Soccer2014DS: a dataset containing player events from the 2014 World Cup

Marcos Roberto Ribeiro, Maria Camila N. Barioni, Sandra de Amo, Claudia Roncancio, Cyril Labbé

Soccer2014DS: a dataset containing player events from the 2014 World Cup

Marcos Roberto Ribeiro[1,2], Maria Camila N. Barioni[2], Sandra de Amo[2], Claudia Roncancio[3], Cyril Labbé[3]

[1] Instituto Federal de Minas Gerais (IFMG), Bambuí, Brazil

[2]Universidade Federal de Uberlândia (UFU), Uberlândia, Brazil

[3]Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000, Grenoble, France

marcos.ribeiro@ifmg.edu.br, {camila.barioni, deamo}@ufu.br,
{claudia.roncancio, cyril.labbe}@imag.fr

Abstract. *The player monitoring has become a common task in many sports. However, there is no public datasets of detailed soccer player events. Thus the creation of such datasets can be useful for diverse research fields such data mining, sports analytics and continuous preference queries. In this paper, we describe the construction of the dataset Soccer2014DS containing player events of the 2014 Soccer World Cup. This dataset is composed by the raw extracted data collected by a web crawler and by derived streams with new calculated attributes. We also explain how we are using this dataset in experiments related to the development of a new query language.*

1. Introduction

The monitoring of sport players during matches started in the end of XX century [Ali and Farrally 1991]. In this period, the researchers were more concerned to collect just health data from players. Since 2000, the monitoring tasks became more sophisticated due to the development of new technologies for GPS devices and softwares for video processing [Baca et al. 2009, Barris and Button 2008]. This new technologies allow to collect complex data to be used on detailed analysis of player events.

Despite the players monitoring be very common in official competitions, there are few public datasets with these information available. If we consider the soccer sport, to the best of our knowledge, there exist no public datasets with detailed player events. Thus, our main goal herein is to present the public dataset *Soccer2014DS* which contains player events of the 2014 Soccer World Cup. This dataset can be useful for diverse research fields such as data mining [Bialkowski et al. 2014, Gyarmati and Hefeeda 2015], sports analytics [Lucey et al. 2013, Perin et al. 2013] and continuous queries [Arasu et al. 2016].

This paper is organized as follows. Section 2 describes the original extracted data. Next, Section 3 presents the data streams derived from the original data. Section 4 discusses about the research opportunities and Section 5 explains the dataset limitations. Finally, Section 6 concludes the paper.

2. Data Sources

The creation of the *Soccer2014DS* dataset started with the extraction of the original data available on the Huffpost Data web site[1] [Boice et al. 2014]. This web site contains information about the 2014 Soccer World Cup provided by the company Opta Sports[2]. These

[1]http://data.huffingtonpost.com/2014/world-cup

[2]http://www.optasports.com/

data is used to display statistics and graphs about player events. Every match has an individual page where a user can delimit a time-line and see the details of this selection. Our first task was to study the source code of the Huffpost Data web site. Based on this study, we developed a web crawler to extract the data. The crawler starts the extraction in the page of the final match and follows the links to the remaining matches to complete the data collecting.

The extraction of the original raw data was performed in 2015. After this task, we organized the extracted data into the relations `Matches`, `Teams` and `Players` and the stream `Events`. The relations have just one instance and the duplicated data from all matches was eliminated. On the other hand, the stream `Events` has 64 instances (one instance per match) preserving all extracted event data. Appendix B presents the logical schema of the dataset (the derived data is addressed in Section 3).

The attributes of the relation `Matches` are `id` (match identifier), `date`, `time`, `venue` and `attendance`. Table 1(a) displays the attributes of the stream `Events`. The player coordinates (`x` and `y`) and final coordinates of the ball (`to_x`, `to_y` and `to_z`) are expressed as a percentage of the field dimensions. The attributes `type` and `outcome` are used to identify the move performed by players. Appendix A presents the events associated to every combination of values for these attributes. The attribute `field_pass` represents the continuous ball possession, t for true and f for false. The attribute `side` is the field side of the team, H for left side and A for right side. When the move is a pass to another player, the attribute `to` assumes the identifier of this player. For streams, we also must associate a `timestamp` for every tuple [Arasu et al. 2016], in the stream `Events` the `timestamp` is calculated using the attributes `min` and `sec` (`timestamp` = `min` $\times$ 60 + `sec`).

