involve
a journal of mathematics

An application of Google's PageRank to NFL rankings

Laurie Zack, Ron Lamb and Sarah Ball

msp

2012 vol. 5, no. 4

INVOLVE 5:4 (2012)
dx.doi.org/10.2140/involve.2012.5.463

An application of Google's PageRank to NFL rankings

Laurie Zack, Ron Lamb and Sarah Ball

(Communicated by Charles R. Johnson)

We explain the PageRank algorithm and its application to the ranking of football teams via the GEM method. We then modify and extend the GEM method with the addition of more football statistics to look at the possibility of using this method to more accurately rank teams. Lastly, we compare both methods by aggregating each statistical ranking using the cross-entropy Monte Carlo algorithm.

1. Introduction

Over the last few decades, abundant research has been done in the mathematics of rankings. There are numerous ranking methods in the field of sports, such as the Massey ratings and Colley matrix, which have been used by the Bowl Championship Series to rank Division I collegiate football teams [BCS 2011]. The search engine Google also uses a mathematical algorithm to compute PageRank, a ranking method used to determine which websites should appear above others in its search results. Google receives 71% of all internet search requests, while the next leading search engine receives only 14% of the requests [SEO 2010], and its PageRank algorithm is one of the main reasons it is the leading search engine on the internet.

There are many factors that determine which websites come up first when you search for something through an internet search engine. On Google, one of those factors is a webpage's PageRank score, and it is this idea of PageRank that set Google apart from other search engines when it was created. The PageRank algorithm assigns a score to each webpage in order to rank the pages according to usefulness. In theory, the most relevant and important pages should come up first in the search results [Wills 2006].

The general concept of the algorithm is to model a random web surfer, starting on one webpage and then clicking on different links to make his or her way through the web. The most "important" webpages are those that have a higher probability of

MSC2010: 15A18, 15A99, 68M01.

Keywords: PageRank algorithm, linear algebra, ranking football teams.

being seen by the random surfer [Wills 2006]. For a page to have higher probability of being seen, either more webpages have to link to that page, or other highly ranked webpages have to link to it.

2. The mathematics behind PageRank

The exact code and formula for Google's PageRank algorithm are kept secret and it is only known what was first used during the development of Google and PageRank. The algorithm that will be used throughout this paper to show how PageRank is calculated is the one that was originally used by Sergey Brin and Lawrence Page [Brin and Page 1998; Page et al. 1999], the creators of Google, and is most likely not the same one used today. To show how Google calculates PageRank let's consider an internet with only four webpages: A, B, C, and D. The web link diagram below shows how the webpages link to each other, where each arrow represents a link from one page to another. For example, webpage C links to both A and D, but not to B.

$$\begin{tikzcd} A \arrow[r] & B \arrow[d] \arrow[dl] \\ D \arrow[u] & C \arrow[l] \arrow[ul] \end{tikzcd}$$

This web link diagram is turned into a web hyperlink matrix H, where

$$H_{ij} = \begin{cases} 1 & \text{if } i \text{ links to } j, \\ 0 & \text{if } i \text{ does not link to } j. \end{cases}$$

Therefore,

$$H = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

for this example.

Next, a row stochastic matrix S is formed from H and is then used to model the random web surfer with the equation $G = \alpha S + (1-\alpha) y v$, where α is defined as the *dampening factor*, y is a column vector of ones, and v is called the *personalization vector*. The vector v is a probability distribution vector, and is currently unknown, but during the development of Google $v = \left(\frac{1}{n} \;\; \frac{1}{n} \;\; \cdots \;\; \frac{1}{n}\right)$ was used [Brin and Page 1998; Page et al. 1999]. The *dampening factor* models the random web surfer's ability to move to a different webpage by means other than following a link, with probability $(1-\alpha)$. The dampening factor used by Brin and Page during early development was $\alpha = 0.85$. In most research done since 1998, values of α range between 0.85 and 0.99 [Wills 2006]. For this example and throughout the paper,

