Learning to Rank Sports Teams on a Graph

Jian Shi ^{1,2} and Xin-Yu Tian ^{2,3,*}¹ School of Statistics, Huaqiao University, Xiamen 361021, China; jshi@iss.ac.cn² Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China³ School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China

* Correspondence: tianxinyu181@mails.ucas.edu.cn

Received: 21 July 2020; Accepted: 20 August 2020; Published: 23 August 2020

Abstract: To improve the prediction ability of ranking models in sports, a generalized PageRank model is introduced. In the model, a game graph is constructed from the perspective of Bayesian correction with game results. In the graph, nodes represent teams, and a link function is used to synthesize the information of each game to calculate the weight on the graph's edge. The parameters of the model are estimated by minimizing the loss function, which measures the gap between the predicted rank obtained by the model and the actual rank. The application to the National Basketball Association (NBA) data shows that the proposed model can achieve better prediction performance than the existing ranking models.

Keywords: sports ranking; graph model; prediction; basketball

1. Introduction

It is often necessary to determine the importance of an alternative compared to others in a group. The process of ordering a list of alternatives based on their relative strengths is referred to as ranking. In this paper, we focus on ranking methods with applications to sports. Ranking of individual players or teams in sports, both professional and amateur, is a tool for entertaining fans and developing sports business. Moreover, it is of great significance to provide relevant personnel with reference to the strengths of competitors.

Recent years have seen increasing interest in ranking elite athletes and teams in professional sports leagues and in predicting the outcomes of games. Several methods have been proposed since the last century. Examples of these ranking methods include the Elo method [1], the Keener method [2], the Massey method [3], the Colley method [4], the logistic regression/Markov chain (LRMC) method [5,6], and the network-based method [7]. More recently, due to the extraordinary success of Google's PageRank, an algorithm originally developed for ranking web pages based on their importance [8], graph-based algorithms have gained more ground in the topic of sports ranking problems. The graph-based methods, such as PageRank and its extensions, have been adopted to rank individual players or teams in tennis [9], soccer [10], basketball [11], football [12], rugby [13], and hockey [14].

Although PageRank considers the indirect effect between games through the graph model, like most ranking models, these efforts, which just use game points and win-loss information, pay little attention to obtaining a comprehensive representation to measure the marginal effect of each game. The performance of players and teams is influenced by a variety of causes [15]. For example, the effects of home wins and away wins should be treated differently [16]. As another example, in basketball games, in addition to points, assists and rebounds are also important factors used to measure the strength of a team [17]. To solve this problem, the idea of learning to rank can be used to learn a comprehensive representation of each game.

Learning to rank is first used in information retrieval where machine learning algorithms are used to learn to rank candidate documents in a query according to their relevance to the query. In learning to rank, ranking is considered as a supervised learning problem, and several machine learning algorithms have been applied to it. The work in [18] provided a detailed overview of learning to rank. Learning to rank on a graph is ranking the results of queries that are executed against graphical data, such as social media, linked data, semantic web data, and a semantic bibliographic network [19]. In these graphs, nodes represent entities, and edges mimic interactions or relationships. The work in [20] extended the well-known PageRank authority flow-based ranking method to entity relationship graphs in order to learn an authority flow weight vector.

The sound data index system has been developed for basketball games, providing the possibility and demand for a more comprehensive ranking system. In this paper, we focus on the ranking methods in basketball. A general framework is proposed to learn the setting parameters in a game graph with various game information and rank the teams by an extended PageRank method. Moreover, due to the model setting, our model has the power to realize FutureRank, predicting the rank of teams in the future. The rest of this paper is organized as follows. Section 2 presents the framework of learning to rank sports teams on a graph. In Section 3, the methodology is applied to the data for the National Basketball Association. More aspects about the empirical results are discussed in Section 4. Concluding remarks and future research are summarized in Section 5.

