Discerning Tactical Patterns for Professional Soccer Teams: an Enhanced Topic Model with Applications

Qing Wang[1]*, Hengshu Zhu[2], Wei Hu[2], Zhiyong Shen[2], Yuan Yao[1]
[1]School of Mathematical Science, Peking University, [2]Big Data Lab, Baidu Research
{wangqing09,zhuhengshu,huwei05,shenzhiyong}@baidu.com, yuany@math.pku.edu.cn

ABSTRACT

Analyzing team tactics plays an important role in the professional soccer industry. Recently, the progressing ability to track the mobility of ball and players makes it possible to accumulate extensive match logs, which open a venue for better tactical analysis. However, traditional methods for tactical analysis largely rely on the knowledge and manual labor of domain experts. To this end, in this paper we propose an unsupervised approach to automatically discerning the typical tactics, i.e., *tactical patterns*, of soccer teams through mining the historical match logs. To be specific, we first develop a novel model named Team Tactic Topic Model (T^3M) for learning the latent tactical patterns, which can model the locations and passing relations of players simultaneously. Furthermore, we demonstrate several potential applications enabled by the proposed T^3M, such as automatic tactical pattern discovery, pass segment annotation, and spatial analysis of player roles. Finally, we implement an intelligent demo system to empirically evaluate our approach based on the data collected from La Liga 2013-2014. Indeed, by visualizing the results obtained from T^3M, we can successfully observe many meaningful tactical patterns and interesting discoveries, such as *using which tactics a team is more likely to score a goal* and *how a team's playing tactic changes in sequential matches across a season.*

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications — Data Mining; H.4.m [**Information Systems Applications**]: Miscellaneous

Keywords

Tactical Patterns; Professional Soccer; Topic Model

*The paper was done when the first author was an intern in Baidu Research – Big Data Lab under the supervision of the fourth author.

KDD'15, August 10–13, 2015, Sydney, Australia.

DOI: http://dx.doi.org/10.1145/2783258.2788577.

1. INTRODUCTION

As the most popular sport in the world, soccer stimulates more than 3 billion loyal fans [2] and tens of billions of dollars in revenue each year [7]. A growing interest along this line is to provide professional tactical analysis for the soccer industry, which can facilitate the decision making of soccer coaches and improve the watching experiences of match audiences. Recently, with the rapid development of wearable devices and sensor technology, the mobility information of players/ball and match events can be recorded as extensive match logs. Indeed, these large-scale and fine-grained data open a venue for better professional tactical analysis, and thus attract a lot of research attentions [18, 26, 4, 24, 17]. However, traditional methods for tactical analysis largely rely on the knowledge and manual labor of domain experts, which are very expensive and unscalable for the entire industry. Therefore, it is very appealing to have an effective and automatic approach for learning the typical tactics, i.e., *tactical patterns*, of soccer teams. Specifically, here we define a tactical pattern as a series of frequently used ball-passing combinations that are characterized by the players who make the passing or receiving actions, and their corresponding locations.

In the literature, while there are some related studies on sports data mining, such as [18, 17, 26, 3, 4, 16, 24, 10, 9], the problem of mining tactical patterns for soccer teams is still under-explored. It is because that the tactical patterns are high-level features and only low-level data can be observed, such as the positions of players and the ball over time. To this end, in this paper we propose an unsupervised probabilistic approach to automatically discerning the tactical patterns of soccer teams through mining the historical match logs. To be specific, we first present a novel generative model named Team Tactic Topic Model (T^3M) for learning the latent tactical patterns, which can model the locations and passing relations of players simultaneously. Furthermore, several potential applications enabled by the proposed T^3M are demonstrated, such as automatic tactical pattern discovery, pass segment annotation, and spatial analysis of player roles. Finally, an intelligent demo system is implemented to empirically evaluate our approach based on the data collected from La Liga 2013-2014. Indeed, by visualizing the results obtained from T^3M, many meaningful tactical patterns and interesting discoveries can be observed, such as *using which tactics a team is more likely to score a goal* and *how a team's playing tactic changes in sequential matches across a season.* Particularly, the contributions of this paper can be summarized as follows.

Table 1: A sampled match log of FC Barcelona, where match segments are separated by horizontal lines.

Event ID	Timestamp	Passer ID	Receiver ID	Start position	End position
1574	02:27	[D]J.Alba	[D]J.Mathieu	(44.2,100.0)	(29.8,92.2)
1575	02:30	[D]J.Mathieu	[D]J.Mascherano	(25.5,81.7)	(17.8,42.9)
1576	02:33	[D]J.Mascherano	[D]D.Alves	(20.9,34.8)	(30.6,5.6)
⋮	⋮	⋮	⋮	⋮	⋮
1588	03:06	[M]I.Rakitic	[D]J.Mathieu	(52.4,14.4)	(45.2,56.7)
1589	03:10	[D]J.Mathieu	[M]A.Iniesta	(46.2,58.5)	(56.5,79.3)
1599	03:54	[G]C.Bravo	[M]S.Busquets	(4.8,40.6)	(21.9,49.8)
1600	04:01	[M]S.Busquets	[D]J.Mascherano	(33.6,49.6)	(36.7,20.7)
1601	04:06	[D]J.Mascherano	[M]I.Rakitic	(46.7,17.1)	(52.8,25.9)
1602	04:07	[M]I.Rakitic	[D]J.Mascherano	(52.8,25.9)	(45.1,17.8)
1611	04:43	[M]A.Iniesta	[D]J.Alba	(60.5,97.7)	(70.2,96.9)
1612	04:44	[D]J.Alba	[F]M.E.Haddadi	(70.2,96.9)	(80.4,82.8)
1613	04:45	[F]M.E.Haddadi	[F]L.Messi	(80.4,82.8)	(77.1,62.5)

- To the best of our knowledge, this work is the first attempt to leverage unsupervised modeling approach for automatically discerning the tactical patterns of soccer teams by mining match logs.
- We propose a novel generative model T^{3}M for learning the latent tactical patterns. Comparing to existing models, our model can learn the locations and passing relations of players simultaneously, yielding more meaningful and spatially localized tactical patterns.
- Further, we build an intelligent demo system to evaluate our approach. By visualizing the results obtained from T^{3}M, we show that our model can promote many related applications, including automatic tactical pattern discovery, pass segment annotation, and spatial analysis of player roles.

Overview. The remainder of this paper is organized as follows. Section 2 introduces the details of the proposed generative model T^{3}M. In Section 3, we further discuss the T^{3}M and introduce several potential applications. Then in Section 4, we report the evaluation results based on our intelligent system and a real-world data set. And Section 5 provides a brief review of related works and our innovations. Finally, we conclude the paper in Section 6.

2. MODELING THE TACTICAL PATTERNS FOR SOCCER TEAMS

In this paper, we are interested in the problem of automatically finding a team's tactical patterns in sequential matches across a season. Specifically, a team may have a number of tactical patterns in terms of attacking and defending, such as *how to break through the defense on the sideline*, and *how to keep possession of the ball*. Thus, in this section, we propose an unsupervised approach to modeling the tactical patterns for soccer teams based on the analysis of match logs.

