EG UK Computer Graphics & Visual Computing (2024)
A. Slingsby and D. Hunter (Editors)

Interplay of Visual Analytics and Topic Modeling in Gameplay Analysis

L. Moussavi[1], G. Andrienko[1,2], N. Andrienko[1,2], and A. Slingsby[1]

[1]City, University of London, UK
[2]Fraunhofer Institute IAIS, Sankt Augustin, Germany

Abstract

Spatio-temporal event sequences consist of activities or occurrences involving various interconnected elements in space and time. Exploring these sequences with topic modeling is a relatively new and evolving research area. We use topic modeling to analyze football games, as an example of complex and under-explored spatio-temporal event data. A key challenge in topic modeling is selecting the most suitable number of topics for the downstream application. Selecting too few topics oversimplifies the data, merging distinct patterns, whereas selecting too many can fragment coherent themes into overlapping categories. We propose a visual analytics technique that uses dimensionality reduction on topics derived from multiple topic modeling runs, each with a different number of topics. Our technique organizes all the topics in a hierarchical layout based on their spatial similarity, making it easier to make an informed decision about selecting the most expressive set of topics that represent distinctive spatial patterns. We apply our visual analytics technique to a football dataset, illustrating how it can be used to select an appropriate set of topics for this data. We then use these topics to represent game episodes, which help us summarize game dynamics and uncover insights into the games.

CCS Concepts

• ***Human-centered computing*** → *Visual Analytics;*

1. Introduction

Topic modeling is a powerful technique that identifies abstract themes or topics within a collection of documents [VK20]. It discovers patterns of words that represent topics, facilitating the categorization of documents based on these identified topics. Originally developed for text mining to uncover hidden semantics in text data, the underlying principles of topic modeling have been adapted to various other domains. This adaptability has enabled the application of topic modeling across a wide range of fields, including image analysis [WG07] and bioinformatics [HHQ12].

Chen et al. [CAA*20] and *Andrienko et al.* [AAH23] have used topic modeling to analyze *event sequences*. Event sequences are activities or occurrences that involve various interconnected elements. These events typically involve multiple interconnected elements, making them challenging to disentangle. Topic modeling provides a level of abstraction that identifies recurring patterns, facilitating effective analysis. *Chen et al.* [CAA*20] have used topic modeling to analyze non-spatial event sequences, such as user commands during a session in a security management system, and spatial event sequences such as visiting behaviors in an amusement park. *Andrienko et al.* [AAH23] have utilized topic modeling to analyze *spatiotemporal event sequences*, where both a position and a timestamp characterize each event. They applied topic modeling to trajectories derived from road traffic movement data. More recently, *Andrienko et al.* [AAS23] have used topic modeling on time intervals called episodes to understand the distribution of multi-attribute dynamic characteristics across different episodes.

In this paper, we use topic modeling to analyze sequences of spatially-referenced events in football games [PM19]. These types of event data are very rich with various complexities (e.g., different behaviors for different teams and players and their effect on game outcomes), which may be possible to capture and model as topics. We split game events into episodes (Section 3) and treat the episodes as documents for topic modeling.

One of the biggest challenges for topic modeling is how to obtain the most suitable combination of well-interpretable and distinctive topics, which requires choosing an appropriate number of topics for the algorithm to extract. Selecting too few topics can oversimplify data by merging distinct patterns, while too many topics can fragment coherent themes into overlapping categories. In this paper, we propose a visual analytics technique (Section 5) for making an informed decision about determining the most suitable number of topics for spatial event data.

Our technique performs topic modeling iteratively (Section 4),

DOI: 10.2312/cgvc.20241213

so that each run extracts a different number of topics from a specified range. Each run's topics are projected into 1D space taking their similarities into account, and the sets from all runs are then arranged in a hierarchical graph layout where edges connect similar topics from different levels of the hierarchy. Individual topics are represented visually by small spatial heatmaps. This makes it possible to compare visually the topics from each run and across the runs, allowing the identification of stable topics that are preserved across multiple runs. Such stability indicates topic significance. We add interactivity to the visualization to facilitate the selection of the minimum sufficient number of topics that represent distinctive spatial patterns.

While our proposed visual analytics technique can work for any spatial event data (as long as we can represent individual topics by interpretable compact images such as spatial heatmaps or glyphs), we have applied it to a case study involving a football dataset [PM19]. We demonstrate how our technique can be used to select the representative set of topics for this data (Section 5). The selected set of topics is then utilized for representing episodes, which helps us summarize game dynamics and reveal insights into a single game (Section 6) or series of games (future work).

In summary, our contributions are A) a visual analytics technique for selecting the most suitable result from multiple runs of topic modeling on spatial event sequences, B) a demonstration of how our technique can be used to discover spatially consistent and easily distinguishable topics from a football dataset, and C) a demonstration of how the obtained topics can be used to summarize football game dynamics and reveal insights into the games.

2. Related Work

Topic Modeling.

Topic modeling is a powerful technique used to identify hidden themes from a collection of documents [VK20]. It is a tool commonly employed in text mining to discover the concealed semantics within text data. Topic modeling aims to discover patterns of words that represent topics, allowing for the categorization of documents based on these topics. Two popular methods used in topic modeling are Latent Dirichlet Allocation (LDA) [BNJ03] and Non-negative Matrix Factorization (NMF) [LNC*17]. They have their distinct mathematical foundations but share the same goal, use same inputs, and produce similar outputs.