2. Materials and Methods

2.1. Data Description

The National Basketball Association (NBA), consisting of two conferences, each with 15 teams, is the organization responsible for professional basketball in the United States. During a regular season in the NBA, where home and away competition system is adopted, each team plays 82 games, resulting in a total of 1230 regular season games. A large amount of statistical data are generated after each game, which provide a rich laboratory for studying various aspects of the sport. Box score data are a structured summary of the results from a basketball game, which contains comprehensive information about the game. The box score lists the game score, as well as individual and team achievements in the game. In our research, NBA box score data from the 2009–2010 season to 2018–2019 season were collected from www.nba.com. An example of the data is shown in Table 1.

Table 1. The box score of the game between Philadelphia 76ers (PHI) and Boston Celtics (BOS) in the 2018–2019 season.

Team	Assists	Rebounds	Score	Fouls	Turnovers
PHI	18	47	87	16	20
BOS	21	55	105	15	20

2.2. Graph Construction

Like the traditional PageRank method, the network of sports teams is considered as a weighted directed graph. We construct the graph from the viewpoint of Bayesian correction, which is similar to the Elo method. It will be shown that the building process can provide a more statistical explanation for the model. The construction process is illustrated as follows. Let $N(a, b)$ represent the normal distribution with mean a and variance b . For a game k between team i and team j , we assume:

$$h(S_j; \gamma_j) \rho(X_i^k, X_j^k; \theta) \sim N(S_i, \sigma^2), \quad (1)$$

where $X_i^k(X_j^k)$ and $S_i(S_j)$ denote the performance vector and strength of team $i(j)$, respectively. $\rho(X_i^k, X_j^k; \theta)$ is set as the link function with parameter θ to integrate game information including the performance of two teams. $h(S_j; \gamma_j)$ is a discount factor dependent on S_j with a team-wise parameter γ_j ,

which allows taking the opponent's strength into account in the measurement of the effect of one game. Furthermore, the strength S_i is considered as a stochastic variable with prior distribution $N(S_i^0, \tau^2)$. Let $R(i)$ denote the set of games that team i has played and $g(k, i)$ denote the opponent team of team i in game k . Then, the posterior mean of S_i can be given by:

$$\hat{S}_i = \frac{\tau^2 \sum_{k \in R(i)} h(S_{g(k,i)}; \gamma_{g(k,i)}) \rho(X_i^k, X_{g(k,i)}^k; \theta) + \sigma^2 S_i^0}{\|R(i)\| \tau^2 + \sigma^2}, \quad (2)$$

where $\|R(i)\|$ is the cardinal number of the game set $R(i)$. The posterior mean of each team's strength can be calculated in the same fashion. Then, we will have a network with strengths that flow from team to team. Suppose there are n teams to be ranked. Let $S = (S_1, S_2, \dots, S_n)^\top$, $S^0 = (S_1^0, S_2^0, \dots, S_n^0)^\top$, $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_n)^\top$, and $H(S, \gamma) = (h(S_1, \gamma_1), h(S_2, \gamma_2), \dots, h(S_n, \gamma_n))^\top$. Then, the rating vector is recursively defined as follows.

$$S = \text{Diag}(\alpha)WH(S; \gamma) + (1 - \alpha)S^0, \quad (3)$$

where $\text{Diag}(\alpha)$ is the diagonal matrix composed of the vector α whose i -th element is equal to $\frac{\|R(i)\| \tau^2}{\|R(i)\| \tau^2 + \sigma^2}$, and matrix W is defined by:

$$W_{ij} = \begin{cases} \frac{1}{\|R(i)\|} \sum_{k \in R(i), g(k,i)=j} \rho(X_i^k, X_j^k; \theta), & \text{team } i \text{ has played with team } j, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

Assume $H(S, \gamma) = \text{Diag}(\gamma)S$, and let $A = W\text{Diag}(\gamma)$. Then, we obtain the generalized PageRank rating system,

$$S = \alpha AS + (1 - \alpha)S^0. \quad (5)$$

To preserve the column stochastic property of matrix A , we design γ_j as a normalization constant in practice, i.e., $\gamma_j = \sum_i A_{ij}$. The rating vector for teams is retrieved from the solution of the linear system. Then, the team's ranking can be obtained directly from the rating vector.

