avui) com a líder en el mercat dels cercadors d'internet.

La paraula "google" és una variació fonètica del terme **googl** amb el qual es denomina (en anglès) el nombre 10^{100} . Els seus fundadors pretenien oferir un cercador ràpid i eficient. De fet, l'objectiu inicial de Brin i Page era que almenys *una* de les deu primeres pàgines que mostrés el cercador contingui informació útil per a la persona que la consulta.

No cal ara us expliqui l'èxit que ha tingut Google des dels seus inicis fins ara; sense cap mena de dubte es tracta del cercador d'internet més utilitzat a tot el món, superant marques de popularitat impensables. Per citar un exemple, el maig de 2011 va aconseguir superar els mil milions de visitants en un mes per primera vegada en la història de la informàtica. És sobrer dir que cap altre cercador d'internet ha aconseguit estar prop d'aquests nombres

Aquest succès també es tradueix òbviament en les finances, ja que l'any 2004 —quan va sortir a cotitzar al mercat de valors— la companyia estava valorada en aproximadament \$ 25.000.000.000, una xifra que anà creixent amb el pas del temps, aconseguint els \$ 37.905.000.000 al darrer balanç de 2011. I tot això per haver aconseguit diagonalitzar unes matrius...

Per intentar explicar breument l'algorisme PageRank i veure com hi apareixen de forma natural els vectors i valors propis, primer hem de veure com es modela matemàticament un cercador d'internet, ja que aquest algorisme forma part fonamental de la seva arquitectura.

3. ELS CERCADORS D'INTERNET

Es podria comparar el treball d'un cercador amb el d'un bibliotecari. Per fer-ho més explícit, diguem que es tracta d'un bibliotecari de les èpoques en què no hi havia ordinadors. En aquelles èpoques, cada cop més llunyanes, quan acudies a la biblioteca intentant de trobar informació d'algun tema en particular, et trobaves amb un fitxer molt voluminós o amb un catàleg enorme, imprecís, que contenien tota la informació existent a la biblioteca fins a la darrera actualització. Amb una mica de sort, hi havia també alguna espècie de catàleg-diccionari que relacionava els llibres amb algunes paraules clau.

Ara suposem m'apropés a una d'aquestes biblioteques antigues a la qual m'han enviat a investigar sobre el tema "girafa". No m'han donat cap referència bibliogràfica, i sé que la informació que em pugui proporcionar un diccionari o una enciclopèdia no em serà suficient. Què havia de fer? En aquells temps, la resposta més simple era: preguntar-ho al bibliotecari i després consultar les referències que t'hagi recomanat. Si no quedés satisfet amb les seves recomanacions, caldria o bé preguntar-li novament donant-li més precisió sobre el tema que estic buscant, o intentar-ho en una altra biblioteca.

Tota aquesta interacció amb el bibliotecari que estic explicant sembla gairebé trivial i podríeu preguntar-vos per què us estic fent perdre el vostre temps explicant aquesta història tan avorrida. Però suposem ara que la biblioteca conté més de mil milions de llibres, i que sota la paraula clau "girafa" hi ha quatre milions de textos que tenen alguna cosa a dir d'aquest tema, i que per enumerar d'un en un tots aquests textos —a raó de un llibre cada 10 segons— el bibliotecari trigaria uns 463 dies. Però és clar que no necessito llegir els quatre milions de referències per fer el treball que em van encarregar, potser amb 10 ja n'hi ha prou. Però llavors..., quins 10? L'algorisme PageRank és justament el que pot ajudar-me (o més aviat, ajudar al bibliotecari) a decidir de quina ha d'ordenar

la llista de sortida, quins són els llibres que m'ha de recomanar de manera que hi pugui trobar informació útil en les primeres referències que em proporcioni.

Essencialment un cercador d'internet és una espècie de “catàleg de biblioteca” juntament amb un “bibliotecari” que et recomana quins llibres cal llegir. L'èxit d'aquest cercador depèn precisament del fet de disposar d'una bona base de dades, ordenada d'acord a paraules clau d'una manera raonable, i també d'un bon criteri de recomanació, ja que cal accedir a aquesta informació de manera ràpida i eficient.

La tasca de “censar” les pàgines webs la fan per uns robots que circulen contínuament per la xarxa. Cal observar que aquest és un procediment dinàmic ja que hi ha milers de pàgines noves que apareixen a la xarxa minut a minut, i diverses (poques respecte de les noves) que desapareixen. I es desitja disposar d'informació actualitzada; així doncs aquest treball és molt important. Un altre element que també cal tenir en compte és que aquesta base de dades és enorme, i creix exponencialment. En 1998, any en què es llançà Google, tenia 26 milions de pàgines. En 2008 va aconseguir (cf. [Goo08]) el bilió (1.000.000.000.000) d'entrades.

