# PageRank Approach to Ranking Football Teams' Network

**Boren Wang**
Southern University of Science and Technology
Shenzhen,Guangdong,518055
China
wangbr@mail.sustc.edu.cn

**Zongwei Luo**
Southern University of Science and Technology
Shenzhen,Guangdong,518055
China
luozw@sustc.edu.cn

## ABSTRACT

Football is one of the world's most favored sports with a huge amount of data that one could inspect, analyze and reach interesting conclusions. In this paper we analyze such football data made available through collecting via world football matches.. Our goal is to rank the teams not just based on directmatch result, but also considering team relationship. For this purpose, we apply the PageRank algorithms with restarting mechanismto a graph built from the games. Several statistics such as matches wining and goals scored are combined in different metrics with weights to the links in the graph. Finally, our results indicate that the Random walk approach with the use of right metrics can indeed produce relevant yet more meaningful rankings comparable to the official ranking.

## KEYWORDS

PageRank; Social Network; Football Team Ranking

## 1 INTRODUCTION

Football is one of the world's most favored sports, drawing people's attention in every field, from the simple means of entertainment to more complex objectives of statistics, research and data analysis. In fact, there is significant amount of football match data that one could inspect, analyze and draw conclusions from.

Having that in mind researchers are tackling problems regarding playing strategy, ranking of teams or performance analysis from different aspects including economic, demographic, cultural and climatic factors. A team's game strategy for example can be observed from graph theory perspective by constructing a network of passes between players. In this context different centrality measures can be used to determine the importance of particular players. Other subject of interest might be modelling football matches in terms of scores during the game. For example, authors discuss a statistical model for scoring times in a match. [1].

Here we address the problem of ranking football teams. Our main task is to use the available statistics, in order to come up with an alternative ranking method for the football teams based on their achievements. There are different rating methods currently in use and they produce relevant results. Often organization ranks the teams by scores. Winner gets 3 points while loser gets 0. If teams have tied, every team gets 1 point. A good ranking method should not only take into account how many times a team has won, but also consider how strong an opponent they have defeated. Victory against stronger opponent is preferable and thus more significant than victory against weaker opponent. Toward this end, we exploit methodology that incorporates such logic with PageRank (Random walk) algorithms, which is applicable to vast varieties of network based problems that require ranking in some way. Other than the well-known problem of rating web-pages, [2] PageRank is also utilized in social network analysis, in tasks such as link prediction, information diffusion and communities detection. In addition, it is used in Natural Language Processing (NLP) for the purpose of text summarization and word sense disambiguation.

The rest of the paper is organized as follows. In next section we present the ranking problem and the PageRank based method for solving it. We also give description and statistics of the data that was available. The obtained results are presented in Section III including a discussion and comparison to the official rankings and the difference between national teams and teams in professional leagues then we conclude the paper in last section.

## 2 DATA AND METHODS

### 2.1 Data

The data we used was obtained from 11v11[3], web-site for football statistics that contains all time figures about the matches of the notional team. On the level of clubs, the official web-site of Premier League [4] show us all statistics about clubs. For each team there is information on which team they have played against, the number of matches won, drawn and lost, as well as the number of scored and conceded goals during all match-ups. Throughout this paper we use the term match-up in context of a single game played between two teams. And a match-up pair are every two teams that have played against each other. The dataset contains 210 countries and statistics on 2335 match-up pairs that have played against one another, or 7141 games in total, during which 20298 goals were scored. The average number of games per match-up pair is 3.0582, and the average number of goals scored per matchup pair is 4.3465. In


*SIMUTOOLS '17, September 11–13, 2017, Hong Kong, China*

ACM ISBN 978-1-4503-6388-4/17/09…$15.00
https://doi.org/10.1145/3173519.3173533

Premier League, we choose matches in 2015-2016 season, every team played against others twice at home and away.

### 2.2 Method

The ranking method explored throughout this paper is the PageRank with restarting algorithm applied to a graph build around the available data. Each team is a single node in the graph and two nodes are linked if the two teams (the match-up pair) have ever competed against each other. The weight of the link is determined by a weighting function that involves one or more metrics such as number of games played between a match-up pair, the number of won, lost and drawn games, or the number of scored and conceded goals. The various weighting functions we have tested are given in Table 1.

