Title
Deriving Locational Reference through Implicit Information Retrieval

Journal
International Conference on GIScience Short Paper Proceedings, 1(1)

Authors
Hervey, Thomas
Kuhn, Werner

Publication Date
2016

DOI
10.21433/B3110tn5s4v9

Peer reviewed

Deriving Locational Reference through Implicit Information Retrieval

T. Hervey, W. Kuhn

Department of Geography and Center for Spatial Studies,
University of California, Santa Barbara, CA 93106
Email: {thomas.hervey; kuhn} @geog.ucsb.edu

Abstract

The often fragmented process of online spatial data retrieval remains a barrier to domain scientists interested in spatial analysis. Although there is a wealth of hidden spatial information online, scientists without prior experience querying web APIs (Application Programming Interface) or scraping web documents cannot extract this potentially valuable implicit information across a growing number of sources. In an attempt to broaden the spectrum of exploitable spatial data sources, this paper proposes an extensible, locational reference deriving model that shifts extraction and encoding logic from the user to a preprocessing mediation layer. To implement this, we develop a user interface that: collects data through web APIs and scrapers, determines locational reference as geometries, and re-encodes the data as explicit spatial information, usable with spatial analysis tools, such as those in R or ArcGIS.

1. Introduction

GIS advancements have produced a growingly complex general-purpose toolbox rather than functionality tailored to domain-specific questions. Frequently, domain scientists including Green (2015: 717) highlight the salient lagging data and tool limitations associated with GIS. As Kuhn (2012: 2267) notes, it is essential to rethink the fundamentals of spatial information while promoting clarity that cuts across technical boundaries and broadens spatial literacy for non-experts. Contributing to the work by Kuhn and Ballatore (2015: 219) and Vahedi *et al*. (2016) to design an intuitive GIS language for question-driven spatial studies, we focus on bridging the gaps between data discovery and spatial analysis tools by broadening the spectrum of exploitable spatial data sources.

Compared to the vast amount of implicit spatial data (hidden location attributes often in the form of metadata, auxiliary place names, and geotagged attributes (Heinzle and Sester 2003: 335)), there remains a relatively limited quantity of online explicit spatial data (georeferenced geometry-based features (Brisaboa *et al.* 2011: 358)). When available, explicit data are typically served from a limited number of administrative portals or require intensive energy and time from a user searching, exporting, encoding and cleaning before being usable (Munson 2013: 65). These preprocessing requirements limit the feasibility of question-driven spatial analysis (Vahedi *et al*. 2016) and force domain scientists to base their studies on data availability.

It is clear that implicit spatial data is an attractive alternative. However, as the numerous research challenges associated with GIR (geographic information retrieval) suggest (Jones and Purves 2008: 219), current methods do not provide adequate solutions to navigate, gather or utilize the mass of heterogeneous implicit spatial data spread across the array of online repositories. Custom-constructed web API requests and scrapers can help retrieve and process this unpublished information. Yet, without technical expertise to build new or use existing

modules, this information remains unobtainable. This begs the question, why so few GIR studies have focused on aiding implicit spatial data retrieval.

To explore this idea, we propose an extensible locational reference deriving model that shifts extraction and encoding logic from the user to preprocessing software. In the following sections, we describe previous work and provide a prototype architecture with applications to test the plausibility of a model that adopts the core concepts of spatial information (Kuhn 2012: 2267).

2. Previous Work & Limitations

There have been many notable efforts to address the difficulties of GIR, but few have focused on the extraction and encoding of implicit spatial data. INSPIRE[1], NSDI[2], and SPIRIT (Jones *et al.* 2004: 125) for example, are spatial data infrastructure and search engine projects that focus on standardizing and indexing web documents rather than broadening access to new faster-growing data repositories (Jones *et al.* 2004: 125; Brisaboa *et al.* 2011: 358).

Google's Fusion Tables[3] have simplified data integration and ArcGIS Online has grown online spatial catalogs through user sharing and publishing. Yet, these tools do not aid in data extraction from external sources. Enterprise products like Crimson Hexagon[4] and Temboo[5] respectively provide users with spatial social media analytics and code snippets to ease web API querying. But these solutions provide neither data source extensibility, nor recognize multiple location types. Furthermore, studies on linked data, the semantic web, and semantic gazetteers (Cardoso *et al.* 2016: 389; Gao *et al.* 2014) propose a traversable data-centric organization of the web (Kuhn *et al.* 2014: 173; Wiegand and García 2007: 355). But until a semantic model broadly leverages new and existing data, significant user effort will remain necessary for collecting and preprocessing. The following architecture aims to address these issues and bridge the gap between discovered data and analysis tools.

