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

Discovering Thematic Patterns in Geo-Referenced Tweets through Space-Time Visual Analytics

Abstract

This paper presents an exploratory study of the potential of geo-referenced Twitter data for extracting knowledge about everyday life of people: their interests, activities, habits and behaviors. This is different from other researches on analysis of Twitter data, which focus on detection of extraordinary events. In our study, we applied several methods of Visual Analytics, which enable combining the power of computers with the power of human vision, thinking, and insight. By example of tweets from residents of the greater Seattle area for two-month period, we demonstrate what kinds of knowledge can be gained by analyzing such data.

1 Introduction

The high popularity of microblogging services such as Twitter in conjunction with the widespread proliferation of personal mobile devices that are able to provide location information has led to the availability of ever increasing volumes of location- and time-referenced data. For the Twitter service alone, users worldwide generate in excess of 340 million tweets each day[*]. Analysis of microblogs is interesting for a number of applications, from the validation of socio-economic theories, localized marketing, to their use as a form of highly distributed 'social sensors' that utilize Twitter users as potential field reporters of extraordinary events or disasters[1].

Microblogs have been investigated by researchers in computer science, social science, and related to data analysis domains. Social scientists analyzed characteristics such as structure and relationships of social networks implied by microblogging activity[2]. Twitter in particular has been used as a source for recommendations[3], event detection and tracking[1, 4] as well as sentiment[5] or hashtag analysis[6].

However, the analysis of this unstructured source is quite challenging: tweets that may be of interest to an analyst are buried in a very large amount of non-related messages, and contents of individual tweets typically contain many abbreviations, slang, typing errors and surprisingly often, just plain nonsense. This high ratio of noise combined with the brevity of individual tweets makes many traditional natural language processing tasks, such as part-of-speech tagging[7], named entity recognition[8], and sentiment analysis[9] much more challenging. Yet, this kind of processing is often required to detect relevant tweets and to extract higher-level, meaningful information from them, such as general topics or sentiments.

Here we describe an approach to the analysis of frequently tweeted words and their spatio-temporal patterns. Interpreting these regular or at least repetitive spatial and temporal distribution patterns

[*] https://business.twitter.com/en/basics/what-is-twitter/

allows us to discover topical (thematic) tweeting behavior of people – both individuals and as a collective – related to their everyday activities and habits. This approach differs from the related works that focus on the detection of extraordinary events in near-real time, e.g. an earthquake[1, 4].

2 Visual Analysis of Seattle Tweets

We explore a number of approaches to space-time visual analytics using a consistent example: that of tweets originating from the greater Seattle area during two month, from August to October 2011. In the following sections, we briefly describe the data acquisition process and the preparatory content analysis processes applied to this data set. Next, we provide a detailed discussion of the obtained results, interpret their visual representations, and discuss insights we as human analysts can gain from them.

2.1 Data Acquisition

We gathered geographically referenced tweets through an interface (API) provided by the Twitter service itself. This real-time, public, and cost-free data stream covers only around 1-2% of the whole stream of tweets but can be parameterized by search terms and additional filters. Because the amount of geographically referenced tweets (1% of total) roughly corresponds to the rate limitation of the steam, by configuring the filter such that it collects just those with a recorded location anywhere on the globe we can actually record almost all tweets (94%) with geo-references.

For the analysis presented here, in particular, we selected only tweets of two month (August 8th to October 8th, 2011) from the greater Seattle area in Washington state, USA; whereby we defined this area as having the east-west extent between 123.05567°W and 121.72083°W longitude; and a north-south extent between 46.94494°N and 48.391205°N latitude (see 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 data set contains 306,326 tweets of 13,752 Twitter users. While this number may seem rather small considering the volume of tweets produced each day, it should be noted again that only about 1% of all tweets are geo-referenced.

As stated earlier we are mainly interested in the thematic tweeting behavior of people reflecting the everyday life. To this end, we want to concentrate our analysis on tweets of locals of the greater Seattle area, thus leaving out tweets of e.g. a visiting tourist. 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).

A manual check of the very few IDs having both $N_1 > 9$ and $N_2 > 9$ revealed that theses tweets were computer-generated messages like foursquare notifications or some games. These types of tweets usually exhibit a defined content pattern. For example, foursquare check-ins are following the pattern “I’m at <place name> (<address>) http://...”, while automated job announcements typically contain hash tags #Job, #Jobs, or end with #TweetMyJOBS. Thus, we can quite easily define additional filters to remove these tweets from the set.

As the final result of the gathering and prefiltering process we obtained a set of 163,203 individual, geo-referenced tweets from 2,607 local Twitter users within the greater Seattle area and selected time period.

2.2 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 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".

2.2.1 Term Usage Cluster Analysis

Our analysis starts with a large set of messages and no additional structure or prior knowledge about the content of this data set. 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). Considering the contents of the tweets 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 stopwords, e.g., "the", "at", "to", "and" ...). After excluding these stopwords from the analysis, we obtain a list of potentially interesting terms (e.g., "seattle", "good", "time", "love", "tacoma", "jobs", "bellevue", "people", "work", "game", "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".

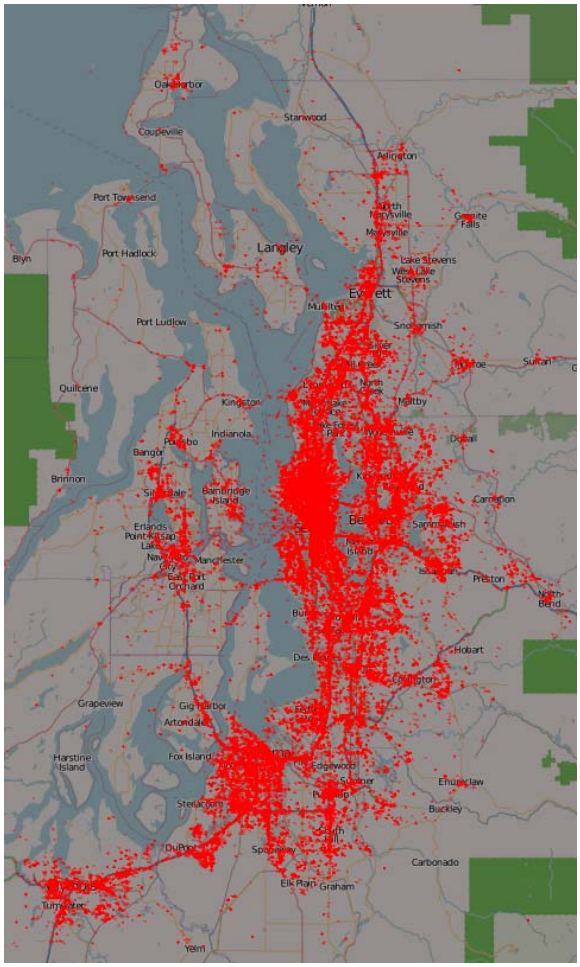

Figure 1: Map of the greater Seattle area under examination with positions of all Tweets in the data set.

