Recuperació de la Informació (REIN)

Grau en Enginyeria Informàtica

Departament de Ciències de la Computació (CS)

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Politècnica Superior d'Enginyeria
de Vilanova i la Geltrú

Índex

Cerca a internet, I

Quan els documents estan enllaçats

Internet és enorme

- 100000 pàgines indexades el 1994
- $\approx 10^{10}$ pàgines indexades al 2013

Cerca a internet, I

Quan els documents estan enllaçats

Internet és enorme

- 100000 pàgines indexades el 1994
- $\approx 10^{10}$ pàgines indexades al 2013
- comproveu-ho ara mateix
  `http://www.worldwidewebsize.com/`

Cerca a internet, I

Quan els documents estan enllaçats

Internet és enorme

- 100000 pàgines indexades el 1994
- $\approx 10^{10}$ pàgines indexades al 2013
- comproveu-ho ara mateix
  `http://www.worldwidewebsize.com/`
- La majoria de consultes retornaran *milions* de pàgines amb una gran similitud.
- El contingut (text) sol, no és suficient per discriminar.

Cerca a internet, I
Quan els documents estan enllaçats

Internet és enorme

- 100000 pàgines indexades el 1994
- $\approx 10^{10}$ pàgines indexades al 2013
- comproveu-ho ara mateix
  `http://www.worldwidewebsize.com/`
- La majoria de consultes retornaran *milions* de pàgines amb una gran similitud.
- El contingut (text) sol, no és suficient per discriminar.
- Aprofitar l'estructura de la xarxa - un graf dirigit.
- Dóna indicacions de la popularitat - utilitat de cada pàgina.

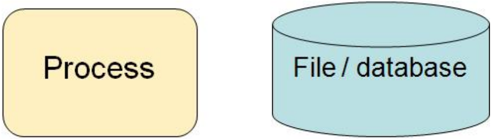
Índex

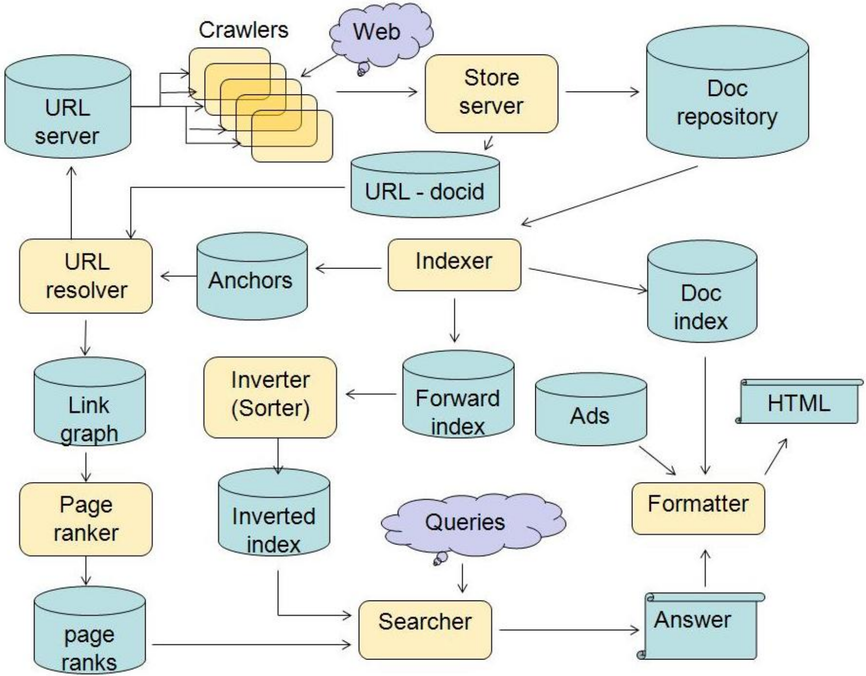
1. 5. Cerca a internet. Arquitectura de sistemes de RI simples
   - Arquitectura d'un sistema de cerca a internet
   - Cerca a internet
   - Algorisme de Pagerank
   - Algorisme HITS

Com funcionava Google al 1998

S. Brin, L. Page: “The Anatomy of a Large-Scale Hypertextual Web Search Engine”, 1998

Notació:




Crawlers
Web
URL server
Store server
Doc repository
URL - docid
URL resolver
Anchors
Indexer
Doc index
Link graph
Inverter (Sorter)
Forward index
Ads
HTML
Page ranker
Inverted index
Queries
Formatter
page ranks
Searcher
Answer

Alguns components

- URL store: URL en espera de ser explorades.
- Doc repository: documents complets, comprimits (`zip`).
- Indexer: Analitza pàgines, separa text (cap al *Forward Index*), enllaços (cap a l'*Anchors*) i info essencial del text (cap al *Doc Index*).
  - El text de l'enllaç és molt rellevant per la pàgina *destinació*

    `<a href="http://page">anchor</a>`
  - El tipus de lletra, la posició en la pàgina, donen rellevància extra a alguns termes.
- Forward index: `docid` $\rightarrow$ llista de termes que apareixen a `docid`.
- Inverted index: terme $\rightarrow$ llista de `docid` que contenen el terme.