To explain the topic modeling process, imagine a scenario with a large, mixed collection of academic papers (documents) spanning three main subjects: history, biology, and mathematics, where the subjects are not initially clear. When topic modeling is applied to these documents, it uncovers distinct themes corresponding to each subject area. This method examines word frequencies and their co-occurrence patterns within the texts. For example, history-related papers might frequently use words like "empire", "medieval", and "revolution". In contrast, biology papers might feature terms such as "DNA", "evolution", and "species", while mathematics papers might include words like "equation", "calculus", and "theorem". *El-Assady et al.* [EASD*18] proposed a visual analytics framework for performing and optimizing topic modeling on text data. It enhances the traditional topic modeling process by incorporating user interactivity and speculative execution, allowing for a more flexible and interpretable approach to text data analysis. In contrast to their approach that focuses on text data, our visual analytics approach is designed for spatio-temporal event sequences.

While topic modeling techniques were originally developed for text data, the underlying principles can be adapted to other domains. This adaptability has allowed topic modeling to be embraced for a broad spectrum of applications such as Image Data [WG07] and Bioinformatics [HHQ12]. Topic modeling on spatio-temporal event sequences involves applying algorithms to uncover hidden topics within a dataset, which helps to understand how these topics/behaviors change across space and time. By analyzing the frequency of topics and their co-occurrence with other situations in various locations and over time, topic modeling can reveal insights into how certain activities or occurrences are distributed and evolve over space and time and in different situations.

Topic modeling employs matrices to represent the relationships between documents, terms, and topics. The input to topic modeling is a document-term matrix storing the distribution of terms (columns) across documents (rows). When applying NMF to the document-term matrix, it decomposes the matrix into two lower-dimensional matrices, W (document-topic matrix) and H (topic-term Matrix). Each element in the W matrix represents the weight of a topic within a document, while each element in the H matrix represents the weight of a term within a topic. Review [VK20] suggests that NMF is more suitable for short texts such as social media messages. Similar findings were reported in [AAS23] when applying NMF to game episodes characterized by multivariate time series. These observations motivated us to utilize NMF in our study.

Selecting the Number of Topics for Topic Modeling on Spatio-Temporal Event Sequences.

Previous research has utilized topic modeling to analyze road traffic movement data by treating each trajectory as a document consisting of terms representing either visited places or moves between them, depending on the chosen representation [AAH23]. They investigated methods for selecting the minimal sufficient number of topics by projecting the outputs of NMF into a 2D space, determining concentrations of similar topics from multiple runs, and using these concentrations to select the desired topics number.

Similar to [AAH23], we run topic modeling multiple times with various numbers of topics. However, we project topics from all runs into a 1D space (rather than 2D). 1D projections are easier to interpret as each data point can be represented on a single axis, making it easier to compare and analyze. In addition, using a 1D representation, we can show a more detailed visualization for each data point (Figure 2).

Topic Modeling on Football Data.

Andrienko et al. [AAS23] developed a general approach for analyzing multivariate temporal data through the identification of episodes and the use of topic modeling techniques. In their analysis, the variation in each attribute's value within an episode is represented by symbols, forming a 'word' (term in our terminology). The combined variations of multiple attributes within an episode create 'words', and these words across different episodes collectively constitute a 'text'. These texts are then analyzed using topic

modeling to identify patterns and recurring themes. Their methodology was applied to two football matches from the German Bundesliga 2019–2020 season. They made episodes of a fixed duration of 10 seconds using a sliding window approach. They also designed an interactive visualization tool to help analysts interpret these topics. In our work, we make episodes based on the semantics of the games (Section 3). Our data included specific event types such as passes, free kicks, and shots. We use the spatial locations of these events in discretized form as our terms.

Wang *et al.* [WZH*15] developed a novel topic modeling method specifically meant for football data. The method is capable of modeling both locations of the players and their passing relationships. The topics are supposed to capture various tactical patterns. The authors use heatmaps for visualizing individual topics and, after assigning distinct colors to the topics, represent the sequence of tactical patterns used during a game by a horizontal arrangement of colored bars. Such representations of different games are placed below one another to enable comparisons between the games. In our work, we used a generic method of topic modeling. Our main contribution is a visual analytics technique supporting comparisons between the results of topic modeling with different parameter settings and choosing the most appropriate one. We also demonstrate how the chosen result can be used for downstream analysis, including exploration of the course of one game. We note that our proposed visual analytics method does not depend on the choice of the topic modeling technique, as it only uses the results, which have the same structure and meaning across different algorithms. We have performed our experiments using NMF, a standard and scalable topic modeling approach. The football-specific algorithm proposed by Wang *et al.* [WZH*15] might not be scalable to our relatively large data, since their method has many parameters such as a Gaussian distribution per player and topic.

3. Extracting Episodes from Football Dataset

We demonstrate our approach using the football dataset published by [PM19]. The dataset contains data from the 2017/2018 season of five top-tier European football leagues (Spain, Italy, England, Germany, and France), and two major international tournaments, namely the 2018 World Cup and the 2016 European Championship. Overall, this collection contains information on seven prominent football competitions.

The dataset covers details on events, teams, matches, players, referees, and coaches [PCR*19], including 1,941 matches with a total of 3,719,995 recorded events (1,917 events per game on average). Previously, this dataset underwent statistical analysis to assess players' performance metrics [PCF*19].

The events in the dataset include the following actions or happenings: *duel*, *foul*, *free kick*, *goalkeeper leaving line*, *interruption*, *others on the ball*, *pass*, *save attempt*, *shot*, and *offside*. Each event is detailed with sub-type (e.g., a *pass* could be a hand pass, cross, etc.), timestamp, involved player(s), and pitch positions for origin and reception.

