EE5137 Lecture 9: Motivating Markov Chains: The PageRank Algorithm

Vincent Y. F. Tan

NUS
National University of Singapore

Department of Electrical and Computer Engineering,
Department of Mathematics,
National University of Singapore

Mar 2021

Web Search Before Google

- Internet searching in the 1990s was very inefficient.

Web Search Before Google

- Internet searching in the 1990s was very inefficient.
- Yahoo or AltaVista would scan pages for your search text, and simply list the results with the most occurrences of those words.

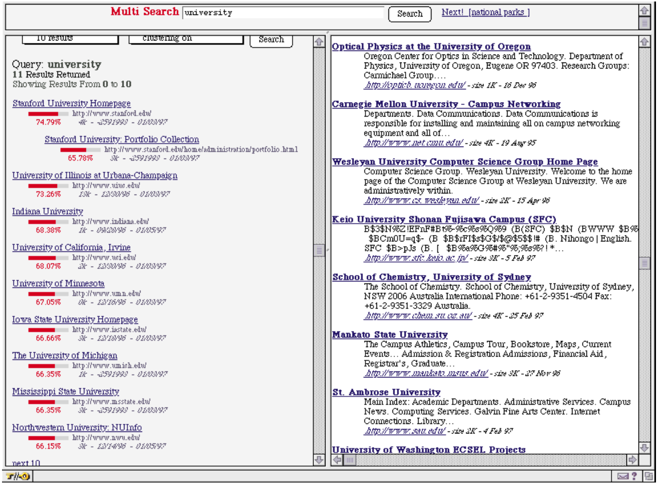
Web Search Before Google

- Internet searching in the 1990s was very inefficient.
- Yahoo or AltaVista would scan pages for your search text, and simply list the results with the most occurrences of those words.
- So if you profess to be an expert in Markov chains, you don't have to write any papers on Markov chains.

Web Search Before Google

- Internet searching in the 1990s was very inefficient.
- Yahoo or AltaVista would scan pages for your search text, and simply list the results with the most occurrences of those words.
- So if you profess to be an expert in Markov chains, you don't have to write any papers on Markov chains.
- Just put 10^6 occurrences of "Markov chains" on your website!

Web Search Before Google



Taken from Page et al. (1999), “The PageRank Citation Ranking: Bringing Order to the Web”

Pagerank Led to Google

- Larry Page and Sergey Brin invented a way to rank pages by their importance.

Pagerank Led to Google

- Larry Page and Sergey Brin invented a way to rank pages by their importance.
- This led to

Pagerank Led to Google

- Larry Page and Sergey Brin invented a way to rank pages by their importance.
- This led to

Google

- Each web page i has an associated importance, or score r_i. This is a positive number.

Intuition Behind Pagerank

- The importance rule: If a page P links to m other pages $Q_1, Q_2, \ldots, Q_m$ then each page Q_i inherits $1/m$ of P's importance.

Intuition Behind Pagerank

- The importance rule: If a page P links to m other pages $Q_1, Q_2, \ldots, Q_m$ then each page Q_i inherits $1/m$ of P's importance.
- In practice, this means:
  - If a very important page links to your page (and not to a billion other ones as well), then your page is considered important.
  - If a billion unimportant pages link to your page, then your page is still important.
  - If only one unknown page links to yours, your page is not important.

Intuition Behind Pagerank

- The importance rule: If a page P links to m other pages $Q_1, Q_2, \ldots, Q_m$ then each page Q_i inherits $1/m$ of P's importance.
- In practice, this means:
  - If a very important page links to your page (and not to a billion other ones as well), then your page is considered important.
  - If a billion unimportant pages link to your page, then your page is still important.
  - If only one unknown page links to yours, your page is not important.
- Random surfer interpretation:
  - A "random surfer" just sits at his computer all day, randomly clicking on links.
  - The pages he spends the most time on should be the most important.
  - Important pages are those where a random surfer will end up most often. This measure turns out to be equivalent to the score.

