Workshop on

GIScience in the Big Data Age

In conjunction with the seventh International Conference on
Geographic Information Science 2012 (GIScience 2012)

Columbus, Ohio, USA. September 18th, 2012

Proceedings

Organizing Committee

Krzysztof Janowicz, University of California, Santa Barbara, USA
Carsten Keßler, University of Münster, Germany
Tomi Kauppinen, University of Münster, Germany
Dave Kolas, BBN Technologies, USA
Simon Scheider, University of California, Santa Barbara, USA

Programme Committee

Benjamin Adams, University of California, Santa Barbara, USA
Boyan Brodaric, Geological Survey of Canada, Canada
Oscar Corcho, Universidad Politecnica de Madrid, Spain
Isabel Cruz, University Of Illionois, USA
Mike Goodchild, University of California, Santa Barbara, USA
Willem Robert van Hage, Vrije Universiteit Amsterdam, NL
Pascal Hitzler, Wright State University, USA
Bin Jiang, University of Gävle, Sweden
Werner Kuhn, University of Muenster, Germany
Jens Lehmann, , University of Leipzig, Germany
Matthew Perry, Oracle, USA
Christoph Schlieder, University of Bamberg, Germany
Claus Stadler, University of Leipzig, Germany
Kristin Stock, University of Nottingham, UK

Table of Contents

Building streaming GIScience from context, theory, and intelligence

Carson J. Q. Farmer, Centre for GeoInformatics, University of St Andrews
Alexei Pozdnoukhov, National Centre for Geocomputation, National University of Ireland, Maynooth

Abstract

In this paper, we suggest that a focus on the current strengths of GIScience, coupled with a strategic view towards the future of data-driven research, can help to propel GIScience forward as a leader in data-intensive social science. With the majority of the world's data being embedded in space, it is only natural for GIScience to take a leadership role in the analysis and understanding of individual, local, regional, and global data by providing context, theory, and intelligence to an otherwise data-centric science. We suggest that in order to avoid the limitations of current data storage, management, and retrieval practices, a focus on real-time, intelligent analysis of data in a streaming framework is the most logical step forward. This type of analytical framework addresses two key concerns of GIScience: scalability and relevance. By focusing on results and models over raw data, we build on the current strengths of GIScience, leading to process-based research that is scalable over the long-run. Furthermore, by developing the methods and theories around streaming spatial data, we ensure that GIScience remains relevant in the increasingly data-intensive world of computational social science research.

Introduction

Data has always been big. Researchers, businesses, and government departments have continually collected and maintained large datasets relevant to their area of expertise. For decades, 'large' has been a moving target, chiefly dictated by the accelerating decrease in processing and storage costs. Despite a long history behind the use of large data sets for decision making and analysis by business and government, 'big data' has only recently emerged as an area of inquiry unto itself. This late emergence of big data is likely a function of several factors, including the fact that our ability to sense, collect, and process data from multiple sources is now far outpacing our ability to store and manage said data [1] (Figure 1). Additionally, we are beginning to see a major shift in the way many scientists are thinking about information and analysis, leading to a more data-intensive social science where hypotheses are generated through an abductive process (i.e., hypotheses are developed to account for observed data).

With increasingly efficient means of generating data from multiple sources, the amount and number of different types of data that industry and government collect on a regular basis has reached critical levels. Additionally, information pertaining to all facets of society, from public databases such as national census', to private customer databases, to community-built open data sources, are increasingly being linked to geographical locations. Indeed, as much as 80% of all information held by business and government may be geographically referenced [2,9], and this number is only likely to grow as more and more organisations realise the importance of locational information [6]. The massive amounts of data being collected, coupled with the additional complexity that spatial data yields, means that the traditional GIS model of data storage and management is no longer sufficient, and that new insights into spatial data management and analysis are required.