2.1 Data Description

The match log of a specific team contains many event records, such as pass event, foul event, ball-out event and tackle event. In this paper, we define a tactical pattern as a series of frequently used ball-passing combinations that are characterized by the players who make the passing or receiving actions, and their corresponding locations. Therefore, here we mainly focus on the records related to the pass events in the match logs. We further split the pass events into several segments by stoppages and turnovers in matches [17]. A stoppage means the game is paused, usually due to a ball-out or a foul; a turnover means the opponent team takes the possession of the ball, without pausing the game. We label the corresponding events in match logs as stoppages and turnovers. Then the passes before and after a stoppage or a turnover are considered as in two disjoint segments. For example, Table 1 shows a sampled match log of FC Barcelona with three different match segments, where each row in the table represents a successful pass event. Specifically, each event record contains a unique *Event ID*, a *Timestamp*, a *Passer* with corresponding *Start Position* and a *Receiver* with corresponding *End Position* (i.e., each position is scaled to 100×100).

2.2 Team Tactic Topic Model (T^{3}M)

Here we introduce the details of the proposed T^{3}M for modeling tactical patterns. Specifically, let us denote a team as a set of players Q, of which the match log can be split into multiple segments D. For the i-th pass in segment d, $i = 1, \ldots, N_d$, we denote the passer as $u_{d,i}$ and the receiver as $v_{d,i}$. Meanwhile, the pass of the ball can be denoted as a directed edge from the passer to the receiver, i.e., $w_{d,i} = (u_{d,i}, v_{d,i})$. And the end position of a pass is denoted as $x_{d,i}$. The end position is used instead of the start position since the end position can better describe the receiver's movement without the ball. To write compactly, a token $f_{d,i}$ in our model is defined by $f_{d,i} = (w_{d,i}, x_{d,i})$.

We assume each team has several tactical patterns $k = 1, \ldots, K$, each consists of a passing pattern and a team positioning. Specifically, a passing pattern is a transition frequency matrix between players that is modeled as a multinomial distribution on the Cartesian product between players $(Q \times Q)$ (i.e., parameterized by ϕ_k); A team positioning indicates where the ball is most likely passed to the receiver v. In particular, we assign a Gaussian distribution parameterized by $\psi_{k,v} = (\mu_{k,v}, \Sigma_{k,v})$ to each receiver $v \in Q$ for each pattern k respectively, resulting in $M = K \times |Q|$ Gaussian distributions in total. Furthermore, for each pass $f_{d,i}$ in segment d, we assume there is a latent variable $z_{d,i} \in \{1, \ldots, K\}$, indicating which tactical pattern the team is currently playing in this pass. And the latent pattern $z_{d,i}$ in segment d is consid-

Table 2: Mathematical notations.

Symbol	Description
K	The predefined number of tactical patterns.
D	The pass segments of a team. $D = \{d_1, \ldots, d_{\lvert D\rvert}\}$.
Q	The squads (set of players) for a team.
ϕ_k	Passing pattern for pattern k.
$\psi_{k,v}$	Gaussian parameters of player v's location in pattern k. $\psi_{k,v} = (\mu_{k,v}, \Sigma_{k,v})$
θ_d	Distribution of tactical patterns for segment d.
$w_{d,i}$	The pass from $u_{d,i}$ to $v_{d,i}$. $w_{d,i} = (u_{d,i}, v_{d,i})$.
$x_{d,i}$	The end position of the pass. $x_{d,i} \in \mathbb{R}^2$.
$f_{d,i}$	The players and the end position of a pass. $f_{d,i} = (w_{d,i}, x_{d,i})$.
$z_{d,i}$	Latent tactical pattern of pass $f_{d,i}$.
α	Hyperparameter of the Dirichlet prior on θ_d.
β	Hyperparameter of the Dirichlet prior on ϕ_k.
γ	Hyperparameter of the NIW prior on $(\mu_{k,v}, \Sigma_{k,v})$. $\gamma = (\mu_0, \Sigma_0, \Lambda_0, \nu_0)$

ered as drawn from the multinomial distribution $\text{Mult}(\theta_d)$. As a result, the actual pass $f_{d,i}$ is regarded as drawn according to $z_{d,i}$. To be more specific, given $z_{d,i} = k$, each $w_{d,i}$ conforms a multinomial distribution parameterized by ϕ_k; and the end position $x_{d,i}$ follows a Gaussian distribution with parameters $(\mu_{k,v_{d,i}}, \Sigma_{k,v_{d,i}})$.

In this paper, we leverage the Bayesian approach for the modeling process, where conjugate priors are placed for respective distributions. Specifically, the generative process of our model is demonstrated as follows:

1. Draw K tactical patterns from its prior distribution.
   (a) Draw ϕ_k from the Dirichlet prior $\text{Dir}(\beta)$ for $k = 1, \ldots, K$.
   (b) Draw $\psi_{k,v}$ from $\text{NIW}(\gamma)$ for $k = 1, \ldots, K$ and $v \in Q$.
2. Draw θ_d from the Dirichlet prior $\text{Dir}(\alpha)$ for $d \in D$.
3. For each observation i in segment d:
   (a) Draw latent assignment $z_{d,i}$ from the segment's specified distribution $\text{Dir}(\theta_d)$.
   (b) Draw the pass $w_{d,i}$ from $\text{Dir}(\phi_{z_{d,i}})$.
   (c) Draw the player location $x_{d,i}$ from $N(\mu_{\phi_{z_{d,i}}, v_{d,i}}, \Sigma_{\phi_{z_{d,i}}, v_{d,i}})$.

In summary, we illustrate the important mathematical notations in Table 2, and the graphical representation of T^{3}M in Figure 1. The full model parameterizations are

$$\begin{aligned} w_{d,i}|z_{d,i} = k, \phi_k &\sim \text{Mult}(\phi_k) \\ x_{d,i}|z_{d,i} = k, v_{d,i} = v, \mu_{k,v}, \Sigma_{k,v} &\sim N(\mu_{k,v}, \Sigma_{k,v}) \\ z_{d,i}|\theta_d &\sim \text{Mult}(\theta_d) \\ \theta_d &\sim \text{Dir}(\alpha) \\ \phi_k &\sim \text{Dir}(\beta) \\ \psi_{k,v} &\sim \text{NIW}(\gamma), \end{aligned}$$

where the NIW denotes the Normal-Inverse-Wishart distribution [19].

2.3 Model Inference

To estimate the model parameters, here we exploit an iterative approach called collapsed Gibbs sampling (CGS) [11], which is a Monte Carlo Markov Chain (MCMC) algorithm to approximate the posterior distribution of model parameters. Specifically, in CGS procedure we first assign a random tactical pattern to each pass. Then in each iteration of CGS, a new tactical pattern for each pass is drawn from the posterior distribution conditioned on other assignments. Finally, under some regularity conditions, the distribution of assignments for each pass converges to the posterior distribution as the number of iteration increases.

