Laboratori de CAIM

Pràctica 4

Walter J.Troiani

Prof: Ignasi Gómez Sebastià

12/11/2023 - 2023/24 Q1

1 PageRank. Detalls d'implementació

Aquest escrit està orientat a donar context al lector de l'algorisme de PageRank emprant la matriu de Google, el qual és l'algorisme implementat en el codi. En aquest cas, però, l'algorisme estarà orientat a aeroports (Nodes) i les seves rutes (Arestes) que formen un graf dirigit, els quals han estat extrets gràcies a la base de dades d'OpenFlights, i no pas en pàgines web. No entrarem en molt més detall del modelat e implementació (Consultar codi font altrament).

L'algorisme que implementarem constarà de repetir una simple formula $\vec{p}(t+1) \approx G * \vec{p}(t)$ on G és la matriu de Google $G = \lambda * M + \frac{1-\lambda}{n} * J$ on J és la matriu de tot uns , M és la matriu d'adjacència del graf dirigit original normalitzat per files del sistema d'avions i λ és el factor d'amortiment. Aquesta fórmula serà aplicada fins a arribar al punt de convergència, que s'arribarà quan la variació de pes entre iteracions sigui menor al factor de convergència ϵ, el qual considerem que una aproximació més fidel a l'algorisme que no pas un nombre fixat d'iteracions. És essencial pel correcte funcionament i convergència de l'algorisme que el graf sigui fortament connex i aperiòdic. La correctesa i eficiència de l'algorisme pot ser consultada a Wikipedia o altres fonts fiables on ho explicaran millor del que podria jo en un paràgraf.

Un problema al qual ens enfrontàrem seran els aeroports "especials", que són aquells que o bé no tenen rutes sortints o bé no tenen rutes entrants o cap de les dues. Com a paràmetres del programa tindrem 3 booleans que permetran l'eliminació de qualsevol dels 3 tipus de nodes, per a poder experimentar amb els diversos resultats.

- Aeroports Aïllats (Isolated): Són aquells que tenen rutes de sortida però no pas d'entrada. No són problemàtics a menys que $\lambda = 1$, aleshores s'impedeix una de les 2 condicions essencials de l'algorisme
- Aeroports Pou (Terminal): Aeroports que tenen rutes d'entrada, però dels quals no es pot sortir. Són problemàtics ja que el pes acaba en aquests nodes (fan de pou).
- Aeroports Inaccessibles (Unreachable): Aeroports els quals no tenen cap ruta d'entrada ni sortida.

Un problema típic del PageRank són els autobucles, però com trivialment cap avió vola al mateix aeroport podem oblidar-nos d'aquest problema. Per solucionar els problemes anteriorment esmentats, farem un petit retoc al codi: Si el node es tracta d'un aïllat o bé un terminal (o ambdós), el sumatori de pesos (de la fórmula del PageRank per un pes p_i) serà $\frac{1}{n}$, així el seu pes no canviaria de l'inicial i garantiria que no desaparegui el pes d'aquest tipus de nodes, aconseguint que tot plegat sumi 1.

2 Experimentació

En aquest algorisme entren en joc un nombre divers de paràmetres i s'han de fer certes decisions (algunes arbitràries i altres obligatòries per garantir l'acabament de l'algorisme). Caldrà revisar la casuística dels nodes inaccessibles, pous o bé ambdós.

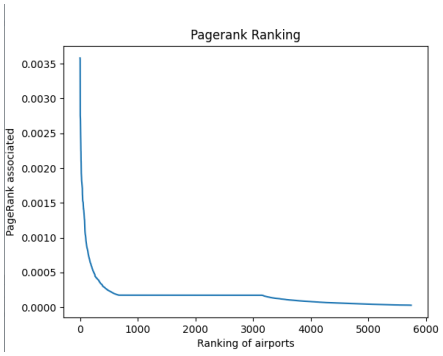
- λ: Factor d'amortiment de la matriu G de Google en el rang $[0, 1]$, que regula la importància del graf original en comparació a un graf complet (i fortament connex per conseqüència)
- ϵ: Constant de la condició d'aturada major a 0. El bucle acaba quan el pes de tots els nodes tenen una variació entre iteracions inferior a aquest factor, és a dir: $\forall i \in \{1, 2, .., n\} : \|p_i(t+1) - p_i(t)\| < \epsilon$ atura el bucle en l'instant t+1.

Al tractar-se d'un nombre tan reduït de variables, és raonable plantejar un estudi multivariable tenint els 2 paràmetres en compte i les dues possibles variables objectiu com poden ser el nombre d'iteracions/temps d'execució o bé el PageRank (mitjana, diferencia entre el màxim o mínim ...), per fer un estudi del temps de convergència entre altres.

