Moving on Twitter: Using Episodic Hotspot and Drift Analysis to Detect and Characterise Spatial Trajectories [*]

Hansi Senaratne
University of Konstanz
Universitätsstr. 10
Konstanz, Germany
hansi.senaratne@uni-konstanz.de

Arne Bröring
Esri Suisse
Josefsstr. 218
Zurich, Switzerland
a.broering@esri.ch

Tobias Schreck
University of Konstanz
Universitätsstr. 10
Konstanz, Germany
tobias.schreck@uni-konstanz.de

Dominic Lehle
University of Konstanz
Universitätsstr. 10
Konstanz, Germany
dominic.lehle@uni-konstanz.de

ABSTRACT

Today, a tremendous source of spatio-temporal data is user generated, so-called volunteered geographic information (VGI). Among the many VGI sources, microblogged services, such as Twitter, are extensively used to disseminate information on a near real-time basis. Interest in analysis of microblogged data has been motivated to date by many applications ranging from trend detection, early disaster warning, to urban management and marketing. One important analysis perspective in understanding microblogged data is based on the notion of *drift*, considering a gradual change of real world phenomena observed across space, time, content, or a combination thereof.

The scientific contribution provided by this paper is the presentation of a systematic framework that utilises on the one hand a Kernel Density Estimation (KDE) to detect hotspot clusters of Tweeter activities, which are episodically sequential in nature. These clusters help to derive spatial trajectories. On the other hand we introduce the concept of *drift* that characterises these trajectories by looking into changes of sentiment and topics to derive meaningful information. We apply our approach to a Twitter dataset comprising 26,000 tweets. We demonstrate how phenomena of interest can be detected by our approach. As an example, we use our approach to detect the locations of Lady Gaga's concert tour in 2013. A set of visualisations allows to analyse the identified trajectories in space, enhanced by optional overlays for sentiment or other parameters of interest.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Research

Keywords

Visual knowledge discovery, time varying data, geospatial visualisaition, feature detection & tracking

1. INTRODUCTION

Volunteered Geographic Information (VGI) as coined by Goodchild [10] is the process where citizens who are most often untrained and have varying expertise, collaboratively contribute spatially referenced information with the help of smart devices. These VGI are classified by Craglia et al. [8] as explicit-VGI and implicit-VGI based on the type of volunteering and the type of geography that is contributed. Examples of explicit-VGI are Openstreetmap or Wikimapia where citizens contribute mapped information explicitly with a spatial reference. Implicit-VGI are contributed on platforms such as Twitter, Flickr, or Wikipedia where these contributions implicitly refer to a location. E.g., on Twitter a user could tweet 'I'm going to see the Lady Gaga concert in *New York*', or the tweet could be 'I'm at the Lady Gaga concert' with an attached geotag as part of the Tweet's metadata. This implicit location information makes it challenging to derive location based services without including additional content. Spatial movement and trajectory detection using VGI is a pressing topic. Research in this area can lead to methods and technologies which are valuable for various applications ranging from marketing (e.g., how is the word about a new product spreading) to disaster management (e.g., what is the path of the hurricane). Thereby, the key benefit of using VGI in such applications is its real-time character.

Spatial trajectory is thereby defined by Zheng [20] as a trace generated by a moving object in geographical space, usually represented by a series of chronologically ordered points (p1, p2,... pn), where each point consists of a geospatial coordinate set and a time stamp such as p=(x,y,t). In relation to the explicit and implicit nature of VGI, trajectories that are derived from VGI can also be categorised into what we call here as *directly observed trajectories* and *indirectly observed trajectories*. Directly observed trajectories have an explicit spatial reference, e.g., tracking of a twittering parcel package as described by [6]), while indirectly observed trajectories have an implicit spatial reference, e.g., extracting the movement of a flood by analysing the information given by contributors on Twitter and other VGI sources as described by [9].