L'inversor (classificador), I

Transforma el *forward index* en un índex invertit.

Idea inicial:

```
for cada document d
  for cada terme t in d
    afegir docid(d) al final de la llista de t;
```

Es perd la localitat, moltes cerques a disc, massa lent.

L’inversor (classificador), II

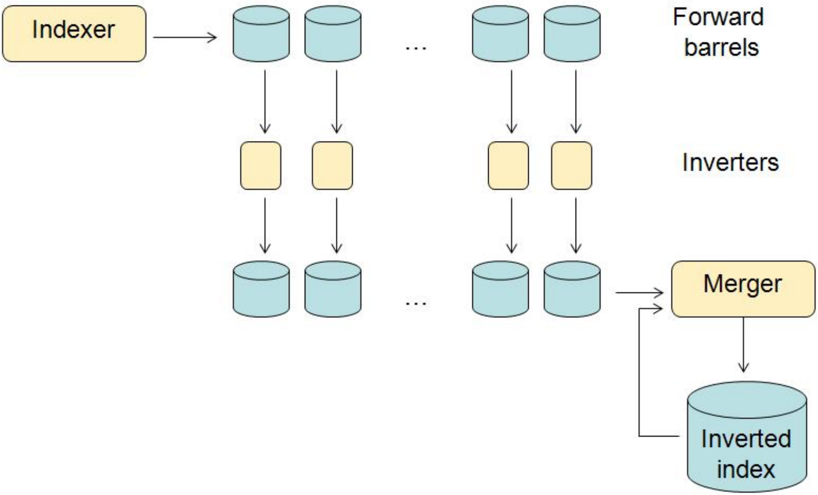
Millor idea per la indexació:

```
crear en el disc un índex invertit buit, ID;
crear en la RAM un índex buit IR;
for cada document d
  for cada t in d
    afegir docid(d) al final de la llista de t a IR;
    if RAM plena
      for cada t, fusiona la llista de t a IR
          amb la llista de t a ID;
```

La fusió de llistes ordenades es fa amb accés seqüencial. Millor localitat. Molts menys accessos a disc.

L'inversor (classificador), III

L'algorisme anterior es pot fer de forma concurrent en diferents conjunts de documents:



L'inversor (classificador), IV

- En els *forward barrels* hi ha fragments del *forward index*.
- Mida d'un *barrel* = el que cap a memòria.
- De forma independent, són invertits a memòria concurrentment.
- Els *inverted barrels* es fusionen en l'índex invertit.
- 1 dia en lloc d'una estimació de mesos.

Índex

Cerca a internet, II

Quan els documents estan enllaçats

Internet és enorme

- 100000 pàgines indexades el 1994
- $\approx 10^{10}$ pàgines indexades al 2013

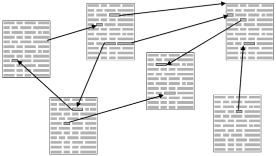
Cerca a internet, II
Quan els documents estan enllaçats

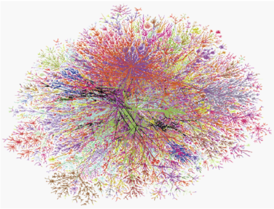
Internet és enorme

- 100000 pàgines indexades el 1994
- $\approx 10^{10}$ pàgines indexades al 2013

Per trobar un contingut, cal buscar-lo.

- Sabem com manegar el contingut de les pàgines web.
- Però. . . què podem treure de l'estructura d'internet?





Cerca a internet, III

Significat d'un hiperenllaç

Quan la pàgina A enllaça cap a la pàgina B, significa

- L'autor d'A pensa que el contingut de B és interessant o important.
- Llavors, un enllaç de A a B, incrementa la reputació de B.

Però no tots els enllaços són iguals. . .

- Si A és molt important, llavors $A \to B$ "compta més".
- Si A no és important, llavors $A \to B$ "compta menys".

