



THEMATIC PATTERNS IN GEOREFERENCED TWEETS THROUGH SPACE-TIME VISUAL ANALYTICS

By Gennady Andrienko, Natalia Andrienko, Harald Bosch, Thomas Ertl, Georg Fuchs, Piotr Jankowski, and Dennis Thom

An exploratory study of the potential of georeferenced Twitter data (using tweets from Seattle-area residents over a two-month period) extracts knowledge about people's everyday life.

For 20 years, analysts have been interested in using authoritative spatial data to support planning and decision making by using the proliferation of geographic information systems (GIS) and their applications. Some of these systems have appeared in *CiSE*.¹ In this article, we present a different kind of spatial data created by lay members of society rather than by authoritative agencies, and we demonstrate how we can visually analyze such data to obtain potentially useful information about the people who created the data.

Analysts now have access to ever-increasing volumes of location- and time-referenced data, because of the high popularity of microblogging services such as Twitter, in conjunction with the widespread proliferation of personal mobile devices that can provide location information. Users worldwide generate in excess of 340 million tweets each day on Twitter alone (<https://business.twitter.com/en/basics/what-is-twitter>). We analyze microblogs because they apply to a number of applications, from the validation of socioeconomic theories and localized marketing, to a form of highly distributed “social sensors” that utilize Twitter users as potential field reporters of extraordinary events or disasters.

Researchers have investigated microblogs in computer science, social science, and related areas. Social scientists analyzed characteristics such as structure and relationships of social networks implied by microblogging activity.² Researchers have particularly used Twitter as a source for recommendation, event detection, and tracking³ as well as sentiment⁴ or hashtag analysis.⁵ However, researchers feel challenged while analyzing this unstructured source. Analysts might have to dig through a large number of non-related messages, abbreviations, slang, typing errors, and surprisingly often, just plain nonsense. This high ratio of noise combined with the brevity of individual tweets make analysts struggle with many traditional natural language-processing tasks, such as part-of-speech tagging,⁶ named entity recognition,⁷ and sentiment analysis.⁴ Yet, analysts often need this kind of processing to detect relevant tweets and to extract higher-level, meaningful information from them, such as general topics or sentiments.

Here, we describe a visual analysis approach for examining frequently tweeted words and their spatiotemporal patterns. We can discover the topical (thematic) tweeting behavior—of individuals and people as a collective—related to their everyday

activities and habits by interpreting these regular or at least repetitive spatial and temporal distribution patterns. This approach differs from related works that focus on the detection of extraordinary events in near-real time, such as earthquakes.³

Visual Analysis of Seattle Tweets

We explore a number of approaches to space-time visual analytics using a consistent example: tweets originating from the greater Seattle area from August to October 2011.

We gathered geographically referenced tweets through an API provided by the Twitter service itself. This real-time, public, and cost-free data stream covers only around 1–2 percent of the whole stream of tweets, but we can parameterize the information that we do receive by search terms and additional filters. The number of geographically referenced tweets (1 percent of the total) roughly corresponds to the rate limitation for the tweet stream. Therefore, by configuring the filter such that it collects just those with a recorded location anywhere on the globe, we can actually record almost all of the tweets (94 percent) with georeferences.

For the analysis presented here, in particular, we selected only tweets

of two months (8 August to 8 October 2011) from the greater Seattle area in Washington, whereby we defined this area as having the east-west extent between $123.05567^{\circ}\text{W}$ and $121.72083^{\circ}\text{W}$ longitude and a north-south extent between 46.94494°N and $48.391205^{\circ}\text{N}$ latitude (see the map in Figure 1). Each tweet consists of a unique tweet identifier, its geographic coordinates, time of tweeting, the tweet text itself, and an (anonymized) identifier of the Twitter user. This raw dataset contains 306,326 tweets of 13,752 Twitter users. Although this number might seem rather small considering the volume of tweets produced each day, it bears repeating that only about 1 percent of all generated tweets are georeferenced.

Because we're mainly interested in the thematic tweeting behavior of people that reflects their everyday life, we want to concentrate our analysis on the tweets of Seattle locals (thus excluding tweets of a visiting tourist, for example). To distinguish locals from visitors, we counted for each unique user ID the days N_1 while being inside the greater Seattle area, and the count of days N_2 being outside during the observation period. We considered a user to be a Seattle local if $N_1 > 9$ days and $N_2 < 9$ days (of the 60-day period).

When we performed a manual check of the very few IDs having both $N_1 > 9$ and $N_2 > 9$, we found that these tweets were computer-generated messages such as Foursquare notifications or other games. These types of tweets usually exhibit a defined content pattern. For example, Foursquare check-ins follow the pattern

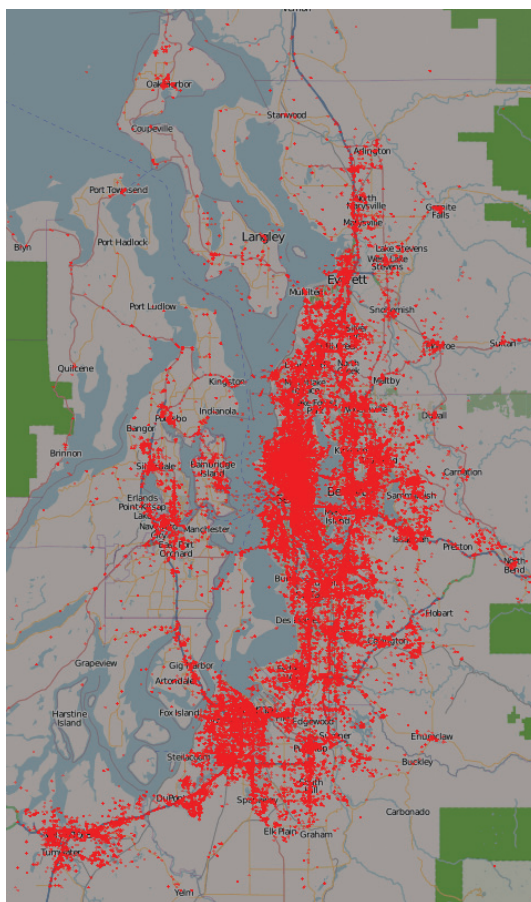


Figure 1. Map of the greater Seattle area under examination with positions for all the tweets in the dataset. Each tweet consists of a unique tweet identifier, its geographic coordinates, time of tweeting, the tweet text itself, and an (anonymized) identifier of the Twitter user.

“I’m at <place name> (<address>) <http://...>,” while automated job announcements typically contain hashtags, such as #JOB, #JOBS, or end with #TweetMyJOBS. Thus, we can quite easily define additional filters to remove these tweets from the set.