We apply topic modeling to episodes of ball movement that contain sequences of events. Using episodes rather than single events makes it possible to extract meaningful information from the interaction between multiple players on different parts of the football pitch and during a period. We pre-process the dataset to extract episodes in a way similar to previous works [AAA*19, AAS23]. We define each episode as the duration during which a team possesses the ball until it loses it. We allow brief interruptions, meaning an episode does not end if the opposing team briefly gains possession for less than 10 seconds. For instance, if team A is moving the ball through a series of events and team B momentarily disrupts possession with a touch, team A's episode will still continue to incorporate subsequent events if the interruption lasts no more than 10 seconds. Simultaneously, an episode for team B will be recorded from the moment of their interruption. With this pre-processing, the average number of episodes for each match is 316 (158 per team), and the average number of events per episode is 6.07.

Table 1 shows an example of the initial episodes from the 6th May 2018 match between Barcelona and Real Madrid, which ended in a 2-2 draw. Thus, Real Madrid's first episode ended with a foul from Barcelona. Each column in the table represents an individual event from the first episodes for each team, labeled R-E1 and B-E1 respectively.

4. Topic Modeling on all Matches of the Football Dataset

In this section, we discuss how we performed topic modeling on all the matches of Pappalardo *et al.*'s football dataset [PCR*19].

Topic modeling (NMF in our experiments) is performed on a collection of documents, where documents are represented with terms. We define our documents as the episodes of ball possessions by teams during matches (Section 3). For each team, we consider a pitch to be oriented upwards in the direction of the team attack. We divide the pitch into grid cells with 15 rows and 12 columns, consisting of 180 grid cells in total. These grid cells will be treated as terms. For each team in a match, we generate a matrix $X_{M\times N}$ that represents the ball positions when controlled by that team, where M is the number of episodes (documents), N is the number of grid cells (terms), and each element X_{ij} equals the number of times that in episode i the ball-related event was recorded within grid cell j.

Our goal is to identify common patterns or topics across all teams and matches, enabling us to compare different teams' behaviors during a match or a season. Applying NMF separately to each matrix X, i.e., one team's episodes during one match, would result in disparate sets of discovered topics for each team, complicating the comparison of teams' behaviors. Even if we employ techniques to align topics, such as assigning similar colors to similar topics (topics that are close in a projection space) or unifying topic numbers, it remains challenging to interpret different colors across many matches and teams. Moreover, the same topic numbers would not necessarily represent the same patterns across different matches and teams. Therefore, we combine the matrices of all teams and matches in all seven competitions in the dataset into a matrix $\mathbf{X}_{\mathbf{M}\times N}$, where $\mathbf{M}$ is the number of all episodes of all teams in all matches. We apply NMF to this combined matrix and thus, our approach finds a document-topic matrix $\mathbf{W}_{\mathbf{M}\times K}$ and a topic-term matrix $\mathbf{H}_{K\times N}$, where $\mathbf{X} \approx \mathbf{W} \times \mathbf{H}$. These matrices are supposed to reflect common tactical patterns across all teams in all leagues.

	R-E1	R-E1	R-E1	R-E1	R-E1	R-E1	R-E1	B-E1	B-E1
Team ID	675	675	675	675	675	675	675	676	676
Player ID	3321	14723	3306	3309	3915	3306	40756	3476	3476
Position 1	(50, 49)	(37, 40)	(30, 23)	(26, 57)	(8, 40)	(11, 14)	(25, 21)	(79,75)	(76, 77)
Position 2	(37, 40)	(30, 23)	(26, 57)	(8, 40)	(11, 14)	(25, 21)	(23, 24)	(76, 77)	(79, 75)
Match Half	1H	1H	1H	1H	1H	1H	1H	1H	1H
Event Second	3.275	5.109	7.110	8.912	11.290	13.934	15.949	16.32	17.35
Event ID	8	8	8	8	8	8	1	1	2
Event Name	Pass	Pass	Pass	Pass	Pass	Pass	Duel	Duel	Foul
Sub-event ID	85	85	85	85	85	85	11	12	20
Sub-event Name	Simple pass	Simple pass	Simple pass	Simple pass	Simple pass	Simple pass	Ground attacking duel	Ground defending duel	Foul
Tags	1801	1801	1801	1801	1801	1801	503, 703, 1801	504, 701, 1802	

Table 1: *First episodes of Real Madrid and Barcelona teams during their match in May 2018. Each column represents an individual event from the first episodes for each Real Madrid and Barcelona team, shown respectively with R-E1 and B-E1. As shown in the table, Real Madrid's first episode consists of seven events (six passes followed by a duel), while Barcelona's first episode contains two events (duel and foul). Each row provides details about a particular event.*

5. Visual Analytics Technique for Selecting the Number of Topics

In this section, we present our visual analytics technique for selecting the most appropriate result from multiple runs of topic modeling with different parameter settings, particularly varying the number of topics to extract. Our goal is to find the smallest possible set of topics that comprehensively encompass all important information. These topics must be spatially consistent and easily distinguishable. Specifically, the topics should remain consistent across different runs, avoiding redundant patterns that can be combined into a single topic. While our approach is applicable to any spatial event data, we demonstrate it using a football dataset for simplicity. We iteratively apply NMF to the data aiming at different numbers of topics $K \in \{K_{\min}, \ldots, K_{\max}\}$. At each iteration for the number of topics K, we obtain a topic-term matrix $\mathbf{H}^K_{N\times K}$, where each row of the matrix contains the weight of one topic.

Previous work has applied dimensionality reduction techniques to project the topic-term matrices into a 2D space [AAH23]. While this is useful in comparing topics obtained from different values of K, we argue that our approach enables a deeper analysis by projecting into a more compact 1D space.