If we denote the latent assignments for the passes in all segments except the i-th pass in segment d as $z_{-(d,i)}$, and the latent assignments for all the passes in segment d except the i-th pass as $z_{d,-i}$ (similar notations hold for x, w, f, etc.), the probability for assigning the tactical pattern k for the i-th pass in segment d conditioned on the other assignments can be computed as

$$\begin{aligned} & P(z_{d,i} = k | x_{d,i}, w_{d,i}, z_{-(d,i)}, x_{-(d,i)}, w_{-(d,i)}) \quad (1) \\ \propto\ & P(z_{d,i} = k | z_{-(d,i)}) \\ & \cdot P(w_{d,i} | z_{d,i} = k, z_{-(d,i)}, w_{-(d,i)}) \\ & \cdot P(x_{d,i} | z_{d,i} = k, z_{-(d,i)}, x_{-(d,i)}), \end{aligned}$$

where we have

$$P(z_{d,i} = k | z_{-(d,i)}) = \frac{n^{(k)}_{(d,-i),\cdot} + \alpha}{(N_d - 1) + \alpha K} \quad (2)$$

$$P(w_{d,i} | z_{d,i} = k, z_{-(d,i)}, w_{-(d,i)}) = \frac{n^{(k)}_{-(d,i),w_{d,i}} + \beta}{n^{(k)}_{-(d,i),\cdot} + \beta |Q|^2} \quad (3)$$

$$P(x_{d,i} | z_{d,i} = k, z_{-(d,i)}, x_{-(d,i)}) = P(x_{d,i} | X_{d,i}). \quad (4)$$

The notation $n^{(k)}_{a,b}$ means the frequency of f_a labeled as tactic k while $w_a = b$. And N_d represents the number of passes in segment d. To have a compact representation, we use the notation $X_{d,i} = \{x_{d',i'} : (d', i') \neq (d, i), z_{d',i'} = z_{d,i}, v_{d',i'} = v_{d,i}\}$. The posterior predictive distribution for $P(x_{d,i}|X_{d,i})$ is a multivariate Student-T distribution with parameters [19]:

$$\begin{aligned} P(x|X) &= t_{\nu_n - 1}(\mu_n, \tfrac{\Lambda_n(\kappa_n + 1)}{\kappa_n(\nu_n - 1)}) \quad (5) \\ \mu_n &= \tfrac{\kappa_0}{\kappa_0 + n}\mu_0 + \tfrac{n}{\kappa_0 + n}\bar{x} \\ \kappa_n &= \kappa_0 + n \\ \nu_n &= \nu_0 + n \\ \Lambda_n &= \Lambda_0 + \textstyle\sum_{j=1}^{n}(x_j - \bar{x})(x_j - \bar{x})^T \\ &\quad + \tfrac{\kappa_0 n}{\kappa_0 + n}(\bar{x} - \mu_0)(\bar{x} - \mu_0)^T \end{aligned}$$

where $n = |X_{d,i}|$ and $\bar{x}$ is the sample mean for $x_j \in X_{d,i}$.

3. APPLICATIONS & DISCUSSION

In this section, we firstly introduce how our T^{3}M can promote various applications in the sports analytic industry, including automatic tactical pattern discovery, pass segment annotation, and spatial analysis of player roles. And then we discuss the improvements of our T^{3}M over basic topic models (e.g., Latent Dirichlet Allocation (LDA) [6]).

3.1 Applications

With T^{3}M, we can study the tactical patterns discovered from a team's passing data, and understand what tactic the team is most likely to use, as well as who is the key player in each tactical pattern. The analysis results of tactical patterns can help soccer coaches to better plan a match.

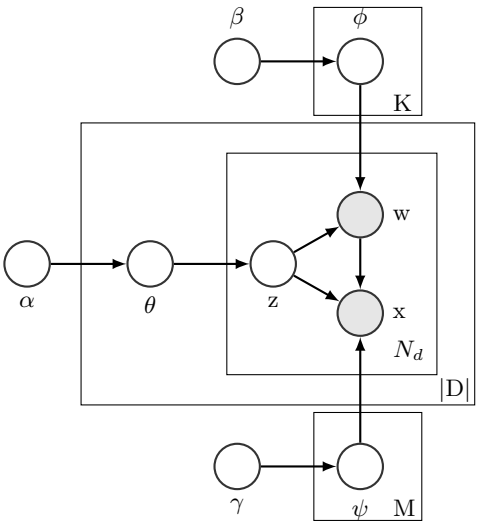

Figure 1: The graphical representation of T³M, where K is the number of tactical patterns, N_d is the number of passes in segment d, $|D|$ is the total number of segments, and $M = K \times |Q|$ is the total number of player position distributions.

3.1.1 Automatic Team Tactical Pattern Discovery

Team tactical pattern discovery is an important topic in soccer data analysis. Existing methods include using entropy maps of ball movement patterns to study the spatial variability of tactics (e.g., [17]), and motif finding method to characterize the passing patterns between players (e.g., [10]). In this work, our method can be viewed as a simultaneous analysis on the passing patterns between players and the spatial patterns in the pitch. For a given team, we can find K tactical patterns with our T³M. And for each pattern, we can know its passing patterns ϕ_k on the graph of players, and receiving locations for each player v. By the visualization of these learned tactics, we can study the playing styles of a team in a higher level. Therefore, we can understand the learned tactical patterns in a semantic way, and help coaches and sports analysts to give quantitative and objective analysis of a team's playing styles. With the help of our system, we can also answer the questions like *what is the most successful tactical pattern of a team (resulting in the most goals)* and *who is the central player (strategic center) of a playing style of this team.* Empirical evaluations of this application are demonstrated in Section 4.2.

3.1.2 Pass Segment Annotation

In addition to automatic team tactical pattern discovery, we can also study the progress of a match as dynamic changes of tactics. For each segment d in a match, we can find the posterior distribution (parameterized by θ_d) of tactical patterns for the passes in this segment. We can then annotate these disjoint segments in a match with different colors according to the dominating tactical pattern

$$k^* = \text{argmax}_k \theta_d. \tag{6}$$

Following the above approach, we can use our unsupervised method to cluster similar playing styles, and color the pass segments in a series of matches of a team, then analyze the dynamic change of a team's favorite tactical pattern across a season. Meanwhile, we can then study the team's tactical changes, as well as the strategic interaction between the team and different opponent teams. Furthermore, the annotation of pass segments also provides a natural summarization and abstraction of match playing progress. Empirical evaluations of this application can be found in Section 4.3.

3.1.3 Spatial Analysis of Player Role

Furthermore, our model is useful for individual analysis as well. Specifically, a joint analysis of team's tactical patterns and the roles a player plays in different tactics can help us define the participation measure of players in different tactics, and players' spatial roles at different regions of the playing field.

