Coeficient d'importància d'una pàgina web

Agafaré totes les pàgines web i els donaré una importància, un coeficient.
Cada una tindrà una etiqueta, els posaré C. C1 és la pàgina web 1. C2 és la pàgina web 2 i així fins a cinc. En farem cinc. Amb les meves d'abans. I farem el següent: com es reparteix la importància que tens?
Agafo això d'aquí i diré: mira, la importància que té en un moment determinat la pàgina 1, anomenada C1, quan l'enllaço és com si part de la meva importància
es repartís entre tots els meus col·legues. I això ho posaré aquí. Fixeu-vos on ho posaré. C1, que és la meva importància, la divideixo entre les pàgines a les que enllaço, C1 partit per 4, a cadascun dels enllaços i li passo... no són diners, una part de la meva importància. Si jo sóc molt important i dividim per quatre, cadascú tindrà una quarta part de la meva importància. I continuo. Aquest és el gran i farem pas a pas pels demés. Perquè ara es complica una mica més la notació,
ja veureu que és un pèl més complicat.

El C2 té importància C2. Com que jo enllaço a 2, es divideix. Ara el P4 tindrà un C2 partit per 2 que es posarà a la butxaca i el P3 també. D'acord? I faré el mateix amb aquest, amb aquest i amb aquest. Ja ho tinc establert. Fixeu-vos, ara estic barrejant coses. Estic barrejant grafs amb una cosa que m'he tret de la màniga.

Una cosa que és molt important a les matemàtiques: si entenc el primer pas, moltes vegades fer el salt al pas general no es molt diferent. Bé doncs, jo. Possessions número 1: tinc les meves pàgines web. Jo reparteixo cromos. Els cromos són les importàncies: C1, C2, C5. La pregunta és, després d'un clic... jo faré un clic de manera aleatòria. Un enllaç, segueixo un enllaç a una pàgina web. On aniré a parar? La importància de la següent pàgina web quina serà? Les C petites són la importància d'una pàgina web amb un pas. Quan he fet un salt, he seguit un clic, quina serà la importància de les pàgines web pel segon pas? Fem-ho. Ja veureu que ara la cosa es barreja una mica perquè estic barrejant moltes coses. El graf és el mateix, és un graf. Fixeu-vos en la primera, mireu la C1. La importància de la pàgina web 1, la importància al cap d'un pas, d'on li ve? Qui li proporciona importància? Doncs tots aquells que li envien una fletxa. C1. Té un zero partit multiplicant, per què té un zero?
Perquè ella mateixa no s'enllaça i per tant, això és un zero. Per aquesta d'aquí també és un zero, la C2. Qui em diria per què? Per què el C2 té un zero multiplicant?
Què penseu? Perquè el C2 no li enllaça. A la P2 no li envia un enllaç el P1. Això d'aquí és la importància de P1. Aquest és aquest. No li enllaça. I només en té un aquí, al C5.

Al P5 sí que li enllaça, però no li passa tota la seva importància. Què li passa? La meitat, sí senyor, li passa només la meitat. Perquè està repartida, el P5 apunta aquí i apunta aquí. Per tant, es posa així. Ara ho complico una mica més, ho poso més matemàtic. Això d'aquí no em va bé per treballar. Ho posaré així, mireu. Això és igual que això per això. Això com es multiplica? Això és el que s'anomena matriu,
i això és un vector que ja el coneixeu, sobre tot els de batxillerat, vosaltres el coneixeu molt. I com s'aplica una matriu a un vector?

Primer per primer, més segon per segon, més tercer per tercer, més quart per quart, més cinquè per cinquè. Justament això. I es posa així. Ho poso de manera esquemàtica. Per què? Perquè això puc fer-ho amb cinc però si ho he de fer amb 2,7 bilions vaig llest.

És impossible. És inviable. Si he de vendre el producte, s'ha de vendre molt bé.
Anem per la segona, això d'aquí, la C2. Penseu tots una mica com és la importància de la pàgina web 2 al cap d'un salt. Qui li envia importància? Mirem tots els que l'enllacen. Qui enllaçarà? A la 2 li enllaça aquesta d'aquí, la P1 i la 4. I no veig cap més. Què li passa P1? Li passa un C1 partit per 4. C1 partit per 4. I després, de la P4, un C4. Li envia sencera, tot. El P4 és molt amic de P1 i de P2. Molt.
Quan té un enllaç, tot va a P2. I té això d'aquí. La importància de la pàgina web 2 ve de la importància de les altres. I ho passem aquí. Què estic construint? Una matriu i una relació però en matrius.