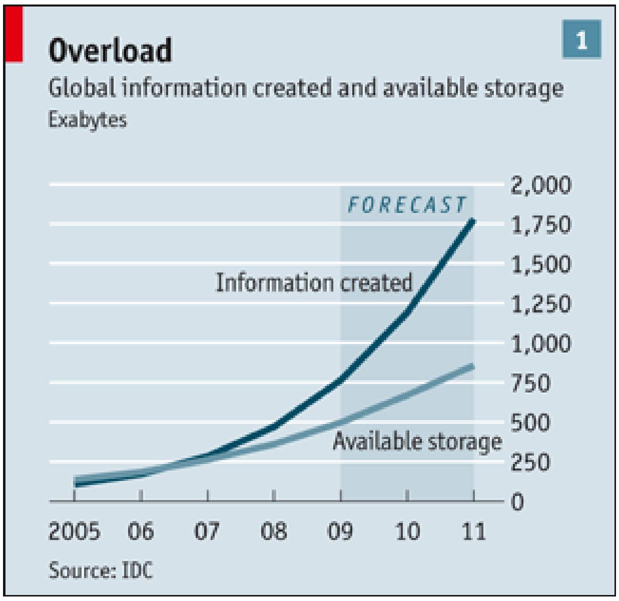

Figure 1: Global information created and available storage space. From The Economist (Feb 25 2012).

In the past, the limiting factor for geographic information science has been data. This is because GIScience has long been an abductive science, and inferring processes from patterns requires detailed information on both the location and attributes of the particular phenomenon under consideration. This data constraint is particularly relevant to GIScience due to the (previously) prohibitive costs associated with collecting data over large spatial extents. As such, despite rapid development of theory throughout the 1980s and 90s in GIS and quantitative geography, many new ideas were left untested due to a lack of tools and/or technology. Currently, GIScience (among other fields) is experiencing the opposite problem: the rapid pace of data collection is exceeding the reach of current theory and methods [8].

From a data management perspective, it is no longer feasible to store data using the the traditional geodatabase design. As GIScience continues to move from a set of tools and techniques for working with geographic data to a science of geographic information [4], our remit is shifting from the development of software and hardware solutions for capturing, storing, managing, retrieving, and disseminating geographic data to one of context, theories, and intelligence. Indeed, the true strengths of GIScience lie in its ability to provide substantive knowledge, develop geospatial thinking/literacy, ask the right questions, interpret data and outputs within the correct context, and understand the implications of research findings. With this in mind, the best way that GIScience can contribute to a data-intensive science is to focus on its strengths, and use its home field advantage in full when forced to enter a tournament with data-intensive computer science.

A way forward

While GIS development and design has, in the past, been done primarily within GIScience, it is unrealistic to assume that GIScientists should continue to develop software and hardware solutions for processing and disseminating geographic data on its own. These are things that computer scientists, engineers, and software developers working on GIS and related information technologies can do, and do well. By focusing on higher-level problems and concentrating on what can be done with data rather than a focus on data itself, GIScientists can provide the context, theory, and intelligence required by other fields to provide solutions to spatial problems.

The theories and techniques behind spatial analysis, GISystems, spatial statistics, spatial data representation, and other developments within the realm of GIScience are being stretched to capacity by modern data sets. In a survey of fifty-eight key researchers in the field of spatial analysis (and more broadly, GIScience), overcoming methodological limitations imposed by large datasets was highlighted as a key challenge for the future [8]. Furthermore, with the ubiquitous adoption of web-based mapping systems and increasing awareness that 'space matters', GIScience as a field is now a net-exporter of methods and ideas [7], and maintaining the home field advantage in terms of spatial analysis will be a key deciding point in the success of GIScience in a data-driven world.

Context, theory, and intelligence

A GIScience focused on the implementation details of GISystems will develop 'users' rather than 'researchers', where users "will see the world through a lens defined by the constraints and principles of database design" [4] and legacy GIS thinking. Instead, we suggest that a more fundamental approach, based on using geographic knowledge and theories to refine our methods and expand our understanding of spatial processes is warranted. Here, the use of geographic information theory is useful, but not sufficient on its own. Process-based theories such as spatial interaction, spatial behaviour, and spatial diffusion should also be considered, and can contribute to a process- or model-based GIScience where models and linked information are used to solve fundamentally geographic problems. The key component to developing geographical intelligence is the integration of context (i.e., when, where, and what spatial patterns were generated) and theory (i.e., why and how do the observed patterns correspond to known processes).

