Automatic Discovery of Tactics in Spatio-Temporal Soccer Match Data

Tom Decroos
KU Leuven
Leuven, Belgium
tom.decroos@cs.kuleuven.be

Jan Van Haaren
SciSports
Amersfoort, Netherlands
j.vanhaaren@scisports.com

Jesse Davis
KU Leuven
Leuven, Belgium
jesse.davis@cs.kuleuven.be

ABSTRACT

Sports teams are nowadays collecting huge amounts of data from training sessions and matches. The teams are becoming increasingly interested in exploiting these data to gain a competitive advantage over their competitors. One of the most prevalent types of new data is event stream data from matches. These data enable more advanced descriptive analysis as well as the potential to investigate an opponent's tactics in greater depth. Due to the complexity of both the data and game strategy, most tactical analyses are currently performed by humans reviewing video and scouting matches in person. As a result, this is a time-consuming and tedious process.

This paper explores the problem of automatic tactics detection from event-stream data collected from professional soccer matches. We highlight several important challenges that these data and this problem setting pose. We describe a data-driven approach for identifying patterns of movement that account for both spatial and temporal information which represent potential offensive strategies. We evaluate our approach on the 2015/2016 season of the English Premier League and are able to identify interesting strategies per team related to goal kicks, corners and set pieces.

CCS CONCEPTS

• **Information systems** → **Data mining**; **Clustering**; **Data stream mining**; • **Computing methodologies** → *Artificial intelligence*; *Unsupervised learning*; *Cluster analysis*;

KEYWORDS

Sports analytics, Eventstream data, Soccer match data, Pattern mining, Tactics discovery

ACM Reference Format:
Tom Decroos, Jan Van Haaren, and Jesse Davis. 2018. Automatic Discovery of Tactics in Spatio-Temporal Soccer Match Data. In *KDD '18: The 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, August 19–23, 2018, London, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3219819.3219832

KDD '18, August 19–23, 2018, London, United Kingdom

ACM ISBN 978-1-4503-5552-0/18/08…$15.00
https://doi.org/10.1145/3219819.3219832

1 INTRODUCTION

When preparing for an upcoming match, video analysts typically spend many hours watching video footage to better understand the tactics of their opponent. This is a very time-consuming and tedious process, which could be considerably sped up and improved by leveraging the large amounts of data that are available nowadays. However, soccer clubs lack the computational methods that can handle both the size and complexity as well as the spatial and temporal aspects of the data in a natural way. Therefore, most soccer clubs restrict themselves to computing simple descriptive statistics such as the number of shots on target or the number of tackles. As soccer is a highly dynamic game, there is obviously much more value in gaining a better understanding of the many complex interactions among players, which simple descriptive statistics cannot capture. Hence, there has been an explosion of interest in applying automated techniques to analyze data collected about sports matches (e.g., [1, 3, 12–14, 18]).

This paper focuses on the problem of detecting strategies from professional soccer matches based on spatio-temporal data. This problem poses a number of significant challenges from a data mining perspective. First, important patterns will involve both spatial (i.e., a location on the pitch) and temporal (i.e., a timestamp or order) components. Second, there will rarely be exact matches in terms of the same set of players performing the same actions in the same order in the same locations. Hence, it is necessary to generalize over both spatial locations and which players perform which action. The combination of the previous two facts can lead to a greatly increased search space. Third, there is rich domain knowledge about soccer (e.g., what events are important or different ways of characterizing passes) that can be exploited to guide the discovery process. Fourth, frequency is not necessarily the most important criteria for interestingness in tactics detection. Certain events such as goals and shots are rare, and sequences involving them are correspondingly more valuable and interesting.

We propose an approach to find patterns in professional soccer matches. On a high level, our approach performs the following five steps. First, the algorithm splits the match into phases, which are uninterrupted sequences of events where one team is in possession of the ball. Second, we cluster these phases by their spatio-temporal characteristics. Third, we rank each cluster according to their expected relevance to the user. Fourth, we search for frequently occurring patterns (i.e., sequences of events) within each cluster. Fifth, based on domain knowledge, we develop a ranking function that orders the discovered patterns in each cluster according to their expected relevance to the user.

We evaluate our approach on data from the 2015/2016 season of the English Premier League, where we have access to event stream

data for all matches. We let a domain expert inspect the discovered patterns and find that our approach is capable of identifying interesting tactics. Furthermore, we evaluate how each of our design choices contributes to the overall performance of the system.

2 RELATED WORK

This paper falls within the emerging area of work that looks at analyzing spatio-temporal sports data. Knauf et al. [12, 13] proposed a novel spatio-temporal kernel for clustering player trajectory data. Their kernel is able to consider multiple different trajectories simultaneously, which is important for capturing tactics. Furthermore, it is based on the solid theoretical foundations of kernels. In several context-specific scenarios, such as play initiation and attacks, the approach identified interesting clusters that are illustrative of differences between two team's playing styles. Another trajectory-based approach, which focuses on scoring opportunities [7], clusters together different scoring opportunities based on hand-crafted features as well as trajectory information about players on both teams. The goal is to assess how effective a team is at creating chances from certain types of situations (e.g., corners). Another approach to characterizing scoring chances is based on inductive logic programming [18], which allows representing rich, relational structure in a domain. This work focuses on discriminative mining of event streams to find patterns of play that are more likely than not to lead to shots on goal, but primarily focuses on capturing (hierarchical) spatial relations. Van Haaren et al. [17] used a similar approach in hockey to detect patterns of play that occur frequently in won rallies.

Another way to analyze tactics is to build occupancy maps based on ball movements [14]. The occupancy maps are then used to assess how predictable the ball's movement is within a given region of the field. While the approach does provide a characterization of team behavior, it does not yield insight into specific trajectories or patterns of movement that a team employs to generate attacks. Another line of work looks at trying to characterize playing style and tactics by looking at passing patterns. One approach is to look at the passing graph and look for common passing sequences, that is, sequences of a given length that involve mostly the same players passing in mostly the same order [11]. However, it ignores the spatial component. Another approach looks at identifying frequent passing patterns by applying dynamic time warping [10]. Beyond these, other tactics analyses include recognizing team formations in soccer (e.g., [1]) or identifying specific plays in American football (e.g., [16]).

