Markov Anniversary Meeting
Charleston, SC
Work in Progress Session

2 PM, June 13, 2006

Gaussian Elimination for the Google PageRank Problem: Insights and Limitations

James Crowson

Computer Science, North Carolina State University

Partially supported by NSF CCR-0204157 (REU)

PageRank Is a Markov Process

- Estimates traffic to websites based on the Web's link structure.
- Models Web surfers with the following behavior:
  - 85% chance: pick a random link from the current page and follow it.
  - 15% chance: jump to a completely random page. This ensures that the resulting matrix will be irreducible.
- This is a Markov process; compute the stationary distribution of Web surfers.

Classical Google PageRank Uses the Power Method

π: The stationary distribution of Web surfers

α: The chance of following a link from the current page (0.85)

$\bar{P}$: The stochastic transition matrix

n: The number of nodes

$\bar{\bar{P}}$: $(\alpha\bar{P} + \frac{(1-\alpha)ee^T}{n})$

- $\pi^T = \pi^T \bar{\bar{P}}$. Compute π using sparse Power Method; the full matrix $\bar{\bar{P}}$ need not be formed.

The Stationary Distribution Can Be Calculated With Gaussian Elimination

- Gaussian Elimination first suggested by Funderlic and Mankin (1981) for stochastic and irreducible matrices.
- Since $\pi^T = \pi^T \bar{\bar{P}}$, $(\bar{\bar{P}}^T - I)\pi = 0$.
- U is rank $n - 1$, since $u_{n,n}$ is inevitably 0.
- Find the 1-dimensional null space of $\bar{\bar{P}}^T - I$ with Gaussian elimination followed by back substitution. L is not needed.
- $\bar{\bar{P}}^T - I$ is diagonally dominant, so no pivoting is needed.

Preserving Sparsity with Bordering

- Since $\bar{\bar{P}} = \alpha \bar{P} + ((1-\alpha)e)(\frac{e}{n})^T$, we can instead force irreducibility by bordering $\bar{P}$ (see handout for details):

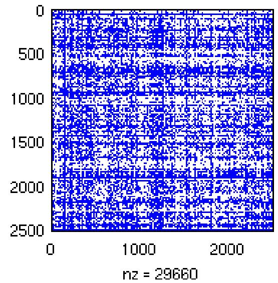
$$\begin{bmatrix} \pi^T & \omega \end{bmatrix} \begin{bmatrix} \bar{P} & (1-\alpha)e \\ (\frac{e}{n})^T & 0 \end{bmatrix} = \begin{bmatrix} \pi^T & \omega \end{bmatrix}.$$

- $\bar{P}$ is sparse, so the bordered matrix is sparse. Proceed with sparse Gaussian elimination.
- Nodes without outgoing links (dangling nodes) do not contribute to fill-in.

Taming Fill-in For Gaussian Elimination

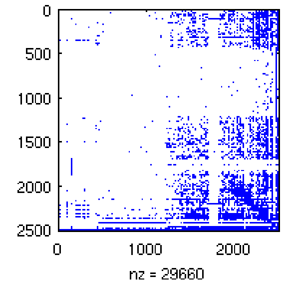
- During Gaussian Elimination, rows and columns only cause fill-in below and to their right, so placing denser rows and columns last tends to decrease fill-in.
- Reorder nodes according to the product of the number of out-links and the number of in-links, ascending.
- Collected links from 2,500 Web pages by breadth-first traversal from www.ncsu.edu.
- On this dataset: without reordering, storage requirements increased by a factor of 29 during Gaussian Elimination; with reordering, storage requirements did not increase.

How Well Does Reordering Work?



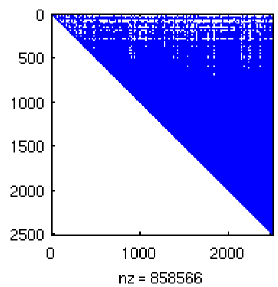
NCSU dataset before reordering

(nnz=29660)

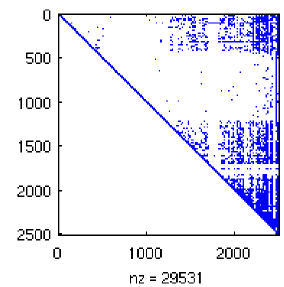


NCSU dataset after reordering

(nnz=29660)



Triangularized NCSU dataset without reordering (nnz=858566)



Triangularized NCSU dataset with reordering (nnz=29531)

Further Work

- Thresholding (discarding small values in U)
- Parallel Sparse Gaussian Elimination (triangularization)
- Updating π without doing Gaussian Elimination again

Results and Conclusions

- It is practical to compute PageRank using Gaussian Elimination if the matrix fits into memory.
- Performance is variable; real Web data produces very little fill-in and performance is within an order of magnitude of the Power Method.
- Unlike the Power Method, with Gaussian Elimination the parameter α may be increased without impact on performance.

Bibliography

- Solution of Homogeneous Systems of Equations Arising from Compartmental Models (R.E. Funderlic and J.B. Mankin) SIAM J. Scientific and Statistical Computations (2) 375–383 (1981)
- Sensitivity of the Stationary Distribution Vector for an Ergodic Markov Chain (R.E. Funderlic and C.D. Meyer, Jr.) Linear Algebra and Its Applications (76) 1–17 (1986)
- Updating LU Factorizations for Computing Stationary Distributions, (R.E. Funderlic and R.J. Plemmons), SIAM Journal Algebraic and Discrete Methods (7), 30–42 (1986)
- Deeper Inside PageRank, (A. N. Langville and C.D. Meyer), Internet Mathematics (1), 335–400 (2004)