El treball de “catalogar”, és a dir d'indexar les dades censades en funció de certes paraules clau també el fan programes informàtics que estudien distribucions estadístiques de paraules, freqüències d'aparició i enllaços a aquesta pàgina per fer aquesta feina.

O sigui que tot cercador d'internet ha de tenir tant un bon catàleg de pàgines indexades com un bon índex pel que fa a les paraules clau. Bé... Com es fa el treball de bibliotecari? En altres paraules, com decideixo quines pàgines mostrar primer quan algú posa en el cercador la paraula “girafa”?

Hi ha milers d'algorismes i programes dedicats a respondre aquesta pregunta i, entre ells, l'algorisme PageRank que és el que va catapultar Google a l'èxit entre els cercadors d'internet. Dedicarem la secció següent a explicar-ho.

4. EL MODEL PAGERANK. VECTORS I VALORS PROPIS

Tal com hem explicat fins ara, per completar el treball del cercador caldria assignar-li una “importància” a cada una de les pàgines web que tinc censades. La *teoria de graf*s ens ajudarà a modelar la nostra situació.

En el model PageRank, l'univers de les pàgines web indexades és un gran *graf dirigit*, on cada pàgina web censada és un node, i hi haurà una “fletxa” (*aresta orientada*) que va de la pàgina p_i a la pàgina p_j si hi ha un enllaç des de la primera pàgina cap a la segona. Per exemple, si el gràfic de la figura 2 és el que a partir d'ara denominarem *el graf d'internet*, llavors del dibuix podríem concloure que —per exemple— la primera pàgina és la més popular ja que hi ha enllaços des de totes les altres cap a ella, i és l'única que compleix aquesta propietat.

En aquest gran graf dirigit, hem d'assignar una “importància” a cada pàgina. Una manera raonable d'assignar importàncies podria ser: quants més enllaços rep una pàgina més important és. Observem l'analogia que hi ha amb el famós —però

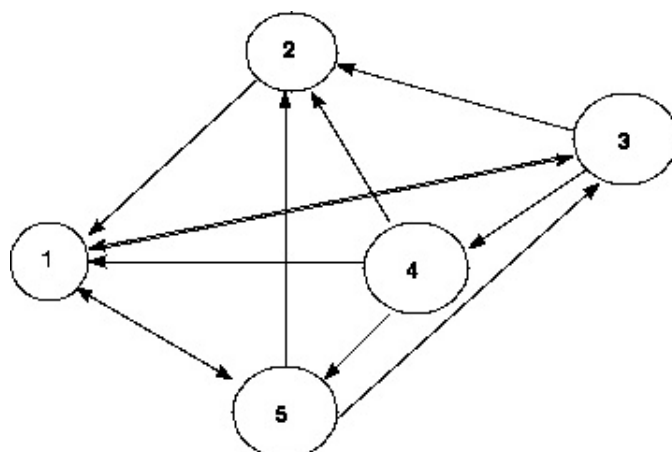


FIGURA 2. Un graf dirigit

no sempre tan benvolgut— índex de cites que intenta de regular el currículum dels investigadors.

El problema que té aquest model, aparentment assenyat, és que hom podria “inflar” ràpidament la importància d’una pàgina web determinada simplement creant diverses pàgines que enllaçades amb ella, i aquest procediment és molt fàcil d’implementar en pocs minuts, la qual cosa faria que tot el sistema fos molt fàcil d’influenciar.

Per evitar-ho, canviarem la funció “nombre de cites” per “importància de les cites”. És a dir, no solament donarem importància a la quantitat de cites que té una pàgina donada, sinó que a més tindrem en compte si la citen pàgines importants. Diguem que —per exemple— si tinc enllaços des de *Amazon.com* o *Microsoft.com*, la meua importància hauria de ser major. En aquest sentit, el graf de les pàgines web seria una mica més semblant al de la figura 3, que mostra una distribució d’importàncies relatives a les importàncies de les pàgines donades. En ell s’entén per què la pàgina “C” és més important que la “F” atès que cadascuna rep un enllaç, però la primera ho està per una pàgina molt més important que la segona.

En paraules, el “postulat” del model PageRank diu el següent:

La importància x_i de la pàgina p_i és directament proporcional a la suma de les importàncies de les pàgines que enllacen amb ella.