In terms of data mining tasks, related areas include trajectory mining [9] and finding frequent spatio-temporal patterns [2]. In contrast to trajectory mining, the transition time between different events is not as important in our case. Both of these approaches also take the typical pattern mining approach of focusing on identifying frequent patterns. For example, a team's defenders may pass the ball among themselves, simply because they are trying to kill time. While this would result in a frequent pattern that represents a tactic ("kill time"), it does not provide significant insight into the more important strategic decisions such as how does a team build up the attack. Furthermore, frequency also ignores the fact that certain sequences are inherently more interesting. In soccer, events like goals, shots, and getting the ball in a dangerous area are very infrequent, but incredibly important. Hence, it is natural to assign more or less weight to a sequence based on how interesting and valuable the individual events within the pattern are. Importantly, spatio-temporal patterns in sports are not the result of a single object moving around. Instead, they arise from a complex, dynamic environment where many factors such as interactions among multiple different players across space and time, and features of the game state (e.g., score, field position, time left, and team quality) influence decision making and tactics. Our approach is able to account for some of these factors.

Table 1: Special types of passes and their frequencies

Event type	Frequency
pass	100.00%
normal pass	67.72%
long ball	16.52%
head pass	8.03%
throw in	4.78%
cross	4.24%
free kick	2.47%
corner	1.11%
through ball	0.28%

3 DATASET

Our dataset consists of event data for the English Premier League for the 2015/2016 season. This event data was manually collected by humans who watch video feeds of the matches through special annotation software. Each time an event happens on the pitch, the human annotates the event with, amongst others, a timestamp, the location (i.e., a (x, y) position), an appropriate type (e.g., foul, pass, or cross) and the players who are involved. Depending on the type of the event, additional information is available. For example, the end location and type of a pass or the outcome of a tackle.

Our dataset contains 652,907 events of 39 different types, which are related to either the flow of the match (e.g., a player substitution, a yellow card, an awarded corner) or the action on the pitch (e.g., a shot, a clearance, or a foul). This corresponds to an average of 1,718 events per match. The most frequent event types in our dataset are "pass" (368,426), "out" (48,046), and "ball recovery" (41,448).

A special type of event are passes. These can sometimes be of a more specific type, such as "cross". This is stored as additional information. Table 1 shows an overview of the type of special passes, along with their frequency. Note that sometimes passes can have multiple specific types. For example, a pass can be both a "corner" and a "cross" at the same time.

Another special type of event are shots. These are registered in the data as four different event types based on their outcome: goal, miss, attempt saved or post (i.e., the ball hit the post). Hence, if we want to analyze the shots in our database, we have to aggregate over these event types.

4 APPROACH

The goal of our approach is to identify common attacking strategies employed by a specific team. Formally, our task can be defined as:

Given: A set of matches, where each match is represented as an event sequence, about a team of interest.
Find: Relevant spatio-temporal patterns that characterize attacking strategies.

This is a challenging task as soccer is a highly dynamic game with many movements and interactions among players across time and space. Concretely, the challenges include the following:

Challenge 1: Strategy involves coping with both a spatial component, as the location where an event occurs is important, as well as a temporal component, as the order of events is important.
Challenge 2: The various events that occur during a game are described by both discrete attributes, such as the type of event and the players involved, as well as continuous attributes, like the event location.
Challenge 3: There is very little exact repetition in sequences of game play. That is, the same players rarely perform the same actions in the same order in the same locations. There will often be variations in the location of players or events, and the players involved in the events. This makes counting occurrences difficult, as the counting must generalize over both spatial locations and which player performed an action.
Challenge 4: There are a significant number of representational choices to consider, particularly as it relates to encoding domain knowledge about soccer. Two examples include information about formations and different ways to define player positions some of which involve hierarchical information.
Challenge 5: Identifying interesting and relevant patterns is a highly subjective decision. Certain sequences of play, such as high quality attempts on goal occur infrequent, but are of high interest.
Challenge 6: Teams do not employ a single tactic. Each team has several different tactics during a game. Furthermore, each tactic has minor variations.
Challenge 7: No uniform definitions for events in soccer matches exist. For example, it is unclear when a shot is considered a shot, or a cross is considered a cross. An intended cross might accidentally end up in the goal, while an intended shot can end up as a crossed assist for another player. Also, the definitions might differ from one human annotator to another.

To tackle these challenges, we perform the following five steps.

(1) Divide the event stream of each match into phases.
(2) Cluster the phases on their spatio-temporal component.
(3) Rank the clusters based on the preferences of the user.
(4) Mine each of the obtained clusters to identify frequent sequential patterns.
(5) Rank the discovered patterns based on the preferences of the user.

The following subsections discuss each of these five steps in more detail.

4.1 Dividing a match event stream into phases

Formally, each match M is represented as a sequence of events $M = e_1, \ldots, e_n$, where each e_i is an event and n is the total number of events. Each event e_i is a tuple $e_i = (t, l, p, et)$, where t is a timestamp, l is the location on the pitch where the event took place as given by x and y coordinates, p is the set of players involved in the event, and et is the event type. We use ET to refer to the set of all possible event types.

In terms of tactics, a sequence representing an entire match represents too coarse of a granularity to consider for analysis. Strategies will manifest themselves as short, consecutive sequences of actions on the pitch such as attacking through the middle or playing a through pass. Therefore, a more natural unit to analyze is what a domain expert may call a "soccer gameplay phase" or simply a phase (e.g., a corner, an attack from the left flank, a turnover). A phase is a sequence of consecutive events that fit together. An added benefit of a phase presentation is that it is easier to find patterns in multiple shorter event sequences than one long event sequence. Therefore, we split the event stream of a match into phases. Figure 1 shows an example of a phase.

