Using the PageRank Algorithm to Rank Football Players in a Game

Aditya Iyer[1] and Shadi Ghiasi[#]

[1]Dhirubhai Ambani International School, India
[#]Advisor

ABSTRACT

There are many methods to rank football players based on their performance in a game or series of games. However, since most methods are subjective, this paper proposes the PageRank algorithm as an objective method to rank players in a football team, where players can be considered as the nodes, and the passes made between them as the edges of a graph. To achieve this, we consider weighting functions, which are based on parameters which consider the number and quality of passes as well as the actions of individual players in the game. In this paper, the game chosen for implementing the rankings Is the 2018 World Cup Final between France and Croatia. The weighting functions are then combined in multiple ways to create different models, which hare implemented in Python to compute the rankings. The models are compared with the official rankings of players during the game with the help of the Kendall's Tau Correlation Coefficient in order to find the distance between the two ranking vectors. While the results may not be highly accurate for the models tested in this paper, a number of additional factors influencing player performance, which official rankings account for, can be considered through more weighting functions. This would lead to more accurate results, thus making the PageRank algorithm a promising and objective tool for ranking football players in a game.

Introduction

Football is one of the most popular sports in the world. The increase in the number of players in football are important for teams' desire to attract better player and choose players based on performance assessments.

The concept of ranking football players is not universal. Instead, different people, coming from different perspectives, have unique ways of determining players who have contributed the most to the team. A large reason for this is because the only score in football is the number of goals scored, which ultimately is seen as the result of a single player. However, most of the times, there are 2 to 4 players that are directly involved in scoring a goal, and their role is often undermined. Additionally, the value of goalkeepers, defenders, and midfielders is undermined, since they do not usually score many goals. Additionally, the most widely used systems to rank players, which consider situational factors such as the assessor's perception and assessment criteria, lack objectivity.

Hence, this paper proposes a new, more objective method to rank the contributions of football players in a game through the use of the PageRank algorithm (Page et al., 1999). Popularly known to rank web pages, this paper demonstrates how this algorithm can be applied to rank players in a football match. The complex, interconnected web graph, can be compared to the graph of a football match, where the web pages represent the players and the links between them represent the passes made between players.

Although ranking of players has received significant attention in sports like basketball (Cooper et al., 2011), tennis (Ruiz et al., 2011), and baseball (Chen & Johnson, 2010), there have been very few studies which analyse the ranking of football players via more objective methods like data envelopment analysis and ordered weighted averaging (Oukil & Govindaluri, 2017).

Additionally, while research has been conducted on the use of the PageRank algorithm to rank national football teams (Lazova & Basnarkov, 2015), the approach of using the algorithm to rank players in a football game has not been studied.

This paper firstly explores the methods for the computation of the PageRank algorithm (e.g. the power method and the steady state approach) in ‘The PageRank Algorithm’. Then, the methods used for the data collection are outlined in ‘Data Collection’. This paper uses weighting functions based on multiple factors to create models to rank the players. The rankings are then compared to the official rankings obtained online using the Kendall's Tau Correlation Coefficient (Shieh, 1998). We also discuss the results obtained, identify the possible flaws in the experiments, and suggests further improvements for future studies in the same topic.

Materials and methods

The PageRank Algorithm

The random surfing model is a graph model which gives the probability of a random user visiting a web page. Mathematically, the random surfer model represents a random walk on the web graph. The probability that a random surfer arrives on a certain page depends on a variety of factors such as the number of links leading to the page, the frequency with which the surfer arrives on pages containing those links, and the number of outgoing links on those pages.

Links from long lists of web pages do not count for much, since the probability of following any one of these links is low. Links from unpopular websites also do not count for much since these pages are not visited often. However, links from popular pages, which do not link to many other pages are valuable and will greatly increase the probability that the surfer visits the linked page. This shows how often a random surfer arrives on a page, which is the idea of PageRank. In the web graph, each website is represented as a node and each link represents an edge between two nodes. For example, if a node (website) A is linked to a node B, it means that A refers to B (Amine, 2020).

