Graph-based Approaches for Analyzing Team Interaction on the Example of Soccer

Markus Brandt[1] and Ulf Brefeld[1,2]

[1] Technische Universität Darmstadt, Darmstadt, Germany
[2] DIPF, Frankfurt am Main, Germany

Abstract. We present a graph-based approach to analyzing player interaction in team sports. A simple pass-based representation is presented that is subsequently used together with the PageRank algorithm to identify the importance of the players. Aggregating player scores to team values allows for turning our approach into a predictor of the winning team. We report on empirical results on five German Bundesliga seasons.

1 Introduction

Soccer is the most popular team sports in the world with 3.3–3.5 billion fans all around the world.[3] As all team sports, the success of a team depends on how well the players cooperate with each other. Team interaction, however, is hardly captured by descriptive statistics such as the number of completed passes.

Nevertheless, passes are predominant means to capture team interaction. Although there are alternative key indicators, including running into space, asking for the ball, etc., these are captured to some extend by focusing on passes as the respective player may, for instance, receive the ball as a consequence of a nice run. In this paper, we interpret soccer players as nodes of a graph and passes between players as (directed) edges. We analyze the passes using simple metrics including the PageRank [4] algorithm on the spanned player graph to measure team interaction. Our empirical results show that a combination of simple features allows for accurately predicting the winning team.

PageRank has previously been used to analyze soccer data. Lazova et al. [3] rank national soccer teams using PageRank for different interactions, such as shots on goal or the number of won matches. They generate a ranking for soccer teams which is comparable to the FIFA official all-time ranking board. Peña and Touchette [5] present an approach to measure team performance using graphs. Similar to our approach, they use pass interactions to generate a pass network, where players are represented by nodes and the passes are represented by edges. Centrality metrics, including closeness, betweenness as well as PageRank, are deployed to determine the performance of a players' contribution to the game. However, these metrics are used to analyze and discuss single matches. An evaluation or comparison like in Lazova et al. [3] is not presented.

[3] `http://sporteology.com/top-10-popular-sports-world/`

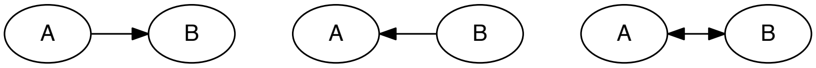

Fig. 1. Different representations of a pass: pass-based (left), receiver-based (center), and interaction-based (right).

The remainder is structured as follows. Section 2 introduces graph-based representations for soccer and metrics that are used in the remainder. Section 3 reports on our empirical results and Section 4 concludes.

2 Viewing Soccer as a Graph

2.1 Representation

We study three representations of passes that differ in the direction of edges. Assume that the event on the pitch is a pass from player A to player B. The first representation is perhaps the most intuitive one and implements the direction of the pass. Thus, if player A passes the ball to player B, the event is represented by a directed edge pointing from A to B. The second representation focuses on the receiving player and the direction of the edge is inverted, so that it points now from B to A. Finally, the third representation measures that there has been an interaction between A and B which is expressed by two directed edges, one pointing from A to B and the other from B to A. We refer to the three representations as *pass*, *receive*, and *interaction*, respectively. Figure 1 shows a visualization.

2.2 Player Metrics

We now introduce simple metrics to compute scores for players. To this end, we compute *pass chains*, that is, we join all successive passes of a team in a single graph. Figure 2 shows an exemplary pass chain that consists of five passes, that is, A → B → C → D → B → E. For simplicity, we denote the number of passes of the chain as the *chain length*; the example in Figure 2 possesses a chain length of five. In the remainder, let C denote the set of all pass chains of a team and $C(p) \subseteq C$ the subset of chains that involve player p.