**Table 1: Set of tested weighting function.**

| | WEIGHTING FUNCTION | INVERSIONS |
|---|---|---|
| 1 | $f_{i,j} = \frac{l_{i,j}}{g_{i,j}} \cdot \frac{1}{G - g_{i,j}+1}$ | 0.032 |
| 2 | $f_{i,j} = \frac{l_{i,j}}{g_{i,j}}$ | 0.038 |
| 3 | $f_{i,j} = \frac{l_{i,j}}{g_{i,j}} + \frac{c_{i,j}}{c_{i,j}+s_{i,j}}$ | 0.040 |
| 4 | $f_{i,j} = l_{i,j}$ | 0.040 |
| 5 | $f_{i,j} = \frac{c_{i,j}}{s_{i,j}}$ | 0.041 |
| 6 | $f_{i,j} = \frac{l_{i,j}}{w_{i,j}}$ | 0.043 |
| 7 | $f_{i,j} = \frac{l_{i,j}}{g_{i,j}} + 0.5 \cdot \frac{d_{i,j}}{g_{i,j}}$ | 0.044 |
| 8 | $f_{i,j} = \frac{c_{i,j}}{c_{i,j}+s_{i,j}}$ | 0.044 |
| 9 | $f_{i,j} = \frac{c_{i,j}}{g_{i,j}}$ | 0.046 |
| 10 | $f_{i,j} = c_{i,j}$ | 0.050 |

Within the functions we use the following notation:

***fi,j*** weight of the link from node i to node j;
***gi,j*** number of games played between the two teams;
***li,j*** number of games lost by team i amongst all the games i and j played;
***wi,j*** number of games won by team i amongst all the games i and j played;
***ci,j*** number of goals conceded by team i during all the games i and j played;
***si,j*** number of goals scored by team i during all the games i and j played;
***di,j*** number of games drawn between the two teams;
***G*** maximum number of games played between any match-up pair;

Another factor that affects the PageRank is the damping factor. The damping factor corresponds to the probability that a random walker would discontinue the walk and jump to a random node [5]. The damping factor other than being necessary as assurance that the random walk would converge to a stationary distribution, it is also intuitive. The intuition behind the use of damping factor within our match-ups network is the following: although the graph is dense not every team have played against every other. So when using weighting metrics such as the loss ratio (weighting function 1 in Table I) the damping factor would mean adding some wining chances to all the teams that have never been played against. It also adds some wining chances to a team that has never won a game within a matchup. The PageRank is calculated using the power method [6] Internet Mathematics, vol. 1, no. 3, pp. 335–380, 2004. This method is an iterative algorithm (eq. 2) that finds the dominant eigenvector, which corresponds to the invariant distribution of the time a random walker spends at a certain node - the PageRank. By normalizing the adjacency matrix A we get the transition probability matrix Q with elements as given in eq. 1.

$$Q_{i,j} = (1-d) \cdot \frac{A_{i,j}}{\sum_{k}^{N} A_{i,k}} + \frac{d}{N} \quad (1)$$

$$\pi^T = \pi^T Q \quad (2)$$

Note that Q is guaranteed to be irreducible and aperiodic as a consequence of the nonzero damping factor d.

### 2.3 EXAMPLE

For the sake of demonstration, let's consider an example that illustrates our goal. Suppose there are 4 teams with given statistics for each pair shown in Table 2.

**Table 2: Number of games played and result**

| PAIR | GAMES | RESULTS |
|---|---|---|
| A-B | 3 | A wins 2, B wins 1 |
| A-C | 3 | A wins 2, C wins 1 |
| A-D | 3 | A wins 3, D wins 0 |
| B-C | 3 | C wins 3, B wins 0 |
| B-D | 3 | D wins 3, B wins 0 |
| C-D | 3 | C wins 1, D wins 2 |

**Table 3: The pagerank of each team**

| TEAM | GAMES | WIN | PAGERANK |
|---|---|---|---|
| A | 9 | 7 | 0.333 |
| C | 9 | 5 | 0.281 |
| D | 9 | 5 | 0.211 |
| B | 9 | 1 | 0.175 |

