

LLIÇÓ INAUGURAL
DEL CURS ACADÈMIC
2012-2013

FACULTAT DE MATEMÀTIQUES

L'àlgebra lineal és darrere dels cercadors d'internet

Carlos D'Andrea

PROFESSOR DEL DEPARTAMENT D'ÀLGEBRA
I GEOMETRIA

Àlgebra lineal per a informàtics

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

- 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 coneixements d'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 en què s'aprèn a (decidir quan es pot) diagonalitzar matrius.

No hi ha res trivial ni sarcàstic en aquesta conclusió. És indubtable que l'àlgebra lineal, en general, i el problema del 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 són presents en diversos processos centrals d'aquesta disciplina. Com a exemples, en 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 Informàtica d'aquesta facultat, alguns d'aquests temes s'impartiran al llarg de la carrera. Naturalment, els alumnes els veuran després d'haver acabat el curs d'Àlgebra. És raonable, doncs, 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 lliçó: 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 vectors i valors propis (per diagonalització) senzill d'enunciar, que s'ha utilitzat recentment i amb molt d'èxit en el món de la informàtica per resoldre un problema d'un dels tipus esmentats més amunt. Es tracta del problema de recomanació que tenen 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.

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 continuat utilitzant fins fa ben poc. Aquest algorisme va produir una veritable revolució en el món del trànsit d'informació en línia. I tot això gràcies a l'àlgebra lineal!

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.

L'algorisme que van dissenyar per a Google s'anomenà PageRank, atès que Page havia començat inicialment el projecte al qual després s'incorporà Brin (cf. [BP98]), i va acabar sent implementat per Google. Al 1998 l'algorisme es va patentar, i al mateix temps va aparèixer a internet el cercador Google, que també va ser dissenyat per Brin i Page. Des del principi, Google va utilitzar amb èxit aquest algorisme per posicionar-se

de manera molt primerenca (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 contingués informació útil per a la persona que feia la consulta.

No cal que ara us expliqui l'èxit que ha tingut Google des dels seus inicis fins a l'actualitat; sense cap mena de dubte és el cercador d'internet més utilitzat a tot el món i ha superat marques de popularitat impensables. Per citar-ne un exemple, el maig del 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 no ha aconseguit estar a prop d'aquestes xifres.

Aquest èxit 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, fins arribar a 37.905.000.000 \$ al darrer balanç del 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.

Els cercadors d'internet

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

Ara suposem que m'apropo 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 haig de fer? En aquells temps, la resposta més simple era: preguntar-ho al bibliotecari i després consultar les referències que em recomanés. Si no quedava satisfet amb les seves recomanacions, calia o bé preguntar-li novament i amb més precisió sobre el tema que estava buscant, o intentar-ho en una altra biblioteca.

Tota aquesta interacció amb el bibliotecari que estic explicant sembla gairebé trivial i podrieu preguntar-vos per què us hi faig perdre el temps. Però suposem ara que la biblioteca conté més de mil milions de llibres, i que amb la paraula clau «girafa» hi ha quatre milions de textos que tenen alguna cosa a dir sobre aquest tema, i que per enumerar d'un en un tots aquests textos —a raó d'un llibre cada deu 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 deu ja n'hi ha prou. Però llavors..., quines deu? L'algorisme PageRank és justament el que pot ajudar-me (o, més aviat, ajudar el bibliotecari) a decidir de quina forma ha d'ordenar la llista de sortida, i quins són els llibres que m'ha de recomanar de manera que hi pugui trobar informació útil en les primeres referències que em proporioni.

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 que es disposa d'una bona base de dades, ordenada d'acord amb 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 web la fan 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 vol 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. Al 1998, l'any en què es llançà Google, tenia 26 milions de pàgines. Al 2008 va aconseguir un bilió d'entrades (cf. [Goo08]).

La tasca de «catalogar», és a dir, d'indexar les dades censades en funció de certes paraules clau també la fan programes informàtics que estudien distribucions estadístiques de paraules, freqüències d'aparició i enllaços a aquesta pàgina.