PageRank is an algorithm developed to compute the ranking of all websites on the Internet, based on the graph of the web, which contains the outgoing and incoming links to each website. This leads us to the simplified PageRank algorithm (Page et al., 1999):

$$PR_i = \sum_{j \in in(i)} \frac{PR(j)}{|out(j)|}$$

The formula is iterative because pageranks can fluctuate, and one iteration is not enough to tell how important a page is. We stop iterating further when each page’s pagerank from the current iteration differs from the previous iteration by less than or equal to a set margin of error. For the 0th iteration, the pages receive a rank of $\frac{1}{n}$, where n is the number of pages/nodes. The following iterations then follow the formula.
The following is an example of the calculation of ranks of pages using the simplified algorithm.

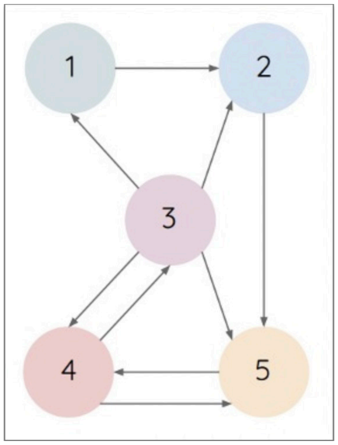

Figure 1. Exemplary calculation of PageRank from graph with nodes and edges

Table 3. Calculation of PageRank from graph in Figure 1

Node	Iteration 0	Iteration 1	Iteration 2	Final PageRank
1	1/5	1/20	1/40	5
2	1/5	5/20	3/40	4
3	1/5	2/20	5/40	3
4	1/5	5/20	15/40	2
5	1/5	7/20	16/40	1

However, this formula does not work in the case of websites which have no outgoing or incoming links, or a group of pages which link to each other, but in no way are connected to the web graph. In reality, a random surfer might jump directly to a random page, ignoring any links. There is also a probability that a random surfer can make a direct jump at any time, which is accounted for the in the real PageRank algorithm using the damping factor d (L. Reeves et al., 2020). Each node gives a d fraction of its pagerank to its neighbours and a $(1-d)$ fraction of its pagerank to every other node in the graph. This means that even pages with no incoming links will receive a pagerank. This results in the formation of the real pagerank algorithm, where d is the damping factor and N is the total number of nodes (Page et al., 1999).

$$PR_i = \frac{1-d}{N} + d \sum_{j \in in(i)} \frac{PR(j)}{|out(j)|}$$

We define a link matrix L, whose entries L_{ij} give the probability that a random surfer will follow a link from page j to page i. (L. Reeves et al., 2020).

$$L_{ij} := \begin{cases} \frac{1}{out(j)} & if\ page\ j\ links\ to\ page\ i \\ 0 & otherwise \end{cases}$$

Additionally, we define d as the probability of making a direct jump and a jump vector e, whose kth component is the probability that a direct jump will lead to page k.

$$e_k := Probability\ that\ a\ direct\ jump\ lands\ on\ page\ k$$

We can now define a transition matrix T, which combines the effects of following links and making direct jumps in the following way (L. Reeves et al., 2020).

$$L_{ij} := \begin{cases} e_i & if\ out(j) = 0 \\ (1-d)L_{ij} + de_i & otherwise \end{cases}$$

The first case corresponds to when a page has no outgoing links, in which case the random surfer always makes a direct jump. The second case combines cases where the random surfer follows links with the probability $(1-d)$ and makes direct jumps with the probability d.

T is a column-stochastic matrix because (L. Reeves et al., 2020).