Chain Scores: The chain length for a player p is given by the average chain score of all chains $c \in C(p)$ he is involved in, that is,

$$cs(p) = \frac{1}{|C(p)|} \sum_{c \in C(p)} length(c). \tag{1}$$

Note that the chain length of a player is oblivious of the actual number of times he receives/completes a pass within a chain. Also note that the representation of

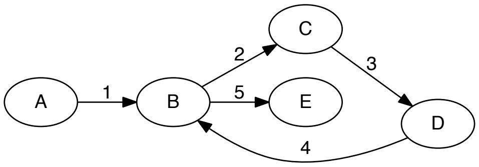

Fig. 2. Example pass chain using the *pass-based* representation with a chain length of five. Numbers attached to edges indicate the temporal order of the passes.

the passes is negligible since only the number of edges are counted, irrespectively of their direction.

PageRank: The PageRank algorithm has originally been devised to analyze the link structure of the Web. Scores for web pages are computed by simulating a random surfer who navigates through the directed graph. Let p be a node (player) and F_p be the set of nodes that p points to and B_p be the set of nodes that point to p. A simplified variant of the PageRank [4] computes scores $R(p)$ for p according to

$$R(p) = c \sum_{q \in B_p} \frac{R(q)}{N_q}. \tag{2}$$

The quantity $R(p)$ of a node corresponds to the sum of the ranks $R(q)$ of all nodes pointing to p, weighted by the amount of total links of q ($N_q = |F_p|$). The factor c is used for normalization. The computation of Equation (2) is repeated until convergence.

Note that some implementations of the PageRank algorithm also include dampening factors in accordance with the metaphor of the behavior of a random surfer [1]. However, modeling random passes is certainly not in the scope of this paper and the thus left out.

2.3 Team Metrics

The previous metrics compute scores for every player. To obtain a team score, the individual player scores need to be aggregated accordingly. For simplicity, we deploy the *average* of all players of a team as well as the *sum* of the individual scores.

3 Empirical Results

Our empirical results are based on data from the German Bundesliga that has kindly been provided by Opta.[4] We use all matches from the seasons 2009/2010 to 2013/2014. Every match is given in form of a temporally annotated sequence of events. We focus on completed passes in the following analysis.

[4] `http://optasports.com`

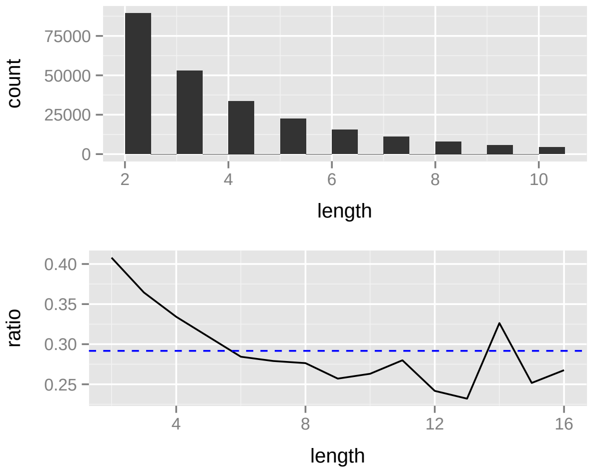

Fig. 3. Top: Lengths of pass chains. Bottom: Probability of an unsuccessful succeeding pass. The dotted line shows the average of 29.17%.

3.1 Chain Lengths vs. PageRank

In total, there are 1530 matches containing 255,231 pass chains ranging from a length from 2 to 65. Figure 3 (top) displays the distribution of all passes. From this distribution, the probability of whether the next pass is successful/unsuccessful can be computed, respectively. Figure 3 (bottom) shows the results for the latter. While the probability of an unsuccessful succeeding pass is 40.79% after the initiating pass, the average probability is 29.17%.

Figure 4 shows the top-ranked players according to chain lengths (left) and PageRank (right). The latter scores are normalized values and computed using the pass-based representation. The lists differ from each other. The chain length favors players that take part in many chains. However, it is unclear if these players play a major role in the game as they could just be passing the ball to each other. PageRank, on the contrary, favors players that are important for team interaction and playmaking. They perform many passes in the chains. Analogous results for receive-based and interaction-based representations lead to an clearer picture, with internationally well-known players as is shown in Figure 5; the displayed rankings are very similar.

