

Keresés és matematika

Tű a szénakazalban

RÓNYAI LAJOS
MTA SZTAKI, BME

Szűk kérdés

Searched the web for *Krúdy pisztráng*.

Results 1 - 10 of about 20. Search took 0.10 seconds.

Krúdy Gyula: Szindbád ...mindmegannyi Gergely pápák, egy piros pisztráng ijedten suhant tova ...
www.mek.iif.hu/porta/szint/human/szepirod/magyar/krudy/szindbad/html/01.htm

Krúdy Gyula: Rákóczi harangja ... Ebédre mindenütt van pisztráng, lazac, fácán, osztriga, pávapástétom, ...
www.mek.iif.hu/porta/szint/human/szepirod/magyar/krudy/rakoczi/html/rakoczi2.htm

Halételek ... Fogas á la Krúdy. ” ... Pisztráng nyárson sütve. 4 pisztráng (kb. ...
www.nepmuvesz.hu/vhaz/receptek/haletelek.htm

Krúdy Gyula: Ifjú évek ... kavarta, nagy kövek bukkantak fel a távolban a víz mélyén, ...

Tág kérdés

Searched the web for *football*.

Results 1 - 10 of about 19,000,000. Search took 0.08 seconds.

NFL.com ... Raiders: Gannon in fourth straight Pro Bowl....

Federation Internationale de Football Association
www.fifa.com/ - 49k

Football.com If it's football, we got it! NFL, NCAA, CFL, AFL, NFLe, XFL, Prep,...
www.football.com/ -

Football news, fun, results, fixtures and features from ... com has the all the latest Premiership, Champions League and Uefa Cup news...
www.football365.com/

Welcome to the FA Home. Sven: I'll pick Davis and Jenas. www.thefa.com/

Fontos információk óriási találati halmazokból

Keresés a világhálón: igen nagy számú, az adott kérdéshez illeszkedő találat.

Melyek a fontosak közülük?

Lehet-e a fontosságot objektíve meghatározni?

Első válasz: NEM! Pl. *csavar* különböző jelentései...

DE MÉGIS... *PageRank* a korai Google módszere (Sergey Brin– LarryPage '98).

Alapötlet: Próbáljuk meg a lapok közötti kapcsolatokat (linkeket) felhasználni erre a célra. A kapcsolatok maguk jelentős tudást, emberi értékítéletet rögzítenek.

Az ötlet kivitelezése

Adott N lap a Hálóról, (pl. a KÖMAL szót tartalmazók). Szeretnénk hozzájuk nemnegatív valós fontossági értékeket rendelni:

$$u = (u_1, u_2, \dots, u_{N-1}, u_N),$$

a lapok rangvektora.

A k . lap fontosabb, mint az l . lap, ha $u_k > u_l$.

Elvárások: ha valakire sokan mutatnak, akkor ő fontos.

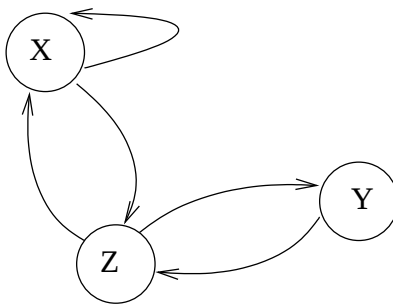
Ha valaki fontos, akkor az általa mutatott lapok is azok.

Egy iteratív megoldás:

1. Kezdetben minden lap fontossága (egy)enlő.
2. Az iteráció egy menetében minden lap egyenletesen szétosztja fontosságát a szomszédai között (és ugyanígy kap másoktól).

Példa:

Legyen $N = 3$ és a laphalmaz a kapcsolataival



Az iteráció első lépései:

$$u_1 = \begin{pmatrix} 1 & 1 & 5/4 & 9/8 \end{pmatrix}$$

$$u_2 = \begin{pmatrix} 1 & 1/2 & 3/4 & 1/2 \end{pmatrix}$$

$$u_3 = \begin{pmatrix} 1 & 3/2 & 1 & 11/8 \end{pmatrix}$$

A vektorok az $u = (6/5, 3/5, 6/5)$ vektorhoz tartanak.

Szemléletes indoklás

SZTOCHASZTIKUS SZÖRFÖLŐ:

Egyenlő eséllyel választ kezdőlapot az N lap közül.

Később pedig egyenlő eséllyel választ a tartózkodási helyéről kimenő élek közül.

