Web graph and PageRank algorithm

Danil Nemirovsky[1]

[1] Department of Technology of Programming
Faculty of Applied Mathematics and Control Processes
St. Petersburg State University

Outline

Outline

1. Web graph
2. Markov theory
3. PageRank
4. Decomposition
5. Aggregation/Disaggregation methods

Basic terminology of graph theory. I

Definition (Directed graph)

A **directed graph** G is a pair $G = (V, E)$, where V is a set of any nature, elements of which is called nodes, E is a set of ordered pairs (u, v) called arcs.

Definition (In-degree and out-degree)

The **out-degree** of a node u is the number of distinct arcs $(u, v) \in E$, and the **in-degree** is the number of distinct arcs $(v, u) \in E$.

Basic terminology of graph theory. II

Definition (Path)

A **path** from node u to node v is a sequence of arcs $(u, u_1), (u_1, u_2), \dots, (u_k, v)$, where $(u, u_1), (u_i, u_{i+1}), (u_k, v) \in E, \forall i = \overline{1, k-1}$.

Definition (Strongly connected component)

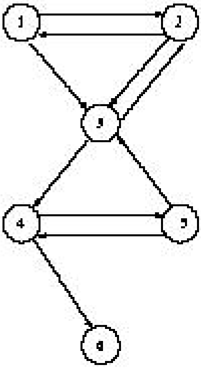
A **strongly connected component** (strong component for brevity) of a graph $G = (V, E)$ is a set of nodes such that for any pair of nodes u and v in the set there is a path from u to v.

Definition (Diameter)

A **diameter** of a graph $G = (V, E)$ is the maximum over all ordered pairs (u, v) of the shortest path from u to v.

Definition of the Web graph.

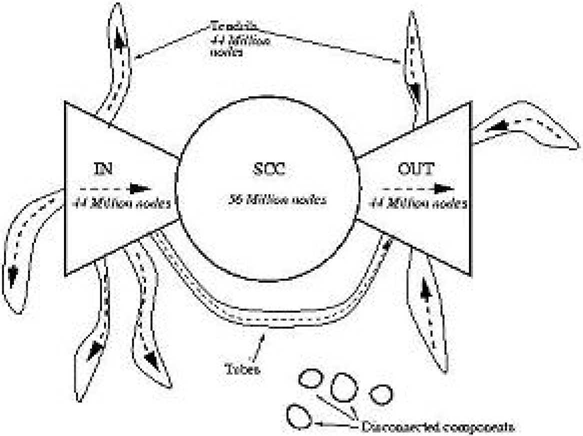
- We consider pages in the Web as nodes.
- Links between pages are arcs.
- We obtain graph called the Web graph.



Properties of the Web graph.

1. Macroscopic structure of the Web graph
2. Diameter of the Web graph
3. In- and out-degree distributions

Macroscopic structure of the Web graph.



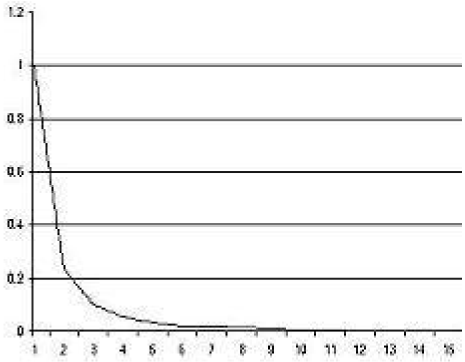
In- and out-degree distributions.

- It is turned out that in- and out-degree are distributed according to power law.
- the probability that a node has in-degree (out-degree) i is proportional to
- $(x > 1)$

$$\left(\frac{1}{i}\right)^{x}$$

In- and out-degree distributions.

In-degree: the exponent of the power law is around 2.1
Out-degree: the exponent of the power law is around 2.72



Outline

1. Web graph
2. Markov theory
3. PageRank
4. Decomposition
5. Aggregation/Disaggregation methods

Markov processes.

Definition (Markov process)

An S-valued **Markov process** is an infinite sequence of random variables $X_k = X_0, X_1, \ldots \in S$ if S is finite and the probability function P satisfies:
$P(X_{k+1} = b | X_0 = a_0, \ldots, X_k = a_k) = P(X_{k+1} = b | X_k = a_k)$ is the same for all $k \geqslant 0$.
Its **transition function** is $\omega(a, b) = P(X_{k+1} = b | X_k = a)$.
Its **initial distribution** is $\sigma(a) = P(X_0 = a)$.

In the Stochastic processes literature, this is technically called a homogeneous, discrete time, finite space Markov process. In applications of the theory, they are often simply called Markov processes or Markov chains.

