Applying theory of Markov Chains to the problem of sports ranking.

A. Govan  C. Meyer

Department of Mathematics
North Carolina State University

AMS Southeastern Section Meeting, March 2007
Outline

Google’s ranking algorithm.

Ranking NFL.

Results and current work.
Basics of PageRank.

- Basic Idea: \[ r(P) = \sum_{Q \in B_P} \frac{r(Q)}{\text{deg}^{-}(Q)} \]

where \( r(P) \) is the rank of a webpage \( P \), \( B_P \) is the set of web pages pointing to \( P \), and \( \text{deg}^{-}(Q) \) is the outdegree of a webpage \( Q \).
Applying theory of Markov Chains to the problem of sports ranking.

Google’s ranking algorithm.

Web digraph.
Applying theory of Markov Chains to the problem of sports ranking.

Google's ranking algorithm.

**Web digraph adjacency matrix.**

WWW digraph is represented by an adjacency matrix $A$.

$$
A = \begin{pmatrix}
    P_1 & P_2 & P_3 & \cdots & P_n \\
    0 & 1 & 0 & \cdots & 1 \\
    0 & 0 & 0 & \cdots & 0 \\
    1 & 1 & 1 & \cdots & 0 \\
    \vdots & \vdots & \vdots & \ddots & \vdots \\
    1 & 0 & 1 & \cdots & 1 
\end{pmatrix}
$$
Applying theory of Markov Chains to the problem of sports ranking.

Google's ranking algorithm.

**Web digraph hyperlink matrix.**

\[
H = \begin{pmatrix}
0 & \frac{1}{\text{deg}^-(P_1)} & 0 & \cdots & \frac{1}{\text{deg}^-(P_1)} \\
\frac{1}{\text{deg}^-(P_2)} & 0 & \frac{1}{\text{deg}^-(P_2)} & \cdots & 0 \\
\frac{1}{\text{deg}^-(P_3)} & \frac{1}{\text{deg}^-(P_3)} & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{1}{\text{deg}^-(P_n)} & 0 & \frac{1}{\text{deg}^-(P_n)} & \cdots & \frac{1}{\text{deg}^-(P_n)}
\end{pmatrix}
\]
Applying theory of Markov Chains to the problem of sports ranking.

Google’s ranking algorithm.

PageRank problem statement.

Basic Idea: \( r(P) = \sum_{Q \in B_P} \frac{r(Q)}{deg^-(Q)} \)

Problem restated:
- \( \pi \) - vector containing the rank scores.
Applying theory of Markov Chains to the problem of sports ranking.

Google’s ranking algorithm.

PageRank problem statement.

- **Basic Idea:** \( r(P) = \sum_{Q \in B_P} \frac{r(Q)}{\text{deg}^-(Q)} \)

- **Problem restated:**
  - \( \pi \) - vector containing the rank scores.
  - \( \pi(0) \) - initial rank vector
Applying theory of Markov Chains to the problem of sports ranking.

Google's ranking algorithm.

PageRank problem statement.

- Basic Idea: \( r(P) = \sum_{Q \in B_P} \frac{r(Q)}{deg^-(Q)} \)

- Problem restated:
  - \( \pi \) - vector containing the rank scores.
  - \( \pi(0) \) - initial rank vector
  - \( \pi^T(k) = \pi^T(k - 1)H \)
Applying theory of Markov Chains to the problem of sports ranking.

Google’s ranking algorithm.

PageRank problem statement.

Basic Idea: \( r(P) = \sum_{Q \in B_P} \frac{r(Q)}{\text{deg}^-(Q)} \)

Problem restated:

\( \pi \) - vector containing the rank scores.
\( \pi(0) \) - initial rank vector
\( \pi^T(k) = \pi^T(k - 1)H \)
\( \pi^T(k) = \pi^T(0)H^k \)
Applying theory of Markov Chains to the problem of sports ranking.

Google’s ranking algorithm.

PageRank problem statement.

► Basic Idea: \( r(P) = \sum_{Q \in B_P} \frac{r(Q)}{\text{deg}^{-}(Q)} \)

► Problem restated:
  ► \( \pi \) - vector containing the rank scores.
  ► \( \pi(0) \) - initial rank vector
  ► \( \pi^T(k) = \pi^T(k - 1)H \)
  ► \( \pi^T(k) = \pi^T(0)H^k \)
  ► \( \pi^T(0)H^k \rightarrow \pi \) ?
Applying theory of Markov Chains to the problem of sports ranking.

Google’s ranking algorithm.

**Google matrix.**

- **Adjacency Matrix** $A$. 

- **Hyperlink Matrix** $H$. 

- **Stochastic matrix** $S$. 

- Replace the zero rows of $H$ with $(1/n)e^T$, where $e$ is a column vector of ones.

- **Google Matrix** $G$. 

- Convex combination: $G = \alpha S + (1 - \alpha)e^T$, where $\alpha \in (0, 1)$, $e^T > 0$ and $e^T e = 1$. 
Applying theory of Markov Chains to the problem of sports ranking.