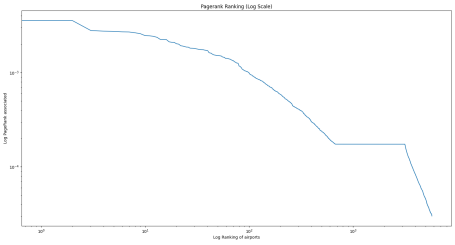
També serà rellevant estudiar els rànquings obtinguts de pes PageRank a cada node i avaluar si tenen sentit amb les dades del món real. Això podem aproximar-ho qualitativament comparant amb estadístiques de com de concorregut es l'aeroport. Finalment, la distribució inicial dels pesos de PageRank també podria ser investigada per tal d'establir alguna conclusió sobre l'efecte que té al resultat dels pesos.

2.1 Estudi de λ i ϵ en el PageRank

Primerament. seria interesant observar quina es la relació entre el rànquing d'un aeroport i el seu pes de PageRank, fixant paràmetres $\lambda = 0.85$ i $\epsilon = 1 * 10^{-12}$ tenim:

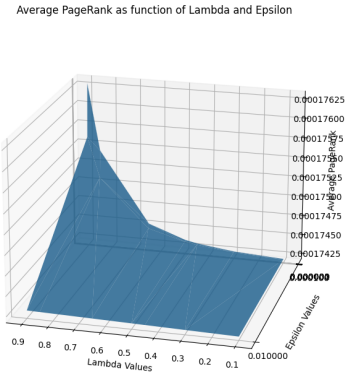


(a) PageRank en funció del Rànquing

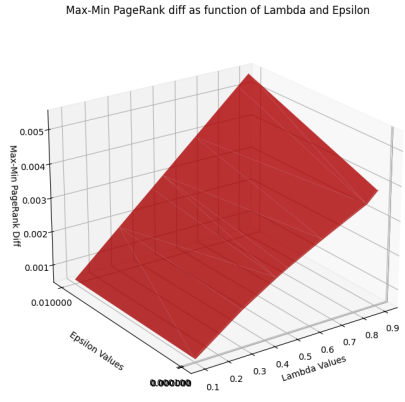


(b) Idem, escala log-log

Es trivial veure que això segueix una llei potencial molt semblant a la llei de Zipf, això és perquè els aeroports més importants són ordre de magnitud més important que els que menys, i així de manera successiva com passa amb les paraules a la llei de Zipf. També seria interessant veure com fluctuen els valors promig i la diferència max-min dels pesos del pagerank p_i:



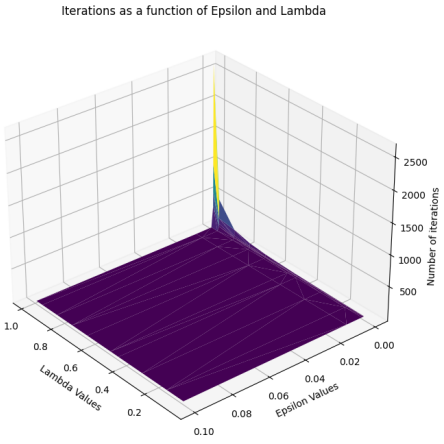
(a) PageRank promig en funció d'ϵ i λ



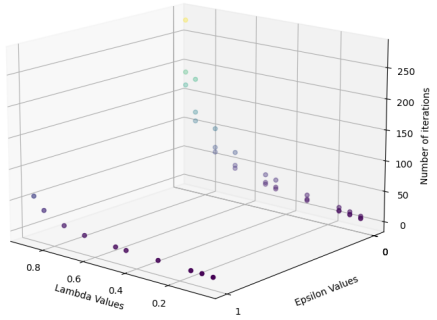
(b) Diferència entre màxim i mínim en funció d'ϵ i λ

2.2 Estudi de λ i ϵ en el nombre d'iteracions

Evidentment, era necessari una anàlisi del nombre d'iteracions, fixant els paràmetres $\lambda = 0.85$ i $\epsilon = 1 * 10^{-12}$, obtenim els resultats següents:

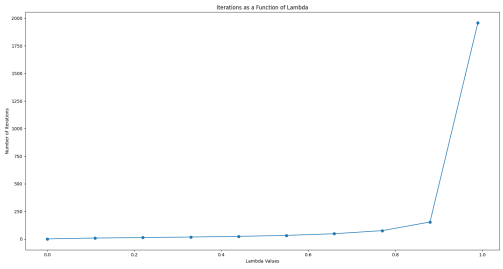


(a) gràfic emprant $\lambda = 0.999$ (outlier)

