CHAPTER 12

Gaining Knowledge from Georeferenced Social Media Data with Visual Analytics

Gennady Andrienko* and Natalia Andrienko
Fraunhofer Institute IAIS, Sankt Augustin, Germany, and City University London, UK
*gennady.andrienko@iais.fraunhofer.de

Abstract

Analysis of the collections of geographically referenced posts published in social media, such as Twitter, Flickr, and YouTube, can bring new knowledge about places, geographical objects, and events interesting to people, and about people's mobility behaviours. Gaining knowledge from large data collections requires combining computational analysis with human interpretation, judgement, and reasoning, which, in turn, require appropriate visual representations of the data and analysis results. Visual analytics integrates computational analysis techniques with interactive visual interfaces to support collaborative human–computer analytical activities. We give a brief overview of visual analytics approaches to extracting various kinds of information and knowledge from georeferenced social media data.

Keywords

Visual Analytics, Social media analysis, trajectories, movement data, temporal data, spatio-temporal clusters

How to cite this book chapter:
Andrienko, G and Andrienko, N. 2016. Gaining Knowledge from Georeferenced Social Media Data with Visual Analytics. In: Capineri, C, Haklay, M, Huang, H, Antoniou, V, Kettunen, J, Ostermann, F and Purves, R. (eds.) *European Handbook of Crowdsourced Geographic Information*, Pp. 157–167. London: Ubiquity Press. DOI: http://dx.doi.org/10.5334/bax.l. License: CC-BY 4.0.

Introduction

Microblogging services, such as Twitter, and services for sharing photo and video, such as Flickr and YouTube, allow the users to supply their posts with geographic coordinates. The high popularity of these services in conjunction with the widespread proliferation of devices capable of providing location information has led to great and constantly increasing volumes of location- and time-referenced data produced by myriads of users. By analysing these data, it is possible to extract interesting new information about various places and events as well as about people's interests, mobility behaviours, and life styles.

Analysis of social media data is currently a popular topic in visual analytics, a research discipline that aims to support synergistic human–computer analytical workflows by combining computational analysis techniques with interactive visual interfaces supporting human interpretation, judgement, and reasoning (Keim et al. 2010). We give a brief overview of the published literature that describes visual analytics approaches to extracting different kinds of information from georeferenced social media data. Most of the works do not focus on extracting a single type of information but deal with several types.

Analysis of georeferenced photo data

The photos published at Flickr, Panoramio, and other photo sharing services are supplied with metadata, which include the dates and times of the shots and may also include titles and/or text tags indicating the contents of the photos. For many photos, the metadata include the coordinates of the locations where the photos had been taken. Collections of metadata records including geographic coordinates were analysed in multiple ways according to the possible analysis foci (space and place or people) and respective tasks (Andrienko et al. 2009). The photo data were considered from two distinct perspectives: as spatial events (independent points in space and time) and as trajectories of people (i.e. of the photo authors).

