Crowdsourcing Geographic Information

Dieter Pfoser
Department of Geography and Geoinformation Science
George Mason University
Fairfax, VA, USA
dpfoser@gmu.edu

Synonyms

volunteered geographic information (VGI), user-generated geospatial content

Definition

The crowdsourcing of geographic information addresses the collection of geospatial data contributed by non-expert users and the aggregation of these data into meaningful geospatial datasets. While crowdsourcing generally implies a coordinated bottom-up grass-roots effort to contribute information, in the context of geospatial data the term volunteered geographic information (VGI) specifically refers to a dedicated collection effort inviting non-expert users to contribute. A prominent example here is the OpenStreetMap effort focusing on map datasets. Crowdsourcing geospatial data is an evolving research area that covers efforts ranging from mining GPS tracking data to using social media content to profile population dynamics.

Historical Background

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 content being offered free on the Internet, the public has also been encouraged to make content available online to everyone. The volumes of such User-Generated Content (UGC) are already staggering and constantly growing. With geospatial content playing an essential UGC role [13], research in this area takes advantage of this explosion in Volunteered Geographic Information (VGI) [6] to produce datasets that complement authoritative datasets rather than replace them. A term frequently mentioned when discussing user-generated content is *crowdsourcing*, which implies a coordinated, bottom-up grass-roots effort to contribute information. VGI refers specifically to geographic content contributed by non-expert users, and does not require some level of coordination among the individuals who are making these contributions [6]. User-generated geospatial content has been categorized in several ways. Web-based services and tools can provide means for users through *attentional* (e.g., geo-wikis, geocoding photos) or *unattentional*, or passive efforts (e.g., GPS traces from their daily commutes) to create vast amounts of data concerning the real world that contain significant amounts of information (crowdsourcing). Another way of looking at user-generated geospatial content is by way of differentiating between explicit and implicit content (cf. [5]). *Explicit content* is generated purposefully in the desired form. An example here is the OpenStreetMap[1] road network. In contrast, *implicit content* reflects derived information as the original, user-generated content was produced with