On microblogs, such as Twitter, indirectly observed trajectories can be observed by keyword filtering over a set of geotagged tweets. This simple technique can be utilised to derive context specific information based on the contributions of Twitter users. For example, taking all geotagged tweets of a particular day, e.g., 11th of February 2013, filtering them for a certain keyword, e.g., 'Lady Gaga', and finally visualizing the geotags on a map can indicate where a concert was held on that day, assuming closer to the location of the concert more tweets are published. However, to derive more reliable spatial trajectories this simple technique is not enough. To effectively detect and characterise spatial trajectories more sophisticated methods need to be applied.

In this work, we tackle the challenge of deriving trajectories more accurately based on the implicit location information of VGI. Therefore, our approach takes a pre-filtered set of geotagged Twitter microblogs for a certain time frame and for a specific keyword (e.g. 'Lady Gaga'). Then, we extract the 10 most often used keywords and select a keyword of interest for the use case (e.g., 'concert'). In the following, we filter the dataset for the chosen keyword and perform a Kernel Density Estimation (KDE) on this dataset. The KDE gives us an estimation of the hotspot clusters of tweeter activity over geographic space and time. These resulting keyword based hotspot clusters show a series of separate events that follow a logical order. This, which we term as *episodic sequential behaviour* helps us to derive spatial trajectories. Finally, we characterise the individual hotspot clusters and thereby derive meaning by looking closely at the *drifts* that occur through various modalities. *Drift* is here considered as a gradual change of a phenomenon and can be observed across space, time, content or a combination thereof. In the developed framework, we consider the drifts that can be observed in sentiment and topics at each of the hotspot clusters that occur over space and time. For the use case of this paper, we utilise this approach to detect the trajectories of a Lady Gaga concert tour in 2013.

The remainder of this paper is organised as follows. In the next section we discuss the related works on trajectory detection on VGI. Following up in Section 3 is a description of our proposed framework for detecting and analysing trajectories in the place of drifts. In Section 4 we present a use case application scenario that implements our framework. Section 5 presents conclusions and future outlook for our approach.

2. RELATED WORK

With the Web 2.0 in place, humans can be considered as virtual sensors who are able to collect and contribute spatially referenced data in the form of maps, text, audio, or video on the Web, making the consumers of data also the producers (VGI). O'Reilly [13]

described such user generated content as the *wisdom of the crowds* thereby emphasising the potential of the data. Surowiecki [18] supports this statement by showing how a group of people may contribute to a solution of a problem that an expert may be unable to solve.

Various researchers utilised such data for movement detection. For example, Sakaki et al. [17] used a classifier that considered features such as keywords, number of words, and the context to approximate the trajectory of a moving Typhoon via Twitter. They utilised a particle filtering to assess the geographic locations of the typhoon path with a weighted average of latitudes and longitudes, and median as a baseline.

Another approach [15] used distance functions to determine the similarity between multiple trajectories, and further introduced a progressive clustering technique which was applied to analyse large sets of trajectories. Fuchs et al. [9] demonstrated how Twitter can be used in combination with other social media data sources and mobile network metrics to determine events that occur through space and time. Andrienko et al.[5] classified position recordings to determine the location of moving phenomena. In [1] and [3], methods were developed to use VGI for exploring the interests, behaviour, and mobility of people. They presented a conceptual framework that allowed the aggregation of movement data, with a focus on situation-oriented and trajectory-oriented movement data. Their research emphasised the importance of aggregating movement data for supporting visual exploration, and [2] discussed the need for appropriate visualisation methods to analyse such movement data. In [4], Andrienko et al. constructed trajectories of Twitter users from tweeting locations by computing the trajectory medoid (i.e., the cluster point of a dataset whose average dissimilarity to all objects in the cluster is minimal) for each spatially referenced tweet.

In this paper, our contribution is a systematic framework that utilises a KDE to observe episodic sequential hotspot clusters of tweeter activity, and uses the concept of drifts through various modalities (sentiment and topic) to characterise the trajectory that results from these episodic hotspot clusters.

