Perron vector Optimization applied to search engines

Olivier Fercoq

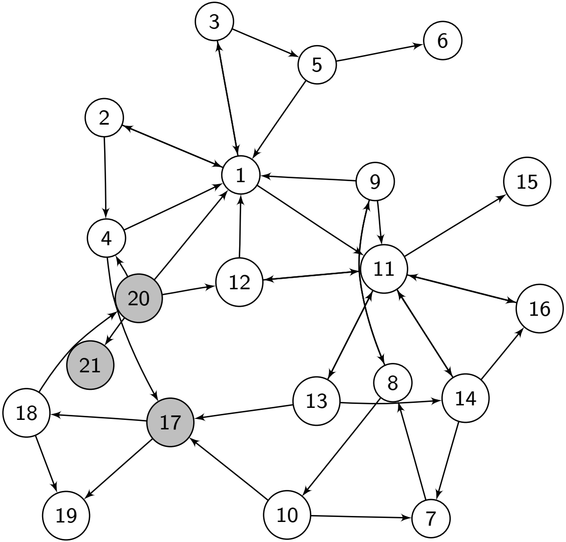
INRIA Saclay and CMAP Ecole Polytechnique

May 18, 2011

Web page ranking

- The core of search engines
- Semantic rankings (keywords)
- Hyperlink based rankings:
  - PageRank [Brin and Page, 1998]
  - HITS [Kleinberg, 1998]
  - Salsa [Lempel and Moran, 2000], HOTS [Tomlin, 2003]
- Webmasters want to be on top of the list

Toy example with 21 pages



Nodes = web pages

Arcs = hyperlinks

21 : controlled page

1 : non controlled page

Definition of PageRank [Brin and Page, 1998]

- An important page is a page linked to by important pages

$$\pi_i = \sum_{j:j\to i} \frac{\pi_j}{N_j} = \sum_j \pi_j S_{ji}$$

- π_i = "popularity" of page i
- Model for the behaviour of a random surfer
- Markov chain model may be reducible
  $\rightarrow$ with probability $1-\alpha$, surfer gets bored and resets
- $P_{i,j} = \alpha S_{i,j} + (1-\alpha) z_j$
- $P_{i,j} > 0, \forall i, j$ ($\alpha = 0.85$)
- PageRank is the unique invariant measure π of P

Definition of HITS [Kleinberg, 1998]

- Given a query, we build a subgraph of the web graph, focused on the query
- Hub and authority scores

$$\rho y_j = \sum_{i:j \to i} x_i \qquad \rho x_i = \sum_{j:j \to i} y_j$$

- A: adjacency matrix of the subgraph: $A^T A x = \rho^2 x$
- Uniqueness guaranteed with $(A^T A + \xi e e^T) x = \rho^2 x$
- Generalization and application to synonyms by [Blondel, Gajardo, Heymans, Senellart and van Dooren, 2004]

The problem

- Maximize ranking
- Controlled pages and non controlled pages

Effort concentrated on PageRank

- Avrachenkov and Litvak, 2006, one page
- Matthieu and Viennot, 2006, unconstrained problem
- de Kerchove, Ninove, van Dooren, 2008, qualitative result
- Ishii and Tempo, 2010, computation with fragile data
- Csáji, Jüngers and Blondel, 2010, optimization algorithm
- OF, Akian, Bouhtou, Gaubert, 2010 (arXiv:1011.2348v1)

PageRank optimization

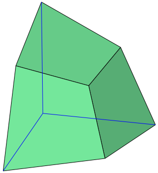
- Obligatory links $\mathcal{O}$, optional links $\mathcal{F}$, prohibited links $\mathcal{I}$ [Ishii and Tempo, 2010]
- Discrete problem
- 2^n hyperlink configurations by controlled page
- Utility $f^{PR}(u, P) = \sum_i r_{i,j} u_i P_{i,j}$
- $r_{i,j}$ is viewed as reward by click on $i \to j$

Admissible transition matrices (PageRank)

Theorem (OF, Akian, Bouhtou, Gaubert)