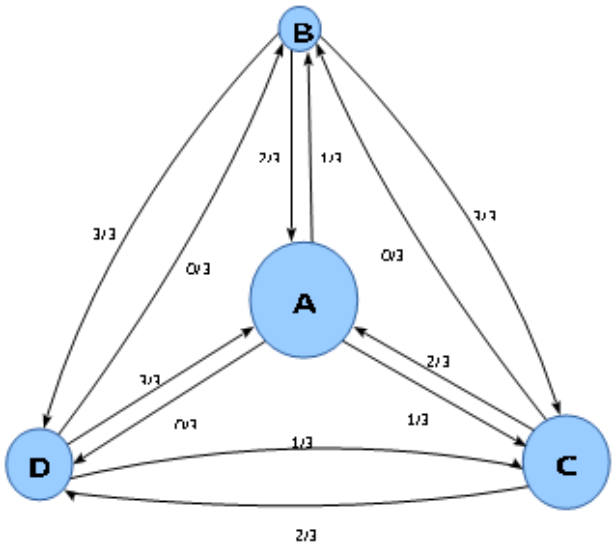


**Figure 1: Graph representation the games played, the size of each node is proportional to it's PageRank**

The graph (Fig. 1) is built using loss ratio as metric (function 2 at Table 1). Therefore the weight of a given link from i to j is the part of the games that i has lost to j. For instance there is a link from A to C with weight of 1/3 and also a link from C to A with weight of 2/3. That means out of 3 matches A and C have played against each other A has won 2 matches, C has won 1 and no matches were drawn. The next step is calculation of the PageRank. Therefore we need transition probability matrix which is calculated according to eq. 1 with a common damping factor value of 0.15. Finally the results are shown at Table 3. A is pointed as highest ranked and B is lowest ranked team as expected. On the other hand, team C and team D both have won 5 games as shown in Table III. However, PageRank takes into account the strength of the defeated opponent not only the number of winnings. As a result, team C is ranked higher since they have won a game against A, considered as strong opponent, in contrast to team D who have winnings only against weaker opponents.

## 3 RESULTS AND DISCUSSION

In order to find the most precise ranking several different weighting functions have been tried and almost all of them delivered similar results. The results were evaluated by comparing the PageRank to the official world cup ranking. We have used normalized number of inversions as evaluation metric, taking the official FIFA all-time rankings as referent ordering. The tested weighting functions and their scores are listed at Table 1. Lower score means the results generated using the corresponding metric are more similar to the official ranking. We only used the top 30 highest ranked teams in the comparison because we wanted to give them higher priority and get their ordering right at the cost of misplacing some of the lower rated teams. The error of the weighting functions also depends on the damping factor. The minimum is achieved when the damping factor value is very small, around 0.05. That is the value we used in the evaluations of the metrics shown in Table 1. Fig. 2 shows errors (in normalized inversions count) for the top 5 metrics as functions of the damping factor. As expected the error increases with the growth of the damping factor.

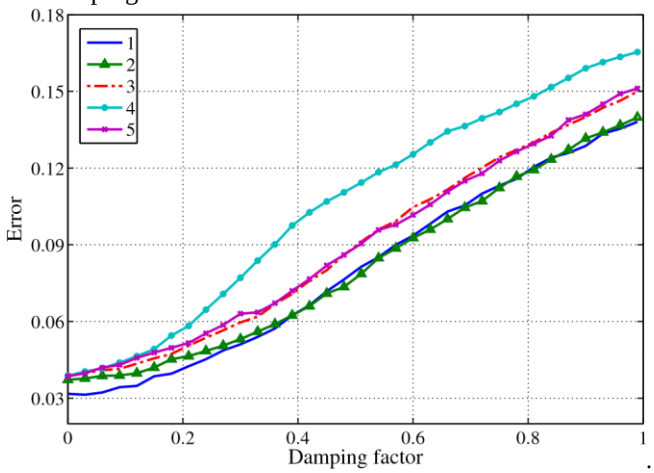


**Figure 2: The error in normalized number of inversions of the first 5 weighting functions in Table I as function of the damping factor**

**Table 4: Top 20 ranked national teams**

| | Country | PageRank | Official |
|---|---|---|---|
| 1 | Brazil | 0.040375 | 1 |
| 2 | Italy | 0.037992 | 3 |
| 3 | Germany | 0.033801 | 2 |
| 4 | Netherlands | 0.031052 | 8 |
| 5 | Argentina | 0.029159 | 4 |
| 6 | England | 0.029100 | 6 |
| 7 | Spain | 0.027904 | 5 |
| 8 | France | 0.025670 | 7 |
| 9 | Czechoslovakia | 0.025155 | NA |

| | | | |
|---|---|---|---|
| 10 | Sweden | 0.022882 | 10 |
| 11 | Mexico | 0.022034 | 13 |
| 12 | Hungary | 0.022014 | 16 |
| 13 | Uruguay | 0.020660 | 9 |
| 14 | Belgium | 0.020255 | 14 |
| 15 | Portugal | 0.020211 | 17 |
| 16 | Poland | 0.019528 | 15 |
| 17 | Denmark | 0.019206 | 25 |
| 18 | Croatia | 0.018993 | 27 |
| 19 | Switzerland | 0.016650 | 21 |
| 20 | Yugoslavia | 0.016466 | NA |