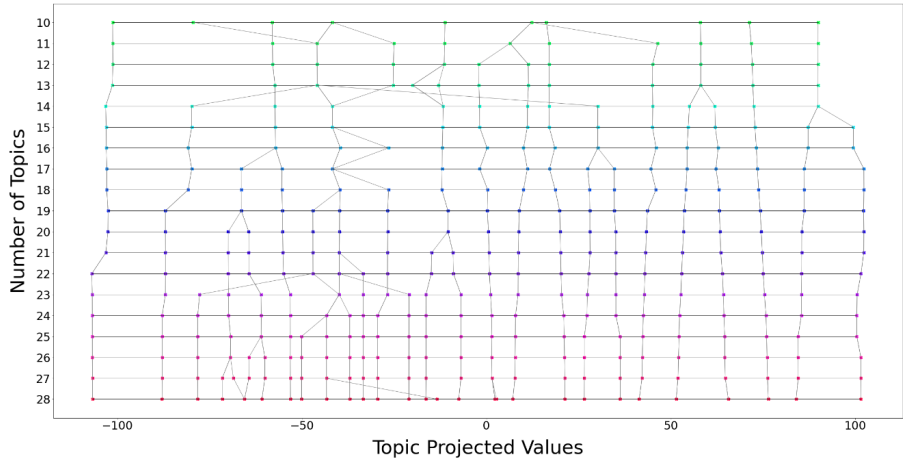
We project the topics onto a shared 1D space across different numbers of topics. For each number of topics K, we add K points onto a line. We then stack the lines in a single image in the order of increasing topic number, so that the top line contains the projected topics for $K_{\min}$ and the bottom line contains the projected topics for $K_{\max}$. Figure 1a shows the projected topics in 1D obtained using the t-SNE dimensionality reduction technique [VdMH08], for K ranging from 10 to 28. We connect each topic in the K topics solution to its most similar topics in $K-1$ and $K+1$ topics solutions. The resulting branching shows how topics split into higher topic number solutions. The most similar topics are determined as the nearest neighbors in the original N-dimensional space.

Inevitably, dimensionality reduction introduces distortions, i.e., the distance between the points in the projection space (1D) might not always correlate with their distance in the original space (N-dimensional). In an ideal scenario, with the absence of distortions, comparable topics in each row should maintain their content and position. In this situation, each point will be connected to one of the closest points on both the top and bottom lines (either to its left or right). Without any distortions, the connecting links remain uncrossed forming a trapezium-like shape. We measure the distortion introduced by dimensionality reduction as the number of crossing links, i.e., links that connect a point to a point other than the closest one in the image space (7 crossing links in Figure 1a).

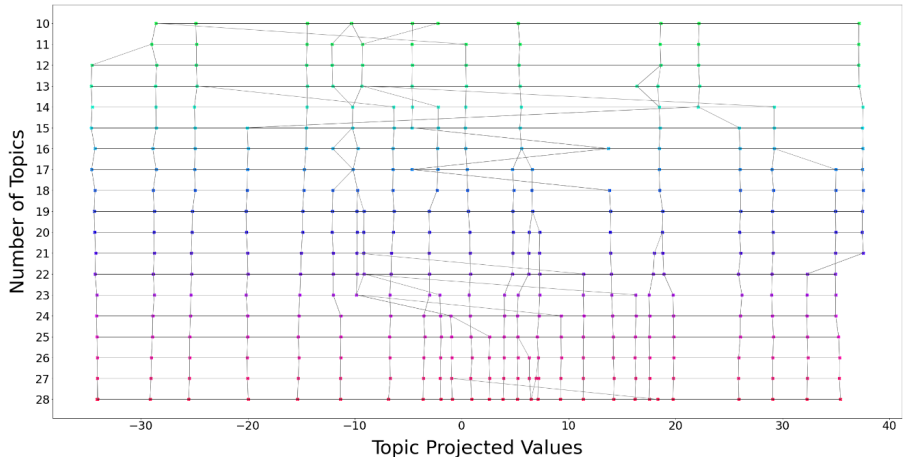
Dimensionality reduction can be performed with different methods. We explored three prominent dimensionality reduction techniques: MDS [Kru64], UMAP [MHM18], and t-SNE [VdMH08]. For both UMAP and t-SNE, we carefully adjusted the parameters to minimize distortion. Specifically, we selected a neighborhood size of 15 for UMAP and a perplexity value of 10 for t-SNE. All three methods produced consistent results, but t-SNE demonstrated the least distortion, yielding the clearest outcomes. Figure 1a presents the results for t-SNE that has the least distortion with 7 crossing links. Figure 1b shows the results for UMAP with 15 crossings that has more distortion than t-SNE, but less than MDS. Figure 1c shows the results for MDS with many crossing links.

5.1. Spatial Visualization supporting Topics Interpretation

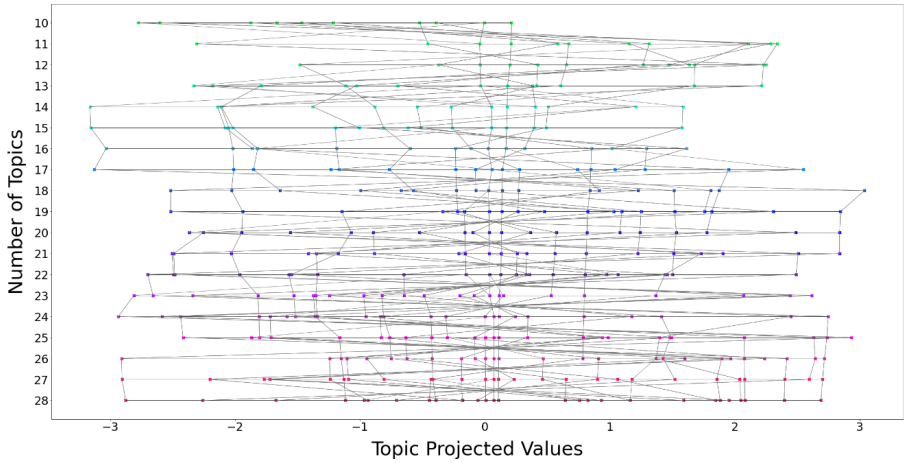
To make an informed decision about the optimal number of topics, it is crucial to interpret the composition of topics within each set from a single run, understand their patterns, and compare results across different runs. To facilitate this, we created the visualization shown in Figure2a which addresses these criteria. This visualization adopts the layout presented in Figure 1a, but replaces the dots with small heatmaps showing topics' spatial signature. The heatmaps' rectangle shape represents a football pitch (with attacking direction upward) and they visualize the topic-term matrix on the pitch, displaying the weight of each topic within each grid cell. In other words, these heatmaps highlight the specific areas of the football pitch that each topic covers. Each topic's vector (180 values for one heatmap) has been normalized so that the values of its grid cells sum to 1. The connecting links are omitted, as they are no longer necessary, and retaining them would only clutter the Figure