Given a player v, we can find in which tactical pattern player v plays a key role. Formally, for a passing pattern $\phi_k \in Q \times Q$ of pattern k, a player v's participation measure is defined as

$$\text{score}(k, v) = \phi_{k,v,\cdot} + \phi_{k,\cdot,v}, \tag{7}$$

which is equal to the sum of the weights for the passes in ϕ_k from or to player v. A visualization of the player scores is given in Figure 6 as an example. Furthermore, denoting the scores for player v as

$$S_v = (\text{score}(1, v), \ldots, \text{score}(K, v)), \tag{8}$$

we can then perform clustering algorithms on S_v to study the different player groups with respect to their playing roles.

Finally, we can also study a player's active regions in different tactical patterns. A player may appear in different locations with different roles to play. To visualize the diversity, we can collect the locations where player v receives the ball in different tactics respectively, and plot the heat map for each pattern as

$$P(x_{d,i}|v_{d,i} = v) = \sum_{k=1}^{K} P(x_{d,i}, z_{d,i} = k|v_{d,i} = v). \tag{9}$$

The distribution of $x_{d,i}$ where $v_{d,i} = v, z_{d,i} = k$, can be seen as the active region for player v in pattern k. Experiment results of this application are given in Section 4.4.

3.2 Discussion: LDA vs. T³M

Actually, T³M can be seen as an extension of classic topic model LDA. Therefore, here we discuss the differences between T³M and LDA in terms of modeling tactical patterns.

Specifically, for modeling passes between players, we note that both LDA and T³M can capture the prominent passing styles in the our data set. However, without using the location information (e.g., fitting with classical LDA), the learned patterns are not spatially clustered. As in Figure 2, we depict the learned tactical patterns by LDA (i.e., top two rows) and T³M (i.e., bottom two rows) respectively. In this figure, heat maps of receiving positions of the passes that assigned to each learned tactical pattern are drawn to represent the spatial distribution of these patterns. From these figures, we can observe that the location information incorporated in our T³M can help to assess the latent playing styles with more compact spatial patterns in the pitch.

In addition, T³M aims to connect the player locations with latent tactical patterns, so that we can find the different roles or locations of a player in different tactics. For example, player A can play the role of attacking midfielder in both

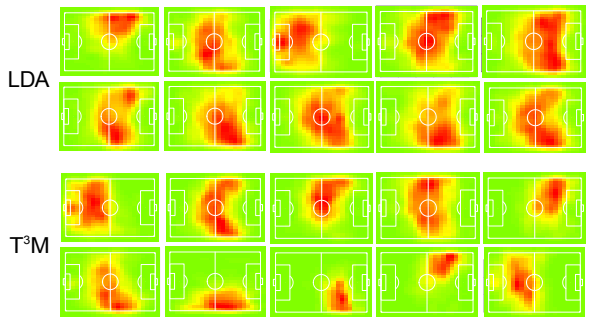

Figure 2: Heat maps for the 10 tactical patterns of Barcelona learned by LDA and T³M.

left side and right side, and player B is a strategic midfielder who usually passes the ball to player A. In classical LDA, the location information of the players is invisible to the model so that the passes from B to A are all categorized as the same tactical pattern. However, with the proposed T³M, the location where player A received the ball is related to the latent tactics he was playing, so it is likely that the passes from B to A will be classified into two distinct patterns, namely one is to attack in the left side and the other in the right side. In this manner, we can give a spatial analysis of each player's roles as described in Section 3.1.3.

4. EXPERIMANTAL RESULTS

In this section, we empirically study several novel applications enabled by our model, including soccer team tactical patterns discovery, team tactics analysis over matches, player role analysis, and a specific case study on Barcelona vs. Real Madrid.

In particular, we have developed a web-based demo system[1] for evaluating different applications. This system can visualize the results obtained from T³M, provide in-depth analysis and objective quantification of team performance, and enhance better watching experience for soccer fans. Figure 3 shows a screenshot of our prototype system for automatic soccer tactic analysis. In the middle of the page, we plot the spatial positions of major players in each tactical pattern, with various statistics below the plot. On the right side, we plot the heatmaps of receiving positions for each tactical pattern.

4.1 Evaluation Setup

For the evaluation of our approach and system, we used match logs from a commercial sports data provider [1], which contains 241 (out of 380) matches in La Liga 2013-2014, resulting in 401,105 events in the match logs. The match logs were preprocessed according to the descriptions in Section 2. To facilitate understanding, we mainly use FC Barcelona, a leading team in La Liga, as an example for evaluating our approach in the following subsections. Specifically, we extracted 24 matches of Barcelona in this season, and split the 13,367 passes into 2,652 playing segments according to the description in Section 2.

In our experiments, the number of CGS iterations was chosen as 10,000 to ensure convergence. The number of tactical patterns was empirically chosen as 10 for evaluation. Note that, although a team may arguably have much more tactics

[1]This demo system will be publicly available soon.

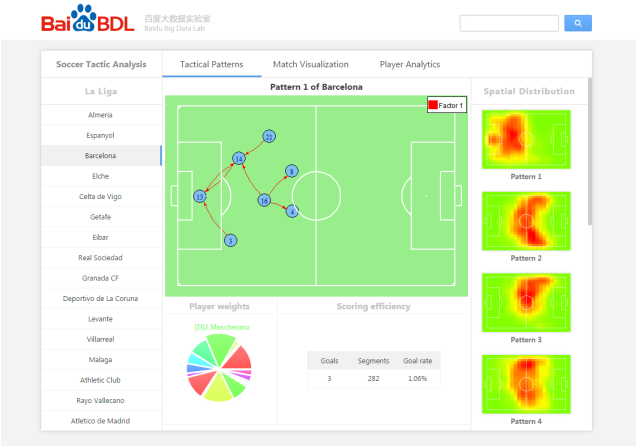

Figure 3: A screenshot of our system for automatic soccer tactic analysis.

than 10, here we only aim to learn the typical tactical patterns. In addition, we set $\alpha = 0.01$ and $\beta = 0.001$ for both LDA and T³M [11]. For NIW prior, we set $\mu_0 = (50, 50)^T$, $\kappa_0 = 2$, $\Lambda_0 = \begin{pmatrix} 100 & 50 \\ 50 & 100 \end{pmatrix}$, $\nu_0 = 2$. Indeed, empirically we find that the results are not sensitive to the choice of NIW prior when we set small values for both κ_0 and ν_0.

4.2 Visualization of Tactics for Barcelona

As one of the most successful soccer clubs in the world, FC Barcelona is well known for its world-class superb players, top-tier records and achievements in various competitions, as well as the famous *Tiki-taka* tactical style. Specifically, Tiki-taka tactical style is characterized by maintaining ball-possession with short passes and movements, which is considered as a diversion from the traditional thinking of formations to a concept derived from zonal play [25]. As an example, here we propose to study the tactical patterns of Barcelona with our approach and system.

First, we fit T³M as described in Section 2 with corresponding match logs. For each pass $f_{d,i}$ in these segments, our model can compute a latent tactical pattern $z_{d,i}$ assigned to the pass. Then, we can collect the passes with the same latent pattern k, and plot the corresponding end positions as a heat map visualized in Figure 2.