Table 1. Relation attributes: (a) `Events` (b) `Players`

(a)

Attribute	Description
`id`	Event identifier
`period`	Period of math
`min, sec`	Minute and second of the event
`displaymin`	Displayed minute
`team`	Team identifier
`player_id`	Player identifier
`x, y`	Player coordinates
`type`	Type of event (move performed)
`outcome`	Result of the event
`field_pass`	Continuous ball possession
`side`	Field side of team
`to_x, to_y, to_z`	Final coordinates of ball
`to`	Player identifier of next move

(b)

Attribute	Description
`id`	Player identifier
`name`	Player name
`real_position`	Detailed position
`real_position_side`	Position side
`known_name`	Known name
`short_name`	Short name
`last_name`	Last name
`first_name`	First name
`middle_name`	Middle name
`team_id`	Team identifier
`preferred_foot`	Preferred foot
`club`	Club
`caps`	Matches played by player team
`goals`	Goals
`jersey_num`	Jersey number
`country`	Birth country
`birth_date`	Birth date
`position`	Position

The relation `Teams` is composed by the attributes `id` (team identifier), `name` and `iso` (ISO acronym). Table 1(b) presents the attributes of the relation `Players`. The values for the attributes `real_position`, `real_position_side`, `preferred_foot` and `position` are shown in Table 2. Please see [Bakker 2015] for

more details about the data gathered by the company Opta Sports.

Table 2. Values for `Players` attributes

Attribute	Values
`real_position`	Attacking Midfielder, Central Defender, Central Midfielder, Defensive Midfielder, Full Back, Goalkeeper, Second Striker, Striker, Wing Back, Winger
`real_position_side`	Centre, Centre/Right, Left, Left/Centre, Left/Centre/Right, Left/Right, Right, Unknown
`preferred_foot`	Both, Left, Mostly Left, Mostly Right, Right, (empty)
`position`	Defender, Forward, Goalkeeper, Midfielder

3. Derived Streams

After the data extraction described in the previous section, we created derived streams by applying cleaning and conversions over the original data. As described in the previous section, in order to know the exact player move, we must check the attributes `type` and `outcome` of the relation `Events`. In addition, the coordinates of the player and the ball, expressed by float values, could not be suitable for some applications where the user has to indicate a region of the soccer field. To deal with this situation, we decided to create new derived streams: `Moves(player_id, place, move)` for the performed moves; and `Places(player_id, place, ball, direc)` for the player positioning. The logical schema presented in Appendix B shows the relationship of the derived streams and the original relations.

The stream `Moves` contains just the moves performed by the players. So, we do not consider events without ball like cards, substitutions, etc. More precisely, the events types 17, 18, 19, 34, 43, 58, 60, 102 and the events with `(type, outcome)` equal to (5, 1), (6, 1), (53, 1) and (57, 1) are ignored. For the remaining move types, we use the mapping from (`Events.type`, `Events.outcome`) to `Moves.move` described in Table 3.

Table 3. Move mapping

`move`	`(type, outcome)`
pass	$(1,1)$, $(59,1)$
bpas (bad pass)	$(1,0)$, $(2,1)$
lbal (lost ball)	$(3,0)$, $(7,0)$, $(44,0)$, $(50,1)$, $(51,1)$, $(57,0)$, $(59,0)$, $(61,0)$
drib (dribble)	$(3,1)$, $(42,1)$
foul	$(4,0)$
fsuf (foul suffered)	$(4,1)$, $(55,1)$
dled (dribbled)	$(45,0)$
bout (ball out)	$(5,0)$, $(6,0)$
brec (ball recovery)	$(7,1)$, $(44,1)$, $(49,1)$, $(56,1)$, $(61,1)$
int (interception)	$(8,1)$, $(74,1)$
gsav (goalkeeper save)	$(10,1)$, $(11,1)$, $(41,1)$, $(52,1)$, $(54,1)$
clea (clearance)	$(12,1)$
wsho (wrong shot)	$(13,1)$, $(14,1)$, $(15,1)$
goal	$(16,1)$
rec (reception)	$(100,1)$
cond (conduction)	$(101,1)$

The possible values for the attribute `place` are defensive area (da), defensive intermediary (di), middle-field (mf), offensive intermediary (oi) and offensive area (oa). Figure 1 displays how we compute these values according to the attribute `Events.x` represented by dashed lines. The attribute `ball` (ball possession) is a mapping from the at-

tribute `field_pass`. If `field_pass` $= t$ then `ball` $= 1$, and if `field_pass` $= f$ then `ball` $= 0$. In order to compute de attribute `direc` (move direction of a player), we consider the previous and the current place of each player. Next, we calculate the horizontal distance (*xdist*) and vertical distance (*ydist*) between these places. When *ydist* $= 0$ and *xdist* $= 0$, the direction is *none* since the player did not move. If *ydist* $>$ *xdist* then the direction is *lateral*. Otherwise, the direction is *backward* (for *xdist* < 0) or *forward* (for *xdist* > 0).