(a) *Topic projection using t-SNE has the least distortion with 7 crossing links.*



(b) *Topic projection using UMAP, with 15 crossing links, shows more distortion than t-SNE but less than MDS.*



(c) *Topic projection using MDS with many crossing links has the most distortion.*

Figure 1: 1*D projections of topics obtained from topic modeling are shown for topic numbers ranging from 10 to 28. Each row represents the projection for a specific topic number. The dimensionality reduction techniques applied in these projections are: (a) t-SNE, (b) UMAP, and (c) MDS. Each topic is connected to its most similar topics in the projections above and below. The projection technique with fewer crossing links (t-SNE in our case with 7 crossings) is the best option for further analysis.*

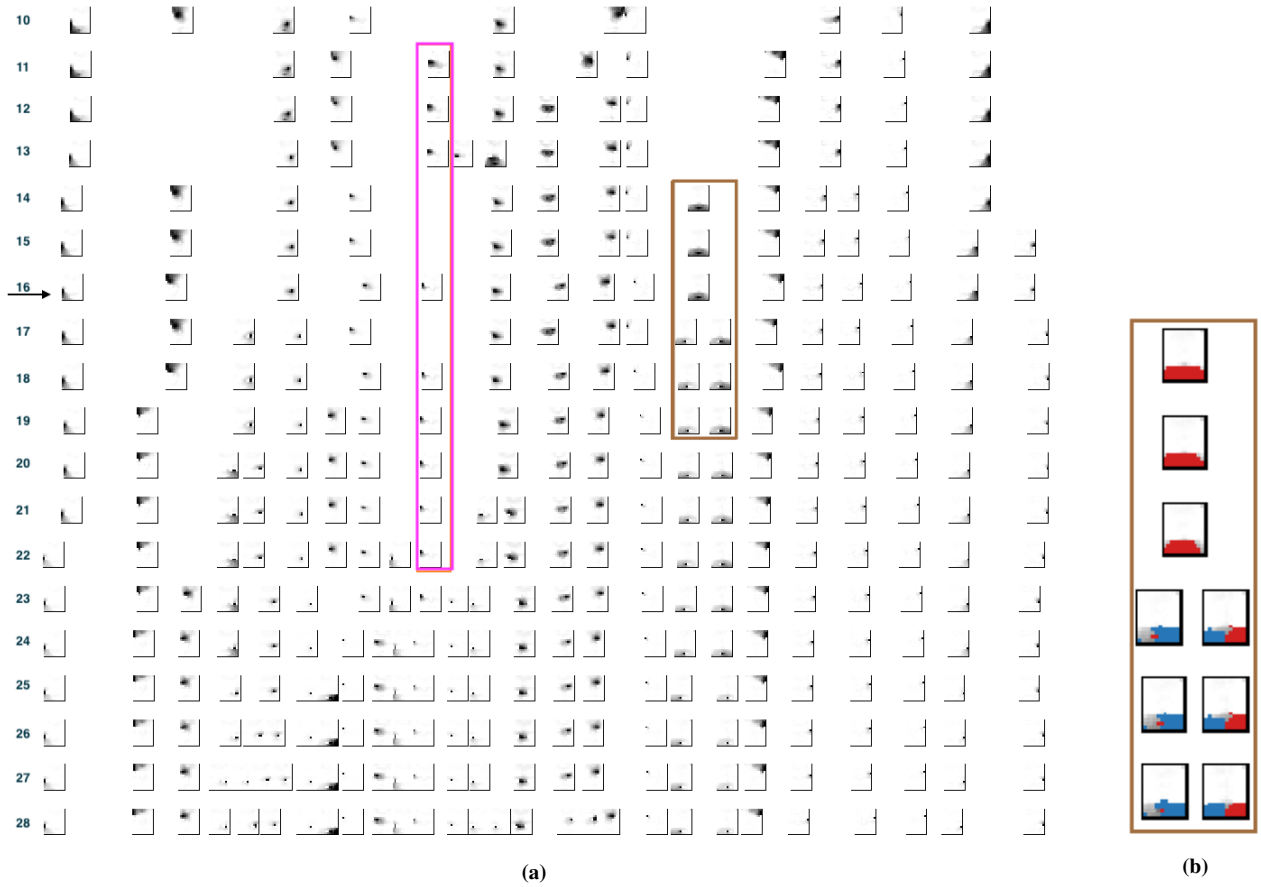


Figure 2: ***Figure (a)*** *presents a comprehensive visualization of topic signatures (spatial heatmaps), derived from topic modeling across a range of topic numbers (10 to 28, as labeled at each line). Each heatmap represents a football pitch, with its grids colored black proportionately according to the topic-term matrix. In all these heatmaps, the attacking direction is from bottom to top. The final selected topic number is 16 (indicated by an arrow). The reason for this selection is that topics in this row are both distinct and spatially consistent. The group in the pink rectangle illustrates what spatial consistency means, while the group in the brown rectangle demonstrates redundancy.*
Figure (b) *provides a detailed view of the topics selected within the brown rectangle. The interactive tool is used to select the topics and color their cells red/blue when they are higher/lower than the average of the selection. The presence of red cells in the first three rows and a mix of red and blue cells in the last three rows suggests that the cell values have been distributed between two similar topics in the branch.*

Now, the heatmaps allow us to visually interpret the topics in each row and compare their content with those above and below. Using Figure 2a, we can examine the similarity of topics within the same row (for instance, moving from left to right in the scenario with 11 topics) and their resemblance to topics in adjacent rows (for example, comparing scenarios with 10 and 12 topics).

