Hybrid geo-information processing:
Crowdsourced supervision of geo-spatial machine learning tasks

Frank O. Ostermann
Department of Geo-Information Processing
Faculty of Geo-Information Science and Earth Observation (ITC)
University of Twente
PO Box 217
7500 AE Enschede, The Netherlands
f.o.ostermann@utwente.nl

Abstract

This paper introduces an approach to crowdsource the supervision of machine learning classification and regression tasks in order to process geo-social media streams. It builds on a review and comparison of four existing approaches to process geo-social media streams in order to identify specific opportunities and challenges. An original conceptual framework situates the machine learning tasks within a geo-information processing workflow. The paper presents and discusses concrete techniques and software solutions for implementing it.

Keywords: crowdsourcing, supervised machine learning, geo-social media streams, user-generated geographic content, volunteered geographic information.

1 Introduction

The new modes of production and consumption of geographic information offer great opportunities for improved near real-time decision making of individuals and organizations, from small daily tasks like where to buy groceries or park your car, to critical tasks like coordinating humanitarian crisis response and saving lives. As has been argued before [5], there are several challenges to tackle before we can harness the full potential of geo-social media streams that contain volunteered geographic information or user-generated geographic content (UGGC). These include the heterogeneity and volatility of data formats and sources, the sheer volume of information streams that can change rapidly, and the credibility and accuracy of the information.

The main objectives of this short paper are threefold. First, to explore the state-of-the-art of comprehensive approaches to processing geo-social media streams. Second, to identify the potential of crowd-sourced supervised learning to improve the handling of UGGC. Third, to develop a conceptual architecture based on a methodological review that serves as a first step to the implementation and testing of a prototype.

2 Processing Geo-Social Media Streams

Numerous studies investigated UGGC, including event detection and information processing for disaster management [17], citizen science observations on phenology [2], neo-cartography [13], health-related inquiries [14], environmental monitoring [6], and the automatic description of places [19].

Although often producing astonishing and useful results, all efforts had their effectiveness reduced by the UGGC's unknown quality - the uncertainty about its uncertainty. UGGC often shows areas of data scarcity and data abundance or even data redundancy [9, 10]. Its availability seems to follow a positive feedback loop, increasing the inequalities and generating false data shadows [21]. Further, content does not arrive in anticipated packages, but as highly varying streams that often need processing in near real-time. UGGC formats are heterogeneous, with content commonly being unstructured. Finally, the credibility of the source and the quality of the data is often obscure, with the source's level of expertise and the motivation for participation only indirectly inferable [7], and the origins of inaccuracies manifold, e.g. conflicting feature classifications between countries or cultures, outdated or wrong information available to the contributor, or even malicious intent. In summary, effective and efficient information retrieval for UGGC remains a difficult task [16].

A common response to these challenges has been the crowdsourcing of curation tasks [23]. Despite encouraging results, the curation by human volunteers might not be as accurate as once hoped for, and the process itself is not replicable [4]. Further, it might not scale up well, with organizational overhead increasing exponentially with volume of data to be curated. Lastly, it faces problems of sustainability, with the risk of external factors (vacations of volunteers, remote affected areas) influencing negatively the numbers of volunteers.

A more recent response tries to address these shortcomings by employing machine learning techniques to automatically select, filter, classify, and enrich UGGC. While unsupervised machine learning focuses on detecting hitherto unknown patterns, supervised machine learning tries to classify unknown data or predict values based on trained (learned) classifiers, mostly involving human annotators to create a training data set. Challenges include the dependency on data quality for unsupervised machine learning [12], overfitting of the learning model [3], and diversity of contexts and tasks [17]. The following paragraphs briefly describe four existing implementations:

The AIDR system [11] uses adaptive aggregation and filtering of Twitter, integrating crowd-sourced labelling to learn rules to filter and classify social media information. It focuses on the content. It is open source and allows near real-time processing. However, currently AIDR relies on a single source (Twitter) and does not consider the geographic semantics of information.

