Social listening of City Scale Events using the Streaming Linked Data Framework

Marco Balduini[1], Emanuele Della Valle[1], and Daniele Dell'Aglio[1], Mikalai Tsytsarau[2], Themis Palpanas[2], Cristian Confalonieri[3]

[1] DEIB – Politecnico di Milano, Italy
marco.balduini@polimi.it, emanuele.dellavalle@polimi.it, daniele.dellaglio@polimi.it
[2] DISI – Universitá degli Studi di Trento, Italy
tsytsarau@disi.unitn.it, themis@disi.unitn.eu
[3] Studiolabo, Italy
cristian.confalonieri@studiolabo.com

Abstract. City-scale events may easily attract half a million of visitors in hundreds of venues over just a few days. Which are the most attended venues? What do visitors think about them? How do they feel before, during and after the event? These are few of the questions a city-scale event manger would like to see answered in real-time. In this paper, we report on our experience in social listening of two city-scale events (London Olympic Games 2012, and Milano Design Week 2013) using the Streaming Linked Data Framework.

1 Introduction

City-scale events are a group of events (usually with a common topic) located in multiple venues around a city. Olympic games, trade exhibitions and white night festivals can be examples of these kinds of events: they can be located in different venues in one or more districts of a city. The scale of these endeavors implies the involvement of different actors, such as city managers, organisers, sponsors, citizens and visitors.

One common problem of the involved actors is the monitoring of the city-scale events: organisers are interested in real-time monitoring of appreciation and popularity of the events; city managers and citizens want to assess the impact on the traffic, pollution, garbage collection; sponsors want to know if their investments are given back in terms of perception and image; visitors want to find more popular events.

The main barrier in monitoring the events is the data collection: available indicators (e.g., capacity of the venues and number of sold tickets) allow to make predictions, but they are not enough to have accurate results. On the other hand, a manual collection of information to perform these kinds of analysis is quite complex and expensive. A cheaper way lies in collecting all the necessary information from the Social Web, e.g., Twitter and Instagram, which provide huge amounts of data.

In this work, we present Streaming Linked Data (SLD), a framework to collect data streams, analyse them and visualise the results in dashboards. SLD exploits several semantic technologies: RDF to model and integrate the data; SPARQL (in particular its extensions for continuous querying) and sentiment mining techniques to process and analyse social data. We report on our experience in designing the framework and on its application to monitoring of two social city-scale events: the London Olympic Games 2012 and the Milano Design Week 2013 (a group of events co-located with the Salone Internazionale del Mobile[4], the largest furniture fair in the world). To summarize the contributions of this paper:

- We describe and analyse the concrete problems and user requirements for social listening of city-scale events (Section 2).
- We describe the Streaming Linked Data (SLD) framework and sentiment mining techniques adapted for streaming (Section 3).
- We report on the pragmatics of deploying and using of SLD to monitor two city-scale events: the London Olympic Games 2012 (Section 4), and the Milano Design Week 2013 (Section 5). These use cases prove the feasibility of our approach based on social listening.
- We assess the pros and cons of implementing, deploying, using, and managing SLD for city-scale events listening based on social media (Section 6).

2 The Problem and User Requirements

The work on SLD started in developing the mobile application BOTTARI [1], but the full requirements of SLD have been elicited through the analysis of other two use cases: the Olympic Games in London 2012 [2], and the Milano Design Week 2013 (MDW).

The analysis of the tweets about London Olympic Games was done at Politecnico di Milano and it is the first experiment with a large amount of data (more than three million tweets) performed within SLD. In this work we focused on the following questions:

1. Is it possible to detect the Olympic Games-related events analysing the Twitter streams?
2. Is it possible to track the movement of the crowds through geo-tagged tweets?

The experience and results we obtained during Olympic Games monitoring served as the basis for the Twitindex Fuorisalone application we implemeted for the Salone del Mobile. The project was realised by Politecnico di Milano and Universitá di Trento, in collaboration with Studiolabo and ASUS Italy. Studiolabo is a Milano-based company that hosts every year Fuorisalone.it, the official portal for the events in MDW; ASUS Italy acted both as an organiser and a sponsor: on one hand, it organised events for new product launches, and on the other hand it sponsored the Fuorisalone.it Web site and the events held in the

[4] Cf. http://www.cosmit.it/en/salone_internazionale_del_mobile

Brera district (grouped under the label *Brera Design District*). Twindex Fuorisalone aims to offer a social listing service for the events, with a particular focus on Brera Design District and the events of ASUS Italy. Studiolabo and ASUS Italy would like to know if it is possible using commodity hardware[5] to visually answer the following questions with an interactive HTML5 web application:

3. Is MDW visible in the social streams posted by people in Milano area?
   If yes in real-time,
   (a) What are the districts from which MDW visitors post the most?
   (b) What are the most frequently used hashtags?
   (c) How do people feel before, during and after the event they join?
4. Is the launch of ASUS products during MDW visible in the social streams posted by people around the world?
   If yes, not necessarily in real-time,
   (a) What are the products that attract more attention?
   (b) What is the global sentiment before, during and after the launch?

Addressing these problems poses the following technical requirements:

R.1 *Accessing the social stream* – all questions require that either the micro-posts of the social stream are brought to SLD or that part of the analysis is pushed to the social stream.

R.2 *Recoding and replaying portions of the social stream* – data streams are unbound and cannot be stored entirely, however it should be possible to record a portion of the data stream and re-play it on demand.

R.3 *Decorating the social stream with sentiment information* – questions 3.c and 4.b require to interpret emotions contained in micro-posts, for this reason it is necessary to decorate (some of) the micro-posts with an indicator of the sentiment they carry. At least for answering 3.c, the decoration has to be performed in real-time.

