Following Soccer Fans from Geotagged Tweets at FIFA World Cup 2014

Eugenio Cesario[#1], Chiara Congedo[*2], Fabrizio Marozzo[¶3], Gianni Riotta[§4],

Alessandra Spada[*5], Domenico Talia[¶6], Paolo Trunfio[¶7], Carlo Turri[*8]

[#]*ICAR-CNR & DtoK Lab Srl*
Rende (CS), Italy
[1]cesario@icar.cnr.it

[*]*TSC Consulting Srl*
Cagliari, Italy
[2]c.congedo@tsc-consulting.com, [5]a.spada@tsc-consulting.com, [8]carlo_turri@hotmail.com

[¶]*DIMES-University of Calabria & DtoK Lab Srl*
Rende (CS), Italy
[3]fmarozzo@dimes.unical.it, [6]talia@dimes.unical.it, [7]trunfio@dimes.unical.it

[§]*Princeton University*
Princeton, NJ, USA
[4]riotta.g@gmail.com

***Abstract*— The world-wide size of social networks, such as Facebook and Twitter, is making possible to analyse the real-time behaviour of large groups of people, such those attending popular events. This paper presents work and results on the analysis of geotagged tweets carried out to understand the behaviour of people attending the 2014 FIFA World Cup. We monitored the Twitter users attending the World Cup matches to discover the most frequent movements of fans during the competition. The data source is represented by all geotagged tweets collected during the 64 matches of the World Cup from June 12 to July 13, 2014. For each match we considered only the geotagged tweets whose coordinates fallen within the area of stadiums, during the matches. Then, we carried out a trajectory pattern mining analysis on the set of the tweets considered. Original results were obtained in terms of number of matches attended by groups of fans, clusters of most attended matches, and most frequented stadiums.**

***Keywords*— Social network analysis, Twitter, Geographical data mining, Trajectory pattern mining, FIFA World Cup.**

I. INTRODUCTION

Understanding how a large community or a population behaves in a major event is a primary question that in the past was extremely difficult to approach on a large scale [1,2]. With the growth of social networks such as Facebook, Instagram, Foursquare and Twitter, it became possible to analyse the real-time behaviour of large groups of people attending popular events [3,4,5]. In fact, one of the leading trends in information processing is the use of social media location-based services. Using geo-localized services of social media like Twitter and Facebook it is possible to link geographic locations to behavioural patterns. Geotagged tweets can be collected today like modern Hansel's pebbles to record movements of people. In particular, we can extract information and patterns from Twitter data to describe mobility behaviours that are useful in applications such as mobility planning and travel route discovery [6,7]. This might reveal underlying economic, social, and political trends and patterns of the complex relations between space and time on both virtual and real life.

In this paper, we present our experience and achieved results based on the use of a methodology designed for the collection and analysis of geotagged tweets. The main goal of this work was to monitor the attendance of Twitter users during the FIFA World Cup 2014 matches to discover the most frequent movements of fans during the competition. The data source is represented by all geotagged tweets collected during the 64 matches of the World Cup from June 12 to July 13, 2014. For each match we considered only the geotagged tweets whose coordinates fallen within the area of stadiums, during the matches. On the geotagged tweets coming from people inside the stadiums, we carried out a trajectory pattern mining analysis. The analysis is based on the search of frequent item sets that allow identifying the groups of matches attended most frequently by spectators. Original results were obtained in terms of number of matches attended by groups of fans, groups of most attended matches, and most frequented stadiums. The number of tweets posted from inside the stadiums during the soccer matches was pretty high, more than half a million, allowing their analysis with well-know data mining techniques, such as the Apriori algorithm, for frequent itemsets computing and association rules discovery.

The methodology adopted to carry out the data analysis task described here, could be re-used in similar scenarios

where groups of people attend social events to understand collective behaviours that are very hard to discover with traditional social analysis techniques.

The rest of the paper is organized as follows. Section II describes the methodology used in the analysis of geotagged tweets. Section III presents a summary of the results obtained. Finally, Section IV includes some comments to the results and methodology.

