Collaborative Spatial Data Acquisition – A Spatial and Semantic Data Aggregation Approach

Sebastian Matyas[1]
[1]Laboratory for Semantic Information Technology,
Otto-Friedrich-University Bamberg, Germany,
sebastian.matyas@wiai.uni-bamberg.de

ABSTRACT

The idea of building a spatial data set by integrating data chunks collected by a large group of voluntary contributors is gaining momentum. Well-known examples are the trail libraries that mountain bikers and other sports enthusiasts have created recently. With Web 2.0 technologies supporting the self-organization of large user communities, the recruitment of contributors seems the least problem of the approach. Much more serious is the issue of data aggregation which can threaten the feasibility of collaborative acquisition of spatial data. Solutions have been proposed to the purely spatial part of the problem (e.g. the aggregation of GPS-tracks) while the semantic part, that is, the integration of conflicting thematic data, has received much less attention so far. We present an approach that uses Semantic Web technologies – formal ontologies – to describe spatial and semantic aggregation methods. More specifically, we present a rule-based language that allows us to express spatial and semantic preconditions for the application of aggregation methods. With a worked example we illustrate which type of problems our approach is able to resolve.

1. INTRODUCTION

With the elevated price of spatial data sets and the unavailability of data meeting the needs of specific user groups it is not surprising that the idea of collecting spatial data by voluntary contributors emerged. Mountain bikers where among the first communities to feel unhappy with the commercially available spatial data products: their cartography focuses on roads and much of the thematic information is useful just to car drivers. As a consequence, several biker communities started to collect GPS track data and published their data through web portals such as the GPS-tour.info portal[1]. Bikers download tours that others have uploaded and use them for planning and navigating their own tour. Another example is the PARAMOUNT project (Loehnert et al., 2001) which provides a location-based service for hikers, where the users in addition to GPS data also can post their touring experiences.

Many communities do not aggregate the spatial data they collect which is why we do not consider these projects to be examples for a *collaborative* acquisition of spatial data. In contrast, data aggregation constitutes a central concern for the digital trail library described by Morris et al. (2004). A spatial data set of biking trails is generated from many individual GPS tracks collected by bikers in the Frank Church River of No Return wilderness area. Aggregation is dealt with at the level of geometrical information using algorithms that compute trails by averaging over many GPS tracks.

Voluntary contributors with handheld GPS devices typically collect more than just track data. Many of them also record some form of thematic data, for instance, points of interest (POI). The aggregation of semantic data such as the feature type of a POI constitutes a challenge in its own. Almost none of the portals for shared GPS data define a vocabulary for POI feature types. But even if such vocabularies were provided, it is unlikely that contributors not having had some training in (geo)information science will appreciate the importance of using the vocabulary. However, by using both sources of data, spatial and thematic, at least some plausibility tests could help to identify suspect data sets and help to decide which data aggregation method to apply.

[1] http://www.gps-tour.info/

We propose to use methods from Semantic Web research, namely formal ontologies, to build a descriptive formal framework for data aggregation. The goal is to have a specification language that permits to describe the spatial and thematic conditions under which different aggregation method deliver plausible results. The remainder of the paper is structured as follows. Section 2 gives a brief overview of approaches for the collaborative acquisition of spatial data by non-experts. Section 3 analyzes the data aggregation problem inherent to data acquisition by communities. We then present our descriptive framework in section 4. The paper closes with a discussion and an outlook on future research in section 5.

2. COLLABORATIVE DATA ACQUISITION

It is fair to say that collaborative data acquisition has received more attention outside the Geographic Information Science community than within it. In 2005, researchers from Artificial Intelligence started to discuss the issue of semantic data integration at the AAAI Spring Symposium on "Knowledge Collection from Volunteer Contributors" (KCVC05). Several of the solutions can be transferred to spatial data acquisition. An approach relevant for the aggregation of POI feature types was described by Ahn and Dabbisch (2005). They proposed a web-based game named ESP which would make it possible to label all images indexed by Google in only 31 days, given an appropriate popularity and usage of the game. Interestingly, this approach does not make use of a predefined vocabulary of tags. Instead, the vocabulary is constructed implicitly by aggregating any agreements of two users about a single tag.

To some extent, a spatial version of the idea to use a game to collect chunks of semantic information has already been realized in the context of pervasive gaming. Pervasive games are played in the geographic environment ("the real world") and the location of a player is used as a main game element. Examples of such games haven been described by Bell et al. (2006), Capra et al., (2005) and Peltola et al. (2006). Beside being a new way to play outdoors, these pervasive games are also used to collect spatial data like WLAN hotspots strength or GPS coverage during the course of the game. In principle, any game that uses a localization technology could be used for that purpose although some games are more adapted for collection POI feature types than others.