El mateix per la C3, la C4 i per la C5. Es veu ara? Ara és el moment clau. Per què és un moment clau? Torno aquí. Com abans. Fixeu-vos, jo abans deia "matriu de connectivitat". Aquesta no és de connectivitat, és diferent. Una matriu de connectivitat és 0 i 1. Aquesta depèn de la quantitat d'enllaços. Aquesta és importàncies. Abans deia: si mirem per fileres, la P1 on va? Si vull saber la popularitat, quina m'envia una fletxa a mi? Ho miro per columnes. Doncs aquí és el mateix. Per fileres. Aquesta d'aquí ens diu d'on ve la importància de la C1. Si m'ho miro per columnes, aquesta d'aquí és la P1, perquè és la primera columna, i ens diu a qui reparteix la seva importància.
Comprovem-ho. La P1, a ella mateixa no es dóna res. I aleshores, dóna un quart, un quart, un quart i un quart que reparteix entre la P2, la P3, la P4 i la P5.
Si ho miro per columnes em diu una cosa i si ho miro per fileres em diu una altra cosa. Que és justament això d'aquí. Per tant, tot aquest quadrat si jo us dono això,
m'heu de posar això i si us dono això, m'heu de reproduir justament el graf.

Que és el que feia Google, això és el que feien en Sergey Brin i en Lawrence Page.
Continuem. La P2, la P3, la P4 i la P5. I tinc la matriu. I ara ve quan el mata. Això vol dir importància. Primera versió. Ara em poso més cap aquí. Ara em poso a la versió de probabilitat. Com ho puc veure en aquesta matriu d'aquí? Imagineu que comencem justament a l'atzar en una pàgina web P1, P2, P3, P4 i P5. I jo, a l'atzar, vaig a una pàgina web. Quina és la probabilitat que jo vagi a una pàgina determinada?

Doncs això d'aquí, el que em donarà justament seran probabilitats.
Per exemple, aquesta d'aquí. Aquesta d'aquí em donarà la probabilitat que tinc
que la pàgina 5 salti a la P1 o a la P4 en un clic. Si jo sóc a la P5, quina probabilitat tinc de saltar a la P1? Un mig. Quina probabilitat tinc de saltar a la P4? Un mig.
Quina probabilitat tinc de saltar a la P3? Zero perquè no hi ha enllaç, no hi ha manera.
Per tant, si ho mires així, són probabilitats. Aquesta és la matriu bona. Aquesta matriu es diu matriu de transició. Tot això d'aquí és el que es va inventar
per fer l'algoritme del rànquing de Google. Això s'anomena PageRank, té un nom i té un copyright. És així, tal qual. Només és que és molt més gran, i vindran problemes, no serà fàcil, però van començar així. De fet podeu llegir a Internet l'article que van publicar, s'hi pot accedir.

Aquesta matriu és una matriu, sobre tot pels que esteu a batxillerat,
és una matriu que està molt bé, que és molt *xula* i té una característica molt important.
Si sumeu els elements de d'aquesta columna, què us dóna? 1. I els d'aquesta? 1. I els d'aquesta? 1. Molt bé, les matrius que quan sumes per columnes donen 1 s'anomenen matrius de Markov i són matrius que apareixen contínuament en molts processos i en moltes disciplines. Apareix aquí, però també a biologia.

A biologia apareixen contínuament. Per què? Perquè serveix per saber com es reparteix una propietat entre els seus enllaços. I la pregunta és: com es calcula?
Doncs ho podem fer amb un llapis i un paper, proveu-ho.

Es poden fer tres iteracions, però quan en porteu 4 us podeu morir. I si la matriu és cinc per cinc encara, però si la matriu és de 2,7 bilions per 2,7 bilions... i en canvi, això dóna de menjar a molta gent. Com ho fan? Com es fa? Ens posem la capa de matemàtics abstractes. Fixeu-vos aquest d'aquí, l'anomeno xn+1, permeteu-me que utilitzi una notació una mica rara. Què vol dir? Al cap d'ena més un pas, com està la foto?

