Tracking Movements and Attention of Crowds in Real Time Analysing Social Streams The case of the Open Ceremony of London 2012

Marco Balduini and Emanuele Della Valle

Dip. di Elettronica e dell'Informazione – Politecnico di Milano, Milano, Italy

Abstract. To manage a big event require tracking in real time the movement of crowds. Solution based on mobile network data analysis are effective, but expensive. Obtaining comparable results by analysing public social stream has a clear commercial potential, especially considering that, being able to access also the content of a micro-post, the analysis can also track the attention of crowds.

1 Introduction

The movement of the crowds in big events can be monitored in a number of ways. The usage of mobile phone network data [1] is one of the most innovative approaches. However, this type of data is very expensive to collect and mobile telecom operators sell them at thousands of euros per hour of analysed data. Accessing the content of SMS and phone calls is, of course, forbidden. In this paper, we present an approach for tracking the movements of the crowds in big events based on the analysis of geo-tagged tweets. Moreover, being public not only the position of the tweet but also the content, we show to be able to also track the attention of the people attending the big event. As a case study we consider the open ceremony of London 2012 Olympic games.

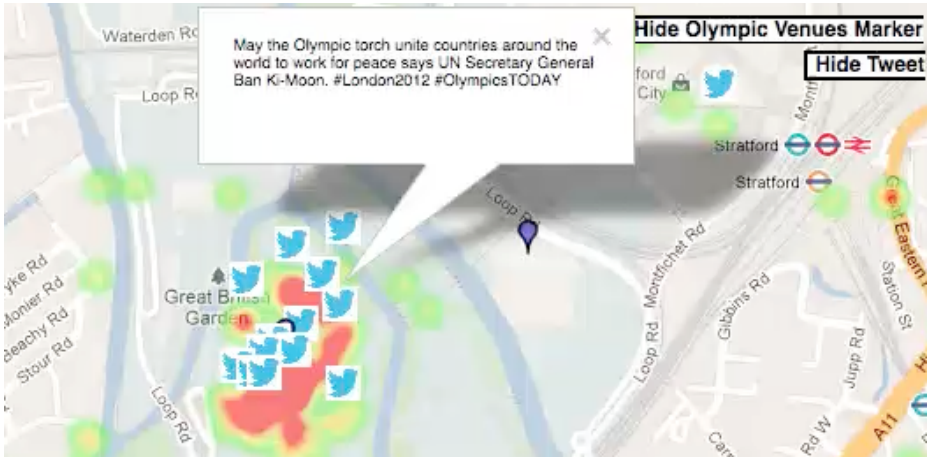

Fig. 1. A screenshot of the application.

Figure 1 shows the interface of the application[1]. It simulates what a user could have observed the day of the opening ceremony. For the convenience of the readers the application time is accelerated so to loop through the tweets posted during the day of the opening ceremony in a little more than half an hour. By clicking on the Twitter icons, users can access the textual content of the tweets. The text appears in a pop-up. The application randomly choose to open a pop-up if not solicited by the user.

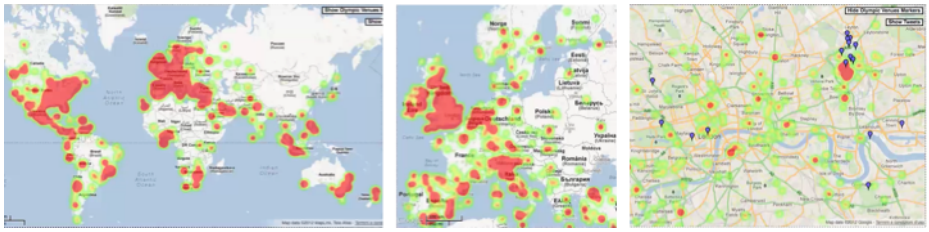

Fig. 2. A screenshot of the application at world, europe and London scale. It allows for tracking the attention the opening ceremony was able to attract all over the world.

An heatmap is used to highlight the presence of a crowd using geotagged tweets as a proxy for micro-bloggers' position. When displaying the map at world, continental and national scale (see Figure 2 and the movies available on the demo website[2]), the heatmap allows for tracking the attention the opening ceremony attracted all over the world.