We observe gradual transitions from one topic to another in both vertical and horizontal directions. When looking vertically, it is interesting to note that the patterns in the rows with a smaller number of topics appear to be combinations of the patterns from the rows with a larger number of topics below them. When looking horizontally, we must remember that some level of distortion is inevitable, so we need to be careful when making sense of positions along the 1D axis. However, despite this, we can still find consistent patterns through different experiments with varying numbers of topics, showing how topics evolve across several rows.

This representation enables finding the most suitable set of topics for further analysis. In particular, we are looking for the smallest number of topics that satisfy two constraints: A) its topics are spatially consistent across a wide range of number of topics, i.e., they appear consistently, with some fluctuations, within a range of rows, and B) its topics are easily distinguishable, i.e., they do not have redundant patterns that can be replaced by one.

To assist the analyst in finding such number of topics that satisfy the above constraints, we added interactivity to this visualization. Our tool enables the analyst to pick a group of topics for a detailed exploration, in which the analyst can compare each of the selected topics against their average. For each topic, grid cells that fall below the average will be colored blue, while those exceeding the average will turn red, making it easier to perceive and understand the differences. As such, the tool color codes the cells to accurately reflect the comparison of heatmap pairs or groups within the entire range of different number of topics. Figure 2b is a screenshot of a

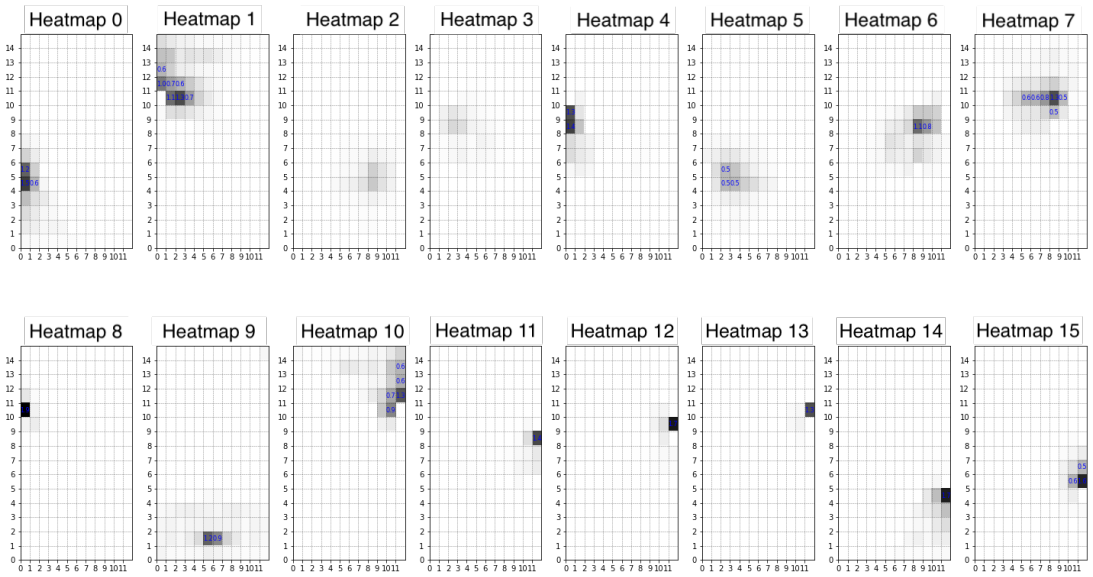


Figure 3: *A detailed view of the selected topic signatures/spatial heatmaps on a football pitch, with 16 topics.*

selection of heatmaps from Figure 2a by our tool. It shows the dominance of cells' red colors for topics in the upper rows, contrasted with the mostly blue cells in the topics of the branching lower rows, and illustrates how a single topic at the top has been divided into multiple topics below. The selection is shown by a brown rectangle in both 2a and 2b figures.

Considering the two constraints, we selected 16 as the most suitable number of topics. Figure 3 displays the topic signatures/heat maps for this selection. The reason for this selection is that the topics in this row show spatial consistency (encompassing all topics or their similar ones that consistently appear in other runs) and distinctiveness (excluding topics that have redundant patterns that can be shown as one). In particular, the selection enclosed in a pink rectangle shows a consistent pattern that does not appear in the result with 15 topics but does appear in many other results, including the one with 16 topics. In addition, the brown selection shows that in the results with 17 or more topics, one topic is split into two topics that have high overlap.

6. Using Topics to Understand Game Dynamics

One potential use case includes illustrating a team's tactics during a game, tracking how topics develop over time, and identifying the key events that correspond to the emergence of different topics. Figure 4 visualizes episodes and their related details for the first half of the Barcelona vs. Real Madrid match, which ended in a 2-2 draw in May 2018. We have employed a standard layout commonly used in football reports and applications. This design has the game timeline on its x-axis starting at 0 and extending beyond 45 minutes to cover the entire first half and provides details for each team's episodes on its y-axis. Barcelona's information is displayed above the dashed divider line and Real Madrid's information is below it, creating two sub-figures.