R.4 *Continuously analysing the social stream* – all our questions require to analyse time-boxed portions of the social stream in order to compute the up-to-date statistics on the fly, even for micro-posts decorated with sentiments.

R.5 *Internally streaming partial results of the analysis* – different continuous analysis may have parts in common; for instance questions 3.a and 3.b share the common need to apply a geo-filter. Moreover, a continuous analysis problem may be naturally split in a number of low-level analyses which detect aggregated events that are further processed in downstream continuous analysis; for instance question 2 requires to identify areas where crowds are assembling and to check if the crowd is moving over time in adjacent areas. So, the system should support the layout of complex analysis as an acyclic directed graph of components connected through internal data streams.

R.6 *Publishing and visualising continuous analysis results* – the results of continuous analysis are tables of data, and effective visualisation is required to allow users to understand the results. Moreover, given that analysis is performed on a server and visualised by HTML5 browsers, a Web-based communication protocol between these two components has to be provided.

[5] €100/month share in a cloud environment: 4 cores, 8GB of RAM, 200 GB of disk.

3 The Machinery

The transient nature of streaming information often requires to treat it differently from persistent data, which can be stored and queried on demand. Data streams should often be consumed on the fly by continuous queries. Such a paradigmatic change have been largely investigated in the last decade by the database community [3], and, more recently, by the semantic technology community [4]. Several independent groups have proposed extension of RDF and SPARQL [5] for continuous querying [6–8] and reasoning [9, 10].

These solutions introduce: *a*) the notion of RDF stream – a continuous flow of triple annotated with a timestamp identified by an IRI –, and *b*) means for continuously analysing RDF streams. These solution cover only the continuous analysis requirement (R.4). For this reason in this paper, we propose the Streaming Linked Data (SLD) framework: a general-purpose, pluggable system that supports the development of applications that continuously analyse RDF streams.

SLD server is designed according with the following three principles:

1. it is a *publish/subscribe* system where senders – the publishers – publish timestamped RDF triples into RDF streams, and receivers – the subscribers – listen to one or more RDF streams, and only receive timestamped RDF triples that are of their interest. Publisher and subscribers do not have to know each other.
2. it is logically a *reliable message-passing* system that guarantees timestamped RDF triples to be delivered in order; and
3. it *minimises latency by using main memory* and avoiding disk I/O bottlenecks.

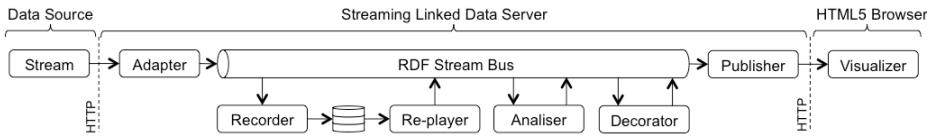

Fig. 1: The architecture of the Streaming Linked Data framework.

Figure 1 illustrates the architecture of the SLD framework. The leftmost column logically contains the *streaming data sources*, the central one the SLD server, and the rightmost one the visual widgets to be embedded in a dashboard.

The *streaming data sources* are assumed to be distributed across the Web and accessible via HTTP. For the scope of this work, we consider only the streaming APIs of Twitter[6]), but a growing amount of data sources exposes information as data stream using a variety of Internet protocols.

The core of the framework is *SLD Server*. It includes components for accessing data stream sources, internally streaming data, registering and replaying

[6] See `https://dev.twitter.com/docs/streaming-api`

portion of data streams, decorating and analysing time-boxed portion of the stream, and publishing the results.

The adapters allow to access data stream resources, possibly delegating filtering operations to the data source, and to translate data items in the stream into set of timestamped RDF triples. Thus, they satisfy requirement R.1. For the scope of this work, we only used the Twitter adapter, but the SLD framework also includes adapters for Instagram, foursquare and several sensor networks. This adapter allows to push to Twitter either geo-spatial filters, which ask Twitter to stream to SLD only tweets posted from given locations, or keyword-based filters, which ask Twitter to stream to SLD only tweets containing one or more of such key-words. Each tweet is internally represented using the extension of SIOC ontology presented in [1].

For instance, hereafter, we represent in RDF the tweet[7] that Tim Berners-Lee posted live from the middle of the Olympic stadium during the opening ceremony of London 2012 Olympic Games:

```
[] sioc:content "This is for everyone #london2012 #oneweb #openingceremony";
   sioc:has_creator :timberners_lee;
   sioc:topic :london2012, :oneweb, :openingceremony .
```

An RDF stream bus supports the publish/subscribe communication among the internal components of SLD. Logically, it is a collection of RDF streams, each identified by an IRI, and takes care of dispatching the timestamped triples injected in an RDF stream to all components that subscribed to it. It addresses, therefore, requirement R.5.

The publishers make available on the Web the content of chosen RDF stream following the Linked Data principles [11] in the Streaming Linked Data format proposed in [12]. The format is based on two types of named RDF graphs: instantaneous Graphs (iGraphs), which contain a set triples having the same timestamp, and stream graphs (sGraphs), which contains triples that point to one or more timestamped iGraphs. The number of iGraphs pointed by an sGraph and their time interval of validity can be configured when instantiating the publisher. Publishers partially address requirement R.6.

The recorders are special types of publishers that allow for persistently storing a part of an RDF stream. As format, we used an extension of the Streaming Linked Data format based on iGraphs and recording graphs (rGraphs). The latter are similar to sGraphs, but they include pointers to all the iGraph recorded and such pointers do not have a time interval of validity. The re-players can inject in an RDF stream what recorded in an rGraph. Recorders and re-players together address requirement R.2.