We hope the link function $\rho(\cdot, \cdot; \theta)$ is set to ensure the non-negativity of matrix A . Hence, an exponential function is considered, i.e.,

$$\rho(X_i^k, X_j^k; \theta) = \exp \left[\theta^T (X_i^k - X_j^k) \right]. \quad (6)$$

With the link function, the problem of dangling nodes in PageRank can be avoided, and such a setting ensures that even the strength of the losing team will be corrected.

In addition to the team's performance, schedule factors should also be considered, the foremost of which is the home advantage. Throughout nearly ten seasons of NBA games, the home team's win rate reaches 59.02%. The advantage deriving from playing at home is commonly taken into account by including a common home effect parameter η for all teams, thus leading to the function:

$$\rho(X_i^k, X_j^k, I_i^k; \theta, \eta) = \exp \left[\theta^T (X_i^k - X_j^k) + \eta I_i^k \right], \quad (7)$$

where I_i^k is an indicator variable, i.e.,

$$I_i^k = \begin{cases} 1, & \text{team } i \text{ is the home team in game } k, \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

It can be seen that the parameter $\exp(\eta)$ acts as a discount factor, which discounts the performance of the home team.

For convenience, we call this extended model TeamRank. Figure 1 shows an example to visualize the 2018–2019 NBA season network constructed with the first three days' games.

Figure 1. The graph for the first three days' games of the 2018–2019 NBA season.

Compared with the traditional PageRank method, the generalized model has the following advantages.

1. Prior information can be incorporated, which allows our model to be aggregated with other ranking models.
2. The hyperparameter, damping factor α , is adjusted adaptively as games continue.
3. The link function considers a multidimensional performance vector, which contains more information than game points or wins.

2.3. Parameter Estimation

The model is used to evaluate the true strength of the team. Therefore, to estimate the parameters, a valid loss function should be designed to measure the gap between the estimated value of the team's strength and the actual value. Assume the list of teams is ranked L times, and then, the parameter estimation can be given by minimizing the loss function,

$$\hat{\Theta} = \arg \min_{\Theta} \sum_{k=1}^L \text{loss}(\hat{S}^{(k)}(\Theta), S^{(k)}), \quad (9)$$

where $\Theta = (\theta, \eta, \frac{\tau^2}{\sigma^2})^\top$ and $\hat{S}^{(k)}(\Theta)$ and $S^{(k)}$ are the rating vector obtained from (5) and the true strength vector of the teams in the k -th ranking, respectively. Considering a measurement of ranking accuracy, the negative Kendall rank correlation is used as the loss function. The function is defined as follows.

$$\text{loss}(\hat{S}^{(k)}(\Theta), S^{(k)}) = -\frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \text{sgn}(\hat{S}_i^{(k)}(\Theta) - \hat{S}_j^{(k)}(\Theta)) \text{sgn}(S_i^{(k)} - S_j^{(k)}), \quad (10)$$

where $\text{sgn}(\cdot)$ is the sign function and the subscript i refers to the i -th element in the vector. This means that the estimation process will choose the parameters that make the team's strength rating most relevant to the true value. Since the objective function is not continuous, it is solved by gradient-free optimization algorithms, such as the genetic algorithm.

3. Results

In this section, we apply the proposed ranking method to the data of NBA collected from the 2009–2010 season to 2018–2019 season. Five continuous variables (points, rebounds, assists, turnovers, and fouls) and one binary variable (home field advantage) are used as the input for the link function of the model. For each season, the game graph is constructed by the first half of the games (615 regular season games), and the total wins of the half are employed as the prior distribution for each team's strength. The number of games a team wins throughout the season is used as a measurement of the team's true strength. In the training process, the rating vector and the loss function are computed for the first six seasons, and the parameters are estimated by minimizing the total loss. Then, the testing process is done over the four seasons left.

The estimated parameters in the link function $\rho(X; \theta)$ are shown in Table 2. The estimated result is in line with common sense. These indicators of points, rebounds, and assists all have a positive effect on the team's strength; at the same time, turnovers and fouls have the opposite effect. In addition, home advantage will bring a $\frac{1}{0.8266}$, 120.97% improvement to the home team's performance, which is similar to the ratio of the home team's winning percentage of 50%.