Vegem com es tradueix això matemàticament i com, en aquest context i d’una forma natural, apareix l’àlgebra lineal. El dibuix d’un graf (dirigit o no) és il·lustratiu i interessant si es tracta de pocs nodes, però el graf d’internet té més d’un bilió de pàgines així que no hi guanyarem gaire si intentem dibuixar-ho (i perdrem molt de temps i tinta). En lloc d’això, considerarem la matriu d’incidència del graf, la qual es defineix com la matriu quadrada de dimensió igual a la quantitat de nodes del graf. En aquesta matriu, posem un 1 en el lloc (i, j) , si hi ha un enllaç des de p_i fins a p_j ; si no ho hi ha, posem un zero. Per

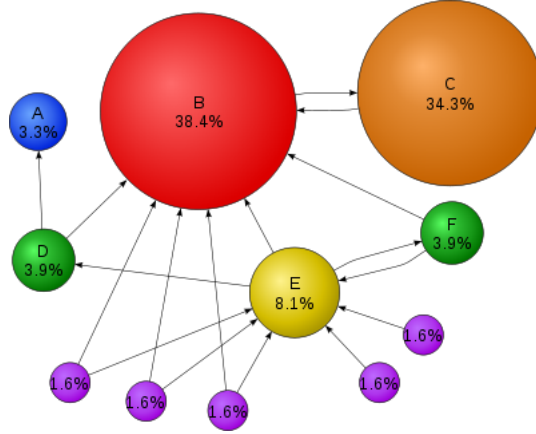


FIGURA 3. El graf d'“importàncies”

exemple, la matriu d'incidència del graf d'internet de la figura 2 és la següent:

$$\mathbb{M}_0 = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \end{pmatrix}.$$

I aquí és on l'àlgebra lineal apareix juntament amb els vectors i valors propis.

Teorema 4.1. *Si $\mathbb{M}_I$ és la matriu d'incidència del graf d'internet, i $\mathbf{x} = (x_1, \dots, x_N) \in \mathbb{R}_{\geq 0}^n$ el vector d'importàncies, llavors*

$$\mathbb{M}_I^t \mathbf{x}^t = \lambda \cdot \mathbf{x}^t,$$

on $\lambda \in \mathbb{R}_{>0}$ és la constant de proporcionalitat.

Aquest és un bon moment per recordar algunes definicions clàssiques de l'àlgebra lineal.

Definició 4.2. Donats una matriu quadrada $\mathbb{M}$ de dimensió n , un vector no nul $\mathbf{x} \in \mathbb{R}^n$ (o $\mathbb{C}^n$) i $\lambda \in \mathbb{R}$ (o $\mathbb{C}$), el vector $\mathbf{x}$ és un *vector propi* de $\mathbb{M}$ amb *valor propi* associat λ si i solament si es verifica

$$\mathbb{M}\mathbf{x}^t = \lambda \cdot \mathbf{x}^t.$$

Corol·lari 4.3. *El vector d'importàncies de les pàgines web és un vector propi (positiu) de la matriu $\mathbb{M}_I^t$, i la constant de proporcionalitat λ és el valor propi associat a aquest vector.*

Exemple 4.4. *Anem a veure com efectivament el que diu el postulat PageRank i el que enunciem en el Teorema 4.1 coincideixen. Diguem que λ és la constant*

de proporcionalitat que apareix en l'enunciat del postulat. Llavors, d'acord amb aquesta afirmació, tenim les següents equacions:

$$\begin{cases} 0 \cdot x_1 + 1 \cdot x_2 + 1 \cdot x_3 + 1 \cdot x_4 + 1 \cdot x_5 = \lambda x_1 \\ 0 \cdot x_1 + 0 \cdot x_2 + 1 \cdot x_3 + 1 \cdot x_4 + 1 \cdot x_5 = \lambda x_2 \\ 1 \cdot x_1 + 0 \cdot x_2 + 0 \cdot x_3 + 1 \cdot x_4 + 1 \cdot x_5 = \lambda x_3 \\ 0 \cdot x_1 + 0 \cdot x_2 + 1 \cdot x_3 + 0 \cdot x_4 + 0 \cdot x_5 = \lambda x_4 \\ 1 \cdot x_1 + 0 \cdot x_2 + 0 \cdot x_3 + 1 \cdot x_4 + 0 \cdot x_5 = \lambda x_5, \end{cases}$$

que en notació matricial és, precisament,

$$\mathbb{M}_0^t \mathbf{x}^t = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \end{pmatrix}^t \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \lambda \cdot \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} = \lambda \cdot \mathbf{x}^t.$$

Observem que el que hem vist en aquest exemple no és casual, ja que si en les files de la matriu d'incidència del graf d'internet s'hi llegeix quants enllaços surten d'una pàgina donada, en les columnes apareixen justament tants uns com enllaços hi hagi cap a la pàgina indexada per aquesta columna. És per això que es necessita transposar la matriu per utilitzar-la en el problema del PageRank.