Importance Matrix

- Consider an internet with n pages. The importance matrix is the $n \times n$ matrix whose (i, j)-entry is the importance that page i passes to page j.

Importance Matrix

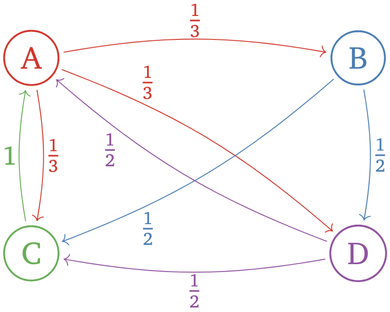
- Consider an internet with n pages. The importance matrix is the $n \times n$ matrix whose (i, j)-entry is the importance that page i passes to page j.
- Observe that the importance matrix is a row stochastic matrix, assuming every page contains a link.

Importance Matrix

- Consider an internet with n pages. The importance matrix is the $n \times n$ matrix whose (i, j)-entry is the importance that page i passes to page j.
- Observe that the importance matrix is a row stochastic matrix, assuming every page contains a link.
- If page i has m outgoing links, then the i-th row contains the number $1/m$ a total of m times, and the number zero in the other entries.

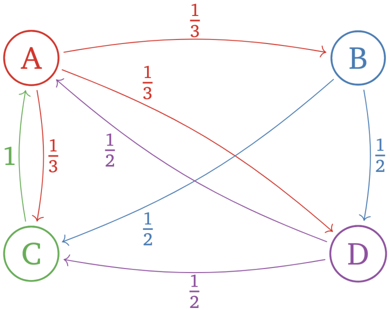
Example of Importance Matrix

Internet with only 4 pages.



Example of Importance Matrix

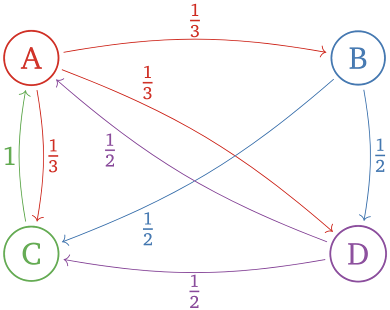
Internet with only 4 pages.



- Page A has 3 links, so it passes $1/3$ of its impt to pages B, C, D;

Example of Importance Matrix

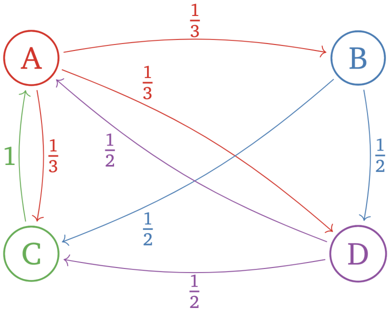
Internet with only 4 pages.



- Page A has 3 links, so it passes $1/3$ of its impt to pages B, C, D;
- Page B has 2 links, so it passes $1/2$ of its impt to pages C, D;

Example of Importance Matrix

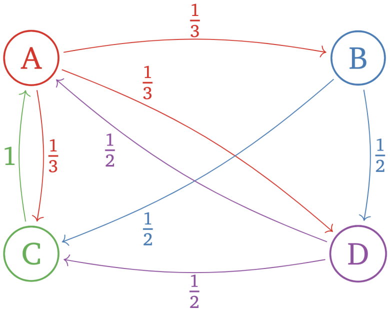
Internet with only 4 pages.



- Page A has 3 links, so it passes $1/3$ of its impt to pages B, C, D;
- Page B has 2 links, so it passes $1/2$ of its impt to pages C, D;
- Page C has one link, so it passes all of its impt to page A;

Example of Importance Matrix

Internet with only 4 pages.



- Page A has 3 links, so it passes $1/3$ of its impt to pages B, C, D;
- Page B has 2 links, so it passes $1/2$ of its impt to pages C, D;
- Page C has one link, so it passes all of its impt to page A;
- Page D has 2 links, so it passes $1/2$ of its impt to pages A, C.