Google’s ranking algorithm.

Google matrix.

- Adjacency Matrix $A$.
- Hyperlink Matrix $H$. 

$G = \alpha S + (1 - \alpha) e^T v^T$, where $\alpha \in (0, 1)$, $v^T > 0$ and $v^T e = 1$. 

$e$ is a column vector of ones.
Applying theory of Markov Chains to the problem of sports ranking.

Google's ranking algorithm.

Google matrix.

- Adjacency Matrix \( A \).
- Hyperlink Matrix \( H \).
- Stochastic matrix \( S \).
  - Replace the zero rows of \( H \) with \((1/n)e^T\), where \( e \) is a column vector of ones.
Google matrix.

- Adjacency Matrix $A$.
- Hyperlink Matrix $H$.
- Stochastic matrix $S$.
  - Replace the zero rows of $H$ with $(1/n)e^T$, where $e$ is a column vector of ones.
- Google Matrix $G$.
  - Convex combination: $G = \alpha S + (1 - \alpha)ev^T$, $\alpha \in (0, 1)$, $v^T > 0$ and $v^T e = 1$. 

Applying theory of Markov Chains to the problem of sports ranking.

Google's ranking algorithm.
Applying theory of Markov Chains to the problem of sports ranking.

Google’s ranking algorithm.

**PageRank vector** $\pi$.

- $G$ is the transition probability matrix.
Applying theory of Markov Chains to the problem of sports ranking.

Google's ranking algorithm.

PageRank vector $\pi$.

- $G$ is the transition probability matrix.
- $G$ is irreducible (and aperiodic).
Applying theory of Markov Chains to the problem of sports ranking.

Google’s ranking algorithm.

PageRank vector \( \pi \).

- \( G \) is the transition probability matrix.
- \( G \) is irreducible (and aperiodic).
- Markov Chains theory implies:
  \[ \pi^T(0)G^k \rightarrow \pi^T \]
  such that \( \pi^T = \pi^T G \)
Applying theory of Markov Chains to the problem of sports ranking.

Google’s ranking algorithm.

PageRank vector $\pi$.

- $G$ is the transition probability matrix.
- $G$ is irreducible (and aperiodic).
- Markov Chains theory implies:
  \[ \pi^T(0)G^k \to \pi^T \]
  such that $\pi^T = \pi^T G$
- $\pi$ is a unique probability distribution vector.
Applying theory of Markov Chains to the problem of sports ranking.

Google’s ranking algorithm.

**PageRank vector** \( \pi \).

- \( G \) is the transition probability matrix.
- \( G \) is irreducible (and aperiodic).
- Markov Chains theory implies:
  \[ \pi^T(0)G^k \to \pi^T \]
  such that \( \pi^T = \pi^T G \)
- \( \pi \) is a unique probability distribution vector.
- \( \pi_i \) is the PageRank score of the web page \( i \).
Applying theory of Markov Chains to the problem of sports ranking.

Ranking NFL.

NFL weighted digraph.
Applying theory of Markov Chains to the problem of sports ranking.

Ranking NFL.

**NFL adjacency matrix.**

\[
A = \begin{pmatrix}
    0 & \cdots & 4 & 0 & 0 & 0 & 0 & \cdots \\
    \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
    0 & \cdots & 0 & 10 & 3 & 0 & 20 & \cdots \\
    0 & \cdots & 0 & 0 & 0 & 12 & 0 & \cdots \\
    \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
    0 & \cdots & 3 & 0 & 0 & 0 & 0 & 14 & \cdots \\
    \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
    0 & \cdots & 0 & 0 & 0 & 0 & 0 & \cdots \\
    \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
    0 & \cdots & 10 & 3 & 0 & 0 & 0 & \cdots \\
    \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\end{pmatrix}
\]
Applying theory of Markov Chains to the problem of sports ranking.

Ranking NFL.

GeM (Generalized Markov Method).

- Adjacency matrix $A$.
- Hyperlink matrix $H(i, j) = \sum_t w_{ij}^t / (\sum_j (\sum_t w_{ij}^t))$
  where $w_{ij}^t$ is the weight on the edge from team $i$ to team $j$ during week $t$.
- Stochastic matrix $S$, dealing with undefeated teams.
- GeM matrix $G = \alpha_0 S + \alpha_1 \mathbf{e}_1 \mathbf{v}_1^T + \ldots + \alpha_k \mathbf{e}_k \mathbf{v}_k^T$
  where $k \geq 1$. 
Applying theory of Markov Chains to the problem of sports ranking.

Ranking NFL.

Feature vectors $v_1, \ldots, v_k$.

- Based on the statistical data of the given season.
Applying theory of Markov Chains to the problem of sports ranking.

Ranking NFL.

Feature vectors $v_1, \ldots, v_k$.

- Based on the statistical data of the given season.
- Must be nonnegative.
Feature vectors $v_1, \ldots, v_k$.

- Based on the statistical data of the given season.
- Must be nonnegative.
- Problem: What statistical data corresponds the most to the performance?
Applying theory of Markov Chains to the problem of sports ranking.