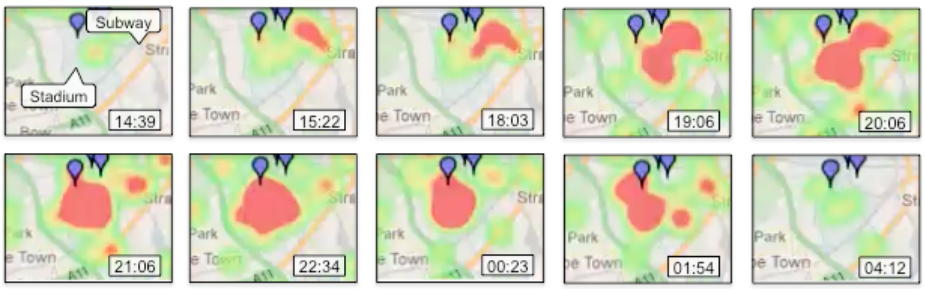

Fig. 3. The sequence of screen-shots show a crowd exiting at Stratford subway and light rail station, funnelling through Stratford walk, entering the stadium, assisting to the open ceremony and leaving the stadium to go back to Stratford.

When zooming to a city district scale, and in particular on the Olympic stadium area, the application also allows for tracking the movements of the crowds. Figure 3 displays a sequence of screenshots taken between 14:39[3] of the day of the opening ceremony and 4:12 of the day after. At 14:39 almost nobody

[1] http://www.streamreasoning.com/demos/sld/london2012/londonolympicevents/London2012Events.html

[2] Interested readers can to view the screencasts of the application at different scale that are available at `www.streamreasoning.org/demos/london2012`. We accelerated them so to observe the day of the opening ceremony in less than 5 minutes.

[3] All times are given in British summer time (BST)

was twitting from the Olympic stadium area. At 15:22 a crowd of twitter users started twitting from Stratford subway and light rail station. The screenshots at 18:03, 19:06, and 20:06 show a continuous flow of people exiting Stratford station, funnelling through Stratford walk, entering the stadium. During the entire ceremony (between 21:00 and 00:46) the crowd only twitted from the stadium. The screenshot at 01:45 shows the presence of a big crowd in the stadium area and a smaller one on Stratford station. By late morning (see screenshot at 04:12) the stadium area was empty again.

The data used by the Web application are also made available as linked data. Interested readers are invited to explore them using the Web pages that explain how the application works "behind the scene"[4]. The machine processable version of the linked data is also available.

2 The Machinery

The application presented in Section 1 is powered by the Continuous SPARQL (C-SPARQL) Engine [2] and the Streaming Linked Data (SLD) framework [3].

C-SPARQL is an extension of SPARQL that brings to SPARQL typical data stream processing concepts [4]. It introduces the RDF stream data type, and it turns the "one-time" operational semantics of SPARQL into a "continuous" one by allowing to register queries over multiple windows opened on multiple RDF streams. The C-SPARQL Engine[5] is an interpreter of C-SPARQL queries. It offers a Java API to create and to consume RDF streams or instantaneous C-SPARQL answers of SELECT/ASK queries.

The SLD framework extends the C-SPARQL Engine with: *a*) an extendible set of adapters (so far, we have implemented adapters for twitter, fourquare, linked sensor data, pachube, and several custom data sources) that transform external (streaming) data sources in RDF streams, *b*) a set of facilities for recording RDF streams and replaying them at variable speed, *c*) an equipment to perform continuous analysis on RDF streams and static RDF graphs using the C-SPARQL Engine (when needed also using C-SPARQL query networks), and *d*) an extendible set of publishers. The standard publisher uses an improved version of the proposal for Streaming Linked Data presented in [5]; additional publishers are available for XMPP, CSV, and console.

Figure 4 visually shows the network of SLD components that underpins the application shown in Figure 1. The application operates on an RDF stream of 35812 geo-tagged tweets that was recorded from Twitter streaming APIs[6] between the 11:58 of the opening ceremony and the 9:21 of the day after. The continuous query named Geotagged extracts the position of all geotagged tweets using a tumbling window of 10 minutes. The results are published as Streaming Linked Data for 30 minutes and renewed every 10 minutes. The heatmaps of the application in Figure 1 are generated using those linked data. The tweets for the pop-ups are extracted by monitoring 7 areas around the stadium. A first query, named Olympic Area selects the geotagged tweets whose position is within a bounding box that covers the Olympic area. Seven downstream queries

