INFORMATION VISUALISATION FOR SOCIAL MEDIA ANALYTICS

Rozenn Dahyot, Conor Brady, Cyril Bourges and Abdullah Bulbul

School of Computer Science and Statistics
Trinity College Dublin
Ireland

ABSTRACT

This paper tackles the audio visual renderings of geolocated datasets harvested from social networks. These datasets are noisy, multimodal and heterogeneous by nature, providing different fields of information. We focus here on the information of location (GPS), time (timestamp) and text from tweets from which sentiment is extracted. We provide two ways for visualising datasets and for which demos can be seen online.

Index Terms— Social Media Analytics, Visualisation

1. INTRODUCTION

Digital content created and shared on the web, in particular social networks, have found unexpected applications beyond simple entertainment for users, by exploiting the strong connection between the virtual world where these data lies and the physical reality [1]. To navigate efficiently into large datasets harvested from the web, it is also important to deploy visualisation capabilities to generate insights and extract valuable information. Several mapping technologies exist for visualising geolocated data. As contributions to the field, we propose web based geolocated audio visualisation tools to summarise information harvested from the web, including an animated and interactive audio-visual tweet sentiment map (Section 3). We start first by presenting relevant prior works in the field.

2. PRIOR WORK

2.1. Social interactions with online content

People interactions with web content can be classified broadly into three classes. The first class is when a user is performing a search on the internet generally by entering keywords or sentences (generating *user search* data). The second class is when the user accesses and presumably read some specific material online (generating *user read* data). Finally the third class refers to when the user creates content online (generating *user create* data). All of these data can be collected and analysed to give insights into the online population.

978-1-4673-8457-5/15/$31.00

User search data captured as search engine queries has been successfully used for instance in detecting influenza epidemic [2]. *User read* data as measured by Wikipedia page view statistics, as well as *user create* data as posted on Twitter for instance, have been used to analyse trending topics [3]. Such user generated data is occasionally geolocated and geolocated tweets has found application in analysing movement patterns of individuals [4]. Similarly, our work focuses on geolocated datasets harvested from social networks (see Section 3) .

2.2. Sentiment Analysis & Deep learning

A common step in social media analysis is in computing a sentiment score. This sentiment score can then be used as a response or explanatory variable for explaining the collected data. For instance, Frank et al. characterize changes in word usage as a function of individuals' movement, and positive sentiment (happiness) extracted from tweets increases with the distance from an individual's average location [4]. Sentiment analysis can also be performed from images as an alternative to text [5].

Deep learning has grown popular in the last decade benefiting from growing computing power and large datasets availability. For instance, Yuan et al. [6] tackle the analysis of social media for retrieval purposes, with latent features automatically learned from a dedicated deep architecture. In Section 3, we propose to perform text sentiment analysis online using the deep architecture provided by CoreNLP library [7, 8].

2.3. Visualisation tools

The use in graphical methods in statistics has been around for more than a century [9]. The argument is that well designed charts are more effective in creating interest and appealing to the attention of the viewers. It also allows time to be saved while absorbing information content [9]. The use of charts and graphs also reveals hidden facts, relationships to trigger analytical thinking for further investigation. 2D geographic map (e.g. with colour encoding information such as population density) [10], or charts with scaling and effects to emphasis differences [9], or using advanced software such as Google

Earth [11, 10] are all very informative visual representation. Of interest in this paper is the sentiment viz project [12] for visualizing sentiment for twitter posts. Colour, brightness, size, and transparency are used for representing various information about the sentiment of a tweet. In a similar fashion, our work presented next proposes online visualisation tools using both visual and audio cues to encode sentiment extracted from prerecorded tweets.

3. GEOLOCATED AUDIO VISUAL RENDERINGS

We first introduced the datasets used in our experiments (Section 3.1) and then present some early work into the rendering of tweets activities (Section 3.3) and sentiment rendering (Section 3.4). These renderings are based on a common mathematical formulation using kernel density estimates explained in Section 3.2.

3.1. Architecture, JSON format and datasets

Data is harvested from the social networks Viddy, Instagram and Twitter using their APIs, and queries to collect the data are formulated using either keywords or hashtags, or GPS locations for a particular period in time. The data is collected in the JSON format (JavaScript Object Notation `http://json.org/`) which has many fields of information available for every post collected (see Figure 1). Tables 1 and 2 present the different fields of information available depending on the social networks used for both a user, and a post. This geolocated, time stamped data harvested from social media is noisy, multimodal and heterogeneous by nature. It is important therefore to have visualisation tools available to get insights about the collected information.