Megmutatható, hogy ekkor u_i -vel arányos annak a $\approx$ valószínűsége, hogy m lépés után i -re kerül – feltéve, hogy m nagy.

Mátrixok

Valós számokból álló téglalap alakú táblázatok

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix},$$

ahol n az A oszlopainak, m a sorainak száma,
 $a_{ij} \in \mathbb{R}$.

Az ilyen mátrixok halmaza: $\mathbb{R}(m, n)$.

Értelmezhető az $A, B \in \mathbb{R}(m, n)$ mátrixok
 $C = A + B \in \mathbb{R}(m, n)$ összege:

$$c_{ij} = a_{ij} + b_{ij}.$$

Sor–oszlop-szorzás

Egy m -hosszú sor- és oszlopvektor szorzata egy szám:

$$(a_1, a_2, \dots, a_m) \cdot \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} = a_1 b_1 + a_2 b_2 + \dots + a_m b_m.$$

Például

$$(5, 2, 3) \cdot \begin{pmatrix} 1 \\ 4 \\ 5 \end{pmatrix} = 28.$$

Legyen u egy m -hosszú sorvektor, és $A \in \mathbb{R}(m, m)$, az oszlopai $v^{(1)}, v^{(2)}, \dots, v^{(m)}$. Ekkor az $u \cdot A$ egy m -hosszú sorvektor lesz:

$$u \cdot A = (u \cdot v^{(1)}, u \cdot v^{(2)}, \dots, u \cdot v^{(m)}).$$

Példák: $m = 3$ és

$$(1, 1, 1) \cdot \begin{pmatrix} 1/2 & 0 & 1/2 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{pmatrix} = (1, 1/2, 3/2).$$

$$(1, 1/2, 3/2) \cdot \begin{pmatrix} 1/2 & 0 & 1/2 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{pmatrix} = (5/4, 3/4, 1).$$

A számok éppen az iterációnál kapott értékek...

Az iteráció mátrixokkal

Tekintsük ismét az N netlapot. Az N -szer N -es A mátrix $a_{i,j}$ eleme $1/n$ ha az i . lap $n > 0$ másikra mutat, és van mutatója a j . lapra, $a_{i,j} = 0$ különben.

Tegyük fel, hogy A -nak nincs csupa 0 sora. Ekkor $\sum_{j=1}^N a_{i,j} = 1, i = 1, \dots, N$.

Tétel: Legyen $u^{(0)} = (1, \dots, 1)$ és $u^{(k)} = u^{(k-1)} \cdot A$. Ekkor

$$u = \lim_{k \rightarrow \infty} u^{(k)}$$

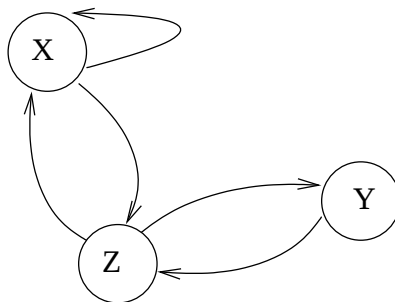
létezik, $u \geq 0$ és $u \cdot A = u$.

Az N lap **rangvektora** u . Az i . lap rangja u_i , ahol

$$u = (u_1, u_2, \dots, u_N).$$

FONTOS ALGORITMIKUS TÉNY: u közelítése,
 $u^{(k)} = u^{(k-1)} \cdot A$ gyorsan számítható.

Példa:



$$A = \begin{pmatrix} 1/2 & 0 & 1/2 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{pmatrix}.$$

Ekkor

$$u^{(0)} = (1, 1, 1),$$

$$u^{(1)} = (1, 1/2, 3/2),$$

$$u^{(2)} = (5/4, 3/4, 1),$$

végezetül

$$u = (6/5, 3/5, 6/5).$$

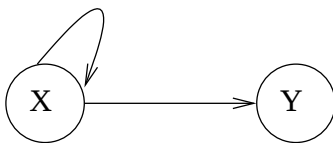
KOMOLY GONDOK: *zsákutca és pókháló.*

Zsákutca

Olyan oldal, amely nem mutat más lapra.

Ekkor A egy sora csupa 0, és a fontosság elillan a rendszerből.

Példa:



A rendszer mátrixa: $A = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 0 \end{pmatrix}$.