Convergence of Markov processes. I

Definition (**Period of state**)

Let $\{X_k\}$ be an S-valued Markov process. The **period** of a state $a \in S$ is the largest d satisfying: $(\forall k, n \in \mathbb{N})$

$$P(X_{n+k} = a | X_k = a) > 0 \Rightarrow d \textit{ divides } n$$

If $d = 1$, then the state a is **aperiodic**.

Definition (**Ergodic Markov process**)

An **ergodic** Markov process is a Markov process $\{X_k\}$ that is both

- **irreducible**: every state is reachable from every other state.
- **aperiodic**: the greatest common divisor of the states' periods is 1.

Convergence of Markov processes. II

Lemma (Ergodic Condition)

An irreducible S-valued Markov process with transition function ω that has $\omega(a, a) > 0$ for some state $a \in S$ is aperiodic, and hence ergodic.

Theorem (Ergodic Convergence)

If $\{X_k\}$ is an ergodic S-valued Markov process, then the probability function converges for all $a \in S$:

$$\lim_{k \to \infty} P(X_k = a) = p_a$$

Transition matrix and stationary distribution.

- If the set of states is finite we can define transition matrix.
- If the Markov chain is ergodic, then it has unique stationary probability distribution

$$P_{ij} = \omega(a_i, a_j), \forall a_i, a_j \in S$$

$$\pi P = \pi \quad \pi e = 1$$

Power method.

- $\|\pi\|_1 = \pi e$
- v is the first approximation
- ε is an accuracy
- rate of convergence $\frac{|\lambda_2|}{|\lambda_1|}$
- If P is row-stochastic matrix then $\lambda_1 = 1,\ 1 \geqslant |\lambda_2| \geqslant |\lambda_3| \geqslant \ldots \geqslant |\lambda_n| \geqslant 0$

$$\pi^{(k+1)} = \pi^{(k)} P$$

function $\pi^{(m)} = PowerMethod(P, v, \varepsilon)$

{

$\quad \pi^{(0)} = v;$

$\quad k = 1;$

$\quad$ **repeat**

$\qquad \pi^{(k)} = \pi^{(k-1)} P;$

$\qquad \delta = \|\pi^{(k)} - \pi^{(k-1)}\|_1;$

$\qquad k = k + 1;$

$\quad$ **until** $\delta < \varepsilon;$

}

Outline

Defining of PageRank.

- A is a page
- c is a damping factor
- T_i is a page, linking to the page A
- $\pi(A)$ is PageRank of a page A
- $l(T_i)$ is the number of outgoing link from T_i

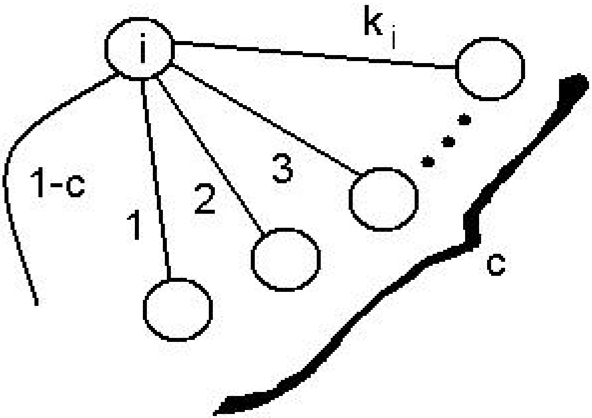
$$\pi(A) = \frac{(1-c)}{n} + c(\pi(T_1)/l(T_1) + \ldots + \pi(T_m)/l(T_m))$$

PageRank vector.

- If we number all pages we can define a PageRank vector as row vector whose every entry is PageRank of some page.
- The PageRank vector is a stationary distribution of specially formed Markov chain

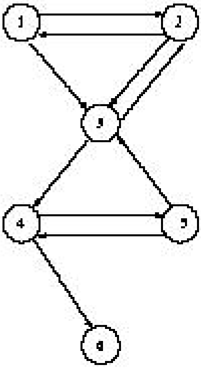
$$
\begin{array}{ccc}
p_1 & \to & \pi_1, \\
p_2 & \to & \pi_2, \\
\ldots & \ldots & \ldots, \\
p_n & \to & \pi_n.
\end{array}
$$

Defining Markov chain.



Transition matrix.

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix}$$



Google matrix and PageRank.

$$G = cP + (1-c)1/nE$$

$$\pi = \pi G$$

$$\pi e = 1$$

- Google: c = 0.85
- About 6 clicks before going to arbitrary page

$$\pi = \frac{1-c}{n} e^t (I - cP)^{-1}$$

Power method for PageRank.

