What are the topics in football? Extracting time-series topics from game episodes

Gota Shirato*
University of Bonn
Fraunhofer Institute IAIS

Natalia Andrienko
Fraunhofer Institute IAIS
City, University of London

Gennady Andrienko
Fraunhofer Institute IAIS
City, University of London

ABSTRACT

We propose an approach to analysis of multivariate time series describing multiple episodes in the development of a dynamic phenomenon or process. We represent variation patterns of different features by symbolic codes and reveal pattern co-occurrences using topic modelling. We apply the approach to episodes of a football game characterised in terms of a novel feature *gate width* reflecting space availability to attackers and space control by defenders.

1 INTRODUCTION

Data describing behaviours of dynamic objects or phenomena usually have the form of multivariate time series, i.e., values of multiple attributes (features) measured or observed at different time moments. It is challenging to synthesize holistic understanding of the behaviours from these elementary data items. Analysts need methods supporting pattern discovery, which means integration of multiple items into constructs that can be considered as units. The integration becomes possible due to relationships existing between data elements, such as ordering and/or distance [3]. In a time series of values of a single numeric feature, the relationships between times and between corresponding feature values create temporal variation patterns: increasing and decreasing trends, peaks, troughs, fluctuations, stability. Such patterns are easy to detect and, very importantly, easy to interpret. However, it is not obvious how to construct unified patterns from time series of several features. In a general case, there are no intrinsic relationships between values of different features that would suggest an intuitive way of integration across time series. What can be exploited is the relationship of temporal co-occurrence of patterns from different time series.

Our idea is to use topic modelling as a tool capable to integrate multiple items (words) into units (topics) solely based on their co-occurrence (in documents, paragraphs, or sentences). To make it applicable to multivariate time series, we transform temporal variation patterns of each time series into sequences of symbols '-', '+', and 'e' that can be treated as "words". A combination of symbolic codes generated for all variables is treated as a "text". Multiple "texts" are obtained by dividing long time series into shorter parts or by extracting time intervals (episodes) of interest.

We have tested this idea by applying it to multivariate time series describing collective movements of players in a football game. It allowed us to discover groups of episodes with similar collective behaviours of the teams. Importantly, the symbolic representation of the feature dynamics was supportive for semantic interpretation of the specifics of the behaviours in each group of episodes.

2 RELATED WORK

A broad survey of approaches to visualization of time-oriented data [1] mentions only a few works dealing with multivariate time series, among them the one by Guo et al. [9] whose approach in different variations re-occurred in several later works. It consists of applying clustering and/or projection to the combinations of values of multiple variables corresponding to the time steps and transforming the time series into sequences of cluster memberships [9], state transition graphs [10], or paths in the projection space [4]. In all these approaches, data are analyzed as sequences of states disregarding dynamics of individual variables.

Probabilistic topic modelling methods, which were originally developed for texts, have recently been applied to other data types, such as movement data and action sequences [5, 6], but we are not aware of attempts to apply such methods to multivariate time series.

3 GENERAL APPROACH

Our approach is intended for data representing temporal variation of values of multiple numeric features on multiple time intervals, called *episodes*. To transform a time series of numeric values into a symbolic code representing the pattern of value variation, we compare the values to the median of the time series and encode values lower than the median by the symbol '-', greater than the median by '+', and approximately equal to the median by 'e'. This transformation can be applied to each individual value when the time series has a small number of time steps; otherwise, the time span is divided into several intervals, and the encoding is applied to the average values on the intervals. A value is treated as "approximately equal" to the median m when it lies within an allowance interval $[m-d_1, m+d_2]$. The values d_1 and d_2 can be chosen based on the overall statistics of the deviations from the median for this feature. Please note that the median m is determined *individually for each episode* whereas the values d_1 and d_2 are *common for all episodes*. This way of encoding produces codes that either solely consist of the symbol 'e' thus indicating a stability pattern or include both '-' and '+' symbols while 'e' may be present or absent. The order of the symbols can represent various patterns of value change; e.g., "-+" means rapid increase whereas "-e+" means more gradual increase.