If $out(j) = 0$, then:

$$\sum_{i=1}^{N} T_{ij} = \sum_{i=1}^{N} e_i = 1$$

If $out(j) \neq 0$, then:

$$\sum_{i=1}^{N} T_{ij} = \sum_{i=1}^{N} (1-d)L_{ij} + de_i = (1-d)\sum_{i=1}^{N} L_{ij} + d\sum_{i=1}^{N} e_i = 1$$

Power method

Finally, we define the pagerank vector R, whose kth component gives the probability that the random surfer is on page k.

$$\mathrm{R_k} := \text{Probability that the current page is page k}$$

When the random surfer clicks on any link, it is considered as an iteration, and the effect can be determined by matrix multiplication. (L. Reeves et al., 2020).

$$R^{(t+1)} = TR^{t}$$

And from an initial R^0 we can find the pagerank vector R.

$$R := \lim_{t\to\infty} T^t\, R^0$$

We can stop iterating when the difference between iterations is less than a margin of error.

$$|R^{(t+1)} - R^{t}| < Tolerance$$

Steady State Approach

The left eigenvectors of the transition matrix T are X such that:

$$TX = \lambda X$$

The entries in the principal eigenvector are the steady-state probabilities of the random walk, including the probability of direct jumps, and thus the PageRank values for the corresponding web pages. If X is the probability distribution of the surfer across web pages, the surfer remains in the steady state distribution X.

We have $\lambda = 1$, given that X is the steady state distribution. Hence we have:

$$TX = 1X$$

Thus, by computing the principal left eigenvector of the transition matrix T, the PageRank values can be obtained.

Data Collection

The data required, which was the number of passes each player made to every other player in the team, was not available online. So, the data was obtained manually from an existing game that had been played. The game chosen is the 2018 world cup final game between France and Croatia, because it would have information available about the performance ratings of the players in the game, thus allowing verification of results.

In order to collect the relevant data, each of the French players were labelled with a number:

1. Hugo Lloris
2. Benjamin Pavard
3. Raphael Varane
4. Samuel Umtiti
5. Lucas Hernandez
6. Kylian Mbappe
7. Paul Pogba
8. Ngolo Kante
9. Blaise Matuidi
10. Antoine Griezmann
11. Olivier Giroud
12. Tolisso Corentin
13. Steven Nzonzi
14. Nabil Fekir

A table was constructed with all the 182 possible passes that could happen between 14 players. Then, from the YouTube video from the official FIFA channel of the full game, each pass that the French players made amongst each other was recorded. Whenever a pass was made, it was noted down in the relevant row. For example, when Hugo Lloris passed the ball to Benjamin Pavard, it was noted down in the 1-2 row.

Experiments

The following table explains the notations used to define the weighting functions.

Table 2. Notation used to design weighting functions

p_i	Total number of passes made by player i
l_i	Number of times player i lost the ball to the opposition
$p_{i,j}$	Number of passes from player i to player j
$l_{i,j}$	Total number of passes between player i and player j

We have designed different weighting functions to conduct our experiments. These functions are as follows:

Weight 1: Player turnovers (Weighted node)

$$f_i = \frac{p_i - l_i}{p_i}$$

Weight 2: Number of passes (Weighted edge)

$$f_{i,j} = p_{i,j}$$

Weight 3: Length of passes (Weighted edge)

For this function, passes are categorised into short, medium, and long, and each of them given a coefficient, represented by A. The coefficient is 0.3 for short passes, 0.6 for medium passes, and 1.0 for long passes, meaning that longer passes are valued more than shorter ones.

$$f_{i,j} = A\,(\frac{p_{i,j}}{t_{i,j}})$$

Weight 4: Value of passes (Weighted edge)

For this function, the positions of players are considered. The 4 positions used in this weighting functions are attacker (A), midfielder (M), defender (D), and goalkeeper (G). There are 15 possible passes between these positions and an assist (a pass leading to a goal). Each of the possible passes is assigned a coefficient, represented by B.

$$f_{i,j} = B\,(p_{i,j})$$

Weight 5: Length and value of passes (Weighted edge)

$$f_{i,j} = A\,(\frac{p_{i,j}}{t_{i,j}}) + B\,(p_{i,j})$$

Using these weighting functions, models were created with different combination of weights. The different models are explained in the table below.