The CrisisTracker system as described in [18] is capable of detecting events and aggregates and filters social media about them, creating stories of clustered social media. However, it uses only one source (Tweets) and focuses entirely on the content for analysis, disregarding geographic semantics. It employs part of AIDR for the Tweet classification and is open source.

The Twitcident [1] aggregates and filters social media around events extracted from emergency broadcasting services. It semantically enriches the incoming information and links it with other external information. However, location only seems to influence the filtering of the information and not the assessment. It seems that a recent extension (Crowdsense) enables it to use several social media sources. No source code could be found.

The GeoCONAVI system [22] is capable of detecting past forest fire events. In contrast to the previous systems, it uses multiple sources (Tweets, Flickr images), and exploits geographic semantics by contextualizing the UGGC geographically, and clustering it spatio-temporally. The processing is done in near real-time, i.e. high-frequency batch processing. The content classification employs decision tree learner trained on an event-specific annotated data set. Case study results [17] show a low false positive rate (high specificity), and a low false negative rate (high sensitivity). It does not examine the source, nor does has it been adapted to other event types yet: The effort in supervising the learning process was substantial, and another context – e.g. different languages, geographic scope, or disaster type – would require new training and supervision.

The following section suggests a novel approach (hybrid geo-information processing) to combine the strengths - adaptability of crowdsourced supervision, and analytic power of geographic semantics - of the shown approaches.

3 Hybrid geo-information processing

This review of approaches suggests that a combination of geographic analysis and crowdsourced supervised machine learning technique have a great potential to improve results.

The specific challenges are how to (i) link the characteristics of geographic information with machine learning class labelling and regression, (ii) provide a multi-modal interface to let human oracles simultaneously label instances, (iii) translate the learner models into nomothetic principles on geographic semantics.

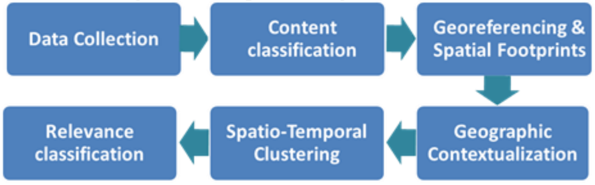
Ad (i): Every UGGC instance needs multi-class labelling on several attributes: The content type (e.g. call for help, offer for help, information on geographic features or processes), any locations mentioned and geographic footprints of locations and/or events (relevant or affected geographic area), distinct event membership (e.g. related to a particular earthquake, or emergency situation, or public event), and credibility based on a combination of the other class labels (e.g. an request for help related to a forest fire event which is coming from a distant desert is unlikely to be true). The learners have to deal with the specific characteristics of geographic information, i.e. spatial autocorrelation, vague boundaries and class memberships, and uncontrolled variance.

Ad (ii): A human oracle has to annotate instances for all model classes described in (i), with queries detailed further below. The responses will not only modify the learners, but also the parameters used for the geographic analysis steps to compute footprints and clusters.

Ad (iii): The resulting models will indirectly encode the semantic similarity of geographic places and concepts. To allow the results to be shared most effectively, the geographic concepts and places will be referenced to the most important linked data repositories such as DBpedia and GeoNames when possible.

Previous work has already explored the possibilities of using crowd-sourced supervision for machine learning tasks within the bigger framework of a Digital Earth Nervous System [15]. This study focuses on the specific geoprocessing steps show below in Figure 1, which are based on the GeoCONAVI approach:

Figure 1: Geoprocessing work flow


Data Collection
Content classification
Georeferencing & Spatial Footprints
Geographic Contextualization
Spatio-Temporal Clustering
Relevance classification