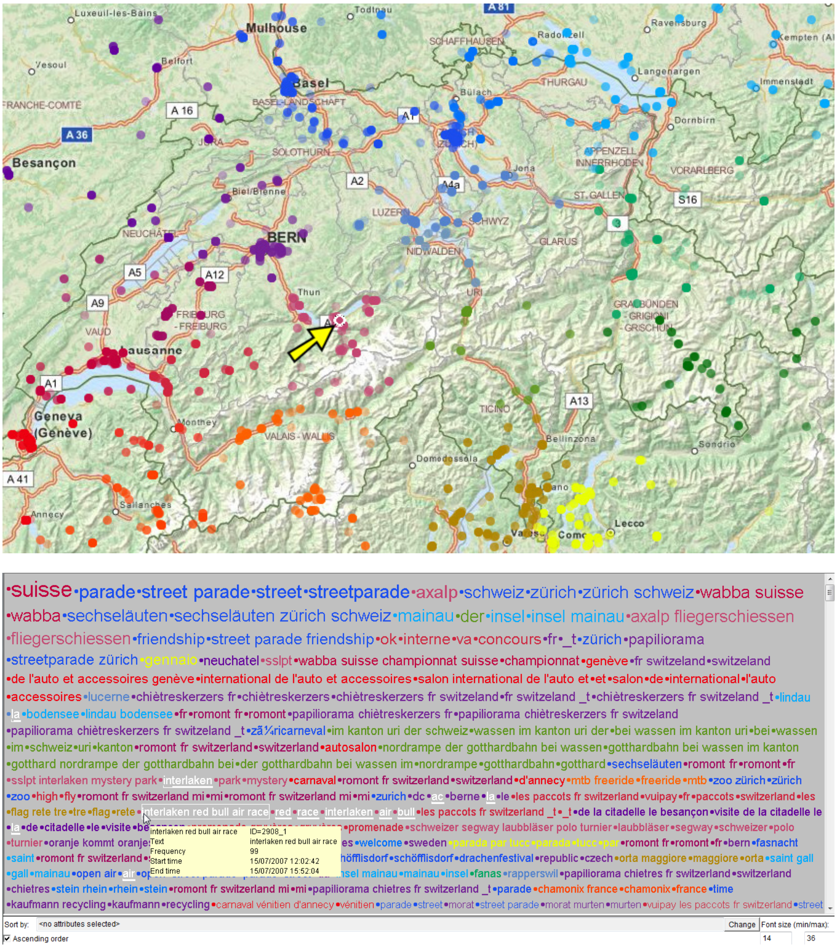
Analysing photo taking events

In analysing the data as spatial events, spatial density-based clustering was used for identifying popular places attracting much attention of the photo authors. Visualisation of the times when the photos had been taken in these places revealed different seasonal patterns of the place visits. To study the spatial distribution of the photos over a territory and compare the temporal patterns of visiting different parts of it, the territory is divided into compartments, e.g. by a regular (Andrienko et al. 2009) or irregular (Jankowski et al. 2010; Andrienko et al. 2012) grid, and the photo taking events are aggregated by these compartments and time intervals. The resulting time series of the event counts are visualised on a map (Andrienko et al 2009) or on a time graph (Jankowski et al. 2010; Andrienko et al. 2012), which is linked to a map display through interactive techniques, including synchronous highlighting, selection, and filtering of corresponding visual objects. By analysing the time series using either mostly interactive (Jankowski et al. 2010) or computationally supported (Andrienko et al. 2012) techniques, the researchers detected places with interesting temporal patterns of visits, such as periodic peaks at particular times of the year, very high irregularly occurring peaks, and significant increase of place popularity starting from a particular time. To understand the reasons for these patterns, the researchers extracted frequently occurring words and word combinations from the titles of the photos that had been taken in the places and times of the peaks or sudden increases of attendance. In most cases, the extracted words referred to various public events (festivals, open-air shows and concerts, etc.), but also to interesting natural phenomena, such as cherry tree blossoming or abundant snowfalls. A different approach to identifying public events and other happenings attracting people's attention is by using spatio-temporal clustering of the photo taking events (section 6.2.3 of Andrienko et al. 2013a) which finds occurrences of multiple photos taken closely in space and time, i.e. spatio-temporal clusters. For the clusters, frequently occurring words and word combinations are extracted and investigated using a text cloud display linked to a map (Figure 1).

Sections 7.2.1-7.2.5 of the book Andrienko et al. (2013a) present an example of an in-depth analysis of time series of the presence of distinct photographers by regions of Switzerland. The analysis includes, among other techniques, visually supported clustering of the time series and interactive generation of models for predicting the number of photographers that can be expected to visit the regions in the future at different times of a year. The time series can also be viewed from a different perspective: as a sequence of spatial distributions of the photographers' presence in different time intervals. To study the temporal patterns of the occurrence of similar and dissimilar spatial distribution patterns, the distributions are clustered by similarity, summarized by the resulting clusters, and compared using multiple map displays and special interactive operations supporting comparisons (section 8.1.1 of Anrienko et al. 2013a). The temporal distribution of the clusters is visually represented on temporal displays. The provided example demonstrates how the analysis reveals an interaction between temporal periodicity and temporal trends in the sequence of the spatial distributions of the presence of Flickr photographers over the territory of Switzerland.