[4] http://www.streamreasoning.com/demos/sld/london2012/londonolympicevents/behind.html

[5] Interest readers may want to download the C-SPARQL Engine from http://streamreasoning.org/download

[6] https://dev.twitter.com/docs/streaming-apis

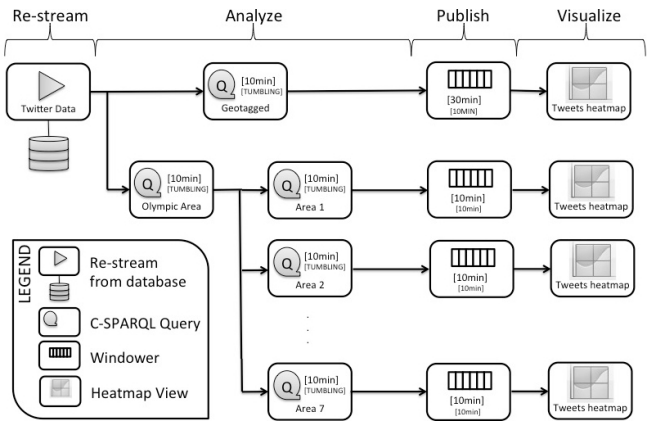

Fig. 4. The network of SLD components underpinning the application in Figure 1.

further select the tweets around Stratford subway station, Stratford International station, Stratford walk, the Olympic stadium, Hackney Wick subway station, Abbey Road light rail station, and Pudding Mill Lane light rail station.

3 Evaluation

The effectiveness of the application in tracking the movements of the crowds is well illustrated by Figure 3: the sequence of screen-shots clearly shows the crowd exiting at Stratford station, funnelling through Stratford walk, entering the stadium, assisting to the opening ceremony and leaving the stadium to go back to Stratford.

In order to prove that our approach is effectively able to also track the attention of the people attending or watching at TV the opening ceremony, we also performed a simple peak analysis of the hashtags that appears in the tweets posted during the open ceremony. We used the C-SPARQL query in Listing 1.1.

```
1  REGISTER STREAM HashtagAnalysis AS
2  PREFIX sioc:<http://rdfs.org/sioc/ns#>
3  CONSTRUCT { ?tag sma:number ?n ;  sma:maxTS ?maxTS . }
4  FROM STREAM <http://streamreasoning.org/LOG2012> [RANGE 2h STEP 2h]
5  WHERE { {
6      SELECT ?tag (COUNT(?tweet) AS ?n) (MAX(?ts) AS ?maxTS)
7      WHERE { ?tweet sioc:topic ?tag ; sioc:created_at ?ts .  }
8      GROUP BY ?tag } }
```

Listing 1.1. A C-SPARQL query that counts the number of times a tag appears in the tweets twitted in a sliding window of two hours that tumbles.

The `REGISTER STREAM` clause at Line 1 asks to register the continuous construct queries that follows the `AS` clause. The queries consider a tumbling window of 2 hours (see clause `[RANGE 2h STEP 2h]` at Line 4) open on the replayed RDF stream of tweets about the open ceremony (see clause `FROM STREAM` at Line 4). The `WHERE` clause at Line 7 matches the hashtags and the creation date of each

tweet in the window. Lines 8 asks to group the matches by hashtag. Finally, Line 6 projects for each hashtag, its IRI, the number of tweets that contains it and the creation date of the most recent tweet that contains the hashtag.

