Lineární algebra a jehla v kupce sena

With all due respect John, I am the head of IT and I have it on good authority that if you type "Google" into Google, you can break the Internet. So please, no one try it, even for a joke. It's not a laughing matter. You can break the Internet.

Jen Barber, head of IT Department, Reynholm Industries

Historie třídění informací

1. ∼ 3.stol.př.n.l.: kartičkový katalog (knihovna v Ninive).
2. 1876: Deweyho desetinné třídění.
3. ∼1960: MARC (MAchine Readable Cataloging).
4. . . .

Historie počítačového vyhledávání informací

1. 1945: popis stroje Memex, připomínající PC a WWW.
   - ☞ Vannevar Bush, As we may think, *Atlantic Monthly*, 176 (1945), 101–108.
2. ∼1960: SMART (vyhledávací systém Cornellovy university).
3. 1989: T. Berners-Lee navrhl jazyk html. Svět se změnil.
4. 1998: S. Brin a L. Page navrhli algoritmus PageRank analysující topologii WWW. Svět se změnil znovu.

Plán povídání

1. Myšlenka vektoru důležitosti a matice hyperlinků.
2. Trocha teorie z počátku 20. století od Oskara Perrona a Ferdinanda Georga Frobenia.
3. Algoritmus Sergeje Brina a Lawrence Page.

Povídání je velmi ovlivněno články

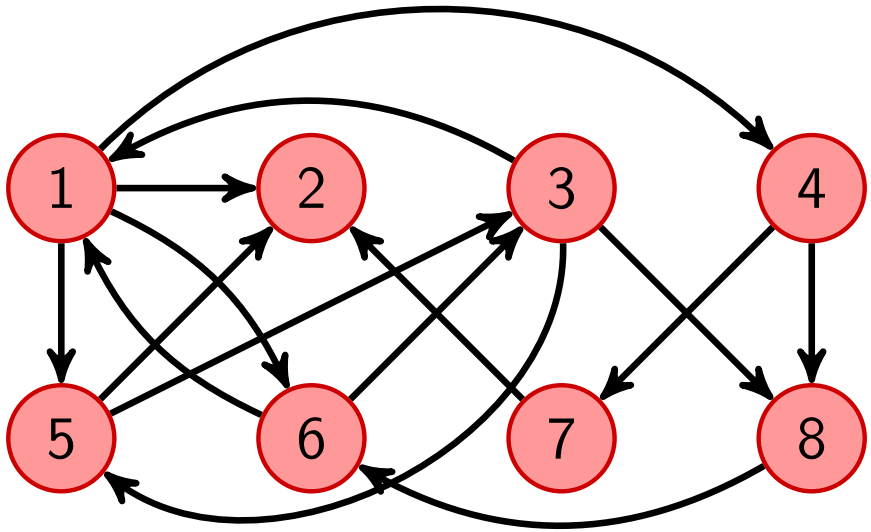
- ☞ David Austin, How Google finds your needle in the web's haystack, *AMS Feature Columns*, 2008.
- ☞ Amy N. Langville a Carl D. Meyer, Deeper inside PageRank, *Internet Math.*, 1 (2005), 335–380.
- ☞ Kurt Bryan a Anna Leise, The $ 25,000,000,000 eigenvector: The linear algebra behind Google, *SIAM Rev.* 48.3 (2006), 569–581.

Jak se na webu vyhledává?

1. Search engine neustále prohledává web a indexuje všechny stránky s veřejným přístupem.
2. Stránky z takto vytvořené databáze search engine indexuje klíčovými slovy a frázemi.
3. Jednotlivým stránkám z databáze search engine dává skóre důležitosti. Pak se stránky, označené jako důležité, objeví mezi prvními.

Budeme mluvit pouze o posledním bodu

WWW budeme modelovat orientovaným grafem



Vrcholy: jednotlivé WWW stránky.
Hrany: hyperlinky mezi stránkami.

Základní myšlenka důležitosti WWW stránky: stránka je důležitá, pokud na ni odkazují důležité stránky.[a]

[a] Tato rekursivní povaha věci bude zdrojem dalších radostí a strastí.