Each match M is subdivided into subsequences $P_1, \ldots, P_m$, where each P_j is a phase and m is the total number of phases in the match. Each phase $P_j = e_{j_1}, \ldots, e_{j_p}$ is a subsequence of consecutive events that appear in the sequence M. A new phase starts if there

(1) Is a pause of at least 10 seconds between events; or
(2) Possession switches from one team to the other (e.g., a successful tackle, the ball goes out of play for a throw-in or corner kick, a goal is scored, or a free kick is awarded).

This approach was shown to lead to interpretable self-contained phases in earlier work [4].

We only consider phases that have at least three events. Phases with only one or two events are usually not very informative of the playing style of a team. Figure 2 shows the distribution of phase lengths for Manchester City in the 2015-2016 English Premier League season. Finally, like in many sports, teams switch which goal they are attacking at halftime. Comparing phases that attack different goals is difficult, so we normalize each phase such that the team of interest is always attacking the same goal.

Another approach to divide the match event stream into phases is to divide the event stream into subsequences of constant length (e.g., windows of length 10 seconds) [3]. While this approach is more straightforward, it has two important drawbacks. First, the time between two consecutive events can differ greatly from one match to another due to a difference in intensity and the unreliability of human annotators. This would lead to many uninformative phases being constructed. Second, using this approach, many phases would contain events of both teams in the match, which makes it harder to infer the tactics of one specific team.

4.2 Clustering phases on their spatio-temporal component

The goal of the second step is to identify similar spatio-temporal phases via clustering. We do this for two reasons. One, this helps reduce the space of possible patterns that we need to search in step four. Two, a team is likely to employ multiple different attacking strategies, such as corners, attacking through the middle, down

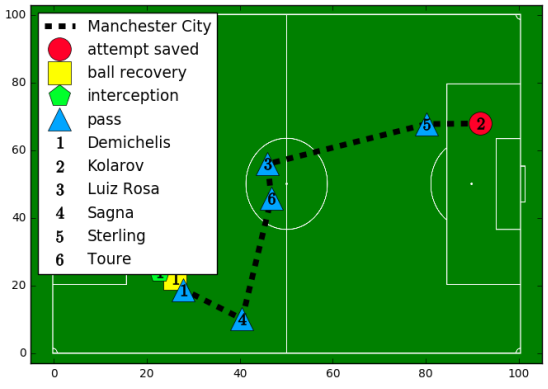

Figure 1: An example phase in our data. A phase is a sequence of consecutive events that fit together according to a domain expert.

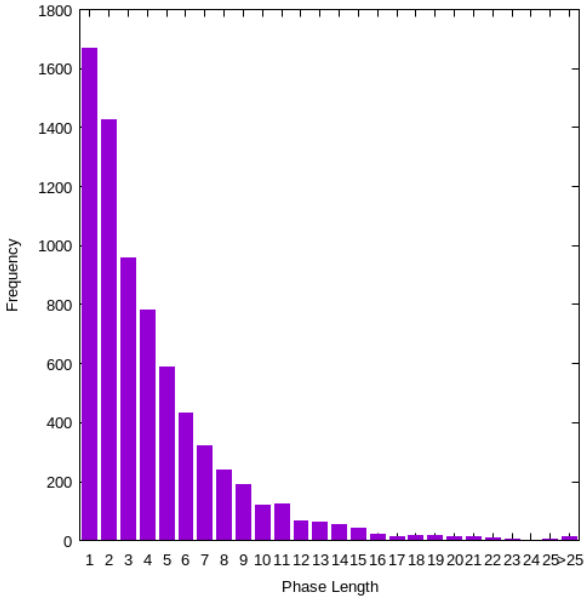

Figure 2: Distribution of phase lengths for Manchester City in the 2015-2016 English Premier League season.

the flank, each of which will be characterized by different spatial characteristics. Clustering gives us a natural way to divide the data along these lines.

In this paper, we use hierarchical agglomerative clustering, which is a popular and simple approach for clustering data [6]. To measure the distance between two clusters, we use the complete-linkage metric. Complete-linkage clustering tends to find compact clusters of approximately equal diameters but not necessarily equal amount of objects [6], which is precisely the type of clustering we want.

The clustering works as follows. First, each element is assigned its own cluster. Next, clusters are iteratively merged together until a stop criteria is met. In each iteration, the two clusters separated by the shortest distance are combined. Complete-linkage clustering computes the distance between two clusters C_i and C_j as the distance between those two elements (one in each cluster) that are farthest away from each other. In this work, we stop once there are k clusters remaining, where k is a user-defined parameter.

A crucial step in clustering objects is choosing the right distance function. In our setting, we have to cope with the fact that phases are of varying length. Furthermore, we want to identify spatially similar phases. One well known way to cope with these two desiderata is to use Dynamic Time Warping (DTW) [15]. DTW computes a cost by finding a one-to-many matching between two sequences. The matching allows for a non-linear warping effect when aligning the two sequences. Consequently, DTW is able to cope with minor mismatches between sequences, such as delays or shifts. A drawback to DTW is that it is not a distance function as it does not satisfy the triangle inequality.

The most natural way to explain and compute the DTW cost is via dynamic programming. In our case, given a phase P_1 of length m and a phase P_2 of length n, the DTW cost can be computed as:

$$D[i,j] = \delta(P_{1_i}, P_{2_j}) + \min(D[i-1,j-1], D[i,j-1], D[i-1,j])$$

where D is a $m \times n$ matrix and $\delta(P_{1_i}, P_{2_j})$ computes the cost of aligning the i^{th} element of P_1 with the j^{th} element of P_2 as:

$$\delta(P_{1_i}, P_{2_j}) = \sqrt{(P_{1_{i,x}} - P_{2_{j,x}})^2 + (P_{1_{i,y}} - P_{2_{j,y}})^2}$$

where $P_{1_{i,x}}$ ($P_{2_{j,x}}$) gives the x coordinate of i^{th} event in P_1 (j^{th} event in P_2) and $P_{1_{i,y}}$ ($P_{2_{j,y}}$) gives the y coordinate of i^{th} event in P_1 (j^{th} event in P_2) y. That is, the cost function is only considering the spatial proximity of the events in two phases. The final DTW cost is given by $D[m,n]$.