Source: The author

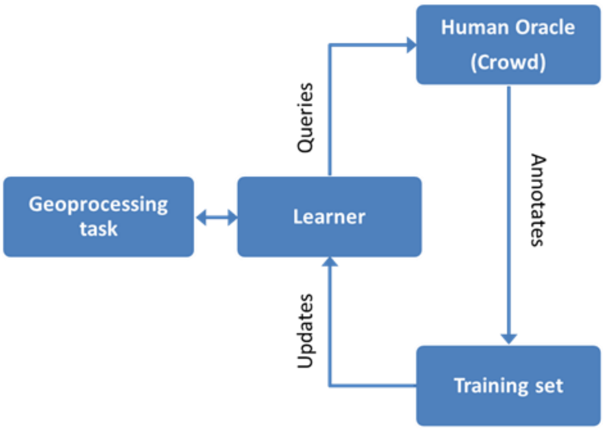
After collecting an UGGC instance, the first task can be an early content classification if the use case is sufficiently well defined so that this step can reduce noise reliably. Then, the processor needs to analyse the UGGC for place names (toponyms) using natural language processing procedures such as named entity recognition. Of particular interest here are toponyms that are natural features or vernacular (non-official). Found toponyms need then disambiguation (since many toponyms exist multiple times throughout the world). If there are multiple locations mentioned - either in the content, or through additional meta-data such as geographic coordinates from a global navigation satellite system – these need to be synthesized into a geospatial footprint, in order to allow geographic contextualization with additional ancillary (authoritative) data. Next, the individual content items can be grouped (clustered) to detect patterns in space in time, e.g. reports on the same incident, in order to remove redundant content, or reinforce the credibility of kept content. Finally, the instance can be classified according to its relevance to a particular topic or use case, integrating the results from previous classifications for an assessment of credibility and information gain.

As outlined above, the idea is to use the feedback from the crowdsourced supervision to help parameterize the geoprocessing tasks. A promising machine learning strategy is active learning, in which a learner chooses instances to be labelled and presents them to the human annotator [20]. Active learning provides an opportunity to maximize the impact of human annotation, while allowing the learner to remain flexible towards new instances, since it does not assume a static pool of instances. [20] reviews several possible types of query strategies that are suitable for a geographic active learner. An actual implementation of the machine learning tasks mostly likely will utilize at first:

- Stream-based selective sampling: The learner samples an instance and decides to query it or not; it is a well-established in practice, e.g. for word sense disambiguation.
- Density-weighted margin-based uncertainty sampling: A frequently employed and well-research strategy, mostly for classification tasks; density-weighting avoids the choice of outliers which are uncertain to classify, but will not improve a model's performance
- Ensembles of decision trees for classification and regression tasks (parameterisation of geographic analysis).

Building on lessons learned in an iterative implementation design, later options include testing of support vector machines or other more advance machine learning techniques. A sample extension of the workflow in Figure 1 is shown in Figure 2 below:

Figure 2: Geo-social media processing extended with machine learner.



Source: The author

The learner improves its performance by asking the following example queries:

- Toponym disambiguation, e.g. by asking “Does this [item] talk about [location A] or [location B], or none, or both?”
- Spatial footprint calculation for vague or multiple geographies, e.g. by asking “Is this spatial footprint for [item] correct? If not, is it too large, too small, or wrong shape, or wrong place?”
- Spatio-temporal clustering, e.g. by asking “Does this [item] belong to a cluster named [event] in [location]? If not, what’s wrong: Event, Location, or both?”

Two challenges are the formulation of the queries, and the development of strategies to deal with multiple, potentially noisy human oracles

Following from the above, there are several desirable criteria for processing platform of geos-social media streams: open source, near real-time processing, multiple sources, geographic semantics (cross-validation with and/or enrichment from external sources of geographic information), extendibility, and flexibility to adapt to new tasks and events.