Analysing trajectories of photo authors

Trajectories of people can be constructed from georeferenced photo data by arranging the records of each individual photographer in a chronological


Luxeuil-les-Bains
Mulhouse
Todtnau
A 81
Ravensburg
Kempten (Al
Vesoul
Belfort
Basel
SCHAFFHAUSEN
Radolfzell
Langenargen
Immenstadt
FRANCHE-COMTÉ
A 16
BASEL-LANDSCHAFT
Bülach
THURGAU
Dornbirn
A 36
A1
Besançon
JURA
SOLOTHURN
ZÜRICH
APPENZELL INNERRHODEN
VORARLBERG
Jona
A2
Biel/Bienne
A4a
ST. GALLEN
S16
LUZERN
NEUCHÂTEL
BERN
SCHWYZ
NIDWALDEN
GLARUS
A5
A12
A9
Thun
URI
GRAUBÜNDEN - GRIGIONI - GRISCHUN
FRIBOURG - FREIBURG
VAUD
Lausanne
A1
TICINO
A13
Geneva (Genève)
Monthey
VALAIS - WALLIS
Bellinzona
Sondrio
Domodossola
A 41
Lugano
Sallanches
Annecy
Lecco
Varese
Como
•suisse•parade•street parade•street•streetparade•axalp•schweiz•zürich•zürich schweiz•wabba suisse
•wabba•sechseläuten•sechseläuten zürich schweiz•mainau•der•insel•insel mainau•axalp fliegerschiessen
•fliegerschiessen•friendship•street parade friendship•ok•interne•va•concours•fr•_t•zürich•papiliorama
•streetparade zürich•gennaio•neuchatel•sslpt•wabba suisse championnat suisse•championnat•genève•fr switzeland•switzeland
•de l'auto et accessoires genève•international de l'auto et accessoires•salon international de l'auto et•et•salon•de•international•l'auto
•accessoires•lucerne•chiètreskerzers fr•chiètreskerzers•chiètreskerzers fr switzeland•fr switzeland _t•chiètreskerzers fr switzeland _t•lindau
•la•bodensee•lindau bodensee•fr•romont fr•romont•papiliorama chiètreskerzers fr•papiliorama chiètreskerzers fr switzeland
•papiliorama chiètreskerzers fr switzeland _t•zã¼ricarneval•im kanton uri der schweiz•wassen im kanton uri der•bei wassen im kanton uri•bei•wassen
•im•schweiz•uri•kanton•romont fr switzerland•switzerland•autosalon•nordrampe der gotthardbahn bei wassen•gotthardbahn bei wassen im kanton
•gotthard nordrampe der gotthardbahn bei•der gotthardbahn bei wassen im•nordrampe•gotthardbahn•gotthard•sechseläuten•romont fr•romont•fr
•sslpt interlaken mystery park•interlaken•park•mystery•carnaval•romont fr switzerland•switzerland•d'annecy•mtb freeride•freeride•mtb•zoo zürich•zürich
•zoo•high•fly•romont fr switzerland mi•mi•romont fr switzerland mi•mi•zurich•dc•ac•berne•la•le•les paccots fr switzerland•vuipay•fr•paccots•switzerland•les
•flag rete tre•tre•flag•rete•interlaken red bull air race•red•race•interlaken•air•bull•les paccots fr switzerland _t•_t•de la citadelle le besançon•visite de la citadelle le
•la•de•citadelle•le•visite•be •promenade•schweizer segway laubbläser polo turnier•laubbläser•segway•schweizer•polo
•turnier•oranje kommt oranje es•welcome•sweden•parada par tucc•parada•tucc•par•romont fr•romont•fr•bern•fasnacht
•saint•romont fr switzerland• höfflisdorf•schöfflisdorf•drachenfestival•republic•czech•orta maggiore•maggiore•orta•saint gall
•gall•mainau•open air•air•op nsel mainau•mainau•insel•fanas•rapperswil•papiliorama chietres fr switzerland•switzerland
•chietres•stein rhein•rhein•stein•romont fr switzerland mi•mi•papiliorama chietres fr switzerland _t•parade•chamonix france•chamonix•france•time
•kaufmann recycling•kaufmann•recycling•carnaval vénitien d'annecy•vénitien•parade•street•morat•street parade•morat murten•murten•vuipay les paccots fr switzerland•street
Change