Example of Importance Matrix

- Let $r = (r_1, r_2, r_3, r_4)$ be the vector of scores of pages A, B, C, D.

Example of Importance Matrix

- Let $r = (r_1, r_2, r_3, r_4)$ be the vector of scores of pages A, B, C, D.
- The importance matrix is

$$[Q] = \begin{bmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \end{bmatrix}$$

Example of Importance Matrix

- Let $r = (r_1, r_2, r_3, r_4)$ be the vector of scores of pages A, B, C, D.
- The importance matrix is

$$[Q] = \begin{bmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \end{bmatrix}$$

- By construction,

$$\begin{bmatrix} r_1 & r_2 & r_3 & r_4 \end{bmatrix} \begin{bmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/2 & 1/2 \\ 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \end{bmatrix} = \begin{bmatrix} r_1 & r_2 & r_3 & r_4 \end{bmatrix}$$

because

$$\begin{aligned} r_1 &= r_3 + r_4/2, & r_2 &= r_1/3, \\ r_3 &= r_1/3 + r_2/2 + r_4/2, & r_4 &= r_1/3 + r_2/2. \end{aligned}$$

Equality expresses the importance rule.

Key Property of Importance Matrix

- Let's look at the equality again

$$r[Q] = r$$

Key Property of Importance Matrix

- Let's look at the equality again

$$r[Q] = r$$

- This says that r is a left-eigenvector of $[Q]$ with eigenvalue 1.

Key Property of Importance Matrix

- Let's look at the equality again

$$r[Q] = r$$

- This says that r is a left-eigenvector of $[Q]$ with eigenvalue 1.
- r is a steady-state vector of $[Q]$

Key Property of Importance Matrix

- Let's look at the equality again

$$r[Q] = r$$

- This says that r is a left-eigenvector of $[Q]$ with eigenvalue 1.
- r is a steady-state vector of $[Q]$
- So Google can construct $[Q]$ to rank webpages by learning r.

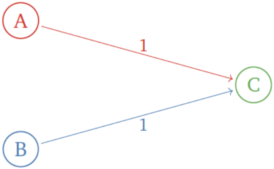
Key Property of Importance Matrix

- Let's look at the equality again

$$r[Q] = r$$

- This says that r is a left-eigenvector of $[Q]$ with eigenvalue 1.
- r is a steady-state vector of $[Q]$
- So Google can construct $[Q]$ to rank webpages by learning r.
- Unfortunately, in real-life there are problems with this approach, e.g., $[Q]$ need not be ergodic!

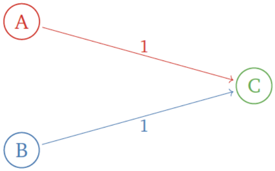
Problem 1: Page with No Links



- For the above internet with 3 pages, the importance matrix here is

$$[Q] = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

Problem 1: Page with No Links



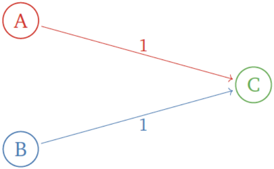
- For the above internet with 3 pages, the importance matrix here is

$$[Q] = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

- The characteristic polynomial is

$$\det([Q] - \lambda[I]) = \det\left(\begin{bmatrix} -\lambda & 0 & 1 \\ 0 & -\lambda & 1 \\ 0 & 0 & -\lambda \end{bmatrix}\right) = -\lambda^3$$

Problem 1: Page with No Links



- For the above internet with 3 pages, the importance matrix here is

$$[Q] = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

- The characteristic polynomial is

$$\det([Q] - \lambda[I]) = \det\left(\begin{bmatrix} -\lambda & 0 & 1 \\ 0 & -\lambda & 1 \\ 0 & 0 & -\lambda \end{bmatrix}\right) = -\lambda^3$$

- 1 is not an eigenvalue because $[Q]$ is not row stochastic!