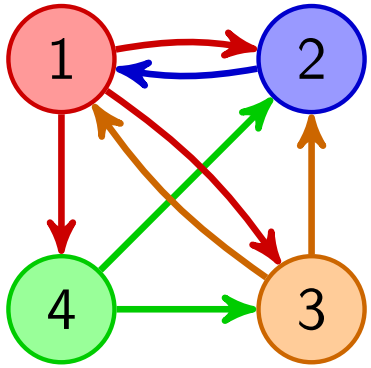
Odbočka: rekursivní povahu důležitosti lze nalézt téměř všude

1. Ekonometrie: firma je výnosná, pokud má zakázky od výnosných firem.
   - ☞ W. W. Leontieff, *The structure of American economy 1919–1929*, Harvard Univ. Press, 1941.
2. Sociometrie: osoba má prestiž, pokud je schvalována prestižními osobami.
   - ☞ J. R. Seeley, The net of reciprocal influence: A problem in treating sociometric data, *The Canadian Journal of Psychology* 3 (1949), 234–240.
3. Bibliometrie: časopis má vliv, pokud je citován vlivnými časopisy.
   - ☞ G. Pinski a F. Narin, Citation influence for journal aggregates of scientific publications: Theory, with applications to the literature of physics, *Inf. Processing & Management* 12.5 (1976), 297–312.

Více detailů lze nalézt například v článku:

- ☞ Massimo Franceschet, PageRank: Standing on the shoulders of giants, arXiv:1002.2858v3, 2010.

Matice hyperlinků



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

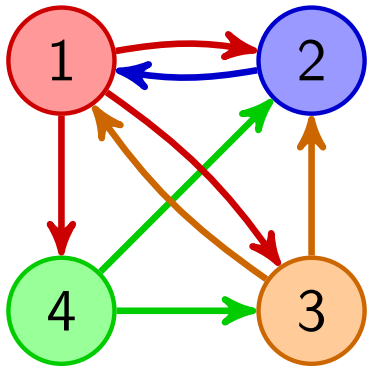
Interpretace (například) čísla 1/3 v 1. sloupci a 3. řádku:

stránka 1 předává 1/3 své důležitosti stránce 3.

Důležité pozorování: součet každého sloupce $\mathbf{H}$ je roven 1.[a]

[a] Vznešená terminologie: $\mathbf{H}$ je sloupcově stochastická. Každý sloupec je pak možné považovat za pravděpodobnostní rozdělení.

Vektor důležitosti



$$\mathbf{g}_0 = \begin{pmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{pmatrix}, \mathbf{H}\mathbf{g}_0 = \begin{pmatrix} 0.37 \\ 0.33 \\ 0.20 \\ 0.08 \end{pmatrix}, \mathbf{H}^2\mathbf{g}_0 = \begin{pmatrix} 0.43 \\ 0.27 \\ 0.16 \\ 0.12 \end{pmatrix}, \mathbf{H}^3\mathbf{g}_0 = \begin{pmatrix} 0.35 \\ 0.29 \\ 0.20 \\ 0.14 \end{pmatrix}$$

$$\mathbf{H}^4\mathbf{g}_0 = \begin{pmatrix} 0.39 \\ 0.29 \\ 0.19 \\ 0.11 \end{pmatrix}, \mathbf{H}^5\mathbf{g}_0 = \begin{pmatrix} 0.39 \\ 0.28 \\ 0.19 \\ 0.13 \end{pmatrix}, \mathbf{H}^6\mathbf{g}_0 = \begin{pmatrix} 0.38 \\ 0.29 \\ 0.19 \\ 0.13 \end{pmatrix}, \mathbf{H}^7\mathbf{g}_0 = \begin{pmatrix} 0.38 \\ 0.29 \\ 0.19 \\ 0.12 \end{pmatrix} = \mathbf{H}^8\mathbf{g}_0$$

Postup se zdá být správným: našli jsme $\mathbf{g}$, pro které (při zaokrouhlení na dvě desetinná místa) platí $\mathbf{H}\mathbf{g} = \mathbf{g}$ a vektor $\mathbf{g}$ je pravděpodobnostní rozdělení.

Požadavky na vektor důležitosti pro matici hyperlinků

Od vektoru $\mathbf{g}$ chceme:

1. $\mathbf{H}\mathbf{g} = \mathbf{g}$.

   Rovnost modeluje dogma: stránka je důležitá, pokud na ni odkazují důležité stránky.

2. Položky vektoru $\mathbf{g}$ tvoří pravděpodobnostní rozdělení.

   Požadavek modeluje relativní důležitost jednotlivých stránek.

