CPSC 6030: Soccer Data Visualization

Akash Indani and Reek Majumder

Abstract

As per multiple online resources like worldatlas [1], totalsportek [2], sportytell [3], etc, soccer is the most popular sports with biggest fan base of 4 billion. This inspired us to pick this topic for data visualization. We picked the two most famous leagues which are English premier league and Spanish first division league - Laliga.

abbreviations: EPL - English premier league, Laliga - Spanish first division league

1 Introduction

We wanted to visualize the relative team performance and how teams are progressing throughout the league. For this purpose, we needed the dataset that contains the matches, teams, and events data. We found an interesting open-source dataset on the Figshare website [4] which contained tournament, matches, and team data for 7 competitions of the year 2017-18. We focus our story around two visualizations - Week-wise team progression and relative team performance. In EPL and Laliga 20 teams play home and away games against each other which results in a total of 38 matches for each team over 38 weeks. This inspired us to visualize the week-wise team's progression if the performance of each team can improve or degrade. We were also interested to see how each team is performing against every other team in a home and away games. We used Line and Network graphs respectively for both of these questions.

2 Dataset description

We got the opensource soccer match event dataset by Luca Pappalardo, all in JSON format. The whole dataset is divided into different parts like competition, matches, and teams dataset which are internally connected with common keys between them. An example screenshot can be seen in Figure 1. The important attributes of our visualizations are the home team, away team, scores, game week, competition. All this data is stored in nested JSON format and that's why we wanted to simplify it. We wrote a python script to simplify the data, take out the important attributes, and ordered it by team names and game weeks. Then we use the game week to plot the line graph by putting the game week on the x-axis and all the matches of the teams on the Y-axis. Similarly, home and away team data with the scores were used in the network graph.

Figure 1: Interlinked json files of matches, competitions and teams

3 Design solutions

We wanted to address two types of questions:

- Week-wise each team's progression based on points and goal differences.
- Relative team performance of each team against every other team.

For both the visualizations we have provided the user a filter to select the league between EPL or Laliga.

3.1 Week-wise team's progression

All 20 teams in the League, play every other team twice every season, playing a total of 38 matches. Every team plays one match every week. A win earns a team 3 points, a draw earns both the teams 1 point and a loss does not earn any point. At the end of 38 weeks, the team with the maximum number of points wins the League. Different teams perform differently in the league. The better teams perform consistently throughout the year, while some teams perform extremely well for one half of the year, but manage to loose in the other half. Also, due to the highly unpredictable nature, any team can win or lose at any time. Therefore, looking at how teams perform throughout the year, can provide some interesting insights.

We have provided two types of visualizations to user: One is compact which have very compact lines and points while the other is subway tracks as it is familiar to users.

1. For compact visualization, we have used points and lines as marks and colors as the channels to differentiate between each team. The horizontal scale explains the weeks for the tournament. The left vertical scale shows the ranking of teams and the right vertical scale sorts the teams based on the final ranking at the end of the tournament. (See Figure 2)
2. The second visualization is an extension of our first visualization inspired by tracks on the subway system where we aim to better visualize a team ranking in the middle of the tournament.(See Figure 3)

We have added Mouse over as an interaction when we highlighted the teams on the right and its initial ranking on the left. Points on this graph refer to teams ranking for a particular week which is denoted as text within it. We have iterated through the matches of a particular tournament (EPL or Laliga), we allocated ranks to teams based on points table (calculated by match won, lost, and tied till that week) and goal difference (the difference between the goal scored and conceded till that week). As a part of interaction with points, we display the scoreline of the match between the home and the away team.