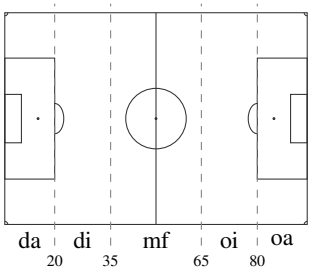

Figure 1. Soccer field division

The full dataset and the importing tool are available for download in a Github repository[3]. This repository also provides additional tools for benchmarking continuous queries with temporal conditional preferences. All tools were developed using the python language. The dataset is stored in the directory *data*. This directory contains the relations `Matches`, `Teams` and `Players` stored in CSV (comma separated values) format. Such relations are the union of the data extracted from all matches. The *data* directory also contains the subdirectories *raw*, *events*, *moves* and *places* to store respectively the extracted raw data, the stream `Events`, the stream `Moves` and the stream `Places`.

The subdirectory *raw* has, for every match, the JSON files *match_id.json*, *match_id-players.json* and *match_id-teams.json* where *match_id* is the match identifier. The subdirectories *events*, *moves* and *places* contain the CSV files *match_id.csv* with the stream data of the matches. Table 4 presents the information about tuples and instances for all relations and streams of the dataset. The streams `Moves` and `Places` have less tuples than stream the `Events` due to the data cleaning and the computation of the new attributes.

Table 4. Data statistics

Relation/Stream	Tuples	Instances	Tuples/Instance
`Matches`	64	1	64
`Teams`	32	1	32
`Players`	736	1	736
`Events`	167,801	64	2621
`Moves`	130,607	64	2040
`Places`	137,621	64	2150

4. Research Opportunities

As we mentioned in the first section, our dataset is useful for a multitude of studies. This section outlines a non-exhaustive list of research fields that can benefit from use our dataset.

[3] `https://streampref.github.io/wcimport/`

Data Mining. In the data mining field, our dataset can be explored for the validation of new techniques aiming the detection of temporal patterns and key events. As the dataset has the player coordinates, we can also discover the special relations between this patterns and some specific field regions. Using additional information about the localization of the matches, it could be possible to find player patterns correlated to environmental variables.

Sport Analytics. Our dataset can be used to make various sport analysis over individual players or teams. Using our dataset, specialists are able to analyze the moves performed, the player positioning, pass distances and many others variables related to a specific player. In addition, the information of all team players can be combined to perform analysis over the team strategy and positioning.

Preference Continuous Queries. An interesting research topic in the field of data streams processing is to incorporate temporal conditional preferences into continuous queries [Ribeiro et al. 2017b]. This new kind of query makes use of the implicit temporal information of data streams to select sequences of elements that best fit user preferences. As the tuples of data streams has a *timestamp*, *we can know the order of the tuples*. So, by using temporal preferences, the user can express wishes like "if there exists a value X in the past then I prefer a value Y to a value X in the current moment".

Unlike traditional databases, data stream applications do not store all data due to limitations of time and space. The existence of datasets for data streams scenarios are important to allow the validation of new techniques for the evaluation of continuous queries. There are works that proposed synthetic data generators [Bifet et al. 2011], but the real datasets are still useful for many specific situations.

In the work [Ribeiro et al. 2017a] we proposed a preference model for reasoning with temporal conditional preferences on data stream scenarios. The *Soccer2014DS* dataset was used in the experiments to demonstrate the effectiveness of our proposed approach. We also used the *Soccer2014DS* dataset in the work [Ribeiro et al. 2017b]. In this latter work, the dataset was used to conduct an extensive set of experiments to compare the performance and the memory usage of algorithms to process continuous queries with temporal conditional preferences.

5. Limitations

The real soccer datasets collected by Opta Sports have additional information provided by an special attribute (`qualifiers`). However, these datasets are not public. The extracted data has this attribute, but it has no meaningful values. So, we drop this attribute from our dataset.

Our dataset does not have the coordinates of all players at every second. Only events related to moves, cards, fouls and ball outs were gathered. So, positioning analysis must take this information into consideration.

We calculated just the attributes `move`, `place`, `ball` and `direc` for the derived streams `Moves` and `Places`. Although, new attributes can be computed using the available information in the stream `Events`, for example, the distance of passes and shots. In addition, we decided to keep each match into an individual stream, but it is possible to join these data into a single stream if it is a requirement of the data analysis task.

6. Conclusion

In this paper we described the creation of the public dataset *Soccer2014DS* containing player events of the 2014 Soccer World Cup. The construction of the dataset started with

the extraction of data from *Internet* using a web crawler. This extracted data was used to create the derived data by applying cleaning and conversions techniques.