3. A FRAMEWORK FOR DETECTING MEANINGFUL TRAJECTORIES IN MICROBLOGGED DATA

Detecting (indirectly observed) spatial trajectories on Twitter is not straightforward. Methods are needed to characterise the approximated trajectories, and thereby support the trajectory analysis process. In our approach we show how we utilise a KDE to detect episodic sequence of hotspot clusters that is used to derive our indirectly observed trajectories, and later use a sentiment and topic drift analysis to characterise these episodic hotspot clusters.

The Figure 1 shows the developed framework for detecting and characterising trajectories. The first step of pre-processing includes the filtration of the Twitter stream to include only those tweets that are geotagged, so that all Tweets taken into account have a spatial reference. These geotags are recorded in the metadata of the tweets, where the device in-built GPS captures the location of the contributor if the location function is turned on. In the next filtration step we choose a time frame for which we want to extract the geotagged tweets. After this stage we run a Term Frequency-Inverse Document Frequency (tf-idf) analysis [14] on the extracted Tweets with

a stop word list specifically for tweets to generate a pre-selection of the frequently used keywords in the dataset. These keywords are then ranked based on their frequency count. After sampling the data as necessary, the user can proceed further by choosing a keyword to generate the Kernel Density Estimation (KDE).

In the following, we describe the two key parts of our *Moving on Twitter* framework: Section 3.1 describes the detection of trajectories and Section 3.2 describes the characterisation of those trajectories. Figure 2 shows our implementation of this framework as a tool that allows users to visually analyse Twitter data sets.

3.1 Detection of Trajectories through Hotspot Cluster Analysis

Contrary to point data mapping which focuses on mapping the location of individual events, hotspot mapping focuses on highlighting areas which have higher than average incidence of events. These hotspot areas can exist in different scales of interest. In order to estimate these hotspots of events corresponding to the chosen keyword, we utilise a smooth, continuous, and differentiable Gaussian Kernel Density Estimation (KDE) at each time step which in principle creates a surface based on the distribution and density of Twitter message geotags. We utilise a heatmap visualisation to display the results of the KDE.

At the next step of the framework, the user is optionally able to specify certain Twitter metadata as parameters, such as status count, follower count, list count, or friend count of Twitter users (e.g., the user can filter out the tweets from popular contributors who have higher numbers of followers) (Figure 2 top left). We chose these parameters as they were found to be the top ranked credibility indicators by Castillo et al. [7], with the help of which the user can derive more trustworthy tweets. These parameters are weighted based on their impact as found by [7], and can be used to smooth a trajectory.

After determining the hotspots at multiple time steps, the episodic sequential hotspots can be observed once visualised in a heat map. Those resulting hotspots can be connected to represent a trajectory. To connect these sequential hotspot clusters, we first compute the average time at each hotspot cluster, based on the assumption that people tweet about an incident around the particular geographic region on the day of the incident. We connect these averaged times at each hotspot cluster sequentially. Then, we utilise arrow heads to show the direction of the trajectory based on the sequentially occurring hotspot clusters. These lines are coloured based on the avearge sentiment of the next occurring hotspot cluster (in a line going from time t1 at hotspot cluster1 to t2 at hotspot cluster2, the colour of the line would represent the sentiment at t2 and hotspot cluster2).

The framework explicitly allows for an iterative trajectory detection. I.e., if a user is not content with the results, the input parameters can be changed. Also, a user can refine the filtration step and adjust the specified time frame, or keywords as input to the KDE. This way, after multiple iterations, the detection of trajectories can be optimised.

3.2 Characterising Trajectories through Drift Analysis

Characterising trajectories and thereby deriving meaning from them is crucial for such spatial analysis. Therefore through the next step

of the framework the user is able to perform drift analyses within two selected modalities: sentiment and topic.

The sentiment drift analysis helps the user to determine the subjective opinion and the emotions of the contributors and how it changes across geographic space over time. The sentiment drift is visually analysed in our approach by linking a word cloud with the above described map visualizations. Therefore, we compute the average sentiment for each hotspot cluster. To achieve this we first classify each tweet with the help of Sander's annotated Twitter data set which has been evaluated by [16] for significant results. By using this dataset, we trained a classifier using the LingPipe Java toolkit[1] which uses computational linguistics for processing the text. The polarity of these tweets were annotated as either positive, negative or neutral sentiments. To obtain the collective sentiment of each hotspot cluster we averaged the sentiments of each tweet belonging to the hotspot clusters. We visualise these sentiments using different colours (Red for negative sentiments, Yellow for neutral sentiments, Green for positive sentiments).