Faig una foto de les 5 pàgines web. Aquesta d'aquí és la foto al cap d'ena, és a dir l'anterior. Quina importància té cadascuna d'elles? Perquè estic mesurant importàncies. Aleshores jo tinc aquesta relació d'aquí. Aquesta és una relació que s'ha de verificar i jo sé que les importàncies, el coeficient d'importància, el pas n+1 és l'anterior multiplicat per la matriu. Això em diu sempre la probabilitat que tinc de saltar a una altre.

Bé doncs fem-ho pas a pas. Primer pas. Començo amb... totes les pàgines web tenen la mateixa probabilitat: 0,2. Que vol dir un 20%. Si jo faig un pas,
quina probabilitat tinc d'estar a cada pàgina web? Doncs tal com hem dit, al cap d'un salt, d'un clic, jo estic aquí, a "A" aplicat a l'anterior, la matriu A que jo he calculat és sempre la mateixa, només la calculo una vegada i la tinc sempre. Aquesta matriu aplicada a aquest vector. I si multiplico em surt això d'aquí. Com es tradueix això?
Tinc un 10% de probabilitats d'estar a la pàgina 1. Tinc un 25% d'estar a la pàgina 2.
Un 15% d'estar a la 3. Un 45% d'estar a la 4, i un 5% d'estar a la 5. De moment qui guanya? De moment guanya la 4. Però això és un salt. A mi què m'interessa?

Una cosa a llarg termini. Realment m'interessa saber què passarà quan hagi passat molt temps. Doncs si tinc paciència i tinc un ordinador, calculo el següent...
Fixeu-vos en una cosa, això d'aquí, si ho sumeu, dóna 1. Per què? Perquè és la probabilitat d'estar en una de les pàgines web, per tant la suma ha de ser igual a 1.
O sóc aquí, o sóc aquí, o sóc aquí, o sóc aquí, o sóc aquí. Això d'on surt? Ve de la matriu, perquè la matriu que tinc és una matriu de Markov, aquesta d'aquí. Com anava dient, la probabilitat d'estar a la P1. Quin és el segon pas? Següent iteració i tinc això d'aquí: 0,25. Tot es pot llegir amb llenguatge de probabilitat. Ànim! En faig moltes, agafeu l'ordinador i ho calculeu. Jo en tinc unes quantes calculades: 2, 3...
després de 10 iteracions, després de 50... 10^{-24} vol dir zero, 23 zeros i un 1; per tant això és molt petit. I després de mil iteracions surt això d'aquí. És a dir, si jo em poso així, a l'atzar, a saltar d'una pàgina web a una altra, quina probabilitat tinc d'aturar-me en cada una de les pàgines web?

Doncs, per la P1 ho tinc negre. Si sou P1 ho teniu malament, ningú us visitarà al cap de molt temps, veneu la pàgina web. La P2, un 0,4 que és un 40%. Això és sempre sobre 1, per tant un 40%. La 3, un 20%, la 4 un 40% i la 5 també, oblideu-vos.
Aleshores si jo aquí hagués de repartir importància, quina és la més important? O quines són les més importants? La 2 i la 4. Fixeu-vos que en la versió de pes, només sortia la 4, i en canvi la 2 és tan important com la 4, igual d'important.

Per tant, si això d'aquí fossin pàgines web relatives a una pregunta que he fet,
quina sortiria primer? Sortirien primer la P4 i la P2, després la P3 i les últimes, pobretes, serien la P1 i la P5. Per tant em dóna un rànquing. Aquest és el rànquing que us surt a Google, però no són 5, són milions i milions i milions de pàgines web. I així és com funciona. Problema: ara ve la part de computació. És a dir, com es calcula de veritat. Això podeu fer-ho i ho podeu simular a l'institut si voleu. Porteu matrius, agafeu un ordinador i es calcula però clar, funciona amb 5, amb 10, però si és una cosa molt gran com s'ha de fer? Perquè això és un monstre. Això és un monstre.
Mireu això que m'ha sortit aquí. Aquest número que m'ha sortit aquí...

per què és important aquest vector? Això és un vector. Per què és important aquesta distribució? Què verifica? Us poseu dins de la matriu i... verifica això. Jo agafo aquesta matriu, la multiplico per aquest vector i em torna a sortir el mateix. Em torna a sortir el mateix. Quan tinc un vector que verifica que agafo aquesta matriu i la multiplico pel vector i em torna a sortir el mateix vector diem que és un vector xulo, un vector propi.
I el número que sortiria aquí, aquí surt un 1 però podria sortir un 3, un 5, un 25, un -10... s'anomena valor propi. Què estic buscant jo? Si buscava una ordenació, estic buscant un vector propi d'aquesta matriu. Tot el problema de Google és: tinc una matriu molt gran i he de buscar un vector propi de valor propi igual a 1.
Problema? Que la matriu és molt gran.