O sigui, que tot cercador d'internet ha de tenir tant un bon catàleg de pàgines indexades com un bon índex de paraules clau. 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 aquests, l'algorisme PageRank, que és el que va catapultar Google a l'èxit entre els cercadors d'internet.

El model PageRank. Vectors i valors propis

Tal com hem explicat, per completar el treball del cercador caldria assignar 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*, en què 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 1 és el que a partir d'ara denominarem *el graf d'internet*, segons el dibuix podríem concloure que —per exemple— la primera pàgina és la més popular ja que hi van enllaços des de totes les altres, i és l'única que compleix aquesta propietat.

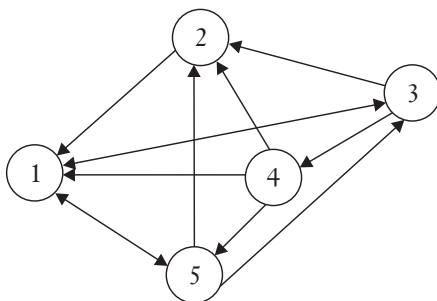


Figura 1. Un graf dirigit

En aquest gran graf dirigit, hem d'assignar una «importància» a cada pàgina. Una manera raonable d'assignar importàncies podria ser: com més enllaços rep una pàgina, més important és. Observem l'analogia que hi ha amb el famós —però no sempre tan apreciat— índex de citacions 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 enllaçades amb aquesta, 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 citacions» per «importància de les citacions». És a dir, no solament donarem importància a la quantitat de citacions que té una pàgina donada, sinó que a més tindrem en compte si l'esmenten pàgines importants. Diguem que —per exemple—, si tinc enllaços des d'Amazon.com o Microsoft.com, la meua importància hauria de ser més gran. En aquest sentit, el graf de les pàgines web seria una mica més semblant al de la figura 2, que mostra una distribució d'importàncies relatives a les importàncies de les pàgines donades. En aquest

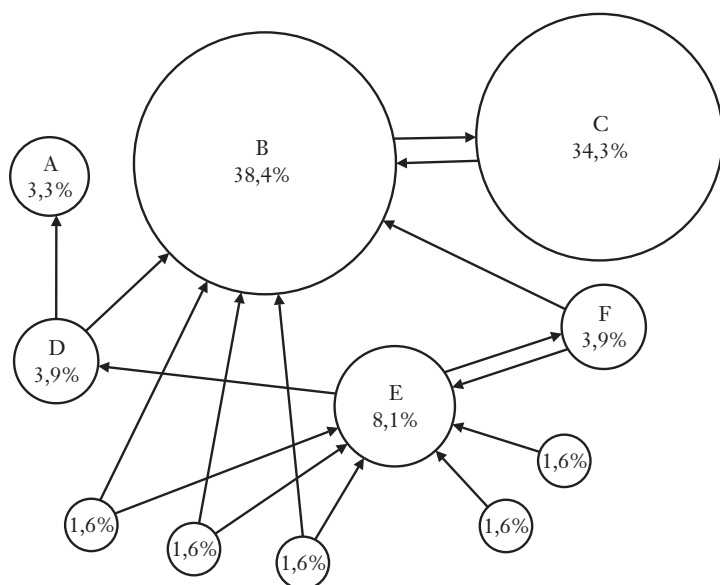


Figura 2. El graf d'«importàncies»

cas s'entén per què la pàgina C és més important que la F , atès que cadascuna rep un enllaç, però la primera està enllaçada 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 hi enllacen.»

Vegem com es tradueix això matemàticament i com, en aquest context i d'una manera 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 i, doncs, no hi guanyarem gaire si intentem dibuixar-lo (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 n'hi ha, hi posem un zero. Per exemple, la matriu d'incidència del graf d'internet de la figura 1 és la següent:

$$M_0 = \begin{pmatrix} 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \end{pmatrix}.$$

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

Teorema 1. Si 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:

$$M_I^t \mathbf{x}^t = \lambda \cdot \mathbf{x}^t,$$

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.