Indeed, the learned tactical patterns are meaningful and interpretable. Combining Figure 2 and Figure 4, we can have a simultaneous analysis of passing patterns and spatial patterns. For example, Pattern 1 can be seen as a defensive pattern, with frequent passes between the defenders (Mascherano, Adriano, Pique) and the goalkeeper (Pinto), while Pattern 2 indicates an attacking pattern in the two wings. Specifically, Pattern 5 and Pattern 9 are offensive styles in the left wing, with Iniesta as the playmaker, while Pattern 8 is an offensive style in the right wing, centered on the midfielder Xavi. It should also be noted that, for most patterns there is a central player in the passing network, usually acting as a strategic center or a playmaker.

Further, statistics of the scoring efficiency for the tactical patterns are shown in Table 3. For each goal in the matches, we find the last segment of passes before the goal, and attribute the goal to the most frequent tactical pattern

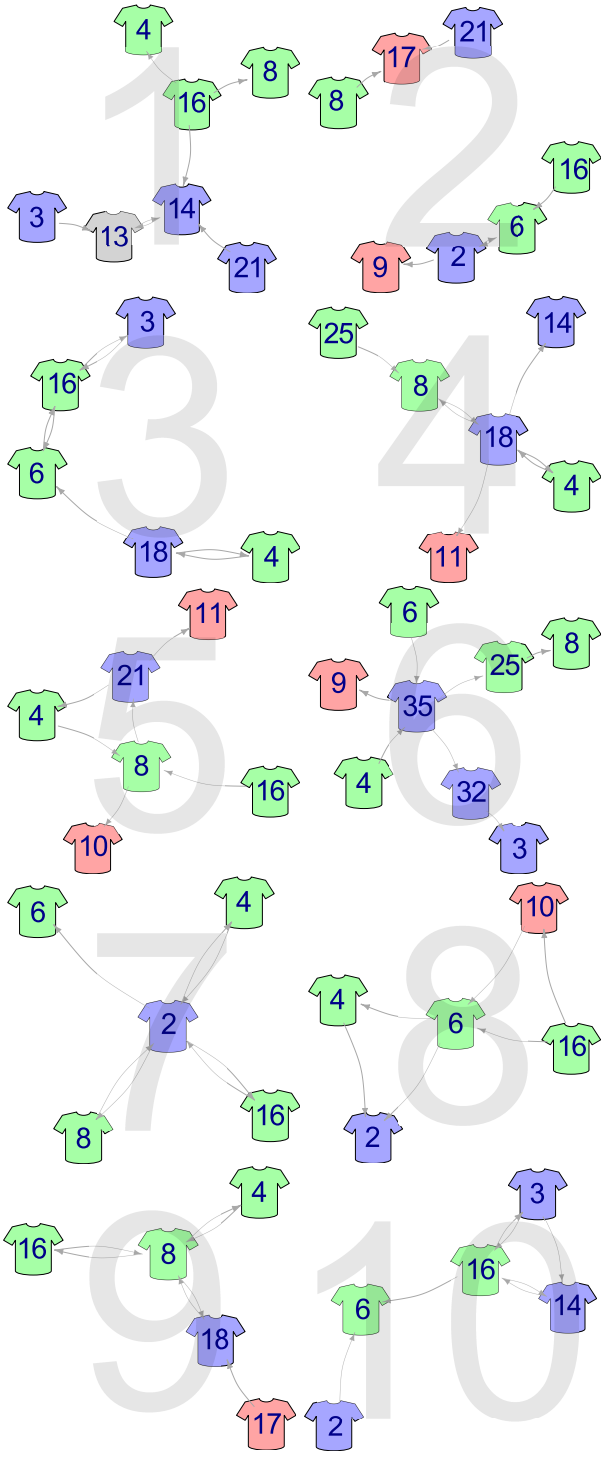

Figure 4: The 10 learned passing patterns for players of Barcelona. Goalkeepers, defenders, midfielders, and forwards are colored in grey, blue, green and red respectively. Players are represented by their jersey numbers (13:[G]Pinto, 35:[D]M.Montoya, 14:[D]J.Mascherano, 21:[D]Adriano, 3:[D]G.Pique, 32:[D]M.Bartra, 18:[D]J.Alba, 2:[D]D.Alves, 4:[M]F.Fabregas, 6:[M]Xavi, 16:[M]S.Busquets, 8:[M]A.Iniesta, 25:[M]A.Song, 11:[F]Neymar, 9:[F]A.Sanchez, 17:[F]P.Rodriguez, 10:[F]L.Messi).

Table 3: Goals statistics for different tactical patterns of Barcelona.

Pattern	# Goals	# Segments	Goal rate
	3	282	1.06%
	4	327	1.22%
	9	280	3.21%
	1	208	0.48%
	13	325	4.00%
	8	280	2.86%
	3	201	1.49%
	6	280	2.14%
	7	189	3.70%
	2	280	0.71%

of the segment. Comparing to existing methods for match statistics, which only attribute shots and goals to teams and players, our model can directly show the scoring statistic for each tactical pattern, which is more meaningful and useful for subsequent analysis.

From the above statistics we can see that some of the tactical patterns are more likely to score a goal. For example, Pattern 5 is an attacking pattern in the left wing, with Adriano and Iniesta as playmakers, passing balls to Neymar, Fabregas, Busquets and Messi. In a total of 325 segments with this pattern as the dominating tactical pattern, there are 13 segments finished with a goal, resulting in a 4.00% scoring rate. The pattern with the second highest scoring efficiency is Pattern 9, which is also an attacking pattern in the left, with Iniesta as the strategic center and ball passes with Fabregas, Busquets, Alba and Pedro. This pattern is also an effective attacking pattern, with 7 of 189 segments ended with goals.

4.3 Visualization of Tactics over Matches

A team's major tactical pattern often changes over time for several reasons, e.g., the coach's experiments with new strategies, or the transfers of core players. Indeed, when deciding which tactic to play before a match, coaches and club managers may also consider the opponent team's major tactics and player conditions of his own team. Therefore, the actual playing tactics usually vary over matches.