II. Methodology

Let $TW=\{tw_1, ... ,tw_N\}$ be a set of geotagged tweets, where each tweet tw_i is described by the following properties: *user* who posted tw_i, *latitude* and *longitude* (coordinates of the place from where tw_i was sent), *source* (device or application used to generate tw_i), *date* and *text*. Let $S=\{s_1,\ldots,s_{12}\}$ be the set of stadiums in which the World Cup matches have been played, where for each stadium s_i are known the four corner coordinates of the rectangle containing it. Finally, let $M=\{m_1, \ldots ,m_{64}\}$ be the 64 matches of the World Cup, where each match m_i is described by the following properties: *stadium*, *date*, $team_1$ and $team_2$ (the two teams playing the match).

The whole process can be described as composed of four main steps: *data acquisition*, *data pre-processing*, *data mining* and *results visualization.*

Data acquisition has been carried out by collecting all the geotagged tweets sent by fans during the World Cup matches, considering only those whose coordinates fallen within the area of stadiums, during the matches. About 526,000 tweets have been collected from June 12 to July 13, 2014.

Pre-processing has been performed to clean, select and transform data to make it suitable for analysis. First, we cleaned collected data by removing all the tweets with unreliable position (e.g., tweets with coordinates manually set by users or applications). Then, we selected only tweets written by users attending the matches, by removing *re-tweets* and *favorites* posted by other users. Finally, we transformed data by keeping one tweet per user per match, because we were interested to know only if a user attended a match or not. The final dataset D contains about 10,000 transactions, each one containing the list of matches attended by a single Twitter user. Formally,

$$D=\{T_1,T_2,\ldots,T_n\}$$

where $T_i=<u_i,\{m_{i1},m_{i2},\ldots,m_{ik}\}>$ and $m_{i1},m_{i2},\ldots,m_{ik}$ are the matches attended by a user u_i.

A data mining task was performed to identify the most attended matches and to extract the most frequent movements of fans during the whole competition, starting from D. This has been done using trajectory pattern mining techniques. The goal of these techniques is discovering trajectory patterns modeling the cumulative behavior of a population of moving people and/or objects. A trajectory pattern is a sequence of geographic regions that, based on the source data, emerge as frequently visited in a given temporal order. The *support* of a pattern, i.e., the number of transactions containing the pattern, is a measure of its strength and reliability.

Our analysis was focused on the discovery of frequent trajectory patterns to identify the groups of matches attended most frequently by fans. In our case, a frequent pattern *fp* with support *s*,

$$fp = <m_i, m_j,.., m_k> (s)$$

is an ordered sequence of matches $m_i, m_j,.., m_k$ where s is the percentage of transactions in D containing *fp*.

Pattern extraction has been performed implementing a custom version of the Apriori algorithm [8]. We used this technique because it is among the most appropriate mining tasks to extract frequent patterns (trajectories) w.r.t. traditional techniques such as classification and clustering [9]. Our algorithm first computes the support of each match in D. Then, it iteratively generates new candidate k-match-sets (i.e., sets of matches of cardinality k) and computes their support, using the frequent $(k-1)$-match-sets found in the previous iteration. In doing so, the algorithm deletes all the candidate match-sets whose support is lower than a given minimum support. The algorithm terminates when no more frequent match-sets are generated.

Finally, results visualization was implemented by the creation of infographics aimed at presenting the results in a way that is easy to understand to the general public, without providing complex statistical details that may be hard to understand to the intended audience. The graphic project has been grounded on some of the most acknowledged and ever-working principles underpinning a 'good' info-graphic piece.

In this case, we followed three main design guidelines: i) preferring a visual representation of the quantitative information to the written one; ii) minimising the cognitive efforts necessary to decoding each system of signs; iii) structuring the whole proposed elements into graphic hierarchies.

Displaying quantitative information by visual means instead of just using numeric symbols - or at least a combination of the two approaches - has been proven extremely useful in providing a kind of sensory evidence to the inherent abstraction of numbers, because this allows everybody to instantly grasp similarities and differences among values. In fact, basic visual metaphors (e.g., the largest is the greatest, the thickest is the highest) enable more natural ways of understanding and relating sets of quantities [10].

