ICC 2017
WASHINGTON
JULY 02-07, 2017

Volunteered Geographic Information: Interpretation, Visualisation and Social Computing (VGIscience)

Dirk Burghardt,[1] Wolfgang Nejdl,[2] Jochen Schiewe,[3] and Monika Sester[4]

1. Technical University Dresden, IfK, Dresden, Germany, dirk.burghardt@tu-dresden.de
2. Leibniz-University Hannover, L3S, Hannover, Germany, nejdl@kbs.uni-hannover.de
3. HafenCity University Hamburg, g2lab, Hamburg, Germany, jochen.schiewe@hcu-hamburg.de
4. Leibniz-University Hannover, IKG, Hannover, Germany, monika.sester@ikg.uni-hannover.de

Abstract: In the past years Volunteered Geographic Information (VGI) has emerged as a novel form of user-generated content, which involves active generation of geo-data for example in citizen science projects or during crisis mapping as well as the passive collection of data via the user's location-enabled mobile devices. In addition there are more and more sensors available that detect our environment with ever greater detail and dynamics. These data can be used for a variety of applications, not only for the solution of societal tasks such as in environment, health or transport fields, but also for the development of commercial products and services. The interpretation, visualisation and usage of such multi-source data is challenging because of the large heterogeneity, the differences in quality, the high update frequencies, the varying spatial-temporal resolution, subjective characteristics and low semantic structuring.
Therefore the German Research Foundation has launched a priority programme for the next 3-6 years which will support interdisciplinary research projects. This priority programme aims to provide a scientific basis for raising the potential of VGI- and sensor data. Research questions described more in detail in this short paper span from the extraction of spatial information, to the visual analysis and knowledge presentation, taking into account the social context while collecting and using VGI.

Keywords: Volunteered Geographic Information, user-generated spatial content, geographic information extraction, geosocial-visual analytics, visual communication, social context

1. Introduction

During the last years the availability of spatial data has rapidly developed, in particular through the diffusion of social networks, Web 2.0 platforms and availability of suitable sensor technologies. A clear indication of this trend is the participation of users to the production of so called "Volunteered Geographic Information", shortly VGI, due to the vast use of smartphones and mobile devices. In the current information society era, these data enable a variety of applications, not only for the solution of societal challenges such as in environment, health or transport fields, but also for the development of commercial products and services.

Goodchild (2007) introduced in conjunction with VGI the concept of "humans as sensor". Thereby the data generation is typically carried out separately from a professional context and without the requirement of formal qualifications. The collected data are freely available and can provide alternative, subjective perspectives. The concept of human sensors has been extended with the terms "collective sensing" and "citizen science". While the first case is related to the analysis of aggregated anonymised data coming from collective networks, the second term describes the utilisation of local experiences and knowledge from interested members of the general public.

In addition there are more and more sensors available that inspect our environment in an always greater spatial and temporal level of detail. Such sensor data are described at least partially with explicit semantic, such as the values of

a temperature sensor or the position of a vehicle. However, they also contain implicit information e.g. about traffic congestion. Both sensor data and user data may be available in the form of data streams with a local reference. Furthermore this data are typically event driven – i.e. they provide evidence that something happened.

Thus, VGI are very up-to-date data available with high spatial resolution, which contain information about the local environment and the behaviour of users in an implicit form. The usability of these data sets in their raw form nevertheless is limited because of the immense degree of heterogeneity, the partially unknown semantics and the varying quality. Furthermore the protection of privacy must be ensured in order to enable sustainable use. In order to unleash the large value and the richness of the data, a custom data processing needs to take place, i.e. the data must be appropriately preselected, text processed, interpreted and possibly semantically annotated. Here, user and task oriented analysis and visualisation of data sets is of particular importance in order to communicate information and support decision making.

The utilisation of VGI and analysis of user generated spatial content requires consideration of social context as well as the characteristics and interactions of the users. Therefore approaches from the new research field of Social Computing are beneficial. Social Computing is an area of Computer Science looking at the interfaces between social behaviour and computer systems; and dealing with methods and systems that support collective acquisition, processing and display of user generated content.