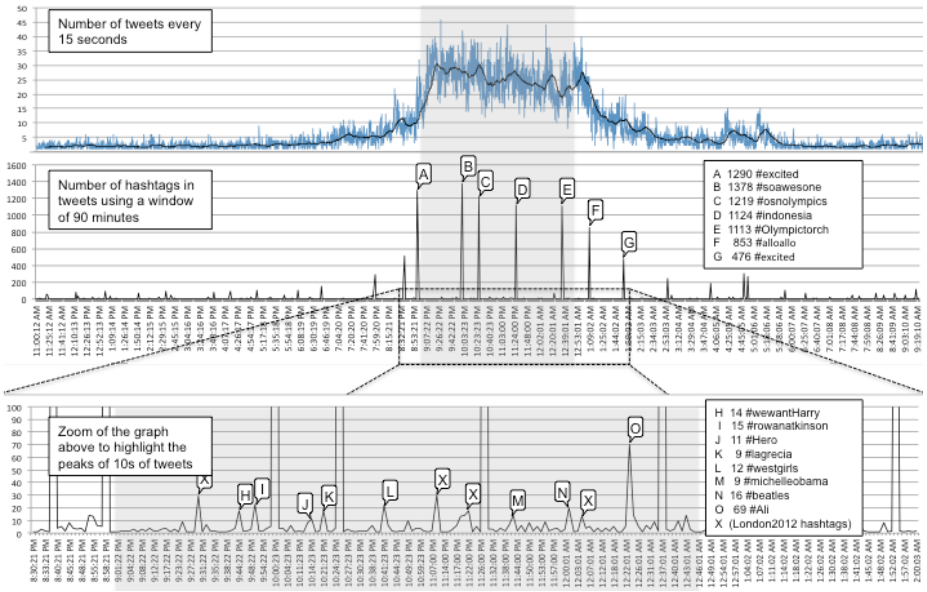

Fig. 5. Results of the analysis of the hashtags that appear in the geo-tagged tweets about London 2012 open ceremony.

Figure 5 shows the results of the analysis of the hashtags that appear in the geo-tagged tweets about the opening ceremony. The area in grey identifies the opening ceremony. The upper graph shows the number of tweets recorded every 15 seconds during the day of the ceremony. The middle and lower graphs plot the results of the query in Listing 1.1. The peaks clearly delimit the open ceremony with a peak right before the ceremony start (see the peak marked with A), and a peak when the crowd left the stadium saying good bye (#alloallo) (see F). Moreover, the peaks identify key moments in the ceremony: B corresponded to the sequence that celebrated British popular culture, H appeared after the sequence that celebrated British children's literature read by J. K. Rowling where Harry Potter was missing, I corresponded to the appearance of Rowan Atkinson in character as Mr Bean, J coincided with the tribute to the victims of the "7/7" 2005 London bombings (on the day after London had been awarded the Games), K matched the entrance of the Greek team, L appeared when the song accompanying the parade was "West End Girls" by Pet Shop Boys, M coincided with the TV framing Michelle Obama, N matched the moment when, after the Parade, the Arctic Monkeys performed The Beatles' "Come Together", O concurred with the moment when the flag paused in front of Muhammad Ali who had lit the flame at the 1996 Atlanta Games, and F coincided with the moment when seven young athletes lit the Olympic cauldron. Only peaks C and D appear unrelated to the ceremony. Apparently, the OSN channel had some audio

problem the night of the open ceremony[7]. We found no explanation for D, but it could be another broadcast problem that interested Indonesia.

The following table, finally, shows the ability of hashtags to capture key moments of the ceremony. We considered all the hashtags produced by the query in Listing 1.1 and we manually check using the video of the ceremony on youtube[8] if the hashtags describe the sequence of the ceremony identified by the creation date of the most recent tweet[9]. Around 38% of the hashtags correctly identify a sequence of the Ceremony and around 18% of them captures the emotional state of the crowd; more than half of the hashtag are relevant in tracking the attention of crowds in big events.

Analysis of the hashtags		
Moments of the ceremony	# of hashtags	Fraction
Total	189	100%
Hashtagged with an emotion state	34	17.99%
Correctly hashtagged	72	38.10%
right on time (1 min tollerance)	50	26.46%
after the event (15 min tollerance)	13	6.88%
before the event (15 min tollerance)	9	4.76%

4 Conclusions and Future Works

The approach presented in this paper appears promising. We intend to extend our analysis to the entire stream of tweets we recorded for London 2012. We intend to show in a quantitative way the ability of the approach to track movements and attention of crowds for all the venues and the events of London 2012.

Acknowledgments

This research is supported by the Search Computing project, funded by European Research Council, under the IDEAS Advanced Grants program.

