Authoritative Sources in a Hyperlinked Environment

Journal of the ACM 46(1999)

Jon Kleinberg, Dept. of Computer Science, Cornell University

Pranav K

Introduction

- Searching on the web is defined as the process of discovering pages relevant to the query
- Relevance is subjective. Quality metrics would require human intervention
- Objective functions that concretely define human notions of quality are missing

Authoritative Sources

Query type and issues

Type	Example	Issues
Specific-topic	*Who is John Galt?*	Scarcity problem
Broad-topic	*Works of American authors*	Abundance problem
Similar-page	*Find similar pages*	Defining similarity

- For broad-topic queries, effective search methods need to identify authorative sources
- Identification of authoritative sources is difficult
  - No apparent *endogenous* attribute of the page that enables it to be identified as an authority.
  - Natural authorities may not use self-identifying terms on their pages.

Analysis of Link Structure

A link based approach

- Paper proposes a link-based model for the conferral of authority.
- The model is based on the relationship that exists between authorities and hubs - pages linking to authorities
- The algorithm operates on focused subgraphs of the web producing a small collection of pages that are most likely to contain the authorative sources for a given topic.

Constructing a focused subgraph

- Graph $G(V, E)$ where
  - V$\leftarrow$ Set of pages
  - A directed edge $(p, q) \in E$ indicates presence of a link from p to q
  - Out degree(p): No. of nodes that p links to
  - In degree(p) No of nodes that link to p
- An induced graph on $W \subset V$ is a graph $G(W, E')$ such that $E' \subset E$ and $\forall (p, q) \in E', p \in W$ and $q \in W$
- Our goal is to identify a collection of pages S_σ with the following traits:-
  1. S_σ is relatively small
  2. S_σ is rich in relevant pages
  3. S_σ contains most of the strongest authorities.

Constructing a focused subgraph (cont.)

- Collect the t-top ranked pages from a search engine that returns text results. The set of pages is referred to as the root set R_σ
- Although the strong authority may not exist in R_σ, it is very likely to be linked to by at least one of the pages in the set R_σ
- The number of strong authorities in the subgraph can therefore be increased by expanding R_σ along the edges.
- The algorithm works in as follows:
  Set $S_\sigma = R_\sigma$
  For each p $\in R_\sigma$:
    Add all pages outgoing from p to S_σ
    Add an arbitrary set of pages d[a] from incoming links to p to S_σ
- The result is a focussed subgraph typically of the size between 1000 - 5000

[a] Limited to a subset since the in degree of p might be very large. In their experiments d was set to 50

Computing hubs and spokes

Preprocessing

- *Domain*: The first level in the url string
- Types of links:-
  - *Transverse*: A link between different domains
  - *Intrinsic*: A link between pages in the same domain. Intrinsic links are assumed typically navigational and hence contribute less information about authorities.
- All intrinsic links from $G[S_\sigma]$ are removed, keeping only the edges corresponding to transverse links.
- In addition, to account for mass advertisements / collusion among referring pages, that allows only up to m pages from a single domain to point to a any given page p. This heuristic is not used in the experiment.
- The result graph is denoted by G_σ

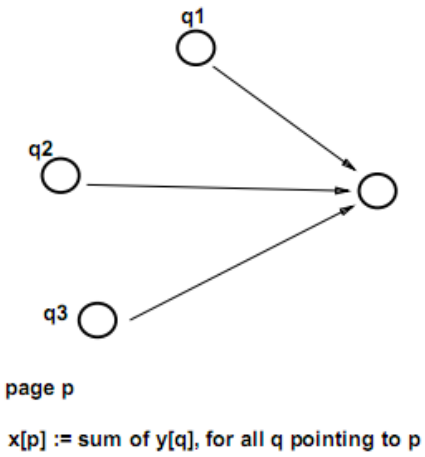
Computing hubs and spokes

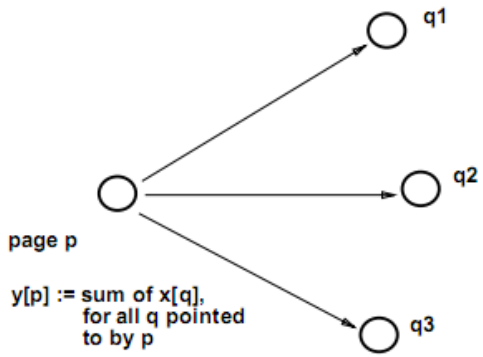
- Authorative pages would have:-
  - A high in-degree: Lots of pages linking to it
  - A considerable overlap in the set of pages that point to it. These pages are called hubs
- Hubs are pages pointing to multiple relevant authorative pages.
- Hubs and authorities exhibit mutually reinforcing relationship - a good hub is a page that points to many authoritative pages and a good authority is one that is pointed to by many good hubs.