Avui veurem dos algorismes basats en aquesta idea:

- *Pagerank* (Brin and Page, oct. 98)
- *HITS* (Kleinberg, apr. 98)

Índex

Pagerank, I

La idea que va consolidar Google

Intuïció:

Una pàgina és important si és apuntada per altres pàgines importants.

- Definició circular...

Pagerank, I

La idea que va consolidar Google

Intuïció:

Una pàgina és important si és apuntada per altres pàgines importants.

- Definició circular... cap problema, matemàticament, té solució!

Pagerank, és l'algorisme que utilitza Google per determinar la posició d'una pàgina web. Aquest algorisme:

1. Mesura el grau d'importància (de forma numèrica) de les pàgines per situar els resultats més fiables en primer lloc.
2. Reflecteix la probabilitat de que un usuari, navegant de forma aleatòria, arribi a una pàgina web concreta.

Pagerank, II

Definicions

Internet és un graf dirigit $G = (V, E)$

- $V = \{1, .., n\}$ són els nodes (és a dir, les pàgines).
- $(i, j) \in E$ si la pàgina i apunta la pàgina j.
- Associem a cada pàgina i un valor real p_i (el *pagerank* de i).
- Imposem que $\sum_{i=1}^{n} p_i = 1$

Com estan relacionades les p_i

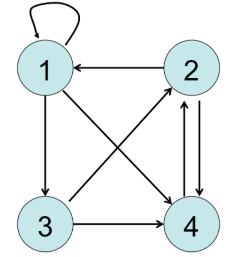
- p_i depèn dels valors p_j de les pàgines j que apunten a i

$$\boxed{p_i = \sum_{j \to i} \frac{p_j}{out(j)}}$$

- on $out(j)$ és el grau de sortida (*outdegree*) de j

Pagerank, III

Exemple



$$p_i = \sum_{j \to i} \frac{p_j}{out(j)}$$

Un conjunt de $n+1$ equacions lineals:

$$p_1 = \frac{p_1}{3} + \frac{p_2}{2}$$
$$p_2 = \frac{p_3}{2} + p_4$$
$$p_3 = \frac{p_1}{3}$$
$$p_4 = \frac{p_1}{3} + \frac{p_2}{2} + \frac{p_3}{2}$$
$$1 = p_1 + p_2 + p_3 + p_4$$

La solució és:

$p_1 = 6/23, p_2 = 8/23, p_3 = 2/23, p_4 = 7/23$

Pagerank, IV

Formalment

Equacions

- $p_i = \sum_{j:(j,i)\in E} \frac{p_j}{out(j)}$ for each $i \in V$
- $\sum_{i=1}^{n} p_i = 1$

on $out(i) = |\{j : (i,j) \in E\}|$ és el grau de sortida del node i

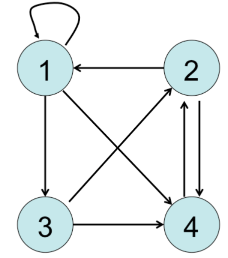
If $|V| = n$

- $n+1$ equacions
- n incògnites

Pot ser resolt, per exemple, per eliminació Gaussiana amb cost $O(n^3)$.

Pagerank, V

Exemple, revisat



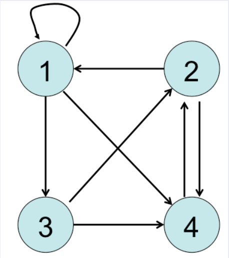
Un conjunt d'equacions lineals:

$$\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 1 \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} \cdot \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix}$$

on: $\vec{p} = M^T \vec{p}$ i a més $\sum_i p_i = 1$

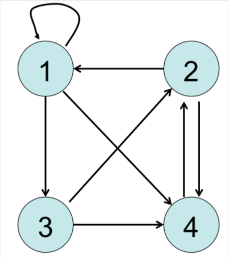
La solució és:

$\vec{p}$ és el vector propi de la matriu M^T associada al valor propi 1.



A què s'assembla M^T?

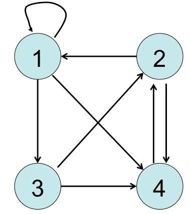
$$M^T = \begin{pmatrix} \frac{1}{3} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 1 \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$



A què s'assembla M^T?

$$M^T = \begin{pmatrix} \frac{1}{3} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 1 \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

M^T és la *transposada* de la matriu d'adjacència, normalitzada per files, del graf!