References

1. Calabrese, F., Colonna, M., Lovisolo, P., Parata, D., Ratti, C.: Real-time urban monitoring using cell phones: A case study in rome. IEEE Transactions on Intelligent Transportation Systems **12**(1) (2011) 141–151
2. Barbieri, D.F., Braga, D., Ceri, S., Della Valle, E., Grossniklaus, M.: Querying rdf streams with c-sparql. SIGMOD Record **39**(1) (2010) 20–26
3. Balduini, M., Celino, I., Dell'Aglio, D., Della Valle, E., Huang, Y., Lee, T., Kim, S.H., Tresp, V.: Bottari: an augmented reality mobile application to deliver personalized and location-based recommendations by continuous analysis of social media streams. Web Semantics: Science, Services and Agents on the World Wide Web **0**(0) (2012)

[7] Searching on Google we found a tweet saying "#osnolympics am I the only having problem with the sound level dropping and coming back?" twitted at 0:38 on 27 Jul 12, see also `https://twitter.com/muscati/statuses/228952514094051328`

[8] `http://www.youtube.com/watch?v=4As0e4de-rI`

[9] We excluded from the analysis the hashtags that refer in general to London 2012 or the Opening Ceremony

4. Babcock, B., Babu, S., Datar, M., Motwani, R., Widom, J.: Models and issues in data stream systems. In Popa, L., ed.: PODS, ACM (2002) 1–16
5. Barbieri, D.F., Della Valle, E.: A proposal for publishing data streams as linked data - a position paper. In C. Bizer et al., ed.: LDOW. Volume 628., CEUR-WS (2010)

Appendix – Addressing Evaluation Requirements

In the following, Table 1 and Table 2 summarize how SLD addresses the minimal and the additional requirements, as they are listed in the Semantic Web Challenge Criteria. We provide a qualitative rating (L = low, M = medium, H = high) and a textual explanation.

Criteria	Rating	Motivation
End-user application	H	The application is thought for the general public and can be embedded in Web sites and Apps about big events like London 2012 open cerimony.
Information sources		
- diverse ownership or control	M	Micro-posts are published through Twitter by hundreds of thousands of different users spread world wide.
- heterogeneous	H	Micro-posts are syntactically homogeneous, but structurally and semantically heterogeneous. The best data structure to represent a micro-posts is a graph, because it allows for representing the variable number of hashtags and links. The semantic of micro-posts is not explicit.
- real-world data	H	The application is based on 35812 geo-tagged micro-posts collected between 27 and 28 July 2012. They are part of a larger collection of around three million tweets recorded between 25 July and 12 August 2012.
Meaning of data		
- Semantic Web technologies	H	All micro-posts are represented as RDF streams using an extension of the SIOC ontology.
- data manipulation/processing	H	The application shows how geo-tagged microposts can be used to track the movements and the interests of the crowds during big events.
- alternative technologies	M/H	The real-time requirement is challenging for RDBMS due to the high rate of updates and existing stream processing solutions, which can cope with the rate in real-time, are better suited for relational streaming data elements, while tweets are better represented as graphs.

Table 1. Minimal requirements

Criteria	Rating	Motivation
attractive and functional Web interface	H	The demonstrator that displays the movements of the crowds around the Olympic Stadium was implemented in HTML 5 using Google Maps Heatmap Layer[10] and JQuery[11]. It offers a flexible and effective visualization of the evolving data. See also the movies at `www.streamreasoning.org/demos/london2012`.
scalable application	H	As readers can assess, the demonstrator loops on the day of the Opening Ceremony analysing 24 hours of microposts in 32 minutes, thus 45 times faster than in real-time
rigorous evaluation	H	The evaluation results are given in this paper; further analysis are being performed.
novelty	M/H	To the best of our knowledge nobody has ever shown to be able to track crowds movements during big events using only twitter as a data source. Tracking public mood using twitter has been shown before, but we specifically show the efficacy of extracting key moments from a big events like the open ceremony of London 2012.
beyond information retrieval	H	The application requires continuous queries of the kind shown in Figure 4.
commercial potential	H	The ability to track movements and attention of crowds for big events has a high commercial potential.
ratings or rankings	H	The application shows: *a*) how spatial proximity and variability of geo-tagged micro-posts can be used to explain movements of crowds and to select representative micro-posts from the most crowded areas; *b*) how temporal proximity and variability of micro-posts can be used to explain evolving attention of crowds and to identify the key moments of a big events.
use of multimedia	–	N.A.
dynamic data	H	Probably the most important feature of the application presented in this paper.
results accuracy	M/H	Our evaluation shows that the hashtags correctly identify key sequences of the open ceremony.
multiple languages and accessibility	M	The user interface is developed in HTML 5 and runs on any HTML 5 enabled browser. We have successfully tested it on Firefox 15, Firefox Mobile 15.0.1, Safari 5.1.7, Safari Mobile for iOS 5, and Chrome for Android.

Table 2. Additional Desirable Features