interlaken red bull air race ID=2908_1
Text interlaken red bull air race
Frequency 99
Start time 15/07/2007 12:02:42
End time 15/07/2007 15:52:04

Sort by: <no attributes selected>
☑ Ascending order
Font size (min/max): 14 36

Figure 1: Top: the frequent occurrences of words and combinations in the photo titles within spatio-temporal clusters of Flickr photos are represented on a map by point symbols coloured according to the spatial positions of the clusters. Bottom: the words and combinations are represented in a text cloud display, the font sizes being proportional to the frequencies and the colours corresponding to the spatial locations, as in the map. One of the word combinations ('Interlaken red bull air race') is selected in the text cloud view by mouse-pointing; the corresponding point is highlighted on the map (marked with an arrow). Source: Andrienko et al. 2013a.

sequence (the same idea applies to any kind of georeferenced data that include identifiers of individuals, in particular, to data from YouTube, Twitter, and other social media). Trajectories of individuals can be aggregated into flows between compartments of a territory division and visualised on flow maps to enable studying of mass movement patterns (Andrienko et al. 2009). The aggregation of trajectories into flows can be done by time intervals for studying seasonal differences between the mass movement patterns (Jankowski et al. 2010). A set of trajectories can also be analysed for discovering frequent sequences of place visits (section 7.3.4 of Andrienko et al. 2013a). The extracted frequent sequences can be explored using a text cloud display combined with an interactive map and a space-time cube. By analysing people's trajectories, one can also detect meetings of two or more individuals, including repeated meetings of the same pairs or groups of individuals, and joint trips of two or more photographers (Andrienko et al. 2009); however, performing such analyses may be unethical, as they may compromise the personal privacy of the individuals.

This overview gives an idea about the diversity of the possible approaches to analysing georeferenced photo data and the kinds of information and knowledge that can be extracted from such data. The same range of approaches is also applicable to georeferenced microblogging data, such as data from Twitter. The types of information that can be extracted from the two different sources of data are the same but the interpretation may be different. Thus, people mostly take photos when they encounter interesting places, objects, or events; besides, not all taken photos but only the best or the most interesting ones may be published. It should also be taken into account that photos are rarely taken in low light conditions, and that there are situations and places in which taking photos is prohibited. Therefore, the photo data cannot be considered representative of people's presence and movements over a territory and of people's everyday activities. Figure 1 shows that photo data may reflect people's leisure activities and touristic travels. However, it would be wrong to assume that this is always the case. The possible relation of the published photos to the author's leisure time, travels, or professional activities can be judged from the temporal frequency and regularity of the photos and from their spatial distribution.