The analysers continuously observe the timestamped triples that flow in one or more RDF stream, perform analyses on them and generate a continuous stream of answers. Any of the aforementioned continuous extensions of SPARQL can be plugged in SLD server and used for the analysis. For the scope of this work, we

[7] See https://twitter.com/timberners_lee/status/228960085672599552

used a built-in engine that executes C-SPARQL queries. The analysers address requirement R.4.

The following C-SPARQL query, for instance, counts for each hashtag the number of tweets in a time window of 15 minutes that slides every minute.

```
1  REGISTER STREAM HashtagAnalysis AS
2  CONSTRUCT { [] sld:about ?tag ; sld:count ?n . }
3  FROM STREAM <http://.../London2012> [RANGE 15m STEP 1m]
4  WHERE { { SELECT ?tag (COUNT(?tweet) AS ?n)
5            WHERE { ?tweet sioc:topic ?tag . }
6            GROUP BY ?tag } }
```

The `REGISTER STREAM` clause, at Line 1, asks to register the continuous queries that follows the `AS` clause. The query considers a sliding window of 15 minutes that slides every minute (see clause `[RANGE 15m STEP 1m]`, at Line 3) and opens on the RDF stream of tweets about the Olympic games (see clause `FROM STREAM` at Line 3). The `WHERE` clause, at Line 5, matches the hashtags of each tweet in the window. Lines 6 asks to group the matches by hashtag. Line 4 projects for each hashtag the number of tweets that contains it. Finally, Line 2 constructs the RDF triples that are streamed out for further down stream analysis.

The decorators are special types of analysers that look for a pattern of triples in a RDF stream. When the pattern matches, the decorators run a computation of the matching and add new triples to the stream. The decorators address requirement R.3.

As one of such decorators for our analysis for MDW, we deployed a sentiment mining component, which runs on the tweets written in English or in Italian that matches specific keywords. Following the identification of a valid tweet, this component adds a sentiment triple to its RDF representation. More specifically, we used a dictionary-based sentiment classifier provided by the Universitá di Trento [13], which was extended by positive and negative emotion patterns. Dictionary-based sentiment classifiers are known to be efficient for short texts concentrating on a single topic, such as tweets. A sentiment dictionary can also be adapted to the particular domain of analysis, since many sentiments are domain-specific. While this method is very suitable for large-scale analysis thanks to its minimal performance requirements, some sentiments (e.g., sarcasms, idioms) require more robust methods.

Last,but not least, the SLD framework includes a library of visual widgets, written in HTML5, that periodically visualises what is published as Linked Data by the publishers. For the scope of this work we used heat maps, bar charts, area charts and dot charts. Publishers and visual widgets together address requirement R.6.

4 London Olympic Games 2012

In the following we describe two of the analyses we developed in the London Olympic Games 2012 application. More informations are available at `http://www.streamreasoning.org/demos/london2012`.

Detecting Events. The first analysis aims to detect the events given the position of a set of venues and socially listening their surroundings. As input data, SLD received all the three million tweets streamed by Twitter between July 25^{th} and August 13^{th} 2012. Additionally, this analysis focused on three venues that represent the big, medium and small venue types of London 2012:

- The Olympic stadium[8] where all the athletic games took place; a prestigious venue with a capacity of 80,000 seats.
- The aquatic centre[9] that was used for the swimming, diving and synchronised swimming events; a medium-size venue that can seat 17,500 people.
- The water polo arena[10]; a 5,000-seat venue that hosted both the men's and women's water-polo competitions.

As ground truth for the experiment we used the calendar of Olympic Games[11]. The analysis relies on the identification of bursts of geo-located social activity. To identify them, we adapt a method that was shown to be effective in identifying bursts in on-line search queries [14]. We model a network of C-SPARQL queries that counts the tweets posted from a given area every 15 minutes and identifies a burst when the number of tweets in the last 15 minutes is larger than the average number plus two times the standard deviation in the last 2 hours. An events is detected in a venue if a burst is detected at public transport stations, then in areas outside the venues and finally in the venues.

Figure 2 visually shows the results of SLD across the 20 days of games in the three venues. Each diamond represents an event detected in the venue. The black line a moving average with a two period. The grey bars represent events scheduled in the Olympic Calendar; light ones are competitions whereas dark ones are finals.

In the stadium, SLD was able to detect all events in the ground truth: the rehearsal for the opening ceremony on July 25^{th}; the opening ceremony on July 27^{th}; the pair of events scheduled (one between 10 am and 1:30 pm, and another between 6 pm and 10 pm) on August 3^{rd}, 4^{th}, and 6^{th} to 9^{th}; the single event on August 5^{th}, 10^{th}, and 11^{th}; and the closing ceremony on August 12^{th}. It is worth to note that the magnitude of the burst is related to the importance of the event, e.g. on August 4^{th} took place the women's 100 metres final, and on August 5^{th} the men's 100 metres final. Moreover, the competitions were absent from the stadium until August 3^{rd}, and in this period our method detected a large number of unscheduled events (i.e., not present in the ground truth) with a little magnitude. Those are, on the one hand, easy to isolate and discard, but they are also interesting because they are spontaneous assembling of people.

In the aquatic arena, which attracts less attention in terms of tweets, our method performed with a high precision[12] (i.e., only three unscheduled events

[8] http://en.wikipedia.org/wiki/Olympic_Stadium_(London)

[9] http://en.wikipedia.org/wiki/London_Aquatics_Centre

[10] http://en.wikipedia.org/wiki/Water_Polo_Arena

[11] http://en.wikipedia.org/wiki/2012_Summer_Olympics#Calendar

[12] With precision, in this context, we mean the fraction of identified events that were actually scheduled.