In each sub-figure, stacked bar charts depict active topics during episodes, placed on a timeline to show the temporal order of the episodes. We chose to show every episode by the same width to prevent episodes from covering each other. Still, with the same length, some short episodes would be covered by their next episode, in such cases we move the the second episode (and the following ones) by an epsilon to prevent masking. There are 16 small bars within each stack, representing the 16 topic IDs. The order of the topics, from bottom to top, is the same as in Figure 3 from left to right. Each bar's color intensity indicates the topic's weight, with green for Barcelona and red for Real Madrid. The topic ID with the most weight (the dominant topic ID) is also written on top of the stack bar for easier spotting.

Additionally, we marked the successful episodes with an "X" above the most dominant row. Successful episodes are characterized by significant progression into the opponent's part of the pitch. An episode is considered successful if its final event is a "shot" or "goal", or if the final event is one of the "pass", "duel", or "others on the ball" events and that event's location is within the last 20% of the pitch. We also defined match key events to be Received a Goal, Scored a Goal, Own Goal, Dangerous Ball Loss, Red Card, Yellow Card, Second Yellow Card, and Substitutions and respectively abbreviated them with (RG), (SG), (OG), (DBL), (RC), (YC), (SYC), (SUB). We use successful episodes and these abbreviations as a tool to help the analyst understand game development and the relationship between topics and match events more easily.

We note that for successful events, usually, at least one of the topics 1 (attack from pitch's left side) or 10 (attack from pitch's right side) are activated; however, they are not necessarily the most

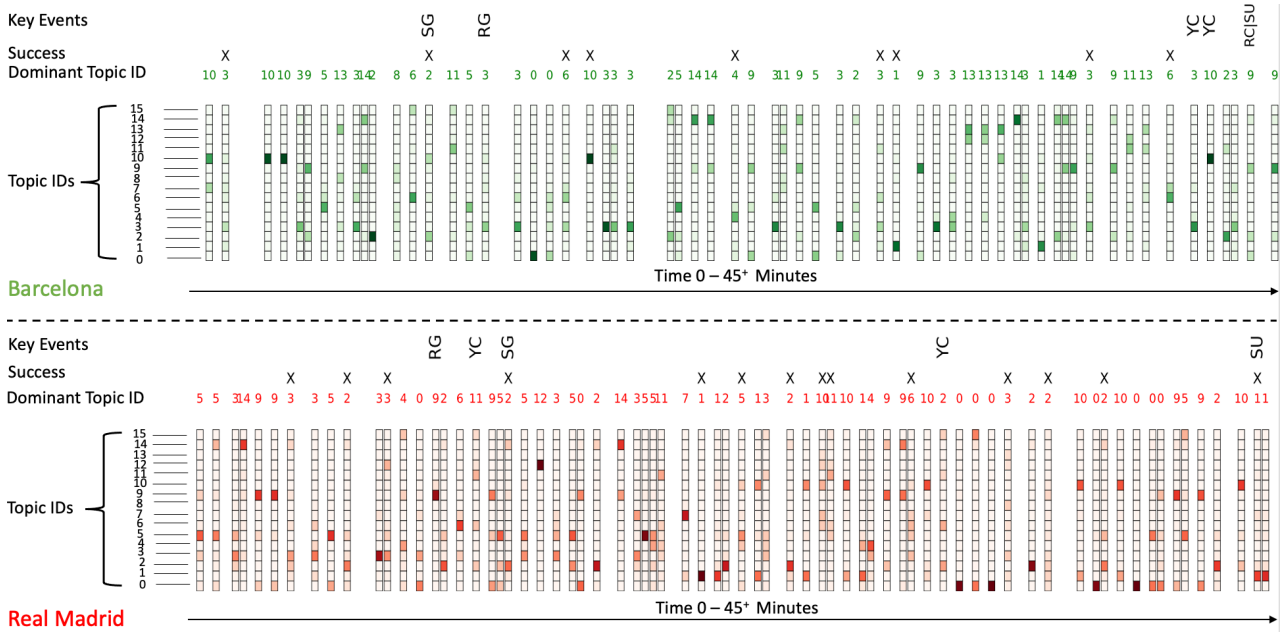


Figure 4: *Analyzing football data using topics. This figure presents data from the first half of the Barcelona vs. Real Madrid match, which ended in a 2-2 draw in May 2018. The x-axis represents time, starting at 0 and extending beyond 45 minutes to cover the entire first half. The development of topics and associated events is illustrated over time, with Barcelona's information displayed above the dashed divider line and Real Madrid's information below it, creating two sub-figures. In each sub-figure, stacked bar charts depict active topics during episodes. These charts are placed on a timeline to show the temporal order of the episodes. Each small bar in the stacked bar charts has a constant width, regardless of the episode's actual length to prevent covering. There are 16 small bars within each stack, representing the 16 topic IDs, labeled from 0 to 15 on the left-hand side. Each bar's color intensity indicates the topic's weight, with green for Barcelona and red for Real Madrid. To facilitate better analysis, the topic ID with the most weight (Dominant topic ID) is displayed at the top of each stack. Additionally, the Success index and key match events are indicated using an "X" and abbreviations, respectively. This helps analysts understand the evolution of topics in relation to these crucial events and successful episodes.*

dominant topic, since those two topics are usually the final active topic of a successful episode. Analyzing the other active topics that lead to successful episodes could give insights into how a team performs its attacks. In addition, the dynamics of topic weights along the timeline show that some topics are consistently appearing with high weights, while others occur infrequently. We can see pairs and even larger groups of topics that are prominently visible together within multiple episodes. For example, topics 9 and 14 are activated together for both teams. Based on the topics' signatures (from Figure 3), this means that they made passes between their goal/penalty area and the right side of their defensive half.