3. Architecture

The proposed model's implementation will be accessible via web interface and GIS plugins. It is not an independent search engine, but rather works as mediation layer once a desirable data source has been found and a user supplies the text parameters noted in Table 1. It does not aim to replace, but rather leverage, existing extraction and encoding methods as well as supply locational reference deriving logic not present in current systems. Illustrated in Figure 1, the mediation layer divides preprocessing (extracting and encoding) into four sequential tasks: extraction, context building, geometry matching, and encoding.

Table 1. User interface and plugin input parameters

Mandatory	Optional
Extraction source (*e.g.* URL or API endpoint)	Basic aggregation (*e.g.* state-level)
Source type (*e.g.* Twitter tweets or hashtags)	Basic conditionals (*e.g.* exclude attributes)
Output geometry type (*e.g.* point, polygon)	Error handling (*e.g.* toponym conflicts)
Output file format (*e.g.* **.**shp, **.**kml, .geoJSON)	

[1] http://inspire.ec.europa.eu/

[2] http://www.fgdc.gov/nsdi/

[3] https://support.google.com/fusiontables/answer/2571232?hl=en

[4] http://www.crimsonhexagon.com/

[5] https://temboo.com/

The Extraction task makes up the majority of the codebase, which will expand as interest in new data sources emerge. It leverages parameterized HTTP requests constructed by web APIs (for database retrieval) and scrapers (for document text retrieval). The result will often be a structured or semi-structured tree of attributes (*e.g.* .json, .xml) presented to the user for accuracy feedback.

The second task searches the data for potential location characteristics in the form of raw geometry or toponyms using a combination of text matching and natural language processing paired with gazetteer lookups (*e.g.* GeoIP, GeoNames, Getty Thesaurus of Geographic Names). When applicable, user feedback will be necessary for place name disambiguation, scoping, and aggregation.

The third task handles geometry pairing and the fourth, output file formatting. Once location characteristics have been determined, associated geometries are retrieved from repositories (*e.g.* TIGER/Line, Open Street Map) and paired with the data. Additionally, rudimentary spatial extent and relationship logic is supported (using ESRI's Java geometry API, GDAL, etc.), such that a *within* operation could aggregate geotagged Flickr photo points to a bounding county polygon. Finally, conversion tools (*e.g.* GDAL ogr2ogr, Python pyshp), encode the data in an array of formats (*e.g. .shp, .kml*)., resulting in a spatially explicit data product usable in analysis.

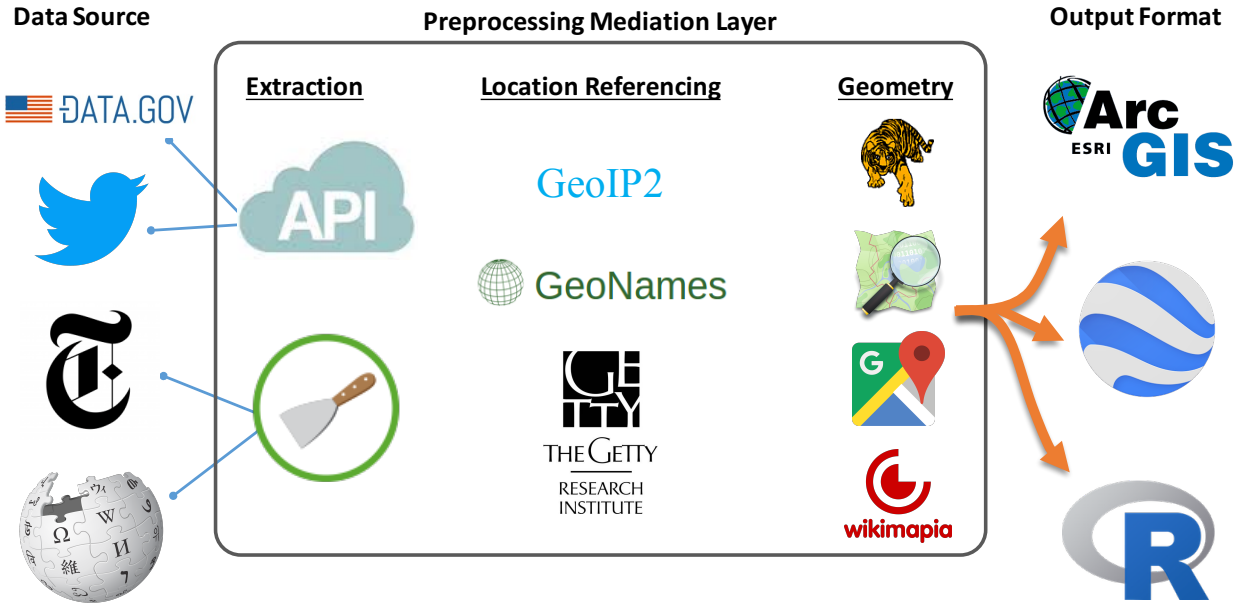

Figure 1. Examples of leveraged services between implicit data and analysis tools