Matriu d'adjacència

$$A = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$M = \begin{pmatrix} 1/3 & 0 & 1/3 & 1/3 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 1 & 0 & 0 \end{pmatrix} \qquad M^T = \begin{pmatrix} 1/3 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 1/2 & 0 \end{pmatrix}$$

(les files sumen 1) (les columnes sumen 1)

Pagerank, VIII

Exemple, revisat

$$A = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} \qquad M = \begin{pmatrix} \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 & 0 \end{pmatrix} \qquad M^T = \begin{pmatrix} \frac{1}{3} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 1 \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

Pagerank, VIII

Exemple, revisat

$$A = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad M = \begin{pmatrix} \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad M^T = \begin{pmatrix} \frac{1}{3} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 1 \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

Pregunta:

Per què necessitem normalitzar per files i transposar A?

$$A = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad M = \begin{pmatrix} \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad M^T = \begin{pmatrix} \frac{1}{3} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 1 \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

Pregunta:

Per què necessitem normalitzar per files i transposar A?

Resposta:

- *Normalitzar per files*: perquè $p_i = \sum_{j:(j,i) \in E} \frac{p_j}{out(j)}$

Pagerank, VIII

Exemple, revisat

$$A = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} \qquad M = \begin{pmatrix} \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 & 0 \end{pmatrix} \qquad M^T = \begin{pmatrix} \frac{1}{3} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 1 \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

Pregunta:

Per què necessitem normalitzar per files i transposar A?

Resposta:

- *Normalitzar per files*: perquè $p_i = \sum_{j:(j,i)\in E} \frac{p_j}{out(j)}$
- *Transposar*: perquè $p_i = \sum_{j:(j,i)\in E} \frac{p_j}{out(j)}$, és a dir,

  p_i depèn de les *arestes entrants* a i.

Pagerank, IX

Resoldre un sistema d'equacions lineals

. . . però

- Com sabem que té solució?

Pagerank, IX

Resoldre un sistema d'equacions lineals

. . . però

- Com sabem que té solució?
- Com sabem que té una solució única?

Pagerank, IX

Resoldre un sistema d'equacions lineals

. . . però

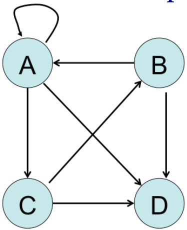
- Com sabem que té solució?
- Com sabem que té una solució única?
- Com podem calcular-la de forma eficient?

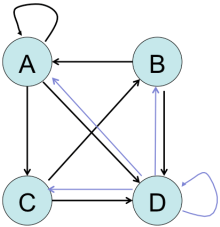
. . . però

- Com sabem que té solució?
- Com sabem que té una solució única?
- Com podem calcular-la de forma eficient?

Pe exemple, el graf de l'esquerra no té solució. . . (proveu-ho!) però el de la dreta sí.

d no apunta a ningu





Pagerank, X

Com sabem que té solució?

Per sort, l'*àlgebra lineal* ens dóna resposta

Definició

Una matriu M és estocàstica, si

- Tots els seus valors es troben en l'interval $[0, 1]$.
- Cada fila suma 1 (i.e., M està normalitzada per files).

Teorema [Perron-Frobenius]

Si M és estocàstica, llavors té almenys un vector estacionari, i.e., un vector diferent de zero p tal que

$$M^T p = p$$

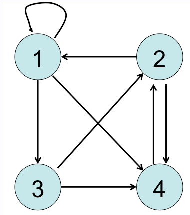
Pagerank, XI

L'altre punt de vista: Camí aleatori per la xarxa

Pagerank, XI

L'altre punt de vista: Camí aleatori per la xarxa

Suposem que M és la matriu de probabilitat de transició entre els estats de G.



$$M = \begin{pmatrix} 1/3 & 0 & 1/3 & 1/3 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

Sigui $\vec{p}(t)$ la probabilitat dels estats a l'instant t

- P.e., $p_j(0)$ és la probabilitat d'estar a l'estat j a l'instant 0

Un navegant salta aleatòriament de la pàgina i a la pàgina j amb probabilitat m_{ij}

- P.e., probabilitat de transició de l'estat 2 a l'estat 4 és $m_{24} = 1/2$

Pagerank, XII

L'altre punt de vista: Camí aleatori per la xarxa

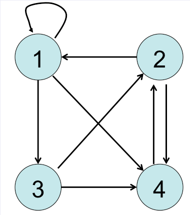
- El navegant comença en una pàgina aleatòria segons la distribució de probabilitat $\vec{p}(0)$.
- A l'instant $t > 0$, el navegant segueix un dels enllaços de la pàgina actual escollit de forma aleatòria

$$\vec{p}(t) := M^T \vec{p}(t-1)$$

- En el límit $t \to \infty$:
  - $\vec{p}(t) = \vec{p}(t+1) = \vec{p}(t+2) = \cdots = \vec{p}$
  - per tant $\vec{p}(t) = M^T \vec{p}(t-1)$
  - $\vec{p}(t)$ convergeix a la solució p perquè $p = M^T p$ (la solució del *pagerank*)!

