Data-Driven Exploration of Real-Time Geospatial Text Streams

Harald Bosch(✉), Robert Krüger, and Dennis Thom

Institute for Visualization and Interactive Systems,
University of Stuttgart, Universitätsstr. 38, 70569 Stuttgart, Germany
{bosch,krueger,thom}@vis.uni-stuttgart.de

Abstract. Geolocated social media data streams are challenging data sources due to volume, velocity, variety, and unorthodox vocabulary. However, they also are an unrivaled source of eye-witness accounts to establish remote situational awareness. In this paper we summarize some of our approaches to separate relevant information from irrelevant chatter using unsupervised and supervised methods alike. This allows the structuring of requested information as well as the incorporation of unexpected events into a common overview of the situation. A special focus is put on the interplay of algorithms, visualization, and interaction.

Keywords: Stream processing · Machine learning · Social media

1 Introduction and Problem Definition

Social media data comprises highly relevant information about events such as natural and technological disasters, crimes, and infectious diseases (e.g., see [1,2]). Separating event-related data from noise such as chatter and speculations, however, is a challenging task. In this work, we summarize the contributions of our former publications [3–6] regarding the interplay of interactive visualization and (un)supervised machine learning approaches for gaining insights into massive, real-time, geolocated message streams. From these contributions, a decision support system was built for (1) keeping an overview over the current situation, (2) organize massive data streams while incorporating unexpected events, and (3) adapt the selectivity and orchestration of filters to the current situation. We further discuss on the synergistic effects between machine learning, information visualization, and human-computer-interaction. The following primarily addresses Twitter as a data source and public safety as the application domain but the approaches are applicable to others streams and applications as well.

2 Stream-Enabled X-means Clustering for Textual Data

As a means to incorporate unexpected incidents into situation awareness, an unsupervised approach is used to find potential real-world events by identifying

A. Bifet et al. (Eds.): ECML PKDD 2015, Part III, LNAI 9286, pp. 203–207, 2015.
DOI: 10.1007/978-3-319-23461-8_14

Fig. 1. Term usage clusters during a large earthquake in August 2011. The label size indicates the cluster's significance. Color indicates term relevance. By zooming in more space becomes available, clusters disaggregate and smaller clusters are shown.

accumulations of similar term usage in social media. The assumption is, that when people suddenly use similar words at the same location and the same time, they might have observed something unusual. We developed a stream-enabled variant of X-means to find such accumulations by clustering spatiotemporal coordinates for each observed term separately [3]. For each term a set of clusters is maintained. When a new message arrives, the cluster of each contained term is updated by adding the message coordinate to the nearest cluster. If the variance of the cluster exceeds a predefined threshold, it is split into two clusters. Because the temporal dimension increases monotonously, old clusters can become 'stale' when they will never be selected as the nearest cluster due to the large temporal difference. At this point, the stale cluster is evaluated for significance to dismiss cluster centroids that only cover sparse regions related to noise in the data. The criteria for significance are a low variance and the participation of multiple message authors as a means to exclude spamming users. This approach guarantees that spatiotemporally dense regions of similar term usage are covered by at least one, usually more cluster centroids and each centroid is related to a location instead of global chatter. The current state of the clustering is a list of locations (cluster centroid) for each term and a significance (the cluster's member count and variance) that can be used directly to place tags on a geographical map, using the significance for the font size (see Figure 1).

Overfitting as a Gift not a Curse. As indicated, our cluster splitting method tends to overfitting by creating too many centroids, especially if a cluster is spatially dense, but temporally elongated. This is an inherent problem of using a fixed variance threshold as splitting criterion because it cannot discriminate between, e.g., a sporting event in a stadium or a natural disaster covering a whole county. Luckily, the explorative data visualization can make use of this in order to offer a structural zooming interaction. In a global view all neighboring tags of a term can be aggregated into a larger tag by summing their significance, thus creating only one but larger tag. When users explore an area by zooming into it, the aggregation is less pronounced and reveals the geographical extent of

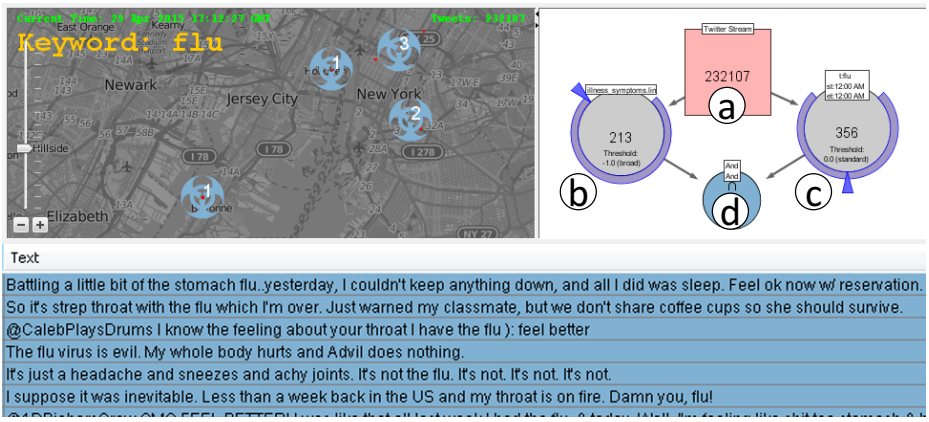

Fig. 2. New York influenza epidemic, January 2013. To filter the stream (a) for messages indicating illness we apply a pre-trained symptom classifier (b). Narrowing the results further with a term-based 'flu' filter (c) reveals all tweets indicating flu symptoms (d).