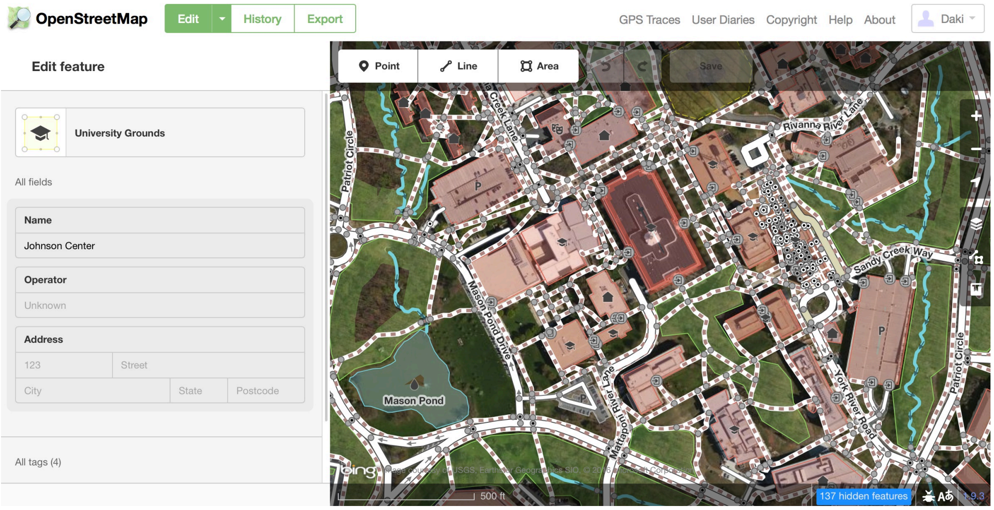
[1] http://www.openstreetmap.org

a different purpose in mind. However, the desired content may be extracted from it nevertheless. An example here would be extracting a road network from GPS tracking data. Implicit content is also often embedded in social media contributions (e.g., blogs, micro-blogs, social multimedia). Here it is referred to as *ambient* (AGI – [15]) geographical information. Overall, while initially VGI research focused on explicit datasets and dedicated applications, the availability of implicit data by means of the smart phone revolution has lead to a surge in data mining research focusing on ambient user-generated (geo)content. It is interesting to see that while Volunteered Geographic Information (VGI) can be large in volume, it is comparatively poor in quality. The full exploitation of crowdsourced information as a means to complement authoritative datasets is contingent on the thorough understanding of its accuracy. This fact has lead to a recent surge in research addressing the quality aspect of VGI.

Scientific Fundamentals

Crowdsourcing geospatial data is a broad research area that spans many different efforts. In the discussion that follows, we will possibly not cover the topic in its entirety, but try to discuss representative contributions. The most well-known crowdsourcing examples of geospatial data include platforms like OSM, Wikimapia[2] and Google MapMaker[3], which allow non-experts to perform basic cartographic tasks, digitizing and editing road networks, building outlines, and Point-Of-Interest (POI) information. OpenStreetMap (OSM) is a collaborative project to create a free editable *map of the world*. It was inspired by the success of Wikipedia and the lack of availability of affordable map data. The project has grown to over 1.6 million registered users, who can collect data using GPS devices, aerial photography, and other free sources. This crowdsourced data is then made available under the Open Database License. Figure 1 shows how simple it is to contribute edits to OSM using the browser-based iD editor. The figure shows a map excerpt of George Mason University with features such as buildings, roads, trees and amenities clearly visible. Wikimapia is also a collaborative mapping project that aims to mark and describe all *geographical objects* in the world. It combines an interactive web map with a geographically-referenced wiki system. As of June 2014, over 23,000,000 objects have been marked by registered users and guests. Google Mapmaker is a commercial effort comparable to OSM. Besides map data, approaches to the collection of geospatial objects include the use of Google (now Trimble) SketchUp[4] and the related 3D Warehouse[5], which creates a geographically tagged database of three-dimensional objects.

[2] http://wikimapia.org
[3] http://www.google.com/mapmaker
[4] http://www.sketchup.com
[5] https://3dwarehouse.sketchup.com



Edit feature

University Grounds

All fields

Name
Johnson Center

Operator
Unknown

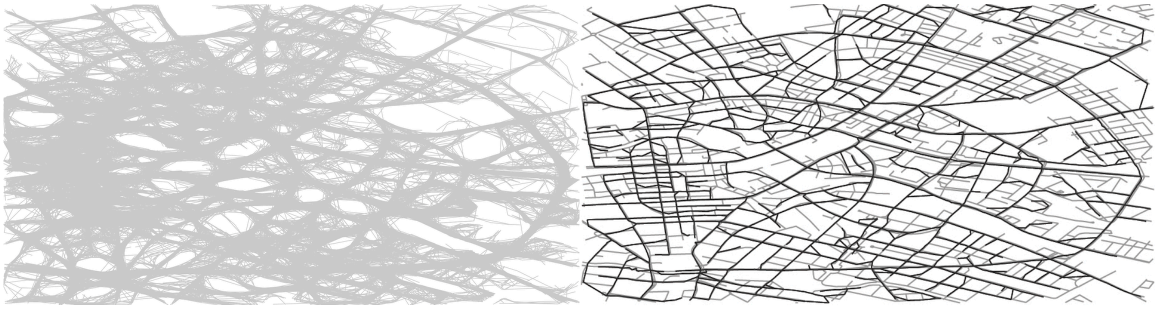
Address
123 | Street
City | State | Postcode

All tags (4)