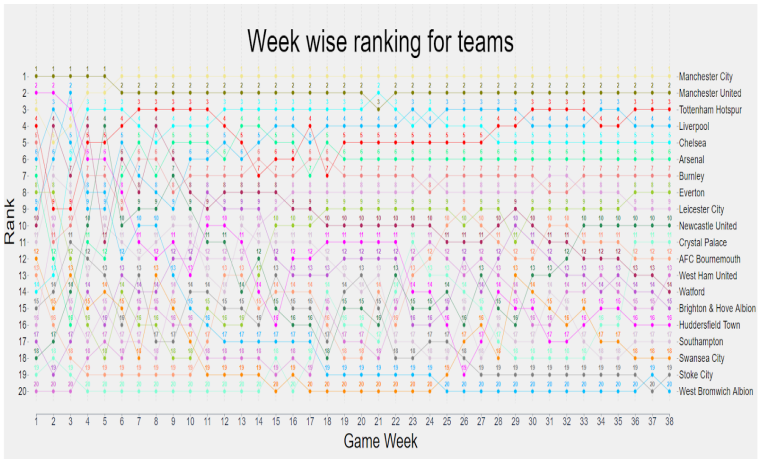

Figure 2: Compact Line graph

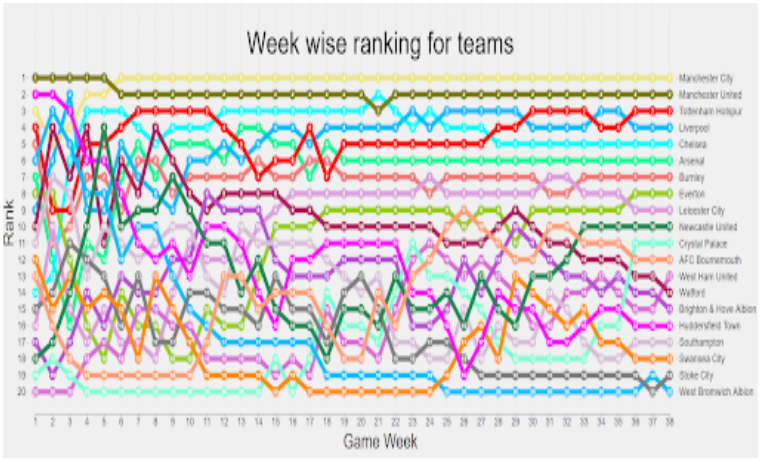

Figure 3: Track Line graph

We have iterated through the matches of a particular tournament (EPL or Laliga) we allocated ranks to teams based on points table (calculated by match won, lost, and tied till that week) and goal difference (the difference between the goal scored and conceded till that week).

3.2 Relative team performance

In the league, teams have different home and away performances, establishing patterns within the league. For example, Liverpool performed very well and was a title contender, but they performed poorly in away games by multiple losses and drawn matches (We can see that in the graph by selecting Liverpool). Such interesting insights can be developed if one views the relative performance of teams with respect to each other.

Initially, we have used Points as our marks and size as our channels. Higher the score of a team at the end of

the tournament higher the radius of the circle (we can see this in figure 4). On double-clicking a node, we introduce lines as new marks. Each line represents a match played. We are showing a relative team performance with colors red, yellow, and green for matches lost, tie, and won respectively. As part of the interaction, we show the match details on the mouse over a link. The match details show the match score between the teams and whether it is a home game or away game with respect to the node selected (we can see this in figure 5).

We have multiple interactions and features like double-click to any node to keep that as source and all the other nodes as destination and show all the matches between the source and the destinations which include all the home and away games. Also, we can drag the nodes to keep that fixated at one (x,y) position and we can analyze the graph as we want.