The topic drift analysis helps the user to identify the most relevant keywords of the tweets without having to read all the tweets. This enables a quick visual analysis of huge amounts of tweets. This part of our approach is related to the Nokia Internet Pulse [12] and Wordle [19] techniques to represent the most relevant keywords in a time frame. However, contrary to [12] where a vertical axis was used, we use a word cloud to represent the keywords in a time frame chosen by the user. This leaves the angles of the keywords in the word cloud and their colour to represent additional features as described below.

The word cloud created as part of our approach[2] shows the most occurring keywords during the chosen time frame. The positions of the words in the word cloud is based on Wordle, which positions the words at a random starting point, and if overlaps occur then the word is moved a step along an increasing spiral. This is repeated until no overlaps occur. The word cloud is drawn at the end when all the words have been positioned (Figure 6). The shown keywords of the word cloud are generated from all tweets that appear in that time frame, by filtering out a list of stop words. We map the average sentiment on a colour scale ranging from Red, over Yellow to Green to indicate the varying emotions of Tweeters regarding the specific keywords. Further, we use the font size to map the number of occurrences of a keyword.Consequently, the most occurring words appear larger than the least occurring words. In addition, we map the optional parameters which are the credibility indicators to the angle of the keywords in the word cloud. The rotation angles are between $0°$implying higher credibility and $90°$implying lower credibility. This design choice was taken to communicate higher credible keywords with ease (when it is more horizontal) and lower credible words with lesser ease (when the keywords are dangling at an angle). The angle of these keywords with credibility is calculated based on a logarithmic scale.

4. APPLICATION OF THE 'MOVING ON TWITTER' FRAMEWORK TO DETERMINE A LADY GAGA CONCERT ROUTE

---

[1]http://alias-i.com/lingpipe/

[2]We used the D3.js library http://www.jasondavies.com/wordcloud/ and combined it with the algorithms in [19] to visualise the tag cloud.

Figure 1: Design of the framework to detect meaningful trajectories.

Figure 2: The tool implementing the designed framework to visually analyse Twitter datasets.

We implemented the systematic framework presented in Section 3 as an HTML5 application (see Figure 2) based on Bootstrap[3] and additional JavaScript libraries, such as D3.js[4], jQuery[5], and the Google Maps API[6]). For storing and processing of large amounts of Tweets we used a MySql[7] database where we store the data with all its metadata.

As a proof of concept for the developed framework, we applied this implemented tool in order to determine a concert route of pop music artist *Lady Gaga*. This artist has over 41.2 Million followers[8] on Twitter, and is known for controversial performances that constantly make headlines on Twitter as well as in other social media platforms. She planned a North American tour between 11.1.-16.3 of 2013. We chose a dataset that represented the tour on Twitter in order to determine her tour trajectory based on what Twitter contributors had to say.

In preparation of the dataset, we initially filtered the Twitter stream[9] for geotagged Tweets, based on the keywords 'lady gaga' and 'ladygaga', as well as for the selected time frame. This resulted in 26,000 tweets that contained any term referring to Lady Gaga. We consider this as sufficient for our initial analysis. We then generated a pre-selection of top-ten most frequently used keywords as a result of a *tf-idf* analysis on the extracted Tweets with a stop word list specifically for tweets. The ranked keywords resulting from this term frequency calculation were: 'Artpop', 'LadyGaga', 'concert', 'Starlight', 'nowplaying', 'Brazil', 'KEPO', 'Center', 'Rihanna', 'show'. For our analysis we chose the keyword 'concert', and therefore the dataset is filtered on a second round based on this selected keyword.