Figure 1: Openstreetmap – browser-based iD editor

Maps and Traffic Data. Crowdsourcing specifically plays an important role for data collection with respect to traffic information and navigation services. Street maps and transportation networks are of fundamental importance in a wealth of applications. In the past, map production was costly and proprietary data vendors such as NAVTEQ (now Nokia), and TeleAtlas (now TomTom) dominated the market. Over the last years, VGI efforts such as OSM have complemented commercial map datasets. However, these efforts still require dedicated users to author maps using specialized software tools. Lately, on the other hand, the commoditization of GPS technology and integration in mobile phones coupled with the advent of low-cost fleet management and positioning software has triggered the generation of vast amounts of tracking data, an example of *un-attentional crowdsourced geospatial map data*. Such tracking information can be collected explicitly through location-based social networks such as Waze[6] or implicitly from GPS enabled devices. In conjunction with the proliferation of GPS technology as an end-user consumer product, GPS devices are being integrated in large-scale enterprise settings for fleet management and asset tracking. One example of this is the emergence of floating car data (FCD) repositories, which refer to using data generated by one vehicle as a sample to assess overall traffic conditions (e.g., “a cork swimming in the river”). Having large amounts of vehicles collecting such data for a given spatial area such as a city (e.g. taxis, public transport, utility vehicles, delivery fleets) can render an accurate picture of the traffic condition in time and space (e.g., [12]). By *map-matching* the tracking data it is possible to derive travel time related to specific portions of the road network, lending itself to a live assessment of the traffic conditions and future trends. Besides the use of such data in traffic assessment and forecasting, there has been a recent surge of actual *map construction algorithms* that derive not only travel time attributes, but also actual road network geometries from tracking data such as shown in Figure 2 (cf. [3]).

[6] https://www.waze.com

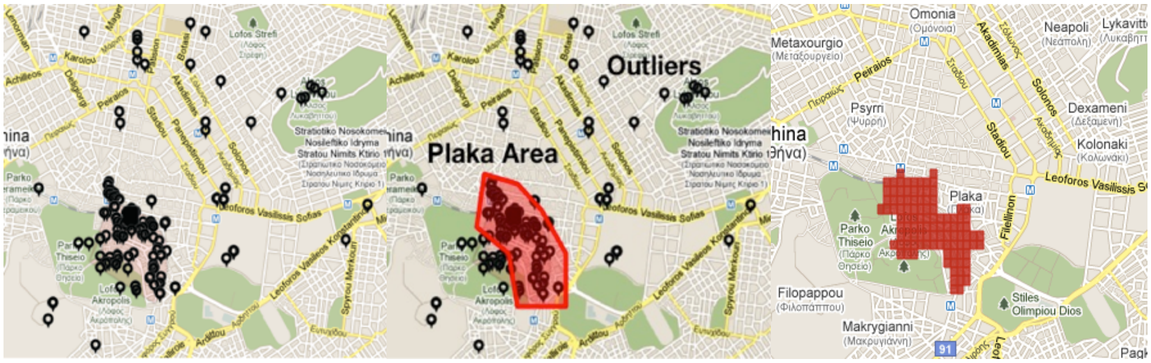


(a) GPS trajectories (b) constructed road network

Figure 2: Map construction example - GPS trajectories of vehicles collected in Berlin, Germany

Social Media. The Web has created a number of services that facilitate the collection of *location-relevant content*, i.e., data for which location is just but one, and most often even not the most important attribute. Prominent examples include (i) photo sharing sites such as Flickr, Panoramio, and Instagram, (ii) social media and specifically microblogging sites with geotagging features, e.g., Facebook, Google+, Twitter as well as related photo sharing sites (twitpic), and (iii) geospatial check-in services such as Foursquare. S*ocial networking and microblogging services* such as Twitter provide a continuous source of data from which useful information can be extracted. Aggregating geo-related blog posts allows one to derive *topic information in relation to space*. They describe for example social activities occurring within a city. Such topics, as captured by social media, are highly dynamic at all scales, as spaces can change with the latest trends, news headlines, or social movements [15]. In [10] a probabilistic topic model is used to decompose the stream of digital traces contained in textual and event-based data from services such as Twitter and Foursquare into a set of urban topics related to various activities of people during the course of the week. Due to the combined use of implicit textual and movement data, one can obtain semantically rich modalities of the urban dynamics and overcomes the drawbacks of traditional methods such as on-site surveys. The results of this work can be used to enrich location- based services with real-time context.

