Matrix and random walk methods in web analysis

Ravi Kumar

Yahoo! Research

ravikumar@yahoo-inc.com

Outline

- Background
- Matrix methods in web information retrieval
  - HITS/Clever
  - PageRank
  - Some applications
- Random walk methods in some web analysis
  - Measurements
  - Communities
  - Proximity
  - Trust

Caveat

- Not meant to an exhaustive survey of known results/applications

Web information retrieval

Web information retrieval

How to find information on the web?

~ 10^{10} web pages
~ 10^{11} hyperlinks
Constantly changing

Classic information retrieval (IR)

Input: Set of documents

Goal: Given a query, find documents that are most relevant to the query

Method:

- Preprocess the documents to build an index
- Search at run-time

Vector space model

- Documents and queries are vectors of terms
- i-th entry = a function of the i-th term occurrence in the document
- Similarity measure between the query and the document
- Order documents based on similarity

What is different about the web?

- Volume (billions)
- Change (23% per day, dynamic content)
- Decay (short half-life)
- Heterogeneity (HTML, AJAX, pdf, images)
- Language variations
- Duplication (exact copying, near-duplication)
- Variable quality
- Spam
- Links (malicious links, 404s, redirects, dead-ends)
- No easy way to evaluate performance

User expectation

- Poor queries (short, imprecise, badly formed, low effort)
- Focus only on top few results
- High quality results matching their inadequately expressed intent
- Instant response

Web information retrieval

Input: Web pages

Goal: Given query, output web pages in order of relevance to the query

At our disposal

Structure: web pages, links, enthusiastic users

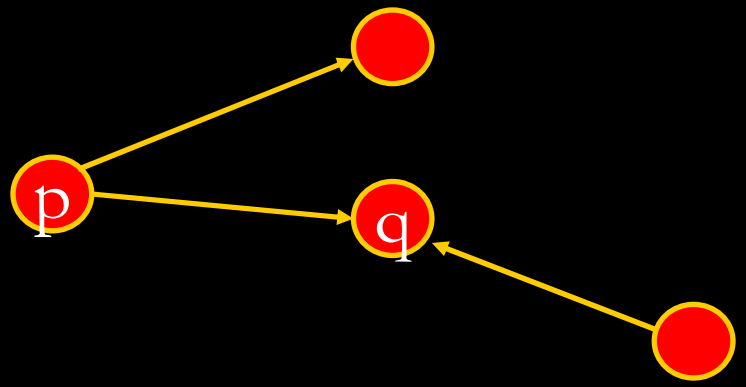
Use: personalization, interaction, feedback

Approaches to web IR

- Text-based ranking
  - Classical IR
  - Query-dependent
- Link-based ranking
  - HITS: query-dependent
  - PageRank: query-independent

Web as a directed graph

- Nodes = (static) web pages
- Directed edges = edge from p to q if page p has a hyperlink to page q



Link-based IR principles

Analyze user behavior and thought process

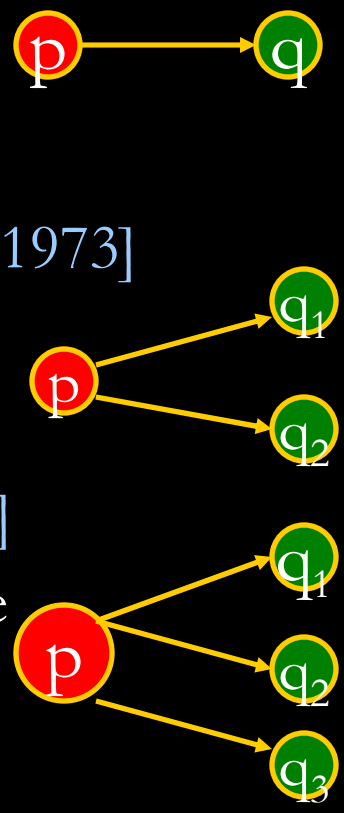
- How do people browse?
  - Random link from the current page
  - Abandon current activity and restart at a random page
- Why do people create links?
  - Confer authority and endorsement
  - Latent indication of trust

Assumptions

- People express judgments by both content and hyperlinks
- More high quality hyperlinks, the better a page

Link-based IR principles …

- Relevant linkage principle
  - p links to q ⇒ q is relevant to p
- Topical unity principle [Kessler 1963, Small 1973]
  - q_1 and q_2 are co-cited in p
    ⇒ q_1 and q_2 are related to each other
- Lexical affinity principle [Maarek et al. 1991]
  - Closer the links to q_1 and q_2 are, stronger the relation between them



Link-based IR principles …

- Page p pointing to page q = endorsement of page q by page p
- Quality of p = # endorsements it gets
- Quality of p depends on
  - # endorsements of p
  - Quality of pages pointing to p
- A recursive definition!

