Big Data Analytics in Sports - Soccer

Rahul Velayutham
Indiana University Bloomington
2661 E 7th Street Apt H
Bloomington, Indiana 47408
rahul.vela@gmail.com.com

ABSTRACT

Big Data is rapidly becoming a crucial component in the majority of the fields, be it from medicine to software. Big data technologies help in processing humongous amounts of data in a rapid manner while enabling us to achieve results fast and accurately. The impact of Big data in the field of sports, in particular, soccer and how they have helped football clubs evolve their business models and operations from a more hands-on approach to applying complex software and ML models to improve tactics, scouting, and training practices. This study takes a look at the technologies that have been used like MiCoach, Tracab and a look at the leading players like Opta and how the data generated from these companies could be put to use. It is hoped that this study will help demonstrate the importance of Big data in sports, its applications, and avenues for improvement in the field.

KEYWORDS

Big Data, Soccer , Scouting

1 INTRODUCTION

2 BIG DATA IN SOCCER

Big Data has become a crucial part of soccer. Data obtained from big data technologies is used to chart training sessions, shape tactics, predict odds for betting and suggest line ups for fantasy premier leagues. Journalists are increasingly using facts obtained from big data to corroborate their stories and often create new stories when normally none could have existed, for example stats may show a playerfis ineffective performance could be masked by the good form of the team mates around him. We will now go into a little detail of how all this is done.

2.1 Big Data in Scouting

2.1.1 Introduction. Two of the biggest commodities in soccer are the clubs and its players. As mentioned previously transfers are now some of the biggest sources of revenue for clubs. Players fetch for as high as 200 million pounds these days[13]. Also, the quest to find the next big star / the hidden gem against proven expensive players is now a mark of success. Clubs cannot freely go and sign whoever they feel are data monsters, restrictions on the number of players they can sign while at the same time the potential costs that may be involved in the transfer force clubs to make sure the investment they make are the right one. It is recommended to have a look at this article to understand what happens behind the scenes at football clubs when it comes to scouting[4].

2.1.2 Data collection. Most clubs either have acquired specific companies for scouting for example Arsenal FC acquired paid over 2 million pounds for the US company StatDNA, whose data has since been used to advise their signings. [8], and or have scouts who obtain the data themselves. As to how clubs obtain the data, most do not divulge such details to protect their strategies but consensus is that popular sites like opta which analyze matches at real time and release statistics for others to make use of[7]. Alternatively, clubs send performance analysts to feeder clubs and they track matches of prospective candidates and create data for themselves. It is also worth noting that big data has led to only to software development but as well as hardware, for example the Adidas MiCoach a device that tracks metrics and displays it to coaches is used during training and potential scouting sessions. The article mentioned provides an example of how the device was used to realize a gem among a batch of superstars[6].

2.1.3 Data Processing. Most articles only explain in theory how they go on about processing the data and even fewer talk about the technical aspect behind it. Corroborating from different sources [10][5][8] a general theoretical summary can be given, In the case of obtaining data from say the internet i.e., mine data from free sites like squawka, whoscored, opta. Data warehousing technologies like pig, Hadoop etc, can be used. Parsing the XML, one can store this data and applying meaningful ML algorithms with defined parameters to filter players. For example, we can mine the data for fields like chances created, distance covered etc for a league and then filter out say midfielders and chances created in order to find the next best attacking midfielder.

For clubs generating their own data, real time analysis of videos using advanced image processing technologies in tandem with their own hands on analysis they could generate data and store it again or say CSV files. These files then could be uploaded to a private databank. From these banks data warehousing can be once again performed and the previous process can be repeated.

2.2 Big Data in Training and Tactics

In todayfis world which is being driven more and more by capitalistic gains, even the worldfis most famous sport i.e., soccer cannot be spared. Sports players command huge transfer fees, MNCs are pumping billions [millions are soon becoming a thing of a past][2], and as such the even the tiniest mistake can lead to millions lost. Hence, now there is a need to augment daily operations from scouting to coaching level with technology. One of the technologies which fast invading the world of soccer is big data. One can never have enough data, data guides tactics, training session, betting, scouting and so much more. Gone are the archaic days of notes and papers and fispecialistsfi [these specialists do have a very important role to play but with the advancing times they may soon become a thing of the past]. The study looks at two crucial aspects in soccer scouting and training, tactics. In tactics, we look at the new sensation known as fantasy leagues.