the event. The same benefit exists in the temporal domain. When the observed time frame is reduced, the amount of centroids used for creating tags is reduced alike.

Statistical Means for Relevance Indication. A problem for the clustering is, that there are 'stable' accumulations which are always dense enough to create significant centroids, most notably the names of touristic sights and cities. Social media authors will always use the terms 'London' and 'Eye' near the aptly named Ferris wheel. While being certainly significant accumulations, they are most often irrelevant due to missing novelty. Here, a temporal analysis can judge the novelty of a cluster. One approach takes the local history of the term usage and extracts the Seasonal and the Trend component using Loess (STL). If the remainder component, not explained by season or trend, is significantly higher than expected (z-score ≥ 2) the anomaly is considered novel [5]. Another approach is a pre-calculation of a smooth measure for location-dependent document densities for each term from a large collection of geolocated messages [4]. With this measure, the current term frequency can be contrasted with the expected value for this location. If deemed novel, the tag is highlighted in the tag map (Figure 1).

3 Orchestrating Filters Based on Classification

Unsupervised learning methods like the described clustering are powerful means to discover the unexpected. However, domain experts already have certain hypotheses or specific topics to search for. For example, a disaster prevention agency would be interested in catastrophe related information, while a health

department might look for epidemics. To structure this a priori known information need we apply supervised learning algorithms [6]. Here, historical data is used to learn the vocabulary employed by social media authors to describe the desired event types. After a related event occurred, the constantly recorded message archive is filtered according to the location and time of the event and are labeled by analysts as relevant or irrelevant regarding the topic, e.g., an earthquake, disease outbreak, or similar. The labeled subset of past observations is then used to train a Support Vector Machine (SVM) classifier. A linear kernel is sufficient due to the high dimensionality of textual data. With a graph-based, interactive user interface the pre-trained classifiers can be loaded and combined ad-hoc during the live-monitoring of a new event to cover a multitude of possible situations (see Figure 2). Here, the combinations of classifiers can be used to assign messages to sets and track them throughout different views of the system to structure the analysis and avoid information overload. To this end, the classifiers' selectivity can be further adapted by the analysts through shifting the SVM decision boundary. This allows an ad-hoc trade-off between precision and recall.

4 Conclusion

Machine learning facilitates the exploration of massive spatiotemporal text streams. While unsupervised techniques are means for aggregated overviews and discovering the unexpected, supervised approaches can narrow down streams to specific topics. Interactive interfaces with suitable visualization can adjust classifications without opening the black-box and make these techniques accessible to domain experts without machine learning background. Moreover, interactive visual means can turn common drawbacks like overfitting and model inflexibility to strength—in our case by interactive structural zooming and model orchestration.

Acknowledgments. This work was supported by the BMBF project *VASA* project and the Horizon 2020 project *CIMPLEX*.

References

1. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: Real-time event detection by social sensors. In: Int'l Conf. WWW, pp. 851–860 (2010)
2. Chew, C., Eysenbach, G.: Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One **5**(11) (2010)
3. Thom, D., Bosch, H., Koch, S., Wörner, M., Ertl, T.: Spatiotemporal Anomaly Detection through Visual Analysis of Geolocated Twitter Messages. In: IEEE Pacific Visualization Symposium, pp. 41–48 (2012)
4. Thom, D., Bosch, H., Ertl, T.: Inverse Document Density: A Smooth Measure for Location-Dependent Term Irregularities. In: COLING Conf., pp. 2603–2618 (2012)

5. Chae, J., Thom, D., Bosch, H., Jang, Y., Maciejewski, R., Ebert, D.S., Ertl, T.: Spatiotemporal Social Media Analytics for Abnormal Event Detection and Examination using Seasonal-Trend Decomposition. In: IEEE Conference on Visual Analytics Science and Technology, pp. 143–152 (2012)
6. Bosch, H., Thom, D., Heimerl, F., Püttmann, E., Koch, S., Krüger, R., Wörner, M., Ertl, T.: ScatterBlogs2: Real-Time Monitoring of Microblog Messages Through User-Guided Filtering. IEEE Trans. Vis. Comput. Graphics **19**(12), 2022–2031 (2013)