After gathering and pre-filtering tweets, we obtained a set of 163,203 individual, georeferenced tweets from 2,607 local Twitter users within the greater Seattle area during the selected time period.

Content Analysis

We use several complementary approaches for gaining insight into peoples’ tweeting behavior: a spatio-temporal *term usage cluster analysis* to find general patterns and topic terms,

as well as a keyword-based categorization (*supervised exploration*) of tweet contents to find out what people tweet about, where, when, and how often. Specifically, the initial term usage analysis may give us an idea about what topic terms are potentially interesting and deserve a subsequent supervised exploration by tweet categories. Finally, we examine some of the findings using an interaction technique called “content lens”.

Term-usage cluster analysis. We begin our analysis with a large set of messages and no additional structure or prior knowledge about the content of this dataset. By marking the locations of all messages on a map of the greater Seattle area, we obtained a large colored area, which outlines the populated places, but provides no additional insight on the tweeting behavior (see Figure 1). Con-

sidering the tweets’ contents instead, we can count the occurrences of single words and have a look into the most prominent terms. Sadly, these terms almost exclusively consist of common English sentence parts not bearing any meaning without their context (so called *stop words*, for example, the, at, to, or and). After excluding these stop words from the analysis, we obtain a list of potentially interesting terms, (for instance Seattle, good, time, love, Tacoma, jobs, Bellevue, people, work, game, or tonight), for which we could again mark their locations on the map individually and compare the results to finally make a statement about the “Seattle tweeting behavior.” Instead of doing this manually for thousands of words, we let the computer do the

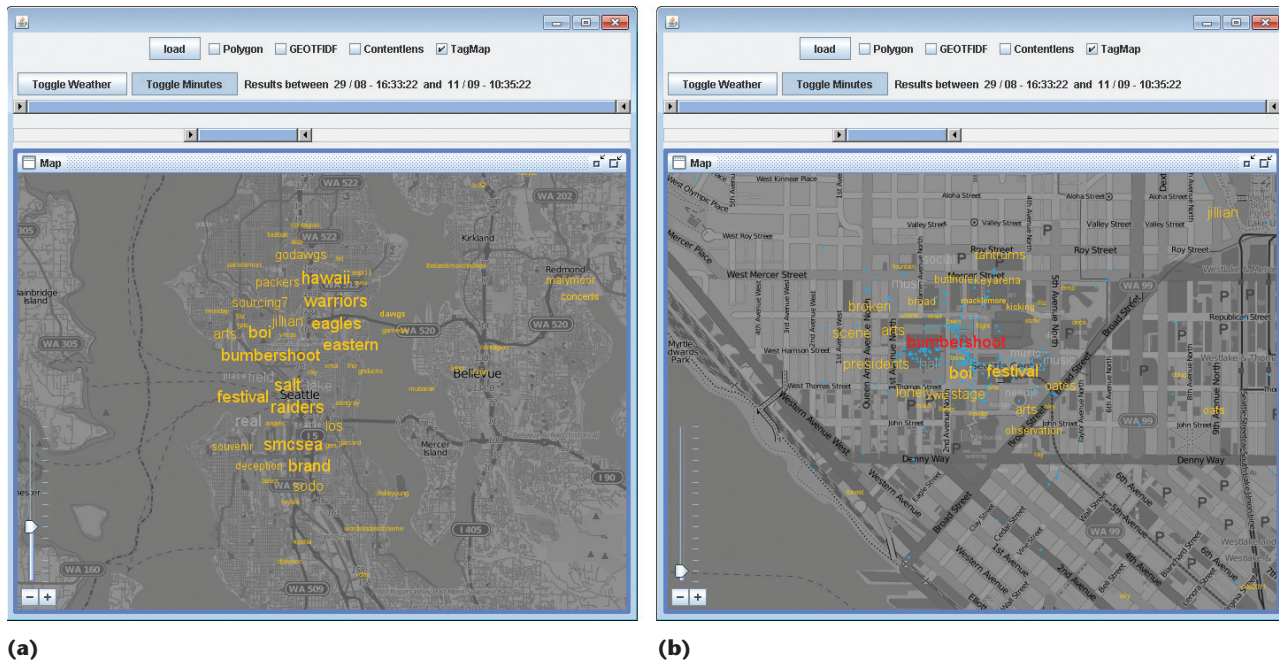


Figure 2. Results of the term-usage cluster analysis, presented as weighted labels on a map of the greater Seattle area. (a) Larger labels refer to denser or larger clusters of messages containing the related word. (b) Users can zoom into the map to reveal smaller subclusters.

heavy lifting. For any word present in the dataset, our procedure collects all messages in which it occurred and extracts their geospatial and temporal positions. Using a modified k -means cluster analysis scheme,⁸ we identify spatiotemporal hotspots for each term and discard all other locations as background noise. As a result, we still have the same list of potentially interesting words, but we've extended them by spatiotemporal positions based on the cluster centroids, describing when and where a word was most interesting. In addition, we can derive an importance weight for each term based on the number and density of elements in the corresponding cluster.

By placing the relevant words as labels sized according to weight on the corresponding spatial and temporal positions of the two-month data, our visualization allows a free exploration of location-specific content as well as the indication of anomalous events. In combination with an automatic layout adjustment to avoid overlapping labels and to aggregate overlapping identical labels, we also achieve a semantic zoom that shows more labels when

more visual space becomes available. For example, in Figure 2 we can observe how people explore the Seattle area, using a time slider and the map, indicating a range of detected events. The term *bumbershoot* corresponds to a large music and arts festival that took place at the Seattle Center. We can select this keyword by clicking on it, which will then highlight all messages associated with it on the map. By zooming into the map, the visualizations show smaller terms used frequently in connection with the event. For example “boi,” “presidents,” and “tantrums,” correspond to artists that performed at the festival.

Categorizing tweets according to content keyword selection. After a general overview of events and frequent keywords in the tweet dataset obtained from the term-usage cluster analysis, we are interested in gaining deeper insight into topic categories and their spatiotemporal occurrences.

Instead of using an automatic machine-learning approach, which might lead to ambiguous results, we preferred to extract the most frequent

topics in tweets. Therefore, we chose a more hands-on approach. To categorize the tweets in accordance with their semantic content, we compiled a list of themes known to be quite common in Twitter messages, represented by topic (category) keywords, including family, home, education, work, transportation, sports, game, love, friendship, music, food, weather, health, fitness, and money. Because people generally associate more than one term with a topic, we further collected lists of related words for each topic keyword (for example we associate *family* with the terms mother, mom, mommy, father, dad, daddy, kids, children, son, sons, daughter, daughters, brother, brothers, sister, sisters, niece, nephew, relatives, uncle, aunt, husband, wife, and folks). This straightforward approach is quite flexible and effective; however, for a more thorough study, we might consider using one or even multiple ontologies/folksonomies in guiding the search of popular keywords—for example, trending tweets as provided directly on <https://twitter.com> or on <http://trendsmap.com>.

