Identifying Locations of Social Significance: Aggregating Social Media Content to Create a New Trust Model for Exploring Crowd Sourced Data and Information

Al Di Leonardo, Scott Fairgrieve Adam Gribble, Frank Prats, Wyatt Smith, Tracy Sweat, Abe Usher, Derek Woodley, and Jeffrey B. Cozzens

The HumanGeo Group, LLC
Arlington, Virginia, United States
{al,scott,adam,frank,wyatt,tracy,abe,derek}@thehumangeo.com

Abstract. Most Internet content is no longer produced directly by corporate organizations or governments. Instead, individuals produce voluminous amounts of informal content in the form of social media updates (micro blogs, Facebook, Twitter, etc.) and other artifacts of community communication on the Web. This grassroots production of information has led to an environment where the quantity of low-quality, non-vetted information dwarfs the amount of professionally produced content. This is especially true in the geospatial domain, where this information onslaught challenges Local and National Governments and Non-Governmental Organizations seeking to make sense of what is happening on the ground.

This paper proposes a new model of trust for interpreting locational data without a clear pedigree or lineage. By applying principles of aggregation and inference, it is possible to identify locations of social significance and discover "facts" that are being asserted by crowd sourced information.

Keywords: geospatial, social media, aggregation, trust, location.

1 Introduction

Gathering geographical data on populations has always constituted an essential element of census taking, political campaigning, assisting in humanitarian disasters/relief, law enforcement, and even in post-conflict areas where grand strategy looks beyond the combat to managing future peace. Warrior philosophers have over the millennia praised indirect approaches to warfare as the most effective means of combat—where influence and information about enemies and their supporters trumps reliance on kinetic operations to achieve military objectives. This approach can assist in the perpetuation and management of peace easier. However, successful indirect means are only as good as the human geographical data assembled.

Aggregating, categorizing, and interpreting geo-referenced or geo-located data on a specific population lies at the heart of contemporary counter-insurgency, counter terrorism, and nation-building and good governance. Western military campaigns in Afghanistan and Iraq are obvious examples as well as intensifying conflict scenarios

G. Meiselwitz (Ed.): SCSM 2014, LNCS 8531, pp. 170–177, 2014.
© Springer International Publishing Switzerland 2014

throughout Africa pose looming future challenges. Isolating insurgents and terrorists from populations and potential recruits is essential to winning at war's 'moral' level—inextricably linked to a grand strategy—but this cannot be accomplished using biased, insufficient, or conversely, unwieldy amounts of socio-cultural data. Unfortunately for military planners, this is often the nature of information found in the open source domain.

This paper discusses how simplified and trustworthy human geography data sources and fusion methodologies can empower U.S. military planning by analyzing Open Source Information/Intelligence (OSINT) from social media to generate significant human geography data on populations amidst a dense stream of Web-based community communication. In so doing, it proposes a new model of trust for interpreting locational data—one that involves applying principles of aggregation and inference to identify locations of social significance and discover 'facts' asserted by a collective of independent data producers. This approach is critical to winning indirect battles and 'doing no harm' in non-Western human terrains vulnerable to exploitation by terrorists and insurgents. This model has other applications such as assisting people and property before, during and after disasters or helping law enforcement agencies mitigate criminal activities by analyzing gang online and offline presence.

2 Value of Crowd-sourced Data

Data sources continually evolve and change. Social movements are rarely characterized by known 'knowns'; shifting online commentaries, or the capture and upload of a singular event, can precipitate sudden and massive human change (the Arab Spring is perhaps the most prominent recent example). By the time data has been validated, verified, synthesized with other government sourced information, it may no longer be relevant to analysts, planners or decision-makers because the event might have al-ready occurred, names have changed, or the meaning of an observation has been completely altered. The shifting social and geopolitical environment is, in sum, constantly redefining and challenging analysts' observational capabilities.

However, crowd-sourced information—that is, social media data used for analysis—is of great analytical value because it presents data snapshots of specific spaces, places and times. Harnessed efficiently and effectively, it allows analysts to identify important socio-cultural landmarks (observations) at any given moment. This analyzed crowd-sourced information can help make sense of a wide variety of human events of interest (such as disaster response) for governments, first responders, politicians, international organizations, and many others.

Present in social media data is the latent capability to update changes to previously defined socio-cultural landmarks, including highly relevant places that may not have been recognized by decision makers, military personnel or diplomats. For instance, captured near real-time crowd source updates can lead to a better understanding of culturally significant artifacts, institutions and landmarks that might help facilitate or strengthen partnerships with regional actors.