For example, in studies of commuting, researchers are often interested in predicting the number of commuters traveling between a particular origin and destination pair, or the destination that a commuter at a particular origin might choose from a range of possible destinations. Here, the context is clear: the phenomenon under investigation is commuting and the spatial setting is (usually) some urban environment. As such, context encompases the problem scenario (i.e., predicting commuting flows), information requirements (i.e., counts, socio-economic factors, distances), required level of detail (i.e., macro vs micro), and target outputs (i.e., maps, model parameters, predictions). While context guides our treatment of the problem, theory provides the means to developing a solution. In the above case of predicting commuting flows, we can incorporate theories of spatial interaction (i.e., macro commuting) or spatial choice (i.e. micro commuting) to inform our models. This provides us with additional information requirements and a framework within which to compare our results (i.e., do these results make theoretical sense?). In this sense, developing geographical intelligence is a synergistic process: intelligence comes from understanding spatial processes, spatial processes can be approximated via models, models are directly informed by theory, and theory is inextricably linked to the context within which it operates.

With these points in mind, we suggest that the most logical step forward for GIScience is a model-centric view on analysis, where the focus is on real-time, intelligent analysis of data in a streaming framework. This type of research framework allows GIScientists to focus on the strengths of GIScience (i.e., ontologies, spatially-aware statistical methods and theories), and avoids problems associated with the storage and retrieval paradigm of traditional geodatabases. In the following section we describe the benefits of streaming analytics, and outline a framework which would allow a model-centric GIScience to contribute to the big data agenda.

Focus on streaming

Applications where real-time analysis of millions of temporally varying spatially referenced samples over wide geographical areas are required are becoming an everyday necessity. In addition to increasing volumes of sensor data produced by city infrastructures, real-time data feeds of users' activities through various applications such as Twitter [12], Flickr, Foursquare and others are becoming increasingly available. Location-aware applications and location-based services have become popular in recent years, such that many data feeds now have a geographic element by default, thus becoming forms of volunteer geographic information. There is much potential for GIScience to explore spatial relationships in such data to understand spatial patterns emerging from low-level human actions and interactions.

Large data volumes and the complex mechanisms behind data generation processes require new analytical approaches which are flexible, non-parametric, computationally efficient, and able to provide interpretable results for modelling non-stationary and non-linear processes in data-rich situations. Machine learning offers a selection of online algorithms designed for streaming data, where it is assumed that every data sample can only be seen once and processed in constant time. Algorithmic solutions for such systems can be borrowed from the signal processing field, where streaming data have been studied for decades and efficient incremental methods for typical optimisation problems (e.g., least squares optimization, matrix inversion and decompositions) have been developed. Such approaches offer straightforward extensions of many spatial statistical models to be applied in real time to temporally-varying data streams.