In earlier work [19], we clustered the phases based on their possession map, which is a grid overlaying the field that shows how often each area of the field was occupied by the players and the ball during a phase. Using DTW as a distance measure has two benefits over our earlier work: DTW is simpler and takes the temporal component of the phases more explicitly into account.

4.3 Ranking clusters

Next, we rank clusters according to their expected relevance to the user. Typically, the quality of clusters is judged by statistics such as average pairwise distance, maximal pairwise distance and minimal pairwise distance [5].

However, these evaluation functions are less likely to be relevant to a domain expert [20]. A soccer coach might be most interested in a cluster with phases that frequently lead to shots and goals. An opponent might be most interested in the clusters with the most phases, in order to identify and anticipate the most frequent patterns of play. Finally, a journalist might be interested in the clusters with the longest phases, as those can be the most interesting for sports fans.

For this paper, we went with the viewpoint of a soccer coach and rank clusters based on the number of shots that they contain.

4.4 Mining patterns

The fourth step involves identifying frequent sequential patterns, that is, time-ordered sequences of events, within each cluster. Typically, sequential pattern miners take as input sequences, where each element in the sequence is an itemset (i.e., unordered set).

An event contains a lot of information, and the key representational challenge is how to convert an event into an itemset. Deciding on an itemset representation requires considering two key questions:

Q1: What information to consider? For example, an end-user may care about knowing which players often play together, in which case the players involved in each event are important. However, teams rotate players between games, players may change positions during a game, and players are substituted within a game. Thus, some users may be interested in abstracting away from the specific players involved when considering a team's tactics. In this case, omitting the players' names from the representation is desirable.

Q2: How to encode the information? Each piece of information contained in an event could be represented in a multitude of different ways. For example, a player's position and the type of pass could be encoded hierarchically. The location could be represented as an exact position or as occurring in a specified zone of the pitch, and furthermore the zones could be organized hierarchically.

When thinking about how to encode the various components of an event, we pay special attention to two aspects of an event: its location and its type. Both aspects require some engineering to obtain good results.

Most pattern mining algorithms are designed to work with discrete data or to convert continuous attributes to discrete ones using a threshold (e.g., checking if the value is less than a threshold). The x and y coordinates of an event are real values, so we discretize the location. Rather than using a standard discretization method such as a grid [19], we divide the soccer pitch in zones based on domain knowledge as shown in Figure 3. Our discretization method leads to more informative patterns than a grid as the zones correspond to areas of the pitch frequently mentioned in soccer experts' discourse. Further discretizing the pitch in a more fine-grained manner had a negative effect on the quality and diversity of the found patterns (i.e., we found many spatially similar patterns with low support).

As mentioned in Section 3, shots and passes require some special care as they are important events and each one has multiple different types. For passes, we augment the itemset by adding any special type of pass as an extra event that is happening simultaneously. We treat shots in an analogous manner. Effectively, this introduces an extensional hierarchy in the data where an itemset can match on either the more generic event (e.g., a pass) or a more specific event (e.g., a through ball).

We want our patterns to be as readable as possible to a domain-expert, therefore we encode them in natural language format. An event of type A at location X is encoded as `[A] AT [X]`, while an event of type B that moves the ball from location $Y$to location Z is encoded as `[B] FROM [Y] TO [Z]`. Table 2 shows the sequence that corresponds to the phase in Figure 1.

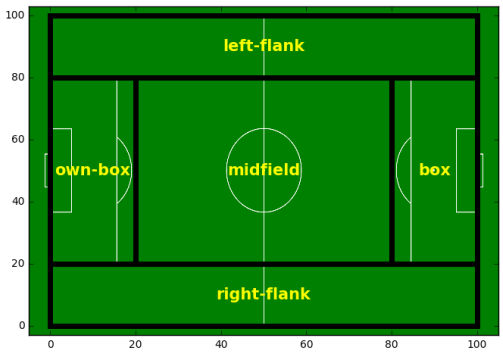

Figure 3: The zones used to discretize (x, y)-locations on the pitch.

Table 2: Sequence representation for the phase in Figure 1

1. Interception AT the right flank
2. Ball recovery AT the right flank
3. A pass FROM the right flank TO the right flank
4. A pass FROM the right flank TO the midfield
5. A pass FROM the midfield TO the midfield
6. A pass OR long ball FROM the midfield TO the box
7. A pass FROM the box TO the box
8. Attempt saved AT the box
9. Shot AT box

Using the discussed processing of locations and events, we consider the following five ways to represent an event as an itemset:

Location. This representation only considers the location of the event as determined by the zone in which it occurred.

Event Type. This representation only considers the type of the event, with the extra processing for passes and shots.

Player. This representation only considers players. Conceptually, each player is represented by a binary variable which takes on a value of 1 if the player participates in the event and 0 otherwise.

Location and Event Type. This representation constructs an itemset by combining (i.e., concatenating) the Location and Event Type representations.

Location, Event Type, and Players. This representation constructs an itemset by combining (i.e., concatenating) the Location, Event Type and Player representations.

Our primary evaluation will focus on the fourth representation and considers Locations and Event Types. The evaluation will explore how each of the representations affects the found patterns.

An alternative approach to mine patterns is inductive logic programming, which allows us to search for a richer set of patterns [18]. However, this approach currently does not scale well to a large volume of data and is thus ill-suited for the event data analyzed in this paper.

4.5 Ranking patterns

Finally, we rank the discovered frequent sequential patterns with respect to their expected relevance to a user. Typically, frequent

patterns are ranked according to their support in the data. For example, Van Haaren et al. [18] ranked patterns according to their m-estimate, which is a smoothed version of precision. However, this evaluation function is less relevant to soccer coaches. Given that most of the action during a soccer match typically happens in the middle of the pitch and that 90% of all events are passes, the top of the ranking is likely to be dominated by patterns describing passing sequences on the midfield.