$\alpha = 0.85$ will be used, and, because there are four webpages in this example, $v = \left(\frac{1}{4}\ \frac{1}{4}\ \frac{1}{4}\ \frac{1}{4}\right)$. Using the equation $G = 0.85S + 0.15yv$ we obtain the Google matrix G:

$$G = \begin{pmatrix} \frac{3}{80} & \frac{71}{80} & \frac{3}{80} & \frac{3}{80} \\ \frac{3}{80} & \frac{3}{80} & \frac{37}{80} & \frac{37}{80} \\ \frac{37}{80} & \frac{3}{80} & \frac{3}{80} & \frac{37}{80} \\ \frac{71}{80} & \frac{3}{80} & \frac{3}{80} & \frac{3}{80} \end{pmatrix}.$$

The PageRank vector π is then found by computing the corresponding left eigenvector satisfying $\pi G = \pi$, and, since G is row stochastic, 1 is the dominant eigenvalue, which means π can always be computed [Bryan and Leise 2006]. The i-th entry of π is known as the PageRank score for webpage i. For this particular matrix, the PageRank vector is approximately $(0.306\ \ 0.297\ \ 0.164\ \ 0.233)$. Therefore, the webpage ranking listed from most important to least important is A, B, D, C.

It should be noted that this method is highly inefficient for large matrices, and in 2010 it was estimated that there were approximately a trillion webpages [Kelly 2010]. With such a large and sparse matrix, the power method can be used fairly efficiently to approximate eigenvectors (i.e., to find the PageRank vector) [Bryan and Leise 2006].

3. Dangling node

With the internet constantly growing, many webpages do not link to the majority of the others. In fact, many of them have no out links at all (e.g, postscript files, images). These webpages are known as dangling nodes, and their prevalence leads to a hyperlink matrix which contains mostly zeros. For example, suppose we have the following web link diagram:

$$\begin{tikzcd} A \arrow[r] & B \arrow[d] \arrow[dl] \\ D & C \arrow[l] \arrow[ul] \end{tikzcd}$$

Webpage D would be considered a dangling node, and, in the hyperlink matrix H, row four would be a row of zeros; therefore the matrix would no longer be row stochastic and 1 would no longer be a possible dominant eigenvalue. To fix this, several options exist, one of which is to insert a *personalization vector*, w, into the dangling node rows. It is unknown what Google actually does, but, for this paper, we model the random web surfer's options when on webpage D by assuming he or she has an equal chance to select any other webpage by typing in its URL or to just stay on webpage D, making $w = \left(\frac{1}{4}\ \frac{1}{4}\ \frac{1}{4}\ \frac{1}{4}\right)$. With D being our dangling

NO	48	PHL	22	NO	10	CAR	23	NO	30	CAR	20
NO	35	ATL	27	NO	26	ATL	23	PHL	38	CAR	10
PHL	34	ATL	7	ATL	28	CAR	20	ATL	19	CAR	28

Table 1. Sample 2009 scores.

node we would obtain the following new web hyperlink matrix:

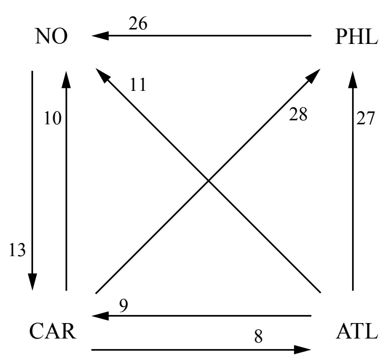
$$H = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}.$$

Calculating as before, the PageRank vector becomes $(0.197\ \ 0.271\ \ 0.219\ \ 0.312)$, producing the ranking D, B, C, A.

4. Using PageRank to rank football teams: GEM 1 method

Applying a similar method to the PageRank algorithm, Govan, Meyer, and Albright [Govan et al. 2008] developed a method called the *GEM method* (which we denoted here by *GEM 1*), using the margin of victory $(v_1 - v_2)$ to weight the "link" between two football teams, where v_1 and v_2 are the teams' scores against each other. As a small sample, the scores in Table 1 were taken from games played in 2009.