The convex hull of the set of admissible transition probabilities is either a simplex or a polyhedron defined by:

$$
\begin{array}{ll}
\forall j \in \mathcal{I}_i \ , & x_j = (1-\alpha) z_j \\
\forall j \in \mathcal{O}_i \setminus \{j_0\} \ , & x_j = x_{j_0} \\
\forall j \in \mathcal{F}_i \ , & (1-\alpha) z_j \leq x_j \leq x_{j_0} \\
 & \sum_{j \in [n]} x_j = 1
\end{array}
$$



- Discrete problem equivalent to a continuous problem
- [Csáji, Jüngers and Blondel, 2010]: graph rewriting

Reduction to ergodic control

Proposition

The PageRank Optimization problem is equivalent to the ergodic control problem:

$$\max_{(\nu_t)_{t\geq 0}} \liminf_{T\to+\infty} \frac{1}{T}\mathbb{E}\Big(\sum_{t=0}^{T-1} r_{X_t,X_{t+1}}\Big)$$

$$\nu_t \in \mathcal{P}_{X_t}, \forall t \geq 0$$

$$\mathbb{P}(X_{t+1} = j | X_t = i, \nu_t = p) = p_j, \forall i,j \in [n], \forall p \in \mathcal{P}_i, \forall t \geq 0$$

where ν_t *is a function of the history* $(X_0, \nu_0, \dots, X_{t-1}, \nu_{t-1}, X_t)$

Corollary

PageRank optimization problem with local constraints is solvable in polynomial time

Resolution by Dynamic Programming

- The ergodic dynamic programming equation
$$w_i + \psi = \max_{\nu \in \mathcal{P}_i} \nu(r_{i,\cdot} + w), \quad \forall i \in [n] \tag{1}$$
has a solution $(w, \psi) \in \mathbb{R}^n \times \mathbb{R}$. The constant ψ is unique and is the value of the ergodic control problem.
- An optimal strategy is obtained by selecting for each state i a maximizing $\nu \in \mathcal{P}_i$.

Resolution by Dynamic Programming

- The ergodic dynamic programming equation
$$w_i + \psi = \max_{\nu \in \mathcal{P}_i} \nu(r_{i,\cdot} + w), \quad \forall i \in [n] \qquad (1)$$
has a solution $(w, \psi) \in \mathbb{R}^n \times \mathbb{R}$. The constant ψ is unique and is the value of the ergodic control problem.
- An optimal strategy is obtained by selecting for each state i a maximizing $\nu \in \mathcal{P}_i$.
- The unique solution of the discounted equation
$$w_i = \max_{\nu \,:\, \alpha\nu + (1-\alpha)z \in \mathcal{P}_i} \alpha\nu(r_{i,\cdot} + w) + (1-\alpha)zr_{i,\cdot}, \forall i \in [n] \quad (2)$$
is solution of (1) with $\psi = (1-\alpha)zw$
- The fixed point scheme for (2) has contracting factor α independent of the dimension

Optimality criterion

Define $\forall P \in \mathcal{P}$, $v(P) = (I_n - \alpha S)^{-1}\bar{r}$, where $\bar{r}_i = \sum_j P_{i,j} r_{i,j}$

- If $r_{i,j} = r_i$, $v(P)$ is the mean reward before teleportation considered by [de Kerchove, Ninove, van Dooren, 2008]
- $(v_j(P) + r_{i,j})\pi_i(P)$ is the derivative of the objective at P

Proposition

$P^* \in \mathcal{P}$ is an optimum of the PageRank Optimization problem if and only if:

$$\forall i, \forall \delta \in \mathbb{R}^n \text{ s.t. } P^*_{i,\cdot} + \delta \in \mathcal{P}_i, \text{ we have: } \sum_j (v_j(P^*) + r_{i,j})\delta_j \leq 0$$

- $v(P^*)$ is solution of the dynamic programming equation.