To sample our dataset we associate each Tweet with the geographic coordinates of the closest larger city which has a population greater than 100,000. This served as a pre-clustering and to remove noise which is necessary for the next step, the KDE.

4.1 Hotspot Analysis

The gathered dataset has a temporal resolution of one day, as there was a minimum time gap of one day between each performance. Further it has a spatial resolution at city level, as the artist performed in different cities in North America on each day of her tour. We ran the KDE based on a Gaussian distribution for the geotagged Tweets point data at each time step (day). The resulting hotspot clusters were visualised as a heat map layer for every day on top of a geographic map. As you can see in resulting map visualisation (Figure 4), higher activity of tweeters are clustered around particular locations.

A time slider helps to navigate through the produced heatmaps at every day of the tour. The user can change the time window of the time slider if she wishes to change the temporal resolution. Time steps are weighted differently on the heat map to enhance the observations of the currently selected visualised time step (i.e., tweets

_____

from previous days have lower weighting).

The resulting KDE for every day already indicates an episodic sequence of hotspot clusters, and we believe these hotspot clusters are an approximation for the cities where the concert took place during the tour as they show the highest densities.

As an alternative to KDE, we clustered the data using the density-based clustering algorithm DBSCAN, where the user can specify the noise in terms of the radius (in km) and the minimum points in a cluster. This allowed us to find the dense regions on top of the kernel density estimation. The resulting clusters were similar to the KDE hotspot analysis results, with this confirmation we continued using KDE for our analysis.

Next, for each of these hotspot clusters, we computed the average time, based on the assumption that people tweet about the concert around that particular geographic region on the day of the concert. We connect these averaged times at each hotspot cluster sequentially, to obtain what represents a trajectory.

Figure 5 shows the actual route of the tour (in Black colour) overlayed on top of the approximated trajectory resulting from our approach. The actual route as well as the determined trajectory start in Las Vegas and correspond in several cities, such as Houston, St. Louis, and Toronto.

4.2 Sentiment and Topic Drift Analysis

The sentiment drift analysis between two subsequent days revealed an interesting finding from our application use case. That was the cancellation of the tour before the performance in Chicago, which was due to the artist's health concerns. Figure 6 shows the topic and sentiment drifts between two days of the concert. The word cloud on the left shows the keyword and sentiment trend before the announcement of the tour cancellation, and the word cloud on the right shows the keyword and sentiment trend after the cancellation announcement. The concert mood has clearly changed from one day to the other.

5. CONCLUSIONS AND FUTURE WORK

The contribution of this paper is a systematic framework for *indirectly observed trajectory* detection in microblogged data using episodic sequential hotspot and drift analysis. As a proof of concept of this framework, we demonstrated a trajectory approximation within geotagged Twitter data for a concert tour by the pop artist Lady Gaga. Realising this use case by following the developed framework, we first filtered and sampled the data stream to perform a hotspot analysis using Kernel Density Estimation over the geotagged Tweets. This step gave us approximations for the cities where the artist performed her concerts. A word cloud visualisation linked with the display of determined hotspots provided further insights and allows to do a sentiment and topic drift analysis. Visually analysing the spatio-temporal drift of the tweet hotspots over averaged times indicated an approximation of the tour path. We could characterise the approximated trajectory of the concert tour, and reveal additional phenomena such as the cancellation of the tour, indicated by negative sentiments.

The developed tool implementing our *Moving on Twitter* framework enables the user to flexibly perform these analyses for different keywords and varying parameter settings to visually explore indirectly observed trajectories in Twitter datasets. Our framework is not limited to be applied in this tool, but it is generic and defines