Here we also give an analysis of Barcelona's playing patterns for evaluation. Based on the match logs of Barcelona, we plotted the tactical patterns in the first half of each match in Figure 5. The matches are arranged from top to bottom according to the natural order of match dates. Passing segments are colored according to the dominating tactical patterns, as described in Section 3.1.2. It can be seen that at the beginning, Barcelona's favorite tactical patterns are **Pattern 6** (vs. Villarreal on 20131215, vs. Elche on 20140105, vs. Sevilla on 20140210, vs. Real Sociedad on 20140223) and **Pattern 3** (vs. Levante on 20140120, vs. Malaga on 20140127,

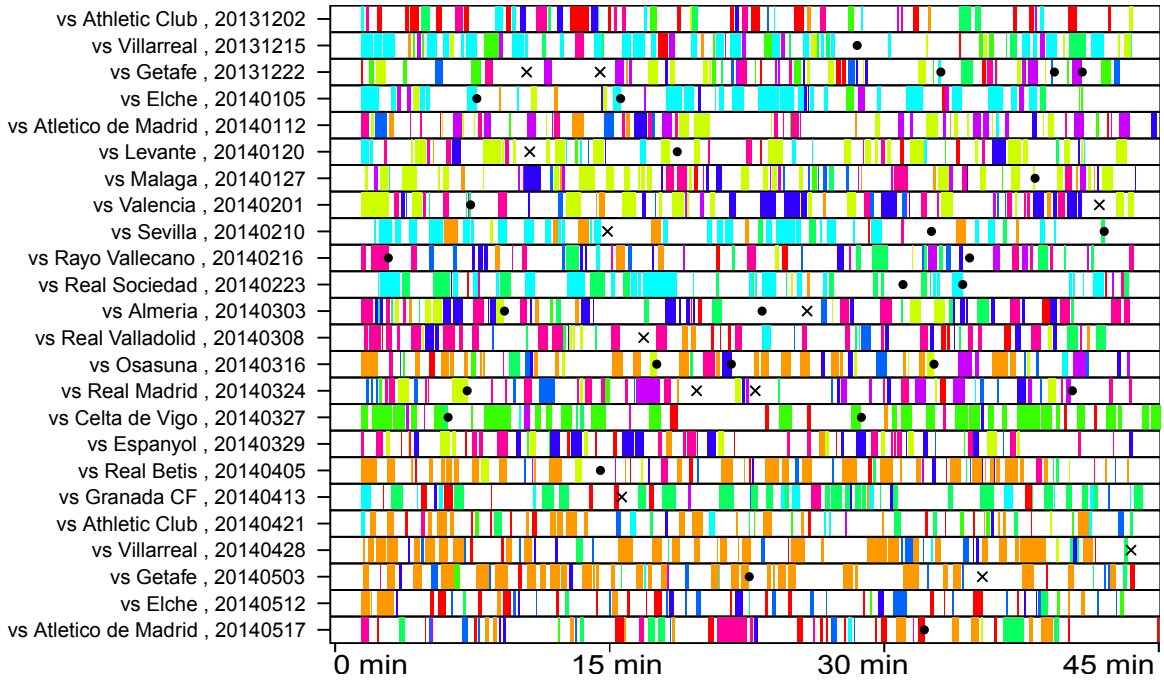

Figure 5: Visualization of tactical patterns over multiple matches, with different colors representing different patterns from 1 to 10 (colored as 1 2 3 4 5 6 7 8 9 10), solid dots representing goals scored, and crossings representing goals conceded.

vs. Valencia on 20140201). And at the end of the season, Barcelona's major patterns are **Pattern 2** (vs. Osasuna on 20140316, vs. Real Betis on 20140405, vs. Athletic Club on 20140421, vs. Villarreal on 20140428, vs. Getafe on 20140503) and **Pattern 4** (vs. Celta de Vigo on 20140327, vs. Granada CF on 20140413). The dynamic of playing patterns reflects the change of tactical strategies of coach Martino, who experimented with different starting lineups for the team during the first stage of the season. From our analysis, we can also see that although the starting lineup changed from match to match, the actual playing styles shared some similarity between matches.

In addition, it should be noted that, Barcelona tends to have more variation on the tactics when playing against stronger teams. From Figure 5, we can see that when playing against teams like Real Madrid and Atletico de Madrid, Barcelona did not stick to a single tactical pattern and played in different styles, as the match summarization bars are colored with more colors. On the other hand, when playing with other teams like Villarreal and Celta de Vigo, the bars are colored with one color mostly, indicating the match playing was dominated by a single tactical pattern.

We find that this kind of visualization can also improve the watching experience of match audiences. The annotation of pass segments to different tactical patterns provides a natural summarization of the match playing progress (an example is given in Section 4.5).

4.4 Player Role Analysis for Barcelona

We can also give a player role analysis with our T^{3}M. For the 10 tactical patterns learned with the passing data of Barcelona, we compute the scores indicating the importance of each player as described in Section 3.1.3. The results for major players are plotted in Figure 6.

From Figure 6, we can see that midfielders often make more passes in most tactical patterns, while forwards usually have less passes comparing to midfielders and defenders. And players with similar positions and roles in a team tend to have similar distributions of scores, that is, having similar representations on the basis of tactical patterns. In Figure 7, we plot the results of Principle Component Analysis (PCA) on the player scores defined in Section 3.1.3. The positions of the players are determined by the top 2 principle components. We can find that the players are roughly clustered according to their roles. Specifically, we know that both Lionel Messi's and Alexis Sanchez's common positions at Barcelona are right forward, while Neymar shows up as left forward more often. And Pedro Rodriguez can play at both sides as attacking winger in different tactics. We find that they have similar positions in the embedding space of PCA in this figure as well.

Furthermore, we can give a spatial analysis of player roles in different tactical patterns. As an example, Pedro's multifacet characteristic is studied here in details. In Figure 8, we plot the heat map of his receiving points in all the tactical patterns as the collection of the receiving points in each tactical pattern (indeed, this is a visualization of equation 9, see Section 3.1.3 for details). Heat maps for the top 4 tactical patterns are plotted, with the percentage of the passes annotated with each tactical pattern labeled above the corresponding plot. By studying these heat maps, we can find Pedro's active regions in different tactical patterns, which can reflect the different roles he actually plays. In the first 3 patterns (Pattern 2, Pattern 9 and Pattern 3), Pedro's active regions are in the left side of the playing field, while in the fourth pattern (Pattern 8) he mostly appears in the right wing. In fact, Pedro is a player who can play well as both left and right winger. Usually he plays as a left winger in

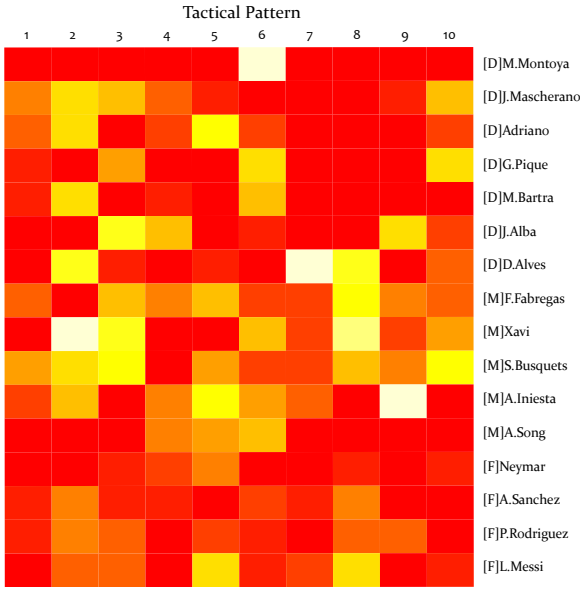

Figure 6: The weights of players in different tactical patterns. Brighter color means the player having a more important role in the tactical pattern (involved in more passes of the pattern).