Instead of doing this manually for thousands of words, we let the computer do the heavy lifting. For any word present in the data set, our procedure collects all messages in which it occurred and extracts their geospatial and temporal positions. Using a modified k-means cluster analysis scheme[10] we identify spatio-temporal 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 it is extended by spatio-temporal positions, based on the cluster centroids, describing when and where a word was most interesting. In addition, we can 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 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 one can observe how the Seattle area is explored 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 shows smaller terms used frequently in connection with the event, e.g. “boi”, “presidents” and “tantrums”, corresponding to artists that performed at the festival.

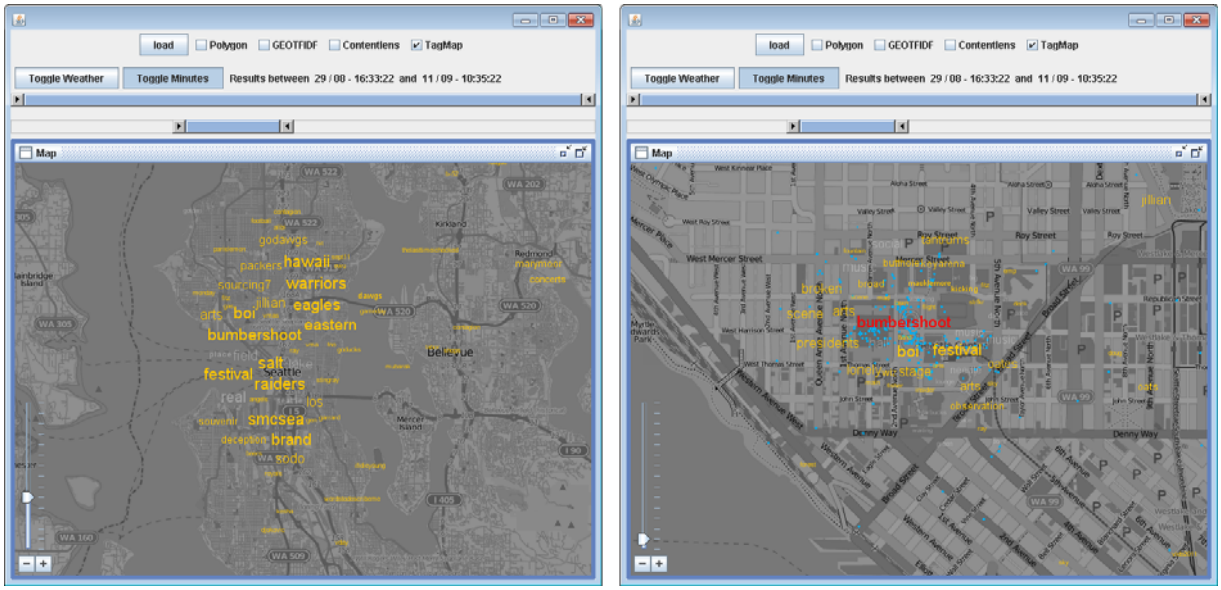

Figure 2: The result of the term usage cluster analysis presented as weighted labels on the map of the greater Seattle area. Larger labels refer to denser or larger clusters of messages containing the related word. Zooming into the map reveals smaller sub-clusters.

2.2.2 Categorization of Tweets according to Content Keyword Selection

After a general overview of events and frequent keywords in the tweet data set obtained from the term usage cluster analysis, we are interested in gaining a deeper insight into topic categories and their spatio-temporal occurrences.

Instead of using an automatic machine learning approach – which may lead to results that are difficult to interpret – to the extraction of the most frequent topics in tweets, we chose a more hands-on approach for our analysis. In order 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, money, etc. Since generally more than one term is associated with a topic, we further collected lists of related words for each topic keyword (e.g. **family** is also associated 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, folks). This straightforward approach is very flexible and quite effective; however for a more thorough study we might consider using one or even multiple ontologies/folksonomies in guiding the search of popular keywords[†].

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. Querying the data base of 163,203 tweets for the

[†] For example, trending tweets as provided directly on https://twitter.com or on http://trendsmap.com/

keywords resulted in selecting 33,343 tweets (20% of the data base) containing one or more of the topic-related keywords (see Table 1).
It should be noted here that 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. Actually, we assume that among the 4,047 tweets with “love” its usage in the latter sense was more frequent than in the former.

Term	Frequency
food	6247
love	4074
family	3767
work	3076
education	2407
home	1954
private event	1928
music	1850
sports	1704
game	1678
friends	1410
health	1358
coffee	1136
transport	1120
fitness	1050
alcohol	981
weather	925
sweets	876
money	524
public event	345
tea	214
wellness	151

Table 1: List of thematic keywords and their frequencies for 163,203 tweets from Seattle and Puget Sound Metropolitan Area collected during mid-August to early October 2011. The dark grey bars represent the relative frequencies (number of occurrences) of each topic.

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, a list of 22 binary keyword presence values called *feature vector* was attached to each tweet, corresponding to the 22 topic categories from Table 1. Hence, for the purpose of content analysis every tweet is represented by its associated feature vector allowing us to abstract from the tweets’ unstructured textual content.
We can then summarize the feature vectors according to different subdivisions (spatial, spatio-temporal, along movement trajectories; explained below) in order to gain insight of spatial and/or temporal patterns, trends and hotspots in users’ tweeting behavior. More specifically, the feature vectors are summed keyword-wise, i.e. for a given subdivision region we compute for each of the 22 keywords the sum of 1-values for all of the tweets originating from that region. Thus, we obtain a *summarization vector* for every subdivision region representing the number of tweets related to each of the 22 the topic categories.