The class of Geogames, for instance, with games such as GeoTicTacToe (Schlieder et al., 2006) and CityPoker (Schlieder, 2005), comes with a specific spatio-temporal synchronization mechanism of the game flow, that causes the players repeatedly to stop for some time at specific locations. This characteristic has been used to integrate feature type classification tasks into the game. In the context of a cultural heritage game, for instance, players had to identify the function and style of historical buildings (e.g. gothic cathedral, neo-renaissance town hall) in order to activate a relevant game action.

Since the players of a Geogame are not experts in spatial data acquisition, the quality of the data they collect is rather heterogeneous. Errors can either be of spatial nature, due to the inaccuracy of the localization technology or semantic, due to mistakes of the players in the categorization task.

3. APPROACHES FOR SPATIAL AND SEMANTIC DATA AGGREGATION

The aggregation problem for collaboratively collected spatial data has two facets: the spatial and the semantic aggregation problem. With POI data, the spatial problem amounts to aggregate several measurements of the geographic position of POI. Spatial averaging based on reasonable assumptions about the positional error distribution will solve this problem. In a similar way, positional data about higher-dimensional features is approached. Morris et al. (2004) or Sayda (2005), for example, we suggest methods for the aggregation of 2D line features. However, the issue of which measurements actually refer to the same feature is more complex as it involves feature type semantics.

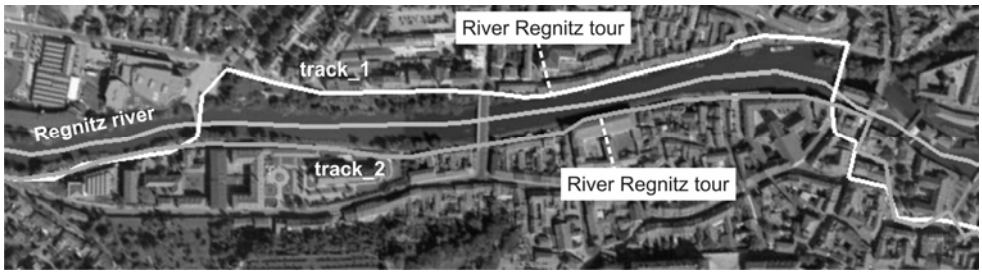

Figure 1: Spatial data gathered and categorized (as "River Regnitz tour") by two biking tourists © aerial photo: Google Earth (http://earth.google.com/).

Consider the following example: In Figure 1 two GPS tracks are shown, which represent bicycle tours recorded by two different tourists. One was taking the route along the left bank of the Regnitz river (track_1), the other the route on the right bank (track_2). Both tourists submit their data to a collaborative location-based service. We assume that the spatial data describing the course of the Regnitz river is already part of the data set. It could have been computed, for instance, from tracks entered by canoe tourists. A simple spatial aggregation algorithm such as the one proposed by Sayda (2005) would just interpolate between the two biking tracks by constructing a line of points with equal distance to each of the two tracks. However, the resulting track would end up in the middle of the river, thereby producing a semantic conflict. A semantic conflict situation is described in form of a rule: the result of spatial aggregation of biking trails may not lie within a water body.

But typically, the data collected for a POI or a higher higher-dimensional spatial feature associates positional information (e.g. lat/long) with information about the feature type (e.g. restaurant). When many contributors help to build the data set, it is almost impossible to avoid cases where a real-world object is categorized differently by different people. A "restaurant" could be categorized as "bistro" by one user and as "tavern" by another. To resolve such semantic conflicts similarity between the feature types has to be measured. Several methods have been proposed to compute similarity values. Examples are described, for instance, by Rodríguez and Egenhofer (2004) or Jones et al. (2001). A typical rule for semantic conflict resolution would state that positional information can be aggregated spatially provided that the difference of the associated semantic information does exceed a given threshold. The same is true when we adopt these methods for the aggregation task of the spatial data in Figure 1. Either of them would aggregate the two tracks to a new one with regard to their semantic similarity, as both are categorized as "Regnitz river tour" by the two biking tourists. But although they are semantically equal, they do not represent the same real world object in Figure 1 and consequently should not be aggregated to one data object.

A comprehensive solution of the data aggregation problem needs a combined approach which takes both, the spatial as well as the semantic data into account. The approaches have been described, for instance, by Uitermark et al. (1999) or Gösseln and Sester (2003 and 2004). However, these approaches use semantic similarity values only as preconditions for a following geometric aggregation. Furthermore they are not able to handle cases where there is a comparatively small variance in the spatial data (track_1 and track_2 in Figure 1) but a huge distance in semantics (land vs. water).