Existence of a master page

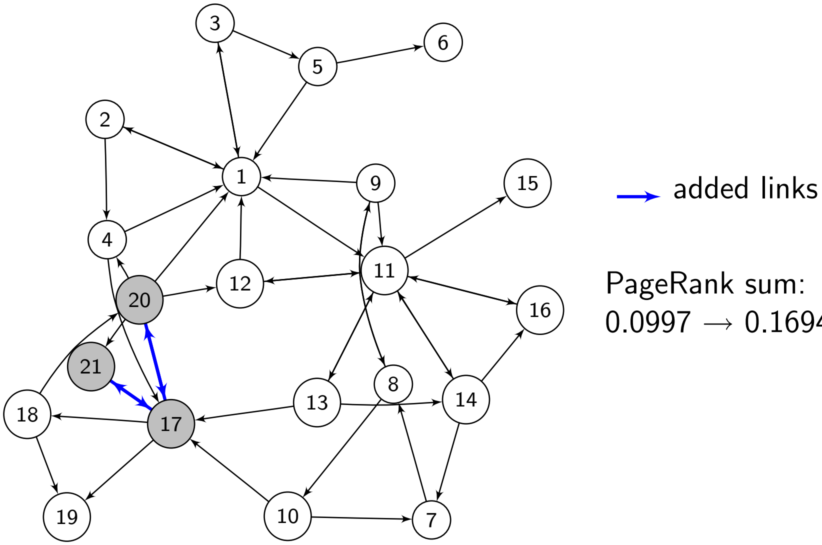
From the optimality criterion, we get:

Theorem (OF, Akian, Bouhtou, Gaubert)

If $\forall i$, $r_{i,j} = r_i$, there is a page to which **all controlled pages** *should link.*

$\longrightarrow$ [de Kerchove, Ninove and van Dooren, 2008]:
similar result with less flexible constraints

Web graph optimized for PageRank



PageRank sum:
$0.0997 \rightarrow 0.1694$

HITS optimization

- A choice of optional links gives an adjacency matrix A
- Authority vector: $\rho u = (A^T A + \xi e e^T) u = h^H(A) u$
- $f^H(u) = \sum_i r_i u_i^2, \quad \sum_i u_i^2 = 1$
- HITS optimization on weighted graphs: still nonconvex
- We abandon global optimality
- But efficient and scalable computation

Perron vector Optimization Problem

We study the following Perron vector Optimization Problem:

$$\begin{aligned} \max \quad & f(u(M)) \\ & M \in h(C) \end{aligned}$$

- $h : \mathbb{R}^m \to \mathbb{R}_+^{n \times n}$ and $f : \mathbb{R}_+^n \to \mathbb{R}$ are differentiable
- $u : \mathbb{R}_+^{n \times n} \to \mathbb{R}_+^n$ is the function that to an irreducible matrix associates its normalized Perron vector
- C convex, $h(C)$ a set of nonnegative irreducible matrices

NP-hardness

Consider the following problems with entries a linear function A and a vector b:

$$(\mathrm{P1})\quad \begin{array}{c} \min \rho(M) \\ A(M) \leq b\,,\ M \geq 0 \end{array} \qquad\qquad (\mathrm{P2})\quad \begin{array}{c} \min f(u(M)) \\ A(M) \leq b\,,\ M \geq 0 \end{array}$$

Proposition

The problems (P1) and (P2) are NP-hard problems.

Proof

LMP, proved NP-hard by Matsui, reduces to (P1) and (P2)

$$(\mathrm{LMP})\qquad \begin{array}{c} \min x_1 x_2 \\ Ax \leq b\,,\ x \geq 0 \end{array} \qquad\qquad M = \begin{pmatrix} 0 & 0 & 1 \\ x_1 & 0 & 0 \\ 0 & x_2 & 0 \end{pmatrix}$$

Derivative for Perron vector optimization

Computing $g_{ij} = \nabla f(u)^T \frac{\partial u}{\partial M_{ij}}$ is enough thanks to chain rule

Proposition (Deutsch and Neumann)

Denote $R = (M - \rho I)^{\#}$ the reduced resolvent at ρ and p the normalization factor at u.

$$\frac{\partial u}{\partial M_{ij}}(M) = -Re_i u_j + (p^T Re_i u_j)u$$

Corollary

Let $w^T = (-\nabla f^T + (\nabla f \cdot u)p^T)R$, then

$$g_{ij} = \nabla f^T \frac{\partial u}{\partial M_{ij}}(M) = w_i u_j$$

Iterative scheme for the derivative

Proposition

Let M be a primitive matrix with Perron vectors u and v.
We denote $\tilde{M} = \frac{1}{\rho} M$, $P = uv$ and $z = \frac{1}{\rho}(-\nabla f^T + (\nabla f \cdot u) p^T)$