3 Analysis Goals and Interpretation of Results

Which spatio-temporal subdivision scheme we use to arrive at the respective summarization vectors depends on the type of analysis question we want to address. In particular, a subdivision scheme can be guided by geographical areas, as well as spatial clusters (ignoring time), spatio-temporal clusters, and movement trajectories. The first option is the most straight-forward since areas are predefined but also often is of limited value because geographic area subdivisions not necessarily correlate with the spatial distribution of the observed phenomenon (tweeting, in our case). Thus in the following, we explore what insight can be gained 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 areas with only a few scattered points can be discarded as 'noise' (i.e., no significant patterns exist there). We can express this density criterion as a minimum required number of tweets occurring within a maximum allowed spatial/spatio-temporal 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 around each seed point a polygon resulting in a mesh of polygons covering the study area (a Voronoi tessellation, see Figure 3 and Figure 5). In the particular algorithm[11] we used, clusters of large spatial extent characterized by a large number of tweets are further split up into smaller regions, as can be observed in Figure 3 for the Seattle downtown area with a comparatively higher tweeting activity than the adjoining areas. This makes descriptive statistics such as counts and averages computed for the resulting regions more readily comparable because they refer to roughly equal absolute numbers of tweets.

3.1 Spatial Patterns of Tweets

To facilitate the analysis of tweets by *area* and *keyword* we aggregated the geo-referenced origin locations of tweets into spatial clusters (Voronoi polygons) of variable size according to the above approach, while we disregard the temporal aspect (timestamps) of tweets for now. The clustering radii varied from 500 m to 5 km depending on the density distribution of tweets with the densest distribution in and around Seattle downtown area. (In other words, tweeting activity per land area is ten times as intense around Seattle downtown compared to the hinterland.) 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).

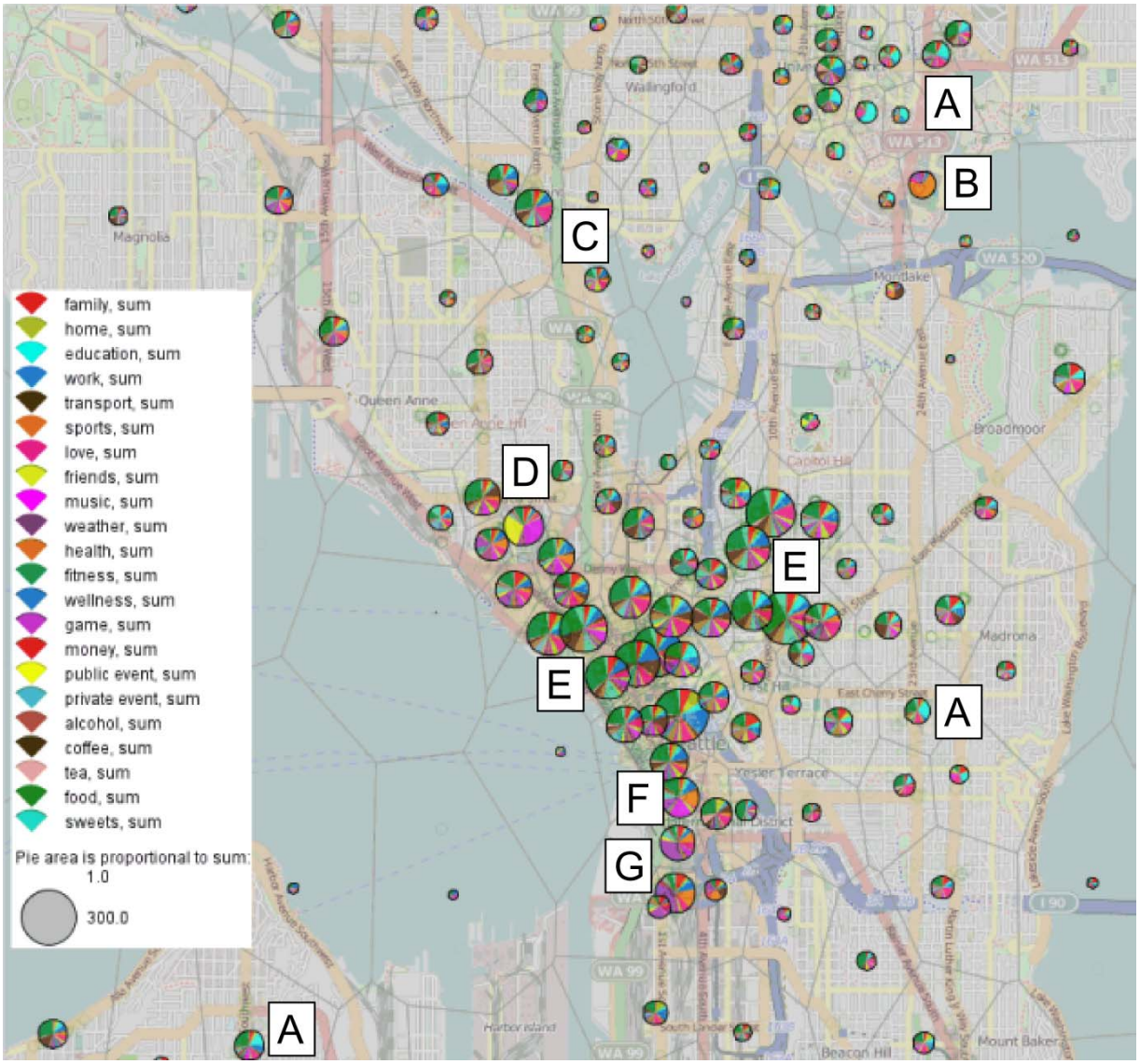

Figure 3: Distribution of keywords by cluster area in and around Seattle downtown. Some of the areas characterized by dominant keywords are: A – high proportion of "education" including the University District (University of Washington) to the North-West and Seattle University (Central-East), B – high proportion of "sports" (University of Washington sports arenas), C – high proportion of "love" (an artsy and Bohemian district of Freemont), D – high proportion of "music" and "public events" (Seattle Center – the location of *Bumbershoot*: Seattle's music & arts festival, the US's largest arts festival), E – high proportion of "coffee" covering the most of Seattle downtown area, F – high proportion of "sports" and "music" (Pioneer Square – the southern end of Seattle's downtown known for its lively bar and club scene), G – high proportion of "sports" and "game" including the CenturyLink Field multi-purpose stadium (American football, soccer) and the Safeco Field baseball park.