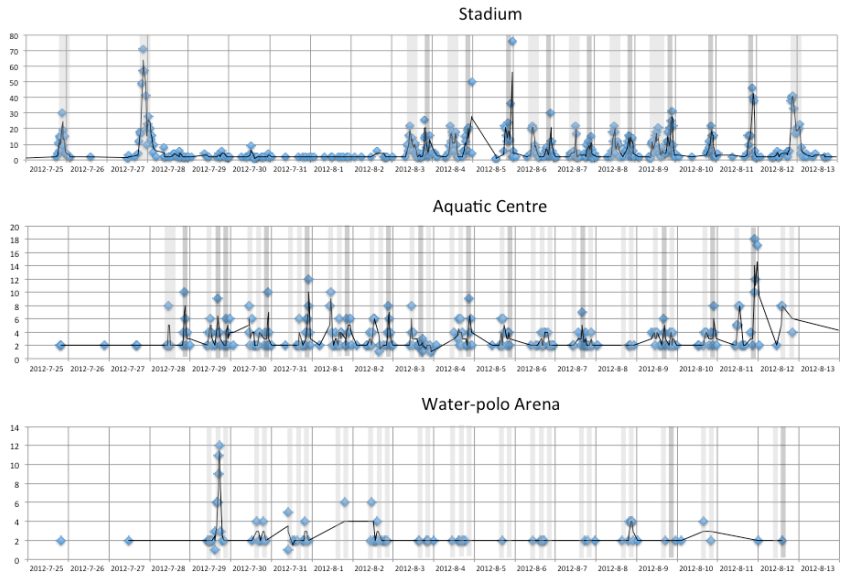

Fig. 2: The results of the event detection experiment.

were detected before the opening ceremony), but with a recall[13] of 76% (32 events out of the 42 planned). Also in this case the magnitude of the burst speaks for the importance of the event: most of the finals have high peaks.

In the water polo arena, which is a small venue hosting a single sport, our method was still precise, but the recall was very low (32%, i.e., 11 events out of the 34 planned). The only event that generated a large burst was on July 29^{th}.

Visualizing Crowd Movements. With the first experiment we give some guarantees about the ability of our machinery to detect crowd assembling to follow an event. The method looks for a sequence of bursts detected first at public transport stations, then in the walkable areas outside the venues and finally in one of the venues.

In this section, we show that this pattern can be visually captured by the means of a time series of heatmaps. Each heatmap highlights the presence of crowds using geotagged tweets as a proxy for Twitter users' positions[14].

We report on two experiments: *a*) on little less than 40 thousands geo-tagged tweets received the night of the Open Ceremony (between July 27^{th}, 2012 at 2 pm[15] and the day after at 6 am), and *b*) on the few thousands tweets collected in a crowded evening at the aquatic centre (between 4 pm and 11 pm on July 31^{th}) where an event that started at 7:30 pm and ended at 9:20 pm.

Figure 3 displays the results we obtained. In the case of the opening ceremony we were able to follow the flow of the crowd. At 2:39 pm almost nobody was

[13] With recall, in this context, we mean event scheduled that are identified.

[14] As in many other studies based on twitter, we are assuming that Twitter user's are uniformly distributed in the crowd.

[15] All times are given in British Summer Time (BST)

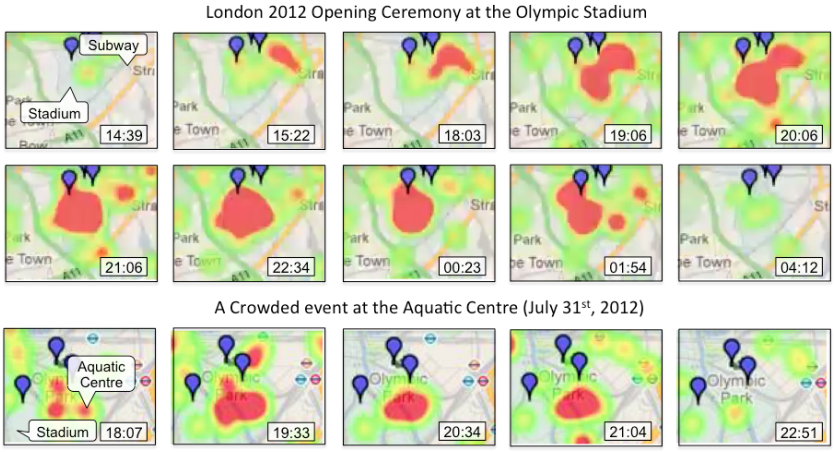

Fig. 3: The sequence of heatmaps visualises the flows of crowd from the public transports to the Olympic venues in two different scenarios.

twitting from the Olympic stadium area. At 3:22 pm a crowd of twitter users started twitting from Stratford subway and light rail station. The heatmaps at 6:03 pm, 7:06 pm, and 8:06 pm show a continuous flow of people exiting Stratford station, funnelling through Stratford walk, and entering the stadium. During the entire ceremony (between 9:00 pm and 00:46 am) the crowd only twitted from the stadium. The heatmap at 01:45 am shows the presence of a big crowd in the stadium area and a smaller one on Stratford station. By late morning (see heatmap at 04:12 am) the stadium area was empty again.

