

Text Technologies

Web Search 2
Link Analysis and Spam

Victor Lavrenko
Some Figures © Addison Wesley, 2008

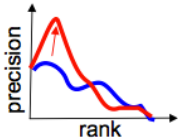
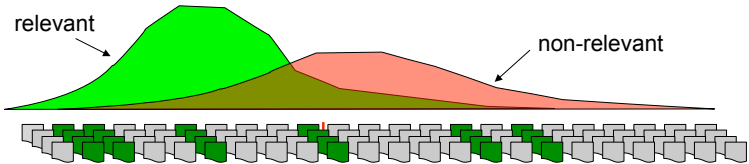
Copyright © 2011, Victor Lavrenko

- Massive amounts of data
- Web query types
- Link analysis
 - PageRank
 - Hubs and Authorities (HITS)
 - Anchor Text
 - Link spam

Copyright © 2011, Victor Lavrenko

Amount of data

- Staggering amount of data
 - Google MapReduce: 20PB/day (2008)
 - challenging... but surprisingly makes some things easier
- Collection ... a (random) subset of what's out there
 - suppose we gathered N documents, precision@10 = 40%
 - what if we gathered 4*N documents? precision@10 = 60%
 - true whenever relevant documents more dense at top ranks
 - same Miss / False Alarm rate, higher average precision



Web queries

- Query types:
 - informational
 - find information on some topic (one or more pages)
 - navigational
 - find particular page seen before, or assumed to exist
 - transactional
 - find a page to perform a task (shopping, downloading)
- Queries are short, not natural-language

What cases have discussed the concept of excusable delay in the application of statutes of limitations or the doctrine of laches involving actions in admiralty or under the Jones Act?

vs. sex

Copyright © Victor Lavrenko, 2011

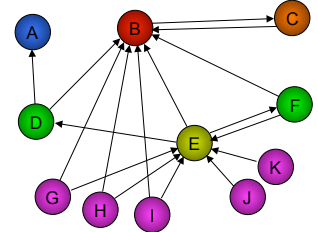
Links between pages

- Google's description of PageRank:
 - relies on the “uniquely democratic” nature of the web
 - interprets a link from page A to page B as “a vote”
- $A \rightarrow B$ means A thinks B is worth something
 - “wisdom of the crowds”: many links mean B must be good
 - content-independent measure of quality of B
- Use as a ranking feature, combined with content
 - but not all pages that link to A are of equal importance
 - a single link from Slashdot or CNN may be worth thousands
- Google PageRank [Brin & Page, 1998]
 - how many “good” pages link to A

Copyright © 2011, Victor Lavrenko

PageRank: random surfer

- Analogy used to derive algorithm:
 - user starts browsing on a random page
 - picks a random out-going link, goes there
 - repeat forever
 - example: $G \rightarrow E \rightarrow F \rightarrow E \rightarrow D \rightarrow B \rightarrow C$
 - with probability $1-\lambda$ goes to a random page
 - otherwise what happens if we go to A?
- PageRank of page x
 - probability of being on page x at a random moment in time
 - formally: eigenvector R that satisfies $A R = c R$
 - A ... adjacency matrix: $A_{y,x} = 1/N_y$ if $y \rightarrow x$, $1/N$ otherwise



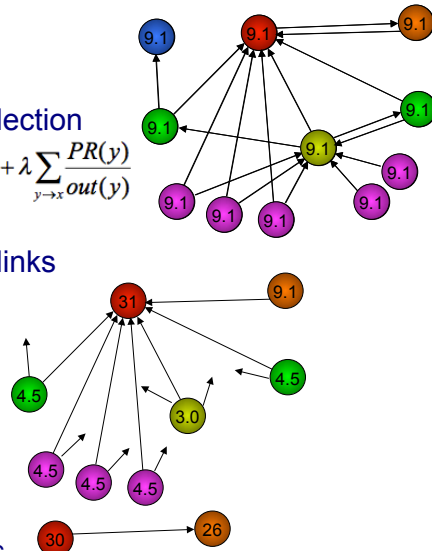
Copyright © 2011, Victor Lavrenko

Computing PageRank

- Initialize $PR(X) = 100\%/N$
 - total number of pages in our collection
- For every page X: $PR(x) = \frac{1-\lambda}{N} + \lambda \sum_{y \rightarrow x} \frac{PR(y)}{out(y)}$
 - y contributes part of its PR to x
 - spreads PR equally among out-links
 - PR scores should sum to 100%
 - use two arrays: $PR_{(t)} \rightarrow PR_{(t+1)}$

Example:

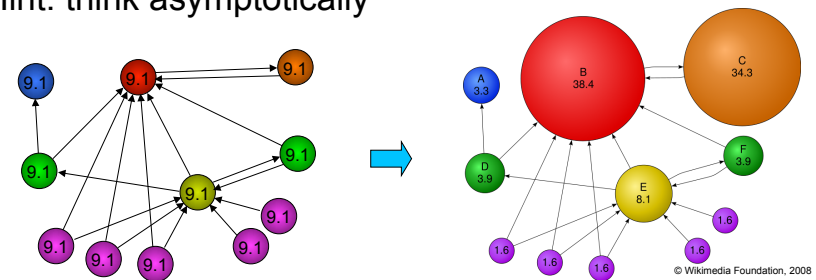
- $PR(B) = 0.18 * 9.1 + 0.82 * [PR(C) + \frac{1}{3} PR(E) + \frac{1}{2} PR(D) + \frac{1}{2} PR(F) + \frac{1}{2} PR(G) + \frac{1}{2} PR(H) + \frac{1}{2} PR(I)] \approx 31$
- $PR(C) = 0.18 * 9.1 + 0.82 * PR(B) \approx 26$