Problem 2: Disconnected Internet

- The importance matrix is

$$[Q] = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1/2 & 1/2 & 0 \end{bmatrix}$$

- Two steady-state distributions

$$r^{(1)} = [1/2 \quad 1/2 \quad 0 \quad 0 \quad 0] \qquad r^{(2)} = [0 \quad 0 \quad 1/3 \quad 1/3 \quad 1/3]$$

Not ergodic! Two recurrent classes.

Page and Brin's Solution: Fix of First Problem

- Replace zero rows by adding a row of $1/n$'s, where n is the total number of pages

$$[Q'] = [Q] + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \quad \text{so} \quad [Q'] = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

Page and Brin's Solution: Fix of First Problem

- Replace zero rows by adding a row of $1/n$'s, where n is the total number of pages

$$[Q'] = [Q] + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 \end{bmatrix} \quad \text{so} \quad [Q'] = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}$$

- Now $[Q']$ is indeed row stochastic.

Page and Brin's Solution: Fix of Second Problem

- Fix a "damping factor" $p \in (0, 1)$.

Page and Brin's Solution: Fix of Second Problem

- Fix a "damping factor" $p \in (0, 1)$.
- The Google Matrix is

$$[P] = (1-p)[Q'] + p[B] \quad \text{where} \quad [B] = \frac{1}{n}\begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}$$

with $p = 0.15$.

Page and Brin's Solution: Fix of Second Problem

- Fix a "damping factor" $p \in (0, 1)$.
- The Google Matrix is

$$[P] = (1-p)[Q'] + p[B] \quad \text{where} \quad [B] = \frac{1}{n}\begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}$$

  with $p = 0.15$.
- In the random surfer interpretation, this matrix says:
  - With probability p, our surfer will surf to a completely random page;
  - Otherwise, he'll click a random link on the current page;
  - Unless the current page has no links, in which case he'll surf to a completely random page in either case.

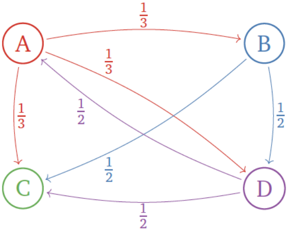
Page and Brin's Solution: Fix of Second Problem

- Fix a "damping factor" $p \in (0, 1)$.
- The Google Matrix is

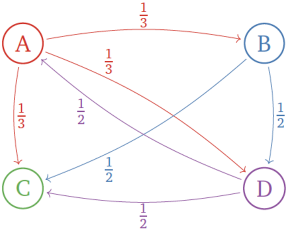
$$[P] = (1-p)[Q'] + p[B] \quad \text{where} \quad [B] = \frac{1}{n}\begin{bmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \dots & 1 \end{bmatrix}$$

  with $p = 0.15$.
- In the random surfer interpretation, this matrix says:
  - With probability p, our surfer will surf to a completely random page;
  - Otherwise, he'll click a random link on the current page;
  - Unless the current page has no links, in which case he'll surf to a completely random page in either case.
- $[P]$ is ergodic $\implies$ Has a unique steady-state vector r or π!

Page and Brin's Solution: Example



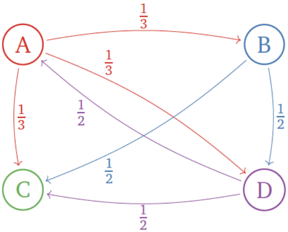
Page and Brin's Solution: Example

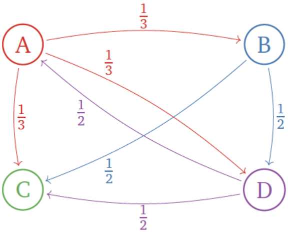


- Fix the first problem

$$[Q'] = \begin{bmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/2 & 1/2 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 & 0 \end{bmatrix}$$