3.2 Predicting the Winning Team

We now turn our approach into a predictor of match outcomes. To this end, we split the data into training (seasons 2009/10–2012/13) and test (season 2013/14) sets. The outcomes are derived in terms of the following features: home team wins (positive difference), draw (no difference), home team loses (negative difference).

Player	Chain Score
S. Benyamina	11.00
P.-E. Höjbjerg	10.17
J. Martínez Aguinaga	8.78
M. Spiranovic	8.45
P. Jensen	8.08
C. Buchtmann	8.07
X. Shaqiri	8.06
D. Sereinig	8.00
M. Titsch-Rivero	8.00
D. Thomalla	8.00

Player	PageRank
P. Lahm	100.00
B. Schweinsteiger	97.02
A. Beck	90.44
M. Schäfer	90.09
J. Arango	89.04
C. Gentner	86.83
S. da Silva Pinto	85.55
H. Westermann	81.95
S. Cherundolo	81.62
M. Reus	80.40

Fig. 4. Top-ranked players by chain length (left) and PageRank (right; normalized, pass-based representation).

Player	PageRank
P. Lahm	100.00
B. Schweinsteiger	84.03
H. Badstuber	70.81
D. Bonfim Costa Santos	56.34
H. Westermann	54.65
N. Subotic	52.37
D. van Buyten	51.75
T. Kroos	50.85
M. Hummels	50.55
J. Boateng	47.84

Player	PageRank
P. Lahm	100.00
B. Schweinsteiger	90.13
H. Badstuber	68.46
H. Westermann	66.72
D. Bonfim Costa Santos	66.48
N. Subotic	60.76
A. Beck	60.38
T. Kroos	58.90
M. Hummels	58.58
S. Reinartz	57.10

Fig. 5. Top-ranked players by PageRank using receive-based (left) and interaction-based (right) representations. All scores are normalized.

To predict the outcome, classifiers are trained to predict one of these three classes.

Note that alternative straw men could be easily computed. For instance, extracting the number of wins/draws/losses yields Table 1 for the training and test data. The table not only confirms that there is a home advantage but the numbers could be turned into a predictor that always picks the majority class, in this case a win for the home team. The accuracy on the test data is 47.39%. Another straight forward straw man is to choose the team that won more games in the training data. Although, the ranking of the teams for training and test set differs slightly, an accuracy of 52.94% is obtained.

We deploy C5.0 [6] and SVMs with RBF kernels [2] as the underlying classifiers. User parameters are optimized by a four-fold cross validation. Since SVMs only support binary classification, the one-against-one strategy is used where models are created for each pair of classes and the prediction is made using a majority vote of these pairs. Both methods are trained on only two features, the team scores of the home and the away team, respectively. To compute team scores, scores of all players of each team are used, irrespectively of whether they

Table 1. Distribution of wins/draws/losses.

	Home team wins	Draw	Away team wins
Train set	43.71%	25.00%	31.29%
Test set	47.39%	20.91%	31.70%

	Baseline			Chain Scores			PageRank		
	pass	receive	interact	pass	receive	interact	pass	receive	interact
C5.0 (mean)	50.33	**53.27**	52.61	41.83	**51.31**	**51.31**	47.06	**53.27**	52.29
C5.0 (sum)	52.29	52.94	**53.59**	47.39	**51.31**	**51.31**	53.27	**54.25**	53.95
SVM (mean)	53.92	**54.25**	**54.25**	40.52	**51.31**	50.65	50.65	53.92	**54.25**
SVM (sum)	55.23	**55.56**	55.23	43.14	45.75	**51.96**	**55.88**	54.90	55.23

Fig. 6. Predictive accuracies in percent.

perform in the match or not. The score of a team is therefore computed by summing/averaging the individual PageRank scores of all players, respectively. As an additional baseline, we include the number of completed passes by every team as an additional baseline (also referred to as *baseline*). This metric is used to evaluate how the proposed graphs metrics compare to simple descriptive statistics of the same attributes. As features for the learning algorithms, the number of the passes, receives and their sum are used, respectively, for the home and away team, so that again two-dimensional vectors are obtained.