   Vektoru $\mathbf{g}$ lze dát i jinou interpretaci: i-tá položka vektoru $\mathbf{g}$ representuje zlomek času, který průměrný surfař na stránce i stráví.[a]

Uvidíme, že matice hyperlinků $\mathbf{H}$ je (pro naše účely) poměrně nevhodný způsob modelování webu.

[a] To by mělo znamenat, že každá položka vektoru $\mathbf{g}$ je nenulová.

Naivní metoda hledání vektoru důležitosti

1. Sestavíme matici hyperlinků $\mathbf{H}$.
2. Zvolíme nenulový iniciální vektor důležitosti $\mathbf{g}_0$.
3. Definujeme posloupnost $\mathbf{g}_k$ předpisem $\mathbf{g}_k = \mathbf{H}^k \mathbf{g}_0$.

Tři otázky

1. Zdalipak posloupnost $\mathbf{g}_k$ vždy konverguje?
2. Zdalipak limita $\mathbf{g}$ posloupnosti $\mathbf{g}_k$ je na $\mathbf{g}_0$ nezávislá?
3. Zdalipak vektor $\mathbf{g}$ representuje to, co jsme chtěli?

Tři odpovědi

Naivní metoda hledání vektoru důležitosti

1. Sestavíme matici hyperlinků $\mathbf{H}$.
2. Zvolíme nenulový iniciální vektor důležitosti $\mathbf{g}_0$.
3. Definujeme posloupnost $\mathbf{g}_k$ předpisem $\mathbf{g}_k = \mathbf{H}^k \mathbf{g}_0$.

Tři otázky

1. Zdalipak posloupnost $\mathbf{g}_k$ vždy konverguje?
2. Zdalipak limita $\mathbf{g}$ posloupnosti $\mathbf{g}_k$ je na $\mathbf{g}_0$ nezávislá?
3. Zdalipak vektor $\mathbf{g}$ representuje to, co jsme chtěli?

Tři odpovědi

1. NE!

Naivní metoda hledání vektoru důležitosti

1. Sestavíme matici hyperlinků $\mathbf{H}$.
2. Zvolíme nenulový iniciální vektor důležitosti $\mathbf{g}_0$.
3. Definujeme posloupnost $\mathbf{g}_k$ předpisem $\mathbf{g}_k = \mathbf{H}^k \mathbf{g}_0$.

Tři otázky

1. Zdalipak posloupnost $\mathbf{g}_k$ vždy konverguje?
2. Zdalipak limita $\mathbf{g}$ posloupnosti $\mathbf{g}_k$ je na $\mathbf{g}_0$ nezávislá?
3. Zdalipak vektor $\mathbf{g}$ representuje to, co jsme chtěli?

Tři odpovědi

1. NE!
2. NE!

Naivní metoda hledání vektoru důležitosti

1. Sestavíme matici hyperlinků $\mathbf{H}$.
2. Zvolíme nenulový iniciální vektor důležitosti $\mathbf{g}_0$.
3. Definujeme posloupnost $\mathbf{g}_k$ předpisem $\mathbf{g}_k = \mathbf{H}^k \mathbf{g}_0$.

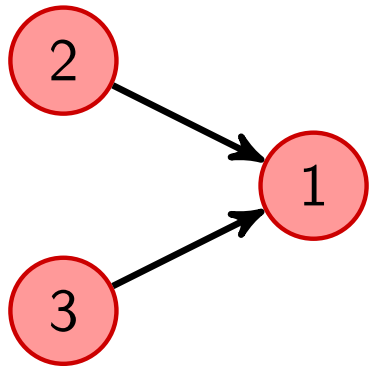
Tři otázky

1. Zdalipak posloupnost $\mathbf{g}_k$ vždy konverguje?
2. Zdalipak limita $\mathbf{g}$ posloupnosti $\mathbf{g}_k$ je na $\mathbf{g}_0$ nezávislá?
3. Zdalipak vektor $\mathbf{g}$ representuje to, co jsme chtěli?