We then used this set of keywords as a minimalistic ontology to categorize the tweets according to the presence of one or multiple topic categories. When we queried the database of 163,203 tweets for the keywords, we found that 33,343 tweets (20 percent of the database) contained one or more topic-related keywords (see Table 1).

However, when interpreting the categorization of tweets, we must also be aware of potential ambiguities stemming from our choice of terms. For example, “love” may be used in the romantic sense as well as an expression of preference or liking something or someone. We assume that among the 4,047 tweets with “love,” tweeters more frequently meant the latter sense.

In particular, we used a binary attribute encoding such that for each keyword on the list; the value of “1” represented the presence and the value of “0” represented the absence of the topic keyword (or one of its related terms) in a given tweet. Thus, we attached a list of 22 binary keyword presence values called a *feature vector* to each tweet, corresponding to the 22 topic categories from Table 1. For the purpose of content analysis, we represent every tweet by its associated feature vector, which lets us abstract from the tweets’ unstructured textual content.

We can then summarize the feature vectors according to different subdivisions (along spatial, spatiotemporal, and movement trajectories, as we explain in the next section) to gain insight of spatial and/or temporal patterns, trends, and hotspots in users’ tweeting behavior. More specifically, we sum up the feature vectors. As an example, we compute for each of the 22 keywords the sum of 1-values for

all of the tweets originating from a given subdivision region. Thus, we obtain a *summarization vector* for every subdivision region representing the number of tweets related to each of the 22 topic categories.

Analysis Goals and Interpreting Results

We choose which spatiotemporal subdivision scheme to use to arrive at the respective summarization vectors depending on the type of analysis question we want to address. In particular, we guide a subdivision scheme by geographical areas, as well as spatial clusters (ignoring time), spatiotemporal clusters, and movement trajectories. The first option is the most straightforward, since we predefine areas, but it’s often of limited value, because geographic area subdivisions don’t necessarily correlate with the spatial distribution of the observed phenomenon (tweeting, in our case). Thus, we explore what insight we can gain using each of the other three subdivision schemes.

A key idea behind all three schemes is that “dense” aggregations of tweets in space or space-time represent significant clusters with respect to tweeting behavior; whereas we can discard areas with only a few scattered points as “noise” (no significant patterns exist there). We can express this density criterion as a minimum required number of tweets occurring within a maximum allowed spatial/spatiotemporal distance of each other. After identifying the clusters, we take a representative object for each class—the tweet with the smallest cumulative space (space-time) distance to all other tweets in the same cluster—as the *seed*, and generate a polygon around each seed point, resulting in a mesh of polygons covering the study area

Table 1. List of thematic keywords and their frequencies for 163,203 tweets.*

Term	Frequency
Food	6,247
Love	4,074
Family	3,767
Work	3,076
Education	2,407
Home	1,954
Private event	1,928
Music	1,850
Sports	1,704
Game	1,678
Friends	1,410
Health	1,358
Coffee	1,136
Transport	1,120
Fitness	1,050
Alcohol	981
Weather	925
Sweets	876
Money	524
Public event	345
Tea	214
Wellness	151

*These tweets are from the Seattle and Puget Sound Metropolitan Area, collected from 8 August to 8 October 2011. The dark gray bars represent the relative frequencies (number of occurrences) of each topic.

(a Voronoi tessellation, see Figure 3). The particular algorithm we used splits up clusters of large spatial areas (characterized by a large number of tweets) into smaller regions for the Seattle downtown area (see Figure 3), with a comparatively higher tweeting activity than the adjoining areas.⁹ This makes descriptive statistics such as counts and averages that we computed for the resulting regions more readily comparable, because they refer to roughly equal absolute numbers of tweets.

Spatial Patterns of Tweets

To facilitate the analysis of tweets by area and keyword, we aggregated

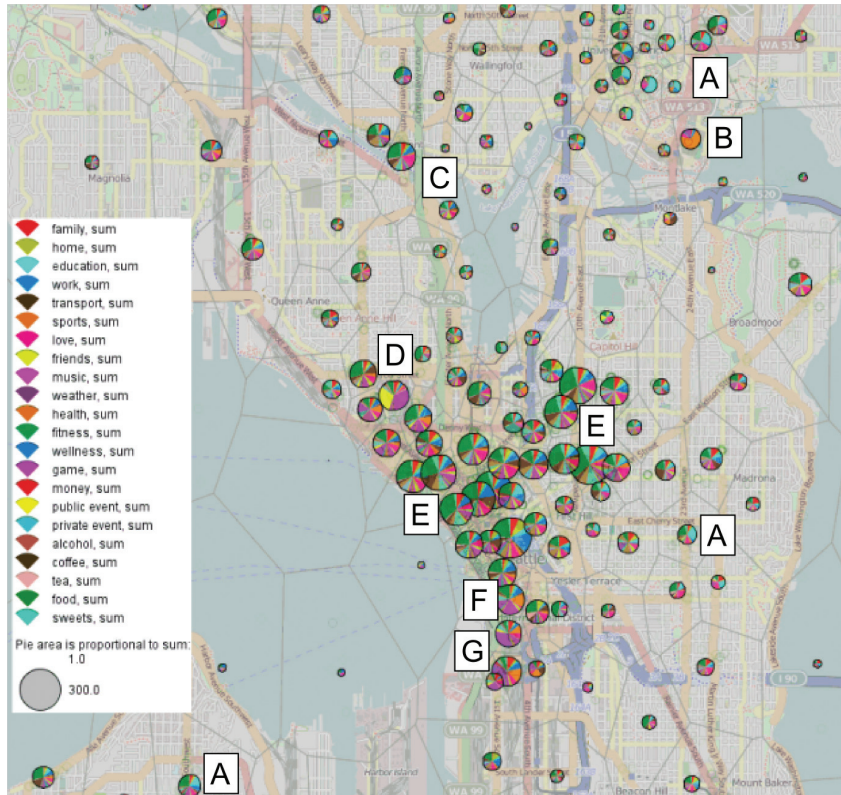


Figure 3. Distribution of keywords by cluster area in and around downtown Seattle. Some of the areas characterized by dominant keywords include the following: (a) a high proportion of “education,” including the University District (University of Washington) to the north-west and Seattle University (central-east); (b) a high proportion of “sports” (the University of Washington sports arenas); (c) a high proportion of “love” (an artsy and bohemian district of Fremont); (d) a high proportion of “music” and “public events” (the Seattle Center, the location of Bumbershoot—Seattle’s music and arts festival); (e) a high proportion of “coffee” covering most of the Seattle downtown area; (f) a high proportion of “sports” and “music” (Pioneer Square, at the southern end of Seattle’s downtown, is known for its lively bar and club scene); and (g) a high proportion of “sports” and “game,” including the CenturyLink Field multipurpose stadium (American football and soccer) and the Safeco Field baseball park.

