Introduction to Search Engine Technology
Introduction to Link Structure Analysis

Ronny Lempel
Outbrain

Navigational Queries & Anchor Text

- Anchor text – the highlighted clickable text of a link
- Physically in page A, but actually describes the content of page B
- Many times, concisely defines pages B
- Anchor text is the dominant factor in ranking navigational queries
  - Queries where the user has a specific home page in mind, and uses the search engine to navigate to that page
- Sometimes (ab)used by a community to mislead search engines in a technique referred to as “Googlebombing”
Retrieval Issues when Using Anchor Text

- Anchor stop words: “home”, “back”, “click here”
- Should intra-site anchor-text be considered as important as inter-site anchor-text?
- In conjunctive queries, users may be dissatisfied when results that do not physically contain the query terms are returned (and counted)
  - Perhaps anchor text should only be used to enhance precision rather than determine recall?
- Anchor-only pages are pages for which the engine has encountered anchor-text but hasn’t crawled the page itself
  1. Are “anchor-only” pages eligible to be returned?
  2. Can effective summaries be constructed for such pages?

Link Analysis - Motivation

Traditional IR text-based ranking methods are not sufficient on the Web, even when considering anchor text:

- The web is huge, with great variation in the quality of the pages
- Queries are usually very short, making differentiation between pages difficult
- The text on many web pages does not sufficiently describe the page
- Keyword spamming

The emphasis of web search is on precision, not recall!
Link Analysis - Motivation

- Beyond anchor-text: the connectivity patterns between Web pages contain a gold mine of information.
- A link from page \(a\) to page \(b\) can often be interpreted as:
  1. A recommendation, by \(a\)'s author, of the contents of \(b\).
  2. Evidence that pages \(a,b\) share some topic of interest.
- A co-citation of \(a\) and \(b\) (by a third page \(c\)) may also constitute evidence that \(a,b\) share some topic of interest.
- However, not all links should be interpreted as such.

(not) All Links are Created Equal

- **The Good**: informative links, in the sense described previously
- **The Bad**: non-informative links, such as:
  1. Intra-domain, navigational links
  2. Commercials, sponsorship links
- **...and the Ugly**:
  1. Link spamming
  2. Non-semantic link exchanges
Mathematical Background - Irreducibility

- A directed graph $G=(V,E)$ is called **irreducible** if for every $i,k \in V$ there is a path in $G$ originating at $i$ and ending in $k$.
- A non-negative $N \times N$ matrix $W$ is called **irreducible** if $\forall \ i,k \in \{1,…,N\}$ there exists a non-negative integer $m$ such that $[W^m]_{i,k}>0$.

The **support graph** $G_W=\{V_W,E_W\}$ of a non-negative $N \times N$ matrix $W$ is a directed graph with $N$ vertices, such that $i \rightarrow k \in E_W$ iff $W_{i,k}>0$.

- Lemma: a non-negative square matrix $W$ is irreducible if and only if $G_W$ is irreducible.
- Lemma: a directed graph $G$ is irreducible if and only if the adjacency matrix of $G$ is irreducible.

Mathematical Background: Perron-Frobenius Theorem (simplified)

- Let $W$ be a non-negative $N \times N$ matrix, and denote by $\lambda_1(W), \lambda_2(W),…,\lambda_N(W)$ the $N$ eigenvalues of $W$, ordered by non-decreasing absolute value (i.e. $|\lambda_1(W)| \geq |\lambda_2(W)| \geq … \geq |\lambda_N(W)|$).
  - $|\lambda_1(W)|$ is the spectral radius of $W$ and will simply be denoted by $\lambda(W)$.
- Perron-Frobenius Theorem for irreducible matrices: let $W$ be an irreducible matrix. Then:
  - $\lambda(W) > 0$
  - $\lambda(W)$ is a simple eigenvalue of $W$ (i.e. it is a simple root of the characteristic polynomial of $W$)
    - In particular, $\lambda(W)$ is a real eigenvalue of $W$
  - $W$ has positive left and right eigenvectors corresponding to $\lambda(W)$ (i.e. all the components in those eigenvectors are greater than zero).
Mathematical Background:
Implications of Perron-Frobenius Theory

- Lemma: let $W$ be an irreducible $N \times N$ matrix. A sufficient condition that guarantees that $|\lambda_1(W)| > |\lambda_2(W)|$ is that for some index $j$, $W_{jj} > 0$ (i.e. $W$ has some non-zero element on its main diagonal).

- Corollary: let $W$ be an irreducible $N \times N$ matrix for which $|\lambda_1(W)| > |\lambda_2(W)|$, and let $v$ be a real eigenvector of $W$ that doesn’t correspond to $\lambda_1(W)$. Then $v$ has both positive and negative entries.