In order to reduce the cognitive load necessary to information decoding and absorbing, several paradigms have been employed. The most relevant of them are: i) aiming at the simplicity of the visual language (by using flat and monochromatic icons, for example); ii) limiting the number of different signs to the necessary; iii) sorting and arranging colours as syntactic elements; iv) using every visual component with coherency throughout each chart [11].

A proper hierarchy of the presented material (graphic images, written text and numbers, symbols, etc.) is a crucial factor that helps the readers in identifying which are the core issues to focus on and what is auxiliary or complementary to them. Since the chart reading process can start randomly everywhere, it is important to create visual affordances capable of attracting the observers gaze and so inducing their understanding pathways to begin from the most proper points. These reading patterns are mostly achieved by adjusting and

combining visual features such as dimension, position, colour, composition [12].

Along with these functional criteria, some aesthetic choices have been made too. Since this project was part of a broader info-graphic production dedicated to Brazil 2014 as well as we were addressing a theme so variegated and entertaining, the overall graphic look was developed in order to retain some of the lightness and playfulness characterising every FIFA World Cup edition.

According to the principles introduced here, we implemented a high-level and very effective visualization model that helped the readers to easily catch the main concepts and the key meaning of the knowledge extracted by the data mining process.

III. Summary of the Results

The analysis was aimed at discovering:

1. The number of matches attended by fans during the competition;
2. The most frequent sequences of matches attended by fans, either in the same stadium or to follow a given soccer team;
3. The most frequent movement patterns obtained by grouping matches based on the phase in which they were played.

In the following, we present an extract of the results.

A. Number of Matches Attended

Table I groups fans based on the number of matches attended during the whole World Cup. The results show that 71.3% of the fans attended a single match, 16% attended two matches, 6% attended three matches, and only 3% attended four matches. It is worth noticing that 3.7% of the spectators attended five or more matches. Looking at the Twitter profiles of the spectators of the latest set, we found that many of them were journalists.

TABLE I
Number of Matches Attended.

No. of matches	Spectators
1	71.3%
2	16.0%
3	6.0%
4	3.0%
5 or more	3.7%

B. Frequent Sequences

Table II provides a general classification of the paths through the Brazil 2014 stadiums followed by fans who attended at least two matches. Focusing on fans who attended two or three matches, the table shows that:

- Among the spectators who attended two matches, 62.9% attended matches played in the same stadium, while 22.2% attended matches played by the same team.
- Among the spectators who attended three matches, 48.8% attended matches played in the same stadium, while 11.8% attended matches played by the same team.

In general, the results show that most of who attended multiple matches did it staying in the same city.

TABLE II
Classification of the Paths Followed by Fans.

No. of matches	Same stadium	Same team
2	62.9%	22.2%
3	48.8%	11.8%
4	41.0%	7.2%
5	37.0%	8.4%
6	33.7%	4.8%

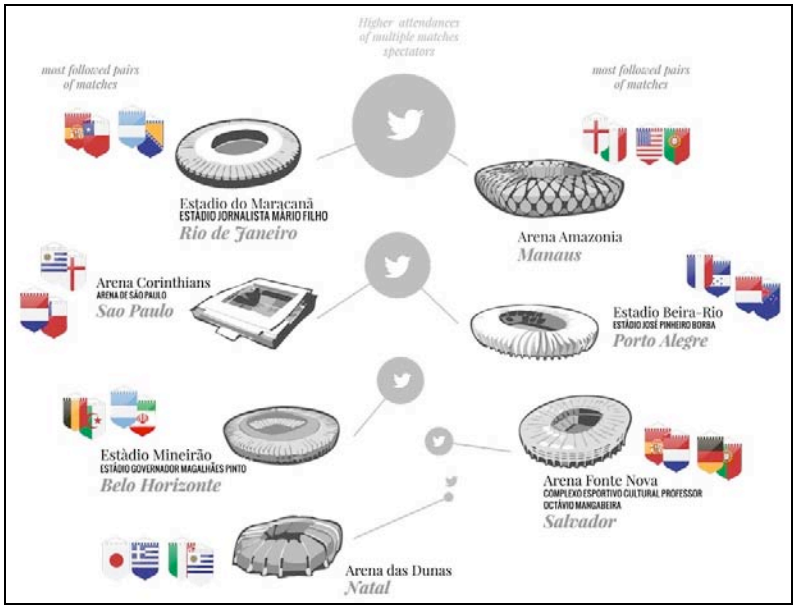

Fig. 1 Infographic illustrating the most frequent 2-match-sets observed during the group stage.