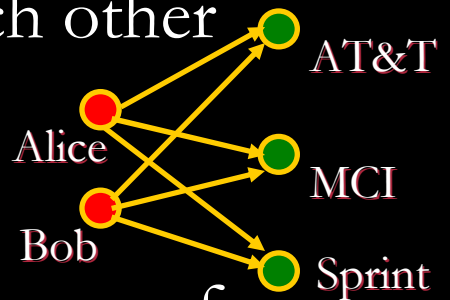
Early link-based IR

- Citation analysis of scholarly publications (Bibliometrics)
- Impact factor [Garfield 1972]
  - Rank by indegree
  - Not all citations are born equal
- Influence weights [Pinksi Narin 1976]
  - $A(i,j)$ = fraction of citations that go from i to j
  - $w(j) = \sum_i A(i,j)\, w(i)$

  $w = A^T w$, the eigenvector of A associated with eigenvalue 1

HITS [Kleinberg, 1998]

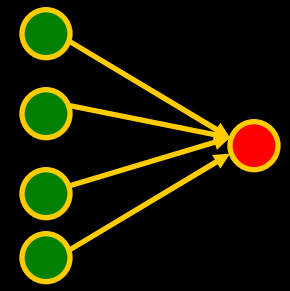
- Page p pointing to page q means p endorses q
- But, two popular pages may not cite each other



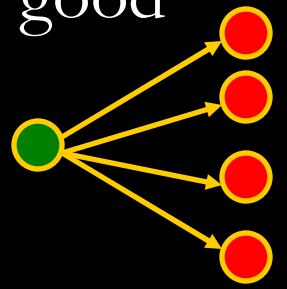
Two-layer model: Hubs and authorities

- Hubs: Pages with pointes to lots of resources for a topic
- Authorities: Representative sources for a topic
- Identify the best hubs and authorities for a given topic

HITS …



- A page is
  - An authority if lots of pages point to it
  - A good authority if lots of pages that are good hubs point to it



- A page is
  - A hub if it points to lots of pages
  - A good hub if it points to lots of pages that are good authorities

A mutually reinforcing and recursive definition!

HITS …

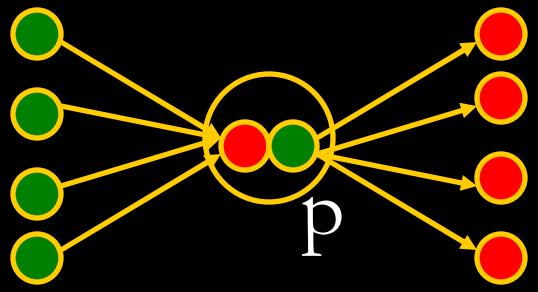
- Each page p has h[p] = a[p] = 1 initially
- Compute hub scores using authority scores

$$h[p] = \sum_{p \to q} a[q]$$

- Compute authority scores using hub scores

$$a[p] = \sum_{q \to p} h[q]$$

- Renormalize scores and repeat



- Output top few hubs and authorities

HITS …

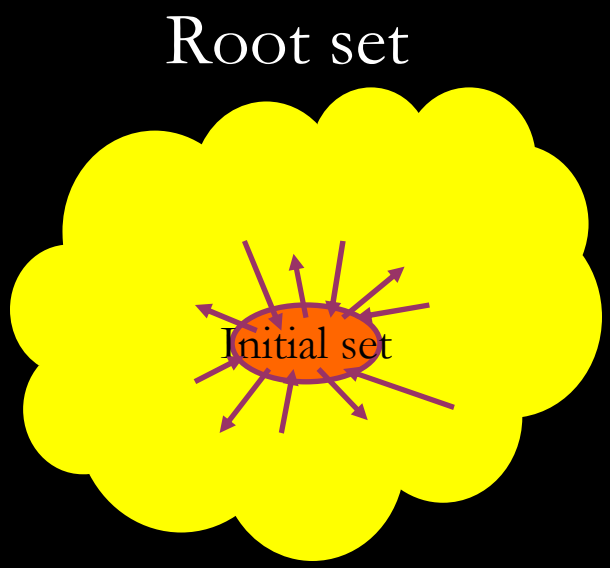
- $a_{i+1}(q) = \sum_{p \to q} h_i(p)$, $h_{i+1}(q) = \sum_{q \to p} a_i(p)$
- $a_{i+1} = A^T h_i$, $h_{i+1} = A\, a_i$
- $a_{i+1} = (A^T A)\, a_i$, $h_{i+1} = (A\, A^T)\, h_i$