We propose an alternative evaluation function that considers the types of the events appearing in a pattern, the length of the pattern and the pattern's support to determine its relevance. More specifically, we first assign a weight to each event type. Higher weights indicate higher relevance. This approach allows the user to define a bias towards a particular type of patterns. Specifically, we use a ranking function of the form:

$$Score(FS) = Supp(FS) \times \sum_{et \in ET} \lambda_{et} \times \#et \in FS$$

where FS is a frequent sequence, $Supp(SP)$ is the support count of the sequential pattern FS, λ_{et} is the weight assigned to event type et, and $\#et \in FS$ is the number of occurrences of event type et in SP. Given that we are mostly interested in goal attempts, we assign a high weight ($\lambda_{shot} = 2$) to shots, a low weight to normal passes ($\lambda_{pass} = 0.5$), and average weights ($\forall et \in ET \setminus \{shot, pass\} : \lambda_{et} = 1$) to all other types of events in our experiments.

5 EXPERIMENTAL STUDY

Our empirical evaluation on the dataset presented in Section 3 addresses the following four research questions:

- **Q1:** Do we discover interesting and relevant patterns?
- **Q2:** Can we characterize the tactics of teams?
- **Q3:** What is the effect of the clustering step?
- **Q4:** What is the best representation for phases?

The first two questions focus on evaluating the quality of our results, whereas the last two questions focus on assessing the impact of our design decisions on the overall results. Next, we discuss the methodology and present the results.

5.1 Methodology

The analysis is performed on a team-by-team basis. That is, all 38 league matches for a given team are used as input to the algorithm. The discussion will focus on the found patterns for Manchester City, Arsenal, and Leicester City. These patterns were evaluated qualitatively by a domain expert, as we do not have access to any ground truth data we can compare our patterns against. Tactics are often kept confidential by soccer clubs, so getting this ground truth data is nearly impossible.

We focus our attention on the top 10 clusters as ranked by the number of shots the cluster contains. We employ the CM-SPADE algorithm in the SPMF toolbox [8] to discover frequent maximal sequential patterns in each cluster. We used a support threshold of 10, and then rank the found patterns according to our score metric. As a default, we consider 100 clusters and use the Location and Event Type itemset representation. For **Q3**, we consider 1, 10, 100, and 500 clusters. For **Q4**, we consider all five ways to convert an event into an itemset discussed in Subsection 4.4.

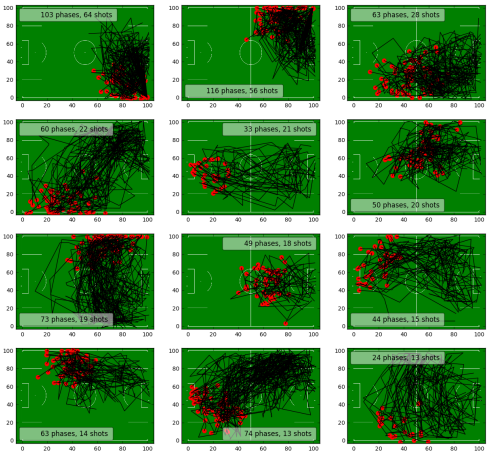

Figure 4: All phases assigned to each of the top 12 clusters for Manchester City. The red dots indicate where the phase begins. Manchester City is attacking to the right. The top-ranked cluster is in the upper left hand corner, and the clusters are ordered from left to right.

All experiments were run on a desktop with an Intel i7-6700 3.40GHz processor with four cores, each having two CPUs. The machine had 32 GB of memory.

5.2 Q1: Do we discover interesting and relevant patterns?

Figure 4 shows the top 12 ranked clusters for Manchester City. The phases that appear within the same cluster exhibit a reasonable degree of spatial coherence. There are identifiable commonalities, such as that the top-right and top-middle clusters contain phases beginning in the opposition's right and left flank. Figure 5 shows a zoomed version of the top-ranked cluster. In this cluster, several patterns were found that show a clear attacking pattern starting from the right flank. This involve actions such as passes followed by a cross, attacks from a corner, and set pieces. Similar patterns were found in the second-ranked cluster.

Figure 6 shows the fourth-ranked cluster, which is also interesting. The highest-ranked pattern in this cluster involved a ball recovery on the right flank, followed by a pass to the midfield, followed by a pass to the left flank. As seen from the cluster, this pattern is capturing a diagonal movement of the ball from the right side of Manchester city's own half to left side of their opponent's half. In the 2015-2016 season, Mauricio Pellegrini commonly employed a formation that aligned Kevin De Bruyne on the right, David Silva in a central role, and Raheem Sterling on the left in support of striker Sergio Kun Aguero. De Bruyne recovers many balls, especially for someone in that role. Sterling is very fast, and hence offers an outlet on the left side for a possible attack.

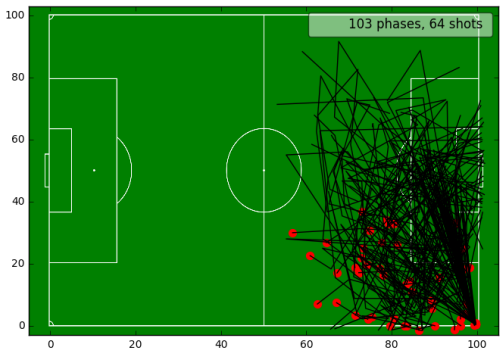

Figure 5: All phases assigned to the top-ranked cluster for Manchester City. The red dots indicate where the phase begins. Manchester City is attacking to the right. This shows a clear attacking pattern starting from the right flank.

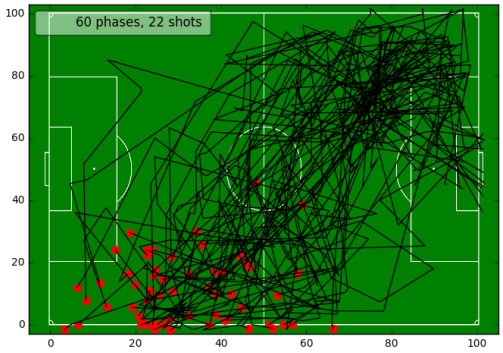

Figure 6: All phases assigned to the fourth ranked cluster for Manchester City. The red dots indicate where the phase begins. Manchester City is attacking to the right.

5.3 Q2: Can we identify team tactics?

To address this question, we compare the patterns found for three different teams: Arsenal, Leicester City, and Manchester City. We discuss this from both a quantitative and a qualitative perspective.