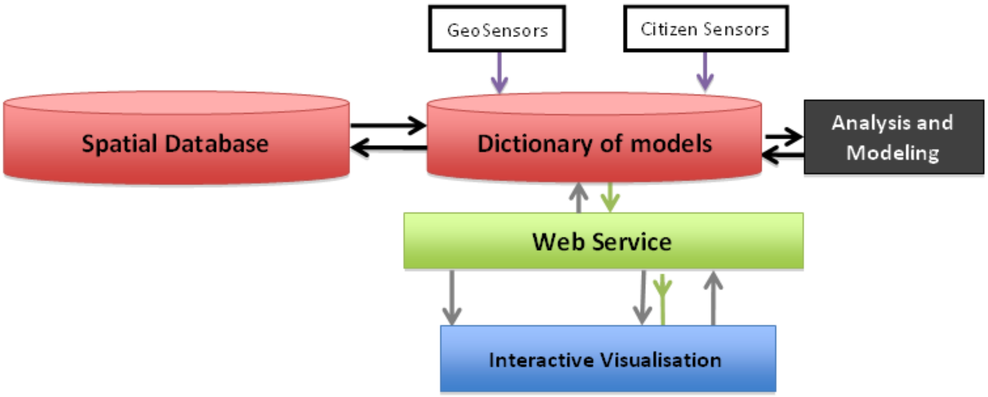

Figure 2: A transition from data- to model-centric design turns a GIS into a spatial knowledge discovery platform.

Excitement surrounding big data is continuing to build (see [1]), yet for the most part, big data analytics for spatial data has focused on data visualisation and descriptive analysis - a return to a data-intensive version of pre-1950s geography [5]. With a large range of spatial theories and domain specific techniques available, GIScience should not be dominated by simplistic description, but maintain its focus on theoretical understanding. A model-centric GIScience returns focus to modelling and understanding the underlying spatial processes rather than ongoing attempts to overcome the storage bottleneck. There is little sense in storing all the data if a model with a limited number of parameters is able to reproduce the phenomenon within a required level of accuracy: “perfection is achieved not when there is nothing more to add, but when there is nothing left to take away[1]". This vision implies a different architecture for information systems which turns a GIS, traditionally centred around a spatial database, into a knowledge discovery system where we "bring the analysis to the data" and an analyst operates with a set of models (Figure 2) rather than with a raw data storage interface and a set of toolboxes. These models, expressed via a set of parameters and hyper-parameters, can be arranged in a dictionary trained on streaming data and capturing the typical (or atypical and thus interesting) state of the observed environment. Stream processing in this architecture is handled by incrementally updating model parameters [10], while particularly interesting samples can be detected and stored in a spatial database.

This focus on process over pattern is important for several reasons. Firstly, a model-centric framework improves interoperability of modelling spatially disparate datasets and simplifies data homogenization. For example, if an analyst is interested in understanding the relationship between temperature, humidity, and measured air pollution derived from different sensor networks, a spatially congruent set of data points is required. In this case, data interoperability is often achieved by querying the relevant databases and interpolating the values over a consistent spatial extent. In a model-centric framework, one gets this interoperability 'for free' by keeping up-to-date models of the various inputs. Similar advantages can be gained when dealing with multiple data streams attributed to spatial units of different types and geometries, and thus requiring polygon-to-point interpolation (see, e.g. [11]). Secondly, the decoupling of storage from modelling provides additional security and privacy by removing direct access to raw data. This means that potentially sensitive datasets may be analysed more freely/openly due to the 'anonymous' nature of models. An additional benefit of this type of analytical framework is its timeliness. Because inputs are analysed as they enter the system, results are continuously updated and therefore instantaneously relevant. The framework is also amenable to comparison, both spatially and temporally. Past models can be stored and retrieved when needed, providing the means to compare models through time, identify similarities in spatial dependencies, and provide quantitative decision support.

Conclusions

In this statement of interest, we have outlined how GIScience can take a leadership role in data-intensive social science research by focusing on its strengths rather than attempting to overcome the storage bottlenecks of legacy GIS systems. The focus here is on computational social science, as this is an area of research with a rich history of geographical theory that can be utilised to inform our models; however, a process-based framework may apply equally well in other GIScience related fields, including human-environment interactions, remote-sensing, and environmental monitoring. Indeed, by returning the focus to modelling and understanding underlying spatial processes, our framework for model-centric analysis provides a 'way forward' for GIScience research in general, that emphasises intelligent analysis, and provides models and tools to effectively explore, analyse, and understand dynamic spatial processes. This intelligent analysis goes a step further than the increasingly common practice of context-free data-driven knowledge discovery by taking advantage of contextual information and linked data to generate models that help us to represent and explain the real-world. Furthermore, a streaming GIScience helps to address concerns of scalability and relevance as we move towards a more data-intensive scientific paradigm. However, our focus on context, theory, and intelligence goes deeper than simply a better way to deal with large datasets; it may also help to address the long-standing concern of GIScientists that our lack of any unifying theories may reduce our field to second-class status in

[1] Antoine de Saint-Exupéry, Terre des Hommes (1939)

the academic community [3]. Changing the prevailing mindset, and focusing on modelling rather than storing, handling, and mining spatial databases will ultimately allow researchers to play to the current and future strengths of GIScience, and transform it into a leading discipline in the area of computational science.