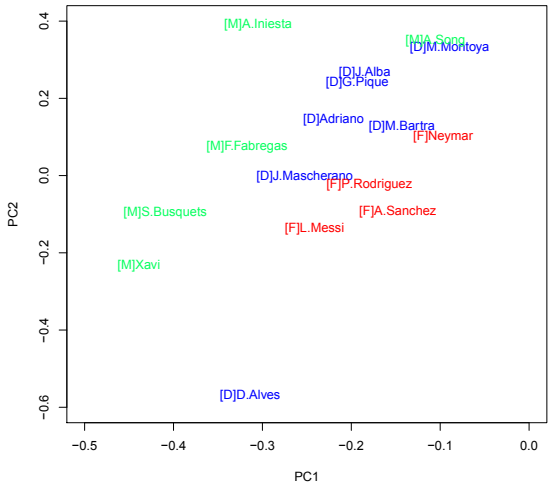

Figure 7: Plot of the players by the top 2 principle components on player scores. Defenders, midfielders and forwards are colored in blue, green and red, respectively.

Barcelona. However he is also considered as a good substitute for Lionel Messi in the right forward field. With T^{3}M, we can automatically recognize Pedro's multi-facet characteristic, which makes him a valuable player in the team.

4.5 Case Study: Barcelona vs. Real Madrid

As an example, here we study how to use T^{3}M to analyze a match between Barcelona and Real Madrid in La Liga 2013-2014. The match finished in 4:3, in favor of Barcelona. A visualization of the first half of the match is given in Figure 9. Both teams scored two goals before the break.

In this match, Iniesta scored first for Barcelona, with Messi's assist from the attacking region outside penalty area. This segment of passes was actually a cooperation between various players in the mid-field (e.g., Fabregas, Alba, Xavi, Busquets), which is recognized by T^{3}M as an offending pattern with these players in the mid-field (see top-left of Figure 9 and Pattern 3 in Figure 2.) In both the 20th and 24th minute, Real Madrid's Benzema finished a Di Maria cross from the left wing of the pitch. Using our model, we successfully recognize the playing pattern of Di Maria setting up chances in the left channel (bottom-left of Figure 9). And just before the break, Messi scored for Barcelona with a shot in the close range. The attack started from the left side by Alba and Iniesta, which is also successfully recognized by our model as an offending tactical pattern in the left wing (Pattern 2 in the bottom-left of Figure 9).

Indeed, the recognized patterns for the segments leading to goals in this match are very effective scoring patterns, with a scoring rate of 3.21% for Pattern 3 and 3.70% for Pattern 9, which are consistent with our analysis in Section 4.2. Furthermore, we can also find that Barcelona possesses the ball more often, noticing that the colored regions are thicker in the top band than the second. In fact, Barcelona is a team good at ball controlling, while Real Madrid is considered to be good at countering back.

5. RELATED WORK

Generally, the related works of this paper can be grouped into two major categories, namely *strategy analysis for soccer matches* and *topic modeling and its applications*.

5.1 Strategy Analysis for Soccer Matches

In almost all team sports, coordination and collaboration between players plays an important role. However, with only low-level observations, it is usually difficult to give a high-level description of a team's playing styles and strategies. We note that there are some related works tackling this problem, which inspired many subsequent works on team playing style analysis and other high-level pattern recognition applications. In a pioneering work [14], Intille *et al.* studied the problem of representing and recognizing complex multi-agent action in American football with probabilistic methods. For soccer playing style analysis, Lucey *et al.* [16, 17] studied the problem of automatic team strategy assessing. In their works, team behaviors were characterized via a study on entropy maps. And in the work of Gyarmati *et al.* [10], the problem of finding playing style in soccer was tackled with motif finding in pass networks, which allowed them to study the playing styles across different teams. For basketball, Fewell *et al.* [9] studied the ball-movement network over player roles. A connected team network was constructed and analyzed to find the probable paths of transition and playing. And in tennis, IBM SlamTracker [12] is a real-time platform providing player statistics and styles of play for a player to win a match.

As the sports data essentially have a spatial-temporal structure, there are also a number of works on analyzing sports activities with locations and/or temporal information. Recently, Yue *et al.* [26] tackled the problem of predictive modeling of dynamic team play in basketball. They used player location information from both teams, and learned semantically meaningful representations of team playing dynamics. For analysis on soccer data, in addition to various

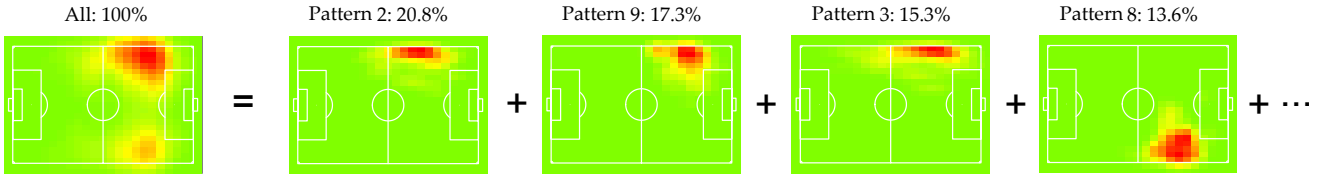

Figure 8: Pedro's active region in the pitch, decomposed by tactical patterns.

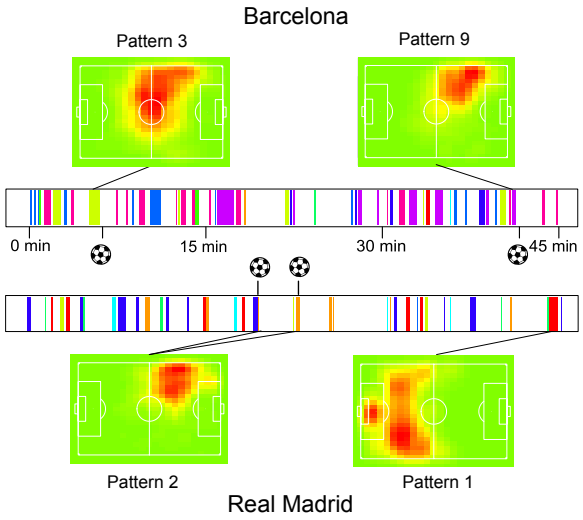

Figure 9: Visualization of the first half of the match between Barcelona and Real Madrid in La Liga on 24th Mar 2014. Segments of play are colored according to the different tactical patterns. We also plot the learned spatial patterns for typical pass segments (e.g., ended with goals).

works discussed above, Bialkowski *et al.* [4, 3] analyzed the player positions over time using an entire season's player and ball tracking data. Using an EM approach, they assigned players to roles and discovered the team's overall role distributions. In another work [24], the authors studied the problem of assigning players to roles at each frame. In addition, they studied the formations used by various teams, and found the most likely formation patterns for a team with spatiotemporal data. There are also some works on players and ball tracking as well as team plays detection with a visual system. We refer interested readers to the corresponding part of Lucey *et al.* [16] for related works.