For these reasons, the German Research Foundation has established a priority program, which will support basic research on the topic „VGI: Interpretation, Visualisation and Social Computing“ in the coming years.

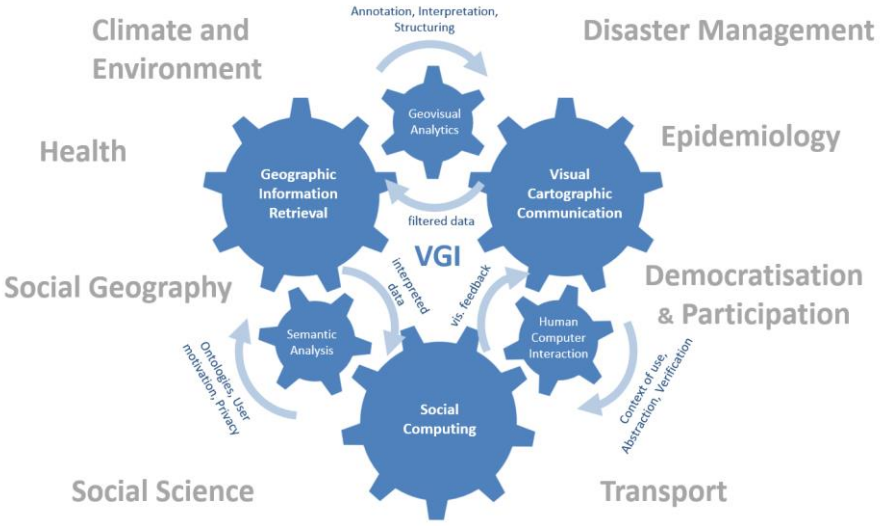

Fig. 1. Priority programme “VGIscience” with research areas and potential application fields

This paper lists research fields and question in summarised form, which are taken from the underlying research proposal. Section 2 identifies research projects with a thematic connection with the priority programme and Section 3 describes major research topics.

2. Related research projects

There are a number of application-specific projects related to the extraction and **interpretation of geographic information**, including:

- *Geographic Privacy-aware Knowledge Discovery and Delivery* (GeoPKDD, EU project, funding 2005 until 2008): The aim of the project was the development of theories, methods and systems for the extraction of

geographic information with consideration of privacy aspects. The focus was set on the analysis of moving objects and trajectory data on the example of car mobility.

- *WeSenseIT: Citizen Observatories of Water* (EU project, funding since 2012): The research project studied heterogeneous sensor networks consisting of physical and social sensors, which were utilised for hydrological modelling, as well as decision making in the area of water management. Therefore data were considered from social networks such Twitter and Facebook.
- *Event Prediction and Decision Support based on Huge Data from Physical-Social Systems* (EPPICS; EU project, funding 2014 until 2017): Within this project methods were developed for modelling, monitoring and prediction of events, through the extraction and combination of data and information from physical and social sensors. Application domains will be the event management for cities as well as flood monitoring and management.
- *Visual Linguistics: Theory and Application of Visualization in Linguistics* (SNF project, funding 2015 until 2018): Aim of the project is the development of a "visual linguistic framework". Therefore visualisation methods from related disciplines should be evaluated as well as their potential related to visual analyses of big (linguistic) data.
- *Citizen Observatory Web* (Cobweb, EU project, funding 2012 until 2016): In this project a system will be developed that will make it easier for citizens to collect environmental data, e. g. through the usage of mobile devices. A special focus will be set on the integration of citizen-sourced data with reference data from public authorities.

The mentioned projects deal particularly with technical and legal aspects (e.g. privacy, computer linguistic, visual linguistic), which certainly need to be taken up in future research (e. g. for the fusion of different data sources) – but in a new context – the processing and interpretation of VGI data for wide range of applications – and therefore also with an integrated perspective.