4. Application

To articulate the model's potential, we present working and hypothetical examples where domain scientists encounter data retrieval problems. In our working example, a social scientist wants to examine the spatial interaction between crime incidents and income across Seattle, USA. She has statistical experience with R and has imported a block-group median household income shapefile, but cannot find explicitly spatial crime data. She has, however, found a well formatted crime statistics table on Seattle's web portal, and wants to extract the *time*, *crime type*, and *charge* attributes by city-block. Traditionally, to integrate this information, she would have to download and import the table, for example, into ArcGIS, remove unwanted attributes, download a block-group shapefile, aggregate the crime points, and handle geocoding anomalies. Alternatively, the previous steps are simplified by supplying the model's web interface with the

table's API endpoint, specifying which attributes to keep at which aggregation level, and specifying polygon shapefile as the output. On the back end, the software retrieves the data as JSON objects, determines street address as its location reference, geocodes and aggregates these points to block-group geometry served by *data.gov*, and produces a shapefile using ogr2ogr.

Our hypothetical example adds complexity when a user requires data from social media. A political scientist wanting to understand spatial voting trends between Californian counties during a presidential primary, decides to use trending Twitter hashtags as predictors. Without programming experience, he is limited to Twitter's simple query building wizards that returns unwieldy and spatially unreferenced tweet data. Instead, by providing twitter API credentials, hashtags of interest, and California counties as text parameters to the interface, the model can filter by hashtag, extract user's location information (through geotags or home locations) and aggregate tweet count to the county level, producing a .kml file from OpenStreetMap geometry.

5. Conclusions

We have presented a locational reference deriving model and associated prototype preprocessing layer that has the potential to promote critical spatial thinking by expanding data source options. Currently, the model's integrity is limited by the credibility of its data retrieval sources, and limited to handling vector data. Therefore, future research will investigate integrating new data sources, analyzing feedback to promote VGI (volunteered geographic information) supported gazetteers, as well as integrating data credibility metrics.

References

Brisaboa N, Luaces M, and Seco D, 2011, New Discovery Methodologies in GIS. *Geographic Information Systems*, 358–376.

Cardoso S, Amanqui F, Serique K, dos Santos J, and Moreira D, 2016, SWI: A Semantic Web Interactive Gazetteer to Support Linked Open Data. *Future Generation Computer Systems*, 54:389–398.

Gao S, Li L, Li W, Janowicz K, and Zhang Y, 2014, Constructing Gazetteers from Volunteered Big Geo-Data based on Hadoop. *Computers, Environment and Urban Systems*.

Green T, 2015, Places of Inequality, Places of Possibility: Mapping "Opportunity in Geography" Across Urban School-Communities. *Urban Review*, 47(4), 717–741.

Heinzle F and Sester M, 2003, Derivation of Implicit Information from Spatial Data Sets with Data Mining. *Cartography*, 35(4)335–340.

Jones C and Purves R, 2008, Geographical information retrieval. *International Journal of Geographical Information Science*, 22(3):219–228.

Jones C, Abdelmoty A, Finch D, Fu G, and Vaid S, 2004, The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. *Geographic Information Science, Proceedings,* 3234:125-139.

Kuhn W and Ballatore A, 2015, Designing a Language for Spatial Computing. *AGILE 2015, Lecture Notes in Geoinformation and Cartography*, 219–234.

Kuhn W, Kauppinen T, and Janowicz K, 2014, Linked Data- A Paradigm Shift for Geographic Information Science. *Proceedings of The 8th International Conference on Geographic Information Science*, Vienna, AUT, 173–186.

Kuhn W, 2012, Core Concepts of Spatial Information for Transdisciplinary Research. *International Journal of Geographical Information Science*, 26(12):2267–2276.

Munson M, 2012, A study on the Importance of and Time Spent on Different Modeling Steps. *ACM SIGKDD Explorations Newsletter*, 13(2):65–71.

Vahedi B, Kuhn W, and Ballatore A, Question Based Spatial Computing - A Case Study. *AGILE 2016 (in press).*

Wiegand N and García, C, 2007, A Task-Based Ontology Approach to Automate Geospatial Data Retrieval. *Transactions in GIS*, 11(3):355–376.