Then the fix point scheme defined by

$$\forall k \in \mathbb{N}, \quad w_{k+1} = (z + w_k \tilde{M})(I - P)$$

converges in geometric speed to

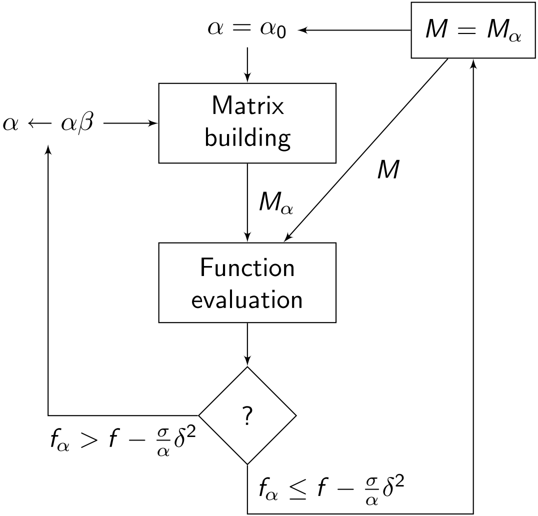
$$w = (-\nabla f^T + (\nabla f \cdot u) p^T)(M - \rho I)^{\#}$$

Descent algorithm

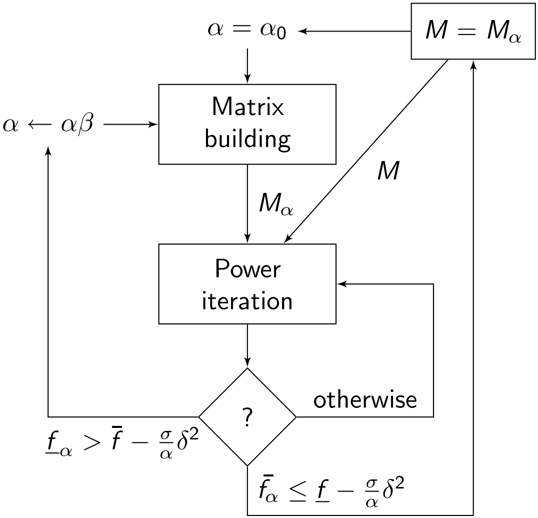
- First order descent method
- Criterion to go to next gradient step before convergence of power iterations
- Computation of value and derivative all together:

$$
\begin{aligned}
u_{k+1} &= \frac{Mu_k}{N(Mu_k)} \\
v_{k+1} &= \frac{v_k M}{v_k M u_{k+1}} \\
z_k &= \frac{1}{\rho_k}(-\nabla f(u_k)^T + (\nabla f(u_k) \cdot u_k) p_k^T) \\
w_{k+1} &= (z_k + \frac{1}{\rho_k} w_k M)(I - u_{k+1} v_{k+1})
\end{aligned}
$$

Coupling power and gradient iterations



Coupling power and gradient iterations



Convergence

Proposition

"Power + gradient" algorithm converges to a stationary point.

- Efficiency depends on the quality of the bounds in the Armijo test
- For HITS,

$$\underline{f} = f(u_k) - \|r\|_\infty \|u - u_k\|_2^2 - 2\|r \circ u_k\|_2 \|u - u_k\|_2$$
$$\overline{f} = f(u_k) + \|r\|_\infty \|u - u_k\|_2^2 + 2\|r \circ u_k\|_2 \|u - u_k\|_2$$
$$\underline{f} \le f(u) \le \overline{f}$$