Fig. 1 shows the most frequent 2-match-sets observed during the group stage, from June 12 to June 26, 2014. The set with the highest support was the couple of matches <*England-Italy, USA-Portugal*> played in Manaus, followed by <*Argentina-Bosnia, Spain-Chile*> played in Rio de Janeiro, and by <*Uruguay-England, Netherlands-Chile*> played in São Paulo.

Fig. 2 shows the most frequent paths of fans who attended two or three matches of the same team during the group stage. The most frequent 2-match-set was <*Colombia-Greece, Colombia-Cote d'Ivoire*>, followed by <*Brazil-Mexico, Croatia-Mexico*>, and by <*Argentina-Bosnia, Argentina-Iran*>, i.e., matches likely attended by fans of Colombia, Mexico and Argentina.

The most frequent 3-match-set was <*Mexico-Cameroon, Brazil-Mexico, Croatia-Mexico*>, followed by <*Brazil-Croatia, Brazil-Mexico, Cameroon-Brazil*>, and by <*Chile-Australia, Australia-Netherlands, Australia-Spain*>. In this case, spectators were likely fans of Mexico, Brazil and Australia. The figure illustrates these frequent movements as directed arcs linking the stadiums in which the matches above were played. Different line sizes are used for the arcs to represent the support of each match-set.

At the end of the group stage, we carried out a specific analysis on the Twitter users who were present at the opening match <*Brazil-Croatia*> played on June 12, 2014 in São Paulo. The results showed that, among these fans:

- 50.4% did not attend other matches after the opening one;
- 13.7%, after the opening match, moved to Rio de Janeiro to attend other matches;
- 9.5% attended other matches in the same stadium in São Paulo;
- 7.0% attended another match played by the Brazilian team, either <*Brazil-Mexico*> played in Fortaleza or <*Cameroon-Brazil*> played in Brasilia;
- 2.8% attended both the following matches played by the Brazilian team, i.e. <*Brazil-Mexico*> and <*Cameroon-Brazil*>.

The results of this analysis are illustrated in the infographic reported in Figure 3.