Copyright © 2011, Victor Lavrenko

PageRank example: result

- Algorithm converges (few iterations sufficient)
- Observations:
 - pages with no inlinks: $PR = (1-\lambda) * 1/N = 18\% / 11 = 1.6\%$
 - same inlinks mean same PR
 - one inlink from high PR >> many from low PR
- Hint: think asymptotically

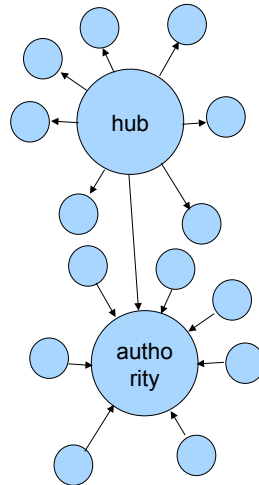


Copyright © 2011, Victor Lavrenko

© Wikimedia Foundation, 2008

Hubs and Authorities

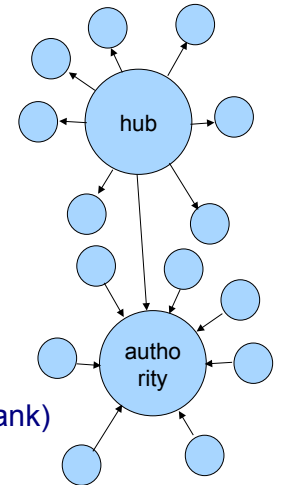
- PageRank: simplistic view of a network
- Network topology: different node types:
 - “hub”: page that points to a lot of others
 - e.g. Yahoo / DMOZ directory
 - “authority”: page that many others refer to
 - authoritative view on some subject
- HITS algorithm [Kleinberg, 1997]
 - Hyperspace Induced Topic Search
 - automatically determine hubs/authorities
 - PageRank is “half” of HITS



Copyright © 2011, Victor Lavrenko

HITS algorithm

- $H(x)$, $A(x)$... hub, authority scores
 - initialize as $N^{-1/2}$, N ... number of pages
 - a good hub links to many good authorities $H(x) = \sum_{y \leftarrow x} A(y)$
 - a good authority is referenced by many hubs $A(x) = \sum_{y \rightarrow x} H(y)$
 - normalize H,A: $\sum_x H(x)^2 = \sum_x A(x)^2 = 1$
- In practice
 - used on result set (not all docs like PageRank)
 - developed for IBM Clever project
 - variant used by Teoma (now Ask.com)



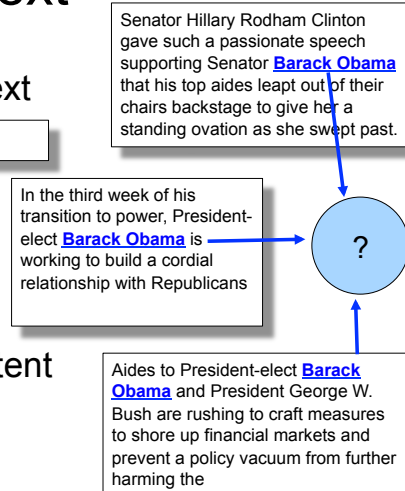
Copyright © 2011, Victor Lavrenko

Anchor text

- HTML links contain anchor text


```
<a href="www.cnn.com">CNN news</a>
```

 - description of destination page
 - short, descriptive, like a query
 - re-formulated in different ways
 - human “query expansion”
- Used in addition to page content
 - together with URL tokens
 - also: surrounding text
 - separate “weights” for every component
- Significantly more effective than PageRank

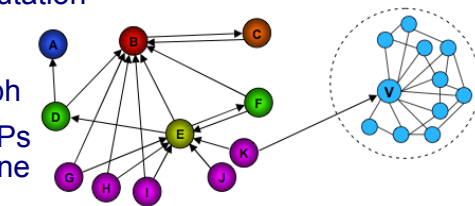
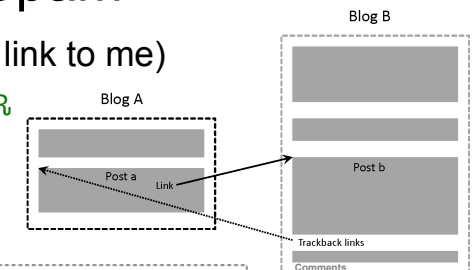


Copyright © 2011, Victor Lavrenko

Link Spam

- Trackback links (blogs that link to me)
 - based on `$HTTP_REFERER`
 - artificial feedback loops
- Links from comments on sites with high PR


```
Excellent post! And there's more evidence to support your argument <a href="medz.com/viagra">here</a>
```
- One solution: insert `rel=nofollow` into links
 - link ignored during PR computation
- Link farms
 - fake densely-connected graph
 - hundreds of web domains / IPs can be hosted on one machine



Copyright © 2011, Victor Lavrenko

Summary

- Massive amounts of data
 - challenging for efficiency, but improves effectiveness
- PageRank
 - probability that random surfer is currently on page x
- Hubs and Authorities (HITS)
 - asymmetric, recursive: good hubs → good authorities
- Anchor Text
 - short, concise description of content on the target page
- Link Spam
 - [trackback links](#), [link farms](#)