Ranking NFL.

Feature vectors $v_1, \ldots, v_k$.

- Based on the statistical data of the given season.
- Must be nonnegative.
- Problem: What statistical data corresponds the most to the performance?
- Start with a matrix containing statistical data for a given season.
Applying theory of Markov Chains to the problem of sports ranking.

Ranking NFL.

Feature vectors $v_1, \ldots, v_k$.

- Based on the statistical data of the given season.
- Must be nonnegative.
- Problem: What statistical data corresponds the most to the performance?
- Start with a matrix containing statistical data for a given season.
- SVD $\rightarrow$ no guaranty on nonnegativity.
Applying theory of Markov Chains to the problem of sports ranking.

Ranking NFL.

**Feature vectors** $v_1, \ldots, v_k$.

- Based on the statistical data of the given season.
- Must be nonnegative.
- Problem: What statistical data corresponds the most to the performance?
- Start with a matrix containing statistical data for a given season.
- SVD $\rightarrow$ no guaranty on nonnegativity.
- NMF (nonnegative matrix factorization)
Feature vectors via NMF

Nonnegative matrix factorization: Given $M_{m \times n} \geq 0$,

$$M = WH$$

such that $W \geq 0$, and $H \geq 0$

$$M_j = \sum W_i h_{ij}$$

Possible uses of NMF:

- Given appropriate $M$ matrix (e.g. teams by stats) feature vectors could be the nonnegative “basis” of columns of $M$, i.e. columns of $W$. 

Applying theory of Markov Chains to the problem of sports ranking.

Ranking NFL.
Results.

GeM ranking method:

<table>
<thead>
<tr>
<th>Sorted Totals '06</th>
<th>Regular</th>
<th>Season</th>
<th>Playoffs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant</td>
<td>Games</td>
<td>Spread</td>
<td>Games</td>
</tr>
<tr>
<td>Colley Ranking</td>
<td>141</td>
<td>2035</td>
<td>11</td>
</tr>
<tr>
<td>Keener Ranking</td>
<td>130</td>
<td>2058</td>
<td>7</td>
</tr>
<tr>
<td>GeM Ranking</td>
<td>130</td>
<td>2246</td>
<td>6</td>
</tr>
<tr>
<td>Govan, Vincent</td>
<td>112</td>
<td>2275</td>
<td>6</td>
</tr>
<tr>
<td>Meyer, Carl</td>
<td>111</td>
<td>2305</td>
<td>5</td>
</tr>
<tr>
<td>Meyer, Bud</td>
<td>110</td>
<td>2325</td>
<td>6</td>
</tr>
<tr>
<td>Kelley, Tim</td>
<td>109</td>
<td>2613</td>
<td>3</td>
</tr>
<tr>
<td>Koh, Gil</td>
<td>106</td>
<td>2039</td>
<td>9</td>
</tr>
<tr>
<td>Glantz-Culver Line</td>
<td>105</td>
<td>2010.4</td>
<td>9</td>
</tr>
<tr>
<td>Rose, Nick</td>
<td>101</td>
<td>2070</td>
<td>3</td>
</tr>
<tr>
<td>Albright, Russ</td>
<td>90</td>
<td>1996</td>
<td>7</td>
</tr>
<tr>
<td>Meyer, Becky</td>
<td>88</td>
<td>1991</td>
<td>8</td>
</tr>
<tr>
<td>Stitzinger, Ernie</td>
<td>83</td>
<td>1886</td>
<td>7</td>
</tr>
<tr>
<td>Massey Ranking</td>
<td>82</td>
<td>1761</td>
<td>7</td>
</tr>
<tr>
<td>Kenney, Holly</td>
<td>69</td>
<td>1410</td>
<td>5</td>
</tr>
<tr>
<td>Kenney, Sean</td>
<td>63</td>
<td>1068</td>
<td>6</td>
</tr>
<tr>
<td>Meyer, Marty</td>
<td>16</td>
<td>316</td>
<td>0</td>
</tr>
<tr>
<td>Laake, Kevin</td>
<td>12</td>
<td>217</td>
<td>0</td>
</tr>
<tr>
<td>Fauntleroy, Amassa</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>
Results.

GeM ranking method:

- (without first two weeks) Basic GeM predicts 70% of the games played correctly during 2004 NFL regular season.
Results.

GeM ranking method:

- (without first two weeks) Basic GeM predicts 70% of the games played correctly during 2004 NFL regular season.
- (without first two weeks) Basic GeM predicts 75.9% of the games played correctly during 2005 NFL regular season.
Results.

GeM ranking method:

- (without first two weeks) Basic GeM predicts 70\% of the games played correctly during 2004 NFL regular season.
- (without first two weeks) Basic GeM predicts 75.9\% of the games played correctly during 2005 NFL regular season.
- (without first two weeks) Basic GeM predicts 62\% of the games played correctly during 2006 NFL regular season.
Summary

- Expanding to bigger data set - NCAA men’s basketball.
- Experimenting with NMF to obtain feature vectors.
- Moving beyond sports (recommendation systems).