Similarly 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 its reputation

of being the coffee consumption capitol of the US (35 coffee shops per 100,000 residents[‡]) the frequency of coffee related tweets in Seattle's downtown area surpasses by far the frequency of tea related messages. This is true with the exception of a few locations 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 in Seattle's international district known otherwise as Seattle's Chinatown – a unique neighborhood where various Asian nationalities and ethnic groups have lived and worked side-by-side.

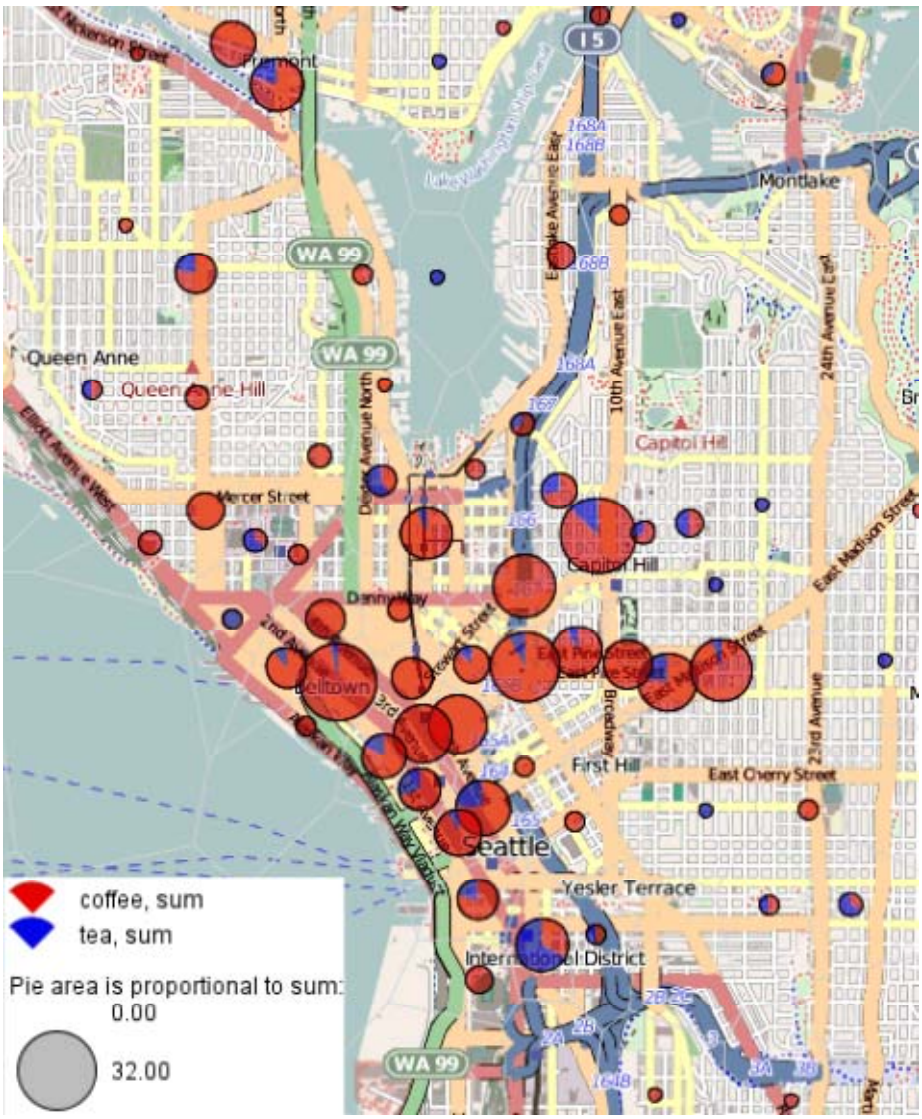

Figure 4: The bivariate distribution of "coffee" and "tea" related tweets. Pie charts are positioned at the location of their respective clusters' representative object.

Mapping distributions of multiple keywords by their frequencies based on absolute numbers reveals one type of pattern. A different type of pattern emerges from mapping keywords by relative numbers

[‡] http://www.thedailybeast.com/galleries/2010/07/26/the-20-most-caffeinated-cities.html#slide1

(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 3 depicts the result of this transformation for the "transportation" category. It is not surprising that 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 enough, between 25% and 40% of all messages tweeted from the ferries contained at least one of the "transportation" keywords.