3 Military Planning Using Social Media: A Matter of Trust

Western militaries' planning lifecycles revolve around geospatial intelligence systems designed to interpret socio-cultural data. However, these systems often demand too much from analysts in order to reach their potential. Many expert analysts do not have the fundamental Geographic Information Systems (GIS) skills (or resource-intensive training) required to ask informed questions and properly interpret GIS system responses. Further, it is much easier for analysts to 'connect the dots' and understand the associations and relationships between various data when using highly structured, well-formatted information. Disparate human geography data derived from social media is a different story. Analysts experience a fundamental trust challenge with social media data at a time when the volume, dispersion, and discrepancies of grassroots information increase by the second.

What is required of the human geography community to effectively support the US military's indirect planning approaches and the 'do no harm' principal when dealing with non-Western foes in their own at-risk communities? HumanGeo has developed a simple, non-parametric approach to socio-cultural data mining built around the National Geospatial-Intelligence Agency's (NGA) "Thirteen Themes of Human Geography" to provide precisely this type of support.[1] HumanGeo's quantitative and predictive solution, is based on a simplified fusion of geography feature data using variable precision data encoding known as "geohash" that allows social patterns to more easily emerge from voluminous data sets [2]. Consequently, this provides immediate answers from crowd-sourced data for planners concerned with generating socio-cultural queries along the lines of "where, when, who, what, why?"

Invented in 2008 by Gustavo Niemeyer, the geohash is a latitude/longitude geocode system that combines decimal degrees latitude and longitude annotations into a single string-variable that defines a certain size box. These synthetic boxes are described as a centroid point and a value range of latitude, with a value range of longitude. By looking at the world as a series of boxes (geohashes) and the attributes that pertain to these boxes (e.g. places of interest (transportation, religious, political, economic, etc), events [sporting, political, violent, economic, etc.], the presence of national interests and/or those of allies, etc.), various types of planners are able to more efficiently fuse disparate data sources into a single, unified view. This approach empowers analysts by providing them with simplified tools that allow them to explore data based on their own expert hypotheses.

---

[1] NGA's thirteen Themes of Human Geography include the following: Transportation, Significant Events, Religion, Medical/Health, Language, Land-Use, Groups, Ethnicity, Education, Economy, Demographics, Communications, and Climate (water).

[2] The best summary description of geohash is found on Wikipedia: `http://en.wikipedia.org/wiki/Geohash` (accessed 13 October 2013).

4 Aggregation Models

HumanGeo has developed a four-step methodology for applying this variable encoding scheme to human geography data. The process is detailed below:

- Aggregate
- Annotate
- Automate
- Analyze

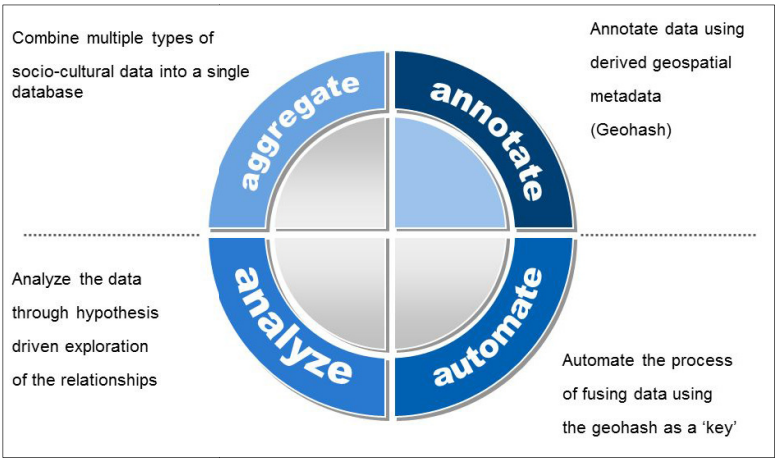

Fig. 1. HumanGeo's Four-Step Process for Social Media Data Retrieval and Enrichment

This four-step process (Figure 1 above) describes how the data was retrieved and enriched in a recent three-country study examining social media in Mexico, Syria, and Nigeria:

4.1 Aggregate

HumanGeo's data aggregation process is unique due to its ability to identify and aggregate disparate geo-located/geo-referenced human geography data from multiple crowd sourcing mediums using algorithms to identify those observations in numerous languages. The aggregation process combines multiple types of data into a single database. Social media data, in its typical state, is otherwise too disparate and cumbersome to be analyzed in aggregation without tools to organize and categorize it—precisely what HumanGeo does. In other words, one cannot simply download Wikimapia, Google Places, or Panoramio to create visualizations or databases for easy assessment.

