a summary of

“Authoritative Sources in a Hyperlinked Environment”

Jon M. Kleinberg, 1997

•••

by Ram Sai Ganesh

Paper linked here

How do we drink from a firehose?

- Information overload on the internet
- Continuous flow of data – “populist hypermedia”
- Web searching – we need to answer string queries with hyperlinks
- Select/return a few out of millions of possible websites.
- Defining a clear objective function – what is a ‘good’ search result?

Network Structure & Querying

- Hyperlinks – natural directed graph edges
  - A form of endorsement. Could also be navigational (intra-website) or advertising-related.
- Queries are strings; could be either specific (words) or broad (topics)
- Popularity vs relevance
  - Most ‘relevant’ sites might not contain the query string.

“We propose a link-based model”

- Attempts to solve the ‘authority’ issue.
- Divide pages into categories: pages sometimes...
  - ...are authoritative about a topic/broad area – referred to as **authorities.**
  - ...contain links to many other authorities – referred to as **hubs.**
- We wish to organize pages by query-relevance (not cluster by topic).

Constructing the Root Set

1. We form an initial collection of pages ‘S(q)’, the ‘root set’.

- Ideally, small, relevant and already containing authorities
- Text-based search, |S(q)| ~ 200
- Usually, S(q) is small & relevant, but doesn’t necessarily contain good *authorities*.
- Authority – most likely pointed to from (or points at) a site in S(q).
- Apply heuristics – filter intrinsic links, and adverts to obtain G(q)

Mathematically Finding Authorities

2. We have a small concentrated graph G(q). To extract authorities:

- We can't just rank pages by their in-degree.
- **Key idea:** Authorities don't *just* have a high in-degree, we also expect a large ***overlap between the sets of pages*** that point to them.
- New type of page: a **hub** contains links to multiple relevant authority pages.
- Mutually reinforcing relationship

Associate Two Attributes With Each Page

3. Assign an (x, y) pair of normalized numerical weights for each page:

- p<x> = authority weight, p<y> = hub weight of a page p.
- A page ‘p’ points should have a high ‘y’ if it points to many pages with a high ‘x’ value and vice versa.
- Mutually reinforcing algorithm, p<x> and p<y> start at 1.

An Iterative Algorithm

4. We perform 2 operations to perform on the graph G(q) iteratively:

- ‘I(p)’ is defined as adding all neighboring hub weights to p<x>
- ‘O(p)’ is defined as adding all neighboring authority weights to p<y>

$$x^{\langle p \rangle} \leftarrow \sum_{q:(q,p)\in E} y^{\langle q \rangle} \qquad y^{\langle p \rangle} \leftarrow \sum_{q:(p,q)\in E} x^{\langle q \rangle}$$

- Iteratively apply I and O to nodes in G, ‘k’ times, normalizing p<x> and p<y> at each step. (p1<x>...pn<x> & p1<y>...pn<y> converge to finite limits over ~20 iterations.)
- Report ‘c’ nodes with the largest hub and authority scores.

Promising Results!

- Filtered result almost entirely consists of desirable authority pages.
- Predominantly, these authorities don't occur in the root set.
- We search the *entirety* of the internet only once, to obtain the initial root set; we still get good authorities from the corpus of millions of pages.
- These results were obtained *disregarding* the textual content of the pages.
- WLOG, can be extended to find-similar-page type queries.
    - Simply return best authorities neighboring current page 'p'.

Connections to Similar Works

- WWW Page rankings:
  - Several others have attempted to solve the problem, like Page and Brin.
  - Their model confers authority directly from one to another; HITS recognizes the existence of intermediary hubs.
  - However, theirs can be precomputed and queried, Hubs and Authorities requires an initial text-based search.
- Quantifying influence and impact in social networks –
  - We can extend the ‘authority’ idea to individual people; out edges are endorsements to varying degrees of one person by the other.

Areas for Potential Improvement

- Mentioned by Kleinberg – HITS doesn't account for the 'people-traffic' movement within a page. Do they often find what they want – or are they going down link 'rabbit-holes'?
- HITS makes intrinsic assumptions about website creators' motives.
  - Do people always publish websites with the intention of creating authorities/hubs?
- Kleinberg feels that querying occurs for "facilitating this [natural exploratory] process rather than for replacing it."
  - Have search engines' motives changed over time? Predictive results
- How would HITS handle new authorities/hubs?
- Does filtering misinformation happen organically?
- How can these ideas be extended to Influence and Passivity?