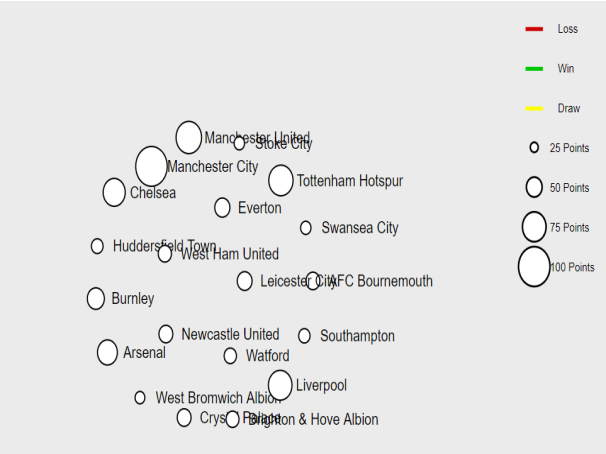

Figure 4: Initial network graph

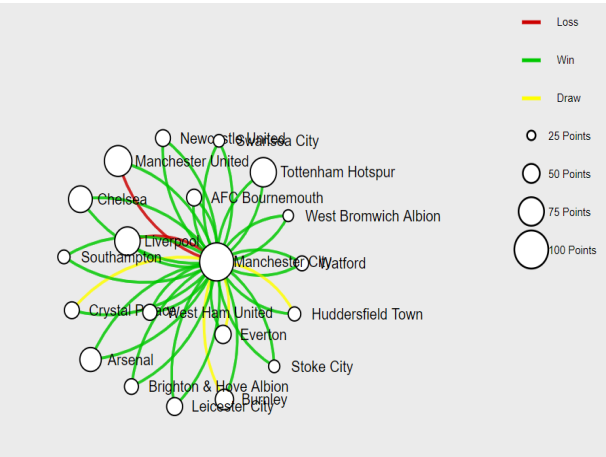

Figure 5: Final network graph on double-click

4 Related work

In the initial stages of the project, we were thinking about different datasets, and the best we found in the paper of Spatio-temporal match events[6], which we are currently using for this project. The major advantage of using this data was that it is open-source and also it is very accurate because along with the automated process of data collection, there was also a manual quality check that removed almost all the errors. Hence, leading to a clean dataset.

We picked up the idea of a week-wise team progression line graph from the paper "State of the Art of Sports Data Visualization" [7]. In this paper, the authors suggested two types of a line graph:

1. The relative team, performance line graph calculates the team's week-wise performance by squeezing it between the max-points and min-points team that week. An issue with this type of visualization was the scale was changing each week, which can confuse the end-user, and teams were colliding in the end. (see figure 6)

2. Rank-based line graph which considers the cumulative points of each team in each week to plot the line graph. We added an idea of goal difference to avoid the clashing of team that end up in same ranks. Finally, as this line graph resolves all of our problems faced in the first one, we went ahead with this option.

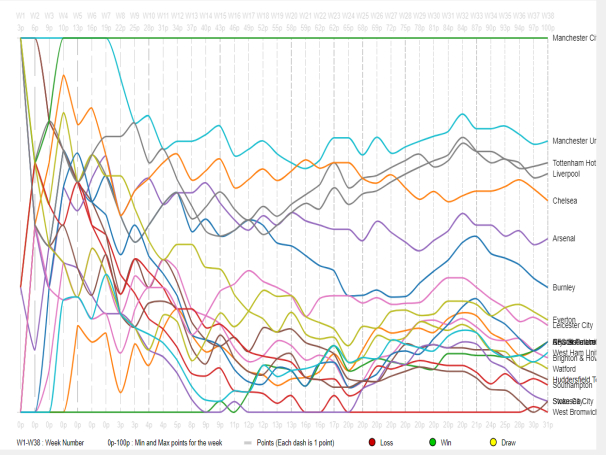

Figure 6: Line graph having changing scale and colliding teams in the end

We referred soccer guru [8] for network graph. We also referred the class materials for building this network graph.

References

[1] The Most Popular Sports In The World (https://www.worldatlas.com/articles/what-are-the-most-popular-sports-in-the-world.html)

[2] 25 World's Most Popular Sports - Ranked by 13 factors (https://www.totalsportek.com/most-popular-sports/)

[3] Top-10 Most Popular Sports In The World 2020 (https://sportytell.com/sports/most-popular-sports-world/)

[4] Soccer match event dataset (https://figshare.com/collections/)

[5] Pappalardo, Luca; Massucco, Emanuele (2019): Soccer match event dataset. figshare. Collection. https://doi.org/10.6084/m9.figshare.c.4415000.v5

[6] Pappalardo, L., Cintia, P., Rossi, A., Massucco, E., Ferragina, P., Pedreschi, D., Giannotti, F. (2019). A public data set of spatio-temporal match events in soccer competitions. Scientific data, 6(1), 1-15.

[7] Perin, C., Vuillemot, R., Stolper, C. D., Stasko, J. T., Wood, J., Carpendale, S. (2018, June). State of the art of sports data visualization. In Computer Graphics Forum (Vol. 37, No. 3, pp. 663-686).

[8] Soccer guru (https://courses.ischool.berkeley.edu/i247/s16/reports/soccerguru.pdf)