Volunteered Geographic Information and Data Quality – The Case of Social Reporting

Olga Yanenko

University of Bamberg, Chair of Computing in the Cultural Sciences

Introduction

Content created by internet users and enriched with geographical footprints is usually referred to as Volunteered Geographical Information (VGI; Goodchild 2007). This new source of information enables the collection of rich data sets that often outperform the data collected by private companies and government agencies in terms of the amount of data producers and their personal motivation to contribute (Goodchild 2008, Goodchild and Glennon 2010).

In spite of the potential of VGI, user-generated data production is often accompanied by problems that have to be overcome to ensure usable and valuable data sets. Four main challenges were identified in literature, namely the motivation of users to contribute, the quality of the resulting data and the spatial and temporal coverage of the latter (Coleman et al. 2009, Feick & Roche 2013, Flanagin & Metzger 2008).

This work studies the issue of data quality in the different stages of VGI data creation, collection and evaluation. It concentrates on social reporting scenarios, in which citizens submit geotagged reports about observations of real-world-events. The following basic research questions are addressed within this work:

1. How can users be motivated to contribute *correct* and *truthful* data?
2. How does *gamification* affect data quality?
3. What (semi-)*automated* approaches can be used to improve and evaluate data quality?
4. How can agreement and disagreement between different data producers be *modeled*, *evaluated* and *interpreted*?
5. What is the difference of *objective* and *subjective* data and what principles can be used to validate both?

 In: A. Comber, B. Bucher, S. Ivanovic (eds.): Proceedings of the 3rd AGILE Phd School, Champs sur Marne, France, 15-17-September-2015, published at http://ceur-ws.org

Related Work

The data quality of VGI highly depends on the motivation of people to collect data and produce reports (Coleman et al. 2009; Feick & Roche 2013; Flanagin and Metzger 2008). Gamification has proven to be an effective method for motivating people to collect data about geographical places (Matyas et al. 2008). There exists considerable research describing how gamification can be used in different contexts to increase the amount of produced data. But there is still only little work studying how gamification affects the quality of the resulting data sets. Since gaming is a competitive situation, the data collected in location-based games is often biased, incomplete or useless and there is a need for finding ways to ensure the data quality without degrading the game experience (Cramer et al. 2011, Cechanowicz et al. 2013).

The most reliable method for validating data is verification on-site as known from classical journalism. In most of the VGI cases this method is not suitable since many spatio-temporal events only exist for a short period of time and it is simply not possible to verify all of the reports on-site and in real-time.

A prominent validation approach is letting volunteers review and rate reports of other volunteers. These ratings are used to compute authority measures for individual reporters (Bishr and Mantelas 2008). Authority methods highly depend on former reputations of a reporter and fail for first-time and infrequent reporters. However, the idea of Bishr and Mantelas (2008) to use a spatial quality criterion for the computation of reputation values by taking the spatial distance between the reporter and the observed object into account, builds the basis for the more generalized spatio-temporal-proximity principle that is part of this work.

Methods and Results

The first part of the project consists in a systematic review of different motivations of reporters in the data creation process and their effect on data quality. The focus is on gamification as a motivational strategy and the development of game principles that are not only attracting people to participate but also motivate them to contribute in an honest way. An example of such an approach are the retesting and confirmation mechanisms included in a mobile VGI game as part of the game play without being recognized as internal control features by the players (Yanenko & Schlieder 2014). These game principles are also dealing with another exiting aspect of VGI, the differentiation between objective and subjective data. While objective or factual data can be confirmed by different individuals, the validation of subjective data is more complicated based on different perceptions.

The second part of the project deals with finding methods and principles for data validation of already existing data sets. Spatio-temporal proximity and social distance were identified as core principles for the integration of different data

sources (Schlieder & Yanenko 2010). The main idea is that the mutual confirmation of two reports with similar content is higher if they were recorded spatially and temporally close to each other – and thus obviously concern the same (spatial) event. Social reporting scenarios also have to address the issue of social bias: the events reported often depend on what social group the contributor belongs to. In such cases, social distance proves to be a useful validation criterion. Reports by contributors from different stake holder groups usually provide higher confirmation than reports from the same group. For computing the confirmation values between different reports, several functions and approaches were tested. Two experiments were performed with phenological data from the BudBurst[1] project: a simple model of the temporal relationship between the different phenophases for identifying incorrect data entries and a constraint-satisfaction approach for restricting the range of possible values for incomplete entries (Yanenko & Schlieder 2012).