From a more quantitative perspective, Arsenal had 3,884 phases in the season containing three or more events and 480 shots occurred in these phases. Leicester City had 4,099 phases in the season containing three or more events and 439 shots occurred in these phases. Manchester City had 3,828 phases in the season containing three or more events and 512 shots occurred in these phases. The number of phases for Arsenal and Manchester City are very similar whereas Leicester has slightly more. One possible explanation could be that Leicester City matches typically involved a lot of duelling in midfield, which can lead to more possession changes. Additionally, Leicester generated around 10% fewer shots than Arsenal, and Arsenal generated about 6.5% fewer shots than Manchester City.

Table 6 gives the number of phases (P), number of shots (S), and the number of frequent sequential patterns (FS) contained within

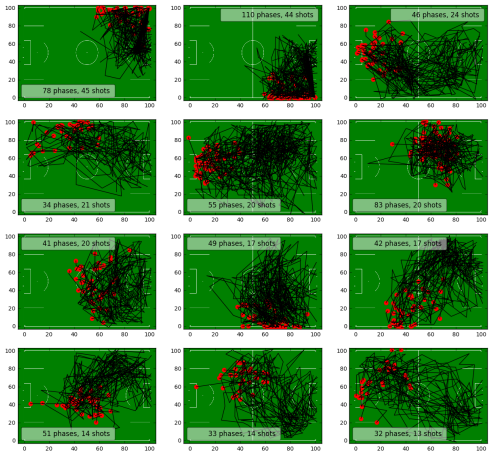

Figure 7: All phases assigned to each of the top 12 clusters for Arsenal. The red dots indicate where the phase begins. Arsenal is attacking to the right.The top-ranked cluster is in the upper left hand corner, and the clusters are ordered from left to right.

each of the top 10 ranked clusters for each team. 59.2% of Leicester City's shots occur in the top 10 clusters, with 16.5% appearing in the top cluster. In contrast, 54.1% of Manchester City's shots occur in the top 10 clusters with 12.5% in the top cluster. For Arsenal, only 50.4% of the shots occur in the top 10 clusters and 9.4% in the top cluster.

For Arsenal, Figure 7 shows the phases assigned to each of the top 12 clusters. Arsenal's play exhibits spatial diversity in how attacks are generated. Like most teams, the top two clusters capture attacks from either flanks (e.g., from corners or crosses). However, the clusters ranked 3 through 5 all capture various phases that start near Arsenal's own goal line, with a large number originating from goal kicks. These account for 65 shots. Their play involves long sequences of passing the ball around, with lots of action through the midfield. Table 3 shows the top-ranked frequent sequences within three clusters for Arsenal.

For Leicester City, Figure 8 shows the phases assigned to each of the top 12 clusters. Leicester City's top two clusters capture attacks from the right or left flank. However, Leicester generates many more shots (71) from the left flank than the right flank (46) whereas other teams are more balanced. The fact that Leicester was more prolific from the left is a bit surprising, as Riyad Mahrez, who won one of the player of the year awards and had a large number of goals (14 from open play) and assists (11), operated on the right. Unlike Arsenal, Leicester City has very few sequences in the top 10 clusters that start with a goal kick. Additionally, Leicester has four clusters where most phases start in the opponent's half of the midfield, and these generate 64 shots. This indicates a direct, counter-attacking style with shorter sequences. Table 3 shows the top-ranked frequent sequences within three clusters for Leicester City.

Table 3: The top-ranked frequent sequences found in the second, third, and ninth ranked clusters for Arsenal.

Cluster	Sequential Pattern
2^{nd} Cluster	1. A pass OR cross FROM the left flank TO the box 2. Shot
3^{rd} Cluster	1. A pass FROM the midfield TO the midfield 2. A pass FROM the midfield TO the midfield 3. A pass FROM the midfield TO the midfield
9^{th} Cluster	1. A pass FROM the midfield TO the midfield 2. A pass FROM the midfield TO the left-flank 3. A pass FROM the left flank TO the midfield 4. A pass FROM the midfield TO the midfield 5. A pass FROM the midfield TO the midfield

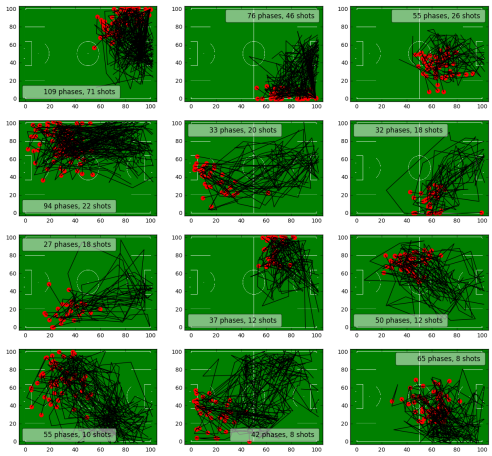

Figure 8: All phases assigned to each of the top 12 clusters for Leicester City. The red dots indicate where the phase begins. Leicester City is attacking to the right. The top-ranked cluster is in the upper left hand corner, and the clusters are ordered from left to right.

Table 4: The top-ranked frequent sequences found in the first, second, and seventh ranked clusters for Leicester City.

Cluster	Sequential Pattern
1^{st} Cluster	1. A pass OR cross FROM the left flank TO the box 2. A shot
2^{nd} Cluster	1. A pass OR cross FROM the right flank TO the box 2. A shot and a Miss 3. Ball goes out of bounds
7^{th} Cluster	1. A ball recovery IN the midfield 2. A shot

Manchester City generates a lot of shots from phases that start on the left or right flank, and the distribution is nearly even with 64 coming from the right and 56 from the left. More generally, Manchester City's style falls somewhere in between Arsenal and Leicester City. On the one hand, there are several clusters with phases starting near midfield that are short and direct. On the other hand, like Arsenal, there are some clusters that show groups of phases starting in Manchester City's half of the field between the penalty box and midfield. However, there are fewer phases initiated with a goal kick. Table 5 shows the top-ranked frequent sequences within three clusters for Manchester City.