Pagerank, XIII

L'altre punt de vista: Camí aleatori per la xarxa



$$M^T = \begin{pmatrix} \frac{1}{3} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 1 \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

- $\vec{p}(0)^T = (1, 0, 0, 0)$
- $\vec{p}(1)^T = (1/3, 0, 1/3, 1/3)$
- $\vec{p}(2)^T = (0.11, 0.50, 0.11, 0.28)$
- ..
- $\vec{p}(10)^T = (0.26, 0.35, 0.09, 0.30)$
- $\vec{p}(11)^T = (0.26, 0.35, 0.09, 0.30)$

Pagerank, XIV

L'algorisme que troba p t.q. $p = M^T p$

Mètode de la potència

- Tria a l'atzar un vector inicial $\vec{p}(0)$
- Repeteix $\vec{p}(t) \leftarrow M^T \vec{p}(t-1)$
- Fins que convergeixi (i.e. $\vec{p}(t) \approx \vec{p}(t-1)$)

Esperem que

Pagerank, XIV

L'algorisme que troba p t.q. $p = M^T p$

Mètode de la potència

- Tria a l'atzar un vector inicial $\vec{p}(0)$
- Repeteix $\vec{p}(t) \leftarrow M^T \vec{p}(t-1)$
- Fins que convergeixi (i.e. $\vec{p}(t) \approx \vec{p}(t-1)$)

Esperem que

- El mètode convergeixi.

Pagerank, XIV

L'algorisme que troba p t.q. $p = M^T p$

Mètode de la potència

- Tria a l'atzar un vector inicial $\vec{p}(0)$
- Repeteix $\vec{p}(t) \leftarrow M^T \vec{p}(t-1)$
- Fins que convergeixi (i.e. $\vec{p}(t) \approx \vec{p}(t-1)$)

Esperem que

- El mètode convergeixi.
- El mètode convergeixi ràpidament.

Pagerank, XIV

L'algorisme que troba p t.q. $p = M^T p$

Mètode de la potència

- Tria a l'atzar un vector inicial $\vec{p}(0)$
- Repeteix $\vec{p}(t) \leftarrow M^T \vec{p}(t-1)$
- Fins que convergeixi (i.e. $\vec{p}(t) \approx \vec{p}(t-1)$)

Esperem que

- El mètode convergeixi.
- El mètode convergeixi ràpidament.
- El mètode convergeixi ràpidament a la solució del *pagerank*.

Pagerank, XIV

L'algorisme que troba p t.q. $p = M^T p$

Mètode de la potència

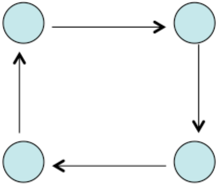
- Tria a l'atzar un vector inicial $\vec{p}(0)$
- Repeteix $\vec{p}(t) \leftarrow M^T \vec{p}(t-1)$
- Fins que convergeixi (i.e. $\vec{p}(t) \approx \vec{p}(t-1)$)

Esperem que

- El mètode convergeixi.
- El mètode convergeixi ràpidament.
- El mètode convergeixi ràpidament a la solució del *pagerank*.
- El mètode convergeixi ràpidament a la solució del *pagerank* independentment del vector inicial.

Pagerank, XV

Convergència del mètode de la potència



Provem el mètode de la potència amb $\vec{p}(0)$:

$$\begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix}, \text{ o } \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \text{ o } \begin{pmatrix} 1/2 \\ 0 \\ 1/2 \\ 0 \end{pmatrix}$$

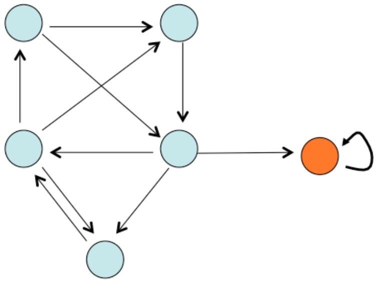
No podem trencar el cicle!

