Numerical Methods for Rapid Computation of PageRank

Gene H. Golub

Stanford University
Stanford, CA
USA

Joint work with Chen Greif
Outline

1 Markov Chains and PageRank
   • Definition

2 Acceleration Techniques
   • Sequence extrapolation
   • Adaptive Computation
   • Other Techniques

3 Arnoldi Based Methods
   • A refined Arnoldi algorithm
   • Sensitivity
   • Numerical experiments
1 Markov Chains and PageRank
   • Definition

2 Acceleration Techniques
   • Sequence extrapolation
   • Adaptive Computation
   • Other Techniques

3 Arnoldi Based Methods
   • A refined Arnoldi algorithm
   • Sensitivity
   • Numerical experiments
Stationary Distribution Vector of a Transition Probability Matrix

We are seeking a row vector $\pi^T$ that satisfies $\pi^T = \pi^T P$ where $P$ is a square stochastic matrix, with nonnegative entries between 0 and 1, and $Pe = e$, where $e$ is a vector of all-ones.

**Theorem**

*Perron(1907)-Frobenius(1912)*: A nonnegative irreducible matrix has a simple real eigenvalue equal to its spectral radius, whose associated eigenvector is a vector all of whose entries are nonnegative.

What happens when $P$ is stochastic and possibly reducible?
What Is PageRank?

**Definition**
Given a Webpage database, the PageRank of the $i$th Webpage is the $i$th element $\pi_i$ of the stationary distribution vector $\pi$ that satisfies $\pi^T P = \pi^T$, where $P$ is a matrix of weights of webpages that indicate their importance.
**What Is PageRank?**

**Definition**
Given a Webpage database, the PageRank of the $i$th Webpage is the $i$th element $\pi_i$ of the stationary distribution vector $\pi$ that satisfies $\pi^T P = \pi^T$, where $P$ is a matrix of weights of webpages that indicate their importance.

**Difficulties**
1. $P$ is too large (size possibly in the billions) for forming any of our favorite decompositions.
2. $P$ could be reducible, contain zero rows, and other difficulties of this sort.
What Is PageRank?

Definition

Given a Webpage database, the PageRank of the \( i \)th Webpage is the \( i \)th element \( \pi_i \) of the stationary distribution vector \( \pi \) that satisfies \( \pi^T P = \pi^T \), where \( P \) is a matrix of weights of webpages that indicate their importance.

Difficulties

1. \( P \) is too large (size possibly in the billions) for forming any of our favorite decompositions.
2. \( P \) could be reducible, contain zero rows, and other difficulties of this sort.

How do we modify \( P \) so that there is a unique solution?
The fundamental idea of Brin & Page: Importance of a webpage is determined not by its contents but rather by which pages link to it. Apply the power method to a web link graph.
**Some issues with web link graphs**

<table>
<thead>
<tr>
<th>Difficulties</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>The existence of <strong>dangling nodes</strong> (correspond to an all-zero row in the matrix): could have very important pages that have no outlinks. (e.g. the U.S. constitution!)</td>
</tr>
<tr>
<td>2</td>
<td><strong>Periodicity</strong>: a cyclic path in the Webgraph. (e.g. You point only to your mom’s webpage and she points only to yours.)</td>
</tr>
</tbody>
</table>

Simple example:

\[ P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \]
Some issues with web link graphs

**Difficulties**

1. **The existence of dangling nodes**: correspond to an all-zero row in the matrix: could have very important pages that have no outlinks. (e.g. the U.S. constitution!)

2. **Periodicity**: a cyclic path in the Webgraph. (e.g. You point only to your mom’s webpage and she points only to yours.)

