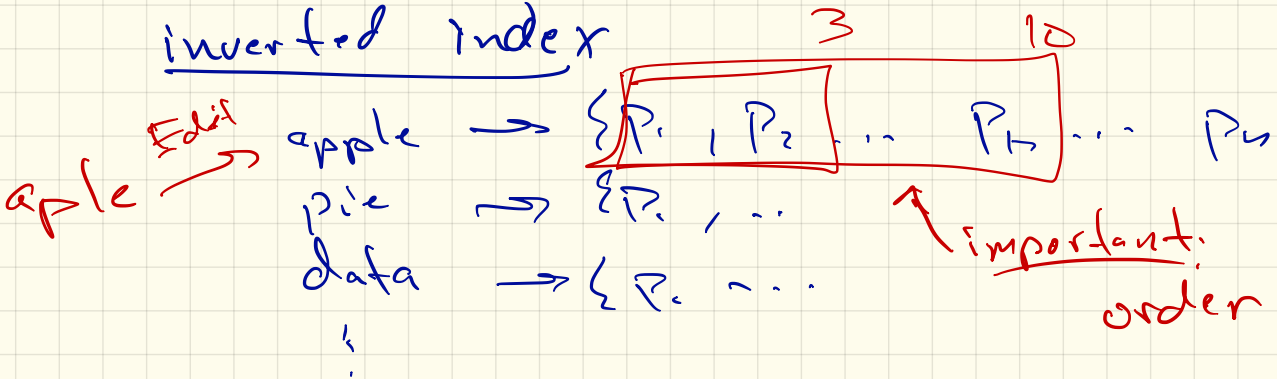
PageRank

and Search Engines

How does a search engine work?

inverted index



Webpage Similarity

pre 1998
pre-PageRank

Altavista, Lycos, Infoseek

webpage → html text → bag-of-words

$p_i \to v_i \in \mathbb{R}^{10,000}$ $\quad v_i = (0, 0, \underset{\text{apple}}{8}, 0, 0 \ldots 0, \underset{\text{pie}}{5}, 0$

search "apple pie" ↓

$q \in \mathbb{R}^{10,000}$ $\quad q_i = (0, \ldots, \frac{1}{\sqrt{2}}, 0 \ldots, \frac{1}{\sqrt{2}}, 0 \ldots)$

highest cosine-sim(v_i, q) → top of list.

... but? "apple pie, apple pie, ..." white

battle: search engine vs. spammers

modify dist

ave(cos, Jaccard)

emphasize certain words

cap word count.

query: (ooo expand)

modify

copy top pages into bottom of your page

Index

Yahoo! / Look Smart

business model : paid placement

1. Google
2. Youtube
3. Facebook
4. Baidu (China)
5. Wiki
6. (18) Tencent QQ
7. Taobao
8. Tmall
9. (6) : Yahoo!
10. (11) Amazon
11. (7) Twitter

15 (10) Instagram

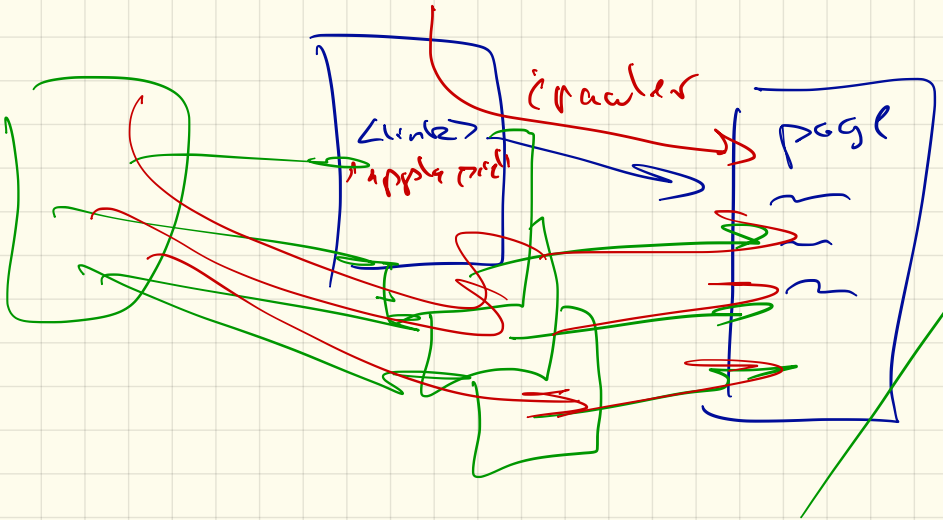
23 (17) Netflix

Crawlers

automated bot

↳ moves around the web

→ follow links



PageRank

Idea 1: important webpages link to important webpages

Idea 2: Important webpages are visited often by a random surfer

Markov Chain

"crawler"

Model web as graph $G = (V, E)$

Vertices = pages
Edges = hyperlinks

Web graph $G = (V, E)$ → Adjacency matrix A

Probability Transition Matrix P

$r: v_i = eig(P, 1) \equiv$ Pagerank vector

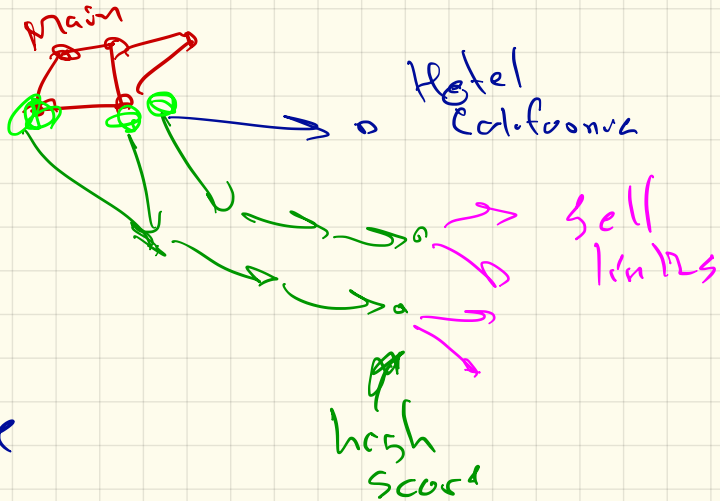
$r(page)$: larger better

$score(page, query) = magic($ $r(page)$, $\cos(v_p, q)$, hyperlinks $)$

pagerank

Is the webgraph ergodic?

- cycles? No
- connected? No (but ok)
- transient?



Fix: Teleportation

$\beta = 15\%$ jump to random page

Compute $q^* = P^* q_0$

option 1 eigs (P) huge

option 2 compute P^n large n

for any q_0: $q_x = P^n q_0$

"small world" P^7 very dense

option 3 for $i = 1$ to $(n = 50)$

$q_{i+1} = P q_i$

return $q_x = q_n$

random jump

$q_{i+1} = ((1-\beta)P + \beta Q) q_i$

$Q = \left(\frac{1}{m}\right)_{n \times m}$

$= (1-\beta) P q_i + \beta \left(\frac{1}{m}\right)$

2015: "truth"

Trust Rank

Trust more: Wikipedia, .edu domain

Run 2 versions of PageRank.

- Regular: teleportation is uniform $r(p)$
- Trust: teleportation is more likely to jump to trusted pages. $t(p)$

$$S(p) = \frac{r(p) - t(p)}{r(p)}$$: larger → more likely spam