Tři odpovědi

1. NE!
2. NE!
3. NE!

Problém viselců



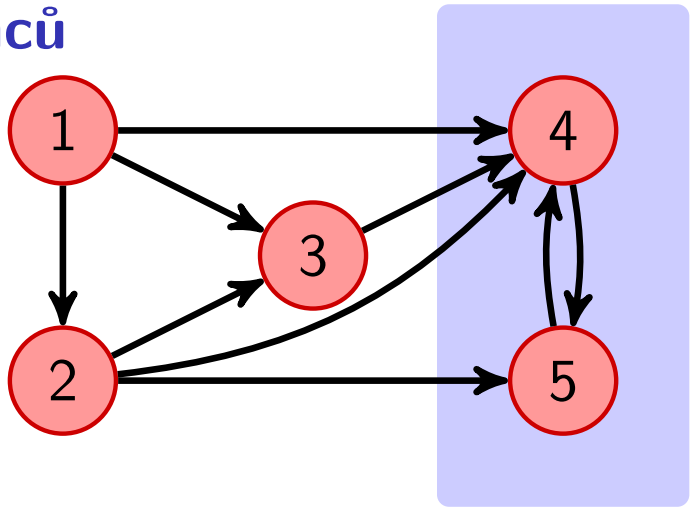
$$\mathbf{H} = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Protože platí $\mathbf{H}^2 = \mathbf{O}$, konverguje $\mathbf{g}_k$ vždy k nulovému vektoru. To je zcela proti naší intuici. Stránka 1 by měla být dvakrát důležitější než každá ze stránek 2 a 3.

Problémem je viselec (dangling node). Stránka 1 postupně vyssává všechnu důležitost ostatních stránek.

Další problém: $\mathbf{H}$ není sloupcově stochastická.

Problém nafoukanců

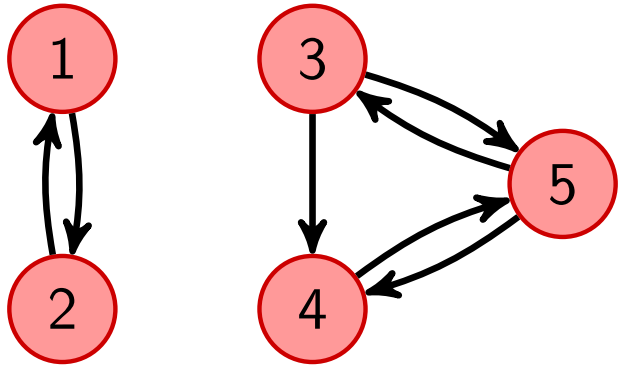


$$\mathbf{H} = \left(\begin{array}{cc|ccc} 0 & 0 & 0 & 0 & 0 \\ 0.33 & 0 & 0 & 0 & 0 \\ 0.33 & 0.33 & 0 & 0 & 0 \\ \hline 0.33 & 0.33 & 1 & 0 & 0 \\ 0 & 0.33 & 0 & 1 & 1 \end{array}\right)$$

Jde o zobecnění předchozího příkladu. Komponenta silné souvislosti $\{4, 5\}$ vyssaje důležitost ostatních stránek.

Matice $\mathbf{H}$ je sloupcově stochastická, ale není ireducibilní (má blok nul). Pak i $\mathbf{g}$ (tj. limita $\mathbf{g}_k$) má blok nul.

Problém chlapců, co spolu nemluví

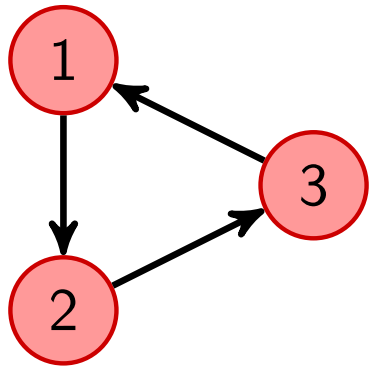


$$\mathbf{H} = \left(\begin{array}{cc|ccc} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{array}\right) = \left(\begin{array}{c|c} \mathbf{H}_1 & \mathbf{O} \\ \hline \mathbf{O} & \mathbf{H}_2 \end{array}\right)$$

Dvě komponenty silné souvislosti webu způsobí existenci dvou nezávislých kandidátů na vektor $\mathbf{g}$ (vzniklých z vektorů $\mathbf{g}^{\mathbf{H}_1}$ a $\mathbf{g}^{\mathbf{H}_2}$ doplněním nul).

Matice $\mathbf{H}$ je sloupcově stochastická, ale není ireducibilní.

Problém marné touhy po uznání



$$\mathbf{H} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

Platí $\mathbf{H}^3 = \mathbf{E}$, takže $\mathbf{g}_k$ obecně nemůže konvergovat.