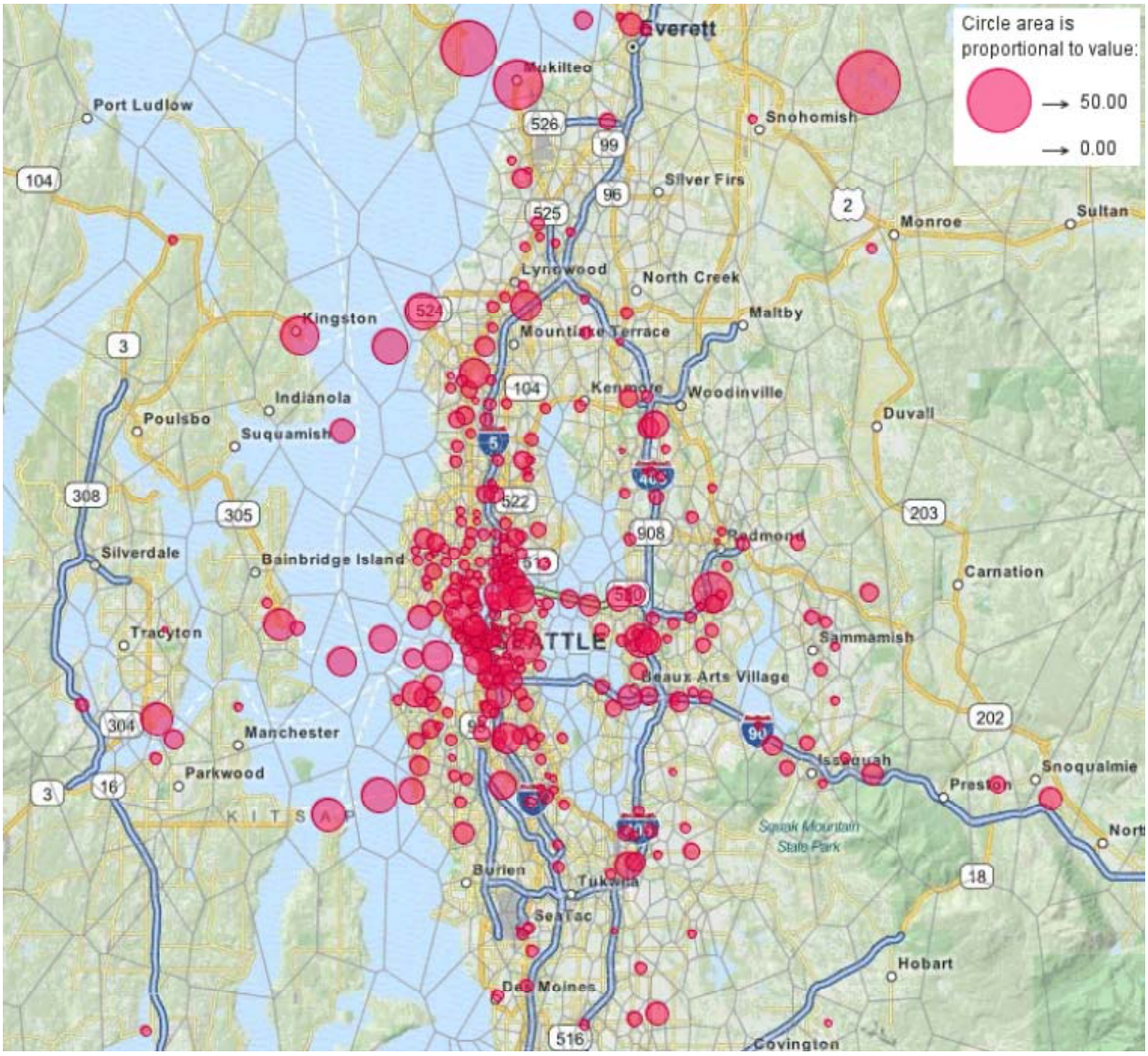

Figure 5: Spatial distribution of tweets with "transportation" keyword. Each circle on the map represents the percentage of transportation-related tweets to all of the tweets originating form a given area.

3.2 Spatio-Temporal Patterns of Tweets

To find out how are the messages containing one or more of the 22 keywords (Table 1) distributed in space *and* time we used the density-based clustering approach on the tweets again, however this time considering *spatio-temporal* distances between tweets[12].

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

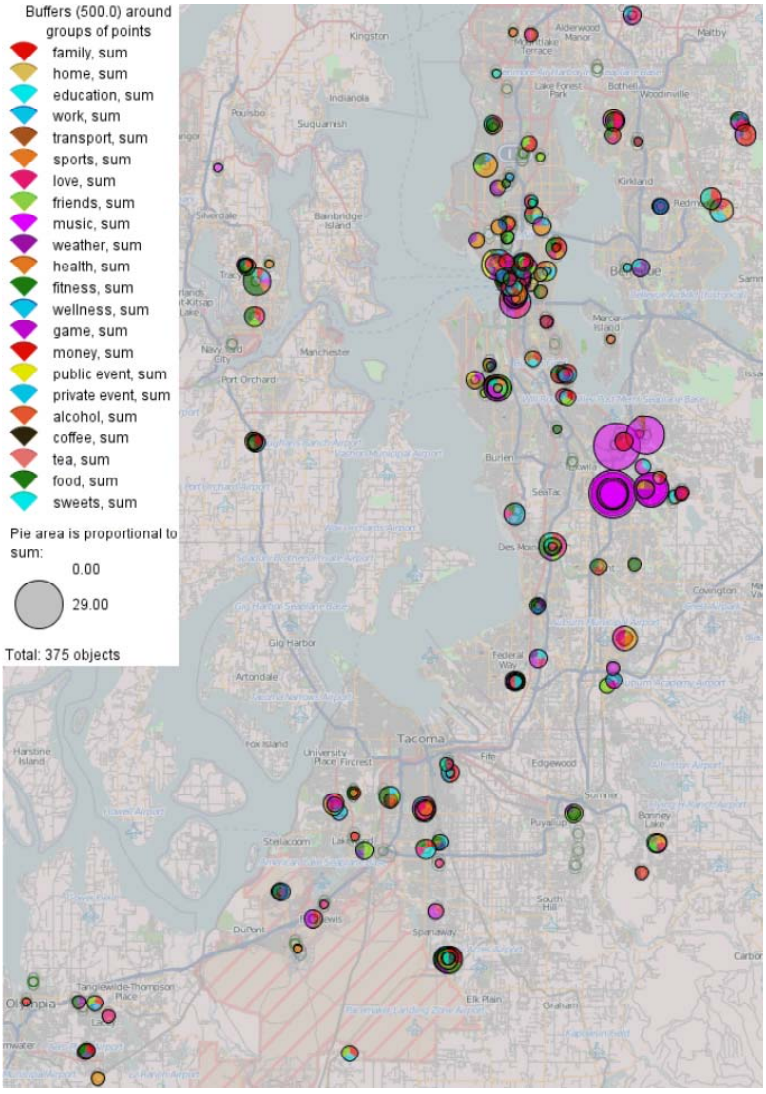

Figure 6: Spatio-temporal clusters (375) of tweets. Note that several clusters, while disjoint in time, overlap spatially.

The *spatial distribution* (locations) of clusters covers the entire study area, with the largest concentration of clusters in and around Seattle downtown, 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 enough, the largest clusters occur in and around the city of Renton (south of Lake Washington) and unlike the diverse clusters covering Seattle downtown, these clusters are comprised almost exclusively of messages about "music". A manual inspection of the tweets from the area (see Section 3.4) lead us to believe that this is an issue with one user spamming about his new "song"(s) that everybody should please listen to. This kind of human evaluation of intermediate findings is one of the key advantages of visual analytics over fully automated approaches – in this case, it allows us to cull tweets from this user from further analysis if deemed appropriate.