Una vegada enunciat el problema... homn vol trobar-ne la solució. Com ho faríem per al nostre graf de la figura 2? Utilitzant alguna de les eines computacionals que tenim a disposició (per exemple, el programa *Mathematica* que és el que s'utilitza en les pràctiques de laboratori de l'assignatura d'àlgebra), ens trobem amb el següent:

- Possibles valors (aproximats) per a λ (valors propis de $\mathbb{M}_0^t$):

2,27; $-0,500 - 0,866i$; $-0,500 + 0,866i$; $-0,635 + 0,692i$; $-0,635 - 0,692i$

- Possibles valors (aproximats) per als vectors propis (ordenats pel que fa als valors propis enumerats a dalt):

(1,74; 1,21; 1,21; 0,532; 1,00),
 (0, 0; $-0,500 - 0,866i$; $-0,500 + 0,866i$; 1,00),
 (0, 0; $-0,500 + 0,866i$; $-0,500 - 0,866i$; 1,00),
 ($-0,469 + 0,101i$; $-0,303 - 0,490i$; $-0,303 + 0,490i$; $-0,166 + 0,591i$; 1,00),
 ($-0,469 - 0,101i$; $-0,303 + 0,490i$; $-0,303 - 0,490i$; $-0,166 - 0,591i$; 1,00)

En aquest exemple, el del graf associat a la figura 2, atès que la constant de proporcionalitat ha de ser real i positiva, ens semblaria que hi ha una única solució al problema:

- $\lambda = 2,27$
- $x_1 = 1,74$, $x_2 = 1,21$, $x_3 = 1,21$, $x_4 = 0,532$, $x_5 = 1$,

la qual cosa sembla ser una resposta raonable ja que la primera pàgina és votada per totes les altres, i és l'única amb aquesta situació. O sigui que mereix guanyar

la competència, i si bé la segona té un vot més que la tercera, aquesta última és votada per la més important (la primera) mentre que l'altra no ho és.

Podríem suposar que aquest fenomen és un fet general: en qualsevol matriu quadrada amb zeros i uns hi haurà un únic valor propi positiu, i associat a ell un sol vector propi positiu que serà la solució al nostre problema. Lamentablement aquesta conjectura no és certa, ja que —per exemple— una matriu tan senzilla com $\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ té com a únic valor propi $\lambda = 0$ únicament. També és fàcil trobar matrius més grans amb més d'un valor propi positiu. Llavors... com ens ho podem fer per resoldre aquesta ambigüitat?

Abans de respondre la pregunta, farem una modificació petita però de vital importància al model. Tal com hem explicat fins a aquí, aquest model fou el que originalment s'utilitzà en el cercador Google durant els seus primers anys. En el decurs del temps i amb l'aparició de les xarxes socials (molt propenses a produir fets “en cadena”, en diversos llocs i al mateix temps) es va trobar una falla en el model —la qual era previsible des d'un primer moment: si una pàgina té un sol enllaç, aquest enllaç val el mateix que qualsevol altre enllaç d'una altra pàgina que produeixi un milió d'enllaços. És —si bé produir enllaços des de la meva pròpia pàgina no augmenta la meva importància— com si quants més enllaços produeix la meva pàgina, més afecta a tota la xarxa.

Per evitar aquest excés d'autoritat, el model es va modificar senzillament de la manera següent: si hi ha un enllaç des de p_i cap a p_j , en el lloc (i, j) de la matriu d'incidència es col·loca el nombre $\frac{1}{\# \text{enllaços des de } p_i}$. D'aquesta manera, cada pàgina té “poder de vot” igual a 1, i aquesta unitat es va distribuint d'acord amb els nombre d'enllaços que té.

Per exemple, la matriu modificada associada al graf de la figura 2, que denotarem amb $\mathbb{M}_{0,E}$, es la següent:

$$\mathbb{M}_{0,E} = \begin{pmatrix} 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 1 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \end{pmatrix}.$$

Notar qu en cada fila de $\mathbb{M}_{0,E}$ hi ha una distribució de nombres no negatius que sumen 1, com si fos una distribució de probabilitats sobre els nodes del graf d'internet. Aquest tipus de matrius es coneix com a *estocàstica per files*, i apareix amb freqüència en el modelatge de processos diversos que esmentarem més endavant.

De moment, calculem els valors i vectors propis de la matriu estocàstica.

- Possibles valors (aproximat) para λ (valors propis de $\mathbb{M}_{0,E}^t$):

$$1,00; -0,333 + 0,471i; -0,333 - 0,471i; -0,167 + 0,289i; -0,167 - 0,289i$$

- *Possibles valors (aproximatats) per als vectors propis (ordenats pel que fa als valors propis enumerats a dalt):*

(1,73; 0,867; 1,20; 0,400; 1,00),
 $(-0,333 + 0,943i; -0,667 - 0,236i; 0,500 - 0,707i; -0,500; 1,00)$,
 $(-0,333 - 0,943i; -0,667 + 0,236i; 0,500 + 0,707i; -0,500; 1,00)$,
 (0; 0; $-0,500 - 0,866i$; $-0,500 + 0,866i$; 1,00),
 (0; 0; $-0,500 + 0,866i$; $-0,500 - 0,866i$; 1,00).