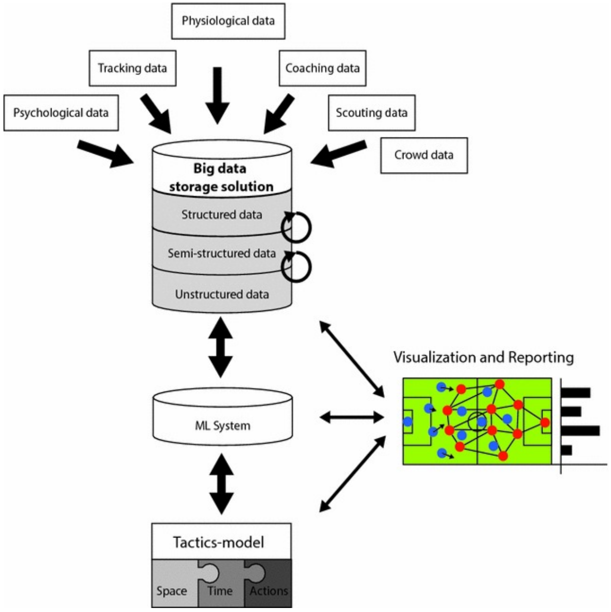
2.2.1 Introduction. Before the advent of big data, coaching was a more personalized hands-on affair, that doesnfit mean it is any less now but the amount is a lot less than before. Preparing for match involved sending scouts and making them watch the match lie and relying on their notes or analyzing videos for hours in hopes of trying to find a weak link. Coaches do spend hours in front of a TV screen but they augment it with software and now look at games from a data sided point of view, an example of this is the former coach of Everton Roberto Martinez [4]. Aside from tactics big data also is slowly invading the field of training sessions big data are being used to create customized training sessions as well as to analyze and mitigate potential injuries.

2.2.2 Data Collection. Data collection here has two aspects to its hardware and software in the previous section the software component was already discussed to a good extent. Now we will shift focus towards the hardware components and their impact/ role in data collection. Below are excerpts from the article [11] which provide excellent insight as to how data is gathered
Athletes are not only monitored by cameras in stadiums, but also by many quirky devices such as accelerometers, heart rate sensors and even local GPS-like systems. for example, the Germans in the world cup held previously in Brazil wore Adidasfi miCoach elite team system during training sessions before and during the competition.[11].The device collects and transmits information directly from the athletefis bodies, including heart rate, distance, speed, acceleration and power, and then display those metrics live on an iPad. All this information is made available live on an iPad to coaches and trainers on the sideline during training, as well as post-session for in-depth analysis. Analysis of the data can help identify the fit players from those who could use a rest.

2.2.3 Data Processing. The article [10] gives a great insight into how data obtained from devices is processed Big data is characterized using the so-called three Vfis: (1) Volume, (2) Variety and (3) Velocity. With respect to tactical analytics in soccer these concepts can be mapped in the following way:
(1) Volume refers to the size of datasets in soccer. For example, a current dataset for positional data typically encoded using Extensible Markup Language (XML) ranges between 86 and 300 megabytes (MB). Thus, storing position, event and video data from a single complete Bundesliga season results in 400 gigabytes of tracking data[10].
(2) Variety refers to different data formats and data sources. Variety can be further distinguished into (a) structured, (b) semi-structured, and (c) unstructured data. Structured data has a clearly predefined schema describing the data. In contrast, unstructured data lacks a definite schema with video data and text messages being typical examples. Semi-structured data falls in between these two extremes and consists of data which lacks a pre-defined structure but may have a variable schema[10].
(3) Velocity describes the speed with which novel data is being generated. In soccer, the velocity varies widely from real-time analysis like in the case of opta to delayed statistics released by journalists etc[10]. From this data generated Machine learning models can be applied to look for anomalies and spot out a weakness in opponents and as well as gauge areas in which the own team requires improvements.

2.3 Tools

In this section, the technological stack and possible tools for implementation are discussed.A candidate big data soccer technological stack for soccer tactics analyses should be organized along several levels.



[10]

First, the necessary infrastructure to collect the data is required. Second, a storage system is required allowing efficient data storage and access. Finally, a processing pipeline has to be established to extract relevant information from the data and to subsequently merge the information to build an explanatory and/or predictive model[10].An in-depth discussion of specific technological solutions is beyond the scope of the present study. A few useful technologies are however discussed, note we will only be discussing the software aspect since the hardware aspects have been discussed in great detail in the previous sections.
It has previously been stressed upon how difficult it is to obtain the details of technologies clubs use to run their daily operations. However, after dissecting it is not all that difficult to make an educated guess on which tools could possibly be used for the above purposes. Let's start with obtaining data in the previous sections we have already seen how opta obtains its data using live analysis[1]. Now we shall explore a new tool called Twitter Heron, which can be used to obtain information from tweets.
One of the problems with opta is that it may not cover leagues/tiers which are not profitable for it. However, football clubs generally have very enthusiastic fan bases and with Twitter being a very convenient social media tool we can try to mine data from tweets to generate our data. Twitter heron is a real-time analytics platform developed by Twitter. It is the direct successor of Apache Storm, built to be backward compatible with Storm's topology API but with a wide array of architectural improvements. Heron supports

Seamless support for different processing semantics, is efficient and scales extremely well. A good blog on why Twitter heron is ideal can be found here.[9]
Previously we touched on the subject of how scouts could use data from opta for analysis, from the internet the best way to obtain such data is to extract from XML. For this many different ways can be used. Some of the most popular manners are using Map Reduce, LogParser and even PIG.
MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. The map takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key/value pairs). Secondly, reduce task, which takes the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.