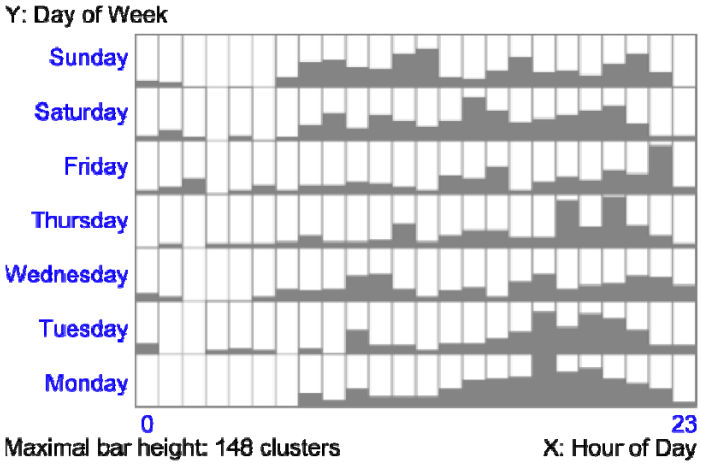

Figure 7: Temporal distribution of clustered tweets. Rows correspond to the days of the week from Monday (bottom) to Sunday (top). Columns of the grid correspond to hours beginning with 0 (midnight) on the left and ending with 23 (11pm) on the right. The bar heights in each cell of the grid are scaled 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.

The *temporal distribution* of clusters is depicted on a grid akin to a calendar sheet (Figure 7). These histograms reveal two different temporal distributions of the tweets; one during the work days of the week (Monday – Friday) and another during the weekend days. The week-day pattern is characterized by a relatively low level of messaging during the morning hours 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 (again Monday, the hour of 17:00 – 18:00 has the highest number of tweets, with 148 message clusters). The weekend pattern has an overall higher and more sustained level of activity with periodic peaks occurring during morning, mid-day, and evening hours. Both of these patterns confirm that tweeting is indeed a form of social activity, in which the majority of the "tweeting public" engages outside their regular work hours. One should note that the empty cells representing eight consecutive hours from Sunday night to Monday morning do not signify absence of tweeting activity. Since we are not looking at individual tweets but at spatio-temporal clusters, the empty cells rather signify that tweeting is more or less evenly distributed over the

area with no hives of activity – probably, a symptom of people coping with the start of the week, getting ready for work and tweeting less than during other day/time periods of the week.

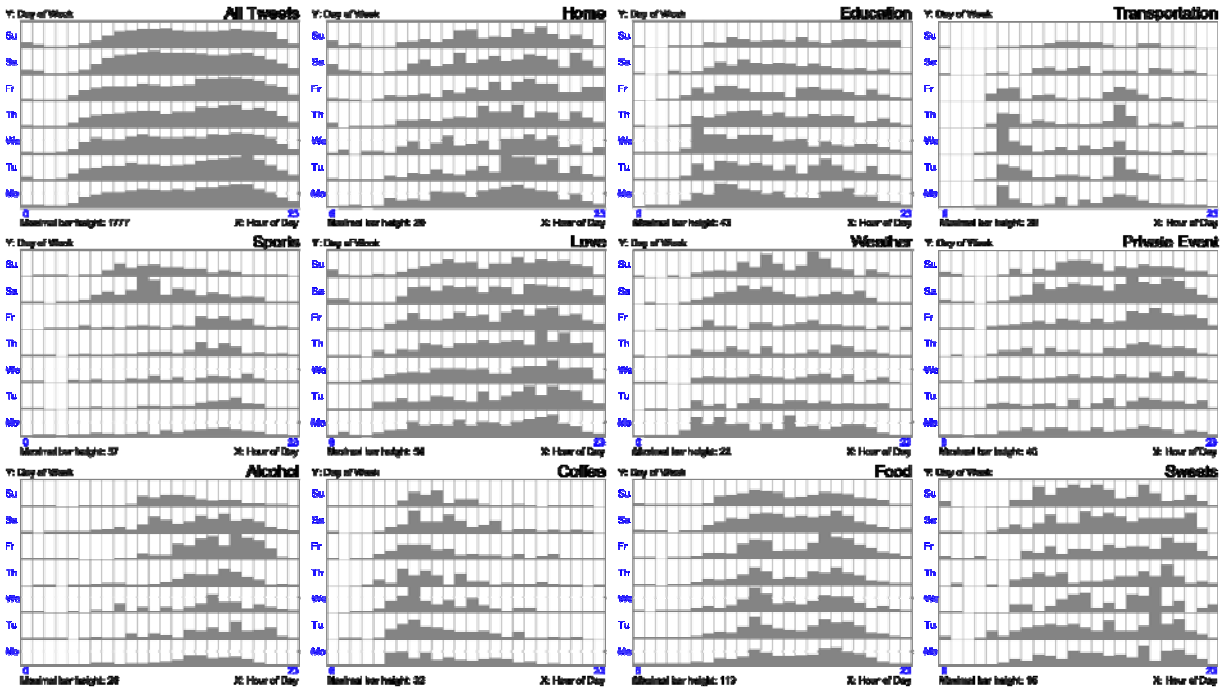

Figure 8: Temporal distributions of absolute tweet counts by topic category. "All Tweets" (top left) is after filtering out auto-generated messages from foursquare etc.

Figure 8 shows a breakdown of the total number of tweets by topic category and aggregated in the same way as explained for Figure 7. The absolute numbers of tweets (sans auto-generated foursquare etc. messages) by days and hours shows again that tweeting activities are highest on the late afternoon and early evening of working days and overall higher sustained activity on the weekends. However, the breakdown by topic category reveals some very interesting patterns of *when* certain topics occupy the peoples' minds. Some exhibit similar, cyclic patterns for all days, such as "food" during lunch and dinner times as well as coffee during/after breakfast and over the forenoon; both are also more distributed and prevalent on weekends. Some show distinct differences between work days and weekends: "transportation" is, not surprisingly, most relevant during working day rush hours, while "private events" largely happen on Friday nights and over the weekends. Yet others – here specifically, "love" – occupy people more or less on every waking hour, although again we need to be aware that we lack the distinction between romantic motives and expression of preference or appeal. To quote George Meredith,[§] "Kissing don't last: cookery do!" – if at all, we could perhaps construe the increased relative prevalence of "love" during party after-hours (Friday and Saturday nights) might indeed be more related to romance-motivated tweets.