En aquest cas sembla també que hi una única solució “real”: $\lambda = 1$ i

$$x_1 = 1,73; x_2 = 0,867; x_3 = 1,20; x_4 = 0,400; x_5 = 1,00.$$

Noteu la diferència subtil que hi ha entre les dues distribucions d'importància. Mentre que en el primer model les pàgines 2 i 3 reben la mateixa importància, en el segon la tercera pàgina supera a la segona, sent que la tercera rep solament dos vots i la segona tres. El motiu d'aquesta diferència pot ser explicat pel fet que la pàgina 3 no solament és votada per la primera pàgina que és la més important, sinó que a més la primera pàgina solament emet dos vots, mentre que totes les que voten a la pàgina 3 emeten 3 vots. És a dir, aquests vots pesen menys que els de la primera pàgina.

Noteu també que el fet que 1 sigui un valor propi de la matriu estocàstica no és casualitat, ja que és fàcil comprovar que el vector $(1, 1, \dots, 1)$ és sempre un vector propi de tota matriu estocàstica per columnes, associat al valor propi $\lambda = 1$.

5. SOLUCIÓ DEL PROBLEMA... UNICITAT?

La resposta al problema que planteja el model PageRank ve de la mà d'una altra branca de la matemàtica que és l'Anàlisi Funcional. El primer resultat en aquesta direcció el donà Oskar Perron a principis del segle passat.

Teorema 5.1 (Perron, 1907).

*Sigui $\mathbb{M}$ una matriu quadrada amb tots els seus coeficients **positius**. Llavors*

- *existeix un valor propi simple $\lambda > 0$ per al qual $\mathbb{M} \cdot \mathbf{v}^t = \lambda \cdot \mathbf{v}^t$, on $\mathbf{v}$ és un vector propi que té totes les seves coordenades **positives**,*
- *aquest valor propi és major, en mòdul, que tots els altres valors propis de $\mathbb{M}$,*
- *qualsevol altre vector propi positiu de $\mathbb{M}$ és un múltiple de $\mathbf{v}$*

Aquest teorema ens comporta una certa unicitat que consistiria a quedar-nos amb l'únic vector propi positiu de la matriu, l'associat al valor propi més gran de tots (en mòdul). Lamentablement, les matrius $\mathbb{M}_I$ estan associades als grafs de pàgines d'internet, i tenen molts zeros. Estan molt lluny de ser positives. Necessitaríem que s'enllacessin totes les pàgines amb totes, fins i tot amb elles mateixes!

Un resultat una mica més general sense restriccions sobre les matrius és impossible com ens ensenya l'exemple de la matriu $\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ que hem vist abans.

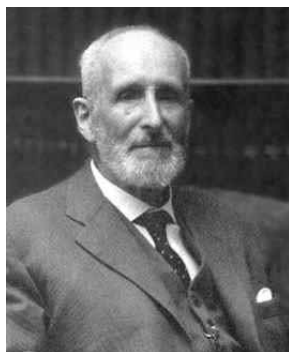


FIGURA 4. Oskar Perron (1880–1975)

No obstant això, Frobenius va aconseguir una versió del Teorema de Perron per a matrius no negatives, sotmeses a una condició addicional sobre la matriu d'entrada. Enunciem primer el resultat i vegem després les restriccions que ens imposa, potser amb l'esperança que el graf d'internet les compleixi.

Teorema 5.2 (Frobenius, 1908–1912).

*Sigui M una matriu quadrada amb tots els seus coeficients **no negatius**. Si M és **irreductible**, llavors*

- *existeix un valor propi simple $\lambda > 0$ per al qual $M \cdot v^t = \lambda \cdot v^t$, on v és el vector propi associat, i té totes les seves coordenades **positives**,*
- *aquest valor propi és **major o igual**, en mòdul, que tots els altres valors propis de M ,*
- *qualsevol altre vector propi positiu de M és un múltiple de v .*



FIGURA 5. Georg Frobenius (1849–1917)

Definició 5.3. Una matriu quadrada M és **irreductible** si no existeix cap permutació de les seves files i columnes que la transformi en una matriu del tipus

$$\begin{pmatrix} M_{11} & M_{12} \\ \mathbf{0} & M_{22} \end{pmatrix},$$

on M_{11} i M_{22} són matrius quadrades.

Proposició 5.4. *Si M és la matriu d'incidència d'un graf dirigit, llavors que sigui irreductible és equivalent al fet que el graf sigui “fortament connex”, és a dir que, donats dos nodes qualssevol del graf, sigui possible trobar una successió d'arestes que portin d'un a un altre.*

Aquesta proposició és de fet interessant ja que no és trivial calcular una descomposició que validi la “irreductibilitat” d'una matriu, però no obstant això —per a un nombre de nodes relativament petit— és immediat verificar si un graf és fortament connex o no. Com a exercici per al lector deixem el de mostrar que el graf de la figura 2 és fortament connex.