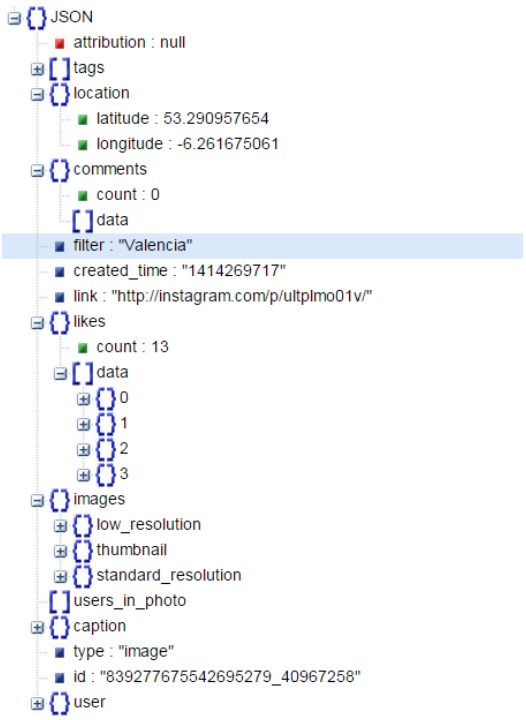


Fig. 1. Example of a post in JSON format.

User	Twitter	Instagram	Viddy
Info			
Username	✓	✓	✓
Fullname	✓	✓	✓
Language	✓		
Statistics (count)			
Followers	✓		✓
Following	✓		✓
Post	✓		✓
Mentioned	✓		

Table 1. JSON information per user.

Post	Twitter	Instagram	Viddy
Info			
Date	✓	✓	✓
Tags or Hashtag	✓	✓	✓
Coordinates or places	✓	✓	✓
Text or Caption	✓	✓	✓
User mention in text or picture	✓	✓	
Language	✓		
Statistics (count)			
Comment		✓	✓
Retweet	✓		
Favorite or Like	✓	✓	✓
Media			
Url	✓	✓	✓
Width	✓	✓	
Height	✓	✓	

Table 2. JSON information per post.

Two datasets have been created. The first, named *Dublin marathon 2014*, contains posts collected on the day of the marathon in Dublin Ireland (on Monday 27th October 2014) with hashtags #dublinmarathon, or #dublinmarathon2014 or with twitter user: @dublinmarathon, or with GPS location covering the marathon route. The second dataset, named *Trinity*, contains posts with GPS locations covering Trinity College Dublin and its neighbouring streets that have been posted over the Christmas break 2014 period.

3.2. Representation with Kernel Density Estimates (KDE)

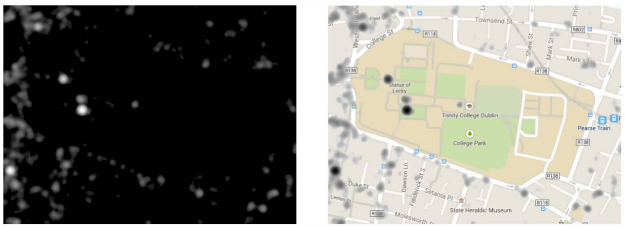
We note the raw dataset $\{d^{(i)}\}_{i=1,\cdots,n}$ as a collection of n documents d that contain several fields $d = (x, t, w, \cdots)$ (see Tables 1 and 2). We note x and t, the GPS location and time stamp respectively. We process the set of words w to provide a sentiment score s. We assume that all n documents have values available for x, t and s,with collected values $\{(x^{(i)}, t^{(i)}, s^{(i)})\}_{i=1,\cdots,n}$. Using kernel density estimators (KDE) [13], estimates of probability densities can be computed, for instance:

$$\hat{p}(x,t) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_x} k\left(\frac{x - x^{(i)}}{h_x}\right) \frac{1}{h_t} k\left(\frac{t - t^{(i)}}{h_t}\right) \quad (1)$$

k is a kernel function that is positive and integrate to 1, and the bandwidths h_x and h_t control the fuzziness level of the KDE. When these bandwidths are set to zero, the kernel k is the Dirac kernel and the estimate $\hat{p}$ is the empirical probability density function. In a similar fashion, estimates such as $\hat{p}(s, x, t)$, and marginals $\hat{p}(t)$ or $\hat{p}(x)$ can also be computed. Section 3.3 shows how $\hat{p}(x)$ can be visualised while Section 3.4 presents a visualisation for $\hat{p}(s, x, t)$.