Analysis of georeferenced microblog data

Posting microblog messages from mobile devices may occur more frequently and spontaneously and in a wider range of places and situations than taking and publishing photos. Besides, there is no time gap between producing and publishing a message, while photo authors may not publish their photos immediately after taking but may do this after some (often quite long) time. Therefore, unlike photos, microblog data are suitable for real time analysis, which may discover information about currently happening events, in particular, abnormal and disastrous events, such as earthquakes or storms (Chae et al. 2012; Andrienko et al. 2014). This requires processing of the message texts. One of the approaches is pre-filtering of the messages for selecting only those that contain analysis-relevant keywords, such as terms denoting extreme weather conditions (Andrienko et al. 2014). Another approach is extracting significant terms, i.e. such words that do not occur frequently in microblog messages in general or in the times (seasons) or places where they have occurred (Chae et al. 2012; Bosch et al. 2013). Each occurrence of a significant term is treated as a separate spatial event. A spatio-temporal concentration (cluster) of events with the same term may indicate that something is happening in this place and time, and the term gives an idea of what may be happening. The significant terms from such spatio-temporal clusters are shown on a map display using the text cloud technique, with the font size being proportional to the number of the term occurrences. The map is constantly updated in real time as new messages appear. By means of an interactive tool called Content Lens, the user can select a particular area and explore in more detail the term occurrences in this area. To increase the relevance of the information that is shown to the user, various user-constructed filters can be applied to the data (Bosch et al. 2013). In the other approach (Anrienko et al. 2014), the message texts are only used for the selection of potentially relevant messages and not used in the further analysis. The work focuses on real time detection of spatio-temporal clusters of relevant events, taking into account only the event locations and times but not the texts, and on tracing the cluster evolution (growing, shrinking, moving, merging, and splitting) over time (Figure 2). The individual events making the clusters and their message texts can be accessed on demand.

An example of an offline investigation of microblog posts related to a disastrous event (an epidemic) is presented in section 6.3.2 of Andrienko et al. (2013a). Although it uses data generated synthetically (based on real data), it shows the principal possibility of using microblog data for identifying the origin and possible cause of an epidemic, the ways of disease propagation, the spatial spread, and the evolution over time.

However, detecting and investigating disastrous or abnormal happenings is not the only possible use case for microblog data. Georeferenced microblog posts, at least those from active bloggers, may to some extent be considered as representative of the people's daily lives and used for studying people's behaviours. Thus, an analysis of a collection of Tweets posted by residents of the Seattle area (USA) revealed interesting patterns of collective and individual behaviours (Andrienko et al. 2013b). For this analysis, the Tweets were classified according to their topics, such as family, work, education, food, sports, etc. based on the occurrences of topic-specific keywords (for example, the topic 'family' is associated with the terms denoting family members: mother, mom, father, daddy, and so on). The researchers explored how much the Tweet topics are related to the locations from which the Tweets were posted and to the times when this happened. For this purpose, they aggregated the Tweets by the topics, areas in space, and time intervals and visually explored the results using maps and time histograms. It was found that there are areas where particular topics