Table 2. Parameter estimation result.

	Points	Rebounds	Assists	Turnovers	Fouls	Home Advantage
θ	0.3303	0.0394	0.2938	−0.1583	−0.0292	−0.1904
$\exp(\theta)$	1.3914	1.0402	1.3415	0.8536	0.9712	0.8266

The testing results are evaluated by three criteria: Kendall rank correlation, prediction accuracy, and playoff accuracy. For each season in the testing sets, the Kendall rank correlation is calculated between the rating vector estimated from the first 615 games and the true strength vector. The winning team of one game is predicted as the team with a higher rating value. Then, the prediction accuracy is computed for the 615 games left. A team ranked in the top eight in the conference will be predicted as a playoff team. Playoff accuracy is defined as the correct proportion of teams that are predicted to enter the playoffs.

Comparisons are done between the proposed model and several popular ranking methods with applications primarily to sports [21], including the Colley method, the Massey method, the PageRank method, and the Elo method. The simplest model in which the team's rating is computed by the current wins is set as the baseline. The results are shown in Tables 3–5.

Table 3. Kendall rank correlation of the results. Red values indicate the highest Kendall rank correlation for each season.

Model	2015	2016	2017	2018	Average
Baseline	0.7823	0.7482	0.6343	0.7618	0.7317
Colley	0.7796	0.7959	0.6242	0.7140	0.7284
Massey	0.8029	0.7681	0.6706	0.6396	0.7203
PageRank	0.6302	0.6242	0.5035	0.5652	0.5807
Elo	0.7562	0.8005	0.6010	0.6721	0.7075
TeamRank	0.8309	0.8191	0.6891	0.6768	0.7540

Table 4. Prediction accuracy of the results. Red values indicate the highest prediction accuracy for each season.

Model	2015	2016	2017	2018	Average
Baseline	0.6504	0.5937	0.6276	0.6146	0.6216
Colley	0.6667	0.6136	0.6488	0.6488	0.6445
Massey	0.6683	0.6153	0.6667	0.6439	0.6485
PageRank	0.6585	0.6036	0.6520	0.6244	0.6347
Elo	0.6715	0.6252	0.6504	0.6569	0.6510
TeamRank	0.6764	0.6169	0.6699	0.6504	0.6534

Table 5. Playoff accuracy of the results. Red values indicate the highest playoff accuracy for each season.

Model	2015	2016	2017	2018	Average
Baseline	0.9375	0.875	0.8125	0.8125	0.8594
Colley	0.875	0.875	0.875	0.8125	0.8594
Massey	0.8125	0.875	0.6875	0.8125	0.7969
PageRank	0.8125	0.875	0.8125	0.8125	0.8281
Elo	0.8125	0.875	0.8125	0.8125	0.8281
TeamRank	0.875	0.9375	0.875	0.875	0.8906

As shown in the results, the proposed model has the best performance with an average correlation of 0.7540, prediction accuracy of 65.34%, and playoff accuracy of 89.06%. In most seasons, the TeamRank model outperforms other models. Although the prediction accuracy of the Elo model is close to that of TeamRank, the correlation and playoff accuracy of TeamRank exceed the Elo model greatly.

4. Discussion

4.1. Correlation Analysis

To discuss the difference between the rank results of these models, a correlation graph is plotted for these models' ranking results in Figure 2.

Figure 2. Kendall correlation matrix for different models' results.

It can be seen that compared with other models, the ranking given by the PageRank method differs greatly from the true ranking. Figure 2 shows that the results of PageRank are not similar to the results of other models. This is because the two NBA teams will play multiple times in one season. The PageRank algorithm reduces the results of multiple games to a total goal difference, ignoring the marginal effect of a single game. For example, consider that team j wins one game with team i by a score of 121 to 100 and loses a second game by a score of 90 to 91. The two margins of victory would be summed and modeled as a single weighted edge of two points from team i to team j in the PageRank model. Consider another example where team j wins one game with team i by a score of 110 to 100 and wins a second game by a score of 100 to 90. The two margins of victory would be summed and modeled as a single weighted edge of 20 points from team i to team j as well. However, obviously, the strength contrast between the two situations is completely different. The TeamRank model is game-wise, which constructs the graph game-by-game. Hence, compared with the traditional PageRank model, this generalized model improves the prediction ability of the graph model.