Ekkor $u^{(k)} = (\frac{1}{2^k}, \frac{1}{2^k})$, ugyanis

$$\left(\frac{1}{2^{k-1}}, \frac{1}{2^{k-1}} \right) \cdot \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 0 \end{pmatrix} = \left(\frac{1}{2^k}, \frac{1}{2^k} \right).$$

Ezért $u^{(k)} \longrightarrow (0, 0)$, ha $k \longrightarrow \infty$.

Pókháló

Lapok olyan rendszere, amelyben minden él csak ezen rendszerbeli lapokra mutat.

Az iteráció során begyűjti (esetleg) a fontosságot.

Visszaélések alapja lehet!

Példa:

A korábbi elrendezést módosítjuk: y a z helyett csak magára mutat. A mátrix:

$$A = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}.$$

A vektorok az első három iterációs lépés után:

$$\begin{aligned} u^{(0)} &= (1, 1, 1), & u^{(1)} &= (1, \frac{3}{2}, \frac{1}{2}), \\ u^{(2)} &= (\frac{3}{4}, \frac{7}{4}, \frac{1}{2}), & u^{(3)} &= (\frac{5}{8}, 2, \frac{3}{8}). \end{aligned}$$

Belátható, hogy a (végső) rangvektor az $u = (0, 3, 0)$ lesz.

Javítás: a **PageRank**

A lapok adóként befizetik fontosságuk ϵ -szorosát ($1 > \epsilon > 0$), ebből az iterációs lépésben egyenletesen adunk mindenkinek. Az iterációt A helyett az alábbi B -vel végezzük:

$$B = \epsilon \cdot U + (1 - \epsilon) \cdot A, \text{ ahol } U = \begin{pmatrix} \frac{1}{N} & \frac{1}{N} \\ \dots & \dots \\ \frac{1}{N} & \frac{1}{N} \end{pmatrix},$$

azaz

$$b_{ij} = \frac{\epsilon}{N} + (1 - \epsilon)a_{ij}.$$

A B egy kedvező tulajdonságú bolyongást definiál.

Pókhálók esetén is értelmes eredményt ad, pl. a végső u vektor pozitív lesz.

A Brin–Page-ajánlás szerint $\epsilon = 0,15 - 0,20$. Az $u^{(k)}$ vektor itt is sebesen számítható.

Intuitív indoklás

SZESZÉLYES SZTOCHASZTIKUS SZÖRFÖLŐ:

Úgy működik, mint a sztochasztikus szörfölő, DE: minden iterációs lépésnél ϵ valószínűséggel rájön a szeszély. Ekkor egyenletesen választ ($1/N$ valószínűséggel) az összes lap közül.

A szeszélyesség kiemeli a rossz kerékvágásból.

A mai PageRank

Keveset tudunk róla (részben titkos).

Komoly erőfeszítések a manipuláció elkerülésére.

A hálózati csomópontoknak is van (előreszámolt) rangjuk. Ez alakítja az ottani lapok rangját.

Vektorteres indexelés

Szövegek kereshető tárolását megvalósító rendszerek. *Szöveg*=szavak halmaza.

Legyen M a szavak száma.

Egy szöveg (szóhalmaz, dokumentum, keresőkérdés) $\mathbb{R}^M$ egy pontja. A d dok. v_d vektorának a t koordinátája a d szöveg és a t szó viszonyát kifejező $w_{d,t} \in \mathbb{R}^+$ szám (súly).

$$v_d = (\dots, w_{d,t}, \dots).$$

Szokásos $w_{d,t}$ súlyok:

1. 1 ha $t \in d$, 0 különben.
2. $f_{d,t}$ a t -k száma d -ben.
3. $\log_2(1 + f_{d,t})$.
4. Általában $glob_t \cdot lok_{d,t}$.

Az N szöveget leíró $A \in \mathbb{R}(N, M)$ sorai a v_d vektorok.

A geometria gráciája: árnyaltabb közelség, távolság, hasonlóság; fűrtök.

Az eredmények *minősége* szempontjából jobb a talált–nem talált-felfogásnál.

A hasonlóság mérése: *koszinusz-mérték*

A d és q dokumentumok hasonlóságának mérőszáma a $v = v_d$ és a $w = w_q$ vektorok szögének koszinusza:

$$\frac{v_1 w_1 + v_2 w_2 + \cdots + v_M w_M}{\sqrt{v_1^2 + v_2^2 + \cdots + v_M^2} \sqrt{w_1^2 + w_2^2 + \cdots + w_M^2}}.$$

A mérték $\in [-1, 1]$, és két szöveg akkor hasonló, ha a vektoruk közel egy irányba mutat.