[§] English Victorian poet and novelist (1828-1909) whose novels are noted for their wit, brilliant dialogue, and aphoristic quality of language

3.3 Spatial Behavior of Twitter Users

People tweet at various times from various locations. To visualize spatial manifestation of tweeting behavior we accounted for each individual represented in the data set and for all his/her tweeting locations. In this way we were able to construct for each individual a trajectory (spatial footprint) representing the sequence of locations from which a given individual tweeted. We then computed the trajectory *medoid* – a central feature for the points comprising the trajectory. The defining property of the medoid is that it has the smallest average distance to all of the other points in the set. The medoid of a Twitter user trajectory can be interpreted as a center of the user's tweeting footprint in geographical space. Figure 9, depicting the distribution of the medoids, shows that communicating via Twitter occurs throughout the greater Seattle area with a few clearly visible clusters (Seattle city center, university district with University of Washington, Freemont district - north of downtown Seattle, and the city of Bellevue) where tweeting seems to be more spatially concentrated than elsewhere.

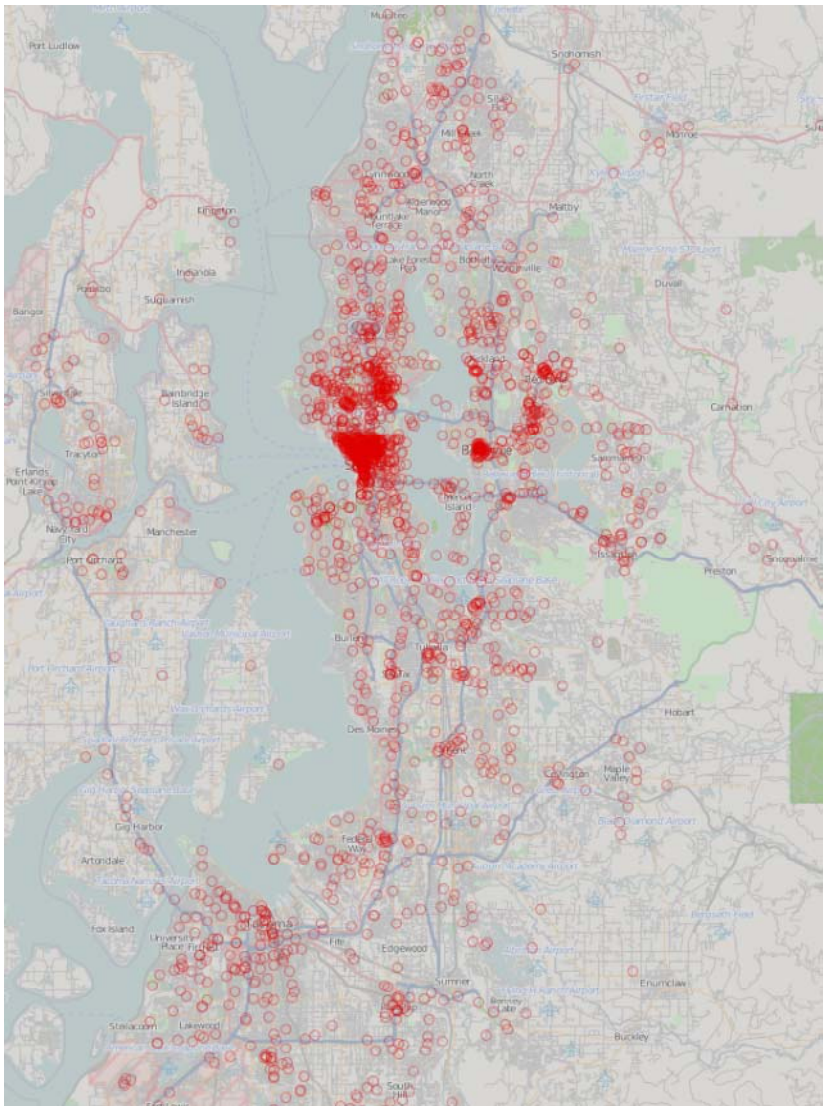

Figure 9: Medoids of Twitter users' trajectories.

The distribution of the medoids representing spatial behavior of Twitter users in the data set is correlated with the 2011 distribution of the population density (US Bureau of Census) in the Puget Sound Metropolitan area (Pearson's r = 0.52, T-test significant at 99%). This means that there is a moderate but significant relationship between the spatial distribution of places where people tweet and the population density (here we used US Census population density data, which does not distinguish between day-time and night-time population).

The medoids serve in our analysis not only a purpose of facilitating the visualization of spatial behavior of Twitter users but also as anchor points for analyzing the spatial patterns of keywords. We summarize the keyword feature vectors of the messages for each Twitter user. The binary attributes are attached 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.

3.4 Interactive Hypothesis Evaluation

To further validate or falsify hypotheses that have been developed using our overview and anomaly indication we provide techniques for an explorative detail analysis of individual messages. Once spatial and temporal points of interest have been identified, an interaction technique called 'Content Lens' can be used to inspect the contents of tweets sent from specific locations by showing either the most prominent or the most unusual terms[13]. This technique can be combined with temporal and textual filters to contextualize messages that have been written in a certain time-range and contain specific keywords.

In order 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 regions of the district. In addition to highlighting the most prominent words, the lens also selects the individual messages and reveals the nature of the chatter when we find a student telling the world how he skips class today to do more sport (see Figure 10 left).

The large "music" cluster around Renton (see Figure 6) seems odd. Placing the Content Lens over the affected regions reveals the single words "song" and "listen" to be among the top terms, which is unusual as music is normally referred to 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 10 right).