The second experiment shows a worst case scenario. It aims at showing the results that can be obtained when some 10 geo-tagged tweets per minute are received. Our methods still adequately shows the assembling of a crowd, but it does not allow to follow its movements. The heatmaps at 6:07 pm shows some activity in the walkable areas in the Olympic park and in the aquatic centre. At 7:33 pm people are still walking down Stratford walk and entering the aquatic centre. The heatmaps at 8:34 pm and 9:04 pm show the crowd in the aquatic centre. By 10:51 pm the venue was empty.

5 Milano Design Week 2013

The Milano Design Week is an important event for the Italian city: every year it attracts more than 500.000 visitors. During that week Milano hosts Salone Internazionale del Mobile – the largest furniture fair in the world – and the Fuorisalone[16] – more than a thousand of satellite events that are scheduled in more than 650 venues around Milano. These events span the field of industrial design from furniture to consumer electronics.

[16] See `http://fuorisalone.it/2013/`

Twindex Fuorisalone is the application we deployed for StudioLabo and ASUS during MDW 2013 using SLD. Interested readers can access the dashboard at `http://twindex.fuorisalone.it` and read more about it at `http://www.streamreasoning.org/demos/mdw2013`. It was planned to be a two-steps experiment. The first one was run in real time during the MDW 2013 on the tweets posted from Milano. An HTML5 dashboard[17] was deployed and it was accessible to organisers and visitors of the event. During this step Twindex Fuorisalone recorded the tweets posted from Milano as well as those posted world-wide that contains 300 keywords related to MDW, Brera district and the products ASUS planned to launch during MDW 2013. The result is a collection of 107,044,487 tweets that were analysed in the second step of the project.

Figure 4 illustrates the lay out of the SLD application that underpins the dashboard shown in Figure 5.(a). Moving from left to right, the leftmost component is the Twitter adapter. It injects tweets represented in RDF using SIOC vocabulary in an internal RDF stream. The sentiment decorator decorates each RDF tweet representation with a value in the range [-1,1] that accounts for the sentiment expressed in the twitter. As vocabulary we used the extension of SIOC proposed for BOTTARI [1]. The decorated tweets are injected in a new internal RDF stream to which a number of components are subscribed. Moving, now, from top to bottom of Figure 4, a publisher, which keeps the last hour of tweets and slides every 15 minutes, makes available that data for the heatmap shown in the topmost position of the dashboard in Figure 5.(a). A continuous query counts the tweets posted from Milano every 15 minutes, isolating those that contains a set of 30 keywords related to MDW, those that carry a positive sentiment (in the range [0.3,1]) and those that carry a negative sentiment (in the range [-1,-0.3]). A publisher listens to the results of this query and makes them available for 2 hours. A bar chart widget is subscribed to such a publisher and displays the number of tweets every 15 minutes broken down in positive, neutral and negative (see the vertical bar chart in Figure 5.(a)). Also an area chart widget is subscribed to the same publisher and shows in black the number of tweets posted in Milano and in yellow the number of those that contains the 30 terms about MDW. A second continuos query extracts the top 10 most frequently used hashtags. Its results are displayed in the horizontal bar chart present in the dashboard. The same analyses are continuously performed also for each area of Milano where MDW events are scheduled.

The real-time experiment was conducted between April 8^{th} and April 17^{th}, 2013. Twindex Fuorisalone was viewed by 12,000 distinct users. The publishers were invoked 1,136,052 times. The SLD server analysed 106,770 tweets with the network of queries illustrated in Figure 4. We spent €25 using at most 2 CPU and 2 GB of RAM of the machine we reserved on the cloud. The most interesting results are shown in Figure 5.(b) and (c) and in Table 1.

As illustrated by Figure 5.(b) MDW 2013 is visible in the volume of observed tweets. On April 8^{th}, 2013 at 18.00 the number of tweets moves from 90/150 every 15 minutes to 180/210 (see point marked with A in the figure). For the entire

[17] See `http://twindex.fuorisalone.it` where the application is still running

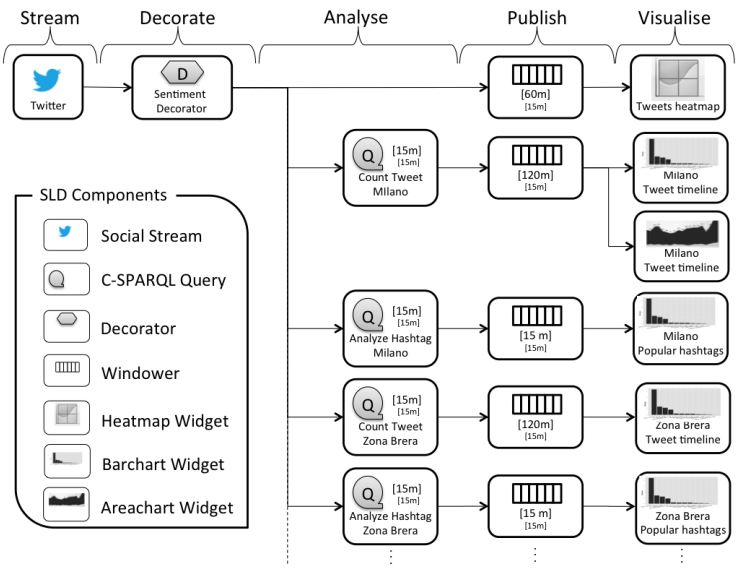

Fig. 4: The lay out of the SLD application that underpins the dashboard shown in Figure 5.(a)

duration of MDW 2013 the volume of tweets is larger than 100 tweets every 15 minutes, while normally is less than 100. During MDW the number of tweets after mid-night is much larger than in the normal days (see point marked with B in the figure). The April 14^{th}, 2013 at 20.00 MDW ends and the volume of tweets rapidly goes back under 100 tweets every 15 minutes (see point marked with C in the figure). The yellow area (the number of tweets that refers to MDW 2013) is more visible during the event that in the following days.