4. A RULE-BASED SPATIAL DATA AGGREGATION LANGUAGE

We propose a rule-based design language to help the designer of a location-based service to formulate preconditions for individual aggregation methods. The language permits to describe the two-step aggregation process for spatial data: (1) the spatial data is checked in order to determine

whether or not a particular aggregation rule is applicable, that is, whether or not its preconditions are satisfied. (2) In case of several applicable aggregation rules, an arbitration mechanism makes a choice – for instance for the most specific rule and applies the rule to the data.

We use a simple relational language that augments spatial SQL relations (see e.g Egenhofer, 1994) by constructs from formal ontologies which permit to specify type information for the relation's arguments. A *relational statement* makes an assertion about a spatial or semantic relation that holds between geographic objects of a certain type. Formally, a relational statement is specified by a spatial or semantic relation symbol and a sequence of typed arguments. The language permits the modeling of both distance relations (e.g. *small_distance*) and directional relations encoding ordering information (e.g. *between*) besides the traditional spatial SQL relations (e.g. *intersects*). Geometric and semantic relations are defined qualitatively rather than with hard metric units. A negation of rules is possible too (e.g. when not … *between*…).

This way a combination of preconditions can control the application of individual aggregation methods for specific spatial data sets. For the example data set in Figure 1 we assume a location-base service designer has defined the following preconditions for his aggregation method (Algorithm of Morris et al. 2004) of choice:

Rules (Preconditions):

when	track_1:BikingTrail	when not	track_1:BikingTrail
	intersects		*between*
	trail_2:BikingTrail		Regnitz_river: WaterBody
when	track_1:BikingTrail		track_2:BikingTrail
	small_distance	when	Regnitz_river_tour: InformationClass
	track_2:BikingTrail		*sufficient_similar*
			Regnitz_river_tour: InformationClass

Aggregation Method:
Algorithm of Morris et al. (2004)

The typed arguments of the precondition statements are defined in a given domain ontology which can be further used to derive T-Box statements from it. In addition from the specified rule set A-Box statements can be defined. In the following the T- and A-Box for the above example rule set is presented (syntax is taken from Baader et al. 2003):

T-Box:	**A-Box:**
BikingTrail $\subseteq$ Concept	BikingTrail(track_1)
WaterBody $\subseteq$ Concept	BikingTrail(track_2)
InformationClass $\subseteq$ Concept	WaterBody(Regnitz_river)
	InformationClass (Regnitz_river_tour)
	InformationClass (Regnitz_river_tour)

The benefit of using a DL as the language for the preconditions is that it enables the service designer to automatically compute an A- and T-Box consistency check with various DL- or ontology tools available. If the rules are consistent according to the A- and T-Box the preconditions can then be applied to available spatial data provided by the users. If the data is checked positively against all defined rules then the specified aggregation method is to be applied. In the case of the example for Figure 1 both newly added data sets would not be aggregated because of the third rule (…*between*…) in the precondition rule set, which is the desired behavior of the location-based service.

5. DISCUSSION AND FUTURE WORK

Although literature about location-based services highlights several interesting fields of research, for example privacy (Gruteser and Grunwald, 2003) or architecture (José et al., 2003) issues, the idea to use the location-based service users, i.e. non-experts for the acquisition of spatial data is getting more widely discussed, see e.g. Sayda (2005). We presented the spatial and semantic aggregation problem on a worked example which occurs in this context. A rule-based design language was introduced, which is used to solve this kind of problems. The language allows the definition of preconditions which enable the control of the usage of aggregation methods for individual spatial data sets.

While the usage of known aggregation methods is a good starting point, a statement set like the one for the preconditions would enable the location-based service designer a more precise control of the aggregation process. These statement sets, respectively the resulting data could then be further mapped to specific quality values in quality dimensions found for example in Navratil (2005), Pipino et al. (2002), or the ISO 19113 norm. The definition of an appropriate statement set is therefore subject of further research.

To evaluate our proposed approach with real world data we are currently implementing an appropriate location-based service. Additionally, we are continuing to specify needed statements for our rule set until it can be shown to be complete.

On the basis of the idea from Von Ahn and Dabbisch (2005), we are further investigating how to use the Geogames framework (see Schlieder et al. 2006) to develop convenient location-based games for collecting spatial data in an entertainment context.