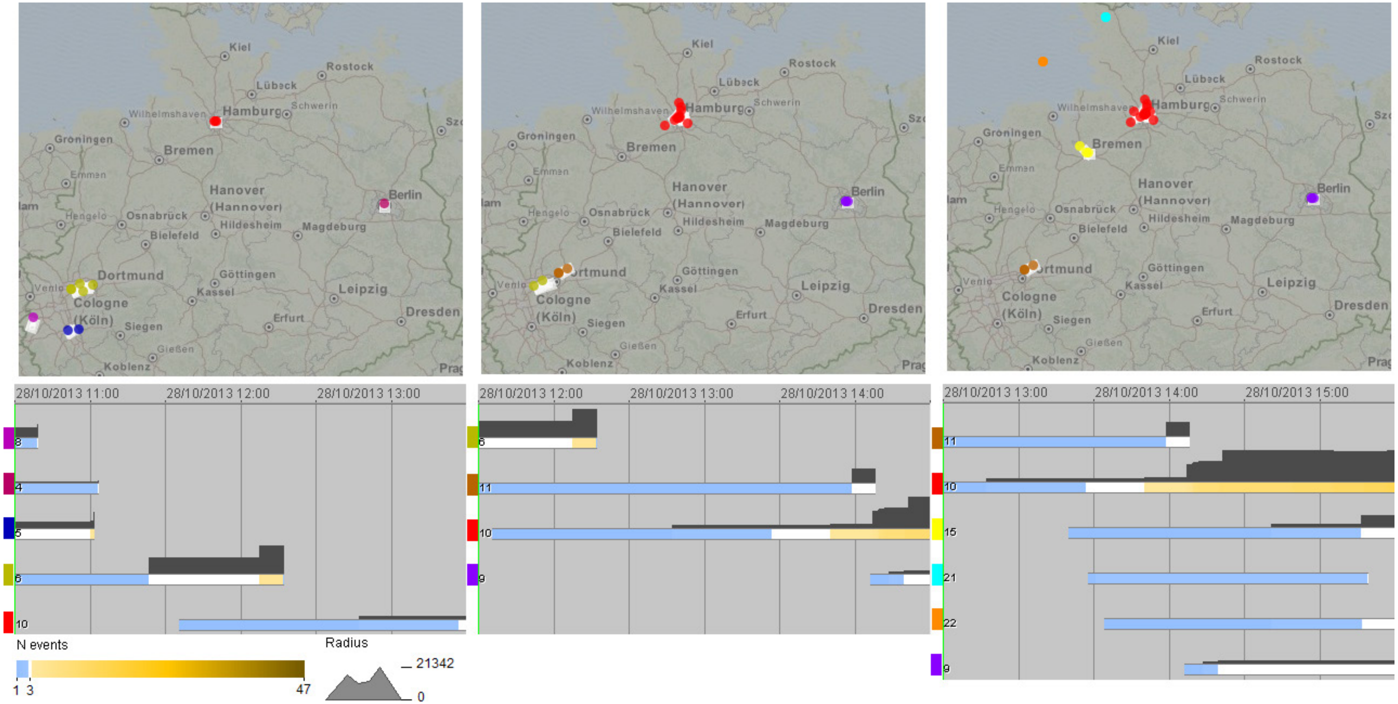


Figure 2: Emergence and evolution of spatio-temporal clusters of georeferenced Tweets related to a hurricane on October 28, 2013. Source: Andrienko et al. (2014).