- Iteration converges to a^*, h^*

- a^*, h^* are eigenvectors of $A\, A^T$, $A^T A$
- a^*, h^* are left and right singular vectors of A

Using HITS for web IR

- Apply keyword search to generate initial set of 200 pages
- Expand initial set into root set by following links
- Compute weights for edges
- Perform HITS iterations
- Output top hubs and authorities
- Teoma is based on HITS



Variants: Clever [Chakrabarti et al 1999]

- Edges in the graph have weights
- Weight is a function of
  - Anchor text vs. query
  - +/- prefixes in the query
  - Source/destination of hyperlink
  - Stop sites
  - Useful sites

PageRank [Brin Page 1998]

Random walk interpretation

- Walk starts at a uniformly chosen web page
- At each step, if currently at page p
  - W/p α, go to a uniformly chosen web page

    α is the teleportation constant
  - W/p 1 - α, go to a uniformly chosen outneighbor of p
- PR(p) = fraction of steps random walk spends at p in the limit

PageRank …

- A = adjacency matrix of web graph

$PR(p) = \alpha/n + (1 - \alpha) \sum_{q|\ (q,p) \in A} PR(q)/outdegree(q)$

- B = normalized adjacency matrix of web graph with no sink nodes

$M = \alpha\ U + (1 - \alpha)\ B$

- PageRank = stationary probability for this Markov chain

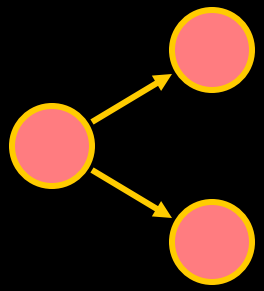
$\alpha = 0.15$

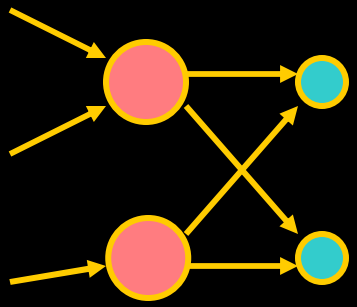
Using PageRank for web IR

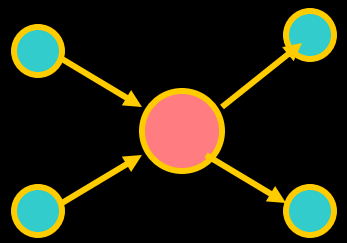
- Query-independent phase
  - Statically rank all pages according to PageRank
- Query-dependent phase
  - Return pages containing the query in order of PageRank
  - More heuristics (title, anchortext, last update, …)
- One of ranking criteria in Google

Axiomatic uniqueness [Altman Tennenholtz 2004]

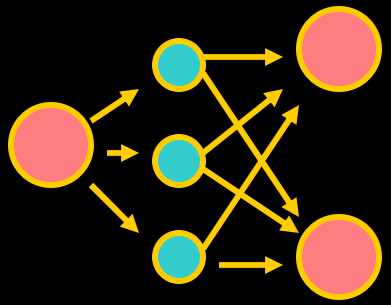
1. Anonymity
2. Adding self-edge to node cannot decrease its rank



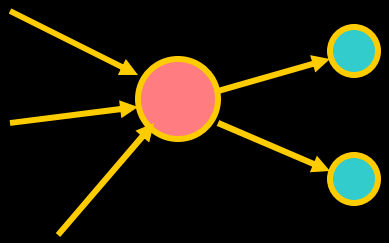




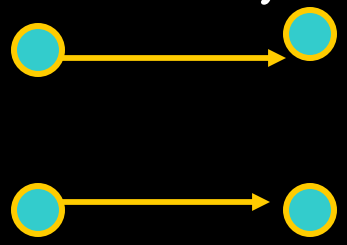
3. Vote by committee



4. Collapsing

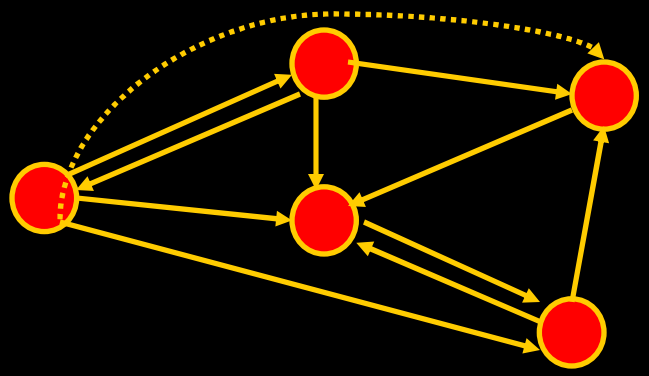


5. Proxy



Monotonicity [Chien Dwork Kumar Simon Sivakumar 2002]