By calculating the margin of victory we can create the following link diagram, where each link has a weight equal to the margin of victory:



For example, if New Orleans (NO) played Philadelphia (PHL) and the score was NO-48 and PHL-22, a directed arrow would point towards NO with a weight of 26. If a pair of teams played two games and the same team won both times, the weight assigned to the link is the sum of the margins of victory for the two games.

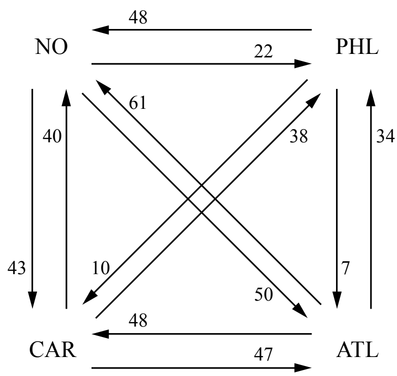
This link diagram then corresponds to the following hyperlink matrix:

$$H = \begin{array}{c} \\ \text{NO} \\ \text{PHL} \\ \text{ATL} \\ \text{CAR} \end{array} \begin{array}{c} \begin{array}{cccc} \text{NO} & \text{PHL} & \text{ATL} & \text{CAR} \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 13 \\ 26 & 0 & 0 & 0 \\ 11 & 27 & 0 & 9 \\ 10 & 28 & 8 & 0 \end{pmatrix} \end{array}.$$

Next, we continue as before to make it row stochastic and follow the PageRank algorithm to get the final ranking. We obtain (0.330 0.252 0.087 0.332) for the PageRank vector, which produces the ranking 1. CAR, 2. NO, 3. PHL, 4. ATL.

5. Ranking football teams: GEM 2 method

We then modified the GEM method to create what we have termed the *GEM 2 method*. Instead of using the margin of victory to weight one arrow for each game, we used both scores to produce two weighted arrows. Since NO scored 48 points against PHL and PHL scored 22 points against NO, the link diagram will now have one arrow directed from PHL to NO with a weight of 48 and another directed from NO to PHL with a weight of 22. If a pair of teams played two games, we summed each team's scores from the two games. Using the data provided in Table 1, we created a new link diagram as follows:



From this diagram the following hyperlink matrix H was then created:

$$H = \begin{array}{c} \\ \text{NO} \\ \text{PHL} \\ \text{ATL} \\ \text{CAR} \end{array} \begin{array}{c} \begin{array}{cccc} \text{NO} & \text{PHL} & \text{ATL} & \text{CAR} \end{array} \\ \begin{pmatrix} 0 & 22 & 50 & 43 \\ 48 & 0 & 7 & 10 \\ 61 & 34 & 0 & 48 \\ 40 & 38 & 47 & 0 \end{pmatrix} \end{array}.$$

The PageRank algorithm gives the PageRank vector (0.317 0.200 0.248 0.335) and the ranking 1. CAR, 2. NO, 3. ATL, 4. PHL.

	Score	Total Yardage	Time of Possession	Turnovers	Actual NFL Ranking
1.	NO	DAL	GB	PHL	NO
2.	NYG	NO	MIN	CAR	MIN
3.	PHL	NYG	DAL	NO	DAL
4.	MIN	MIN	NO	GB	GB
5.	ATL	ATL	NYG	SF	PHL
6.	GB	GB	CAR	TB	ARI
7.	CAR	PHL	ATL	CHI	ATL
8.	DAL	CAR	DET	DET	CAR
9.	CHI	CHI	TB	ATL	SF
10.	ARI	TB	ARI	ARI	NYG
11.	TB	WAS	WAS	DAL	CHI
12.	SF	SEA	CHI	NYG	SEA
13.	WAS	ARI	STL	MIN	WAS
14.	DET	DET	SF	WAS	TB
15.	SEA	STL	SEA	SEA	DET
16.	STL	SF	PHL	STL	STL

Table 2. Final rankings compared to actual rankings using GEM 2.