Now, each episode can be represented by the combination of the variation codes of the multiple variables. A method for probabilistic topic modeling, such as Latent Dirichlet Allocation (LDA), is applied to these combinations, which are treated as "texts" where the variation codes are "words". The resulting "topics" show which variation patterns of different variables tend to occur together in the same episodes. The topic modelling method also assigns vectors of topic probabilities to the episodes. Using these vectors, the episodes can be clustered and/or arranged in a projection space according to similarities of the variation patterns and further explored as it is done in the previous state-oriented approaches mentioned in Section 2.

4 APPLICATION TO FOOTBALL

Creating and closing space is a recurrent subject in football [7]. A defending team tries to limit space on the pitch while the other team seeks available space. Existing approaches to analysing the distribution of the pitch space, e.g., based on a Voronoi diagram [8], do not give special attention to spaces between players, whereas these spaces greatly affect the possibilities to move or pass the

*e-mail: gota.shirato@iais.fraunhofer.de

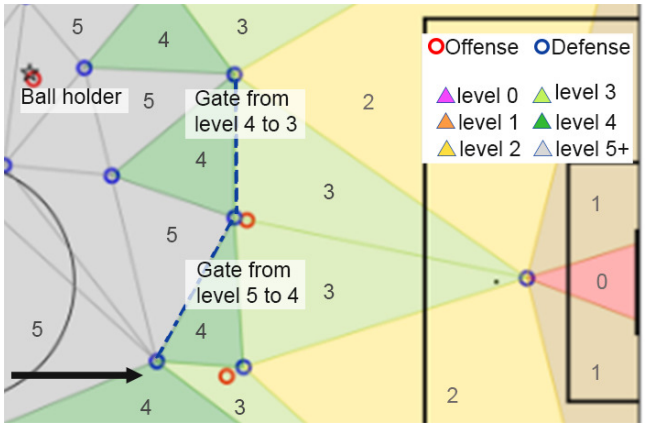

Figure 1: A *gate* is a triangle edge through which the separation from the goal decreases. The separation is expressed through the *levels* of the positions on the pitch. The level is the number of triangles one has to cross before entering the triangle containing the goal.

Table 1: Sequences with the highest weights for each level (row) in each topic (column) for the home team. One cell may contain multiple sequences if they have the same weight (e.g. level 1 in topic 4).

	level 0	level 1	level 2	...	level 6	level 7
...						
topic 4	`--++`	`--++` `ee++`	`--++`	...	`++--`	`eeee`
...						
6	`+e-e`	`+e--`	`++--`	...	`--++`	`-+++`
...						

ball for attackers and to prevent or obstruct attackers' activities for defenders. We introduce a novel feature *gate width* to analyse inter-player spaces and their dynamics.

Figure 1 explains the concept of a gate. Gates are defined by applying the Delaunay triangulation to the positions of the defenders at a given time moment. In this study, we investigate how gate widths at different levels change during multiple game episodes. For this purpose, we describe each episode by a multivariate time series of gate widths at different levels. We demonstrate our approach on positional data from a professional football game. The data contain positions of players and the ball recorded at 25Hz frequency. We focus on the transition episodes around changes of ball possession [2]. We extracted 85 such episodes of the duration $10 (= 5 + 5)$ seconds consisting of $10 \times 25 = 250$ time steps. For each time step of an episode, the width of each gate was encoded by the symbol '-', 'e', or '+' as described in Section 3. To facilitate the interpretation, we compressed the time series by dividing the length into four equal intervals, two before the transition and two after it. For each interval and each gate level, the most frequent symbol was taken. In this way, for each episode, we obtained a set of four-symbol codes corresponding to different levels. We applied LDA to these sets of codes trying different numbers of topics n and evaluating the results by projecting the vectors of topic probabilities assigned to the episodes by LDA onto a 2D plane using t-SNE (Fig. 2). We found that $n = 8$ works the best in terms of grouping similar episodes.

For example, for transitions of a team from defence to attack, topic 4 (see Table 1) has increasing trends ("`--++`" and "`ee++`") near the goal (levels 0 to 2) and a decreasing trend ("`++--`") far from the goal (level 6). The distributions of centroids are wider compared to topic 6 (Figure 3). These indicate that the team retrieves the ball, reducing the attacking opponents' spaces, and subsequently lets spaces between forwards and defenders larger in counter-attacks.