Bé doncs, tenim això aquí. Si parlem d'importància, què sortiria a Google? Sortiria això. Aquests són els més importants, P3 ,P1 i P5. Aquesta seria l'ordenació de Google. La manera de calcular-ho és, amb l'ordinador, calculeu: X1 què és? "A" aplicat a X0. Què és X2? "A^2", A per A, aplicat a X0. I així successivament. Agafeu la matriu A i feu potencies de l'A. Això d'aquí és simplement el mètode de la potència.
És un mètode numèric, ja entrem en el càlcul numèric. Fixeu-vos que per fer això d'aquí hem passat per la teoria de grafs, per la teoria de la probabilitat, per àlgebra lineal, que són matrius, i això d'aquí és càlcul numèric. Una cosa és què vull calcular i una altra és com ho calculo.

A matemàtiques, 1+1 no sempre és igual a 2. S'ha de vigilar, depèn de com calcules les coses. Part final de l'algoritme. Tinc un petit gran problema, que és amb el que es trobaven en Sergey Brin i en Lawrence Page. Penseu una cosa, tota la teoria en la que es basa Google per cercar és del 1907. No és moderna, fa 100 anys que tenim aquesta teoria, el que passa és que ara s'aplica. Aquesta d'aquí... bé què és el que jo vull? Que aquesta ordenació, que aquest vector propi, sigui únic. Per quin motiu? Perquè imagineu que jo us dic: ordeneu les 10 pel·lícules que hi ha en cartell que més us agraden. I comenceu a xerrar, xerrar, xerrar i em feu 30 ordenacions diferents.
No guanyo res perquè jo vull que em doneu una única ordenació que sigui objectiva.
Clar, el problema és que jo vull que aquest vector, que aquesta ordenació final sigui única i això en general no passa.

Hi ha un teorema molt important, el teorema de Perron, de 1907 que diu: "si la matriu A té tots els números positius, segur que és únic". Bé, i aleshores per què falla? Falla perquè hi ha zeros. I els zeros ho espatllen. Aleshores aquest teorema que ens salvaria la vida no funciona. Aquest va ser el primer problema amb el que es van trobar en Sergey Brin i en Lawrence Page. Uns quants anys més tard, 5 anys més tard, va sortir un altre teorema, que ells coneixien, el teorema de Frobenius, una extensió del teorema de Perron. Aquest teorema diu "tranquils! Si tots els números de la matriu són més grans o iguals a zero, cosa que m'agrada molt, només cal que la matriu sigui irreduïble".

Ara la pregunta és: què vol dir irreduïble? Us ho explico amb versió graf, amb fletxes, perquè veieu realment què pot passar. Irreduïble vol dir que hi ha dues coses que no poden passar: no pot haver pous. Què vol dir pous? Doncs que per exemple...
Tu com et dius? Vladimir.En Vladimir té una pàgina web que tothom l'enllaça perquè és molt popular però ell passa dels altres i no enllaça a ningú.
En Vladimir és un pou. Des del punt de vista de les pàgines web. Què vol dir? Que s'ho menja tot i no treu res. En Vladimir, amb tots els respectes, és el P3.
Si jo tinc una pàgina web que no és social, mal rotllo. Per què? Perquè mireu què passa, tinc una columna que és tot zeros. Ja no és de Markov i això s'espatlla.
Per tant això està prohibit, no pot haver-hi pous. Pous, no Vladimirs. Vladimirs sí, pous no.

Segon problema: els teus companys Vladimir. Anem a veure per aquí. Tu com et dius? Sergi i...Germán. Doncs ara imagineu que tothom esta connectat, tothom apunta cap a ells, però ells dos es fan molt amics i només s'apunten entre ells. Què passa? Que de tota la xarxa tinc elements desconnectats. Això d'aquí. Aquests podrien estar aquí, però el que importa és que aquests d'aquí només s'apunten entre ells. Això d'aquí no val. Si això d'aquí passa, la matriu no és irreduïble per tant això d'aquí no pot passar.