Table 3. Models

Model	Weighted node	Weighted edge
$Model_B$		$f_{i,j} = p_{i,j}$
$Model_{BWN}$	$f_i = \frac{p_i - l_i}{p_i}$	$f_{i,j} = p_{i,j}$
$Model_{LP}$		$f_{i,j} = A\,(\frac{p_{i,j}}{t_{i,j}})$
$Model_{LPWN}$	$f_i = \frac{p_i - l_i}{p_i}$	$f_{i,j} = A\,(\frac{p_{i,j}}{t_{i,j}})$
$Model_{VP}$		$f_{i,j} = B\,(p_{i,j})$
$Model_{VPWN}$	$f_i = \frac{p_i - l_i}{p_i}$	$f_{i,j} = B\,(p_{i,j})$
$Model_{LVP}$		$f_{i,j} = A\,(\frac{p_{i,j}}{t_{i,j}}) + B\,(p_{i,j})$

$Model_{LVPWN}$	$f_i = \frac{p_i - l_i}{p_i}$	$f_{i,j} = A\,(\frac{p_{i,j}}{t_{i,j}}) + B\,(p_{i,j})$

The calculation of the PageRank values was done in Python through the networkx module (NetworkX — NetworkX Documentation, n.d.). The weights were an input for each node and edge and a graph was generated. Below, you can find the graph of $Model_B$.

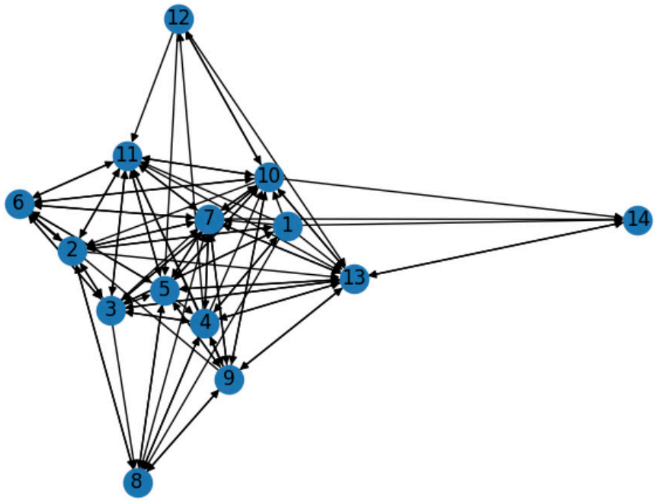

Figure 3. A representation of the directed graph from $Model_B$.

Results and Discussion

The following table outlines the results of the implemented models, in comparison with the official rankings.

Table 4. Rankings obtained from each model.

$Model_B$	$Model_{BWN}$	$Model_{LP}$	$Model_{LPWN}$	$Model_{VP}$	$Model_{VPWN}$	$Model_{LVP}$	$Model_{LVPWN}$	Official ranking
7	7	13	13	10	10	10	10	10
10	10	11	11	7	7	7	7	6
13	13	7	7	6	6	13	13	3
6	6	5	5	13	13	6	6	5
11	11	10	10	11	11	11	11	7
9	9	14	14	9	9	9	9	9
2	2	3	3	14	14	14	14	1
5	5	6	6	2	2	5	5	2
3	3	2	2	5	5	2	2	4
4	4	4	4	3	3	3	3	11
8	8	9	9	4	4	4	4	12
14	14	8	8	8	8	8	8	13
12	12	12	12	12	12	12	12	14
1	1	1	1	1	1	1	1	8

Now, we apply the Kendall's Tau correlation (Magiya, 2019), in order to compute the distance between the ranking vectors obtained from the designed models and the official ranking.

Table 5. Kendall's Tau correlation of model rankings with official rankings

Model	**Kendall's Tau Correlation**
$Model_B$	0.29670
$Model_{BWN}$	0.29670
$Model_{LP}$	0.29670
$Model_{LPWN}$	0.29670
$Model_{VP}$	0.01099
$Model_{VPWN}$	0.01099
$Model_{LVP}$	-0.03297
$Model_{LVPWN}$	-0.03297

In order to find the best football players in a match, different weighting functions were designed, taking into account parameters that affect the rankings of football players. Weighted nodes and weighted edges were constructed and several combinations of weights were used to create different models. The models tested were compared with the official ranking of players. The game chosen was that of the World Cup 2018 Final between France and Croatia, since the results could be verified with existing data on the game. The results show that the Models 1, 2, 3, and 4 have the highest Kendall's Tau Correlation (=0.29670), which means that they are the most accurate model for predicting the rankings of players for this particular match.