Figure 5.(c) shows the *hot points* visually identified by the heatmap during a night of MDW 2013 (on the left) and in a night after MDW (on the right). Normally few geo-tagged tweets are posted from Brera, during MDW a number of hot points were detected. The two most popular venues were Cesati antiques & works of art and Porta nuova 46/b; 16,653 and 13,416 tweets were, respectively, posted in their proximity. [1000-10000] tweets were posted in the proximity of a group of 6 venues that includes Circolo Filologico, Adele Svettini Antichitá, ALTAI, Bigli19, Dudalina and Galleria DadaEast. [100-1000] tweets were posted around another group of 10 venues. The venues around which [10-100] tweets were posted are 62. Around the remaining 81 only few tweets were posted.

Table 1 allows to compare the top-5 most frequently used hashtags in Milano in a late afternoon during MDW 2013 and one after MDW. Normally the geo-tagged tweets of Milano in the late afternoon talks about football, whereas during MDW 2013 the most frequently used hashtags were related to the ongoing event.

The post-event analysis considered the 107,044,487 tweets registered with SLD between April 3^{rd} and April 30^{th}, 2013 asking Twitter to send to SLD tweets containing 300 words related to MDW, ASUS and its products. Figure 6

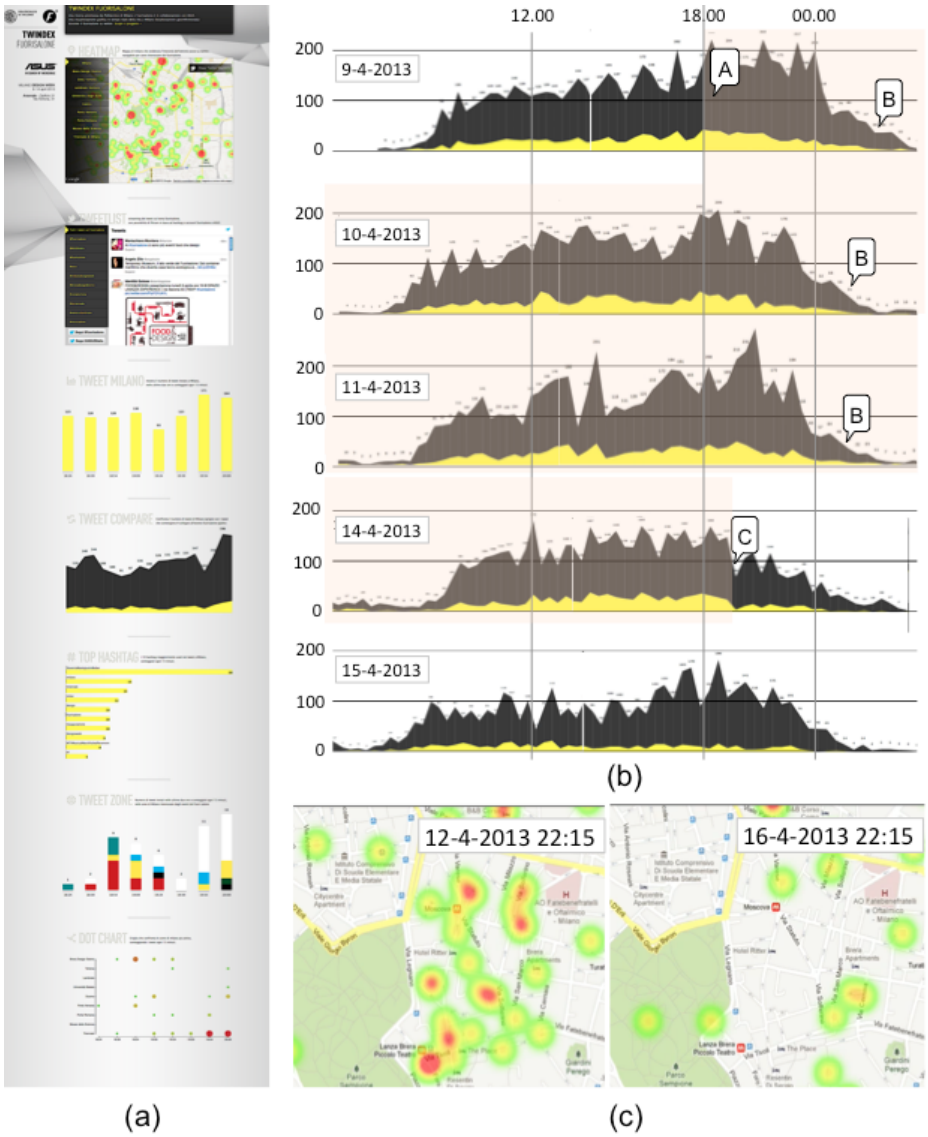

Fig. 5: The figure illustrates: (a) a screenshot of Twindex Fuorisalone (for a the running system visit `http://twindex.fuorisalone.it`, while a detailed explanation is available at `http://www.streamreasoning.org/demos/mdw2013`); (b) a series go area charts that plot the number of tweets posted every 15 minutes in Milano during MDW 2013 (the yellow area is the fraction of tweets that contains keywords related to MDW) where MDW opening (point marked with A), overnight events (B) and closing (C) are well visible; and (c) the comparison on an heatmap between the hot spots visualised in a night during MDW 2013 (on the left) and in a normal day (on the right).

Table 1: A comparison between the top-5 hashtags used in geo-tagged tweets in Milano during a late afternoon of MDW 2013 and one after MDW.