The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model.[12]
Log Parser is a free command line utility for Windows that allows you to perform queries against a variety of file types including things like log files, CSV files, and XML files. This utility can even parse data sources such as the Active Directory or the Windows Event Logs. Log Parser is extremely flexible, but it is not a utility for novices. Using Log Parser requires experience with custom queries as well as with working from the command line. An example of PIG XML parsing can be found in this blog [3].We can Use Spark SQL for querying data from DBs so that it can be used to extract features and clean up data.

3 CONCLUSIONS

It can be seen the huge impact Big Data has in soccer. It has become a multi-million business. The acquisition of StatDNA by Arsenal for 2 million is proof of that. Also nowadays more and more clubs are being run entirely on big data proof of this is FC Midtjylland (Denmark) and also Brentford FC (England). Matthew Benham and Rasmus Ankersen are the pioneers in data analysis and have completely revolutionized their scouting departments. OPTA is the global leader in stats generation and is rated above 60 Million plus. Aside from just scouting potentials, it is used to shape tactics and also understand the strengths and weakness of players. While it may appear that the industry seems to have less scope of development this is only true for the top-ranked clubs. Most of the mid-table and lower league clubs still make use of traditional methods. The scope for open source software which provides a detailed scouting analysis has a huge market potential.

REFERENCES

[1] Carl Bialik. 2014. The People Tracking Every Touch, Pass And Tackle in the World Cup. *oline* (2014). https://fivethirtyeight.com/features/the-people-tracking-every-touch-pass-and-tackle-in-the-world-cup/

[2] Julie Cooling. na. Investing in Soccer. *na* (na). https://www.forbes.com/sites/juliecooling/2017/03/23/investing-in-soccer/#8b404ce2ec97

[3] learnbigdataanalytics. 2000. Pig XML parsing. *online* (2000). https://learnbigdataanalytics.wordpress.com/hadoop-eco-systems/pig/practice/xml-parsing/

[4] Tim lewis. 2012. How computer analysts took over britains top clubs. *guardian* (2012). https://www.theguardian.com/football/2014/mar/09/premier-league-football-clubs-computer-analysts-managers-data-winning

[5] Will Luca. 20000. Data Driven Football. *online* (20000). http://data-speaks.luca-d3.com/2017/10/data-driven-football.html

[6] Wired micoach. 2000. Big Data devices in football. *online article* (2000). https://www.wired.com/2012/09/major-league-socccer-micoach/

[7] Optasports. 2000. generating live time data. *video* (2000). http://www.optasports.com/about/how-we-do-it/how-we-package-the-data.aspx

[8] Outsideoftheboot. 2000. An insight into the data analysis of football. *newspaper article* (2000). http://outsideoftheboot.com/2015/09/24/insight-into-data-analysis-in-football/

[9] Karthik Ramasamy. 2000. Why Heron? *online* (2000). https://streaml.io/blog/why-heron

[10] Robert Rein and Daniel Memert. 2016. Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *PMC* (2016). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4996805/

[11] Jure Rejec. 2000s. How Big Data is Changing the World of Football. *online* (2000s). https://datafloq.com/read/how-big-data-is-changing-the-world-of-football/1796

[12] Tutorialspoint. 2000. map reduce definition. *online* (2000). https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm

[13] Wikipedia. na. Expensive transfers. *na* (na). https://en.wikipedia.org/wiki/List_of_most_expensive_association_football_transfers

4 BIBTEX ISSUES

Warning–no number and no volume in Bialik2014

Warning–page numbers missing in both pages and numpages fields in Bialik2014

Warning–no number and no volume in cooling

Warning–page numbers missing in both pages and numpages fields in cooling

Warning–no number and no volume in learnbigdataanalytics2000

Warning–page numbers missing in both pages and numpages fields in learnbigdataanalytics2000

Warning–no number and no volume in 2012

Warning–page numbers missing in both pages and numpages fields in 2012

Warning–no number and no volume in Luca20000

Warning–page numbers missing in both pages and numpages fields in Luca20000

Warning–no number and no volume in 2000

Warning–page numbers missing in both pages and numpages fields in 2000

Warning–no number and no volume in Optasports2000

Warning–page numbers missing in both pages and numpages fields in Optasports2000

Warning–no number and no volume in Outsideoftheboot2000

Warning–page numbers missing in both pages and numpages fields in Outsideoftheboot2000

Warning–no number and no volume in Ramasamy2000

Warning–page numbers missing in both pages and numpages fields in Ramasamy2000

Warning–no number and no volume in Rein2016

Warning–page numbers missing in both pages and numpages fields in Rein2016

Warning–no number and no volume in Rejec2000s

Warning–page numbers missing in both pages and numpages fields in Rejec2000s

Warning–no number and no volume in Tutorialspoint2000

Warning–page numbers missing in both pages and numpages fields in Tutorialspoint2000

Warning–no number and no volume in wikipedia

Warning–page numbers missing in both pages and numpages fields in wikipedia

(There were 26 warnings)

5 ISSUES

5.1 Formatting

Incorrect number of keywords or HID and i523 not included in the keywords

Other formatting issues: No introduction?

5.2 Details about the Figures and Tables

Capitalization errors in referring to captions, e.g. Figure 1, Table 2