The problem of ignoring the marginal effect of a single game also exists in the models of Colley and Massey, which makes them unable to measure the impact of the opponent's strength following a specific game. Like the TeamRank model, the Elo method considers the influence of the opponent's strength on the team rating. However, the Elo method only focuses on the points, and the update following one game is non-recursive, which makes it impossible to modify the impacts of previous games. Although the predictive ability of Colley, Massey, and Elo is not as good as the TeamRank model, the high correlations with the results of the baseline model show that the three models are hindsight models, which indicates that they have good backtesting ability.

4.2. Dynamic Ranking

In addition to the prediction of the half-season, the game graph can be updated according to the progress of the games, thereby updating the team's rating vector to achieve a dynamic ranking. We present the dynamic ranking results updated every 300 games in the 2018–2019 season in Figure 3. Figure 4 presents the change of rank in four seasons from 2015 to 2018 for eight NBA teams. It can be seen that some teams, such as San Antonio Spurs, have maintained a stable strength for four seasons. On the contrary, some teams have large fluctuations in their rankings during one season, such as Miami Heat and Los Angeles Clippers.

Figure 3. Ranking results for the 2018–2019 season. The four subplots represent the ranking results of 300, 600, 900, and 1200 games, respectively.

Figure 4. Ranking results for eight teams from 2015 to 2018.

5. Conclusions

In this study, we described a generalized PageRank model to rank teams in a sports league. In the proposed model, a game graph is constructed from the viewpoint of Bayesian correction, which allows the strength to flow from team to team after games. The adjacency matrix of the graph is calculated by a link function that synthesizes all aspects of each game, including game results and schedule information. The unknown parameters in the model can be learned under the guidance of the actual ranking of teams. The loss function of a negative Kendall correlation is minimized to estimate the parameters. The application of this model was implemented with the NBA data. In each season, with the data

of half a season, the strength of each team is evaluated, and the team's ranking of the total season is predicted. Our model outperforms other widely used models in three criteria: Kendall rank correlation, prediction accuracy, and playoff accuracy. The model in this paper improves the prediction ability of the PageRank-like models and provides coaches, media, and fans with a more accurate reference to the ranking of teams' strengths.

The proposed model takes into account the marginal impact of each game on the team's strength assessment. However, the time factor is not taken into account, that is the effect of one game may weaken over time. The game can be weighted by the length of time intervals in future work. Moreover, there may be more effective function forms for the link function and the loss function. More complicated forms can be tried for improvement. Besides basketball, the proposed method can be applied to rank teams or athletes in other sports fields.