- . . . cal que el graf sigui aperiòdic
  - no hi hagi cap enter $k > 1$ que divideixi la longitud d'un cicle

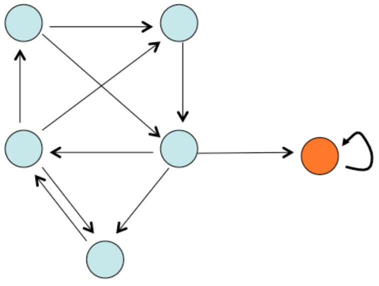
Pagerank, XVI

Convergència del mètode de la potència

Què passa amb el *pagerank* d'aquest graf?



Què passa amb el *pagerank* d'aquest graf?



L'aglutinador acumula tot el *pagerank*!

- cal trobar una manera de sortir de l'aglutinador
- . . . imposarem que els grafs estiguin fortament connectats.

Pagerank, XVII

Un teorema de la teoria de les cadenes de Markov

Teorema

Si una matriu M està fortament connectada i és aperiòdica, llavors:

- $M^T\vec{p} = \vec{p}$ té exactament una solució diferent de zero tal que $\sum_i p_i = 1$
- 1 és el valor propi més gran de M^T
- el mètode de potència convergeix cap a $\vec{p}$ satisfent $M^T\vec{p} = \vec{p}$, des d'un vector inicial diferent de zero $\vec{p}(0)$
- és més, la convergència és ràpida (és exponencial).

Pagerank, XVII

Un teorema de la teoria de les cadenes de Markov

Teorema

Si una matriu M està fortament connectada i és aperiòdica, llavors:

- $M^T\vec{p} = \vec{p}$ té exactament una solució diferent de zero tal que $\sum_i p_i = 1$
- 1 és el valor propi més gran de M^T
- el mètode de potència convergeix cap a $\vec{p}$ satisfent $M^T\vec{p} = \vec{p}$, des d'un vector inicial diferent de zero $\vec{p}(0)$
- és més, la convergència és ràpida (és exponencial).

Per garantir una solució, haurem d'assegurar que les matrius amb les que treballem estiguin fortament connectades i siguin aperiòdiques.

Pagerank, XVIII

Garantir aperiodicitat i forta connectivitat

Definició (la matriu de Google)

Donat un factor d'*amortiment (damping)* λ tal que: $0 < \lambda < 1$:

$$G = \lambda M + (1 - \lambda)\frac{1}{n}J$$

on J és una matriu $n \times n$ tota plena de 1

Pagerank, XVIII

Garantir aperiodicitat i forta connectivitat

Definició (la matriu de Google)

Donat un factor d'*amortiment (damping)* λ tal que: $0 < \lambda < 1$:

$$G = \lambda M + (1 - \lambda)\frac{1}{n}J$$

on J és una matriu $n \times n$ tota plena de 1

Observeu que:

Pagerank, XVIII

Garantir aperiodicitat i forta connectivitat

Definició (la matriu de Google)

Donat un factor d'*amortiment (damping)* λ tal que: $0 < \lambda < 1$:

$$G = \lambda M + (1 - \lambda)\frac{1}{n}J$$

on J és una matriu $n \times n$ tota plena de 1

Observeu que:

- G és estocàstica

Definició (la matriu de Google)

Donat un factor d'*amortiment (damping)* λ tal que: $0 < \lambda < 1$:

$$G = \lambda M + (1 - \lambda)\frac{1}{n}J$$

on J és una matriu $n \times n$ tota plena de 1

Observeu que:

- G és estocàstica
  - ...perquè G és la mitjana ponderada de M i $\frac{1}{n}J$, que també són estocàstiques

Pagerank, XVIII

Garantir aperiodicitat i forta connectivitat

Definició (la matriu de Google)

Donat un factor d'*amortiment (damping)* λ tal que: $0 < \lambda < 1$:

$$G = \lambda M + (1 - \lambda)\frac{1}{n}J$$

on J és una matriu $n \times n$ tota plena de 1

Observeu que:

- G és estocàstica
  - ...perquè G és la mitjana ponderada de M i $\frac{1}{n}J$, que també són estocàstiques
- per tot enter $k > 0$, hi ha un camí de longitud k de tot estat a qualsevol altre de G amb una probabilitat diferent de zero

Pagerank, XVIII

Garantir aperiodicitat i forta connectivitat

Definició (la matriu de Google)

Donat un factor d'*amortiment (damping)* λ tal que: $0 < \lambda < 1$:

$$G = \lambda M + (1 - \lambda)\frac{1}{n}J$$

on J és una matriu $n \times n$ tota plena de 1

Observeu que:

- G és estocàstica
  - ...perquè G és la mitjana ponderada de M i $\frac{1}{n}J$, que també són estocàstiques
- per tot enter $k > 0$, hi ha un camí de longitud k de tot estat a qualsevol altre de G amb una probabilitat diferent de zero
  - ...que implica que G està fortament connectada i és aperiòdica