De totes maneres, és altament improbable que el graf d'internet sigui fortament connex. De fet, un estudi fet en 1999 (cf. [Bro99]) mostrava una distribució del graf d'internet com es veu en la figura 6. De 203 milions de pàgines que s'havien censat en aquell moment, un 90 % estava en una gegantesca component connexa, i solament 56 milions estaven connectats “fortament”.

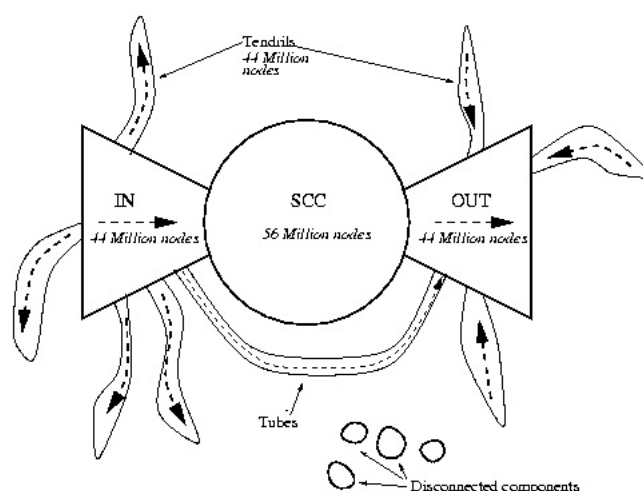


FIGURA 6. Components connexes del graf d'internet en 1999

O sigui que ni el Teorema de Perron ni el de Frobenius es poden aplicar directament al nostre problema. Què fem, llavors? Què fa Google? La sortida a aquest aparent atzucac ve de la mà d'una “pertorbació”, una operació molt freqüent en la *matemàtica computacional* i l'àlgebra lineal numèrica, on és habitual treballar amb dades aproximades. Aquí el que farem serà una cosa molt ingenu però eficient, reemplaçarem la nostra matriu estocàstica que denotarem $M_{I,E}$ per una matriu a la qual farem positius tots els coeficients sumant-li una matriu positiva convenient. El principi subjacent a aquesta idea és que la funció “importància” és contínua, i si puc calcular-la “a prop” de la situació on em trobo, ja em serveix per al que vull (que és ordenar les importàncies i no realment calcular-les).

En símbols, donat $\varepsilon > 0$, molt petit, definim

$$\mathbb{M}_{I,E}^\varepsilon := (1 - \varepsilon) \cdot \mathbb{M}_{I,E} + \frac{\varepsilon}{n} \cdot \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}.$$

Noteu que aquesta operació retorna una matriu amb totes les entrades positives (majors o iguals que 0), que satisfà les hipòtesis del Teorema de Perron, i llavors “declarem” que “la” solució al problema del PageRank és la que s’obté segons aquest enunciat per a un ε prefixat (en la pràctica, Google utilitza $\varepsilon = 0,15$).

Exemple 5.5. *Calculem la matriu “pertorbada” del nostre graf 2:*

$$\mathbb{M}_{0,E}^\varepsilon = \begin{pmatrix} 0,03 & 0,03 & 0,455 & 0,03 & 0,455 \\ 0,88 & 0,03 & 0,03 & 0,03 & 0,03 \\ 0,313 & 0,313 & 0,03 & 0,313 & 0,03 \\ 0,313 & 0,313 & 0,03 & 0,03 & 0,313 \\ 0,313 & 0,313 & 0,313 & 0,03 & 0,03 \end{pmatrix}.$$

Si ara demanem a **Mathematica** que calculi el vector propi positiu d’aquesta matriu (associat al valor propi $\lambda = 1$) obtenim

$$(0,67259; 0,363478; 0,463318; 0,194141; 0,403921).$$

Aquest vector d’importàncies induïx el mateix ordre que el que produeix la matriu sense pertorbar. És a dir, que abans i després de la pertorbació apareix aquest ordre entre les pàgines:

$$x_1 > x_3 > x_5 > x_2 > x_4.$$

6. LA SOLUCIÓ COMPUTACIONAL

Tot sembla molt bonic i agradable quan fem càlculs amb una matriu 5×5 , però si volem calcular la solució “real” al problema de Google, amb matrius amb més d’un bilió d’entrades, com es fa? Com ho fa Google?