Matice $\mathbf{H}$ je sloupcově stochastická, ale není primitivní (tj. neexistuje m tak, že $\mathbf{H}^m$ má všechny položky kladné).

Závěr

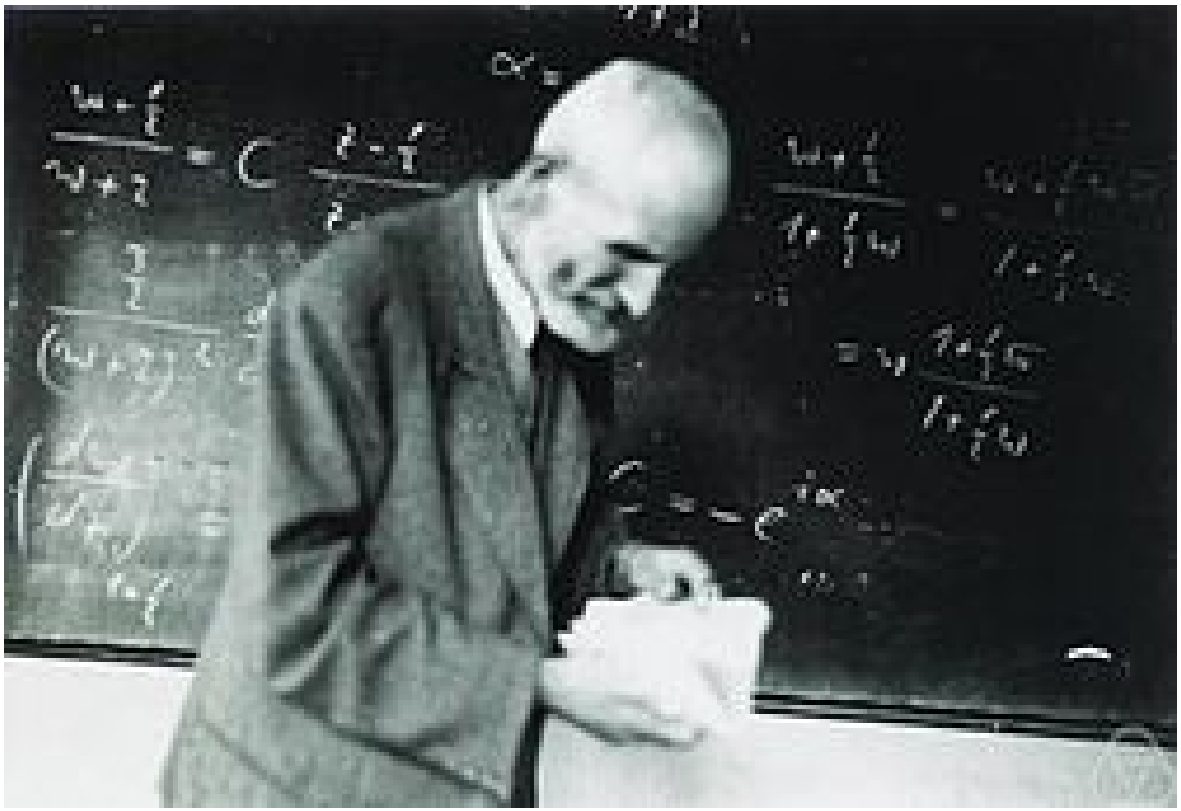
Je zjevné, že matice hyperlinků $\mathbf{H}$ musí splňovat jisté podmínky, aby dynamický proces $\mathbf{g}_k = \mathbf{H}^k \mathbf{g}_0$ splňoval naše požadavky.

Potřebujeme tedy analýzu chování obecných matic $\mathbf{M}$ s nezápornými reálnými položkami.

Naštěstí takovou teorii máme již dlouho k disposici:

- ☞ Oskar Perron, Zur Theorie der Matrices, *Math. Ann.* 64 (1907), 248–263.
- ☞ Ferdinand Georg Frobenius, Über Matrizen aus positiven Elementen I, *Sitzungber. Königl. Preuss. Akad. Wiss.* (1908), 471–476.
- ☞ Ferdinand Georg Frobenius, Über Matrizen aus positiven Elementen II, *Sitzungber. Königl. Preuss. Akad. Wiss.* (1909), 514–518.
- ☞ Ferdinand Georg Frobenius, Über Matrizen aus nicht negativen Elementen, *Sitzungber. Königl. Preuss. Akad. Wiss.* (1912), 456–477.