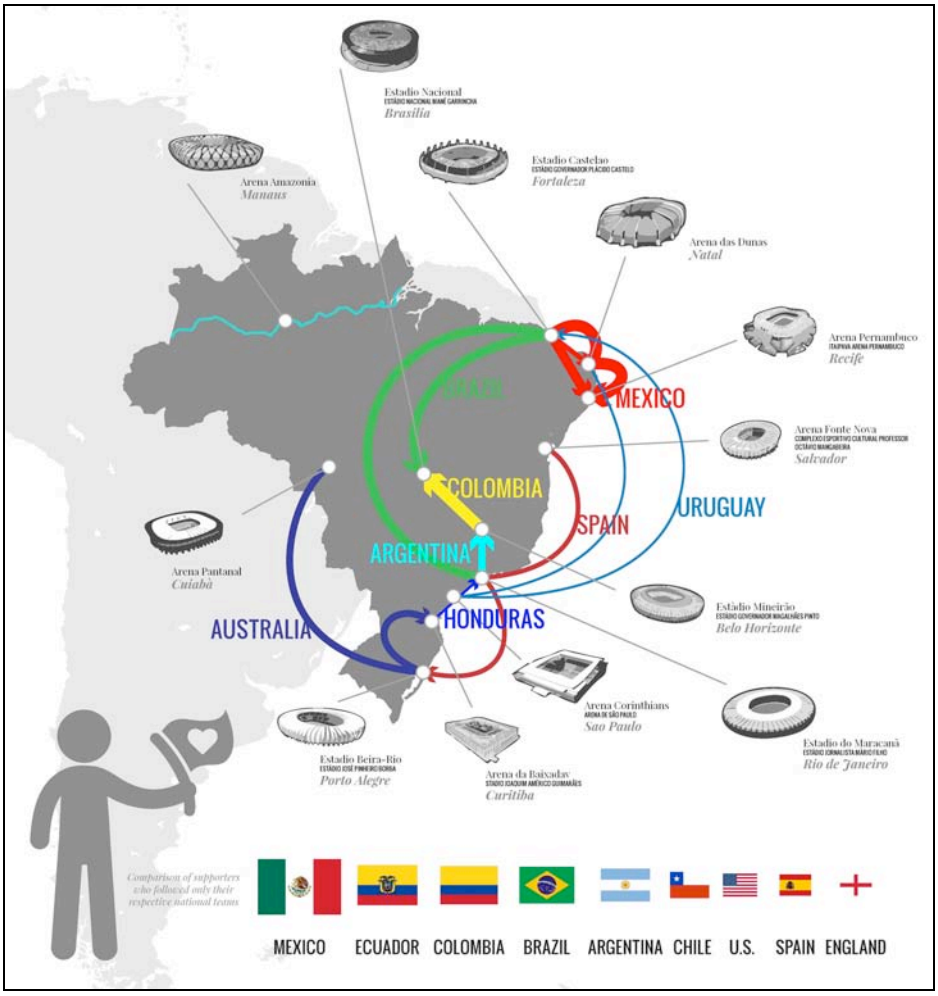

Fig. 2 Most frequent movements of fans who attended matches of the same team during the group stage.

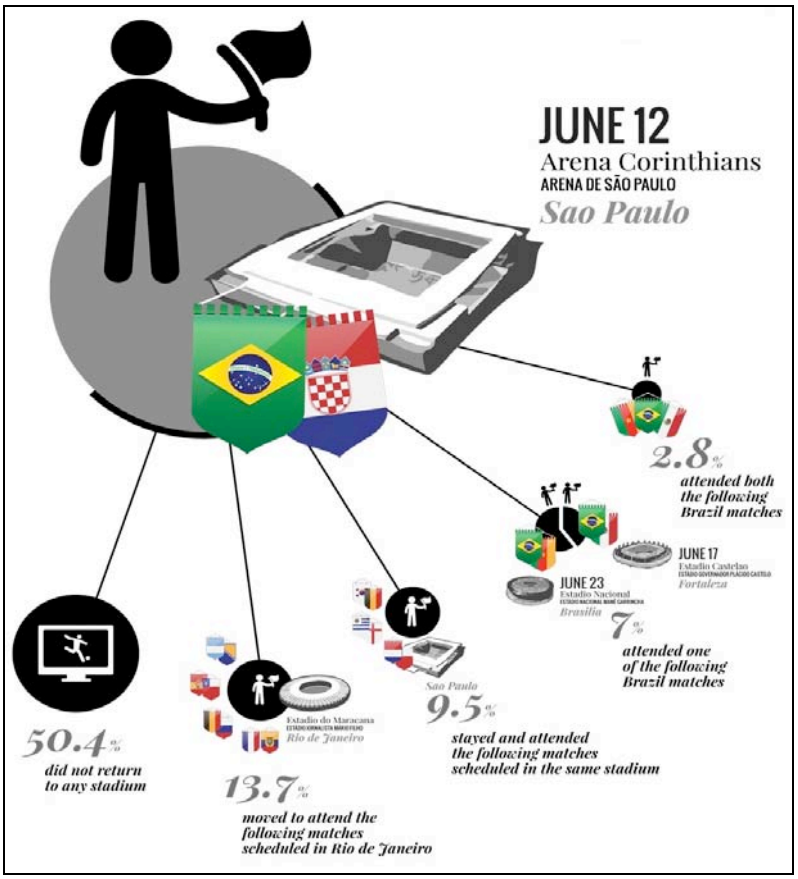

Fig. 3 Infographic showing the movements of fans who attended the opening match, during the group stage

C. Aggregate Analysis

In addition to the analysis described above, we also carried out an aggregate mobility analysis by grouping the Wold Cup matches into the following phases:

- *Opening match* (match no. 1);
- *Group stage* (matches no. 2-48);
- *Round of 16* (matches no. 49-56);
- *Quarter finals* (matches no. 57-60);
- *Semi-finals* (matches no. 61-62);
- *Final* (match no. 64).