We expect our approach to characterising dynamics of multivariate time series using topic modelling to be applicable to a variety of

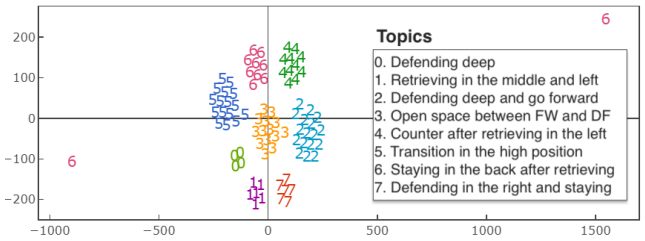

Figure 2: Episodes from a football match are represented by dots in a 2D projection space obtained by applying t-SNE to the vectors of the topic weights. The dot colours represent the dominant topics.

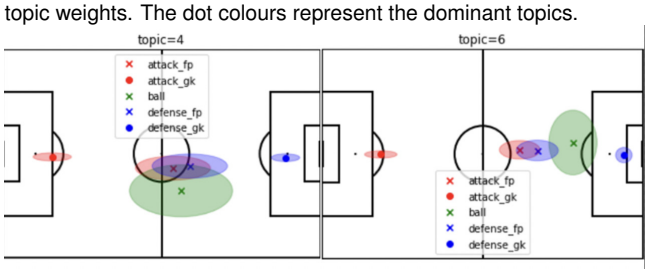

Figure 3: Distributions of centroids after transitions for topic 4 and topic 6. Colors represent the objects (blue: defense, red: attack, green: ball). The cross (x) means the average position of centroids and the ellipse means the standard deviation of centroids for each topic.

events and processes in football and beyond.

References

[1] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of Time-Oriented Data*. Springer Verlag, London, UK, 1st ed., 2011. doi: 10.1007/978-0-85729-079-3

[2] G. Andrienko, N. Andrienko, G. Anzer, P. Bauer, G. Budziak, G. Fuchs, D. Hecker, H. Weber, and S. Wrobel. Constructing spaces and times for tactical analysis in football. *IEEE Transactions on Visualization and Computer Graphics*, 27(4):2280–2297, 2021. doi: 10.1109/TVCG.2019.2952129

[3] N. Andrienko, G. Andrienko, S. Miksch, H. Schumann, and S. Wrobel. A theoretical model for pattern discovery in visual analytics. *Visual Informatics*, 5(1):23 – 42, 2021. doi: 10.1016/j.visinf.2020.12.002

[4] B. Bach, C. Shi, N. Heulot, T. Madhyastha, T. Grabowski, and P. Dragicevic. Time curves: Folding time to visualize patterns of temporal evolution in data. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):559–568, 2016. doi: 10.1109/TVCG.2015.2467851

[5] S. Chen, N. V. Andrienko, G. L. Andrienko, L. Adilova, J. Barlet, J. Kindermann, P. H. Nguyen, O. Thonnard, and C. Turkay. LDA ensembles for interactive exploration and categorization of behaviors. *IEEE Transactions on Visualization and Computer Graphics*, 26(9):2775–2792, 2020. doi: 10.1109/TVCG.2019.2904069

[6] D. Chu, D. A. Sheets, Y. Zhao, Y. Wu, J. Yang, M. Zheng, and G. Chen. Visualizing hidden themes of taxi movement with semantic transformation. In *2014 IEEE Pacific Visualization Symposium*, pp. 137–144, March 2014. doi: 10.1109/PacificVis.2014.50

[7] J. Fernandez and L. Bornn. Wide open spaces : A statistical technique for measuring space creation in professional soccer. *MIT Sloan Sports Analytics Conference*, pp. 1–19, 2018.

[8] J. Gudmundsson and M. Horton. Spatio-temporal analysis of team sports. *ACM Computing Surveys*, 50(2), 2017. doi: 10.1145/3054132

[9] D. Guo, J. Chen, A. MacEachren, and K. Liao. A visualization system for space-time and multivariate patterns (VIS-STAMP). *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1461–1474, Nov. 2006. doi: 10.1109/TVCG.2006.84

[10] L. Stopar, P. Skraba, M. Grobelnik, and D. Mladenic. Streamstory: Exploring multivariate time series on multiple scales. *IEEE Transactions on Visualization and Computer Graphics*, 25(4):1788–1802, 2019. doi: 10.1109/TVCG.2018.2825424