Most of the analytical tasks (geoprocessing and active learning) could be performed in a stream processing framework like Apache Storm. The tasks will have to be disaggregated into atomic ones. The crowdsourced supervision is possible within the Pybossa framework, constantly updating a training data set. It is important to not only rely on simple crowdcrafting-style interfaces – use everything that's possible, including text messages and gamification elements, e.g. MicroMappers

Preferably, the full workflow should be implemented on a cloud computing platform in order to benefit from the cloud computing characteristics of scalability, elasticity, reliability, and availability. Scalability and elasticity allow a process to adapt the varying volume of UGGC streams. Increased reliability provides fail-safe mechanism for critical applications such as crisis responses. Finally, cloud computing platforms improve the availability of crowdsourcing tasks by reducing the bandwidth needed by the user, i.e. an end-user does not need to have full access to UGGC streams s/he is about to help processing.

4 Discussion and outlook

The majority of the systems presented in the previous section focus on the content to process the information. However, human activities and events take place in geographic space and time, and spatio-temporal context of information provides valuable clues to the information’s credibility, accuracy, and relevance for a particular task.

All of the approaches presented rely on a combination of human and machine intelligence. The involvement of the human intelligence varies greatly between systems, however. Most require humans to create a set of rules or train a machine learning algorithm for filtering and aggregating at the beginning (e.g. Twitcident, GeoCONAVI, AIDR), and leave the interpretation of the resulting information to humans. While Twitcident allows a user to interactively adapt the interface and search for related information, only the AIDR system currently uses crowd-sourcing to adapt the filtering and classification rules during run-time, with promising results for the detection of valuable information over the course of a changing event.

Such a combined approach could help to solve some of the particular challenges that the inclusion of spatio-temporal location entails, as [17] has shown: Geocoding of the geo-social media streams will remain necessary even if more information comes geo-referenced, because the content might be about one or more other locations than the information’s origin. Geocoding itself faces the problem of toponym disambiguation, a task which is complicated because of the brevity and lack of structure of much geo-social media [8]. Further, many of the geographical references might be in vernacular or abbreviated form. Crowd-sourced supervised learning shows promise for situational toponym resolution. An unresolved problem of any statistical inference process is also that a high sensitivity (needed to detect or follow small events) has a reduced specificity as trade-off. Crowd-sourcing can help to increase specificity by eliminating false positive event clusters (or stories). Finally, a crowd-sourcing effort can help to update geographic datasets used in the contextualization, introducing a positive feedback loop to improve classification and quality assessment.

Two future implementation strategies emerge: (i) extension of AIDR with GeoCONAVI functionality, or (ii) extension of GeoCONAVI with facilities to crowd-source the supervision of machine learning tasks and the parameterization of analysis function. The next step will be to decide on a concrete strategy, followed by a step-wise, iterative implementation and testing of the geoprocessing tasks described in this paper.

References

[1] Abel, Fabian, Claudia Hauff, Geert-Jan Houben, Richard Stronkman, and Ke Tao. 2012. "Twitcident: Fighting Fire with Information from Social Web Stream." In *International Conference on Hypertext and Social Media, Milwaukee, USA*. ACM.

[2] Brunsdon, C., and L. Comber. 2012. "Assessing the Changing Flowering Date of the Common Lilac in North America: A Random Coefficient Model Approach." *GeoInformatica* 16 (4): 675–90.

[3] Butler, Declan. 2013. "When Google Got Flu Wrong." *Nature*, February 13.

[4] Camponovo, Michael E., and Scott M. Freundschuh. 2014. "Assessing Uncertainty in VGI for Emergency Response." *Cartography and Geographic Information Science* 41 (5): 440–55.

[5] Craglia, M., F. Ostermann, and L. Spinsanti. 2012. "Digital Earth from Vision to Practice: Making Sense of Citizen-Generated Content." *International Journal of Digital Earth* 5 (5): 398–416.

[6] D'Hondt, Ellie, Matthias Stevens, and An Jacobs. 2013. "Participatory Noise Mapping Works! An Evaluation of Participatory Sensing as an Alternative to Standard Techniques for Environmental Monitoring." *Special Issue on Pervasive Urban Applications* 9 (5): 681–94.