These grouping was arranged to study the movements of fans during the different phases of the competition.

Fig. 4 shows the patterns of movements based on the grouping above, and the relative frequency (support) of these patterns.

The most frequent pattern is represented by fans who attended at least one match of the group stage and one match of the round of 16. The second most frequent pattern are fans who attended a match of the group stage and one of the quarter finals. The third most frequent pattern includes spectators of at least one match of group stage, group of 16 and quarter finals.

The relative frequency of each pattern is represented by a circle: the larger the circle, the higher the frequency. The least frequent pattern was that of fans who attended one semi-final and the final match.

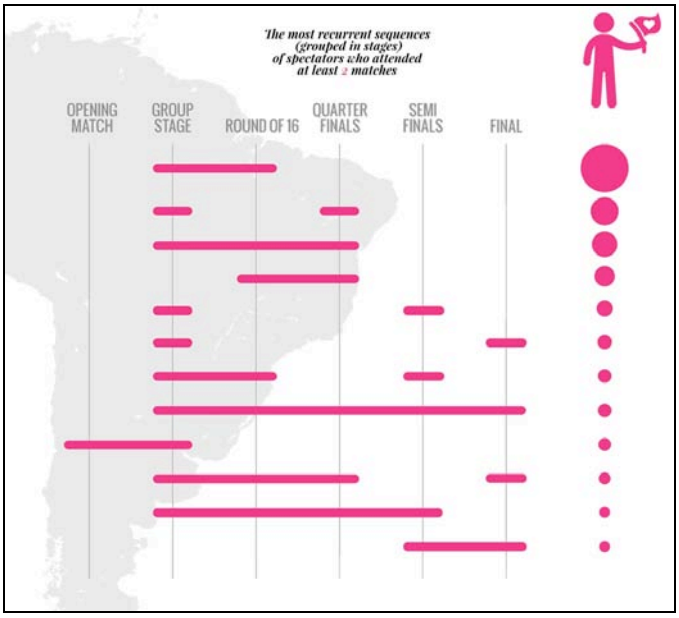

Fig. 4 Patterns of movements by grouping matches in phases.

Finally, Fig. 5 compares the flows of the semi-finalists fans, who attended at least two matches of their team. Also in this case, the figure uses different line sizes to represent the relative number of users in each flow.

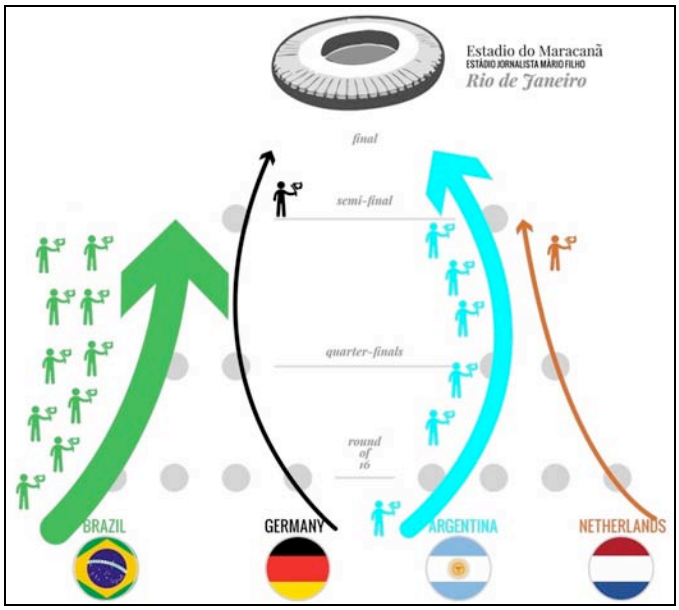

Fig. 5. Comparison among the flows of the semi-finalists fans.

IV. Comments to the Results

This research work allowed a detailed analysis of the movements of groups of people during the FIFA World Cup 2014. The aim of the study was to discover the most frequent movements of fans during the competition. It is an example of how social data analysis can help to know how people behave in big events.