Pagerank, XVIII

Garantir aperiodicitat i forta connectivitat

Definició (la matriu de Google)

Donat un factor d'*amortiment (damping)* λ tal que: $0 < \lambda < 1$:

$$G = \lambda M + (1 - \lambda)\frac{1}{n}J$$

on J és una matriu $n \times n$ tota plena de 1

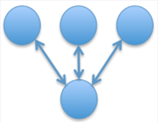
Observeu que:

- G és estocàstica
  - ...perquè G és la mitjana ponderada de M i $\frac{1}{n}J$, que també són estocàstiques
- per tot enter $k > 0$, hi ha un camí de longitud k de tot estat a qualsevol altre de G amb una probabilitat diferent de zero
  - ...que implica que G està fortament connectada i és aperiòdica
- i, per tant, el mètode de la potència convergirà amb G i de pressa!

El significat de λ

- Amb probabilitat λ, el navegant aleatori segueix un enllaç de la pàgina actual.
- Amb probabilitat $1 - \lambda$, el navegant aleatori salta a una altra pàgina aleatòria del graf (teleportació).

Calculeu el valor de *pagerank* per cada node del graf següent suposant un *factor d'amortiment* de $\lambda = 2/3$:



Pista: resoleu el sistema d'equacions següent usant $p_2 = p_3 = p_4$

$$\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix} = \left[\frac{2}{3} \begin{pmatrix} 0 & 1 & 1 & 1 \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 \end{pmatrix} + \frac{1}{3} \cdot \frac{1}{4} \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \right] \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix}$$

Pagerank sensible al context, I

Observeu que el pagerank d'una pàgina és independent de la consulta de l'usuari

- Avantatges
  - Es pot calcular off-line.
  - El valor depèn de la reputació col·lectiva.
- Desavantatges
  - Insensible a les necessitats concretes de l'usuari.

Pagerank sensible al context, II

Suposem que hi ha un conjunt reduït de K temes (esports, ciència, política. . .)

- Cada tema $k \in \{1, .., K\}$ està definit per un subconjunt de pàgines T_k.
- Per cada k, es calcula el pagerank del node i pel tema k:

$$p_{i,k} = \text{“pagerank del node } i \text{ amb teleportació reduïda a } T_k\text{”}$$

- Finalment, es calcula el rànquing de la pàgina i donada la consulta q

$$rank(i, q) = \sum_{k=1}^{K} sim(T_k, q) \cdot p_{i,k}$$

Índex

1. 5. Cerca a internet. Arquitectura de sistemes de RI simples
   - Arquitectura d'un sistema de cerca a internet
   - Cerca a internet
   - Algorisme de Pagerank
   - Algorisme HITS

HITS, I

Hypertext-Induced Topic Search

L'interès d'una pàgina web és degut a dos aspectes diferents

- si el contingut de la pàgina és interessant (*authority* o grau d'autoritat), i
- si la pàgina enllaça a d'altres pàgines interessants (*hub* o pàgina guia/recurs).

Fonament principal de HITS

- els *hubs* són importants si enllacen *authorities* importants
- les *authorities* són importants si són enllaçades des de *hubs* importants

HITS, I

Hypertext-Induced Topic Search

L'interès d'una pàgina web és degut a dos aspectes diferents

- si el contingut de la pàgina és interessant (*authority* o grau d'autoritat), i
- si la pàgina enllaça a d'altres pàgines interessants (*hub* o pàgina guia/recurs).

Fonament principal de HITS

- els *hubs* són importants si enllacen *authorities* importants
- les *authorities* són importants si són enllaçades des de *hubs* importants
- . . . tornem a tenir una definició circular. . . però tampoc és un problema!

HITS, II

Associar a cada pàgina i un valor d'*authority* a_i i un valor de *hub* h_i

- el vector amb tots els valors *authority* és $\vec{a}$
- el vector amb tots els valors *hub* és $\vec{h}$

Mantenir aquests vectors normalitzats (usant norma L2!)

- $\|\vec{a}\| = \sum_i a_i^2 = 1,$ i $\|\vec{h}\| = \sum_i h_i^2 = 1$

Amb constants d'escala apropiades c i d

- $a_i = c \cdot \sum_{j \to i} h_j,$ i $h_i = d \cdot \sum_{i \to j} a_j$

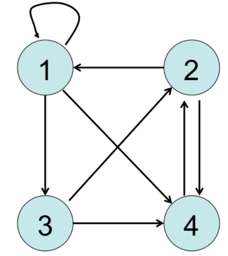
Observeu que ja no es tracta d'un sistema lineal

- …però que encara es pot resoldre amb una variant del mètode de la potència.