How does adding a new edge affect PageRank?



Adding a new link to page p can only

- Improve the PageRank value of p
- Improve the PageRank ordinal of p

Parametrized teleportation

[Boldi Santini Vigna 2005]

- $\alpha \to 1$
  - More uniform jumps, less informative
- $\alpha \to 0$
  - Closer to the actual web graph
  - Computational obstacles
  - Sink nodes absorb all the rank
- $PR'(\alpha) = PR(\alpha)\ (\ B\ (I - \alpha\ B)^{-1} - (1 - \alpha)^{-1}\ 1)$
- Compute higher-order derivatives during power method
- Use MacLaurin expansion to compute $PR(\alpha)$ for any α

Variant: Personalized PageRank

[Brin Motwani Page Winograd 1998, Haveliwala 2003]

- Captures notion of importance wrt given topic
- Instead of jump to a random page, jump to a page w/p proportional to its relevance to the topic
- p is the preference/personalization vector
- W/p $\alpha \cdot p_v$, jump to v, where

  p_v = relevance of v to the topic
- Can precompute small set of relevant pages and set p_v to be uniform among these pages

PageRank vs. HITS

PageRank

- Query-independent
- Offline computation
- Large graph
- Additional query-time step
- Harder to spam
- More stable

HITS

- Query-dependent
- Per-query computation
- Small graph
- Outputs both hubs and authorities
- Easier to spam
- Quality depends on seed

General computational issues

- Web graph is huge, sparse, changing
- Simple power iteration
  - Popular, few iterations sufficient, parallelizable
- Exploiting block and web graph structure
- Gossip algorithms [Achlioptas McSherry 2004]
- Other numerical methods [Stewart 1999, Langville Meyer 2004]

Special cases of computation

- Personalized computation [Jeh Widom 2003]
  - PageRank is linear in the preference vectors
  - A structural decomposition theorem based on a given "hub" set of interest
- Incremental computation [Chien et al 2001]
  - Locate the changed nodes
  - Expand the seed set
  - Recompute PageRank for expanded set
  - Propagate values to the rest of the graph

Other methods: SALSA
[Lempel Moran 2000]

Given a set of pages

- Out-step (O): Go to a uniform out-link
- In-step (I): Go to a uniform in-link

- Authority scores = fixed point of O-I chain
- Hub scores = fixed point of I-O chain
- If p is in component V_p with E_p links

$$\alpha(p) = |V_p|/|V| \cdot \text{indegree}(p)/E_p$$

Applications: Web page analysis

Spam [Gyongi, Garcia-Molina, Pedersen 2004]

- Personalized PageRank with teleportation to known non-spam pages

Crawling policies [Cho Garcia-Molina 2000]

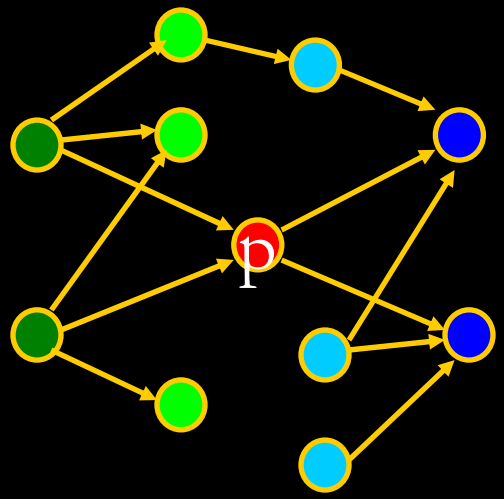
- A variant of PageRank to determine the order and frequency of web page crawling

Application: Finding related pages

Find web pages that are related to a given page

Link-based algorithm [Dean Henzinger 1999]

- Build a neighborhood graph around the input p
- From p, go back, forward, back-forward, forward-back
- Run HITS on this graph
- `Query-less' mode



