On User-Generated Geocontent

Dieter Pfoser

Institute for the Management of Information Systems
Research Center Athena
G. Mpakou 17, 11524 Athens, Greece
pfoser@imis.athena-innovation.gr

Abstract. Spatiotemporal reasoning is a basic form of human cognition for problem solving. To utilize this potential in the steadily increasing number of mobile and Web applications, significant amounts of spatiotemporal data need to be available. This paper advocates user-generated content and crowdsourcing techniques as a means to create rich, both, in terms of quantity and quality, spatiotemporal datasets.

1 Problem Description

Effective discovery, integration, management and interaction with spatiotemporal knowledge is a major challenge in the face of the somewhat recent discovery of the spatial "dimension" of the World Wide Web. In many contexts, space in connection with time is used as the primary means to structure and access information simply because *spatiotemporal* (ST) reasoning is essential to everyday problem solving. In combination with the already staggering number of spatially-aware mobile devices, we are faced with a tremendous growth of user-generated content and demand for spatiotemporal knowledge in connection with novel applications and challenges.

With the proliferation of the Internet as the primary medium for data publishing and information exchange, we have seen an explosion in the amount of online content available on the Web. Thus, in addition to professionally-produced material being offered free on the Internet, the public has also been allowed, indeed encouraged, to make its content available online to everyone. The volumes of such User-Generated Content (UGC) are already staggering and constantly growing. Our goal has to be to tame this data explosion, which applied to the spatial domain translates to massively collecting and sharing knowledge to ultimately digitize the world. Our ambition has to be to go beyond considering traditional ST data sources and to include any type of available content such as narratives in existing Web pages in a ST data collection effort. We view all available content that has a ST dimension as a potential data source that can be used for computation. When utilized, this vast amount of data will lead to a digitized ST world view beyond mere collections of co-ordinates and maps.

One could argue that as early maps were traces of people's movements, i.e., view representations of people's experiences, digitizing the world in the this context relates to collecting pieces of knowledge gained by a human individual tied not only to space and time, but also to her context, personal cognition, and experience.

To realize this vision, (i) a proper understanding of ST language expressions will allow us to utilize non-traditional sources, e.g., narratives containing spatial objects and their relationships, (ii) appropriate representations of the collected data and related data management techniques will enable us to relate and integrate data, (iii) data fusion techniques will combine individual observations into an integrated ST dataset and (iv) adequate algorithms and queries will take (the continuously changing) uncertainty of the data into account. In addition, ST data is typically delivered and queried through map-based interfaces. With a thorough ST language understanding, (v) novel user interfaces that abstract and represent ST information in qualitative human terms might become a reality and provide for effortless and naive natural language interaction.

Overall, these ideas promote the GeoWeb 2.0 vision and advance the state of the art in collecting, storing, analyzing, processing, reconciling, and making large amounts of semantically rich user-generated geospatial information available.

2 Research Directions

In order for the proposed vision to become a reality, several research challenges spanning multiple disciplines have to be addressed.

User experience related to ST information is currently directly linked to the representation of the data; geographic co-ordinates pinpoint locations on maps, routing algorithms determine the best route based on distance, etc. In contrast, human interaction with the world is based on experience, learning and reasoning upon loosely coupled, qualitative entities, e.g., spatial relationships such as "near/far". The challenge will be on devising means to better understand people's perception of space by *deciphering how people express* ST *concepts in natural language terms* by means of a Rosetta-stone-equivalent tool for deciphering the ST component of (a range of) natural languages and, if possible, define the underlying building blocks, i.e., cognitive concepts inherent in ST reasoning. To improve *natural language processing*, a key aspect will be to engage the user in providing her conceptions of space in natural language terms. For example, *games-with-a-purpose* (GWAP) [10] can be used as a vehicle to record and analyze natural language descriptions of known ST scenarios, i.e., to provide spatial descriptions of scenes by means of text or audio during the course of the game. GWAP will be the motivator for crowdsourcing of spatial scene descriptions for a multitude of languages and to finally produce a ST language corpus. This corpus can then be used to extract ST knowledge from non-traditional content sources such as narratives in travel blogs.

Data capture focusses then on amassing user-generated ST data from various sources. Existing *attentional information* can be exploited by data mining user-generated ST content, e.g., point cloud data such geocoded flickr images [2] and by extracting ST data from text/audio narratives using NLP techniques, e.g., translating the phrase "the hospital is next to the church" to two spatial objects and respective relationship. In addition, *specifically designed tools* can support the user in the creation of geospatial content (cf. “Geoblogging” [7]). A means for the collection of *un-attentional data* is *ubiquitous positioning*, i.e., using a large number of complementary positioning solutions (GPS, WiFi, RFIDs, indoor positioning) to relate content to absolute spatial (coordinates) and temporal (time stamps) values. Data capture will not only produce quantitative data, i.e., spatial objects and their locations, but also qualitative data in terms of spatial relationships.