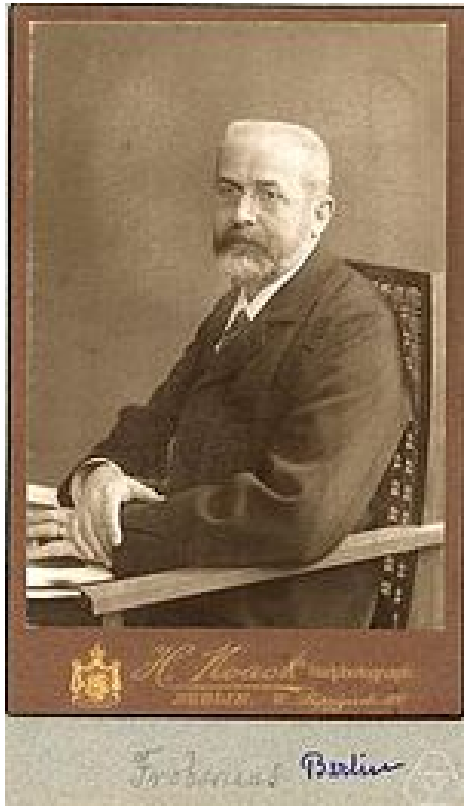
Oskar Perron



Oskar Perron (7. května 1880 – 22. února 1975)

- Narozen ve Frankenthalu (Pfalz).
- 1898–1902: Universität München, PhD u Ferdinanda von Lindenmanna (téma: rotace pevných těles).
- Poté profesury: Tübingen, Heidelberg, München.
- Práce: diferenciální rovnice, řetězové zlomky, teorie integrálu.
- Přednášet přestal v 80 letech, ve svých 84–93 letech publikoval 18 článků.

Ferdinand Georg Frobenius



Ferdinand Georg Frobenius (26. října 1849 – 3. srpna 1917)

- Narozen v Charlottenburgu (předměstí Berlína).
- 1870: PhD u Karla Weierstrasse (téma: diferenciální rovnice).
- 1875–1892: Zürich (Eidgenössische Polytechnikum).
- Od 1893: Universität Berlin.
- Práce: teorie grup (důkazy Sylowových vět), teorie čísel, teorie matic (obecný důkaz Cayley-Hamiltonovy věty)

Definice

Ať **M** je čtvercová matice s reálnými položkami. Řekneme, že

1. **M** je nezáporná, když všechny položky matice **M** jsou nezáporné.
2. **M** je positivní, když všechny položky matice **M** jsou kladné.
3. **M** je primitivní, když je nezáporná a pro nějaké $k \geq 1$ je $\mathbf{M}^k$ positivní.
4. **M** je ireducibilní, když je nezáporná a příslušný orientovaný graf[a] je silně souvislý.

[a] Hrana vede z i do j právě tehdy, když $m_{ij} > 0$.

Připomenutí: L_1-norma v $\mathbb{R}^n$

Pro vektor $\mathbf{v}$ z $\mathbb{R}^n$ je $\|\mathbf{v}\|_1 = \sum_{i=1}^{n} |v_i|$.

Věta (Perron-Frobenius)

Ať $\mathbf{M}$ je čtvercová positivní sloupcově-stochastická matice. Potom platí:

1. $\mathbf{M}$ je primitivní a ireducibilní.
2. $\mathbf{M}$ má vlastní hodnotu 1 násobnosti 1 a dimense vlastního podprostoru $\ker(\mathbf{M}-\mathbf{E})$ je 1.

   V $\ker(\mathbf{M}-\mathbf{E})$ existuje jediný positivní vektor $\mathbf{v}$, $\|\mathbf{v}\|_1 = 1$.

   Platí $\mathbf{v} = \lim_{k\to+\infty} \mathbf{M}^k\mathbf{v}_0$, pro jakékoli positivní $\mathbf{v}_0$, $\|\mathbf{v}_0\|_1 = 1$.
3. Pro jakoukoli další vlastní hodnotu λ matice $\mathbf{M}$ platí $|\lambda| < 1$.
4. Limita $\lim_{k\to+\infty} \mathbf{M}^k$ existuje, je to ortog. projekce na $\ker(\mathbf{M}-\mathbf{E})$.

Poznámka

Celá Perronova-Frobeniova teorie je daleko obecnější: uvažuje nezáporné ireducibilní matice, ne nutně sloupcově-stochastické.