The data source consists of geotagged tweets collected during the matches of the World Cup, from June 12 to July 13, 2014. For each match we collected more than half a million of geotagged tweets whose coordinates fallen within the area of stadiums, during the matches. Then, we carried out a trajectory pattern mining analysis on this set of tweets. The analysis was based on the search of frequent item sets and allowed us to identify the groups of matches attended most frequently by fans. Original results were obtained in terms of number of matches attended by groups of supporters, clusters of most attended matches, and most frequented stadiums.

With the use of new facilities, like social media location-based services, and complex data mining techniques it was possible to aggregate tweets posted by people attending a popular event and process them to understand soccer fans movements and preferences. In particular, our analysis was able to discover all the movement patterns of fans among the stadiums. From them we then extracted the number of matches attended by fans during the competition, the most frequent sequences of matches attended by fans, either in the same stadium or to follow a given soccer team, and the most frequent movement patterns obtained by grouping matches based on the phase in which they were played.

As a result of this work, we learned that the analysis of online Twitter data allows:

1. the collection of real-time event information,
2. the understanding of users behaviours, and
3. the community relationships.

The use of location-based services allowed us to identify a community of individuals that became object of mathematical study. With the described approach, a community can be partitioned into various components. Groups and individuals can be observed on their movements, actions and feelings. Therefore, it is possible to investigate the social dynamics and relationships within the community at different scales.

The use of data mining algorithms applied to advanced Twitter functionalities creates a data processing methodology that can be applied to the study of future events. Social data applications, such as the one presented in this paper, can help the organization of future events, providing full access to a wide range of information that can be decisive for the monitoring and management of key services like transports, security, logistics, and others.

Using the methodology discussed here, large communities of people can be effectively analysed to understand complex human behaviours and social dynamics. This approach is very promising, as it provides critical information and high-quality knowledge that are fundamental for the growth of organization systems. Finally, if very large data sets must be analysed, cloud-based approaches could be exploited to reduce the execution time [13].

References

[1] E. Adar, D. S. Weld, B. N. Bershad, and S. S. Gribble, “Why we search: visualizing and predicting user behavior”, in *Proc. WWW '07*, 2007, pp. 161-170.

[2] N. Mamoulis, H. Cao, G. Kollios, M. Hadjieleftheriou, Y. Tao, and D. W.Cheung, “Mining, indexing, and querying historical spatiotemporal data”, in *Proc. KDD’04*, 2004, pp. 236-245.

[3] Z. Yin, L. Cao, J. Han, J. Luo, T. S. Huang, “Diversified Trajectory Pattern Ranking in Geo-tagged Social Media”, in *Proc. SDM’11*, 2011, pp. 980-991.

[4] T. Kurashima, T. Iwata, G. Irie, and K. Fujimura, “Travel route recommendation using geotags in photo sharing sites”, in *Proc. CIKM’10*, 2010, pp. 579-588.

[5] T. Schreck and D. Keim, “Visual Analysis of Social Media Data”, *IEEE Computer*, vol. 20, no. 5, pp. 68–75, 2013.

[6] A. Ahmed, L. Hong, and A. J. Smola, “Hierarchical geographical modeling of user locations from social media posts”, in *Proc. WWW '13*, 2013, pp. 25-36.

[7] E. Cesario, C. Comito, and D. Talia, “Towards a Cloud-Based Framework for Urban Computing, The Trajectory Analysis Case”, in *Proc. CGC’13*, 2013, pp. 16-23.

[8] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules in large databases”, in *Proc. VLDB’94*, 1994, pp. 487–499.

[9] F. Giannotti, M. Nanni, F. Pinelli, and D. Pedreschi, “Trajectory Pattern Mining”, in *Proc. KDD’07*, 2007, pp. 330-339.

[10] E.R. Tufte, *The Visual Display of Quantitative Information*, Cheshire, CT: Graphics Press, 1997, pp. 161-171.

[11] J. Maeda, *Laws of Simplicity*, Cambridge MA: MIT Press, 2006, pp. 1-9.

[12] E.R. Tufte, *Envisioning Information*, Cheshire, CT: Graphics Press, 1998, pp. 53-65, 81-85.

[13] F. Marozzo, D. Talia, P. Trunfio, “Using Clouds for Scalable Knowledge Discovery Applications“, in *Proc. Euro-Par Workshops*, 2012, pp. 220-227.