References

[1] Cukier, K. (Feb 25 2010) Data, data everywhere. The Economist, Retreived from: http://www.economist.com/node/15557443

[2] Franklin, C. (1992). An introduction to geographic information systems: Linking maps to databases. Database, 15(2) p12.

[3] Goochild, M. F. (2010) Twenty years of progress: GIScience in 2012. Journal of Spatial Information Science, 1:3-20.

[4] Goodchild, M. (2006) GIScience Ten Years After Ground Truth. Transactions in GIS, 10(1): 687-692.

[5] Golledge, R. G. (2008) Behavioral Geography and the Theoretical/Quantitative Revolution. Geographical Analysis, 40: 239-257.

[6]Hill, K. (Feb 5 2011) Internal Google Emails Shed Light On Importance Of Mobile Location Information. Forbes, Retreived from: http://www.forbes.com/sites/kashmirhill/2011/05/02/internal-google-emails-shed-light-on-importance-of-mobile-location-information/

[7] Longly, P. (2000) The academic success of GIS in geography: Problems and prospects. Journal of Geographical Systems, 2:37-42.

[8] Nelson, T. A. (2012) Trends in Spatial Statistics. The Professional Geographer, 64(1): 1-12.

[9] OGRIP. Advisory Committee's First Year Report. Columbus, OH: Department of Administrative Services, State of Ohio, 1990.

[10] Pozdnoukhov A., Kaiser C. Scalable Local Regression for Spatial Analytics, Proc. Of the 19th ACM SIGSPATIAL GIS'2011, 2011.

[11] Pozdnoukhov A., Kaiser C. Area-to-point Kernel Regression on Streaming Data, Geostreaming workshop at 19th ACM SIGSPATIAL GIS'2011.

[12] Pozdnoukhov A., Kaiser C. Space-Time Dynamics of Topics in Streaming Text, Location Based Social Networks workshop at 19th ACM SIGSPATIAL GIS'2011.

Semantics and Ontologies For EarthCube

Gary Berg-Cross[1]*, Isabel Cruz[2], Mike Dean[3], Tim Finin[4], Mark Gahegan[5], Pascal Hitzler[6], Hook Hua[7], Krzysztof Janowicz[8], Naicong Li[9], Philip Murphy[9], Bryce Nordgren[10], Leo Obrst[11], Mark Schildhauer[12], Amit Sheth[6], Krishna Sinha[13], Anne Thessen[14], Nancy Wiegand[15], and Ilya Zaslavsky[16]

[1] SOCoP
[2] University of Illinois at Chicago
[3] Raytheon BBN Technologies
[4] University of Maryland, Baltimore County
[5] University of Auckland, New Zealand
[6] Wright State University
[7] NASA Jet Propulsion Laboratory
[8] University of California, Santa Barbara
[9] University of Redlands
[10] Rocky Mountain Research Station, USDA Forest Service
[11] MITRE
[12] NCEAS at University of California, Santa Barbara
[13] Virginia Tech
[14] Marine Biological Laboratory
[15] University of Wisconsin-Madison
[16] San Diego Supercomputer Center

Abstract. Semantic technologies and ontologies play an increasing role in scientific workflow systems and knowledge infrastructures. While ontologies are mostly used for the semantic annotation of metadata, semantic technologies enable searching metadata catalogs beyond simple keywords, with some early evidence of semantics used for data translation. However, the next generation of distributed and interdisciplinary knowledge infrastructures will require capabilities beyond simple subsumption reasoning over subclass relations. In this work, we report from the EarthCube Semantics Community by highlighting which role semantics and ontologies should play in the EarthCube knowledge infrastructure. We target the interested domain scientist and, thus, introduce the value proposition of semantic technologies in a non-technical language. Finally, we commit ourselves to some guiding principles for the successful implementation and application of semantic technologies and ontologies within EarthCube.