the georeferenced origin locations of tweets into spatial clusters (Voronoi polygons) of variable size according to the aforementioned approach, while we disregarded the temporal aspect (time stamps) of tweets for now. The clustering radii varied from 500 meters to 5 kilometers, depending on the density distribution of tweets, with the densest distribution in and around Seattle’s downtown area. (Tweeters are 10 times more active per land area around downtown Seattle than in the surrounding area.) We then summarized the feature vectors for the tweets originating from each polygon area. Hence, each polygon is characterized by the

frequency distribution of keywords (see Figure 3).

Similar to a bivariate map displaying a spatial relationship between two variables, we can map two semantically related keywords and visualize their relationship in geographical space. Figure 4 presents a bivariate distribution pattern of tweets with “coffee” and “tea” keywords. Consistent with Seattle’s reputation of being the coffee-consumption capitol of the US, having 35 coffee shops per 100,000 residents (www.thedailybeast.com/galleries/2017/26/the-20-most-caffeinated-cities.html#slide1), tweeters posted about the coffee in Seattle’s downtown area far more

frequently than tea-related messages. We found a few exceptions scattered throughout the city, as well as one particular larger area located immediately south of downtown Seattle. This area, where tea-related tweets dominate coffee-related tweets, is Seattle’s international district known as Seattle’s Chinatown—a unique neighborhood where various Asian nationalities and ethnic groups traditionally have lived and worked side-by-side.

We reveal one type of pattern by mapping distributions of multiple keywords by their frequencies, based on absolute numbers. We develop a different type of pattern by mapping keywords by relative numbers (ratios). To prepare data for a map depicting a keyword distribution relative to other keywords, we divided the count of tweets with a particular keyword by all of the tweets carrying out this transformation for each Voronoi polygon. Figure 5 depicts the result of this transformation for the “transportation” category. Unsurprisingly, people tweeted about “transportation” along the main transportation corridors, including the interstate highways I-5, I-405, I-90, and the ferry lines across Puget Sound (from north to south: Mukilteo Ferry, Kingston Ferry, Bainbridge Island Ferry, and Fauntleroy Ferry). Interestingly, people in the ferries tweeted at least one of the transportation keywords in 25–40 percent of all their messages.

Spatiotemporal Patterns of Tweets

We used the density-based clustering approach on the tweets again to find out how the messages containing one or more of the 22 keywords (see Table 1) was distributed in space *and* time. However, this time we

considered *spatiotemporal* distances between tweets.¹⁰

Using a neighborhood size of 10, (the minimum number of other tweets within both spatial and temporal thresholds to form a new cluster) with a 500-m distance threshold (each tweet belonging to a cluster must be within 500-m geographic distance from another tweet that already is a member of the cluster), and a 15-minute temporal threshold (the time separation to another tweet in the cluster must be no longer than 15 minutes), we discovered 375 dense spatiotemporal clusters of messages containing 37,336 (23 percent) out of 163,203 tweets used in the analysis (see Figure 6).

The *spatial distribution* (locations) of clusters covers the entire study area, with the largest concentration of clusters in and around downtown Seattle, again signifying it as an area of frequent and concentrated messaging. Note that because this time we incorporated the time dimension in the clustering process, we obtained multiple clusters that are disjoint in space-time, but actually overlap in geographic space. Interestingly, the largest clusters occur in and around the city of Renton (south of Lake Washington) and unlike the diverse clusters covering downtown Seattle, these clusters are comprised almost exclusively of messages about “music.” After conducting a manual inspection of the tweets from the area, we believe that this is an issue with one user spamming everyone, asking them to please listen to his new song(s). This kind of human evaluation of intermediate findings is one of the key advantages of visual analytics over fully automated approaches. In this instance,

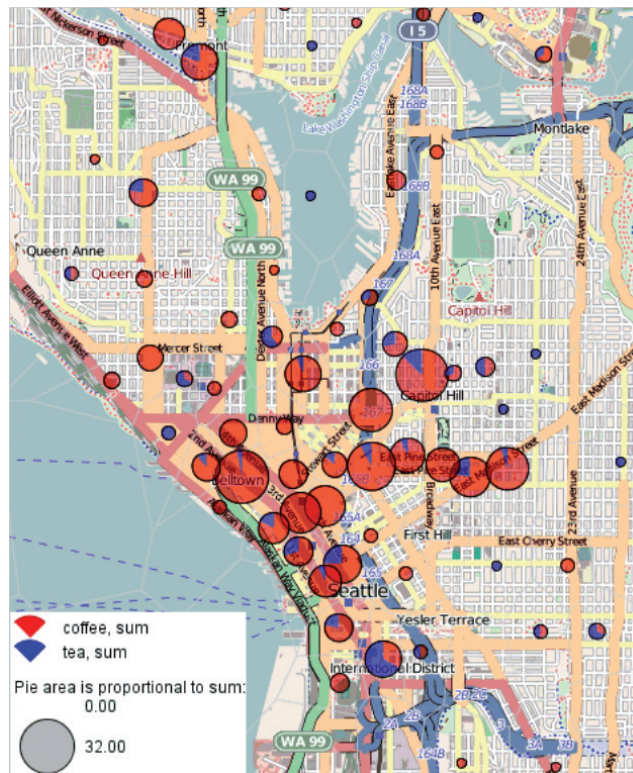


Figure 4. The bivariate distribution of coffee- and tea-related tweets. Pie charts are positioned at the location of their respective clusters' representative object.