VGI data often have characteristics of **big data** (volume, heterogeneity, uncertainty, etc.). Thus there are also thematic intersections with related projects, such as:

- *INtelligent Synthesis and Real-tIme Response using Massive StreaminG of HeTerogeneous Data* (INSIGHT, EU project, funding 2012 until 2015): In this project a prototype has been developed for the fast situation assessment of emergency situations based on the analysis of streaming data. Data sources are heterogeneous ranging from twitter to traffic flow sensing as well as mobile phone data.
- *Algorithms for Big Data* (DFG-SPP 1736, funding since 2013): This priority programme will develop improved software tools and data structures for big data set. Applications will deal for example with 3D image analysis, efficient semantic search, processing of time series, exploration of large networks etc.

The focus of the big data projects is given either on general frameworks or very isolated method developments. Less attention has been paid on user generated spatial data as well as the intersection with the field of social computing.

In the context of **visual communication** of VGI also approaches from **visual analytics** will be very relevant – therefore the following projects provide important input:

- *Scalable Visual Analytics: Interaktive visuelle Analysesysteme für komplexe Informationswelten* (DFG-SPP 1335, funding 2008 until 2014): This priority programme dealt with the theoretic foundations of visual analysis algorithms, the development and practical implementation of scalable visual analysis techniques as well as their integration and evaluation.
- *Visual Analytics for sense-making in Criminal Intelligence Analysis* (VALCRI**,** EU project, funding 2014 until 2017): Aim of the project is the development of an interactive prototype for crime analysis for Europe based on different data source (text, image, video).
- *Visual Analytics for Security Applications* (VASA, funding 2011 until 2014): The aim of these german-american joint project was the development of visual analytics methods and processes to improve disaster prevention and crisis response with focus on critical infrastructures, digital networks and power grids.

Visual analytics approaches from the mentioned project will have great potential to be applied also on user generated spatial content. Focus of future developments will be given on social, thematic, spatial, and temporal aspects.

Related to **Social Computing** the following projects provide relevant insights:

- *Enhance Environmental Awareness through Social Information Technologies* (EveryAware; EU project, FET Open, funding: 2011 until 2014): Within this project a platform was developed to integrate low-cost sensor technologies, network applications and data processing which will help citizens to capture, share and understand their environment. Applications focused on noise and air pollution data. Additional research was carried out in the project to examine the processes of opinion formation and their influence on behaviour changes.
- *Multi-Objective Decision Making Tools through Citizen Engagement* (Consensus, EU project, funding 2013 until 2015): Aim of the project is the development of tools which will support policy makers with high level results based on a number of relevant criteria. Therefore two real world use-cases were implemented: one dealing with Biofuels and Climate Change (EU Renewable Energy Directive), and the other dealing with the Trans-European Transport Networks (TEN-T).

3. Research questions and specific challenges

3.1 Research questions on geographic information extraction from user-generated spatial content

a) Extraction of the spatial, temporal and thematic reference

While a large number of user generated content contains references to location, only small percentage has explicit geo-coordinates. Thus geoparsing with toponym extraction from natural language and geocoding with assignment of lat/long coordinates is required. A major challenge is to resolve ambiguities, which can be caused for example through the usage of identical names for different places. For further use of the data, it is also necessary to identify temporal references. If no timestamp is available, this has to be inferred from contextual information. Besides explicit availability of spatial and temporal reference the automated selection, categorisation and interpretation of data content is required. To this end methods from computer linguistics and data mining can be utilised such as topic modelling, clustering, classification, rules, decision trees, correlation etc.