Applications: Metasearch

For a query, given 100 results from k different search engines, produce a consensus ordering of the results

Markov chains [Dwork Kumar Naor Sivakumar 2001]

- States = web pages in the union of results
- Transitions = function of ordering (probabilistically switch to a better candidate)
- Final ranking = order of stationary probabilities

Web analysis

Size of the web

Web as seen by a search engine

- Size is not a measure of quality, but indicates coverage, comprehensiveness, etc
- Perform a random walk to pick a random page and check if we know about it
  - Directed graph
  - Start state bias
  - Highly reachable web sites are visited more often

Sampling by random walks [Bar-Yossef Gurevich 2006]

- Define a graph G over the indexed web pages
  - $(x,y) \in E$ iff $queries(x) \cap queries(y) \neq \emptyset$
  - $w(x, y) = \sum_{q \in queries(x) \cap queries(y)} 1/pages(q)$
- Perform a random walk on G according to these weights
  - Limit distribution = degree distribution
- Use MCMC methods to make limit distribution uniform
    - Rejection sampling
    - Metropolis--Hastings

Communities [Andersen Lang 2006]

Given a cohesive subset of web pages, enlarge the subset to find their enclosing community

- Useful in HITS
- Finding related pages
- Classifying web pages according to topics
- Identifying spam pages
- Powerful primitive in general

Expanding seed set

- Truncated random walk
  - Perform usual random walk from the seed set
  - Remove probabilities if they are below threshold
- Assuming the seed set is in a community
  - The community will contain much of the probability mass
  - Threshold prevents leak outside the community
  - Local computation, very efficient
  - Truncated random walk will find the community whp

Proximity to a subset

Given a small arbitrary subset S of web pages, determine how close is a given page to this subset S

Candidate notions of proximity

- Shortest path distance
- Flow
- Electrical resistance (random walks)

Application of promixity: Decay
[Bar-Yossef Broder Kumar Tomkins 2003]

- Web changes every day
  - Average half-life of a page is quite short
  - Web littered with lots of broken/dead links
  - Changes are unpredictable
- How do we know if a page is not up-to-date?
  - Last modified date (not reliable)
  - Topics discussed on the page are out-dated
  - Dead links! (easy to detect)

Decay process

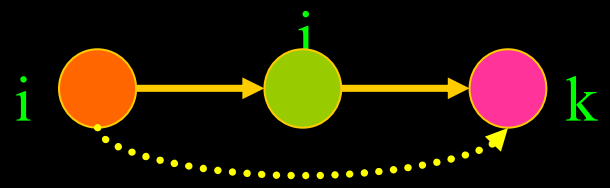
- A random surfer process with absorption
- Absorption can be into a live page or into a dead page
- If the current page p is dead the process terminates with a dead absorption
- If the current page is alive
  - With probability α the process terminates with a live absorption.
  - With probability $1-\alpha$ the surfer moves to a link out of p chosen u.a.r.
- decay(p) is the probability that a random surfer starting at p is absorbed into a dead state
- Captures the idea that we look at how many links are dead in a within pages accessible from p, exponentially discounted for distance

Trust and reputation [Guha Kumar Raghavan Tomkins 2004]

- Nodes are people
- Each directed edge (u,v) has a real-valued trust(u,v)
  - Positive values mean trust
  - Negative values connote distrust
- We are given trust(u,v) only for some edges (u, v)
- Can we infer missing values?

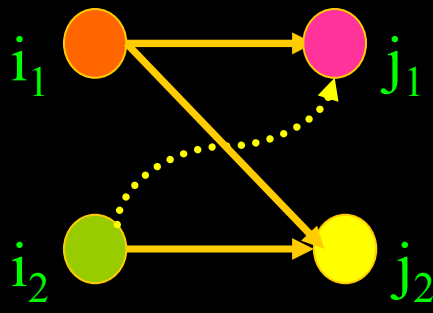
Atomic propagation operators

Direct propagation



- b_{ij} = how much i trusts j
- b_{ik} is a function of how much i trusts j and how much j trusts k
- $b_{ik} = \sum_j b_{ij} \cdot b_{jk}$
- $b \mapsto b \cdot (b)$

Co-citation



- Propagate i_2's trust of j_2 `backward' to infer i_2's trust of i_1
- Propagate i_2's trust of i_1 forward by using i_1's trust of j_1
- Infer about i_2's trust of j_1
- $b \mapsto b \cdot (b^T b)$

Thank you!

ravikumar@yahoo-inc.com