Table 5: The top-ranked frequent sequences found in the first, third, and seventh ranked clusters for Manchester City.

Cluster	Sequential Pattern
1^{st} Cluster	1. A pass OR cross OR corner FROM right flank TO box 2. A shot
3^{rd} Cluster	1. A pass FROM left flank TO left flank 2. A pass FROM left flank TO midfield 3. A pass FROM midfield TO right flank
7^{th} Cluster	1. A ball recovery IN the left flank 2. A pass FROM the left flank TO the midfield

Table 6: A comparison of the clusterings found for Arsenal, Leicester City and Manchester City. The clusters are sorted by the number of shots each one contains and focuses on the 10 clusters that contain the most shots. Within each cluster, the number of phases (P), number of shots (S), and the number of frequent sequential patterns (FS) are shown.

Cluster Number	Arsenal			Leicester City			Manchester City		
	P	S	FS	P	S	FS	P	S	FS
1	78	45	143	109	71	227	103	64	141
2	110	44	159	76	46	134	116	56	127
3	46	24	47	55	26	11	63	28	82
4	34	21	20	94	22	34	60	22	72
5	55	20	187	33	20	17	33	21	19
6	83	20	26	32	18	7	50	20	27
7	41	20	40	27	18	13	73	19	165
8	49	17	150	37	12	12	49	18	10
9	42	17	116	50	12	16	44	15	12
10	51	14	40	55	10	41	63	14	18

5.4 Q3: What is the effect of the clustering step?

In this question, we compare 1 cluster (i.e., no clustering), versus 10, 100, and 500 clusters. We focus the analysis on Manchester City. Table 7 provides statistics on the number of phases (P), number of shots (S), and the number of frequent sequential patterns (FS) contained within each of the top 10 ranked clusters when considering 10, 100, and 500 clusters. From a quantitative standpoint, considering only 10 clusters resulted in 225,118 frequent sequences which is substantially more than for 100 clusters (4,557) or 500 clusters

(676). When looking at 100 clusters, 54.1% of all shots appear in the top 10 clusters, and for 500 clusters this number drops to 30.5%.

When performing no clustering, essentially all found patterns involve passing patterns of differing length within the midfield, with an occasional pass to one of the flanks. There are patterns involving shots or the box in the top 100 ranked frequent sequences.

When clustering into 10 clusters, seven of the clusters generate patterns that contain almost only passes within midfield. There is one cluster for attacks from the right flank and one from the left. Finally, one cluster contains phases starting near Manchester City's own box. Hence there is little diversity in the found patterns.

Figure 9 shows the top 12 ranked clusters when the phases are clustered into 500 clusters. The clusters are typically very spatially similar, but contain very few phases. This makes it difficult to find interesting patterns that have enough support in the data. Consequently, looking at 100 clusters seems to be a good tradeoff between diversity, spatial coherence, and sufficient data to find patterns with the desired support.

Table 7: The effect of the number of clusters with three cluster sizes considered: 10, 100, and 500. The clusters are sorted by the number of shots each one contains and focuses on the 10 clusters that contain the most shots. Within each cluster, the number of phases (P), number of shots (S), and the number of frequent sequential patterns (FS) are shown.

Cluster	10 Clusters			100 Clusters			500 Clusters		
Number	P	S	FS	P	S	FS	P	S	FS
1	424	108	488	103	64	141	57	43	86
2	353	89	314	116	56	127	45	27	88
3	344	85	77,699	63	28	82	25	16	8
4	539	56	6,812	60	22	72	20	13	12
5	144	47	136,212	33	21	19	25	13	3
6	305	46	1,702	50	20	27	17	10	3
7	199	46	908	73	19	165	15	9	3
8	355	35	353	49	18	10	17	9	0
9	763	0	520	44	15	12	14	8	1
10	402	0	110	63	14	18	13	8	0

5.5 Q4: What is the best phase representation for mining patterns?

Next, we consider the effect of the five different ways to convert an event into an itemset representation. We focus on the top 10 ranked patterns within each of the top 10 clusters. The two approaches (location and event type; player, location, and event type) that include multiple pieces of information result in very similar patterns. In both cases, around two thirds of the patterns are length one or two and one third are length three or greater. A drawback to including the player information is that it greatly expands the search space of possible patterns that need to be considered.

Only considering one aspect of an event in the itemset representation is quite limiting. Particularly, for just the location or just the event type, the patterns contain little context about what is happening. For just location or just event type, it is possible to find quite long sequences. For locations, it is common to find sequences of length five, with the longest being of length eight. For event type,

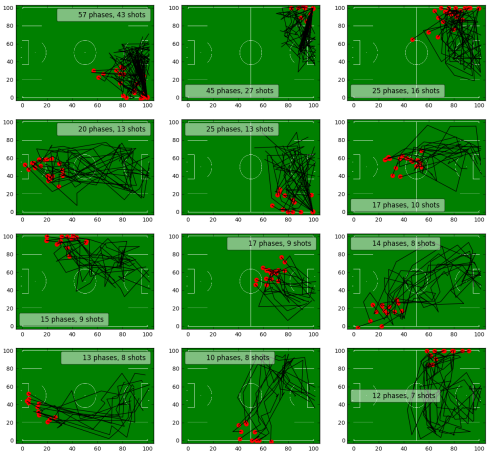

Figure 9: All phases assigned to each of the top 12 clusters from Manchester City when considering 500 clusters. The red dots indicate where the phase begins. Manchester City is attacking to the right.

sequences of up to length 13 are found. When considering just players, most patterns are very short, most are length one or two, and nothing longer than length three. A common finding is a pattern where the same player is involved in multiple consecutive events. These most likely indicate a dribbling sequence followed by a pass. But again, more context is needed to get a good understanding.

5.6 Impact on soccer industry

The patterns generated by our approach were presented to a company that provides data-driven advice to soccer clubs and soccer associations with respect to player recruitment and opponent analysis. The company has expressed interest in building a product based on this approach and implementing it in the near future to be included in their services.