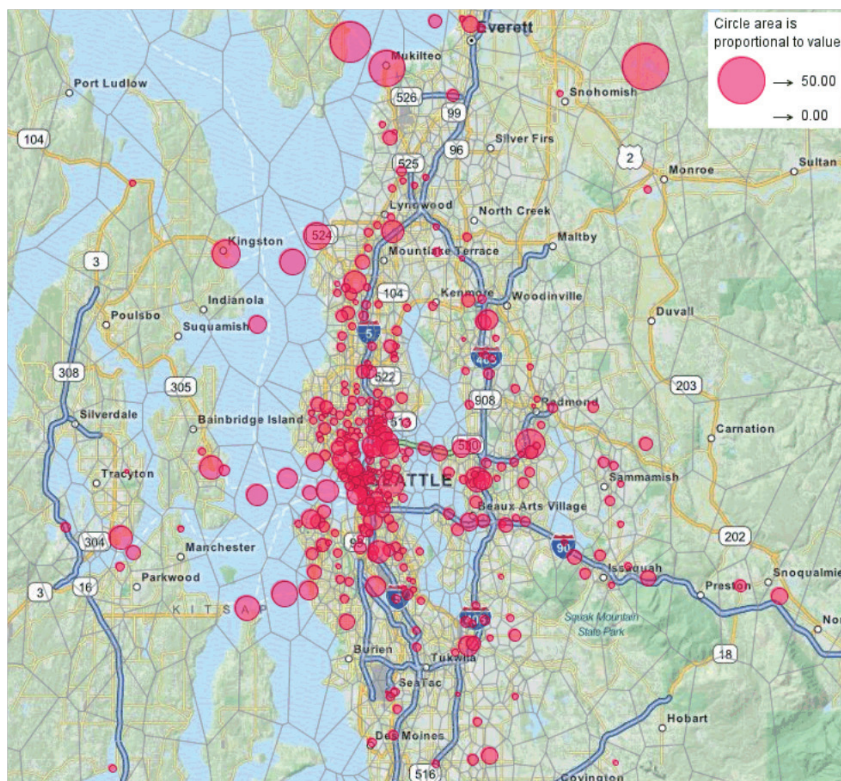


Figure 5. Spatial distribution of tweets with the “transportation” keyword. Each circle on the map represents the percentage of transportation-related tweets to all of the tweets originating from a given area.

we could have culled tweets from this user from further analysis, if deemed appropriate, without necessitating the comparatively complex definition of algorithmic spam filters.

We depicted the *temporal distribution* of clusters on a grid akin to a calendar sheet (see Figure 7). These histograms reveal two different temporal distributions of the tweets—one during the workdays of the week (Monday through Friday) and another during the weekend days. A relatively low number of people were messaging during the morning hours of the weekdays, with Monday morning (back to work after weekend) being conspicuously low in tweeting activity and afternoon and evening hours being the “prime tweeting time,” as people catch up with friends and family. Most people tweeted Monday, during the hour of 17:00–18:00, with 148 message clusters. During the weekend, people overall tweeted more. They had a more sustained level of activity, with periodic peaks occurring during morning, mid-day, and evening hours. We can use both of these patterns to confirm that tweeting is indeed a form of social activity, in which the majority of the “tweeting public” engages outside their regular work hours. Note the empty cells representing eight consecutive hours from Sunday night to Monday morning don’t signify an absence of tweeting activity. Because we’re not looking at individual tweets but at spatiotemporal clusters, the empty cells signify that tweeting is more or less evenly distributed over the area with no hives of activity—probably

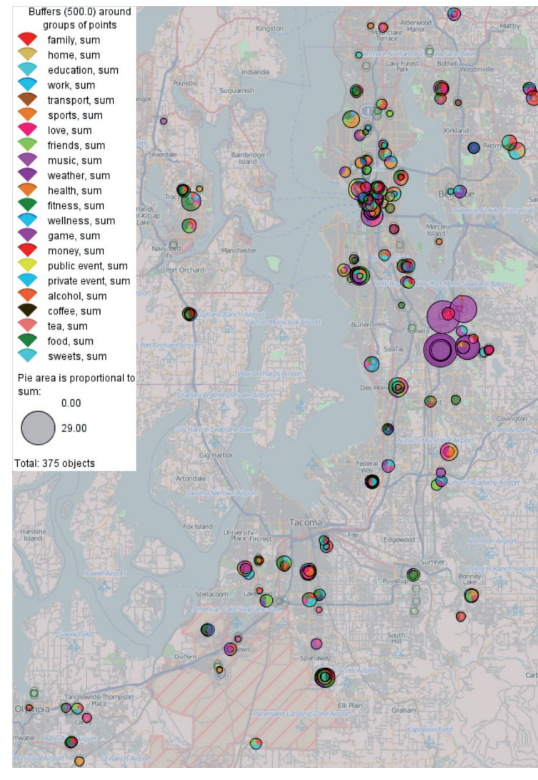


Figure 6. Spatiotemporal clusters of tweets. While disjoint in time, several of the 375 spatiotemporal clusters overlap spatially.

people coping with the start of the week and getting ready for work, so they’re tweeting less than during other day/time periods of the week.

In Figure 8, we show a breakdown of the total number of tweets by topic category, aggregated in the same way as explained for Figure 7. We organized the absolute numbers of tweets (sans auto-generated messages for Foursquare and such) by days and hours to show again that tweeting activities are highest in the late afternoon and early evening of working days, with overall higher sustained activity on the weekends. However, the breakdown by topic category reveals some interesting patterns of *when* certain topics occupy peoples’ minds. Some exhibit similar, cyclic patterns for all days, such as “food” during lunch and dinner times, as well as “coffee” during or after breakfast and over the afternoon. People tweet about both more prevalently on weekends. Some show

distinct differences between work days and weekends. For instance, people tweet about “transportation” most during workday rush hours, while they discuss private events more on Friday nights and over the weekends. Yet people tweet other keywords—specifically “love”—more or less every waking hour. Once again, we need to be aware that we lack the distinction between romantic motives and the expression of preference or appeal. To quote English Victorian poet and novelist George Meredith (1828–1909), whose novels are noted for their wit, brilliant dialogue, and aphoristic quality of language, “Kissing don’t

last: cookery do!” If anything, we could perhaps construe that the increased relative prevalence of “love” during party after-hours (Friday and Saturday nights) might indeed be more related to romance-motivated tweets.

Spatial Behavior of Twitter Users

People tweet at various times from various locations. To visualize a spatial manifestation of tweeting behavior, we accounted for each individual represented in the dataset and for all of his or her tweeting locations. In this way, we constructed a trajectory (spatial footprint) representing the sequence of locations from which each given individual tweeted. We then computed the trajectory *medoid*—a central feature for the points comprising the trajectory. The medoid has the smallest average distance to all of the other points in the set. We can interpret the medoid of a Twitter user’s trajectory as a center of the user’s tweeting footprint

in geographical space. In Figure 9, we depict the distribution of the medoids by showing that communicating via Twitter occurs throughout the greater Seattle area, with a few clearly visible clusters (such as the Seattle city center, university district around the University of Washington, Fremont district, north of downtown Seattle, and the city of Bellevue), where tweeting seems to be more spatially concentrated than elsewhere.