Bounds

Lemma

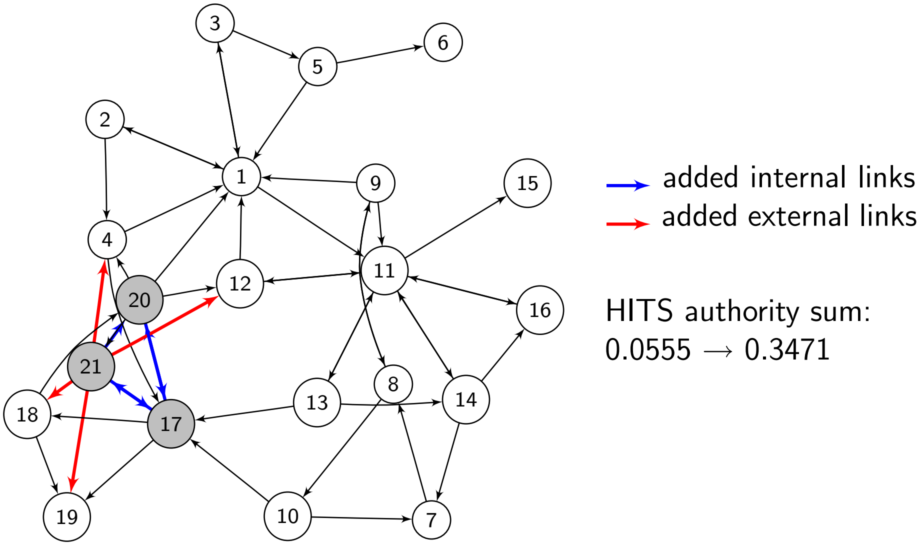
M nonnegative, symmetric matrix, $Mu = \rho u$

$u^0 > 0,\ \|u^0\| = 1,\ u^1 = \frac{Mu^0}{\|Mu^0\|},\ a = \frac{(\max_i \frac{(Mu^0)_i}{u^0_i})^2 - \|Mu^0\|^2}{\|Mu^0\|^2 - \sigma_2^2}$

$$\frac{\|u - u^1\|^2}{\|u^1 - u^0\|^2} \leq \frac{(\frac{\sigma_2}{\rho})^2}{(1 - \frac{\sigma_2}{\rho})^2 - \frac{13}{2}a}$$

Proof uses [Collatz, 1942]:
$M \geq 0,\ u > 0 \Rightarrow \min_i \frac{(Mu)_i}{u_i} \leq \rho(M) \leq \max_i \frac{(Mu)_i}{u_i}$

Web graph optimized for HITS



HITS authority sum:
$0.0555 \rightarrow 0.3471$

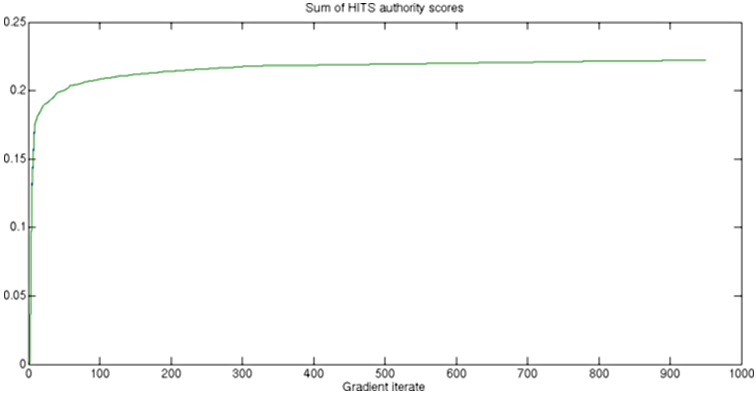
Experimental results

A fragment of the real web graph
with 1,500 pages (nodes) and 18,398 hyperlinks (arcs)
controlled web site has 49 pages

	PageRank	HITS authority
Method	Dynamic programming	Gradient algorithm
Initial value	0.0185	0.00000378
Value with clique	0.0693	0.0000169
Final value	0.0830	0.2266
Execution time	1.27 s	1600 s

Sequential Matlab code on an Intel Xeon CPU at 2.98 Ghz

Numerical experiment for HITS optimization



10,725 descent iterations, 10,776 matrix buildings,
1,737,677 power iterations, 0.21 reached in 150 iterations

Conclusion

- PageRank:
  - Global optimality, discrete solution
  - Very fast and scalable algorithm
  - Qualitative results on the optimal solutions

- HITS authority:
  - No global optimality result
  - No qualitative results on the optimal solutions
  - Non discrete local optimal solutions
  - Efficient and scalable descent algorithm
- Application in population dynamics

  with Billy, Clairambault, Gaubert and Lepoutre