Simple example:

\[ P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \]

**Solution**

Set \( M(c) = cP + (1 - c)E \), where \( E \) is a positive rank-1 matrix.
The matrix $M(c)$

- We have $M(c) > 0$ which yields a unique solution. But what is the significance of the stationary probability vector?
- $M(c)$ is a Markov chain with positive entries, and

$$M(c)z(c) = z(c).$$

Therefore for $c < 1$, $z(c)$ is unique (under proper scaling).
For the identity matrix, $P = I$, no unique stationary probability distribution, but for $M(c) = cl + (1 - c)ee^T/n$ we are converging to

$$z(c) = \frac{1}{n} e.$$
The significance of the parameter \( c \)

- \( c \) is the probability that a surfer will follow an outlink (as opposed to jump randomly to another Webpage).
- \( c = 0.85 \) was the choice in the Brin & Page model.
- Like regularization: small value leads to a more stable computation, but further away from true solution.
For Google, it all boiled down originally to solving the eigenvalue problem

\[ x = Mx \]

using the power method

\[ x^{(k+1)} = Mx^{(k)} . \]
Let $Mz_i = \lambda_i z_i$. For $|\lambda_i| \neq |\lambda_j|$ we have

$$x^{(0)} = \sum \alpha_i z_i,$$

and

$$x^{(k)} = \sum \alpha_i \lambda_i^k z_i,$$

with $\|x^{(k)}\|_1 = 1$ and $x \geq 0$.

After normalization, for $\lambda_1 = 1$ we have

$$x^{(k)} = z_1 + \sum_{j=2}^{n} \beta_j \lambda_j^k z_j.$$
The Eigenvalues of the PageRank Matrix

**Theorem**

*(Elegant proof due to Eldén)*

Let $P$ be a column-stochastic matrix with eigenvalues \{1, \lambda_2, \lambda_3, \ldots, \lambda_n\}. Then the eigenvalues of $M(c) = cP + (1 - c)ve^T$, where $0 < c < 1$ and $v$ is a nonnegative vector with $e^Tv = 1$, are

\[
\{1, c\lambda_2, c\lambda_3, \ldots, c\lambda_n\}.
\]

This implies

\[
\frac{|\lambda_j|}{|\lambda_1|} \leq c.
\]
Outline

1 Markov Chains and PageRank
   - Definition

2 Acceleration Techniques
   - Sequence extrapolation
   - Adaptive Computation
   - Other Techniques

3 Arnoldi Based Methods
   - A refined Arnoldi algorithm
   - Sensitivity
   - Numerical experiments
Quadratic Extrapolation
(Kamvar, Haveliwala, Manning, G.)

Slowly convergent series can be replaced by series that converge to the same limit at a much faster rate.

**Idea:** Estimate components of current iterate in the directions of second and third eigenvectors, and eliminate them.
Suppose $M$ has three distinct eigenvalues. The minimal polynomial is given by

$$P_M(\lambda) = \gamma_0 + \gamma_1 \lambda + \gamma_2 \lambda^2 + \gamma_3 \lambda^3.$$  

By the Cayley-Hamilton theorem, $P_M(M) = 0$. Hence for any vector $z$,

$$P_M(M)z = (\gamma_0 + \gamma_1 M + \gamma_2 M^2 + \gamma_3 M^3)z = 0.$$
Set $z = x^{(k-3)}$ and use the fact that $x^{(k-2)} = Mx^{(k-3)}$ and so on. Thus,

$$(x^{(k-2)} - x^{(k-3)})\gamma_1 + (x^{(k-1)} - x^{(k-3)})\gamma_2 + (x^{(k)} - x^{(k-3)})\gamma_3 = 0.$$ 

Defining

$$y^{(k-j)} = x^{(k-j)} - x^{(k-3)}, \quad j = 1, 2, 3,$$

and setting $\gamma_3 = 1$ (to avoid getting a trivial solution $\gamma = \mathbf{0}$), get

$$(y^{(k-2)} y^{(k-1)})[\gamma_1 \quad \gamma_2]^T = -y^{(k)}.$$

Now, since $M$ has more than three eigenvalues, solve a least squares problem.
The dynamic nature of the web

This problem involves a matrix which is changing over time.