7. Conclusions and Future Work

In this study, we explore the interplay of visual analytics and topic modeling to understand game dynamics in the football domain. One of the challenges in using topic modeling is selecting the most suitable number of topics such that they are spatially consistent and easily distinguishable. To address this, we propose a visual analytics technique. We apply this technique to a complex and relatively unexplored football dataset. The obtained topics were interpreted through spatial heatmaps and then used to visualize game dynamics. Our study demonstrates that incorporating visual analytics at various stages of topic modeling enhances its effectiveness and interpretability. Additionally, the results of topic modeling can be leveraged in visual analytics to increase its utility.

In the future, the obtained topics for the football data could be used to answer several questions, including the impact of active topics on game outcomes and the variation in topics under different conditions, such as facing stronger or weaker opponents and playing at home or away venues. Additionally, these topics can also be analyzed across multiple games of the same team to identify patterns and trends. Visualizing the average of topics for each team within a league can provide insights into their behavior throughout the season. Besides, using different vocabulary of terms for the same dataset (e.g., terms that capture the roles of involved players and/or order of events) could yield different topic types, allowing us to explore different aspects of the game, such as player interactions. We believe our approach will enable a deeper understanding of team tactics and facilitate comparisons across different teams and leagues in various contexts. Moreover, by using different spatiotemporal event sequences, we can explore how visual analytics and topic modeling can complement each other in analyzing various domains. A promising future research direction could involve exploring the practical applications of the visualizations and the insights they yield. In addition, future work could conduct a user study with domain experts to evaluate the effectiveness of the approach.

References

[AAA*19] ANDRIENKO G., ANDRIENKO N., ANZER G., BAUER P., BUDZIAK G., FUCHS G., HECKER D., WEBER H., WROBEL S.: Constructing spaces and times for tactical analysis in football. *IEEE Transactions on Visualization and Computer Graphics 27*, 4 (2019), 2280–2297. doi:10.1109/TVCG.2019.2952129. 3

[AAH23] ANDRIENKO G., ANDRIENKO N., HECKER D.: Extracting movement-based topics for analysis of space use. In *EuroVA: International Workshop on Visual Analytics* (2023), vol. 2023, The Eurographics Association. doi:10.2312/eurova.20231091. 1, 2, 4

[AAS23] ANDRIENKO N., ANDRIENKO G., SHIRATO G.: Episodes and topics in multivariate temporal data. *Computer Graphics Forum 42*, 6 (2023), e14926. doi:10.1111/cgf.14926. 1, 2, 3

[BNJ03] BLEI D. M., NG A. Y., JORDAN M. I.: Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, Jan (2003), 993–1022. 2

[CAA*20] CHEN S., ANDRIENKO N., ANDRIENKO G., ADILOVA L., BARLET J., KINDERMANN J., NGUYEN P. H., THONNARD O., TURKAY C.: Lda ensembles for interactive exploration and categorization of behaviors. *IEEE Transactions on Visualization and Computer Graphics 26*, 9 (2020), 2775–2792. doi:10.1109/TVCG.2019.2904069. 1

[EASD*18] EL-ASSADY M., SPERRLE F., DEUSSEN O., KEIM D., COLLINS C.: Visual analytics for topic model optimization based on user-steerable speculative execution. *IEEE transactions on visualization and computer graphics 25*, 1 (2018), 374–384. 2

[HHQ12] HOLMES I., HARRIS K., QUINCE C.: Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PloS one 7*, 2 (2012), e30126. doi:10.1371/journal.pone.0030126. 1, 2

[Kru64] KRUSKAL J. B.: Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. *Psychometrika 29*, 1 (Mar. 1964), 1–27. doi:10.1007/BF02289565. 4

[LNC*17] LUO M., NIE F., CHANG X., YANG Y., HAUPTMANN A., ZHENG Q.: Probabilistic non-negative matrix factorization and its robust extensions for topic modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2017), vol. 31. 2

[MHM18] MCINNES L., HEALY J., MELVILLE J.: Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018). doi:10.48550/arXiv.1802.03426. 4

[PCF*19] PAPPALARDO L., CINTIA P., FERRAGINA P., MASSUCCO E., PEDRESCHI D., GIANNOTTI F.: Playerank: data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM Transactions on Intelligent Systems and Technology (TIST) 10*, 5 (2019), 1–27. doi:10.1145/3343172. 3

[PCR*19] PAPPALARDO L., CINTIA P., ROSSI A., MASSUCCO E., FERRAGINA P., PEDRESCHI D., GIANNOTTI F.: A public data set of spatio-temporal match events in soccer competitions. *Scientific data 6*, 1 (2019), 236. doi:10.6084/m9.figshare.9711164. 3

[PM19] PAPPALARDO L., MASSUCCO E.: Soccer match event dataset. figshare. collection, 2019. 1, 2, 3

[VdMH08] VAN DER MAATEN L., HINTON G.: Visualizing data using t-sne. *Journal of machine learning research 9*, 11 (2008). 4

[VK20] VAYANSKY I., KUMAR S. A.: A review of topic modeling methods. *Information Systems 94* (2020), 101582. doi:10.1016/j.is.2020.101582. 1, 2

[WG07] WANG X., GRIMSON E.: Spatial latent dirichlet allocation. *Advances in neural information processing systems 20* (2007). 1, 2

[WZH*15] WANG Q., ZHU H., HU W., SHEN Z., YAO Y.: Discerning tactical patterns for professional soccer teams: an enhanced topic model with applications. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015), pp. 2197–2206. doi:10.1145/2783258.2788577. 3