6 CONCLUSIONS

Advanced data collection techniques are becoming more and more commonplace in sports and they generate rich, complex spatio-temporal data. Automatically analyzing team tactics from these data is an interesting and challenging problem. This paper tackled one aspect of this task by trying to automatically discover interesting attacking strategies from event data collected from professional soccer matches. The paper proposed a five-step pipeline to analyze such data. An analysis of the 2015-2016 English Premier League season identified several differences in style of play between different teams. It also identified some relevant, reoccurring patterns of play.

There are several important directions for future work. One is to continue to tackle the representational issues associated with performing pattern mining in a mixed discrete and continuous space. In conjunction with this, the ability to generalize to nearly identical

commonly occurring sequences could also allow finding additional interesting patterns. Another direction is evaluating whether our results are consistent over time, i.e., if the tactics inferred for a given team after a set of matches, carry over to a subsequent set of matches. Finally, it would be interesting and much more informative to have full optical-tracking data for all players and the ball. However, tackling such a setting would require radically different techniques.

ACKNOWLEDGEMENTS

Tom Decroos is supported by the Research Foundation-Flanders (FWO-Vlaanderen). Jesse Davis is partially supported by the KU Leuven Research Fund (C14/17/070, C32/17/036), FWO-Vlaanderen (SBO-150033) and Interreg V A project NANO4Sports.

REFERENCES

[1] A. Bialkowski, P. Lucey, P. Carr, Yisong Yue, S. Sridharan, and I. Matthews. 2014. Identifying Team Style in Soccer Using Formations Learned from Spatiotemporal Tracking Data. In *Proceedings of the Workshop on Spatial and Spatio-Temporal Data Mining*. 9–14.

[2] Huiping Cao, Nikos Mamoulis, and D.W. Cheung. 2005. Mining Frequent Spatio-temporal Sequential Patterns. In *Proceedings of the 5th International Conference on Data Mining*. https://doi.org/10.1109/ICDM.2005.95

[3] Tom Decroos, Vladimir Dzyuba, Jan Van Haaren, and Jesse Davis. 2017. Predicting soccer highlights from spatio-temporal match event streams. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*. 1302–1308.

[4] Tom Decroos, Jan Van Haaren, Vladimir Dzyuba, and Jesse Davis. 2017. STARSS: A spatio-temporal action rating system for soccer. In *Machine Learning and Data Mining for Sports Analytics ECML/PKDD 2017 workshop*.

[5] Brian S. Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. 2011. *Cluster Analysis*. John Wiley and Sons, Ltd. https://doi.org/10.1002/9780470977811.index

[6] Brian S Everitt, Sabine Landau, Morven Leese, and Daniel Stahl. 2011. Hierarchical clustering. *Cluster Analysis, 5th Edition* (2011), 71–110.

[7] T Fernando, X Wei, C Fookes, S Sridharan, and P Lucey. 2015. Discovering Methods of Scoring in Soccer Using Tracking Data. In *Proceedings of the Workshop on Large-Scale Sports Analytics*.

[8] P. Fournier-Viger, A. Gomariz, T. Gueniche, A. Soltani, C. Wu., and V. S. Tseng. 2014. SPMF: A Java Open-Source Pattern Mining Library. *Journal of Machine Learning Research* 15 (2014), 3389–3393. http://www.philippe-fournier-viger.com/spmf

[9] Fosca Giannotti, Mirco Nanni, Fabio Pinelli, and Dino Pedreschi. 2007. Trajectory Pattern Mining. In *Proceedings of the 13th International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, 330–339. https://doi.org/10.1145/1281192.1281230

[10] Laszlo Gyarmati and Xavier Anguera. 2015. Automatic Extraction of the Passing Strategies of Soccer Teams. *arXiv:1508.02171* (2015).

[11] Laszlo Gyarmati, Haewoon Kwak, and Pablo Rodriguez. 2014. Searching for a Unique Style in Soccer. *arXiv:1409.0308* (2014).

[12] Konstantin Knauf and Ulf Brefeld. 2014. Spatio-Temporal Convolution Kernels for Clustering Trajectories. In *Proceedings of the Workshop on Large-Scale Sports Analytics*.

[13] Konstantin Knauf, Daniel Memmert, and Ulf Brefeld. 2016. Spatio-Temporal Convolution Kernels. *Machine Learning* 102, 2 (2016), 247–273. https://doi.org/10.1007/s10994-015-5520-1

[14] Patrick Lucey, Dean Oliver, Peter Carr, Joe Roth, and Iain Matthews. 2013. Assessing Team Strategy Using Spatiotemporal Data. In *Proceedings of the 19th International Conference on Knowledge Discovery and Data Mining*. 1366–1374.

[15] Meinard Müller. 2007. *Dynamic time warping*. Springer, Chapter 4, 69–84.

[16] David J Stracuzzi, Alan Fern, Kamal Ali, Robin Hess, Jervis Pinto, Nan Li, Tolga Konik, and Daniel G Shapiro. 2011. An Application of Transfer to American Football: From Observation of Raw Video to Control in a Simulated Environment. *AI Magazine* 32, 2 (2011), 107–125.

[17] Jan Van Haaren, Horesh Ben Shitrit, Jesse Davis, and Pascal Fua. 2016. Analyzing volleyball match data from the 2014 World Championships using machine learning techniques. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 627–634.

[18] Jan Van Haaren, Vladimir Dzyuba, Siebe Hannosset, and Jesse Davis. 2015. Automatically Discovering Offensive Patterns in Soccer Match Data. In *International Symposium on Intelligent Data Analysis (Lecture Notes in Computer Science)*, Elisa Fromont, Tijl De Bie, and Matthijs van Leeuwen (Eds.), Vol. 9385. Springer, 286–297. https://doi.org/10.1007/978-3-319-24465-5

[19] Jan Van Haaren, Siebe Hannosset, and Jesse Davis. 2016. Strategy discovery in professional soccer match data. In *Proceedings of the KDD-16 Workshop on Large-Scale Sports Analytics*. 1–4.

[20] Matthijs Van Leeuwen. 2014. Interactive data exploration using pattern mining. In *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*. Springer, 169–182.