April 9^{th}, 2013 at 18.00		April 15^{th}, 2013 at 18.00	
fuorisalone	30	inter	20
designweek	28	diretta	11
nabasalone	20	cagliari	6
milano	9	milan	4
design	6	seriea	3

illustrates the results we obtained analysing the tweets related to *ASUS*, *FonePad* – a product ASUS launched during MDW – and *VivoBook* – a product ASUS presented for the first time in Italy during MDW.

As illustrated in Figure 6, the volume of tweets posted worldwide related to the topic *ASUS* slightly increases during MDW 2013 where it launched its *FonePad*, it started the pre-sales of the *FonePad* in Italy, and *b*) it presented to the Italian market its *VivoBook*. Those launches and presentations are also visible in the volume of tweets about the two products. It is worth to note that, while VivoBook was already on the market, the FonePad is a new product. The volume of tweets about VivoBook had a burst to 150 tweets/hour the first day of MDW and then went back to tens of tweets per hour, while the volume of tweets about FonePad steadily increased during the observation period with a high burst during MDW, for the launch in Japan and when the online reviewing started.

The sentiment expressed in the tweets about ASUS was mostly positive. The contraction level during such periods was also high due to concern expressed by some users. A similar phenomenon was also uncovered by our analysis when the online reviews of *FonePad* and *VivoBook* started. Reviews of these products, although very positive, caused a lot of discussions in the media, where mixtures of positive and negative sentiments were expressed, resulting in more contradicting distributions. Analysing the micro-posts on FonePad during the contradictory time intervals we discovered that the negative sentiments mostly concern its unusual large size while the positive sentiments are all about its affordable price and concept novelty. As expected, the method did not handle sarcasm in a satisfactory manner: some tweets about FonePad contained sentences like *"wanna buy it so bad!"*, which were classified as negative, but in reality were expressing positive sentiment.

6 Conclusions

In this section, we first elaborate on pros and cons of using Semantic technologies for social listening and then on cost and benefits of our approach w.r.t. traditional ones (e.g., volunteers, CCTV and mobile telephone data analysis).

SLD is an extensible framework based on Semantic technologies to process data streams and visualise the results in dashboards. The usage of RDF to model a micro-post is straight forward. Tweets are small graphs: a user posts a short text containing zero or more hashtags, including zero or more links, referring to

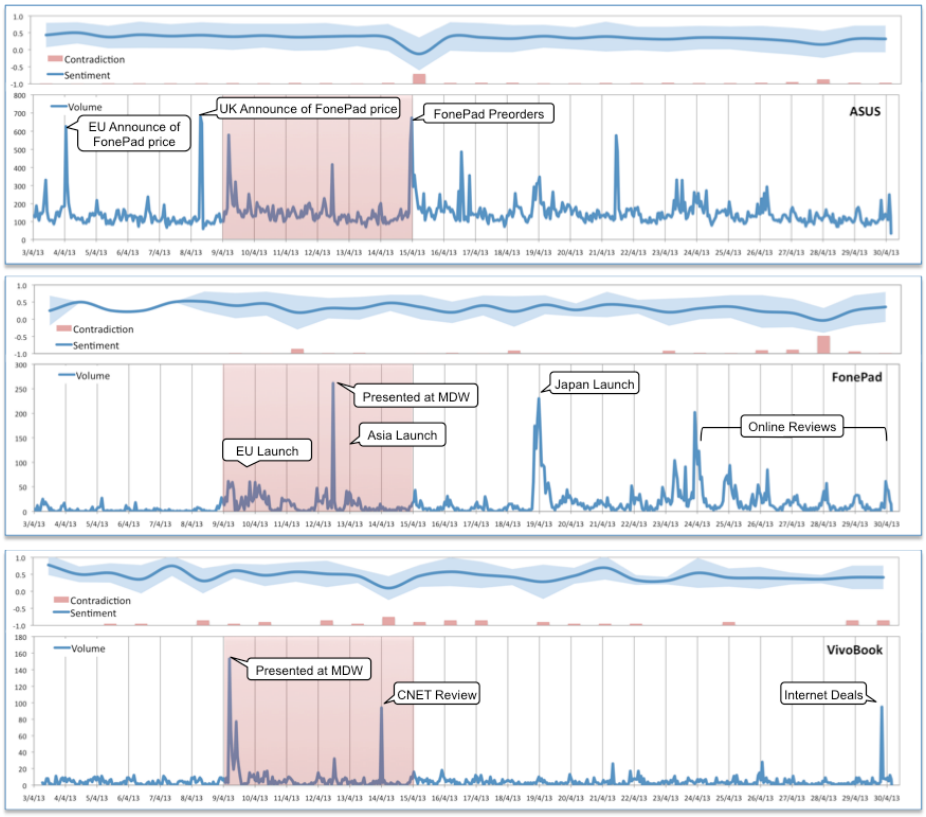

Fig. 6: Results of the sentiment analysis carried out on the tweets about ASUS and two of its products: FonePad and VivoBook.

zero or more users, potentially retweeting another tweet and reporting her location. Using the relation model to represent a tweet is less natural since it requires using denormalised relations. The usage of C-SPARQL to encode analyses is certainly a barrier, but using a continuous relational query language like EPL (i.e., the event processing language used in Oracle CEP and other stream processing engines) is at least as difficult as using C-SPARQL. Moreover, SLD allows the introduction of custom code where needed, both the decorators and the analysers are abstract components to be implemented. For the scope of this work, the sentiment mining component was inserted in SLD as a decorator with minimal effort. Finally, SLD offers a set of visualisation widgets based on (Semantic) Web technologies that simplify the creation of dashboards and decouple presentation from analysis. Such a decoupling was shown to be effecting in realising Twindex Fuorisalone where Politecnico di Milano and Universitá di Trento worked on the analysis, while Studiolabo prepared the dashboard assembling and customising SLD visualisation widgets.