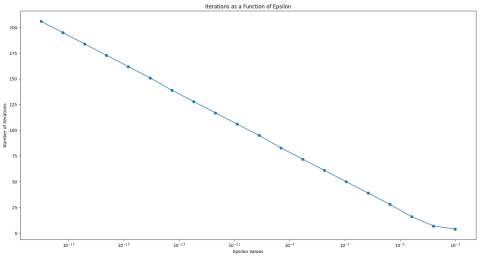


(b) gràfic normalitzat

Es evident que a mesura que la λ s'aproxima a 1 cada cop el nombre d'iteracions és major (Ja que el graf original no complia les condicions necessàries per la convergència de l'algorisme, si ens apropem a 1 res ens garanteix la convergència d'aquest), i també que amb valors més petits d'ϵ creix una mica el nombre d'iteracions necessàries. Llavors hi ha una relació directa entre λ, ϵ i el nombre d'iteracions. Això es pot observar de manera més clara en el gràfic següent on hem estudiat les dues variables per separat:



(a) El nombre d'iteracions creix linealment amb λ i al final asimptòticament



(b) Amb ϵ veiem una correlació indirecta perfecta (escala log)

És interessant veure l'efecte que valors grans de λ te en la mitjana i la diferència entre màxim-mínim, però això té sentit tenint en compte el paper normalitzador en igualador que té el factor d'esmortiment (Contra menys factor d'esmortiment més exagerades seran les diferències de rellevància dels diversos aeroports.

2.3 Estudi d'Aeroports Especials

En aquest apartat farem un petit incís per comentar l'efecte dels aeroports especials, diverses estadístiques, dades rellevants i conclusions, $\lambda = 0.85$ i $\epsilon = 1 * 10^{-12}$.

Com bé es pot apreciar, per a varies execucions, ens adonem que efectivament tots els pesos sumen 1.000007, el qual és aproximadament 1 (tenint en compte errors de precisió de coma flotant).

$$\sum_i P_i = 1$$

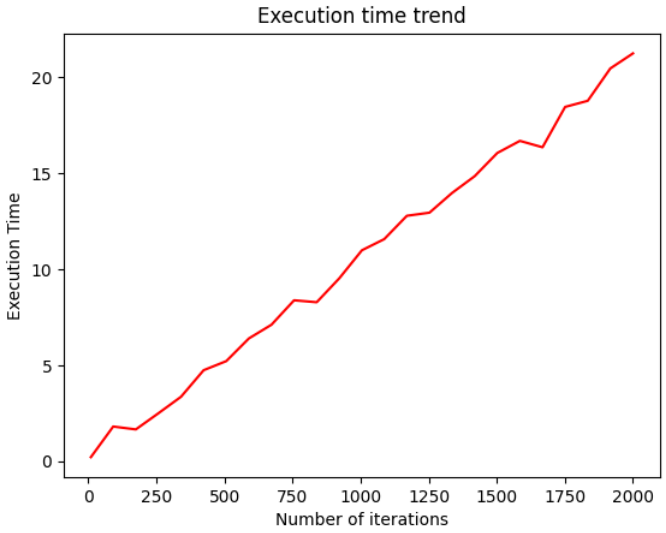
La quantitat d'aquests aeroports no és pas menyspreable (Si no els tenim en compte, el sumatori de pesos dona valors de l'estil 0.633) i hem fet bé de no ignorar-lo. Si computem quants aeroports hi ha de cada tipus la gran majoria són inaccessibles (lo qual implica que siguin aïllats i terminals) però purament terminals o aïllats són ben pocs (19 i 10 respectivament).

Aeroports	Aeroports Aïllats	Aeroports Terminals	Aeroports Inaccessibles
Dades	2436 + 10	2436 + 19	2436

Ens hem adonat que seguint la implementació proposada, a pesar que funciona, li dona un avantatge als aeroports especials exagerat, ja que el seu pes sempre es manté $\frac{1}{n}$ (Això sí, donant-li una avantatja injusta als altres nodes a pesar de la seva ridícula importància). Concloem que a pesar que és una solució la proposada, podria millorar-se, potser, repartint equitativament aquest pes $\frac{1}{n}$ entre tots els altres aeroports normals seria una millor opció.

2.4 Estudi del temps d'execució

En aquest apartat, breument comentarem com evoluciona el temps d'execució $\lambda = 0.85$ i $\epsilon = 1 * 10^{-12}$ d'entrada i provant diversos valors d'iteracions o bé nombre d'aeroports (fent mostres de diverses mides).



Efectivament, veiem que hi ha una gran correlació directa entre el nombre d'iteracions i el temps d'execució com era de suposar. Això per transitivitat ens permet fer suposicions de l'estil, a major lambda major temps d'execució amb tota classe de garantia.