BIBLIOGRAPHY

Baader, F., Calvanese, D., McGuinness, D., Nardi, D. and Patel-Schneider, P., 2003. The Description Logic Handbook – Theory, Implementation and Applications, Cambridge University Press, ISBN 0-521-78176-0

Bell, M., Chalmers, M., Barkhuus, L., Hall, M., Sherwood, S., Tennent, P., Brown, B., Rowland, D. and Benford, S., 2006. *Interweaving Mobile Games With Everyday Life*, CHI 2006, ACM Press, pp. 417-426

Capra, M., Radenkovic, M., Benford, S., Oppermann, L., Drozd, A., and Flintham, M., 2005. *The multimedia challenges raised by pervasive games*. .Proc. of the 13th Annual ACM international Conference on Multimedia, MULTIMEDIA '05. ACM Press, pp. 89-95.

Christopher, B. J., Alani, H. and Tudhope, D., 2001. *Geographical information retrieval with ontologies of place*. LNCS 2205, Spatial Information Theory Foundations of Geographic Information Science, D. Montello (ed), COSIT 2001

Egenhofer, J. M., 1994. Spatial SQL: A Query and Presentation Language, *IEEE Transaction and Data Engineering* 6 (1), 1994, pp. 86-95

Gösseln, G. and Sester, M, 2003. *Semantic and Geometric Integration of Geoscientific Data Sets with ATKIS – Applied to Geo-objects from Geology and Soil Science*, ISPRS Commission IV Joint Workshop "Challenges in Spatial Analysis, Integration and Visualization II", pp. 111-116.

Gösseln, G. v. and Sester, M., 2004. *Integration of geoscientific data sets and the german digital map using a matching approach*, International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences, Vol. 35, ISPRS

Gruteser, M. and Grunwald, D. 2003. *Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking*. In Proceedings of the 1st international Conference on Mobile

Systems, Applications and Services (San Francisco, California, May 05 - 08, 2003). MobiSys '03. ACM Press, New York, NY, 31-42.

José, R., Moreira, A., Rodrigues, H., and Davies, N. 2003. *The AROUND architecture for dynamic location-based services.* Mob. Netw. Appl. 8, 4 (Aug. 2003), 377-387.

Loehnert E., Wittmann E., Pielmeier J., Sayda F., 2001: PARAMOUNT- Public Safety & Commercial Info-Mobility Applications & Services in the Mountains, 14th International Technical Meeting of the Satellite Division of The Institute of Navigation, ION GPS 2001, September 11 - 14, Salt Lake City, Utah, USA

Morris, S., Morris, A., and Barnard, K. 2004. *Digital trail libraries*. .Proc. of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '04. ACM Press, pp. 63-71.

Navratil, G., 2005. Influences of Technology, Law, and Usability on Data Quality, GeoInfo Series Vienna (Ed. Andrew U. Frank), Institute for Geoinformation and Cartography, ISBN 3-901716-32-7

Peltola, J. and Karsten, H., 2006. *When play is not enough: Towards actually useful applications for digital entertainment*, Helsinki Mobility Roundtable 2006, Helsinki School of Economics

Pipino, L. L., Lee, Y. W. and Wang, R. Y., 2002. Data Quality Assessment, *Communications of the ACM*, April 2002/Vol. 45, No. 4

Rodríguez, A. and Egenhofer, M., 2004. Comparing Spatial Entity Classes: An Asymmetric and Context-Dependent Similarity Measure, *International Journal of Geographical Information Science* 18 (3), pp. 229-256

Schlieder, C., Kiefer, P., Matyas S., 2006. Geogames - Designing Location-based Games from Classic Board Games, *IEEE Intelligent Systems, Special Issue on Intelligent Technologies for Interactive Entertainment*, Sept/Okt 2006, pp 40-46

Schlieder, C., 2005: Representing the Meaning of Spatial Behavior by Spatially Grounded Intentional Systems, In: M.A. Rodríguez et al. (Eds.): Geospatial Semantics, LNCS 3799, pp. 30 – 44, Berlin: Springer.

Sayda, F. (2005): Involving LBS users in data acquisition and update. In: AGILE 2005: Conference on geographic information science, ed. by F. Toppen and M. Painho. Lisboa: AGILE, Universidade Nova de Lisboa, 2005.

Uitermark, H. T., van Oostertom, P. J. M,. Mars, N. J. I. and Molenaar, M., 1999. *Ontology-Based Geographic Data Set Integration*, Proc. of the International Workshop on Spatio-Temporal Database Management, STDBM'99, LNCS 1678. Springer, Berlin, pp. 60-78.

Von Ahn, L. and Dabbish, L., 2005. *ESP: Labeling Images with a Computer game*, AAAI 2005 Spring Symposium Knowledge Collection from Volunteer Contributors (KCVC05)