[7] Flanagin, Andrew, and Miriam Metzger. 2008. "The Credibility of Volunteered Geographic Information." *GeoJournal* 72 (3): 137–48.

[8] Gelernter, Judith, and Nikolai Mushegian. 2011. "Geo-Parsing Messages from Microtext." *Transactions in GIS* 15 (6): 753–73.

[9] Graham, M., B. Hogan, R.K. Straumann, and A. Medhat. 2014. "Uneven Geographies of User-Generated Information: Patterns of Increasing Informational Poverty." *Annals of the Association of American Geographers* 104 (4): 746–64.

[10] Haklay, M. 2010. "How Good Is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets." *Environment and Planning B: Planning and Design* 37 (4): 682–703.

[11] Imran, Muhammed, Carlos Castillo, Ji Lucas, Patrick Meier, and Jakob Rogstadius. 2014. "Coordinating Human and Machine Intelligence to Classify Microblog Communications in Crises." In *Proceedings of the 11th International ISCRAM Conference*. ISCRAM.

[12] Kanevski, M., A. Pozdnoukhov, and V. Timonin. 2008. "Machine Learning Algorithms for GeoSpatial Data. Applications and Software Tools." In *Proceedings of the 4th Biennial Meeting of iEMSs*.

[13] Liu, Sophia B., and Leysia Palen. 2010. "The New Cartographers: Crisis Map Mashups and the Emergence of Neogeographic Practice." *Cartography and Geographic Information Science* 37 (1): 69–90.

[14] Mooney, Peter, Padraig Corcoran, and Blazej Ciepluch. 2013. "The Potential for Using Volunteered Geographic Information in Pervasive Health Computing Applications." *Journal of Ambient Intelligence and Humanized Computing* 4 (6): 731–45.

[15] Ostermann, Frank O., and Sven Schade. 2014. "Multi-Sensory Integration for a Digital Earth Nervous System." In *Proceedings of AGILE 2014*. Castellon, Spain: AGILE.

[16] Ostermann, Frank O., Martin Tomko, and Ross Purves. 2013. "User Evaluation of Automatically Generated Keywords and Toponyms for Geo-Referenced Images." *Journal of the American Society for Information Science and Technology* 64 (3): 480–99.

[17] Ostermann, Frank, and Laura Spinsanti. 2012. "Context Analysis of Volunteered Geographic Information from Social Media Networks to Support Disaster Management: A Case Study On Forest Fires." *International Journal of Information Systems for Crisis Response and Management* 4 (4): 16–37.

[18] Rogstadius, J., M. Vukovic, C.A. Teixeira, V. Kostakos, E. Karapanos, and J.A. Laredo. 2013. "CrisisTracker: Crowdsourced Social Media Curation for Disaster Awareness." *IBM Journal of Research and Development* 57 (5): 4:1–4:13.

[19] Ross Purves, Alistair Edwardes, and Jo Wood. 2011. "Describing Place through User Generated Content." *First Monday; Volume 16, Number 9 - 5 September 2011*.

[20] Settles, Burr. 2009. *Active Learning Literature Survey*. Computer Sciences Technical Report 1648. Madison: University of Wisconsin.

[21] Shelton, Taylor, Ate Poorthuis, Mark Graham, and Matthew Zook. 2014. "Mapping the Data Shadows of Hurricane Sandy: Uncovering the Sociospatial Dimensions of 'big Data.'" *Geoforum* 52 (0): 167–79.

[22] Spinsanti, Laura, and Frank Ostermann. 2013. "Automated Geographic Context Analysis for Volunteered Information." *Applied Geography* 43 (0): 36–44.

[23] Sui, Daniel, Sarah Elwood, and Michael F Goodchild, eds. 2012. *Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice*. Berlin: Springer.