3.3. Tweet Activity Visualisation

Figure 2 shows the Kernel density estimate $\hat{p}(x)$ computed with *Trinity* dataset. The Gaussian kernel is used with a preset bandwidth. Figure 2(a) presents the density values $\hat{p}(x)$ appearing white for high values and black for low ones over the GPS domain of x covering Trinity College Dublin. Figure 2(b) presents a negative image of Figure 2(a) superimposed on a Google map capture of the area. High tweeting activities correlate with tourist attraction locations in Trinity College Dublin such as the Campanile and the Book of Kells.



(a) (b)

Fig. 2. Tweet activity map as a kernel density estimate $\hat{p}(x)$ (a), superimposed on a map (b) (*Trinity* dataset).

The tweet activity map can also be rendered using the heatmap functionality in Google API for rendering tweets counts overlayed over Dublin (cf. Figure 3) [14]. Areas of higher intensity (values $\hat{p}(x)$) are colored red, and areas of lower intensity appear green. This functionality is shown on our prototype webserver[1].

[1] http://graisearch.scss.tcd.ie/online-platform

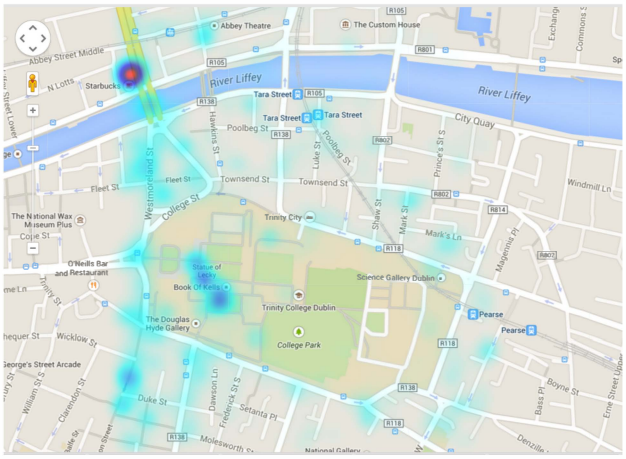


Fig. 3. Tweet Activity Saliency map visualized with Google API (*Trinity* dataset).

3.4. Audio-visual rendering of sentiments

A more challenging goal is to visualise the density $\hat{p}(s, x, t)$ because the function $\hat{p}$ takes scalar values over a 4 dimensional space (i.e. $\dim s = 1$, $\dim t = 1$, $\dim x = 2$). To deal with the temporal domain, we propose an animated sequential rendering. The density value $\hat{p}$ of $\hat{p}(s, x, t)$ is not presented, but instead we show the sentiment score s encoded with different colours and sound. Figure 4 shows a still of our interactive and animated map (*Dublin Marathon 2014* dataset). A clock is initialised at 8am and progresses through the day of the marathon. Sentiment scores are computed automatically from tweets using the Stanford Core NLP library [7, 8]. Posted tweets appear on the map as coloured bubbles accompanied with a sound encoding the sentiment of the text of the tweets (e.g. green high pitch for happy, and red low pitch for sad, and black for neutral). The map is also interactive as each bubble is clickable to visualise the posted document.

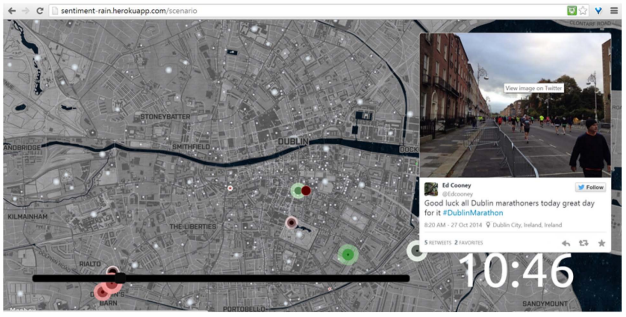


Fig. 4. Tweet Sentiment analysis with deep learning for *Dublin marathon 2014*.

The *Dublin Marathon 2014* dataset, is stored in an Orientdb database on our server [15], that is accessed live. Our

webdemo is available online[2] and Heroku is used for hosting (see Figure 5). While this is done live in our demo, once computed, sentiment scores could be stored in our database to avoid latencies.