Page and Brin's Solution: Example





- Fix the second problem

$$[P] = 0.85 \begin{bmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 1/2 & 1/2 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/2 & 0 & 1/2 & 0 \end{bmatrix} + 0.15 \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix}$$

$$\approx \begin{bmatrix} 0.0375 & 0.3208 & 0.3208 & 0.3208 \\ 0.0375 & 0.0375 & 0.4625 & 0.4625 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.4625 & 0.0375 & 0.4625 & 0.0375 \end{bmatrix}$$

Learned Score Vector

- The Google Matrix is

$$[P] = \begin{bmatrix} 0.0375 & 0.3208 & 0.3208 & 0.3208 \\ 0.0375 & 0.0375 & 0.4625 & 0.4625 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.4625 & 0.0375 & 0.4625 & 0.0375 \end{bmatrix}.$$

Learned Score Vector

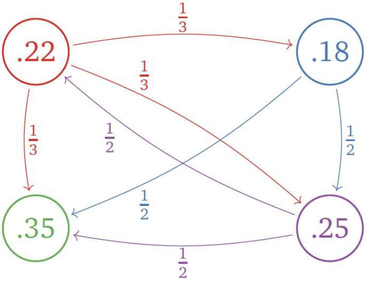
- The Google Matrix is

$$[P] = \begin{bmatrix} 0.0375 & 0.3208 & 0.3208 & 0.3208 \\ 0.0375 & 0.0375 & 0.4625 & 0.4625 \\ 0.25 & 0.25 & 0.25 & 0.25 \\ 0.4625 & 0.0375 & 0.4625 & 0.0375 \end{bmatrix}.$$

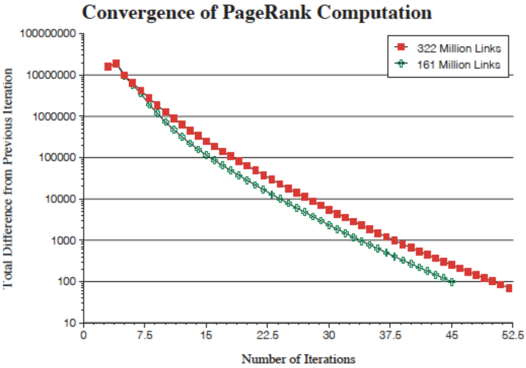
- The left-eigenvector with eigenvalue 1 is

Pagerank solution $\quad \pi = r = \begin{bmatrix} 0.2192 & 0.1752 & 0.3558 & 0.2498 \end{bmatrix}.$

- and the ranking of the webpages is C, D, A, B.

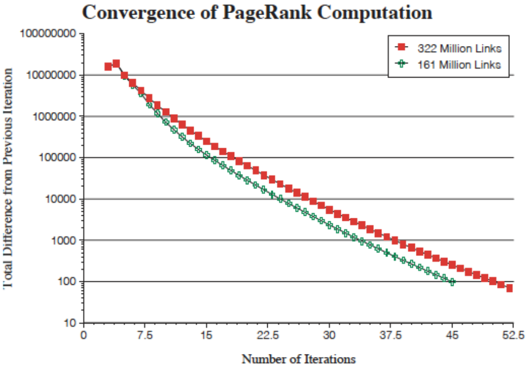


Rate of Convergence



Taken from “The PageRank Citation Ranking: Bringing Order to the Web”

Rate of Convergence



Taken from “The PageRank Citation Ranking: Bringing Order to the Web”

- Fast convergence due to large spectral gap – difference between largest eigenvalue 1 and $1 - p = 0.85$.

Moral of the Story

Google found the 25 billion dollar eigenvector.

Moral of the Story

Google found the 25 billion dollar eigenvector.

Moral of the Story

Google found the 25 billion dollar eigenvector.

"Beautiful math tends to be useful; useful things tend to have beautiful math." ... Statistics is often where it comes together.

References

- https://textbooks.math.gatech.edu/ila/stochastic-matrices.html
- http://statweb.stanford.edu/ tibs/sta306bfiles/pagerank/ryan/01-24-pr.pdf