2.5 Estudi qualitatiu

Finalment, es farà un estudi comparatiu, tenint en compte les dades d'aeroports proporcionades per la mateixa Wikipedia (Ens fiarem que cap xinès ni yankee ha inflat les dades), podem estimar si les conclusions que extrèiem dels aeroports que tenen millor rànquing són mínimament fidels a la realitat. El rànquing obtingut pel nostre script és:

Si ens fixem en el top 10 de Wikipedia per nombre de passatgers: Pequín, Londres, Atlanta, Chicago, Tòquio, Los Angeles, París, Dallas, Frankfurt, Hong Kong. Ens fixem que la major part del nostre top també són a aquest top (excepte Hong Kong i Tòquio). Ara bé, si mirem el rànquing segons el nombre de vols: Atlanta, Chicago, Dallas, Denver, Los Angeles, Charlotte-

IATA Code	City	Country	Score
LAX	Los Angeles	USA	0.00358
ORD	Chicago	USA	0.00358
DEN	Denver	USA	0.00356
LHR	London	UK	0.00280
SIN	Singapore	Singapore	0.00276
ATL	Atlanta	USA	0.00274
CDG	Paris	France	0.00271
PEK	Beijing	China	0.00270
FRA	Frankfurt	Germany	0.00265
SYD	Sydney	Australia	0.00259
DFW	Dallas/Fort Worth	USA	0.00247
JFK	New York City	USA	0.00246
AMS	Amsterdam	Netherlands	0.00243
DME	Moscow	Russia	0.00237
ICN	Seoul	South Korea	0.00225
YYZ	Toronto	Canadà	0.00225
DXB	Dubai	UAE	0.00224
PVG	Shanghai	China	0.00212
IST	Istanbul	Turkey	0.00210
BCN	Barcelona	Spain	0.00210

Douglas, Pequín, Las Vegas, Houston, París, Londres ... Observem que la majoria d'aeroports del nostre rànquing top 10, es troben en alguna d'aquestes dues llistes. Totes aquestes dades són antigues de 2014, potser OpenFlights fa servir dades més actuals i més antigues, d'aquí la diferència que es pot apreciar. Aleshores podem estar contents que els PageRanks obtinguts tenen força semblança a la realitat. Caldria fer un estudi més extensiu, però tenint en compte que hem basat el PageRank en el nombre de rutes i no pas en el nombre de vols o passatgers, és molt bo obtenir resultats tan similars (Ja que són conceptes relacionats).

3 Conclusions i Reptes

Els problemes més comuns van ser errors en la implementació de les classes, o bé lectures dels CSV (al ser manual) incorrectes. Evidentment com sem-

pre, he fet la pràctica completament sol i això implica una menor quantitat de recursos humans, que han sigut compensats amb temps i esforç. També l'experimentació va ser un treball llarg d'implementar i verificar, però els resultats han valgut la pena sobretot tenint en compte que no sabia com fer gràfics 3D.

Per manca de temps (He estat a una hackató a Finlàndia i he fet això a l'avió i creuer, essent honest) no vaig poder implementar l'experimentació de distribucions inicials diferents, és a dir, tots els experiments només han sigut amb la distribució uniforme ($\forall airport_i, p_i = \frac{1}{n}$). També per manca de temps no he pogut provar més modes de redistribuir el PageRank perdut dels aeroports especials (Probablement no calia tenir en compte els aeroports aïllats per aquest experiment).

Les conclusions amb més detall han estat documentades a l'apartat 2. Però de manera resumida tenim una correlació positiva entre els aeroports que tenen més tràfic al món real i els aeroports amb major PageRank al nostre script. A més sabem que el PageRank d'un aeroport en concret és inversament proporcional (Seguint una llei potencial semblant a Zipf) al seu rànquing, lo qual és curiós si menys no, però té una estreta relació, i és que aquest algorisme classifica aeroports de manera que els més importants, si apunten a altres importants, elevaran molt la seva rellevància, de manera exponencial. Com passa amb les primeres paraules, que cada cop comencem a decaure més i més, igual que el PageRank amb els aeroports. Un cop més la llei de Zipf s'ha complert de manera natural. Altres conclusions rellevants són que contra major és el valors de λ més exponencialment gran és el nombre d'iteracions (Valors molt petits d'ϵ i valors propers a 1 de λ) i al revés amb ϵ. També podem observar que el valor promig de p_i... i la diferència mínim vs màxim s'accentua a mesura que λ creix o ϵ disminueix

Concloem que el factor d'esmortiment té una correlació directa amb el temps d'execució i una distribució més irregular dels pesos (Com a mínim en aquest graf en concret si). Cal un valor gran per no distorsionar molt la rellevància original del graf , ja que $\lambda = 0$ implica que tots són iguals i $\lambda < 1$, propers asseguren les difèrencies originals, però a un cost temporal elevat. Si aquest es un problema, amb valors $\lambda <= 0.95$ en serà prou. Hi ha un clar tradeoff entre qualitat i temps. També arribem a la satisfactòria conclusió que el modelatge és prou fidel a la realitat.