We correlated the distribution of the medoids representing spatial behavior of Twitter users in the dataset with the 2011 distribution of the population density (from the US

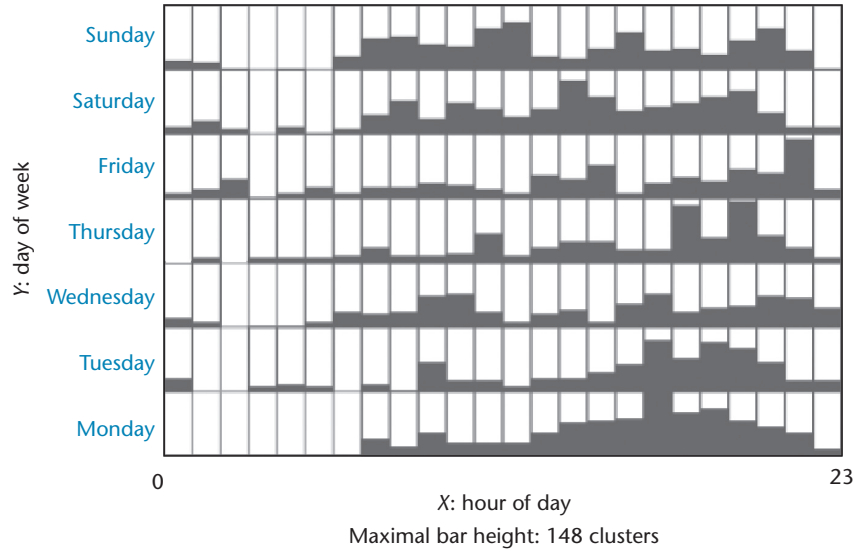


Figure 7. Temporal distribution of clustered tweets. Rows correspond to the days of the week from Monday (bottom) to Sunday (top). Columns of the grid correspond to hours beginning with 0 (midnight) on the left and ending with 23 (11pm) on the right. We scaled the bar heights in each cell of the grid proportionally to the number of clusters in a given hourly interval and day of the week. As indicated, a completely filled cell represents the global maximum value of 148 clusters for an hourly interval.

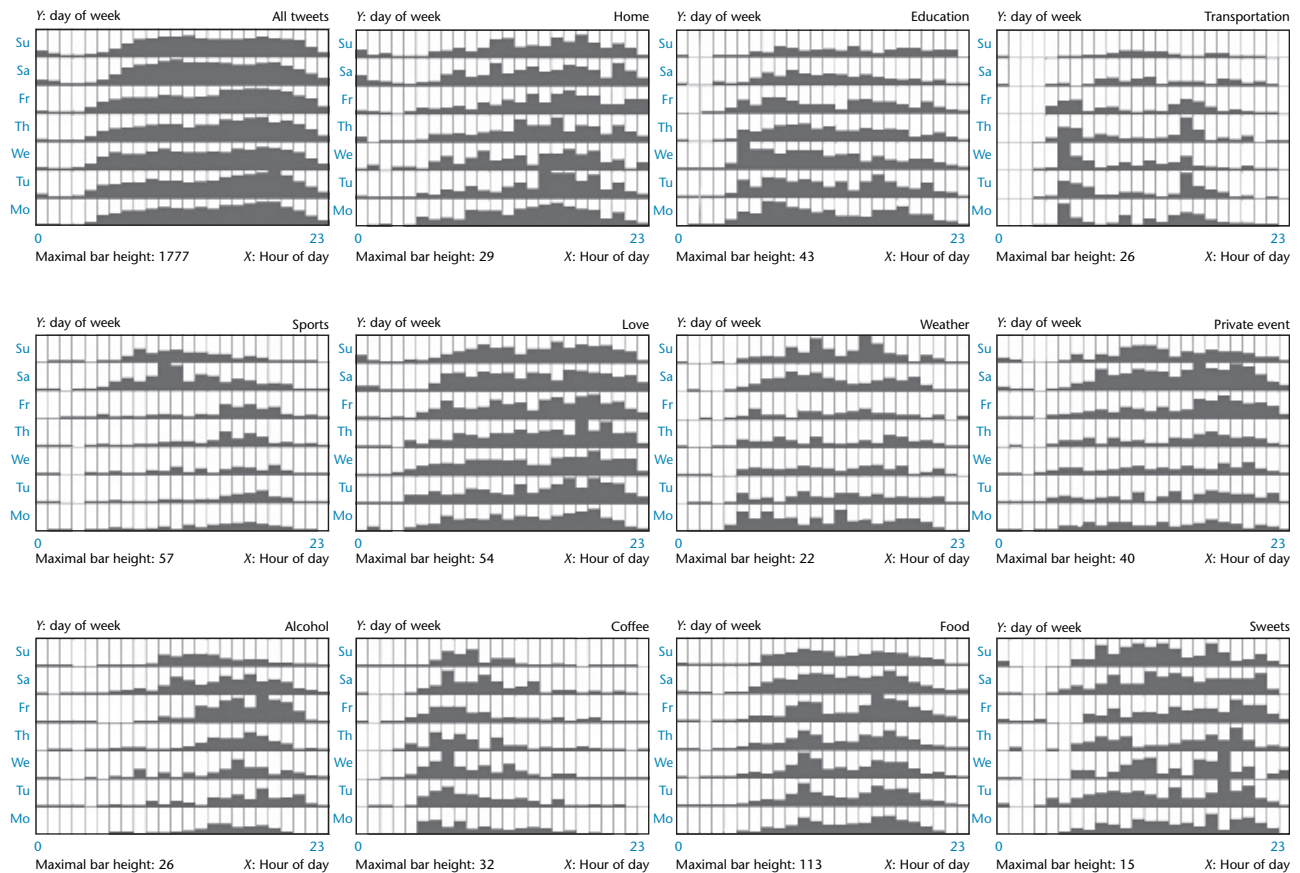


Figure 8. Temporal distributions of absolute tweet counts by topic category. “All tweets” (top left) is after filtering out auto-generated messages or notifications from Foursquare and other games.

Census Bureau) in the Puget Sound Metropolitan area (Pearson's $r = 0.52$, t-test significant at 99 percent). This means there's a moderate but significant relationship between the spatial distributions of places where people tweet and the population density. Here, we used US census population density data, which doesn't distinguish between the daytime and nighttime population.

We use medoids in our analysis not only to facilitate the visualization of spatial behavior of Twitter users, but also to analyze the spatial patterns of keywords as anchor points. We summarize the keyword feature vectors of the messages for each Twitter user. We attach the binary attributes to the medoids of the trajectories to facilitate the visualization of tweeting behavior related to specific keywords. We then normalize the attributes by dividing them by the number of positions (points) in each trajectory.