- $v = (1/n, 1/n, \dots, 1/n)$ is the first approximation
- ε is an accuracy
- $PowerMethod(G, v, \varepsilon)$
- Rate of convergence = c
- $c = 0.85 \Rightarrow$ about 100 iterations
- $c = 0.99 \Rightarrow$ about 1000 iterations

Outline

Decomposition a Google matrix.

$$P = \begin{pmatrix} P_{11} & P_{12} & \dots & P_{1N} \\ P_{21} & P_{22} & \dots & P_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ P_{N1} & P_{N2} & \dots & P_{NN} \end{pmatrix}$$

where $N < n$. The PageRank vector is

$$\pi = (\pi_1, \pi_2, \dots, \pi_N)$$

where π_I is row vector with $dim(\pi_I) = n_I$ and

$$\sum_{I=1}^{N} n_I = n$$

Block-diagonal case.

$$P = \begin{pmatrix} P_1 & 0 & \dots & 0 \\ 0 & P_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & P_N \end{pmatrix}$$

$$G_I = cP_I + (1-c)1/n_I E$$

$$\pi_I = \pi_I G_I$$

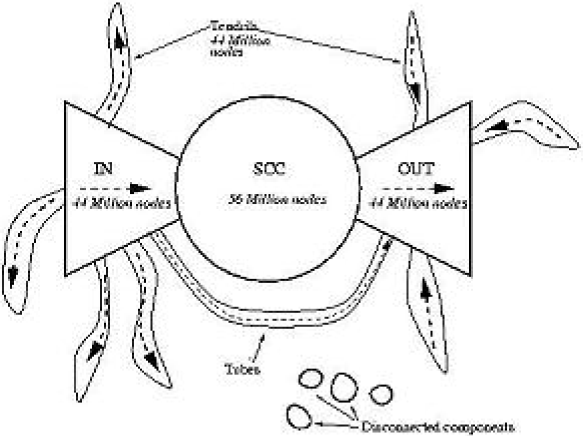
$$\pi_I e = 1$$

Theorem

The PageRank π is given by

$$\pi = \left(\frac{n_1}{n}\pi_1, \frac{n_2}{n}\pi_2, \dots, \frac{n_N}{n}\pi_N \right)$$

Macroscopic structure of the Web graph.



2 × 2 case.

$$P = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}, \pi = (\pi_1, \pi_2)$$

$$\pi(I - P) = 0.$$

$$I - P = LDU$$

$$L = \begin{pmatrix} I & 0 \\ -P_{21}(I - P_{11})^{-1} & I \end{pmatrix}$$

$$D = \begin{pmatrix} I - P_{11} & 0 \\ 0 & I - S \end{pmatrix}$$

$$U = \begin{pmatrix} I & -(I - P_{11})^{-1} P_{12} \\ 0 & I \end{pmatrix}$$

$$S = P_{22} + P_{21}(I - P_{11})^{-1} P_{12}$$

$$\pi LD = 0$$

$$\pi_2 S = \pi_2 \qquad \pi_1 = \pi_2 P_{21}(I - P_{11})^{-1}$$

$$\sigma S = \sigma, \ \sigma e = 1$$

$$\pi_2 = \rho\sigma \ \pi e = 1$$

Outline

Aggregation/Disaggregation methods.

The Power Method converges for components with different rate and we do more then need iteration for the components.

$$\pi = (\pi_1, \pi_2, \dots, \pi_N)$$

$$G = \begin{pmatrix} G_{11} & G_{12} & \dots & G_{1N} \\ G_{21} & G_{22} & \dots & G_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ G_{N1} & G_{N2} & \dots & G_{NN} \end{pmatrix}$$

Blockrank method.

$$\pi_i, \; i = \overline{1, N}$$

$$\pi_i = PowerMethod(G_{ii}, \frac{1}{n}e^t, \varepsilon)$$

$$A_{ij} = \pi_i G_{ij} e$$

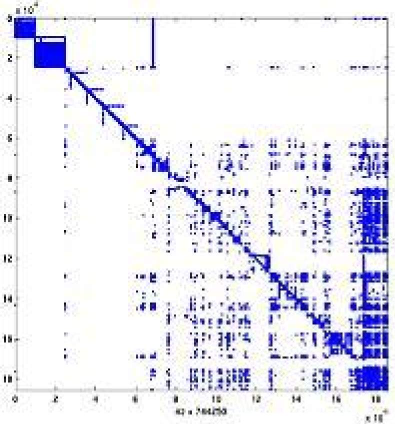
$$\nu A = \nu$$

$$\widetilde{\pi} = (\nu_1 \pi_1, \dots, \nu_N \pi_N)$$

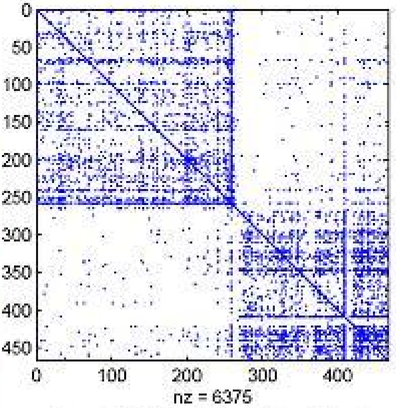
$$\pi = PowerMethod(G, \widetilde{\pi}, \varepsilon)$$