What we’ve learned: let $W$ be an irreducible $N \times N$ matrix with some non-zero element on its main diagonal. Then:

- $\lambda(W) = \lambda_1(W) > \lambda_2(W)$
- There is a unique positive unit eigenvector of $W$ that corresponds to $\lambda(W)$ – denote it by $v_{\lambda(W)}$
- Let $e$ be any non-negative (but not zero) vector. Then, $\langle e, v_{\lambda(W)} \rangle > 0$
- Any real eigenvector of $W$ that doesn’t correspond to $\lambda(W)$ has both positive and negative entries.
Mathematical Background: The Power Method

The Power Method is an iterative algorithm to compute (under certain conditions) the eigenvalue of largest absolute value and the corresponding eigenvector of a general NxN matrix M.

- The conditions:
  - M should have N linearly independent eigenvectors
  - $|\lambda_1(M)| > |\lambda_2(M)|$
  - One can choose an initial (unit) vector $v_0$ that is not orthogonal to the eigenvector that corresponds to $\lambda_1(M)$

- The iteration:
  1. $v_{k+1} = Mv_k$
  2. Scale $v_{k+1}$ to be a unit vector

As $k \to \infty$, $v_k$ will approach the eigenvector corresponding to $\lambda_1(M)$.

Hubs and Authorities

- Notions proposed by J. Kleinberg in his 1998 paper "Authoritative Sources in a Hyperlinked Environment"
- Two types of quality Web pages that pertain to a given topic
  - Hubs: resource lists, containing links to many authorities.
  - Authorities: pages containing authoritative content on the topic of interest

Hubs and authorities exhibit a mutually reinforcing relationship.
Authorities - Authoritative pages

Here is an authority for the query “movies”:

Hubs - Pages Which Link to Many Authorities

Here is a hub for the query “movies”:
Hypertext Induced Topic Search

**HITS**: an algorithm proposed by Kleinberg to identify hubs and authorities, given a topic of interest \( t \)

1. Identifies a subgraph of the Web where many hubs and authorities pertaining to \( t \) should exist
2. Assigns each page \( s \) an authority weight \( a(s) \) and a hub weight \( h(s) \) s.t. the following Mutual Reinforcement properties holds:
   \[
   a(s) \text{ is proportional to } \sum_{x \rightarrow s} h(x)
   \]
   \[
   h(s) \text{ is proportional to } \sum_{x \rightarrow s} a(x)
   \]
- Such weights exist, and can be efficiently computed
- Based on the Perron-Frobenius theory of non-negative matrices

HITS: Assembling the Collection

- Given a query, use a term based search engine to compile the *root set* - an initial set of (somewhat on-topic) pages
- Expand the root set to its bidirectional neighborhood of radius 1

The sub-graph induced by this process will be denoted by \( G \)
HITS: Analyzing the Graph $G$

Assign each page $s$ (=node in $G$) an authority weight $a(s)$ and a hub weight $h(s)$ according to the following iterative algorithm:

- $a(s) \leftarrow 1$, $h(s) \leftarrow 1$ for all pages $s$.
- Repeat the following three operations, until an equilibrium is reached (i.e. until convergence):
  - The “I” operation: for all pages $s$, $a(s) \leftarrow \sum\{x| x \rightarrow s\} h(x)$
  - The “O” operation: for all pages $s$, $h(s) \leftarrow \sum\{x| s \rightarrow x\} a(x)$
  - Normalize the resulting weights vectors

---

HITS: Matrix Notation

Let $W$ denote the adjacency matrix of $G$, $a=(a(s_1), \ldots, a(s_n))^T$, $h=(h(s_1), \ldots, h(s_n))^T$

- The “I” operation: $a \leftarrow W^T h$.
- The “O” operation: $h \leftarrow W a$.

Hence (without normalizing):

- $a_{k+1} \leftarrow W^T(W a_k) = (W^T W) a_k$
- $h_{k+1} \leftarrow W(W^T h_k) = (W W^T) h_k$
HITS – Algebraic Interpretation

- Repeating the three-step iteration $(I, O, Norm)$ is equivalent to applying the power method, which computes the principal eigenvector of a matrix.

\[
l_{mk \rightarrow \infty} a_k = \text{principal eigenvector of } W^TW \\
l_{mk \rightarrow \infty} h_k = \text{principal eigenvector of } WW^T
\]

- The principal community of authorities is defined as the $m$ pages with the highest authority weights in $a$.
- The principal community of hubs is similarly defined in terms of $h$.