Interactive Hypothesis Evaluation

To further validate or falsify our hypotheses, we provide techniques for an explorative detailed analysis of individual messages, using our overview and anomaly indication. Once we identify spatial and temporal points of interest, we can use an interaction technique called *Content Lens* to inspect the contents of tweets sent from specific locations, by showing either the most prominent or the most unusual terms.¹¹ We can combine the technique with temporal and textual filters to contextualize messages that people have written in a certain time range that contain specific keywords.

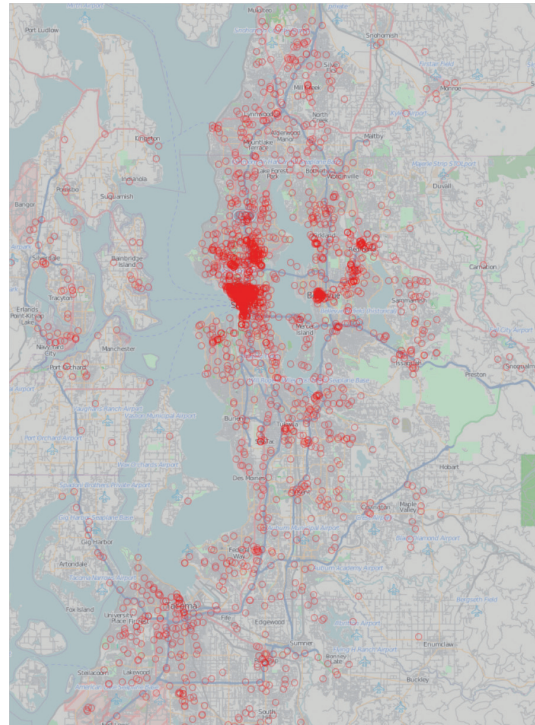


Figure 9. Medoids of Twitter users' trajectories. A few clearly visible clusters (such as the Seattle city center, university district around the University of Washington, Freemont district, north of downtown Seattle, and the city of Bellevue) seem to be more spatially concentrated than elsewhere.

To investigate the higher frequency of education-related keywords in the university district, we adjust the zoom level and the Content Lens size to cover the relevant area. Immediately, words such as university, class, and hall start to appear near the lens and change into homework, papers, and science as we brush further over the different district regions. In addition to highlighting the most prominent words, the lens also selects individual messages and reveals the nature of the chatter when we find a student telling the world how he skipped class today to play more sports (see Figure 10a).

As we mentioned previously, the large “music” cluster around Renton (see Figure 6) seems odd. When we place the Content Lens over the affected regions, we find the single words “song” and “listen” to be among the top terms, which is unusual, because people normally refer to music by many different terms that form a significant signal only when

combined. By applying a filter for these two words and selecting the messages in the region, we find a single user promoting his new song with more than 1,000 tweets (see Figure 10b).

By using the overview and anomaly indication as the first phase and the Content Lens as the second phase of an analysis loop, we can easily generate new findings, validate simple hypotheses, and also discover unexpected aspects of the data.

Our study and visual analytics method shows that we can use georeferenced messages posted by ordinary citizens as a source of interesting information about people and the space where they live. City planners, social scientists, advertisers, and other businesses will find this information potentially valuable. Analysts could use visual analytic approaches in creating “Smart Cities,” where they can aggregate data from many heterogeneous sources into actionable information, such as socioeconomic indicators. For example, the Urban Audit project (www.urbanaudit.org) collects more than 250 indicators in nine categories for 321 major European cities.

On a more general level, our method can assist human analysts in gaining insights from large and unstructured body of data. Analysts find this significantly relevant to big data-related topics, particularly regarding data variety and veracity in business applications. As per IBM's definition (see www-01.ibm.com/software/data/bigdata), big data spans four dimensions: volume (data size), velocity (real-time stream processing), variety (heterogeneous

information about people and the space where they live. City planners, social scientists, advertisers, and other businesses will find this information potentially valuable. Analysts could use visual analytic approaches in creating “Smart Cities,” where they can aggregate data from many heterogeneous sources into actionable information, such as socioeconomic indicators. For example, the Urban Audit project (www.urbanaudit.org) collects more than 250 indicators in nine categories for 321 major European cities.

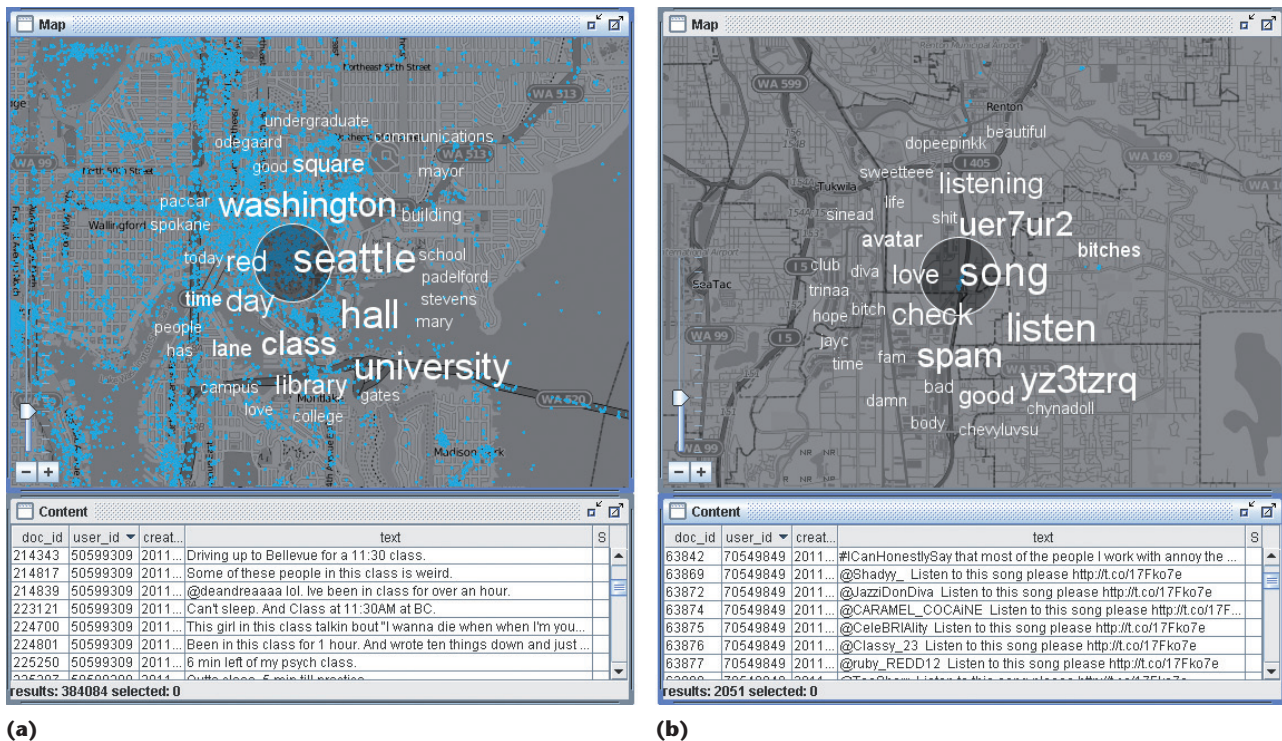


Figure 10. Zooming into the individual tweets for further analysis. (a) Investigating the details of the findings using the Content Lens (circles on map) and (b) the tweet contents (table).

data sources), and veracity (data provenance and trustworthiness).