- States increase and decrease, i.e. new websites are introduced and old websites die.
- Websites are continually changing. $M$ is a function of time and so is its dimension.
Adaptive Computation
(joint with Kamvar and Haveliwala)

Most pages converge rapidly. Basic idea: when the PageRank of a page has converged, stop recomputing it.

\[
\begin{align*}
    x_N^{(k+1)} &= M_N x^{(k)}; \\
    x_C^{(k+1)} &= x_C^{(k)}. 
\end{align*}
\]

- Use the previous vector as a start vector.
- Nice speedup, but not great. Why? The old pages converge quickly, but the new pages still take long to converge. Web constantly changes! Addition, deletion, change of existing pages...
- But, if you use Adaptive PageRank, you save the computation of the old pages.
Example: Stanford-Berkeley, $n \approx 700000$
Other Effective Approaches

- Aggregation/Disaggregation. (Stewart, Langville & Meyer, .....)
- Approaches related to permutations of the Google matrix. (Del Corso et. al., Kamvar et. al.)
- Linear system formulation. (Arasu et. al.)

and more...

Survey paper:
*A survey of eigenvector methods of Web information retrieval*
by Amy Langville and Carl Meyer.

*Stability and convergence analysis:* Ipsen & Kirkland.
Outline

1. Markov Chains and PageRank
   - Definition

2. Acceleration Techniques
   - Sequence extrapolation
   - Adaptive Computation
   - Other Techniques

3. Arnoldi Based Methods
   - A refined Arnoldi algorithm
   - Sensitivity
   - Numerical experiments
Arnoldi method:
The Arnoldi method is generally used for generating a small upper Hessenberg that approximates some of the eigenvalues of the original matrix. When $Q$ is orthogonal,

$$Q^T MQ(Q^T x) = \lambda(Q^T x).$$

1. Find $H = Q^T MQ$ upper Hessenberg, then perform the computations for $H$ instead of $M$.

2. $M$ is $n$-by-$n$ and is huge, but we terminate the process after $k$ steps. Resulting $H$ is $(k + 1)$-by-$k$. 

Using the Arnoldi method for PageRank (joint with Chen Greif)
Computational Cost

1. Main cost: One matrix-vector product (with original large matrix) per iteration.
2. Inner products and norm computations.
3. Power method cheaper but not by much if matrix-vector products dominate.
Similar to computing *refined Ritz vectors* (Jia, Stewart), but pretend largest eigenvalue stays 1 in smaller space, i.e. we do not compute any Ritz values.

Set initial guess $q$ and $k$, the Arnoldi steps number

Repeat

.....$[Q, H] = \text{Arnoldi}(A, q, k)$

.....Compute $H - [I; 0] = U\Sigma V^T$

.....Set $v = V(:, k)$

.....Set $q = Qv$

Until $\sigma_{\text{min}}(H - [I; 0]) < \varepsilon$
Advantages

- Orthogonalization achieves effective separation of eigenvectors.
- Take advantage of knowing the largest eigenvalue.
- Largest Ritz value could be complex, but if we set the shift to 1 then no risk of complex arithmetic.
- Smallest singular value converges smoothly to zero (more smoothly than largest Ritz value converges to 1).
- Stopping criterion with no computational overhead:

\[ \| Aq - q \|_2 = \sigma_{\min}(H - [I; 0]). \]
Disadvantages

- More complicated to implement.
- A single iteration is more expensive than a power iteration; must converge within fewer iterations.
Sensitivity of the PageRank Vector

\[ M(c) = cP + (1 - c)ev^T; \quad e = [1, \ldots, 1]^T, \quad v = \frac{e}{n}. \]

\[ M(c)x(c) = x(c); \]

\[ M'x + Mx' = x'; \]

\[ M' = P - ev^T = \frac{1}{c}(M - ev^T); \]

\[ (I - M)x' = M'x = \frac{1}{c}(x - v). \]

