Searching for

Authority on the WWW

(Not just relevance or popularity...)

Ido Rosen

<ido@cs.uchicago.edu>

Sources of Information on the WWW

- Textual content

- Images, sounds, multimedia content

- <u>Hyperlink digraph</u> (network structure)

  - Pages are vertices, links are arcs

    - Refinement: URLs are nodes

Nature of the WWW

- Local organization may be a priori.

- Global organization "utterly unplanned."

- Billions of agents (users, spiders).

- Millions of publishers.

- Trillions of vertices, at least.

- Too big for simple search.

Searching the WWW

- **Quality** of search method defined by utility of results.

- **Utility** requires human evaluation.

- Utility is closely correlated to **relevance**.

- Algorithmic and storage **efficiency** are a concern: interactivity/response time.

Search: Queries

- Searches are initiated by a user-supplied **query**.

- Three types of queries discussed:

  - **Specific** queries.

  - **Broad-topic** queries.

  - **Similar content** queries.

Search: Problems.

- Specific queries: **Scarcity**.

  - Required information is scarce and pages are hard to find.

- Broad-topic queries: **Abundance**.

  - We only want the **authoritative** pages. (i.e.: Wikipedia itself, not ad-clones.)

Search: Authorities

- Possible measures of authority:

  - **Frequency** of search term on page.

    - Problem: **Self-descriptive**.

  - **Popularity** of page. (rank by links in)

    - Problem: Obfuscation by **hubs**.

  - Analysis of link structure...

Hyperlinks

- Claim:
  Hyperlinks indicate **conferred authority**.

- Claim:
  Hyperlinks solve self-descriptive problem.

- What about navigational links?

- What about paid advertisements?

Popularity

- In some cases, most authoritative pages aren't self-descriptive.

- Universally popular pages would be considered highly authoritative w.r.t any query string, when they are not.

Step 1: Constructing Focused Subgraph

- Obtain root set, **R**, from textual search.

    - Relatively small, rich in relevant pages, but doesn't contain most or many of strongest authorities.

    - Extremely few intra-R links.

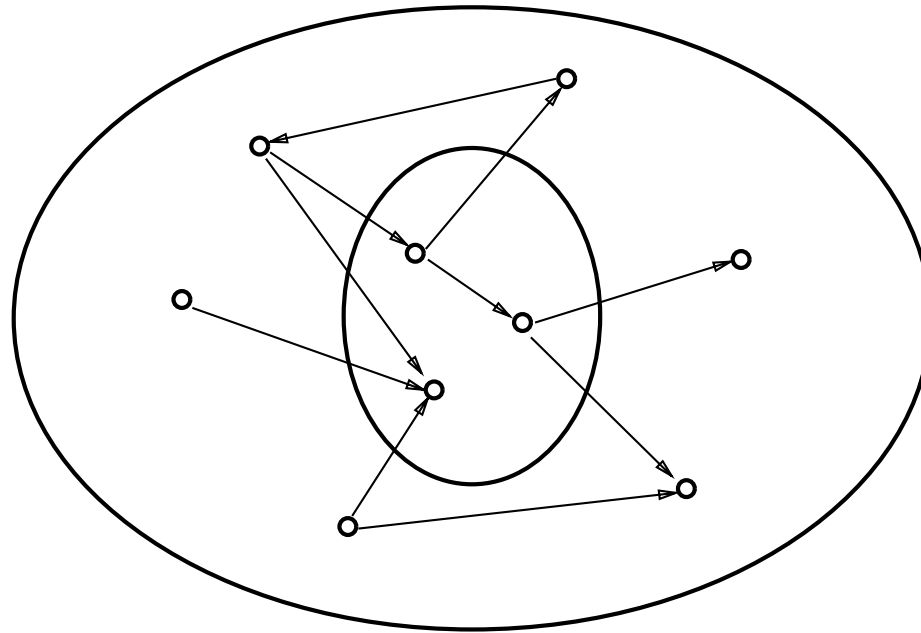- Obtain base set, **S**, from R by adding any pages pointing to or pointed from R.

Figure 1: Expanding the root set into a base set.

$$\mathbb{R} \rightarrow \mathbb{S}$$

- What about navigational links?

    - Transverse vs. intrinsic links.

    - Delete all intrinsic links.

        - Caveats?

- What about "Google Bombing"?

- Set limitations on in-degree or out-degree on a per-domain basis.

Step 2: Computing Hubs and Authorities

- Given our focused subgraph G, now what?

  - Popularity ranking by in-degree?

    - Popularity ≠ relevance.

  - **Hub**: links to multiple relevant authorities.

  - **Authorities**: high in-degree and overlap.