Computing hubs and spokes

An iterative algorithm to assigning weights to pages

- Associated with a page p are:-
  - A non-negative authority weight $x^{<p>}$
  - A non-negative hub weight $y^{<p>}$
  - The weights are normalized so that $\sum_{p \in S_\sigma} (x^{<p>2}) = 1$ and $\sum_{p \in S_\sigma} (y^{<p>2}) = 1$
- Two operations on the weights:
  - $\mathcal{I}$: Updates x-weights as $x^{<p>} \leftarrow \sum_{q:(q,p) \in E} y^{<q>}$
  - $\mathcal{O}$: Updates y-weights as $y^{<p>} \leftarrow \sum_{q:(q,p) \in E} x^{<q>}$





Computing hubs and spokes

An iterative algorithm to assigning weights to pages

Iterate(G, k):

G: A collection of n linked pages

k: A constant

$x_0 = y_0 = 1, 1, 1...1$

For $i = 1...k$

Apply $\mathcal{I}$ operations to (x_{i-1}, y_{i-1}) obtaining new weights x'_i

Apply $\mathcal{O}$ operations to (x'_i, y_{i-1}) obtaining new weights y'_i

Normalize x'_i and y'_i to obtain x_i and y_i respectively

End

Return (x_k, y_k)

Computing hubs and spokes

Obtaining the top c authorities and hubs

Filter(G, k, c)

G: A collection of n linked pages

k, c: natural numbers

(x_k, y_k) = Iterate(G, k)

The pages with the c largest co-ordinates from x_k and y_k are the top authorities and hubs.

- $c \approx$ 5-10
- The Iterate procedure converges as k increases arbitrarily. In their experiments k = 20 provided sufficient convergence.

Basic Results

(java) Authorities

.328	http://www.gamelan.com/	*Gamelan*
.251	http://java.sun.com/	*JavaSoft Home Page*
.190	http://www.digitalfocus.com/digitalfocus/faq/howdoi.html	*The Java Developer: How Do I...*
.190	http://lightyear.ncsa.uiuc.edu/~srp/java/javabooks.html	*The Java Book Pages*
.183	http://sunsite.unc.edu/javafaq/javafaq.html	*comp.lang.java FAQ*

Interpretation of the results

- Besides the initial 'black-box' call to the search engine, the analysis ignored the textual content of the pages. This indicates a considerable amount can be accomplished using a 'pure' link analysis approach.
- It is possible to reliably estimate certain types of global information about the www using a local method of analysis on a small focussed subgraph.
- By modifying the root set to begin with R_p to constitute t pages that point to p, the algorithm can be made to perform similar-page queries.

Brushing up on some linear algebra

Eigen Vectors and Eigen Values

- An eigen vector is a column vector $X_{[n,1]}$ such that $M_{[n,n]}X_{[n,1]} = \lambda X_{[n,1]}$
- e.g. For M $= \begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix}$
  - $\begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ 0 \end{pmatrix}$
  - $\begin{pmatrix} 2 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \end{pmatrix} = 1 \begin{pmatrix} -1 \\ 1 \end{pmatrix}$
  - The scalars $\lambda = \{2, 1\}$ are called eigen values
- If the eigen values are ordered on decreasing eigen values i.e. $|\lambda_i|$, then the eigen vector corresponding to $|\lambda_0|$ is called the principal eigen vector and the rest are called (*surprise!*) non-principal eigen vector

Multiple Sets of Hubs & Authorities

- A number of scenarios where one may be interested in finding several densely linked collections of hubs and authorities. Each collection could be relevant but disconnected from each other:-
  - Ambigious / several meanings: e.g. *Java* - software platform, island, coffee
  - Highly polarized issues (pages that do not cross link): e.g. *abortion*
  - Queries that may encompass multiple communities e.g. *randomized algorithms*
- Multiple sets of hubs & authorities can be identified by considering non-principal eigen vectors to identify additional densely linked collections of hubs and authorities from the base set S_σ

Standings, Impact and Influence

Social Networks & Bibliometrics

- Relative standing of individuals in an implicitly defined social network can be measured by methods similar to the ones discussed for pages so far.
- Bilbliometrics is the study of written documents and their citation structure. It is considered an important measure of the importance and *impact* of individual scientific papers.
- Both these are analogous to the concept of authorities.
- The problems oculd be modeled as a graph $G = (V, E)$ where an edge (i, j) implies an endorsement of j by i
- The critical difference between these and link analysis is that in these scenarios, authorities would tend to endore other authorities (e.g. important research papers citing other important papers)

Diffusion & generalization

- When the initial query string σ specifies a query that is not broad, there will not be enough relevant pages in G_σ from which to extract a subgraph of authorities and hubs.
- Authorative pages competing with *broader* topics will compete and win out in the algorithm.
- This is termed as diffusing from the initial query
- Although, diffusion produces incorrect results for such a class of queries, they are useful to identify a natural generalization of a given topic.
- This problem is alleviated by the use of textual context in addition to the link structure and was is part of the paper's future work.

Evaluation

Yahoo! versus CLEVER versus AltaVista

- CLEVER is an implementation of the link analysis algorithm implemented independently. It improvises on the algorithm by using contextual link text information as additional heuristics.
- At the time the evaluation was performed, *Yahoo!* was a hierarchical manually maintained directory and provided the best measure of human judgement of authority.
- Users were asked to compare the results provided by the three yielding 1369 responses in all
-

Yahoo! and CLEVER Equivalent	31%
CLEVER better	50%
Yahoo! better	19%