$$G = \begin{pmatrix} G_{11} & G_{12} & \dots & G_{1N} \\ G_{21} & G_{22} & \dots & G_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ G_{N1} & G_{N2} & \dots & G_{NN} \end{pmatrix}$$

Blockrank method.



IBM



Stanford/Berkeley Host Graph

Iteration aggregation/disaggregation method.

$function\ \pi^{(m)} = IAD(G, v, \varepsilon)\{$

$\pi^{(0)} = v;$

$k = 1;$

repeat

$A^{(k)}_{ij} = \pi^{(k)}_i G_{ij} e;$

$\nu^{(k)} A^{(k)} = \nu^{(k)};$

$\tilde{\pi}^{(k)} = (\nu^{(k)}_1 [\pi^{(k)}_1], \ldots, \nu^{(k)}_N [\pi^{(k)}_N])$

$\pi^{(k+1)} = \tilde{\pi}^{(k)} G^m$

$\delta = \|\pi^{(k+1)} - \pi^{(k)}\|_1;$

$k = k + 1;$

until $\delta < \varepsilon;$

$\}$

$$[\pi_i] = \frac{\pi_i}{\pi_i e}$$

$$\pi = (\pi_1, \pi_2, \ldots, \pi_N)$$

$$G = \begin{pmatrix} G_{11} & G_{12} & \ldots & G_{1N} \\ G_{21} & G_{22} & \ldots & G_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ G_{N1} & G_{N2} & \ldots & G_{NN} \end{pmatrix}$$

Summary

- The World Wide Web was represented as a directed graph and properties if the Web graph was considered.
- PageRank algorithm and different methods of finding PageRank are discussed.

- Outlook
  - Convergence of Iteration aggregation/disaggregation method will be researched.

Thank you for your patience and attention!

References I

- K.Avrachenkov and N.Litvak. Decomposition of the Google PageRank and Optimal Linking Strategy. Inria Sophia Antipolis, University of Twente, 2004.
- E.Behrends. Introduction to Markov Chains (with Special Emphasis on Rapid Mixing). Vieweg Verlag, 1999.
- A.Berman and R.J.Plemmons. Nonnegative Matrices in the Mathematical Sciences. SIAM Classics In Applied Mathematics, SIAM, Philadelphia, 1994.
- M.Bianchini, M.Gori, and F.Scarselli. Inside PageRank. ACM Trans, Internet Technology, In press, 2002.

References II

- A.Broder, R.Kumar, F.Maghoul, P.Raghavan, S.Rajagopalan, R.Stata, A.Tomkins, J.Wiener. Graph structure in the web. Proc. WWW9 conference, 309-320, May 2000. http://www9.org/w9cdrom/160/160.html
- A.Clausen. Online Reputation Systems: The Cost of Attack of PageRank. 2003
- G.H.Golub and C.F.V.Loan. Matrix Computations. The Johns Hopkins University Press, Baltimore, 1996.
- T.H.Haveliwala and S.D.Kanvar. The second eigenvalue of the Google matrix. Tech. Rep. 2003-20, Stanford University, March 2003. http://dbpubs.stanford.edu/pub/2003-20
- C.F.Ipsen and S.Kirklad. Convergence analysis of the Langville-Meyer PageRank algorithm.

References III

- S.Kamver, T.Haveliwala, C.Manning, and G.Golub. Exploiting the block structure of the web for computing PageRank. Tech. Rep. SCCM03-02, Stanford University, http://www-sccm.stanford.edu/nf-publications-tech.html, 2003.
- A.N.Langville and C.D.Meyer. Deeper Inside PageRank. Preprint, North Carolina State University, 2003.
- C.Meyer. Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems. SIAM Rev., 31 (1989), pp. 240-72.
- C.D.Moler and K.A.Moler. Numerical Computing with MATLAB. SIAM, 2003.

References IV

- L.Page, S.Brin, R.Motwani, and T.Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.
- J.H.Wilkinson. The Algebraic Eigenvalue Problem. Oxford University Press, Oxford, 1965.