Point Clouds. Users generate data by means of many different applications in which geospatial data is simply used to index and access the collected data. This data may not always be in the format, quality and amount we expect it, but by employing intelligent data collection and *mining algorithms*, we are able to discover valuable insights into *urban function*. As mentioned, prominent dataset examples here are photo-sharing sites and micro blogging services using geotagging features. To illustrate the potential for geospatial data generation, consider the use of Flickr data for the computation of feature shapes of various spatial objects including city scale, countries and other colloquial areas [4]. The approach is based on computing primary shapes for point clouds that are grouped together by Yahoo! GeoPlanet WOEIDs (Yahoo! Where On Earth IDs), which are part of the Flickr metadata. Flickr data is used to identify *places* and also *relationships* among them (e.g., containment) based on tag analysis and spatial clustering in [9]. A system that combines browser-based computing with crowdsourcing [11] derives colloquial geospatial objects, or, POIs from point-cloud data such as Flickr image locations. The user provides the search terms (e.g., the name of the tourist area "Plaka, Athens") and respective point databases are queried based on their tag information and using web service APIs to retrieve a point cloud that characterizes this location. This point cloud represents the collective notion – or the colloquial footprint of the area of interest, which then needs to be aggregated to derive the actual location of the sought geospatial object. The specific approach uses a hierarchical grid-based filter-and-refinement approach to retrieve a minimal dataset from the Web data sources and to still produce adequate spatial object geometries.



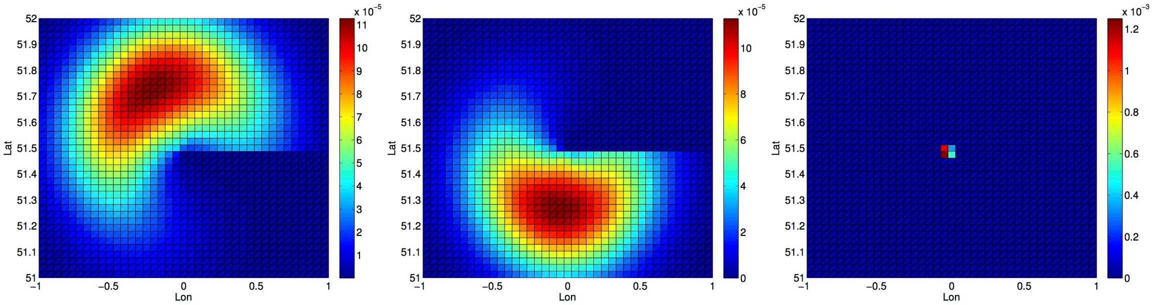
(a) Flickr image locations (b) Plaka common knowledge extent (c) clustering result

Figure 3: Crowdsourcing colloquial geospatial objects from user-contributed content.

Qualitative Geospatial Data Sources. In the geospatial domain, authoring content typically involves quantitative, coordinate-based data. While technology has helped a lot to facilitate geospatial data collection, e.g., all smart phones are equipped with GPS positioning sensors, yet authoring quantitative data requires specialized applications (often part of social media platforms) and/or specialized knowledge, e.g., OSM. This fact hinders the widespread adoption of VGI as an even bigger, large-scale geospatial data source. The broad mass of users contributing content on the Internet are much more comfortable with using *qualitative information*. People do not use coordinates to describe their spatial experiences (trips, etc.), but rely on qualitative concepts in the form of toponyms (landmarks) and spatial relationships (near, next, etc.). With spatial reasoning being a basic form of human cognition, narratives expressing such geospatial experiences, e.g., travels blogs, would provide an even bigger source of geospatial data. Typically, spatial information extraction from texts is associated with georeferencing of texts, which typically involves the identification and geocoding of toponyms. Several commercial software packages and services, e.g., Google Places API[7] and Yahoo! BOSS Geo Services [8] do exist. Using this approach, travel blogs (e.g., travelpod.com, travelblog.org) have been mined to give researchers a person's conceptualization of place based on geo-referenced text to construct regions of thematic saliency [1]. However, the *extraction of qualitative spatial data* in the form of relationships from texts requires the