Efficient *data management* techniques will be of outmost importance when tapping into large amounts of geospatial data streams. The focus in this research will be on distributed data management schemes such as cloud computing. Issues to be addressed are spatial indexing and query processing (cf. geospatial data management using Apache Cassandra [5]). In addition, one needs to investigate novel concepts such as dataspaces [4] and linked data [1] for the specific case of geospatial information. *Mobile devices* are increasingly used as Web infrastructure nodes and, hence, will play an important role not only in data collection but also in distributed geospatial data management.

Data Fusion presents us with the problem of processing diverse incoming information and to specify the *relationships and correspondences between data and/or metadata* of different sources so as to reconcile them. Such a framework for matching and mapping different data streams of user-generated content involves identifying related data and generating better mappings by developing specific tools that involve the user in the process. The final goal is then to reconcile and fuse user-generated data to arrive at a single *ST* dataset. *Spatial uncertainty* has been described as “the Achilles’ Heel of GIS, the dark secret that once exposed will bring down the entire house of cards” [3]. Uncertainty will be essential to the process of correlating and fusing previously unrelated *ST* data sources. In part, data fusion can be seen similar to the problem of adjustment computation in surveying engineering, in which for a number of observations the best fitting (mathematical) model and its parameters are determined, i.e., to derive the true values based on observations. Research needs to focus on relating and mapping qualitative data (relative *ST* data as denoted by topological, metric and directional spatial relationships) to uncertainty. To achieve this task, techniques based on graph similarity, spatial reasoning [11] and Bayesian statistics [8] need to be investigated.

Working with user-generated spatiotemporal data sources has the drawback that there are no “final” datasets, i.e., all datasets are affected by a varying degree of uncertainty, which any kind of *computation* has to take into consideration. An additional aspect is the evolving nature of the data. Given that each observation increases the scope and/or reduces the uncertainty, algorithms have to accommodate this fact (cf. Canadian Traveller Problem [6]).

Current *spatial information visualization and interaction* is typically map-based. Even in novel geographic fusion services over the Web, e.g., Google Maps, the dependence on map-based interfaces for querying and delivering information is dominant and quite often lacking in expressiveness and usability. While recently systems based on augmented reality concepts have been developed, improved ST language understanding may lead to alternative text and audio-based interfaces to consume ST information (cf. [9]).

3 Summary

The presented vision targets user-generated geocontent and turning it into a viable data source that will complement ST knowledge created by expert users. New means to harness, aggregate, fuse and mine massive amounts of ST data are needed in order to achieve best possible coverage and extract hidden knowledge. The task at hand is nothing less than taming this semantically rich user-generated geodata tsunami and addressing the challenge of transforming the data into meaningful chunks of information obtained with simplicity and speed comparable to that of Web-based search.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme - Marie Curie Actions, Initial Training Network GEOCROWD (http://www.geocrowd.eu) under grant agreement No. FP7-PEOPLE-2010-ITN-264994. The author also thanks Spiros Athanasiou for many productive discussions on the various subjects.

References

1. C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int'l J. on Semantic Web and Information Systems*, 5(3):1–22, 2009.
2. A. Cope. The Shape of Alpha. Where 2.0 conf. presentation, http://where2conf.com/where2009/public/schedule/detail/7212, 2009.
3. M. Goodchild. Uncertainty: the Achilles Heel of GIS? *Geo Info Systems*, 8(11):50–52, 1998.
4. A. Halevy, M. Franklin, and D. Maier. Principles of Dataspace Systems. *Proc. 25th PODS conf.*, pages 1–9, 2006.
5. M. Malone. Building a Scalable Geospatial Database on top of Apache Cassandra. http://www.youtube.com/watch?v=7J61pPG9j90, 2010.
6. C. Papadimitriou and M. Yannakakis. Shortest paths without a map. *Journal Theoretical Computer Science*, 84(1):127–150, 1989.
7. D. Pfoser, C. Lontou, E. Drymonas, and S. Georgiou. Geoblogging: User-contributed geospatial data collection and fusion. In *Proc. 18th ACM GIS conf.*, pages 532–533, 2010.
8. H. Richardson and L. Stone. Operations Analysis during the underwater search for Scorpion. *Naval Research Logistics Quarterly*, 18(2):141–157, 1971.
9. S. Strachan, J. Williamson, and R. Murray-Smith. Show me the way to Monte Carlo: density-based trajectory navigation. In *Proc. CHI'07 conf.*, pages 1245–1248, 2007.
10. L. von Ahn. Games with a purpose. *IEEE Computer*, 39(6):92–94, 2006.
11. J. Wallgruen, D. Wolter, and K.-F. Richter. Qualitative matching of spatial information. In *Proc. 18th ACM GIS conf.*, pages 300–309, 2010.