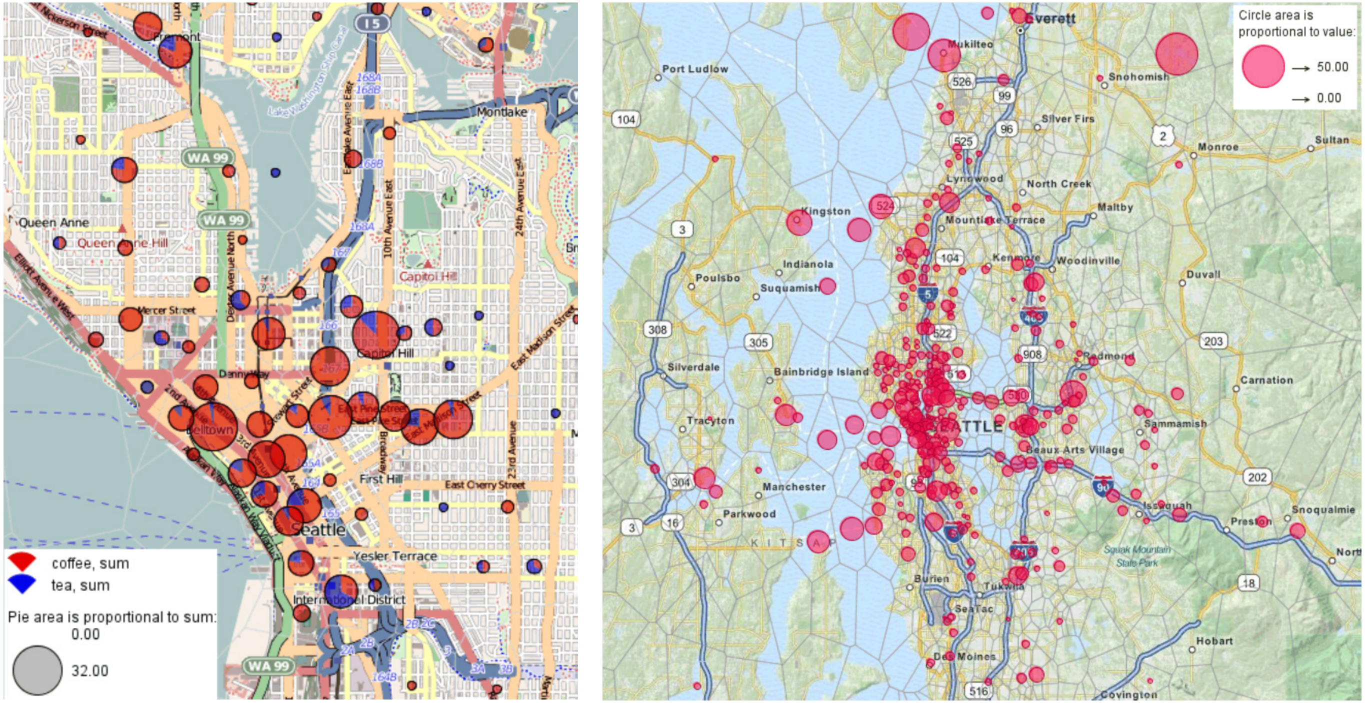


Figure 3: Left: the spatial distribution of the Tweet topics 'coffee' and 'tea' in the central area of Seattle. Right: the spatial distribution of the topic 'transportation'. Source: Andrienko et al. (2013b).

prevail, which may be related to the kinds of objects or facilities located in the areas (e.g. a university or a stadium) or to the characteristics of the population (e.g. an international district; see Figure 3, left). The researchers also looked at the spatial distributions of the different topics and found that some of them are correlated with the distribution of certain kinds of objects or facilities. Thus, the topic 'transportation' occurs along the main transportation corridors (Figure 3, right). Regarding the temporal distributions of the Tweet topics, the researchers found several very interesting patterns of *when* certain topics occupy the peoples' minds. Thus, 'food' occurs more frequently during lunch and dinner times, 'coffee' during/after breakfast and over the forenoon, 'transportation' during working day rush hours, and 'sports' and 'alcohol' in the evenings and over the weekend.

Although the study shows that the contents of some microblog posts are related to the places the authors visit and/or the activities they perform, these data in general contain a large proportion of noise, which includes texts with unidentifiable topics and texts with topics that are not relevant to the places of message posting (thus, a person may Tweet about work while being at home or about food while travelling in public transport). In fact, the proportion of noise outweighs the proportion of potentially relevant data. Therefore, it makes sense to analyse the topic distribution in space and time at the level of a large population of microbloggers, to have a sufficiently large amount of potentially relevant data and to be able to use valid statistical summaries. At the level of individuals, the message texts can hardly be indicative of the individuals' activities or purposes for visiting different places.

In analysing mobility behaviours of individuals, it is reasonable to look not at the message texts but at the temporal patterns of visiting different places (Andrienko et al. 2015). Significant (repeatedly visited) personal places are extracted from the collection of posts of each individual by spatial clustering of the post locations. Place semantics (i.e. the meanings, purposes for visiting, or activities performed in the places) can be determined based on the times over the weekly cycle when the individuals were present in the places. Thus, a place where a person is present in the evenings and nights of all days can be identified as the person's home place. However, separate consideration of the data of each individual is unfeasible and harmful for the personal privacy. The paper of Andrienko et al. (2015) proposes a privacy-respecting approach, in which data of a large number of Twitter users are analysed all together using a combination of computational techniques and visualisations presenting the data and analysis result in aggregated form. After extracting personal places and identifying their meanings in this manner, the original georeferenced data are transformed to trajectories in an abstract semantic space. The semantically abstracted data can be further analysed without the risk of re-identifying people based on the specific places they attend. The paper presents an example of analysing mobility behaviours of Twitter users in the area of San Diego (USA).