[7] https://developers.google.com/places/
[8] https://developer.yahoo.com/boss/geo/

utilization of efficient natural language processing (NLP) tools to automatically extract and map phrases to spatial relations. This has been addressed to some extent in the literature, e.g., [16], but efforts have always been limited by the unclear mapping of spatial language expressions to spatial relationships such as metric, directional and topological relations. In addition, while computing using a qualitative representation of spatial data has been an active research topic for quite some time, the quantitative, i.e., coordinate-based representation is still dominant. Hence, a research topic has been the quantitative representation of qualitative relationship data. The authors of [14] introduce a basic Expectation Maximization/Gaussian Mixture Model approach to quantify qualitative spatial data extracted from crowdsourced travel blog narrative. The method automatically extracts qualitative spatial data from texts, quantifies the relations using probability distribution functions, and introduces a location estimation method based on spatial relation fusion. This approach bridges the gap between qualitative and quantitative representation of spatial relations using efficient machine learning techniques and introduces an actual text-to-map application. Figure 4 gives three quantified relationships examples crowdsourced from textual narrative.



(a) North (b) South (c) Near

Figure 4: Probabilistic heat maps for three basic spatial relationships obtained by crowdsourcing spatial relationships from textual narratives.

Image-based Scene Generation. Algorithmic advancements have further fostered the collection of 3D data in urban environments based on crowdsourcing. For example, large collections of geotagged imagery can be stitched together to reconstruct 3D scenes using extensions of photogrammetric and computer vision principles. Geo-tagged images from Flickr are used in [2] to build 3D models of Rome. User-contributed imagery has also been overlayed and embedded in Google Streetview. This feature has been replaced by Photo Sphere[9], a mobile phone app that allows for the creation of 360-degree immersive panoramas. Microsoft Photosynth[10] provides similar functionality, in which a set of overlapping images showing the same scene are analyzed based on matching features to create a 3D point cloud and a model of the captured scene. 3D panoramas and models have been popularized with the emergence of smartphones that not only provide image capturing capabilities, but also have the computing power to generate the 3D scenes as well as the connectivity to communicate the results to respective repositories in real time.

[9] https://www.google.com/maps/about/contribute/photosphere/

[10] http://photosynth.net

Data Quality. Volunteered data is usually provided with little to no information on mapping standards, quality control procedures, and metadata in general. Understanding and measuring the data quality of information provided by volunteers who may have unreported agendas and/or biases is a significant problem in geography today. Recent studies have already started to assess the *quality of user-generated geospatial content* and here specifically OSM content by comparing it to established authoritative mapping organizations such as the UK's Ordnance Survey road datasets [7]. These studies conclude that such data is comparable to more authoritative sources. Road centerlines in OSM were shown to be within few meters of their Ordnance Survey equivalents. The interesting issue of the localness of the GI contributed data is addressed in [8]. By examining two major websites, Flickr and Wikipedia, the authors find that more than half of Flickr users contribute local information on average, while in Wikipedia the authors' participation is less local.

Key Applications

Mapping
User-generated geospatial data directly contributes to the enrichment of existing map datasets. Data includes automatically generated road networks, point-of-interest data, and colloquial geospatial objects.