Problema: Això d'aquí no és veritat? És a dir, no coneixeu ningú que tingui una pàgina web que no apunta a ningú? Segur que sí. Això a la vida real no passa. Cagada pastoret. Què vol dir? Que tot el que us he explicat no funciona a la vida real perquè hi ha pàgines web que s'apunten només a elles, o pàgines web entre cosinets, que només s'apunten entre ells.

Aleshores, on està la gràcia de l'algoritme? Què es van inventar perquè això d'aquí pugui passar? Es van treure de la màniga una matriu, aquesta d'aquí.
Aquesta A és la meva. Quina és aquesta B? Aquesta B és una matriu que... la "n" vol dir que té "n" elements. Si jo tinc 5 pàgines web tinc una matriu de 5x5 tota d'uns.
I això que seria? 1 partit per 5. Invento aquesta matriu d'aquí. Dono un valor a aquesta P. A Google, el valor original és 0,15. Per tant això d'aquí és 0,85, 0,15, A i B. I aquesta és la matriu M. Aquesta és la matriu de Google. La famosa matriu de Google és aquesta d'aquí. Què li passa a aquesta matriu?

Primer us explico per què va bé i després explico què vol dir. Aquesta matriu ja no pot tenir elements zero perquè com que sempre estic sumant aquesta d'aquí i aquesta d'aquí té uns, els elements són sempre positius. Per tant puc aplicar el teorema de Perron, que em diu que el que jo trobi tindrà una distribució única. Fantàstic. Què vol dir això? Té sentit matemàtic? Sí. Vol dir el següent: de tant en tant, i us passa a vosaltres i em passa a mi, i ens passa a tots i és que se'ns creua un cable i si estic buscant sabates dic: ostres! El viatge de no sé què. I salta una pàgina web completament diferent. S'estima que aproximadament hi ha un 15% de salts completament a l'atzar a una pàgina web que no té res a veure amb el que estem fent.
I això d'aquí és el que mesura aquesta P per P. Jo faig un salt a una cosa que no té res a veure amb el que estic buscant. No segueixo un enllaç d'aquesta pàgina sinó que me'n vaig a una altra directament. No us ha passat mai? Que esteu buscant una cosa i dieu "ostres, les sabates!" i canvieu.

Això vol dir això. Molt bé doncs aquesta és la matriu amb la qual es treballa.
Aquesta matriu ho unifica tot, ho calcula tot i és la que fa servir Google. Ja per acabar, això és el Google clàssic. Però el Google clàssic és com la Coca-Cola.
Això d'aquí és la teoria matemàtica. De veritat funciona així? Sí i no. Això és la versió pública. Què és la versió no pública? Doncs que sempre hi ha coses que afavoreixen una línia determinada. No us ha passat mai que entreu a una pàgina web, per exemple a buscar un vol a easyJet, i curiosament, esteu una setmana rebent correus d'easyJet amb ofertes d'última hora? No és molt curiós? Què passa? Que el rànquing de Google es personalitza. És a dir, la part matemàtica és aquesta d'aquí, aquest és el Google clàssic, del 98. Però està *tunejat*. Com? Per adaptar-se a vosaltres. Si Google detecta, te informació, que vosaltres sou molt amants de l'automobilisme, segurament bona part de la informació dels *banners*, dels enllaços que us arriben quan feu una cerca a Google tenen molt a veure amb allò que voleu.

Per tant una cosa és la versió matemàtica i una altra cosa és com es fa servir de veritat.

De fet Google té una lluita continua perquè la idea de Google és buscar una cosa que sigui bastant correcta, quan detecten que hi ha un servidor que intenta modificar el rànquing de Google el penalitzen. Segur que si busqueu a Internet, trobareu empreses que t'asseguren que si els encarregues la teva pàgina web ells et garanteixen un bon rànquing. I el modifiquen, o ho intenten, és una lluita continua.

Per què? Perquè és molt poder. Diem "som independents, fem allò que volem"... relativament, quan busquem a Google, en realitat estem seguint les línies que Google ens marca. Per tant, vigileu, sapigueu que tots estem manipulats, sigueu conscients d'això. Això d'aquí és un bon model per veure com s'apliquen les matemàtiques de veritat al món real.