Figure 3: Clustered routes before averaging the time. The arrows indicate the sequential direction from one hotspot to the other.
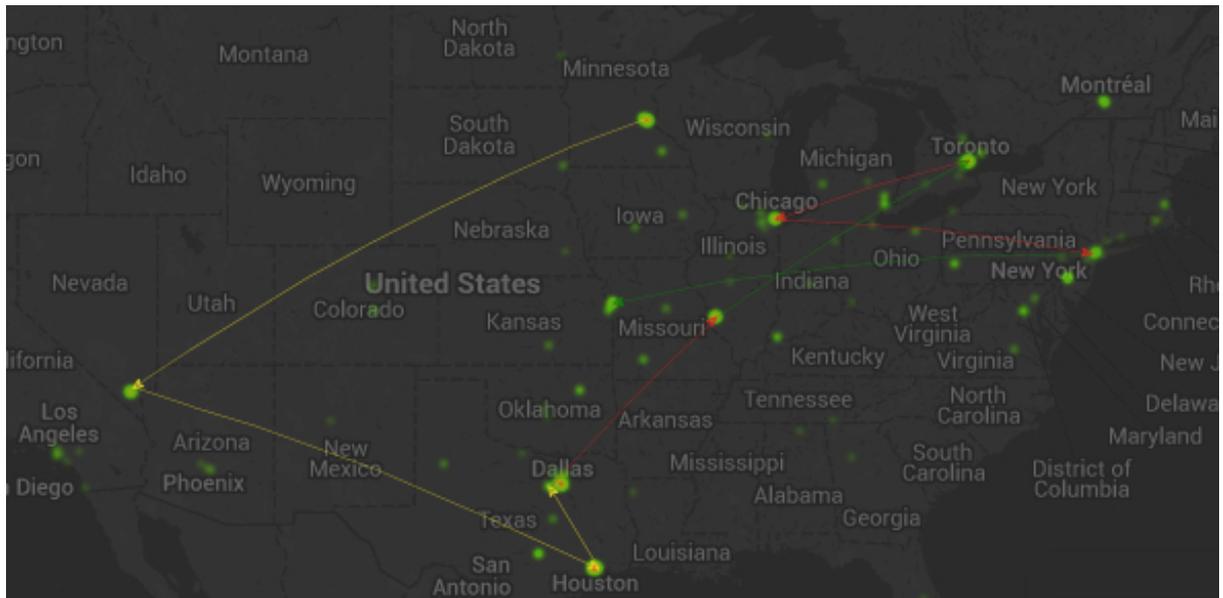