Understanding Urban Form and Function
Crowdsourcing geospatial data from social media applications provides for a better understanding of urban form and function, i.e., the existing infrastructure and how it is utilized. Such data is directly relevant to applications such as geomarketing, which try to understand socioeconomic processes.

Traffic Assessment
Using mobile phones of users as traffic sensors allows us to compute actual traffic conditions in terms of travel time. Such data is then used to optimize navigation solutions and to improve routing services.

3D Scene Reconstruction
Crowdsourced images can be used to develop 3D models of places. This effort is complementary to dedicated collection campaigns of, e.g., Google.

Cross References

- Crowd mining and analysis
- Human factors modeling in crowdsourcing
- Geographic Information System
- Geography Markup Language
- Text Mining
- Linked Open Data
- Geo-targeted web search
- Spatial Data Mining

- Spatio-temporal Data Mining

Recommended Reading

1. Adams, B. and McKenzie, G. Inferring Thematic Places from Spatially Referenced Natural Language Descriptions, in Sui, D., Elwood, S. and Goodchild, M. Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (eds.), Crowdsourcing Geographic Knowledge: Volunteered Geographic Information (VGI) in Theory and Practice, Springer Verlag, pp. 201-221, 2013.
2. Agarwal, S., Snavely, N., Simon, I., Seitz, S.M. and Szeliski, R. Building Rome in a Day. 12th IEEE Computer Vision conf., pp. 72-79, 2009.
3. Ahmed, M., Karagiorgou, S., Pfoser, D., and Wenk, C. Map Construction Algorithms, Springer International Publishing, 2015.
4. Cope, A. The Shape of Alpha. Available at http://code.flickr.net/2008/10/30/the-shape-of-alpha/, 2008.
5. Crooks, A.T., Pfoser, D., Jenkins, A., Croitoru, A., Karagiorgou, S., Efentakis, A., Lamprianidis, G., Smith, D. and Stefanidis, A. Crowdsourcing Urban Form and Function, Int'l Journal of Geographical Information Science 29(5):720-741, 2015.
6. Goodchild, M.F. Citizens as Sensors: The World of Volunteered Geography, GeoJournal, 69(4): 211-221, 2007.
7. Haklay, M. How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets', Environment and Planning B, 37(4): 682-703, 2010.
8. B. Hecht and D. Gergle. On the localness of user-generated content. Proc. ACM Computer Supported Cooperative Work conf. , pp. 229–232, 2010.
9. Intagorn, S., Plangprasopchok, A., Lerman, K. Harvesting geospatial knowledge from social metadata. In Proc. 7th ISCRAM conf., 2010.
10. Kling, F., Pozdnoukhov, A. When a city tells a story: urban topic analysis. Proc. ACM SIGSPATIAL GIS conf., pp. 482-485, 2012.
11. Lamprianidis, G., Pfoser, D. Collaborative Geospatial Feature Search. Proc. ACM SIGSPATIAL GIS conf., pp. 169-178.
12. Pfoser, D., Brakatsoulas, S., Brosch, P., Umlauft, M., Tsironis, G. and Tryfona, N. Dynamic Travel Time Provision for Road Networks. Proc. ACM SIGSPATIAL GIS conf., pp. 475-478, 2008.
13. Pfoser, D., On User-Generated Geocontent. Proc. 12th SSTD Symp., pp. 458-461, 2011.
14. Skoumas, G., Pfoser, D., Kyrillidis, A. and Sellis, T. Location Estimation Using Crowdsourced Spatial Relations. ACM Transactions on Spatial Algorithms and Systems, in press, 2016.
15. Stefanidis, T., Crooks, A.T. and Radzikowski, J. Harvesting Ambient Geospatial Information from Social Media Feeds. GeoJournal, 78(2): 319-338, 2013.
16. Zhang, Z., Zhang, C., Du, C., Zhu, S. SVM-based extraction of spatial relations in text. Proc. IEEE ICSDM conf., pages 529 –533, 2011.