Des d’ara desaconsellem a qualsevol optimista intentar d’utilitzar les tècniques apreses (o per aprendre) en el curs d’àlgebra que involucrarien calcular un polinomi característic de grau major que un bilió, trobar-ne totes les arrels, i detectar entre totes elles la que té mòdul màxim, i després resoldre un sistema lineal enorme per calcular el vector d’importàncies. Fins i tot si coneixem el valor de λ (que en els casos estocàstics és 1), resoldre un sistema lineal de l’ordre d’un bilió és una tasca monstruosa que no pot ser realitzada en poc temps ni tan sols pels ordinadors més ràpids que hi ha disponibles en aquest moment.

Google utilitza el que es coneix com el *mètode de les potències*, en aparença massa fàcil d’enunciar però computacionalment molt efectiu. Es basa en el següent fet bastant simple: si una matriu quadrada $\mathbb{M}$ és diagonalitzable y té tots els vectors propis $\{\mathbf{v}_1, \dots, \mathbf{v}_n\}$ enumerats de manera que els valors propis corresponentstinguin la propietat següent:

$$\lambda_1 > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_n|,$$

Començant amb un $\mathbf{v}_0 > 0$ qualsevol que

$$\mathbf{v}_0 = \alpha_1 \mathbf{v}_1 + \dots + \alpha_n \mathbf{v}_n,$$

amb $\alpha_1 \neq 0$, llavors

$$\mathbb{M}^k \mathbf{v}_0 = \alpha_1 \lambda_1^k \mathbf{v}_1 + \dots + \alpha_n \lambda_n^k \mathbf{v}_n.$$

D'ací doncs,

$$\lim_{k \rightarrow \infty} \frac{\mathbb{M}^k \mathbf{v}_0}{\lambda_1^k} = \alpha_1 \mathbf{v}_1,$$

és a dir, en el límit aconseguim un múltiple no trivial del vector propi que estem buscant.

Aquest és el mètode que utilitza Google per tal d'ordenar les seves pàgines d'internet, i amb resultats bastant raonables. Una anàlisi amb més detall de la velocitat de convergència i aspectes relacionats es pot trobar a [LSW09, S-C05, Wil07].

7. GOOGLEPÍLEG

Avui dia, l'algorisme PageRank és marca registrada de Google, i està patentat als Estats Units. A causa de temes legals, la patent està assignada a la Universitat de Stanford i no pas a Google. No obstant això, la companyia d'internet té drets exclusius sobre aquesta patent, i Stanford va rebre 1.800.000 accions de Google per permetre-li'n l'ús exclusiu.

Una versió modificada del PageRank fou proposada recentment (cf. [BRSR06]) com a alternativa al polèmic factor d'impacte elaborat pel ISI. Una implementació d'aquesta proposta es pot trobar en <http://www.eigenfactor.org>. També s'ha aplicat per predir concentracions humanes en carrers o places (cf. [Jia06]), en models d'evolució en ecosistemes (cf. [AP09]), en altres tipus de cerques d'internet, i fins i tot en anàlisi de xarxes de proteïnes (cf. [IG11]).

PageRank va ser utilitzat per Google fins fa ben poc. El febrer de 2011 la companyia va començar a fer proves d'un nou algorisme de cerca denominat "Google Panda", que essencialment té la capacitat de modificar la importància de seccions senceres de la llista, i no solament de pàgines individuals. Google Panda va reemplaçar definitivament el PageRank l'abril de 2011, i continua ajudant a Google a liderar el mercat de cercadors d'internet fins al dia d'avui.

8. QUÈ HEM APRÈS AVUI?

Un podria sentir-se una mica enganyat després d'aquesta presentació de l'algorisme PageRank, atès que al principi hem promès veure una aplicació senzilla dels vectors i valors propis, que ens ha estat d'utilitat per modelar el problema del PageRank. Però després, per resoldre aquest problema, el nostre camí s'ha convertit en un veritable "tour de force" per diverses branques de la matemàtica: teoria de grafs, anàlisi funcional, càlcul numèric, matrius estocàstiques, matemàtica computacional...

Segons la Wikipedia [Wiki], un enginyer és algú que resol *problemes que afecten l'activitat quotidiana de la societat*. I més endavant afegeix *l'enginyeria és l'activitat de transformar el coneixement en alguna cosa pràctica*.

El treball de Brin i Page és un bon exemple d'això, ja que qualsevol persona que vulgui treballar resolent problemes necessita de recursos, d'eines. I cadascuna d'aquestes àrees de la matemàtica i de la informàtica ha de ser per a l'enginyer precisament això, una eina. I com més eines tinguem, millor.

8.1. Per saber més... i per tal d'aprofundir les idees i resultats matemàtics i computacionals que es troben al voltant de l'algorisme PageRank, suggerim els articles [BL06, Fer04, Gim11, LM06, MGF06, Wil06], i molts més que es troben en les referències bibliogràfiques d'aquests treballs. I segurament si es cerca en Google alguna d'aquestes paraules clau... trobareu molt més per llegir!