Our work differs from these previous works as we use an unsupervised probabilistic model on the natural pass segments split by turnovers and stoppages to study the high-level tactical patterns. In addition, we take both the structure of the pass graph between players and the players' locations into our model, so that we can have a simultaneous analysis of *what kind of tactic a team prefers to play*, and *where on the pitch this tactical pattern is usually performed*. With these improvements and innovations, our approach can give realistic results of soccer team tactic analysis, which can promote various applications as described in Section 3 and potential future directions.

5.2 Topic Modeling and Its Applications

Topic models are tools for clustering co-occurrent words into various topics. Many topic models have been proposed, such as probabilistic latent semantic indexing (pLSI) [13], Latent Dirichlet Allocation (LDA) [6], author-topic model [21], correlated topic model (CTM) [15], and dynamic topic model (DTM) [5], etc. By assuming a latent assignment for each word to a topic, these models can perform unsupervised clustering analysis of various kinds of data sets. In addition to the traditional applications of text retrieval and information extraction, topic models have also been widely used for unsupervised analysis in several domains, e.g., activities discovery with sequential data and videos, object categorization in images, and sentiment analysis, etc. For example, Shen *et al.* [22] studied the problem of temporal activities discovery, and showed that with a topic modeling approach, they could obtain improved next activity prediction and sequence clustering results. For action discovery in videos, Wang *et al.* [23] applied unsupervised hierarchical Bayesian models to the problem of activity perception in video scenes, and tackled the problem of high-level semantic analysis and typical activities summarizing with their video based system. Also, in [20], Niebles *et al.* studied the problem of automatic human action categorization from video sequences. For applications in object categorization, Fergus *et al.* [8] applied a topic model (TSI-pLSA) to visual words in images and obtained competitive performance on standard test set with existing method. And for sentiment analysis, Zhu *et al.* [27] proposed several novel models (eToT,meToT,eDTM) based on topic models and demonstrated the effectiveness of these models.

In a similar manner, we use topic models to perform unsupervised clustering of soccer passing data in match logs, which is to our knowledge, the first work on applying topic model to soccer tactic analysis. In addition, enhanced by location information, the proposed model T^{3}M can learn the players' pass graphs and the spatial patterns of passes simultaneously, with hybrid data of discrete passing events and continuous receiver locations.

6. CONCLUSION AND FUTURE WORK

In this paper we proposed an unsupervised approach to automatically discerning the typical tactics, i.e., tactical patterns, of soccer teams through mining the historical match logs. To be specific, we first developed a novel model named Team Tactic Topic Model (T^{3}M) for learning the latent tactical patterns. A unique perspective of T^{3}M is that it can model the locations and passing relations of players simultaneously. Furthermore, we also demonstrated several potential applications enabled by the proposed T^{3}M, such as automatic tactical pattern discovery, pass segment annotation, and spatial analysis of player roles. Finally, we conducted extensive empirical evaluations for our approach based on the data collected from La Liga 2013-2014. In fact, by visualizing the results obtained from T^{3}M, many meaningful tactical patterns and interesting discoveries have been obtained, which validated the effectiveness of our approach.

In the future, we plan to develop more novel applications based on T^{3}M, such as winning rate prediction, and integrate them into our soccer intelligent system.

Acknowledgments

We are especially grateful to the Baidu Big Data Group for their assistance with the collection of our data. We would also like to thank the anonymous reviewers for their helpful feedback and suggestions. This work was supported in part by National Natural Science Foundation of China: 61402019, and in part by China Postdoctoral Science Foundation: 2014M550015.

References

[1] Opta. http://optasports.com/.

[2] World's most popular sports by fans, 2015. http://www.topendsports.com/world/lists/popular-sport/fans.htm [Online; accessed Feb 11, 2015].

[3] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. Matthews. Identifying team style in soccer using formations learned from spatiotemporal tracking data. In *ICDM, Workshop on Spatial and Spatio-temporal Data Mining (SSTDM)*, 2014.

[4] A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. Matthews. Large-scale analysis of soccer matches using spatiotemporal tracking data. In *ICDM*, 2014.

[5] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *ICML*, 2006.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 2003.

[7] H. Collignon, N. Sultan, and C. Santander. The sports market - major trends and challenges in an industry full of passion, 2015. http://www.atkearney.com/documents/10192/6f46b880-f8d1-4909-9960-cc605bb1ff34 [Online; accessed Feb 11, 2015].

[8] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *Tenth IEEE International Conference on Computer Visio (ICCV)*, 2005.

[9] J. H. Fewell, D. Armbruster, J. Ingraham, A. Petersen, and J. S. Waters. Basketball teams as strategic networks. *PLoS ONE*, 2012.

[10] L. Gyarmati, H. Kwak, and P. Rodriguez. Searching for a unique style in soccer. *arXiv preprint arXiv:1409.0308*, 2014.

[11] G. Heinrich. Parameter estimation for text analysis. Technical report, University of Leipzig, Germany, 2005.

[12] D. Henschen. IBM serves new tennis analytics at wimbledon, Jun 22, 2012. http://www.informationweek.com/software/information-management/ibm-serves-new-tennis-analytics-at-wimbledon/d/d-id/1104987 [Online; accessed Feb 11, 2015].

[13] T. Hofmann. Probabilistic latent semantic indexing. In *22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999.

[14] S. S. Intille and A. F. Bobick. A framework for recognizing multi-agent action from visual evidence. In *National Conference on Artificial Intelligence (AAAI)*, 1999.

[15] J. D. Lafferty and D. M. Blei. Correlated topic models. In *NIPS*, 2005.

[16] P. Lucey, A. Bialkowski, P. Carr, E. Foote, and I. Matthews. Characterizing multi-agent team behavior from partial team tracings: Evidence from the english premier league. In *Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI)*, 2012.

[17] P. Lucey, D. Oliver, P. Carr, J. Roth, and I. Matthews. Assessing team strategy using spatiotemporal data. In *KDD*, 2013.

[18] A. Miller, L. Bornn, R. Adams, and K. Goldsberry. Factorized point process intensities: A spatial analysis of professional basketball. In *ICML*, 2014.

[19] K. P. Murphy. *Machine learning: a probabilistic perspective*. The MIT Press, 2012.

[20] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision (IJCV)*, 2008.

[21] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *20th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2004.

[22] Z. Shen, P. Luo, Y. Xiong, J. Sun, and Y. Shen. Topic modeling for sequences of temporal activities. In *ICDM*, 2009.

[23] X. Wang, X. Ma, and W. E. L. Grimson. Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2009.

[24] X. Wei, L. Sha, P. Lucey, S. Morgan, and S. Sridharan. Large-scale analysis of formations in soccer. In *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2013.

[25] Wikipedia. Tiki-taka - wikipedia, the free encyclopedia, 2015. http://en.wikipedia.org/wiki/Tiki-taka [Online; accessed Feb 11, 2015].

[26] Y. Yue, P. Lucey, P. Carr, A. Bialkowski, and I. Matthews. Learning fine-grained spatial models for dynamic sports play prediction. In *ICDM*, 2014.

[27] C. Zhu, H. Zhu, Y. Ge, E. Chen, and Q. Liu. Tracking the evolution of social emotions: A time-aware topic modeling perspective. In *ICDM*, 2014.