Figure 4: The kernel density estimation of the Tweets relating to the Lady Gaga concert tour, and the clustered routes after averaging the time. Cities (e.g., Las Vegas, Dallas, Houston, or Toronto) where the concert took place are already visible as hotspots. The arrows in the clustered routes indicate the sequential direction from one hotspot to the other. The colours of the trajectories depict the average sentiments from the respective hotspots.
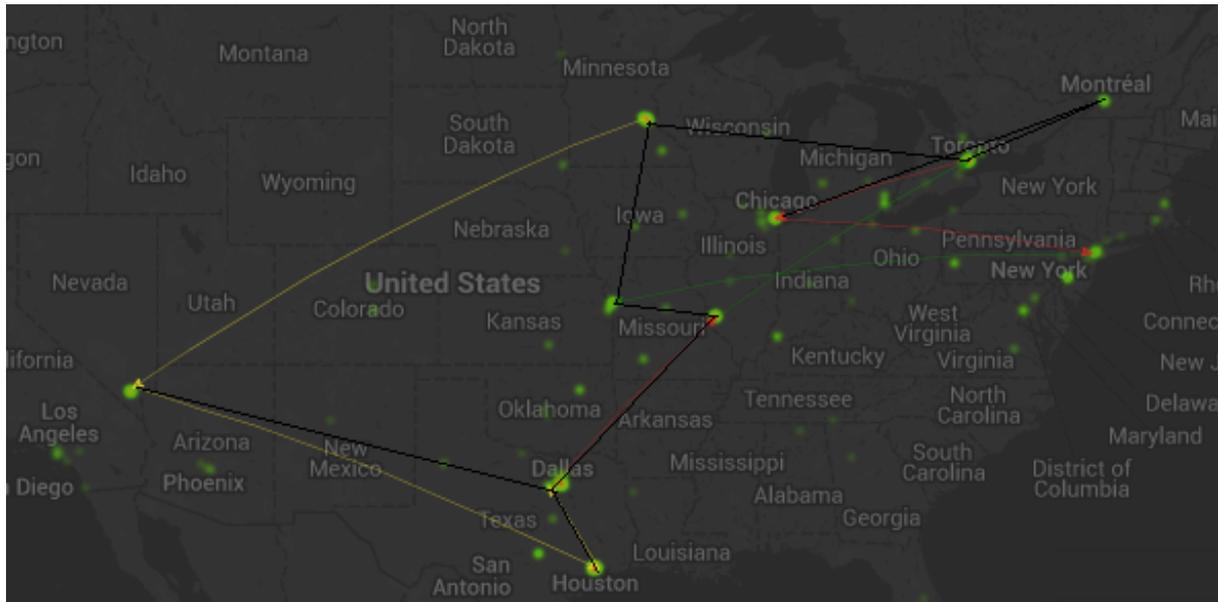
Figure 5: Actual route (in Black colour) over the approximated trajectory (in Green, Yellow, and Red colour that depicts the averaged positive, neutral and negative sentiments from the respective hotspots). The route indicates the following concerts: Las vegas (NV) on 25.01., Dallas (TX) on 29.01., Houston (TX) on 31.01., St. Louis (MO) on 02.02., Kansas city (MO) on 04.02., St. Paul (MN) on 06.02., Toronto (ON) on 08.02., and Montreal (QC) on 11.02., before the concert got cancelled for the remaining leg of the tour starting from Chicago (IL) which was supposed to take place on the following 13.02.
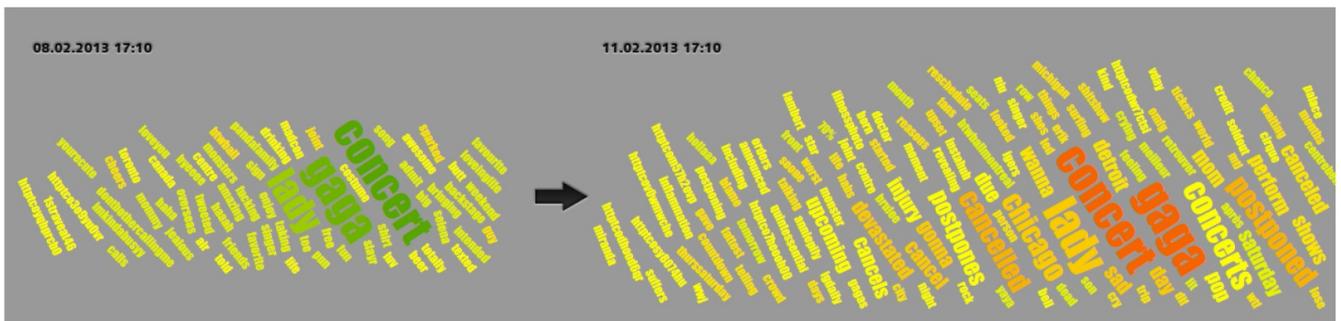

Figure 6: The topic and sentiment drift between two days. This example shows how the mood of Tweeters have changed between these two days, which reveals the cancellation of the concert in Chicago midways of the tour. Notice the words 'cancelled', 'devastated', 'chicago', 'postponed' that all may refer to the cancellation of the tour starting from a planned concert in Chicago

clear extension points for more comprehensive trajectory detection in the future.

Through our proof of concept implementation, we have experienced limitations that we will cope with next. To start with, in the sampling step within our usecase we have mapped tweets to the closest city with over 100,000 population in order to even out the population of tweeters, and also thereby reduce noise and random talks that occur on Twitter. In future we will improve this step by taking the Tweets per population ratio at different location. This will also help us to avoid heavy clustering around particular locations. The presented trajectory is still an approximation, and we need to further improve this method, e.g., with edge-bundling methods (e.g. [11]) to refine the detected trajectories. Also, we have not yet implemented the last part of the model, which is an automated filtering and ranking of candidate trajectories in order to compute an interestingness measure. These will be tackled with variance measures or euclidean distance measures. Further, the implementation of our tool will be adjusted and improved to enable visual exploration at different spatial and temporal resolutions within other datasets. In addition we will explore and visualise the other types of drift, such as text similarity drift, and we will look into the different structures of trajectories that can be found in microblogged data (such as cyclic and back and forth). As a final remark, the tool will be evaluated in future to test its usability within appropriate end users, and further validate the methods we have used to derive and characterise the trajectory.