Conclusion

To summarise, georeferenced data from social media can be analysed as spatial events (i.e. independent points in space and time) and as trajectories of people. To analyse such data, visual analytics proposes a number of approaches combining computational techniques (clustering, aggregation, statistical summarisation, pattern detection, etc.) with interactive visualisations. With these approaches, it is possible to extract interesting information and gain new knowledge about places, events, and people's interests, behaviours, and habits. Metadata of the photos published through photo sharing services can reveal people's interests to tourist attractions, public events and other happenings, or natural phenomena and patterns of touristic behaviour. Georeferenced microblog posts can be analysed in real time for early detection of abnormal or disastrous events. It may also be useful to analyse the evolution of such events by looking at the spatio-temporal distribution of the event-related posts. Besides the information concerning unusual happenings, microblog data may be a source of knowledge about everyday mobility and activities of people. As both the popularity of the social media and the interest to analysing social media data are growing, we can expect the appearance of new analysis methods and new use cases for information that can be extracted by these methods.

Acknowledgements

This research was supported by European Commission within projects SoBigData RI (grant agreement 654024) and VaVeL "Variety, Veracity, VaLue: Handling the Multiplicity of Urban Sensors", grant agreement 688380.

References

Andrienko, G., Andrienko, N., Bak, P., Kisilevich, S., & Keim, D. 2009. Analysis of community-contributed space- and time-referenced data by example of Panoramio photos. In: *Proc. VMV – Vision, Modelling, and Visualization Workshop*, Braunschweig, Germany, November 2009.

Andrienko, G., Andrienko, N., Mladenov, M, Mock, M., & Poelitz, C. 2012. Identifying Place Histories from Activity Traces with an Eye to Parameter Impact. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, *18*(5): 675–688.

Andrienko, G., Andrienko, N., Bak, P., Keim, D., & Wrobel, S. 2013a. *Visual Analytics of Movement*. Springer.

Andrienko, G., Andrienko, N., Bosch, H., Ertl, T., Fuchs, G., Jankowski, P., & Thom, D. 2013b. Discovering Thematic Patterns in Geo-Referenced

Tweets through Space-Time Visual Analytics. *Computing in Science and Engineering*, *15*(3): 72–82.

Andrienko, N., Andrienko, G., Fuchs, G., & Stange, H. 2014. Detecting and Tracking Dynamic Clusters of Spatial Events. In: *Proc. IEEE Visual Analytics Science and Technology (VAST)*, Proceedings, pp. 219–220. DOI: http://dx.doi.org/10.1109/VAST.2014.7042499

Andrienko, N., Andrienko, G., Fuchs, G., & Jankowski, P. (2015). Scalable and Privacy-respectful Interactive Discovery of Place Semantics from Human Mobility Traces, *Information Visualization, 15*(2):117–153. DOI: http://dx.doi.org/10.1177/1473871615581216.

Bosch, H., Thom, D., Heimerl, F., Püttmann, E., Koch, S., Krüger, R., Wörner, M., & Ertl, T. 2013. ScatterBlogs2: Real-Time Monitoring of Microblog Messages Through User-Guided Filtering. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, *19*(12): 2022–2031.

Chae, J., Thom, D., Bosch, H., Jang, Y., Maciejewski, R., Ebert, D., & Ertl, T. 2012. Spatiotemporal Social Media Analytics for Abnormal Event Detection using Seasonal-Trend Decomposition. In: *Proc. IEEE Visual Analytics Science and Technology (VAST)*, 143–152. DOI: http://dx.doi.org/10.1109/VAST.2012.6400557

Jankowski, P., Andrienko, N., Andrienko, G., & Kisilevich, S. 2010. Discovering Landmark Preferences and Movement Patterns from Photo Postings. *Transaction in GIS*, *4*(6): 833–852.

Keim, D. A., Kohlhammer, J., Ellis, G., & Mansman, F. (eds.) 2010. *Mastering the Information Age – Solving Problems with Visual Analytics*, Eurographics.