A lot of research has been carried out on event detection, thus next step might be a more in-depth analyses of event evolution over time and identification of correlated events. Another strength of user generated content is the implicitly contained information about the impact of events on people and their behaviour. Possible research questions are, which social groups react to an event - how and where? Amongst others, this requires the analysis and visualisation of the reactions of people in terms of judgements, responses, attitudes and emotions. Related research might deal with the development of methods to analyse the process of information spreading and opinion forming and to understand the temporal evolution of people´s relation and interaction with their physical contexts. All components should provide quality measures. The subjectivity and possible deliberate falsification of data requires special methods of evaluation and derivation of confidence measures. This will be the basis to ensure that from user-generated data reliable conclusions can be drawn and errors or misstatements be identified and minimised.

b) Fusion of data from various sources and temporal and spatial resolution

The heterogeneity and diversity of VGI data represents a unique source of information but also provides a lot of challenges. Methods of data fusion, as they exist for structured geodata are not applicable directly, instead approaches have to be developed which can integrate data of varying level of abstraction and interpretation. The heterogeneous spatial and temporal distribution requires the development of measures to quantify completeness and consistency, but

also approaches that perform the evaluation based on comparable situations. Furthermore methods of interpolation and extrapolation have to be developed or adapted, e.g. through utilisation of suitable geostatistical approaches.

VGI data are often spatially referenced by point locations such as images, microblogging content or POIs. In order to associate appropriate spatial and semantic context integration with topographic reference data can be beneficial. Thus, methods of data matching are needed, which are able to take constraints and semantics into account. Particularly challenging is the integrated treatment of data from different sources either user-generated or authoritative, but also from multiple scales, which might require matching at higher levels of aggregation.

c) Identification of correlations and pattern within large amount of data and data streams

News and feedback in social media often provide insight into opinions and political orientation, which could be extracted by methods from the field of sentiment analysis. This derived information can be correlated with reference data, additional thematic information (e.g. from news channels) as well sensor data (e.g. about noise, air or water quality) to identify potentially important factors that have impact on quality of life and development of regions. The combination and comparison of data from urban development, epidemics and political developments on one hand and information from social media on the hand might provide input for predictive models (e. g. through techniques of machine learning or time series analysis) and can support the creation of early warning systems.

Scope, quality and bias of data provided by the user in the context of social web and citizen science, are highly dependent on their intentions and predispositions. Therefore, a primary goal is to identify these intentions. Based on that the relationships between subjective intentions, opinions, moods etc. and factual information respective sensory data can be analysed.

Processing of large data streams requires customised data structures and algorithms. They have to guarantee efficient access and scalability. Furthermore they should support problem oriented aggregation hierarchies and visual inspection in top-down approaches. For this purpose methods can be extended from model and cartographic generalisation.

3.2 Research questions related to geovisualisation and cartographic communication

a) Development of innovative, adaptive visualisation metaphors

The special characteristics of user-generated spatial data requires further development of classic cartographic presentation methods, as well as underlying concepts and modelling approaches. Varying quality and heterogeneity of data, different spatial resolution and semantic structure or permanent temporal changes require the derivation of visualisation metaphors, which are able to reflect these characteristics and the meta-information associated with the user-generated data. An example of a new visualisation metaphor, which supports analyses and presentation of less structured data are georeferenced word clouds. There is a need to develop also methods for the visualisation of variability in addition to methods of uncertainty visualisation. Alternative presentation methods are required for the integration of different media types (text, image, video and audio) for example in the form of patchwork maps. Concrete research questions are: How can geovisualisation methods be adapted to the characteristics of user-generated data? How can subjectivity and varying quality in data visually represented?

b) Real-time visualisation, abstraction and interactive user interfaces

Since VGI is subject to high update rates, a major challenge relate to the development of real-time visualisation methods. Therefore dynamic, animated visualisation has to combine automation and interactivity. Cartographic presentation methods are required which are based on automated processing, but also support adaptation to task and user requirements. Direct visual feedback also encourages motivation in the generation of user-generated data. Thus the following research question has to be answered: how can automated (real-time) visualisation be combined with interactive visual analysis? How can animated representations visualise data streams, which are subject to permanent changes?

c) Empirical verification and further development of theoretical foundations

A challenge is to involve different user groups in the exchange of information through VGI as well as on the participation in decision making processes. Therefore empirical studies should examine the suitability of different

interactive visualisation methods in dependence of tasks and user groups – with the aim of extending theoretical foundations of cartography and geovisualisation. Research questions are for example: To what extent do classical requirements on legibility and minimal dimension remain? Do tiny objects (e.g. in web maps or perspective views) represent disturbing noise or a useful indicator for zoom and pan interaction? How can classical cartographic methodological knowledge be applied or extended for interactive, dynamic visualisation of user-generated data?