Sergey Brin a Larry Page

Definice (The Google Matrix)

Označme $n =$ počet stránek webu. Ať $\mathbf{e}$ je vektor v $\mathbb{R}^n$ s položkami 1 a ať $\mathbf{p}$[a] je jakýkoli positivní vektor z $\mathbb{R}^n$, kde $\|\mathbf{p}\|_1 = 1$.

Google matrix je matice $\mathbf{G}_\alpha = \alpha(\mathbf{H} + \mathbf{A}) + (1 - \alpha)\mathbf{T}$ rozměrů $n \times n$, kde

1. $\mathbf{H}$ je matice hyperlinků webu.
2. $\mathbf{A}$ je adjustační matice. Matice $\mathbf{A}$ má jako položky samé nuly, kromě sloupců, ve kterých $\mathbf{H}$ má samé nuly: v nich má $\mathbf{A}$ vektor $\mathbf{p}$.[b]
3. $\mathbf{T} = \mathbf{e}\mathbf{p}^T$ je teleportační matice.[c]
4. α je reálné číslo, $0 \leq \alpha \leq 1$. Číslo α udává pravděpodobnost toho, že surfař se řídí hyperlinky.[d]

---

[a] Vektoru $\mathbf{p}$ se říká personalisation vector, původní volba je $\mathbf{p} = (1/n)\mathbf{e}$.

[b] Tj., $\mathbf{H} + \mathbf{A}$ je sloupcově stochastická. Matice $\mathbf{A}$ „odstraňuje“ viselce.

[c] Tj., $\mathbf{T}$ má v každém sloupci vektor $\mathbf{p}$, je tedy sloupcově stochastická.

[d] Brin a Page použili $\alpha = 0.85$. Tj., průměrný surfař se zhruba v 5/6 případů řídí strukturou webu, zhruba v 1/6 případů se „teleportuje“ na jinou stránku. Pravděpodobnostní rozdělení teleportace je dáno vektorem $\mathbf{p}$.

Věta

Pro jakoukoli hodnotu $\alpha < 1$ je matice $\mathbf{G}_\alpha$ kladná a sloupcově stochastická. Tudíž pro jakoukoli hodnotu $\alpha < 1$ platí:

1. Matice $\mathbf{G}_\alpha$ má vlastní hodnotu 1 násobnosti 1 a vlastní podprostor $\ker(\mathbf{G}_\alpha - \mathbf{E})$ má dimensi 1.

   Pro vlastní hodnotu 1 existuje jediný positivní vlastní vektor $\mathbf{g}$, pro který platí $\|\mathbf{g}\|_1 = 1$.

   Platí $\mathbf{g} = \lim_{k\to+\infty} (\mathbf{G}_\alpha)^k \mathbf{g}_0$,[a] pro jakékoli positivní $\mathbf{g}_0$, $\|\mathbf{g}_0\|_1 = 1$.
2. Pro jakoukoli další vlastní hodnotu λ matice $\mathbf{G}_\alpha$ platí $|\lambda| \leq \alpha$.
3. Limita $\lim_{k\to+\infty} (\mathbf{G}_\alpha)^k$ existuje, je to ortogonální projekce na $\ker(\mathbf{G}_\alpha - \mathbf{E})$.

---

[a] Rychlost této konvergence je rychlost konvergence $\lim_{k\to+\infty} \alpha^k = 0$. Tj., pro $\alpha = 0.85$ a (například) přesnost 10^{-10} vektoru $\mathbf{g}$ stačí 142 iterací.

Literatura k Perronově a Frobeniově teorii

- ☞ Carl D. Meyer, *Matrix analysis and applied linear algebra*, SIAM Publishers, 2001.

Literatura k algoritmu PageRank

- ☞ Sergey Brin a Lawrence Page, The anatomy of a large-scale hypertextual Web search engine, *Computer networks and ISDN systems* 33 (1998), 107–117.
- ☞ Kurt Bryan a Anna Leise, The $ 25,000,000,000 eigenvector: The linear algebra behind Google, *SIAM Rev.* 48.3 (2006), 569–581.
- ☞ Amy N. Langville a Carl D. Meyer, Deeper inside PageRank, *Internet Math.*, 1 (2005), 335–380.
- ☞ Amy N. Langville a Carl D. Meyer, *Google's PageRank and beyond: The science of search engine rankings*, Princeton Univ. Press 2006.