In this work, we discussed the pragmatics of using SLD to analyse city-scale events through two use cases: the London Olympic Games 2012 and the Milano Design Week 2013.

In the case of London Olympic Games, we addressed the problem of detecting the assembling and tracking the movements of crowds during city-scale events. These problems have been solved in a number of ways already. Available solutions include the traditional employment of volunteers and CCTV, and the innovative usage of mobile phone network data [15]. However, only big event organisers can afford either the huge human effort or the high cost[18] of these solutions. On the contrary, social listening is also affordable for city scale events like MDW.

The most critical issues is determining when enough tweets have been observed. The assumption that tweeter users are dense in the crowd does not always hold. However, an interesting fact we noted is that the size of the input data affects the recall more than the precision. As we discussed in Section 4, more data is available, higher is the recall: in a venue like the Olympic stadium our approach identifies nearly 100% of the events in the ground truth, while in water polo arena only 32%. However, the input size is not the only important feature to be considered. The hot spots identified by Twindex Fuorisalone in the Brera district are in close proximity of MDW venues, thus they allow to identify the events even if the number of tweets for venue are less than those related to the water polo arena in London. The size of the venue, the length of the event, probably also the nature of the event also matter. We plan to investigate more on these topic in our future works.

In the case of Milano Design Week 2013, we also address the problem of detecting what attracts the attention of crowds and what are their feelings. It is worth to note that the analysis of mobile phone data is not sufficient to address this second problem. Accessing the content of SMS and phone calls rises serious privacy issues and it is, thus, forbidden. In the case of social streams like Twitter, those who post are aware that the content of their micro-posts is public and both hot topic and sentiment can be extracted from the short text. The results presented in Section 5, positively answer to the question risen in Section 2. Hot spots appear in proximity to the MDW venues in areas from where nobody tweets in other days (answering question 3.a). The most frequently used hashtags during MDW were related to the ongoing event, while in other days topics like football dominates the top-5 hashtags (answering question 3.b). We were able to explain bursts of the tweets volume corresponding to launches and presentations of ASUS products during MDW (answering question 4.a). Moreover, we detected that public sentiments being initially less positive during the anticipation of announcement, transited to more positive during- and after the corresponding events (answering questions 3.c and 4.b).

Social listening proved to be a powerful approach to use in city-scale events, where huge amount of people (usually with common interests) are in the same locations at the same time. However, those that tweets may not be uniformly

[18] Aggregated mobile phone data are sold by telecom at thousands of euros per hour of analysed data.

distributed among the visitor of an event, while mobile phones certainly are. Our future work is centred on the combination of social listening and mobile phone data analysis using SLD. We want to assess if data from social streams and mobile data carry different information, and if they complement each other. As example, before and after a concert people call, while they prefer to use Twitter or Facebook to update their statuses during the play.

Acknowledgments. We thank ASUS Italia for supporting this initiative.

References

1. Balduini et al.: BOTTARI: An augmented reality mobile application to deliver personalized and location-based recommendations by continuous analysis of social media streams. J. Web Sem. **16** (2012) 33–41
2. Balduini, M., Della Valle, E.: Tracking Movements and Attention of Crowds in Real Time Analysing Social Streams – The case of the Open Ceremony of London 2012. Semantic Web Challenge at ISWC 2012
3. Garofalakis, M., Gehrke, J., Rastogi, R.: Data Stream Management: Processing High-Speed Data Streams. Springer-Verlag New York, Inc. (2007)
4. Della Valle, E., Ceri, S., van Harmelen, F., Fensel, D.: It's a Streaming World! Reasoning upon Rapidly Changing Information. IEEE Intelligent Systems **24**(6) (2009) 83–89
5. Prud'hommeaux, E., Seaborne, A.: SPARQL Query Language for RDF. `http://www.w3.org/TR/rdf-sparql-query/`
6. Barbieri, D.F., Braga, D., Ceri, S., Della Valle, E., Grossniklaus, M.: Incremental Reasoning on Streams and Rich Background Knowledge. In: ESWC. (2010)
7. Le-Phuoc, D., Dao-Tran, M., Xavier Parreira, J., Hauswirth, M.: A native and adaptive approach for unified processing of linked streams and linked data. In: ISWC. (2011) 370–388
8. Calbimonte, J.P., Corcho, O., Gray, A.J.G.: Enabling ontology-based access to streaming data sources. In: ISWC. (2010) 96–111
9. Barbieri et al.: C-SPARQL: a Continuous Query Language for RDF Data Streams. Int. J. Semantic Computing **4**(1) (2010) 3–25
10. Anicic, D., Fodor, P., Rudolph, S., Stojanovic, N.: EP-SPARQL: a unified language for event processing and stream reasoning. In: WWW. (2011) 635–644
11. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. Int. J. Semantic Web Inf. Syst. **5**(3) (2009) 1–22
12. Barbieri, D.F., Della Valle, E.: A proposal for publishing data streams as linked data - a position paper. In: LDOW. (2010)
13. Tsytsarau, M., Palpanas, T., Denecke, K.: Scalable Detection of Sentiment-Based Contradictions. In: DiversiWeb workshop, WWW, Hyberabad, India (2011)
14. Vlachos, M. et al.: Identifying similarities, periodicities and bursts for online search queries. In: SIGMOD Conference. (2004) 131–142
15. 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