From Table 4 and Table 5 we notice that the rankings, and hence the value of the correlation coefficient does not change depending on the existence of weighted nodes. This shows that the weighted nodes have a negligible effect on the rankings of players.

Additionally, model 1 and 2, which are the baseline models having weights equal to the number of passes made between players, have the highest correlation coefficient while models 7 and 8 have a negative correlation, even though they involve multiple weighting functions. This suggests that models relying solely on the number of passes made by players are more accurate predictors than models.

This should not be the case, as factors affecting the worth of any pass should be accounted for, which was attempted through the weighting functions 3, 4, and 5.

Possible reasons for this could be that only a few factors were taken into account. A number of additional factors such as goals scored, possession time, number of tackles by a player, can be considered by forming multiple weighting functions, and combining them to form a comprehensive model, resulting in more accurate rankings. Furthermore, the models were tested only for one game. Using multiple games would allow us to find flaws in the models, as well as come up with new ones based on the comparisons from several games.

Conclusion

In conclusion, this paper proposed the PageRank algorithm as a method to rank football players in a game. While this paper employed the use of only a few weighting functions, future studies could consider the use of more weights, accounting for the possession time, tackles made, goals scored, etc. This could make the PageRank algorithm an objective and promising method to rank football players.

Acknowledgments

I would like to thank my advisor for the valuable insight provided on this topic.

References

Amine, A. (2020, December 20). PageRank algorithm, fully explained. Medium. https://towardsdatascience.com/pagerank-algorithm-fully-explained-dc794184b4af

Chen, W.-C., & Johnson, A. L. (2010). The dynamics of performance space of Major League Baseball pitchers 1871–2006. Annals of Operations Research, 181(1), 287–302. https://doi.org/10.1007/s10479-010-0743-9

Cooper, W. W., Ramón, N., Ruiz, J. L., & Sirvent, I. (2011). Avoiding Large Differences in Weights in Cross-Efficiency Evaluations: Application to the Ranking of Basketball Players. Journal of CENTRUM Cathedra: The Business and Economics Research Journal, 4(2), 197–215. https://doi.org/10.7835/jcc-berj-2011-0058

L. Reeves, B. Pant, & Ramirez. (2020). Google PageRank Explained via Power Iteration. ASU School of Mathematical and Statistical Sciences. https://math.asu.edu/sites/default/files/reeves_lee_apm_505_project_2_math_ma_portfolio_fall_2019-publish.pdf

Lazova, V., & Basnarkov, L. (2015). PageRank Approach to Ranking National Football Teams. Arxiv. https://doi.org/10.48550/arXiv.1503.01331

Magiya, J. (2019, November 23). Kendall Rank Correlation Explained. Medium. https://towardsdatascience.com/kendall-rank-correlation-explained-dee01d99c535

NetworkX — NetworkX documentation. (n.d.). Networkx.org. https://networkx.org/

Oukil, A., & Govindaluri, S. M. (2017). A systematic approach for ranking football players within an integrated DEA-OWA framework. Managerial and Decision Economics, 38(8), 1125–1136. https://doi.org/10.1002/mde.2851

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank Citation Ranking: Bringing Order to the Web. - Stanford InfoLab Publication Server. Stanford.edu. https://doi.org/http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

Ruiz, J. L., Pastor, D., & Pastor, J. T. (2011). Assessing Professional Tennis Players Using Data Envelopment Analysis (DEA). Journal of Sports Economics, 14(3), 276–302. https://doi.org/10.1177/1527002511421952

Shieh, G. S. (1998). A weighted Kendall's tau statistic. Statistics & Probability Letters, 39(1), 17–24. https://doi.org/10.1016/s0167-7152(98)00006-6