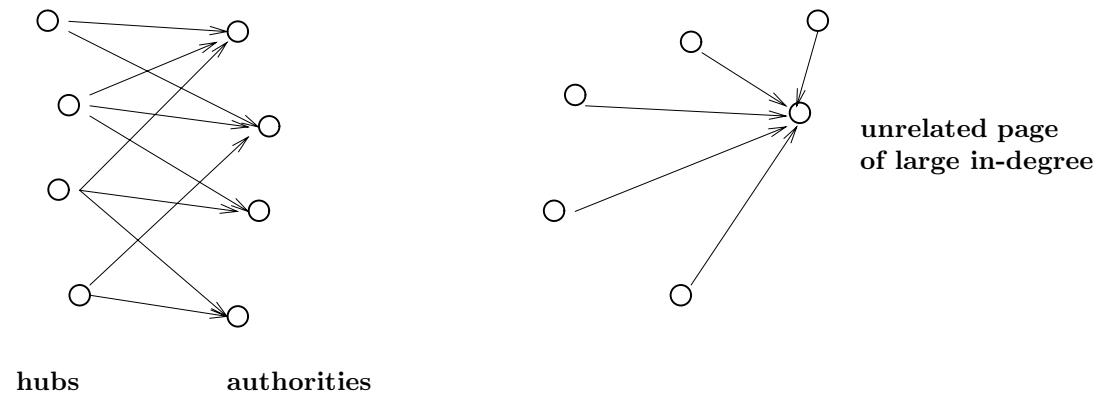 - Hubs & Authorities: Mutually reinforcing.

hubs   authorities

Figure 2: A densely linked set of hubs and authorities.

Iterative Algorithm

- Subgraph G = (V, A).

- Normalized weights, x<p> & y<p>.

- Update operations, I & O.

- Mutually reinforcing:

  - I: $x<p> = \sum y<q> \; \forall \; q : (q, p) \in A.$

  - O: $y<p> = \sum x<q> \; \forall \; q : (p, q) \in A.$

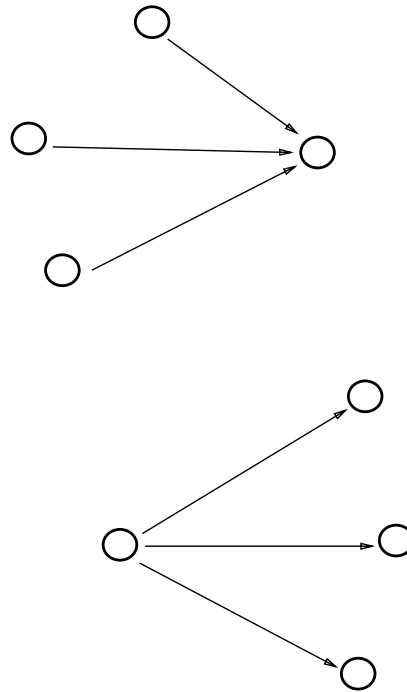Figure 3: The basic operations.

I & O

- x is a vector containing all x\<p>

- y is a vector containing all y\<p>

- Iterate(k): apply I & O and normalize.

- Filter(c): obtain c largest coordinates.

- Optimization of k is trivial:

  - x and y converge eventually. (3.1)

Iterate(G,k)

    G: a collection of n linked pages

    k: a natural number

    Let z denote the vector $(1, 1, 1, \ldots, 1) \in \mathbf{R}^n$.

    Set $x_0 := z$.

    Set $y_0 := z$.

    For $i = 1, 2, \ldots, k$

        Apply the $\mathcal{I}$ operation to (x_{i-1}, y_{i-1}), obtaining new x-weights x_i'.

        Apply the $\mathcal{O}$ operation to (x_i', y_{i-1}), obtaining new y-weights y_i'.

        Normalize x_i', obtaining x_i.

        Normalize y_i', obtaining y_i.

    End

    Return (x_k, y_k).

Iterate

Filter(G,k,c)

    G: a collection of n linked pages

    k,c: natural numbers

    $(x_k, y_k) := \mathtt{Iterate}(G, k)$.

    Report the pages with the c largest coordinates in x_k as authorities.

    Report the pages with the c largest coordinates in y_k as hubs.

Filter

Method Quirks

- Textual search as black box.

- Only probabilistically global.

- Does not address scarcity problem.

Similar-Page Queries

- "similar:www.example.com"

- Very little modification necessary!

- Obtain root set from in-pages search.

  - R = t pages pointing to p.

- In-degree still not a good ranking.

Related Work

- Standing in social networks.

- Influence in scientific citation networks.

- PageRank. (i.e.: WWW indices, no hubs)

Multiple Sets of H&A

- What about ambiguous query terms? (Terms with several meanings.)

- What about different contexts?

- What about polarized issues? (Groups that won't link to one another, but are debating the same topic.)

- Clusters exist.

Diffusion and Generalization

- **Diffusion**: pages corresponding to "broader" topics than the query string are returned, or reference page has insufficient in-degree.

  - Was the query string too specific?

- Possible solutions?

  - Non-principal eigenvectors.

  - Textual approaches (i.e.: term-matching)

Conclusions

- **Abundance** problem is harder each day.

  - Calls for search engines to consider more than simple relevance and clustering.

- Growth of WWW makes **indexing** harder.

- WWW search results must be **global**, WWW search process doesn't have to be.

- **Quality** of results is critical, more so as the WWW grows and becomes polluted.

Conclusions

- WWW is social.
  (Social organization is represented.)

- Further avenues:

  - User traffic pattern analysis.

  - Eigenvector-based heuristics. (LSA)

  - Link-based methods for other queries.