HITS, III

Exemple

L'antic graf



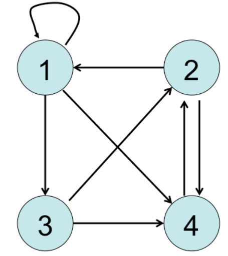
Matriu d'adjacència

$$A = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$a_1 = c \cdot (h_1 + h_2)$ *// aquí usem la primera columna d'A*

$$a_1 \propto (1, 1, 0, 0) \cdot \begin{pmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \end{pmatrix} = (1, 1, 0, 0) \cdot \vec{h}$$

L'antic graf



Matriu d'adjacència

$$A = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$h_2 = d \cdot (a_1 + a_4)$ *// aquí usem la segona fila d'A*

$$h_2 \propto (1,0,0,1) \cdot \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = (1,0,0,1) \cdot \vec{a}$$

HITS, V

Regla per actualitzar $\vec{a}$ i $\vec{h}$

Escrit en forma matricial compacta

- Per actualitzar els valors *authority*
  - $\vec{a} := A^T \cdot \vec{h}$
  - normalitzar després $\vec{a} := \frac{\vec{a}}{\|a\|}$ de manera que $\|a\| = 1$
- Per actualitzar els valors *hub*
  - $\vec{h} := A \cdot \vec{a}$
  - normalitzar després $\vec{h} := \frac{\vec{h}}{\|h\|}$ de manera que $\|h\| = 1$

Donada la matriu d'adjacència A

- Inicialitzar $\vec{a} = \vec{h} = (1, 1, .., 1)^T$
- Normalitzar $\vec{a}$ i $\vec{h}$ de manera que $\|a\| = \|h\| = 1$
- Repetir fins convergir
  - $\vec{a} := A^T \cdot \vec{h}$
  - normalitzar $\vec{a}$ de manera que $\|a\| = 1$
  - $\vec{h} := A \cdot \vec{a}$
  - normalitzar $\vec{h}$ de manera que $\|h\| = 1$

Resposta a la consulta amb l'algorisme HITS

- Llegir la consulta q i llençar-la en un cercador basat en la concordança del text buscat.
- Agafar les k primeres pàgines i formar el $RootSet$.
- Formar el $BaseSet$ estenent el $RootSet$ amb totes les pàgines enllaçades des de pàgines del $RootSet$ i amb les pàgines que apuntin cap a pàgines del $RootSet$ (fins un llindar, p.e. 50).
- Calcular els valors *hub* i *authority* pel subgraf de la xarxa induïda pel $BaseSet$.
- Ordenar les pàgines del $BaseSet$ d'acord a $\vec{a}$, $\vec{h}$ i contingut.

HITS, VIII

Algorisme HITS il·lustrat

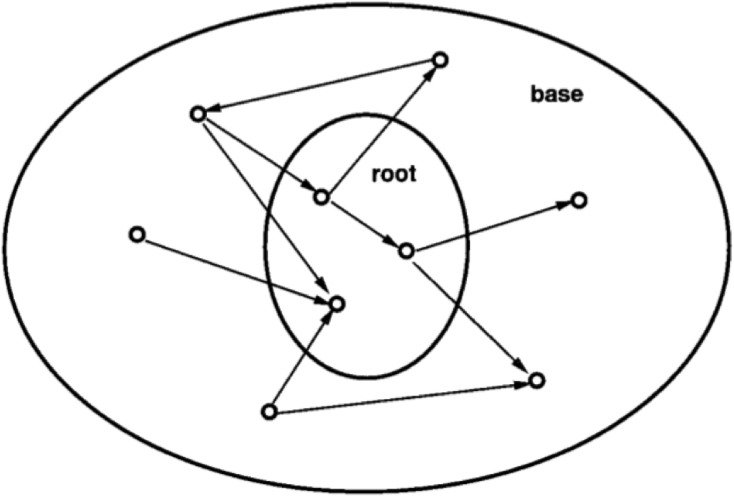


FIG. 1. Expanding the root set into a base set.

HITS vs. Pagerank

Pros de HITS vs. Pagerank

- Sensible a les consultes dels usuaris.

Cons de HITS vs. Pagerank

- Càlcul online, no off-line!
- Més vulnerable al *web spamming* (p.e., afegint molts enllaços de sortida de la nostra pàgina).