6. Extended GEM 1 and GEM 2 methods

We collected data on the score, total yardage, turnovers, and time of possession for each regular season game for all 16 teams in the NFL National Football Conference in 2009 [ESPN 2009]. We created four separate H matrices, one for each of the statistics, then proceeded as in Section 5 following the GEM 2 method and the PageRank algorithm using $v = (1/16 \ \ 1/16 \ \ \cdots \ \ 1/16)$ as our personalization vector. Following the same process as before, we produced a ranking for each statistic collected. However, when calculating turnovers, since it is a negative statistic, we chose to orient the directed arrows in the reverse direction.

Table 2 shows the final rankings for each statistic using the GEM 2 method, and also includes the actual end of the regular season rankings.

In comparison, Table 3 shows the final rankings for each statistic also compared with the actual end of the regular season rankings using the original GEM 1 method.

7. Results

For both GEM 1 and GEM 2, we compared the Kendall rank correlation for each statistic versus the actual rankings which are shown in Table 4. The Kendall rank correlation is defined by $r = (n_c - n_d)/(n(n-1)/2)$, where n_c is the number of

	Score	Total Yardage	Time of Possession	Turnovers	Actual NFL Ranking
1.	DAL	GB	GB	PHL	NO
2.	GB	CHI	DAL	NO	MIN
3.	PHL	MIN	CAR	DAL	DAL
4.	MIN	DAL	MIN	TB	GB
5.	CAR	CAR	SEA	CAR	PHL
6.	NYG	PHL	CHI	GB	ARI
7.	NO	ARZ	NO	NYG	ATL
8.	ARZ	NYG	ARZ	CHI	CAR
9.	TB	NO	ATL	SF	SF
10.	SEA	TB	TB	STL	NYG
11.	ATL	DET	NYG	MIN	CHI
12.	SF	SEA	STL	WAS	SEA
13.	CHI	STL	DET	ARZ	WAS
14.	WAS	SF	WAS	ATL	TB
15.	DET	ATL	PHL	DET	DET
16.	STL	WAS	SF	SEA	STL

Table 3. Final rankings compared to actual rankings using GEM 1.

Statistic	Correlation
SCORE1	0.63
SCORE2	0.60
YARD1	0.38
YARD2	0.67
TIME1	0.32
TIME2	0.35
TURN1	0.32
TURN2	0.25

Table 4. Kendall rank correlations versus actual rankings.

concordant pairs and n_d is the number of discordant pairs in the two rankings. For simplicity, labels of the form STAT1 refer to the GEM 1 method and labels of the form STAT2 refer to the GEM 2 method. Based on the r-values, we can see that each method performed better than the other in different statistics. The ranks were then aggregated using the cross-entropy Monte Carlo algorithm with the distance measure equal to the Kendall tau distance, as this algorithm promotes combining several ordered lists in a proper and efficient manner [Pihur et al. 2009; de Boer et al.

	Aggregate GEM 1	Aggregate GEM 2	Aggregate* GEM 1	Aggregate* GEM 2	Actual NFL Ranking
1.	GB	NO	GB	NO	NO
2.	DAL	GB	DAL	NYG	MIN
3.	CAR	NYG	MIN	MIN	DAL
4.	MIN	MIN	CAR	GB	GB
5.	PHL	DAL	CHI	DAL	PHL
6.	NO	CAR	NO	ATL	ARZ
7.	NYG	ATL	ARZ	CAR	ATL
8.	CHI	CHI	TB	ARZ	CAR
9.	ARZ	ARZ	NYG	CHI	SF
10.	TB	TB	SEA	TB	NYG
11.	SEA	PHL	PHL	PHL	CHI
12.	SF	DET	ATL	WAS	SEA
13.	ATL	SF	DET	DET	WAS
14.	STL	WAS	STL	SF	TB
15.	DET	STL	SF	SEA	DET
16.	WAS	SEA	WAS	STL	STL
r-value	0.53	0.55	0.47	0.60	-

Table 5. Aggregated rankings for both GEM 1 and GEM 2 vs. actual rankings.