3.3 Research question on social context

The acquisition and also the usage of VGI data are carried out with different motivations and based on various applications – thus showing a different user context. In particular data generation is dependent on language and dialect of the person which generates the data. This results in a number of relevant research questions:

a) Quality and generalisability of information: subjective classification vs. general ontologies

One issue concerns the evaluation of the subjective information of individual users in combination with information from other users. Exemplary research questions are: which conclusions can be drawn from (larger or smaller) variability of the same situation? How can uncertain information be dealt with and cooperative control mechanism be instantiated?

Related to that it is largely unsolved, which constants are existing for the description of spatial conditions. For example, the estimation of absolute values strongly differs from one another while relative values are evaluated much more consistently and uniformly. Research subject is therefore the identification of other such constants and their usage for interpretation, evaluation and quality assessment of user data.

Further importance should be given to studies that identify the relevance of data: typically the relevance is determined through statistical methods, however this is not suitable for every application. Therefore it is necessary to develop new measures that work with small samples and heterogeneous data and can integrate social context.

b) Context dependency of data acquisition, abstraction and interaction

The granularity of VGI data is highly dependent on the context and the objectives that existed during capturing. Specific user-generated data often contain more useful information than those that have been acquired for general purposes. The task driven acquisition takes place at a certain level of abstraction, which is subjective, influenced by the target group and influenced by local knowledge and context (e. g. the annotation of comparable images with either "photo from Canada" vs. "photo from downtown Toronto"). Thus a question might be how context of can be identified and classified for the analysis of user intentions during data acquisition. Furthermore, it is of interest how can graphical user interfaces can be designed to encourage and support data collection beyond current usage purpose without unacceptable overhead. Therefore the complexity of the data acquisitions process should be kept away as far as possible from the user. Here the usage of appropriate ontologies can be foreseen which might generate more general usable data without the explicit intervention of the user.

c) Motivation and intention for participation, privacy and trustworthiness

If information should be produced specifically for certain applications, the motivation of the user might be subject of investigation as well as the mechanism to engage participation. Motivation may arise because of professional experience, a "green/social conscience", the aim on elimination of shortcomings, etc. In addition the suitability of incentives can be analysed e.g. as part of gamification and possible side effects on the generated data should be investigated as well.

Since the data generation is carried out directly by the individuals, a direct reference is given to the individual personality. This applies particularly to user position, which can be tracked over space and time. Therefore mechanisms are required to abstract and generalise privacy related data in order to ensure anonymity. Going even further, decentralised processing methods might be applied. Finally social groups can ensure that certain information is known only within the group and will be passed to the outside only in aggregated form.

4. Conclusions

In an information society, which is confronted with more and more data streams, large amounts of data and social networks, forms "seeing the basis for understanding". Thus the interplay of analytical processes with interactive visual (map) representations supports discussions and solution finding of space-time related questions related to environmental change, traffic congestion, population growth or migration all with local and global relevance. The utilisation of user-generated spatial data opens potentials at regional and local level by providing detailed, current, on-site information. Additionally, VGI can be seen as an expression of global networking were volunteers generate collaboratively up-to-date information e.g. to support humanitarian organisations with real-time crisis mapping and situational awareness support. As a vision it may be possible to work on societal challenges with focus on sustainability through interplay of authoritative information provider and community initiatives of volunteers.

Acknowledgements

This work was supported by the German Research Foundation as part of the priority programme "Volunteered Geographic Information: Interpretation, Visualisation and Social Computing" (VGIscience, SPP 1894).

References

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

DFG priority programme "Volunteered Geographic Information: Interpretation, Visualisation and Social Computing" (VGIscience) https://vgiscience.org/