Agraïments: Agraïixo a David Coix per haver-me introduït en aquest fascinant món de l'àlgebra lineal de Google, i també els suggeriments i comentaris de Teresa Cortadellas, Emiliano Gómez, Gabriela Jerónimo, Pablo Mislej, Adrián Paenza i Juan Pablo Pinasco a una versió preliminar d'aquestes notes.

REFERENCIAS

- [AP09] Allesina, Stefano; Pascual, Mercedes. *Googling Food Webs: Can an Eigenvector Measure Species' Importance for Coextinctions?* PLOS Computational Biology, Issue 9, Volume 5, September 2009. <http://dx.plos.org/10.1371/journal.pcbi.1000494>
- [BRSR06] Johan Bollen, Marko A. Rodriguez, and Herbert Van de Sompel.; Rodriguez; Van De Sompel. *Journal Status* . Scientometrics 69 (3): 1030, December 2006). <http://arxiv.org/abs/cs/0601030>
- [BP98] Brin, S. and Page, L. *The Anatomy of a Large-Scale Hypertextual Web Search Engine*. In: Seventh International World-Wide Web Conference (WWW 1998), April 14–18, 1998, Brisbane, Australia.
- [Bro99] Broder, A. et all. *Graph structure in the web*. <http://www9.org/w9cdrom/160/160.html>
- [BL06] Bryan, Kurt; Leise, Tanya. *The \$ 25,000,000,000 eigenvector: the linear algebra behind Google*. SIAM Rev. 48 (2006), no. 3, 569–581
- [CS10] Cicone, Antonio; Serra-Capizzano, Stefano. *Google PageRanking problem: the model and the analysis*. J. Comput. Appl. Math. 234 (2010), no. 11, 3140–3169.
- [Fer04] Fernández, Pablo. *El secreto de Google y el Álgebra lineal*. Bol. Soc. Esp. Mat. Apl. Nro. 30 (2004), 115–141.
- [Gim11] Gimbert, Joan. *The mathematics of Google: the PageRank algorithm*. Butl. Soc. Catalana Mat. 26 (2011), no. 1, 29–55, 97.
- [Goo08] Google official Blog. *We knew the web was big...* <http://googleblog.blogspot.com.es/2008/07/we-knew-web-was-big.html>
- [IG11] Ivan, G; Grolmusz, V. *When the Web meets the cell: using personalized PageRank for analyzing protein interaction networks*. Bioinformatics (Vol. 27, No. 3. pp. 405–407) 27 (3): 405–7, 2011.
- [IW06] Ipsen, Ilse C. F.; Wills, Rebecca S. *Mathematical properties and analysis of Google's PageRank*. Bol. Soc. Esp. Mat. Apl. No. 34 (2006), 191–196.
- [Jia06] Jiang, B. *Ranking spaces for predicting human movement in an urban environment*. International Journal of Geographical Information Science 23 (7), 2006, 823–837.

- [LM06] Langville, Amy N.; Meyer, Carl D. *Google's PageRank and beyond: the science of search engine rankings*. Princeton University Press, Princeton, NJ, 2006. x+224 pp. ISBN: 978-0-691-12202-1; 0-691-12202-4
- [LSW09] Lin, Yiqin; Shi, Xinghua; Wei, Yimin. *On computing PageRank via lumping the Google matrix*. J. Comput. Appl. Math. 224 (2009), no. 2, 702–708.
- [MGF06] Madrid de la Vega, Humberto; Guerra Ones, Valia; Flores Garrido, Marisol. *The numerical linear algebra of Google's PageRank*. Papers of the Mexican Mathematical Society (Spanish), 33–52, Aportaciones Mat. Comun., 36, Soc. Mat. Mexicana, México, 2006.
- [S-C05] Serra-Capizzano, Stefano. *Jordan canonical form of the Google matrix: a potential contribution to the PageRank computation*. SIAM J. Matrix Anal. Appl. 27 (2005), no. 2, 305–312.
- [Wiki] <http://es.wikipedia.org/wiki/Ingenier%C3%ADa>
- [Wil06] Wills, Rebecca S. *Google's PageRank: the math behind the search engine*. Math. Intelligencer 28 (2006), no. 4, 6–11.
- [Wil07] Wills, Rebecca S. *When rank trumps precision: Using the power method to compute Google's PageRank*. Thesis (Ph.D.) – North Carolina State University. 2007. 110 pp. ISBN: 978-0549-19626-6

UNIVERSITAT DE BARCELONA, DEPARTAMENT D'ÀLGEBRA I GEOMETRIA. GRAN VIA 585,
08007 BARCELONA, SPAIN

E-mail address: cdandrea@ub.edu

URL: <http://atlas.mat.ub.es/personals/dandrea/>