HITS: the Role of the Matrices

- $W^TW$ is the co-citation matrix of the collection
  - $[W^TW]_{ij}$ is the number of pages that link to both $i$ and $j$
- $WW^T$ is the bibliographic coupling matrix of the collection
  - $[WW^T]_{ij}$ is the number of pages to which both $i$ and $j$ link
- Both matrices are non-negative and symmetric
  - Thus their eigenvalues are all real, and they have $N$ linearly independent eigenvectors that span $\mathbb{R}^N$
  - If $W$ is non-zero, both matrices will have at least one non-zero element on their main diagonal
  - If the matrices are irreducible, they both have a unique unit positive principal eigenvector, and all non-principal eigenvectors contain both positive and negative entries.
HITS: Non-Principal Communities

- Denote by $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_N$ the eigenvalues of $W^T W (W W^T)$ and by $a_i(h_i)$ the corresponding eigenvectors.
- Weights derived from all pairs $(a_i; h_i)$ exhibit the Mutual Reinforcement property.
- When $i > 1$, eigenvectors contain both positive and negative entries.
- The non-principal communities are determined by $a_2, \ldots, a_m, h_2, \ldots, h_m$ (for some small $m$).
  - Each non-principal eigenvector yields two communities, corresponding to the positive and negative entries of highest absolute value.
  - Experiments show that in ambiguous/polarized topics, the polar communities of a non-principal eigenvector correspond to opposing/orthogonal aspects of the topic.
- $a_2, \ldots, a_m, h_2, \ldots, h_m$ are computed by applying the power method with Gram-Schmidt steps.

Mathematical Background:
Stochastic and Primitive Matrices

- Definition: a non-negative N$x$N matrix $P$ is stochastic if the sum of every row in $P$ is 1.
- Definition: the period of a directed graph $G$ is the greatest common divisor of the lengths of all the cycles in $G$.
  - $G$ is called aperiodic if it has a period of 1.
- Definition: a non-negative N$x$N matrix $M$ is called primitive if its support graph $G_M$ is aperiodic.
Mathematical Background: Ergodic Theorem

Ergodic Theorem: let $P$ be an irreducible and primitive stochastic matrix

- $\lambda(P) = \lambda_1(P) = 1$, and any other eigenvalue of $\lambda^*$ of $P$ satisfies $|\lambda^*| < 1$
- There is a unique distribution row-vector $\pi$ which satisfies $\pi P = \pi$
- $\pi$ is the principal eigenvector of $P$ and is the stationary distribution of the Markov Chain defined by the transition matrix $P$
- For any distribution row-vector $q$, $\lim_{k \to \infty} q P^k = \pi$

PageRank (Brin & Page, 1998)

- Named after Google’s co-founder, Larry Page
- A global, query independent importance measure of Web pages
- A page is considered important if it receives many links from important pages
- Based on Markov chains and random walks
PageRank: “Random Surfer” Model

A random surfer moves from page to page. Upon leaving page $p$, the surfer chooses one of two actions:

1. Follows an outgoing link of $p$ (chosen uniformly at random), with probability $d$
   - See next slide for discussion of pages that have no outlinks
2. Jumps to an arbitrary Web page (chosen uniformly at random), with probability $1-d$

The vector of PageRanks is the stationary distribution of this (ergodic) random walk.

PageRank – Handling Dangling Nodes

PageRank as stated in the previous slide is not well defined with respect to exiting pages that have no outgoing links (dangling nodes).

There are three accepted approaches for treating pages with no outgoing links:

1. Eliminate such pages from the graph (iteratively prune the graph until reaching a steady state)
2. Consider such pages to link back to the pages that link to them
3. Consider such pages to link to all web pages (effectively making an exit out of them equivalent to a random jump)
**PageRank: Steady State Equations**

The PageRanks obey the following equations:

\[ R(p) = \frac{(1-d)}{N} + d \sum_{j \in I(p)} \frac{R(j)}{D(j)} \]

- \( R(p) \) – The PageRank of page \( p \).
- \( d \) – A damping factor, \( 0 < d < 1 \)
- \( N \) – Number of Web pages
- \( I(p) \) – The set of pages that point to \( p \)
- \( D(j) \) – Number of out-links (out-degree) of page \( j \)

**PageRank – Algebraic Notation**

- Let \( W \) denote the NxN adjacency matrix of the Web’s link structure, after some form of handling dangling nodes

- Let \( W_{\text{norm}} \) denote the matrix that results by dividing each row \( j \) of \( W \) by \( j \)’s out-degree (row \( j \)’s sum)

- Let \( T \) by an NxN matrix whose entries are all equal to \( 1/N \)

- Define \( M = (1-d)T + dW_{\text{norm}} \)
  - \( M \) is the transition matrix corresponding to PageRank’s random walk
  - \( M \)’s principal eigenvector is the vector of PageRanks
PageRank – Alternative Interpretation

- Repeat the following:
  - Sample a value $L$ from a Geometric distribution with parameter $1-d$, i.e. $P(L) = (1-d)d^{L-1}$
  - Starting from a random Web page, walk a random path of length $L-1$, where each page is exited by one of its outlinks chosen uniformly at random

- Essentially, PageRank can be viewed as a weighted average of the number of paths leading to each node, where the average is weighted by:
  - Length of the path (probability decreases geometrically)
  - Each edge is traversed with probability inverse to the out-degree of the node from which it originates

\[
R(p) = \frac{1-d}{N} \sum_{L=1}^{\infty} d^{L-1} \sum_{v \in G} \text{Prob}(\text{paths } v \to p \text{ of length } L-1)
\]