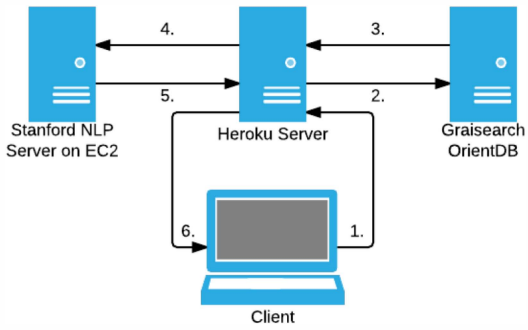


Fig. 5. Achitecture for our web based demonstrator http://sentiment-rain.herokuapp.com/ with dataset *Dublin Marathon 2014* hosted on our GRAISearch server http://graisearch.scss.tcd.ie/.

4. CONCLUSION AND FUTURE WORKS

We have proposed several ways of visualising datasets harvested from social media. These datasets are used for estimating probability density estimates that are then rendered using various techniques such as overlaying information over a map, using colour and sound, displayed in a sequential manner and also providing interactivity to visualise documents in the datasets. Note that this can be applied to any dataset having observations captured in the space or spacetime domains. The techniques presented in this paper are not specific to datasets harvested from social media, or for visualising sentiment s. Indeed similar ideas could be used for visualising information collected from distributed sensors measuring environmental factors for instance (e.g. pollution or noise level).

Beyond visualisation, our interest is in analysing these datasets and in providing models to explain these observations. For instance, marathons are sporting events that often occur in various cities around the world. These events can be disruptive (e.g. transportation network) and its planning can be demanding to insure the best experience possible for people participating in the event (e.g. runners, supporters, security and medical staff for the event) and the public. Analysing datasets harvested from social media may help in getting feedbacks from such events for best preparing the next ones. Similarly collecting information from social media on a particular location such as Trinity College may help in understanding how the different populations (e.g. students, staff, tourists) use the space over time. Future work will investigate using information theory [16] as a mathematical framework to summarise and compare our KDEs for giving insights into these datasets.

[2] `http://sentiment-rain.herokuapp.com/`

Acknowledgements

This work has been supported by the European project GRAISearch FP7-PEOPLE-2013-IAPP (612334), 2014-2018.

5. REFERENCES

[1] Taylor Shelton, Ate Poorthuis, Mark Graham, and Matthew Zook, "Mapping the data shadows of hurricane sandy: Uncovering the sociospatial dimensions of big data," *Geoforum*, vol. 52, no. 0, pp. 167 – 179, 2014.

[2] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, pp. 1012–1014, 2009.

[3] T. Althoff, D. Borth, J. Hees, and A. Dengel, "Analysis and forecasting of trending topics in online media streams," in *Proceedings of the 21st ACM International Conference on Multimedia*, New York, NY, USA, 2013, MM '13, pp. 907–916, ACM.

[4] M. R. Frank, L. Mitchell, P. Sheridan Dodds, and C. M. Danforth, "Happiness and the pattern of life: A study of geolocated tweets," *Scientific Reports*, vol. 3, 2013.

[5] S. Siersdorfer, E. Minack, F. Deng, and J. Hare, "Analyzing and predicting sentiment of images on the social web," in *Proceedings of the International Conference on Multimedia*, New York, NY, USA, 2010, MM '10, pp. 715–718, ACM.

[6] Z. Yuan, J. Sang, C. Xu, and Y. Liu, "A unified framework of latent feature learning in social media," *IEEE Transactions on Multimedia*, vol. 16, no. 6, October 2014.

[7] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2014, pp. 55–60.

[8] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2013.

[9] Stephen E. Fienberg, "Graphical methods in statistics," *The American Statistician*, vol. 33, no. 4, pp. 165–178, 1979.

[10] L. Marek, P. Tuček, and V. Pázto, "Using geovisual analytics in google earth to understand disease distribution: a case study of campylobacteriosis in the czech republic (20082012)," *International Journal of Health Geographics*, vol. 14, no. 7, 2015.

[11] T. Hengl, *A Practical Guide to Geostatistical Mapping*, 2009.

[12] C. Healey, "Sentiment viz: Tweet sentiment visualization," accessed 2015-07-06, `http://www.csc.ncsu.edu/faculty/healey/tweet_viz/tweet_app/`.

[13] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall, 1986.

[14] "Google maps javascript api,`https://developers.google.com/maps/documentation/javascript/heatmaplayer`," accessed 2015-07-06.

[15] Z. Zdziarski, J. Mitchell, P. Houdyer, D. Johnson, C. Bourges, and R. Dahyot, "An architecture for social media summarisation," in *Irish Machine Vision and Image Processing Conference*, Derry-Londonderry, Northern Ireland, 27-29 August 2014, pp. 187–188.

[16] F. Escolano, P. Suau, and B. Bonev, *Information theory in Computer Vision and Pattern Recognition*, Springer, 2009.