6. REFERENCES

[1] G. Andrienko and N. Andrienko. A general framework for using aggregation in visual exploration of movement data. *The Cartographic Journal*, 47(1):22–40, 2010.

[2] G. Andrienko, N. Andrienko, P. Bak, D. Keim, and S. Wrobel. *Visual analytics of movement*. Springer Publishing Company, Incorporated, 2013.

[3] G. Andrienko, N. Andrienko, P. Bak, S. Kisilevich, and D. Keim. Analysis of community-contributed space-and time-referenced data (example of flickr and panoramio photos). In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 213–214. IEEE, 2009.

[4] G. Andrienko, N. Andrienko, H. Bosch, T. Ertl, G. Fuchs, P. Jankowski, and D. Thom. Thematic patterns in georeferenced tweets through space-time visual analytics. *Computing in Science & Engineering*, 15(3):72–82, 2013.

[5] N. Andrienko, G. Andrienko, N. Pelekis, and S. Spaccapietra. Basic concepts of movement data. In *Mobility, Data Mining and Privacy*, pages 15–38. Springer, 2008.

[6] A. Bröring. Live and web-based parcel monitoring with low-cost sensors. In D. Vandenbroucke, B. Bucher, and J. Crompvoets, editors, *Online Proceedings of the 16th AGILE International Conference on Geographic Information Science. Leuven, Belgium. 15.-17. May 2013*, 2013.

[7] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, pages 675–684. ACM, 2011.

[8] M. Craglia, F. Ostermann, and L. Spinsanti. Digital earth from vision to practice: making sense of citizen-generated content. *International Journal of Digital Earth*, 5(5):398–416, 2012.

[9] G. Fuchs, N. Andrienko, G. Andrienko, S. Bothe, and H. Stange. Tracing the german centennial flood in the stream of tweets: first lessons learned. In *Proceedings of the Second ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information*, pages 31–38. ACM, 2013.

[10] M. F. Goodchild. Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007.

[11] C. Hurter, O. Ersoy, S. Fabrikant, T. Klein, and A. Telea. Bundled visualization of dynamic graph and trail data. *IEEE Transactions on Visualization and Computer Graphics*, 2007.

[12] J. Kaye, A. Lillie, D. Jagdish, J. Walkup, R. Parada, and K. Mori. Nokia internet pulse: a long term deployment and iteration of a twitter visualization. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pages 829–844. ACM, 2012.

[13] T. O'Reilly. *What Is Web 2.0 - Design Patterns and Business Models for the Next Generation of Software*. O'Reilly Media, September 2005. Online available: http://oreilly.com/web2/archive/what-is-web-20.html; last accessed 03/2014.

[14] O. Phelan, K. McCarthy, and B. Smyth. Using twitter to recommend real-time topical news. In *Proceedings of the Third ACM Conference on Recommender Systems*, pages 385–388, New York, NY, USA, 2009. ACM.

[15] S. Rinzivillo, D. Pedreschi, M. Nanni, F. Giannotti, N. Andrienko, and G. Andrienko. Visually driven analysis of movement data by progressive clustering. *Information Visualization*, 7(3-4):225–239, 2008.

[16] H. Saif, M. Fern, Y. He, and H. Alani. Evaluation datasets for twitter sentiment analysis a survey and a new dataset, the sts-gold. 2013.

[17] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[18] J. Surowiecki. *The wisdom of crowds*. Random House LLC, 2005.

[19] F. B. Viegas, M. Wattenberg, and J. Feinberg. Participatory visualization with wordle. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):1137–1144, 2009.

[20] Y. Zheng and X. Zhou. *Computing with spatial trajectories*. Springer, 2011.