4.2 Annotate

Annotation involves synthesizing metadata with each dataset using the geohash to encode location data, and adding additional geospatial metadata to each point (i.e. observation or raw data). This is where HumanGeo adds the *Thirteen Themes of Human Geography* to the individual observations (e.g. an identified geo-located or geo-referenced social cultural landmark).

4.3 Automate

This step refers to human geography theme categorization by using software automation to "tie data together" with the geohash to create a grid system for connecting data elements. Every observation is then categorized as one of the defined *Thirteen Themes of Human Geography* whenever possible. The Aho-Corasick sub-string matching algorithm is used to help categorize observations as one of thirteen themes by matching the title and/or name of the observation.[3] The title is translated and transliterated against a list of known themes or category terms. If the match contains a known "stop" word, it is ignored. For instance, "ward" indicates a possible medical/hospital designation, but "forward" does not; therefore, "Forward" is in the set of medical/hospital stop words. If a match is found, the first matching category is used. Observations were not categorized if the title did not match one of the designated *Thirteen Themes of Human Geography*. This process helps analysts because it can aggregate observations and possibly identify clusters of activities and/or relationships between several observations.

4.4 Analyze

This step allows for the analysis of data through hypothesis-driven exploration of the data relationships, enabled by Google Earth Keyhole Markup Language (KML) exports of the fused/combined data layers. As in step three (automate), each dataset is simplified into a "data layer" that is represented as a series of geohash boxes. These data layers are then combined into a "context lens"—user-specific data layers related to a particular analytic hypothesis.

This process, in summary, provides a mechanism for annotating disparate human geography datasets with a geohash format, fusing the data together (at different levels of precision) using the geohash as a connecting key, and projecting patterns and relationships of the data into Open Geospatial Consortium (OGC) formats for visual analytics. For example, one could portray macro-level indicators of violent events by breaking the world up into squares of approximately 150 km each (geohash 3), affording analysts a view of the attributes of each square to enable inference-making. Used differently, one could examine criminality in subdivisions of major cities by encoding crime locations into a geohash string of length six or seven, thereby breaking up the city up into squares of approximately 610 meters or 118 meters,

[3] See Aho, Alfred V.; Margaret J. Corasick (June 1975). "Efficient string matching: An aid to bibliographic search". *Communications of the Association for Computing Machinery* 18 (6): 333–340. More details at: http://xlinux.nist.gov/dads/HTML/ahoCorasick.html Accessed on 18 September 2013.

respectively, then analyzing the event characteristics of each square. Further, the application can draw out (and visualize) critical socio-cultural anomalies in specific regions in a scalable manner using crowd-source data, as the Syria case demonstrates.

Case Study: Syria. HumanGeo applied its crowd source data analysis methodology to Syria in September 2013. The scalable test assessment generated and visualized critical socio-cultural observations on 92,012 resolved entities (combinations of similar observations) at the national, regional and local levels using data from Wikimapia, Google Places and Panaramio.

The national level assessment visualized social media activity throughout Syria, noting spikes in usage along the Mediterranean coast (as anticipated), as well as in historic tourist destinations like Palmyra—essentially an oasis of social media in an otherwise barren area. This highlights the methodology's ability to identify socio-cultural anomalies in a region or country.

Regionally, the data used highlighted a cluster of social media activity along the Euphrates river in the Dayr az Zawr region—another seemingly desolate locale. However, visualization of social media in the form of bar graphs created a pictorial narrative highlighting the importance of the Euphrates as a metric for mapping social development over time. Such data can also lead to making inferences about why social media growth is occurring in the absence of additional infrastructure development.

Data sources used were from the town of Palmyra. The visualizations demonstrated their utility in identifying landmarks of socio-cultural significance that may have been previously unknown. Such discoveries could enhance previously weak knowledge of important local artifacts, institutions or traditions, thereby facilitating or strengthening relationships with local actors.