A mérték ésszerűen, bár nem túl gyorsan megkapható.

Segít, hogy v, w többnyire ritka vektorok, és a nevező előre számítható.

Gond: szinonimák és asszociatív kapcsolatok kezelése.

Pl. a *zsivány* szóval kérdezünk. Megeshet, hogy a *betyár*, *útonálló*, *haramia*, *Sobri*, *rabló*, *lator*, *himpellér*, *Zorro* tartalmú *d*-k elsikkadnak, mert nincs bennük *zsivány*.

Új idea: Dumais, Deerwester, Berry, Landauer és munkatársaik a 90-es évek elején: mögöttes szemantikájú indexelés (LSI).

Csípős tézisük: A keresés szempontjából *zajnak*, *bizonytalanságnak* tekinthető az a sokféleség, amit a nyelv lehetővé tesz ugyanannak a tartalomnak a kifejezésére.

Így a v_d sorvektorokból álló (a szövegek és szavak viszonyát leíró) A mátrixban sokkal több szabadsági fok van, mint amennyit a tárolt szövegek jelentése indokolna.

Itt a *szabadsági fok* az A mátrix *rangja*.

Mi egy mátrix rangja?

Az A mátrix rangja a legkisebb k egész, amelyre igaz, hogy az A minden oszlopa kifejezhető k oszloppal.

A *kifejezés* az adott vektorokból nyújtással/zsugorítással, összeadással, kivonással kapható.

Példa:

$$A = \begin{pmatrix} 1 & 0 & 2 & 10 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 3 \\ 0 & 0 & 1 & 3 \end{pmatrix}$$

rangja 3. Az első három oszloppal a negyedik kifejezhető:

$$4v^{(1)} - v^{(2)} + 3v^{(3)} = v^{(4)}.$$

Pl. $N = 70\,000$ szöveg és $M = 90\,000$ alapszó esetén A rangja közel maximális (azaz $70\,000$). Az „értelmes” szabadsági fokok száma kb. 200-300-ra tehető.

RECEPT: *A helyett vegyük a hozzá legközelebbi legfeljebb 300-rangú A_{300} mátrixot.*

A szövegeket (és a kérdéseket is) 300 fontos mesterséges fogalomból keverjük ki.

A rang-redukció csökkenti a szemantikus zajt, kiemeli a fák közül az erdőt.

Pl. a szinonimák hasonlóak a redukált térben akkor is, ha nem, vagy csak ritkán fordulnak elő egyazon szövegben.

Jelentős a közös rész a szövegkörnyezetükben.

Lehet, hogy az emberi emlékezet működésében is vannak hasonló egyszerűsítő-kiemelő folyamatok.

LSI-alkalmazások

Indexelés, keresés, másnyelvű keresés

A Landauer–Dumais-kísérlet

Feldogozták (beindexelték) a *Grolier's*-t.

Kb. 30 000 szócikk, kb. 61 000-elemű szótár felett. Az A mátrixnak ekkor nagyjából $N = 30\,000$ sora és $M = 61\,000$ oszlopa van.

Képezték A_{300} -at és ennek vizsgálták a szinonimafelismerő képességét.

A TOEFL 80-kérdéses szinonima-tesztjeit használták.

Pl. *imposed, believed, requested, correlated* közül kell kiválasztani, hogy melyik jelentése hasonlít leginkább a megadott ötödik szóéra (*levied*).

A 300-dimenziós LSI-térben kiszámították az ötödik szó hasonlóságát (koszinusz-mérték) az első négyhez.

A példánál kapott értékek:

imposed: 0,70, *believed*: 0,09, *requested*: 0,05,
correlated: $-0,03$.

A (helyes) válasz (*imposed*).

Összességében 64%-ban kaptak helyes választ.

*Ennyi a diákok átlagos eredményessége a hivatalos
TOEFL-vizsgákon!*

Befejezésül

Gazdag vadászmezők várják az alkalmazott matematika és a számítástudomány híveit.

NÉHÁNY ÉRDEKES CÍM:

PageRank és LSI:

<http://www.cs.bme.hu/~bodon/magyar/index.htm>

<http://lsi.research.telcordia.com/>

Magyar nyelvű kereső a SZTAKIban:

<http://keres.sztaki.hu>

<http://www.ilab.sztaki.hu/websearch/index.hu.html>

A BME Matematikai Intézete:

<http://www.math.bme.hu>