The dataset and all the developed tools are available for download in a public repository. So, the dataset can be used on the development of new research works related to data mining, sports analytics and continuous queries. We already use the *Soccer2014DS* dataset on our previous works [Ribeiro et al. 2017a, Ribeiro et al. 2017b] and we are still using this dataset on new researches about continuous queries with temporal conditional preferences.

References

Ali, A. and Farrally, M. (1991). Recording soccer players' heart rates during matches. *Journal of Sports Sciences*, 9(2):183–189.

Arasu, A., Babcock, B., Babu, S., Cieslewicz, J., Datar, M., Ito, K., Motwani, R., Srivastava, U., and Widom, J. (2016). *STREAM: The Stanford Data Stream Management System*, pages 317–336. Springer, Berlin, Germany.

Baca, A., Dabnichki, P., Heller, M., and Kornfeind, P. (2009). Ubiquitous computing in sports: A review and analysis. *Journal of Sports Sciences*, 27(12):1335–1346.

Bakker, L. F. B. C. (2015). Visualizing football team strategies and player performance. Master's thesis, Eindhoven University of Technology, Eindhove, Netherlands.

Barris, S. and Button, C. (2008). A review of vision-based motion analysis in sport. *Sports Medicine*, 38(12):1025–1043.

Bialkowski, A., Lucey, P., Carr, P., Yue, Y., Sridharan, S., and Matthews, I. (2014). Large-scale analysis of soccer matches using spatiotemporal tracking data. In *International Conference on Data Mining (ICDM)*, Shenzhen, China.

Bifet, A., Holmes, G., Pfahringer, B., Read, J., Kranen, P., Kremer, H., Jansen, T., and Seidl, T. (2011). MOA: A real-time analytics open source framework. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, pages 617–620, Athens, Greece.

Boice, J., Fung, H., and Bycoffe, A. (2014). URL: `http://data.huffingtonpost.com/2014/world-cup` (visited on 23/04/2015).

Gyarmati, L. and Hefeeda, M. (2015). Estimating the maximal speed of soccer players on scale. In *Machine Learning and Data Mining for Sports Analytics Workshop*, Porto, Portugal.

Lucey, P., Oliver, D., Carr, P., Roth, J., and Matthews, I. (2013). Assessing team strategy using spatiotemporal data. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1366–1374, Chicago, IL, USA.

Perin, C., Vuillemot, R., and Fekete, J.-D. (2013). Soccerstories: A kick-off for visual soccer analysis. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2506–2515.

Ribeiro, M. R., Barioni, M. C. N., de Amo, S., Roncancio, C., and Labbé, C. (2017a). Reasoning with temporal preferences over data streams. In *International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, Marco Island, Florida.

Ribeiro, M. R., Barioni, M. C. N., de Amo, S., Roncancio, C., and Labbé, C. (2017b). Temporal conditional preference queries on streams. In *International Conference on Database and Expert Systems Applications (DEXA)*, Lyon, France. (to appear).

Appendix

A. Events Associated to Values of Attributes Type and Outcome

type	outcome	Event description
1	0	Non completed pass
	1	Completed pass
2	1	Pass to offside player
3	0	Dribble losing ball
	1	Successful dribble
4	0	Foul committed
	1	Foul suffered
5	0	Ball out
	1	Wined throw-in or goal kick
6	0	Ball out on goal line
	1	Wined corner kick
7	0	Dispossessed opponent without possession
	1	Dispossessed opponent with possession
8	1	Interception
10	1	Goalkeeper save shot
11	1	Goalkeeper catches crossed ball
12	1	Clearance (shot out defensive zone)
13	1	Miss (shot out goal)
14	1	Post (shot on goal frame)
15	1	Attempt saved by other player
16	1	Goal
17	1	Card
18	1	Player substituted
19	1	Player comes on (as substitute)
34	1	Player line up and formation
41	1	Goalkeeper punches ball
42	1	Skill on the ball
43	1	Deleted event
44	0	Lost aerial duel
	1	Wined aerial duel
45	0	Player dribbled
49	1	Ball recovery
50	1	Player dispossessed
51	1	Error (causing ball dispossession)
52	1	Goalkeeper picks up the ball
53	1	Cross not claimed by goalkeeper
54	1	Goalkeeper advances and wins possession
55	1	Offside provoked
56	1	Shield ball from opponent
57	0	Player causes throw-in reversion
	1	Player wins throw-in reversion
58	1	Goalkeeper faced to penalty kick
59	0	Goalkeeper clears ball without possession
	1	Goalkeeper clears ball with possession
60	0	Goal chance missed
61	0	Player touches ball and without possession
	1	Player touches ball and with possession
74	1	Accidental blocking
100	1	Ball reception
101	1	Ball conduction
102	1	Start playing

B. Soccer2014DS Logical Schema