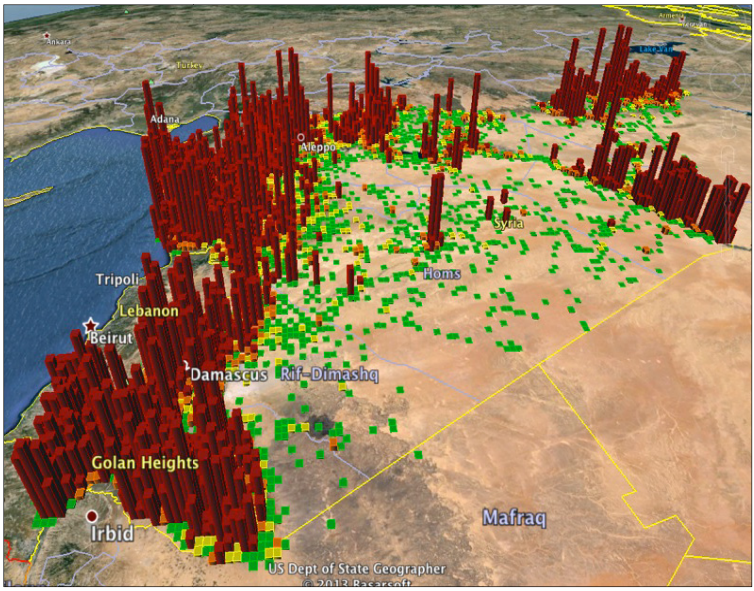

Fig. 2. HumanGeo's National level assessment of crowd source data in Syria (Google Earth view). Palmyra is highlighted in the circle.

5 Geospatial Entity Resolution

Lise Getoor and Ashwin Machanavajjhala note that entity resolution is essentially a method of entity disambiguation.[4] Entity resolution, for the purposes of HumanGeo's social media data sets and enrichment process, also means identifying similar entities—an entity being a set of one or more geo-located observations—and combining and categorizing them with confidence for purposes of inference-making. HumanGeo's geospatial entity resolution algorithm and resulting confidence score methodology—developed from a number of academic resources, especially the Stanford Entity Resolution Framework[5]—is unique in this respect.

When applying HumanGeo's methodology for entity resolution, observations from multiple sources undergo an entity resolution process using the following algorithm: every defined observation is allocated a Geohash bounding box and is at the center of nine Geohash bounding boxes. The surrounding Geohash bounding boxes are called 'neighbors'.

Step one involves finding all observations within a 3x3 Geohash box. Once all obser-vations have been defined within the box, the jaro-winkler string-matching algorithm is used to compute the similarity of titles (names) between the observations. Any observation with a similarity equal to or greater than 0.9 is resolved by combining similar observations as one.

Spatial data can be analyzed to verify and validate geospatial coordinates, and conversion occurs to make spatial coordinates available in latitude/longitude, Military Grid Reference System (MGRS), and geohash formats. Data that is inherently non-spatial in nature is aggregated and fused, but no spatial attributes are assigned to the records. However, for non-spatial full-text data sources, HumanGeo applies place-name extraction software to isolate place-names when referenced within the text. For example, if a full-text field of a data source references "Washington, DC," or "White House," that place-name will be added as a geospatial attribute of the data and will include latitude/longitude, MGRS and geohash coordinates associated with it.

Measuring confidence in entity resolution

HumanGeo measures confidence in the entity resolution process by multiplying three factors: the number of sources, the average jaro-winkler[6] score, and the dispersion—the maximum distance between two observations within an entity. A weighted value system has been applied to account for the number of sources used: one source identifying an entity has a confidence interval rating of 0.6; two sources identifying an entity has a confidence interval rating of 0.8; three sources identifying an entity has the highest confidence interval rating of 1.0.

---

[4] See Lise Getoor and Ashwin Machanavajjhala, "Entity Resolution: Tutorial" (2012), at: `http://www.umiacs.umd.edu/~getoor/Tutorials/ER_VLDB2012.pdf`

[5] See Stanford Entity Resolution Framework (Stanford University), at: `http://infolab.stanford.edu/serf/#motivation`. Accessed 18 September 2013.

[6] See `http://xlinux.nist.gov/dads/HTML/jaroWinkler.html` (accessed 18 September 2013).

6 Conclusion

A simplified and trustworthy human geography analysis methodology as employed by HumanGeo can empower end users (based on their legal authorities) in using OSINT from social media to generate significant human geography data on populations amidst a dense stream of Web-based community communication. This enables analysts to put rich socio-cultural data at their fingertips, highlighting significant social activity and previously unknown, but germane, socio-cultural nodes. This approach is strategically important to helping win indirect battles and 'doing no harm' in non-Western human terrains vulnerable to exploitation by terrorists and insurgents.

Beyond the social media retrieval, enrichment, and visualization capabilities discussed here, companies such as HumanGeo are developing other systems to capture enriched (i.e. with administrative boundaries, language identification, sentiment, etc.) social media data in near-real-time. HumanGeo's system also tags data with one or more event topics from a set of event categories tailored to client analysts using simple keyword matching. This event detection capability is pushing back boundaries using unstructured topic learning as a means of developing even greater insights. This system makes it easier and faster for analysts to find and assess data, since data are pre-categorized in a variety of ways optimized for searching by end-users. Cascading political and security-related events in the Middle East and Africa demonstrate the requirement for such a capability and furthering this line of research.