We also identified several directions for future research. One obvious challenge we faced in the context of big data is the scalability of our approaches. Twitter and other sources generate truly massive amounts of data. We already conducted research toward scalable spatiotemporal clustering¹² that we could integrate with the methods presented here. We're going to consider data at larger temporal scales, and perform comparisons between different cities or regions. We'd like to include further mobility characteristics, such as trajectory patterns and users' significant or personal places in the analysis,¹⁰ as we would gain an even better understanding of the spatiotemporal phenomena we observed. However, this might also raise privacy issues.¹³

Further, we used a pre-collected, static dataset in the current analysis. We want to extend our approach so that it works directly on real-time tweet streams, and this has several implications on the analysis methods—topic

modeling in particular. We also intend to include spatiotemporal sentiment analysis in general, with respect to specific topics.

Acknowledgments

The German Federal Ministry for Education and Research (BMBF) partially funded this work as part of the VASA project (www.va-sa.net).

References

1. J. Wang et al., "Using Service-Based GIS to Support Earthquake Research and Disaster Response," *Computing in Science & Eng.*, 2012, vol. 14, no. 5, pp. 21–30.
2. A. Java et al., "Why We Twitter: Understanding Microblogging Usage and Communities," *Proc. 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, ACM, 2007, pp. 56–65.
3. J. Chae et al., "Spatiotemporal Social Media Analytics for Abnormal Event Detection Using Seasonal-Trend Decomposition," *Proc. IEEE Conf. Visual Analytics Science and Technology (VAST)*, IEEE CS, 2012, pp. 143–152.
4. H. Saif, Y. He, and H. Alani, "Alleviating Data Sparsity for Twitter Sentiment Analysis," *Proc. Making Sense of Microposts (MSM2012)*, vol. 838, paper 1, 2012; http://ceur-ws.org/Vol-838/paper_01.pdf.
5. S. Carter, M. Tsagkias, and W. Weerkamp, "Twitter Hashtags: Joint Translation and Clustering," *Proc. ACM Web Science*, ACM 2011, pp. 1–3.
6. K. Gimpel et al., "Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments," *Proc. Human Language Technologies, Assoc. for Computational Linguistics (ACL)*, 2011, pp. 42–47.
7. X. Liu et al., "Recognizing Named Entities in Tweets," *Proc. Human Language Technologies, ACL*, 2011, pp. 359–367.
8. D. Thom et al., "Spatiotemporal Anomaly Detection through Visual Analysis of Geolocated Twitter Messages," *Proc. IEEE Pacific Visualization Symp.*, IEEE CS, 2012, pp. 41–48.
9. N. Andrienko and G. Andrienko, "Spatial Generalization and Aggregation of Massive Movement Data," *IEEE Trans. Visualization and Computer*

Graphics, 2011, vol. 17, no. 2, pp. 205–219.

10. G. Andrienko et al., "Movement Tracks through Events to Places: Extracting and Characterizing Significant Places from Mobility Data," *Proc. IEEE Visual Analytics Science and Technology*, IEEE, 2011, pp. 161–170.
11. D. Thom, H. Bosch, and T. Ertl, "Inverse Document Density: A Smooth Measure for Location-Dependent Term Irregularities," *Proc. Int'l. Conf. Computational Linguistics*, ACL, 2012, pp. 2603–2618.
12. G. Andrienko et al., "Scalable Analysis of Movement Data for Extracting and Exploring Significant Places," *IEEE Trans. Visualization and Computer Graphics*, 2013, vol. 19, accepted for publication.
13. G. Andrienko and N. Andrienko, "Privacy Issues in Geospatial Visual Analytics," *Proc. 8th Symp. Location-Based Services (LBS)*, Springer, 2011, pp. 239–246.

Gennady Andrienko is a lead scientist at Fraunhofer Institute for Intelligent Analysis and Information Systems. His research interests include visual analytics and geovisualization. Andrienko has a PhD in computer science from Moscow State University.

Contact him at gennady.andrienko@iais.fraunhofer.de.

Natalia Andrienko is a lead scientist at the Fraunhofer Institute for Intelligent Analysis and Information Systems. Her research interests include visual analytics and geovisualization. Andrienko has a PhD in computer science from Moscow State University. Contact her at natalia.andrienko@iais.fraunhofer.de.

Harald Bosch is a research associate at the University of Stuttgart. His research interests include visual analytics, text analysis, and information visualization. Bosch has an MSc in information systems from the University of Stuttgart and Universität Hohenheim. Contact him at harald.bosch@vis.uni-stuttgart.de.


Thomas Ertl is a full professor of computer science at the University of Stuttgart, and the head of the Visualization and Interactive Systems Institute (VIS) and the Visualization Research Center of the University of Stuttgart (VISUS). His research interests include visualization, computer graphics, and human-computer interaction, with a focus on volume rendering, flow visualization, multiresolution analysis, and parallel and hardware-accelerated graphics for large datasets. Ertl has a PhD in theoretical astrophysics from the University of

Tuebingen. Contact him at Thomas.ertl@vis.uni-stuttgart.de.

Georg Fuchs is a research scientist at the Fraunhofer Institute for Intelligent Analysis and Information Systems. His research interests include visual analytics, geovisualization, and computer graphics. Fuchs has a PhD in computer science from Rostock University. Contact him at georg.fuchs@iais.fraunhofer.de.

Piotr Jankowski is a professor of geography at San Diego State University. His research interests include spatial decision support systems, participatory geographic information systems (GIS), and visual analytics. Jankowski has a PhD in geography from the University of Washington. Contact him at pjankows@mail.sdsu.edu.

Dennis Thom is a research associate at the University of Stuttgart. His research interests include visual analytics, machine learning, and information mining. Thom has a Diploma (MSc equivalent) in computer science from the University of Stuttgart. Contact him at dennis.thom@vis.uni-stuttgart.de.

 Selected articles and columns from IEEE Computer Society publications are also available for free at <http://ComputingNow.computer.org>.