The software part of the project consists of a generic report integration tool. Its functionality includes the construction of a confirmation graph between reports based on the principles described above with the possibility of individual adjustments. The flexible architecture allows for use-case-based selection of appropriate confirmation functions and methods that will be applied to construct the edges of the graph and compute the agreement values between two pieces of information (vertices of the graph). Finally, the individual values for singular reports are computed by an aggregation mechanism that combines all edges connected to a report into one veracity value. The tool will implement some white-spread algorithms such as PageRank (Page et al. 1999) but can also be customized for specific interests.

References

Bishr, M., Mantelas, L. (2008). A trust and reputation model for filtering and classifying knowledge about urban growth, *GeoJournal* 72(3), pp. 229-237.

Cechanowicz, J., Gutwin, C., Brownell, B., Goodfellow, L. (2013). Effects of gamification on participation and data quality in a real-world market research domain. In: *Proceedings of the First International Conference on Gameful Design, Research, and Applications* (Gamification'13). ACM, New York, NY, USA, pp. 58-65.

Coleman, D. J., Georgiadou, Y., & Labonte, J. (2009). Volunteered geographic information: The nature and motivation of produsers. In: *International Journal of Spatial Data Infrastructures Research*, 4(1), pp. 332-358.

Cramer, H., Ahmet, Z., Rost, M. and Holmquist, L. E. (2011). *Gamification & Location-sharing: some emerging social conflicts*. Gamification Workshop at CHI'11.

Feick, R., & Roche, S. (2013). Understanding the Value of VGI. In: *Crowdsourcing geographic knowledge*, Springer Netherlands, pp. 15-29.

Flanagin, A. J., Metzger, M. J. (2008). The credibility of volunteered geographic information. In: *GeoJournal*, 72(3-4), pp. 137-148.

[1] http://budburst.org/

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4), pp. 211-221.

Goodchild, M. F. (2008). Commentary: Whither VGI? *GeoJournal*, 72, pp. 239-244.

Goodchild, M. F., Glennon, J. A. (2010) Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth* 3(3), pp. 231-241.

Matyas, S., Matyas, C., Kiefer, P., Schlieder, C., Mitarai, H. and Kamata, M. (2008). Designing Location-based Mobile Games with a Purpose - Collecting Geospatial Data with CityExplorer. In: ACM International Conference Proceeding Series, Vol. 352, *Proceedings of the 2008 International Conference on Advances in Computer Entertainment Technology*, Yokohama, Japan, pp. 244-247.

Matyas, S., Kiefer, P., Schlieder, C., & Kleyer, S. (2011). Wisdom about the crowd: Assuring geospatial data quality collected in location-based games. In: *Entertainment Computing–ICEC 2011*. Springer Berlin Heidelberg, pp. 331-336.

Page, L., Brin, S., Motwani, R., Winograd, T. (1999). *The PageRank citation ranking: bringing order to the Web*.

Schlieder S., Yanenko O. (2010). Spatio-temporal Proximity and Social Distance: a Confirmation Framework for Social Reporting. In: *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks* (LBSN'10). ACM, New York, NY, USA, pp. 60-67.

Yanenko O., Schlieder C. (2012). Enhancing the Quality of Volunteered Geographic Information: A Constraint-Based Approach. In: Bridging the Geographic Information Sciences, *Lecture Notes in Geoinformation and Cartography*, Part 8, pp. 429-446.

Yanenko O., Schlieder C. (2014). Game Principles for Enhancing the Quality of User-generated Data Collections. In: AGILE'14 *Workshop on Geogames and Geoplay*. Castellón, Spain, June 3rd 2014.