The results of the classifiers for the baseline, chain lengths, and PageRank are depicted in Table 6. Chain scores perform worst for all representations and classifiers and stay clearly below the second straw man. The best results are 51.31% by the C5.0 classifier and 51.96% by the SVM, both using average of player chain lengths. In contrast to the chain length, the baseline surprisingly shows that simple descriptive statistics can effectively be exploited by the learning algorithms. The best results are 53.27% for C5.0 and 55.56% for the SVM. Both surpass the two straw men but like the chain length predictions, the prediction of the classifiers does not contain draw games, even if they are trained to predict three classes.

Similarly, the results for PageRank provide higher accuracies than using the chain lengths. Accuracies of 54.25% for C5.0 and 55.88% for the SVM are observed. Note that identical values for the baseline and PageRank in Table 6 are caused by normalization that leads to identical representations for interaction-based representations.

Although the accuracy of C5.0 is generally lower than that of the SVM, the resulting decision trees reveal an interesting fact. By focusing on only the PageRank, received-based, sum-aggregated scores, of the away team, an accuracy of 53.59% is obtained. This may be an indicator for the home advantage as a certain PageRank of the away team is seemingly required to beat the home team.

Player	PageRank
P. Lahm	100.00
B. Schweinsteiger	97.02
H. Badstuber	66.71
F. Ribéry	61.36
T. Müller	58.21
A. Ottl	50.72
D. van Buyten	46.86
A. Tymoshchuk	41.33
D. Alaba	39.43
A. Robben	35.25

Player	PageRank
P. Lahm	100.00
B. Schweinsteiger	84.03
H. Badstuber	70.81
D. van Buyten	51.75
F. Ribéry	44.70
A. Tymoshchuk	39.78
T. Müller	37.08
D. Alaba	30.11
A. Robben	26.73
A. Ottl	23.87

Fig. 7. Top 10 PageRank of the *FC Bayern München* using pass-based (left) and receive-based (right) representations. All scores are normalized.

The decision rule chosen for the feature can be stated as

$$\text{away win} = \begin{cases} true, & \text{receive PR sum} > 281.5 \\ false & \text{otherwise,} \end{cases} \tag{3}$$

where the maximal and minimal values for the sum-aggregated, receive-based PageRank are 766 and 44, respectively; the average value is 255. Recall that in the receive-based representation the player that passes gets the credit from the player the ball has been passed to. In sum, the result underlines that successful passing is very important as no other single feature in our study achieves an accuracy in this range. However, the PageRank of players can also be utilized for other purposes than predicting outcomes such as establishing a ranking of players as shown in Figure 7.

3.3 Combining Features

As mentioned earlier, the feature representation of baseline and PageRank can be identical for the interaction-based representation when normalized while the pass- and receive-based representations are always different. This means that both metrics can be combined, despite the fact that they have been generated from the same data. Using an augmented three-dimensional feature representation assembled by the number of successful passes of the home team (interaction-based representation), the summed PageRanks of the home team (again using the interaction-based representation) and the summed PageRanks for the away team (receive-based representation), the predictive accuracy could be improved to 57.19%.

4 Conclusion

We presented a graph-based approach to soccer by viewing passes as edges between player nodes. Depending on the direction of the edges, we showed that either the passing or the receiving player benefits from the pass in the analysis. Empirically, we compared several variants including the PageRank algorithm with appropriate baselines on soccer data from five Bundesliga seasons. Turning the approach into a predictor of the winning team, we showed that the best results are obtained with only three features and observed accuracies of more than 57%.

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. 30(1-7), 107–117 (1998)
2. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Trans. Intell. Syst. Technol. 2(3), 27:1–27:27 (2011)
3. Lazova, V., Basnarkov, L.: PageRank approach to ranking national football teams. CoRR abs/1503.01331 (2015)
4. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Tech. Rep. 1999-66, Stanford InfoLab (1999)
5. Peña, J.L., Touchette, H.: A network theory analysis of football strategies. In: C. Clanet (ed.), Sports Physics: Proc. 2012 Euromech Physics of Sports Conference. pp. 517–528 (2012)
6. Quinlan, R.: Data Mining Tools See5 and C5.0 (2004)