2005]. With the aggregate rankings, the GEM 2 method performed only slightly better than the original GEM 1 method, with respective Kendall rank correlations of $r = 0.55$ and $r = 0.53$.

We then decided to take out the least-correlated statistic and aggregate the rankings again. We aggregated twice with the GEM 1 method, once without TIME and once without TURN, since both had equally low r-values, and for the GEM 2 method, we aggregated without TURN. When ignoring the least-correlated statistic, the GEM 2 method performed considerably better, with a Kendall rank correlation of $r = 0.60$, compared to the GEM 1 method, $r = 0.45$ when omitting TURN and $r = 0.47$ when omitting TIME. The aggregated rankings when TIME is omitted from GEM 1 and TURN is omitted from GEM 2 are shown in Table 5 along with the original aggregated rankings and the actual end of season rankings and are denoted by Aggregate*.

There is plenty of other variability in the overall approach to this application of PageRank. We could use more statistics or choose different statistics which are better predictors of overall outcome. We also could use a different dampening factor or modify the personalization vector which could improve the rankings as well.

Nonetheless, it is possible to use this method to produce and compute rankings for any sport or anything else from which a link structure can be created.

References

[BCS 2011] Bowl Championship Series, "Bowl championship series official website", webpage, 2011, http://www.bcsfootball.org.

[de Boer et al. 2005] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method", *Ann. Oper. Res.* **134** (2005), 19–67. MR 2006f:90053 Zbl 1075.90066

[Brin and Page 1998] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine", *Computer networks and ISDN systems* **30**:1 (1998), 107–117.

[Bryan and Leise 2006] K. Bryan and T. Leise, "The $25, 000, 000, 000 eigenvector: The linear algebra behind Google", *SIAM Rev.* **48**:3 (2006), 569–581. MR 2008b:15030 Zbl 1115.15007

[ESPN 2009] ESPN NFL, NFL schedule for 2009, 2009, http://espn.go.com/nfl/schedule/_/year/2009.

[Govan et al. 2008] A. Y. Govan, C. D. Meyer, and R. Albright, "Generalizing Google's PageRank to rank National Football League teams", in *Proceedings of the SAS Global Forum*, SAS Global Users Group/SAS Institute, Cary, NC, 2008.

[Kelly 2010] K. Kelly, *What technology wants*, Viking, New York, 2010.

[Page et al. 1999] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web", tech report, Stanford InfoLab, 1999, http://ilpubs.stanford.edu:8090/422/.

[Pihur et al. 2009] V. Pihur, S. Datta, and S. Datta, "RankAggreg, an R package for weighted rank aggregation", *BMC Bioinformatics* **10** (2009), 62.

[SEO 2010] SEO Consultants Directory, "Top search engines for 2010", webpage, 2010, http://www.seoconsultants.com/search-engines/.

[Wills 2006] R. S. Wills, "Google's PageRank: The math behind the search engine", *Math. Intelligencer* **28**:4 (2006), 6–11. MR 2272767

Received: 2012-01-10 Accepted: 2012-06-16

lzack@highpoint.edu *High Point University, 833 Montlieu Avenue, High Point, NC 27262, United States*

rlamb@highpoint.edu *High Point University, 833 Montlieu Avenue, High Point, NC 27262, United States*

ball.sarah.elizabeth@gmail.com *High Point University, 833 Montlieu Avenue, High Point, NC 27262, United States*

involve

msp.org/involve

See inside back cover or msp.org/involve for submission instructions. The subscription price for 2012 is US $105/year for the electronic version, and $145/year (+$35, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to MSP.

Involve (ISSN 1944-4184 electronic, 1944-4176 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840, is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

Involve peer review and production are managed by EditFLOW® from Mathematical Sciences Publishers.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

involve

2012 vol. 5 no. 4