**Table 5: 20 ranked club teams**

| | Team | PageRank | Official |
|---|---|---|---|
| 1 | Leicester City | 0.0742 | 1 |
| 2 | Arsenal | 0.0739 | 2 |
| 3 | West Ham United | 0.06835 | 7 |
| 4 | Manchester United | 0.0679 | 5 |
| 5 | Southampton | 0.06695 | 6 |
| 6 | Tottenham Hotspur | 0.0647 | 3 |
| 7 | Liverpool | 0.05905 | 8 |
| 8 | Manchester City | 0.0533 | 4 |
| 9 | Chelsea | 0.0515 | 10 |
| 10 | Stoke City | 0.05055 | 9 |
| 11 | Swansea City | 0.04655 | 12 |
| 12 | West Bromwich | 0.04485 | 14 |
| 13 | Newcastle United | 0.0428 | 18 |
| 14 | Bournemouth | 0.04215 | 16 |
| 15 | Everton | 0.0385 | 11 |
| 16 | Crystal Palace | 0.0371 | 15 |
| 17 | Watford | 0.0346 | 13 |
| 18 | Sunderland | 0.03365 | 17 |
| 19 | Norwich City | 0.03295 | 19 |
| 20 | Aston Villa | 0.016 | 20 |

Table 4 shows the top 20 teams (for brevity), according to our best weighting function. Table 5 shows the teams in premier League. The 4-th column contains the positions for each team at the official rankings board. The position is marked green if the team holds the same place in both ours and the official world cup rankings. The position is marked with red if there is a large displacement (Denmark and Croatia). If a team is not found in the official ranking (Czechoslovakia and Yugoslavia in our case) their position is marked with NA. Fig 3 shows the match-ups graph. Each team is a node in the graph represented by their national flag and the size of each node is proportional to it's PageRank.

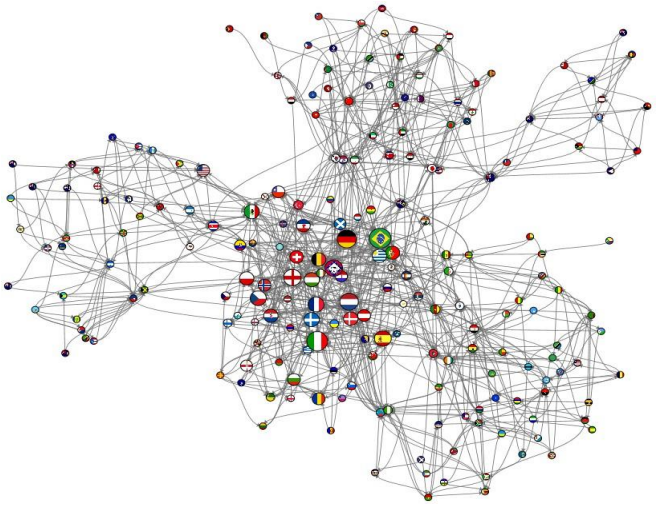

**Figure 3: The graph of the match-ups built with the combination of loss ratio and number of games the two teams played as weighting function (function 1 in Table 1). The size of each node corresponds to their PageRank (damping factor of 0.05 used). For the sake of clarity only the strongest links coming out of each node are shown.**

In the figure a portion of the links are omitted for the sake of clarity, thus the real graph is much denser than it appears.

Possible issue when using PageRank as ranking method might be the following: A node can obtain a high PageRank score if it has a high ranked neighbor from which it can receive significant amount of votes or if it has many low ranked neighbors. In our example, if a national team is highly ranked then they must have either defeated many low ranked teams or achieved remarkable results against a highly ranked opponent. This property of the Random Walk affects our results especially since we treat all matches equally, without taking into account whether it is qualification round or final game. As a result there might be teams that have received high ranking only because they have played and won against many low ranked opponents in less significant qualification matches.

## 4 CONCLUSIONS

Throughout this paper we explored the PageRank method for ranking world football teams. Our results showed that even with weighting functions such as ratio of the goals scored or matches won, the PageRank algorithm derives promising results. The rankings this method produced similar to the official FIFA all-time rankings are more meaningful as they consider team strength/weakness played against each other. However, it is difficult to evaluate whether the PageRank with use of more sophisticated weighting function and more features within the dataset could lead to a better ranking scheme than the official one. Anyway, under the assumption that the FIFA ranking system is proper and accurate, Random walk despite the simple dataset and weighting metrics can generate similar results in a great deal with factors fair to each team.

## ACKNOWLEDGMENTS

This paper is partially supported by SUSTech fund (05/Y01051814, 05/Y01051827, 05/Y01051830, 05/Y01051839).

## REFERENCES

[1] M. Dixon and M. Robinson, A birth process model for association football matches, Journal of the Royal Statistical Society: Series D (The Statistician), vol. 47, no. 3, pp. 523–538, 1998.
[2] L. Page S. Brin R. Motwani and T. Winograd , “The pagerank citation ranking: Bringing order to the web.” 1999.
[3] www.11v11.com
[4] www.premierleague.com
[5] S. Brin and L. Page, “The anatomy of a large-scale hypertextual web search engine,” Computer networks and ISDN systems, vol. 30, no. 1, pp. 107–117, 1998.
[6] A. N. Langville and C. D. Meyer, “Deeper inside pagerank,”