Get the exact same matrix, \( I - M \): singular consistent linear system. Goal: identify ‘sensitive’ vs. ‘insensitive’ components. Difficulty: How do we compute it?
<table>
<thead>
<tr>
<th>name</th>
<th>size</th>
<th>nz</th>
<th>avg nz per row</th>
</tr>
</thead>
<tbody>
<tr>
<td>sg</td>
<td>3,685</td>
<td>32,445</td>
<td>8.8</td>
</tr>
<tr>
<td>bs</td>
<td>19,566</td>
<td>133,535</td>
<td>6.8</td>
</tr>
<tr>
<td>Stanford</td>
<td>281,903</td>
<td>2,312,497</td>
<td>8.2</td>
</tr>
<tr>
<td>Stanford-Berkeley</td>
<td>683,446</td>
<td>7,583,376</td>
<td>11.1</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>1,104,857</td>
<td>18,265,794</td>
<td>16.5</td>
</tr>
<tr>
<td>edu</td>
<td>2,024,716</td>
<td>14,056,641</td>
<td>6.9</td>
</tr>
</tbody>
</table>

Thanks for David Gleich and Yahoo! Inc.
Effect of the damping factor $c$

Typical behavior for the test matrices: difference in convergence rate is significant.
Matrix-vector products for various values of the damping factor $c$, for the $281903 \times 281903$ Stanford matrix. The stopping criterion was $\|x^{(k)} - Ax^{(k)}\|_1 < 10^{-7}$.
Ordering is a function of $c$ (a few rankings in Wikipedia)

<table>
<thead>
<tr>
<th>Entry</th>
<th>$c = 0.85$</th>
<th>$c = 0.90$</th>
<th>$c = 0.95$</th>
<th>$c = 0.99$</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Race (U.S. Census)</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>20</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>France</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>2005</td>
<td>5</td>
<td>5</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>2004</td>
<td>6</td>
<td>6</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>2000</td>
<td>7</td>
<td>15</td>
<td>20</td>
<td>29</td>
</tr>
<tr>
<td>Canada</td>
<td>8</td>
<td>10</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>Category: culture</td>
<td>12</td>
<td>9</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>Category: politics</td>
<td>13</td>
<td>7</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>Category: wikiportals</td>
<td>18</td>
<td>8</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Italy</td>
<td>28</td>
<td>27</td>
<td>31</td>
<td>40</td>
</tr>
<tr>
<td>Sweden</td>
<td>80</td>
<td>92</td>
<td>94</td>
<td>100</td>
</tr>
</tbody>
</table>
Observations

- Top ranked entry stays on top throughout.
- Countries generally lose ground as $c$ goes up; categories make gains.
- Second ranked entry for $c = 0.85$ [Race (U.S. census)] is ranked 20th for $c = 0.99$.
- On the other hand 18th ranked entry for $c = 0.85$ is ranked third for $c = 0.99$. [Wikiportals are pages functioning as a portal for a particular subject area.]
- Other entries also show change in ranking as a function of the damping factor.
Changes at the top as a function of $c$

<table>
<thead>
<tr>
<th>$c$</th>
<th>top 5</th>
<th>top 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.85</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>0.90</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>0.95</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>0.99</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Match of webpages in the top rankings. The top 5 (first column) and top 10 (second column) pages for $c = 0.85$ were taken, and in the table the numbers indicate how many of them appear in the top 5 and 10 for other values of $c$. 
Summary

- Decomposition-free methods are necessary.
- Techniques for convergence acceleration prove effective.
- For $c = 0.85$ power method seems good enough, but not for higher values of $c$.
- Arnoldi approach seems a natural way to go and proves effective.

Challenges

- How to determine the reliability of PageRank by means of sensitivity.
- Efficient methods for a large value of $c$. 