Author Contributions: Conceptualization, J.S. and X.-Y.T.; methodology, J.S. and X.-Y.T.; software, X.-Y.T.; validation, J.S. and X.-Y.T.; formal analysis, X.-Y.T.; investigation, X.-Y.T.; resources, X.-Y.T.; data curation, J.S. and X.-Y.T.; writing, original draft preparation, X.-Y.T.; writing, review and editing, J.S. and X.-Y.T.; visualization, X.-Y.T.; supervision, J.S.; project administration, J.S. All authors read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: We would like to thank the Editor and reviewers for their constructive comments and suggestions, which considerably improved this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Elo, A.E. *The Rating of Chessplayers, Past and Present*; Arco Pub.: New York, NY, USA, 1978.
2. Keener, J.P. The Perron-Frobenius Theorem and the Ranking of Football Teams. *SIAM Rev.* **1993**, *35*, 80–93. [[CrossRef](#)]
3. Massey, K. *Statistical Models Applied to the Rating of Sports Teams*; Bluefield College: Bluefield, VA, USA, 1997.
4. Colley, W. Colley's Bias Free College Football Ranking Method. Ph.D. Thesis, Princeton University, Princeton, NJ, USA, 2002.
5. Kvam, P.; Sokol, J.S. A logistic regression/Markov chain model for NCAA basketball. *Nav. Res. Logist.* **2006**, *53*, 788–803. [[CrossRef](#)]
6. Kolbush, J.; Sokol, J. A logistic regression/Markov chain model for American college football. *Int. J. Comput. Sci. Sport* **2017**, *16*, 185–196. [[CrossRef](#)]
7. Park, J.; Newman, M.E.J. A network-based ranking system for US college football. *J. Stat. Mech. Theory Exp.* **2005**, *2005*, P10014. [[CrossRef](#)]
8. Page, L.; Brin, S.; Motwani, R.; Winograd, T. *The Pagerank Citation Ranking: Bringing Order to the Web*; Technical Report; Stanford InfoLab: Stanford, CA, USA, 1999.
9. Radicchi, F. Who is the best player ever? A complex network analysis of the history of professional tennis. *PLoS ONE* **2011**, *6*, e17249. [[CrossRef](#)] [[PubMed](#)]
10. Govan, A.Y.; Meyer, C.D.; Albright, R. Generalizing Google's PageRank to rank national football league teams. In Proceedings of the SAS Global Forum, San Antonio, TX, USA, 16–19 March 2008; Volume 2008.
11. Xia, V.; Jain, K.; Krishna, A.; Brinton, C.G. A network-driven methodology for sports ranking and prediction. In Proceedings of the 2018 52nd Annual Conference on Information Sciences and Systems (CISS), Princeton, NJ, USA, 21–23 March 2018; pp. 1–6.
12. Lazova, V.; Basnarkov, L. PageRank approach to ranking national football teams. *arXiv* **2015**, arXiv:1503.01331.
13. Cintia, P.; Coscia, M.; Pappalardo, L. The Haka network: Evaluating rugby team performance with dynamic graph analysis. In Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), San Francisco, CA, USA, 18–21 August 2016; pp. 1095–1102.
14. Swanson, N.; Koban, D.; Brundage, P. Predicting the NHL playoffs with PageRank. *J. Quant. Anal. Sport.* **2017**, *13*, 131–139. [[CrossRef](#)]

15. Morales, J.A.; Sánchez, S.; Flores, J.; Pineda, C.; Gershenson, C.; Cocho, G.; Zizumbo, J.; Rodríguez, R.F.; Iñiguez, G. Generic temporal features of performance rankings in sports and games. *EPJ Data Sci.* **2016**, *5*, 33. [\[CrossRef\]](#)
16. Ribeiro, H.V.; Mukherjee, S.; Zeng, X.H.T. The advantage of playing home in NBA: Microscopic, team-specific and evolving features. *PLoS ONE* **2016**, *11*, e0152440. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Zhang, S.; Gomez, M.Á.; Yi, Q.; Dong, R.; Leicht, A.; Lorenzo, A. Modelling the Relationship between Match Outcome and Match Performances during the 2019 FIBA Basketball World Cup: A Quantile Regression Analysis. *Int. J. Environ. Res. Public Health* **2020**, *17*, 5722. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Chapelle, O.; Chang, Y. Yahoo! Learning to rank challenge overview. In Proceedings of the 27th International Conference on Machine Learning (ICML 2010), Haifa, Israel, 26 August 2010; pp. 1–24.
19. Agarwal, S. Learning to rank on graphs. *Mach. Learn.* **2010**, *81*, 333–357. [\[CrossRef\]](#)
20. Raschid, L.; Sayyadi, H.; Hristidis, V. Learning to Rank in Entity Relationship Graphs. *INFORMS J. Comput.* **2019**, *31*, 671–688. [\[CrossRef\]](#)
21. Vaziri, B.; Dabadghao, S.; Yih, Y.; Morin, T.L. Properties of sports ranking methods. *J. Oper. Res. Soc.* **2018**, *69*, 776–787. [\[CrossRef\]](#)

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).