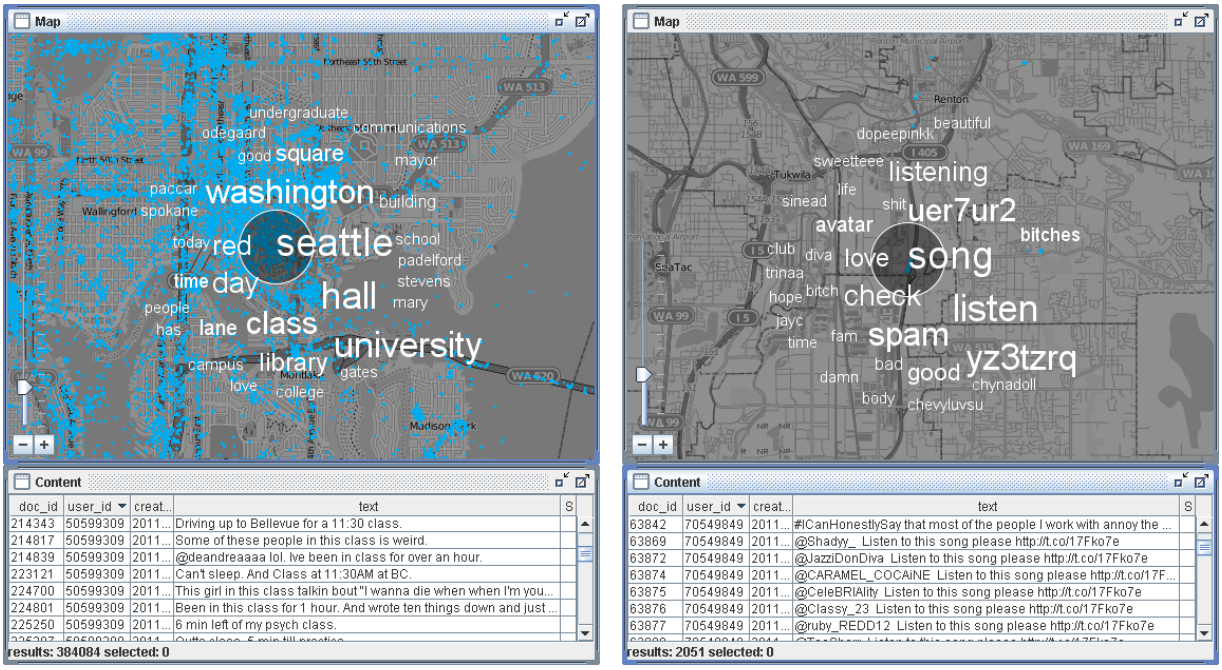

Figure 10: Investigating the details of the findings using the Content Lens (circles on map) and the tweet contents (table)

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

4 Summary

To understand the information potential of large collections of georeferenced text messages posted through public microblogging services such as Twitter, we explored a set of Twitter messages collected for a two-month period in the greater Seattle area. We used Visual Analytics methods, which are particularly suitable for exploratory analyses. We have achieved very good results by addressing analysis questions not covered by other studies before. Rather than trying to detect abrupt or abnormal events, we explored the normal life and daily routines of city residents. In particular, we made several interesting findings and got a feeling for the various aspects of life in Seattle in space and time. Generally, our study shows that georeferenced messages posted by ordinary citizens can be a source of interesting information about people and the space where they live. This information is potentially valuable for city planning, advertizing and other businesses, and for social sciences.

We also identified several directions for future research. One obvious challenge is the scalability of our approaches – Twitter generates truly massive amounts of data. We already conducted research towards scalable spatio-temporal clustering[14, 15] that could be integrated with the methods presented here. We are going to consider data at larger temporal scales as well as perform comparisons between different cities or regions. Including further mobility characteristics – such as trajectory patterns, significant and/or personal places of users[12] – in the analysis will lead to an even better understanding of the spatio-temporal phenomena we observed; although this may also raise privacy issues[16].

Further, the analysis presented here has been conducted using a pre-collected, static data set. We want to extend this to work on the real-time stream of Tweets directly, which has several implications on the analysis methods, topic modeling in particular. We also intend to include spatio-temporal sentiment analysis in general, and with respect to specific topics.

Acknowledgements

This work was partially funded by the German Federal Ministry for Education and Research (BMBF) as part of the VASA project (http://www.va-sa.net/).

References

1. T.Sakaki, M.Okazaki, and Y.Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. Proceedings of the *19th international conference on World wide web* (WWW '10), 2010, pp. 851-860
2. A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. Proceedings of *WebKDD/SNA-KDD'07*, 2007, pp. 56-65
3. Z. Qu and Y. Liu. Interactive group suggesting for Twitter. Proceedings of *HLT '11*, 2011, pp. 519-523
4. J.Chae, D.Thom, H.Bosch, Y.Jang, R.Maciejewski, D.Ebert, and T.Ertl. Spatiotemporal Social Media Analytics for Abnormal Event Detection using Seasonal-Trend Decomposition. Proceedings of *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2012
5. E. Kouloumpis, T. Wilson, and J. Moore. Twitter Sentiment Analysis: The Good, the Bad and the OMG! *Artificial Intelligence*, 2011, pp. 538-541
6. S. Carter, M. Tsagkias, and W. Weerkamp. Twitter hashtags: Joint Translation and Clustering. Proceedings of the *ACM WebSci'11*, 2011, pp. 1-3
7. K. Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. Part-of-speech tagging for Twitter: annotation, features, and experiments. Proceedings of *HLT'11*, 2011, pp. 42-47
8. X. Liu, S. Zhang, F. Wei, and M. Zhou. Recognizing named entities in tweets. Proceedings of *HLT'11*, 2011, pp. 359-367
9. H. Saif, Y. He, and H. Alani. Alleviating Data Sparsity for Twitter Sentiment Analysis. Proceedings of *Making Sense of Microposts (MSM2012)*, 2012
10. D. Thom, H. Bosch, S. Koch, M. Wörner and T. Ertl. Spatiotemporal Anomaly Detection through Visual Analysis of Geolocated Twitter Messages. Proceedings of IEEE Pacific Visualization Symposium, 2012
11. N.Andrienko, G.Andrienko. Spatial Generalization and Aggregation of Massive Movement Data. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2011, v.17 (2), pp.205-219
12. G.Andrienko, N.Andrienko, C.Hurter, S.Rinzivillo, S.Wrobel. From Movement Tracks through Events to Places: Extracting and Characterizing Significant Places from Mobility Data. Proceedings of *IEEE Visual Analytics Science and Technology (VAST), 2011*, pp.161-170
13. D. Thom, H. Bosch and T. Ertl. Inverse Document Density: A Smooth Measure for Location-Dependent Term Irregularities. Proceedings of International Conference on Computational Linguistics (COLING), 2012

14. I.Peca, G.Fuchs, K.Vrotsou, N.Andrienko, and G.Andrienko.Scalable Cluster Analysis of Spatial Events. Proceedings of *International Workshop on Visual Analytics*, 2012, pp.19-23
15. G.Andrienko, N.Andrienko, C.Hurter, S.Rinzivillo, S.Wrobel. Scalable Analysis of Movement Data for Extracting and Exploring Significant Places. *IEEE Transactions on Visualization and Computer Graphics*, 2013, v. 19 (accepted)
16. G.Andrienko and N.Andrienko. Privacy Issues in Geospatial Visual Analytics. Proceedings of the *8th Symposium on Location-Based Services (LBS)*, 2011, pp.239-246