The semantic annotation of data and semantics-enabled search in metadata catalogs are part of many scientific workflow systems, e.g., Kepler [1]. In the past, semantic technologies have shown great potential in many biologically-focused cyberinfrastructures for both data annotation and semantic translation. There is some preliminary evidence to suggest that similar approaches would also add value in the geosciences, e.g., in the context of GEON. However, there appears to be some confusion about the role that

* Author names are listed in alphabetic order.

semantics can play within distributed next-generation knowledge infrastructures such as NSF's EarthCube[1]. Indeed, current Semantic technologies require knowledge of formal logic that is unfamiliar to most Earth scientists. There is, however, a simple way to understand how semantics can contribute greatly to the interoperability [2] of data, models, and services within EarthCube: simply put, by linking scientific observations and other data to terms drawn from ontologies or other forms of vocabularies, one can gain insights from how those terms are linked to other definitions in the ontology. This all happens *behind the scenes*. For example, if a scientist has collected observations of salinity measurements from the sea surface at location *X*, she can automatically link the data to terms like: chemical concentrations, oceanographic measurements, measurements (e.g., sea surface temperature) from 0m depth, and correlated measurements from locations situated near to *X* – all become accessible through the potential relationships revealed through ontologies. Thus, scientists searching for those general terms are more likely to find and potentially reuse the data. This capability will be invaluable to any scientist doing integrative or synthetic research that benefits from finding complementary data that others (e.g. potential collaborators) might have collected [3]. Even more, in an interdisciplinary setting the same terms may have different meanings and data may be collected and published following different measurement procedures and scientific workflows. Ontologies help to make such hidden heterogeneities explicit and, thus, support scientists in understanding whether a certain dataset fits their models [4]. Finally, to a certain degree, ontologies can also automatically translate data to make them interoperable and also reveal differences in the used classification systems [5].

If EarthCube promotes common vocabularies for annotating and describing data using terms drawn from ontologies, the value added will far exceed what can be expected from annotation using simple metadata, or worse, annotation using completely uncontrolled and not structured vocabularies. All the formal semantic processing and reasoning will be automatically accomplished behind the scenes for the scientists, in the same way that a Web browser nicely renders a page for a human to read. As a research community, we need to learn to be flexible, to develop techniques for *hardening* ontologies from looser semantics, to infer connections to more formal semantics, more generally to start with what is available whilst encouraging the development of more formal semantics where it is practical to do so. Google, Apple, the New York Times and Best Buy all use ontologies to support their content management systems or for other purposes related to sharing and managing of data. Thus, we believe that EarthCube should use semantic technologies as well. A key benefit of adopting Semantic technologies is that a vast number of repositories, ontologies, methods, standards, and tools that support scientists in publishing, sharing, and discovering data, is already available.

Semantic technologies provide new capabilities for formally and logically describing scientific facts and processes that may be as transformative as the introduction of the relational model was for organizing and accessing data over the past three decades. While a number of exciting semantic technology developments are underway, perhaps the area with greatest immediate applicability to EarthCube is the Semantic Web. The Semantic Web is a research field that studies how to foster the publishing, sharing, dis-

[1] See http://www.nsf.gov/geo/earthcube/ and the community page at http://earthcube.ning.com/ .

covery, reuse, and integration of data and services in heterogeneous, cross-domain, and large-scale infrastructures. It consists of two major components.

(i) Ontologies and knowledge representation languages that restrict the interpretation of domain vocabulary towards their intended meaning and, thus, allow us to conceptually specify scientific workflows, procedures, models, and data, i.e., the body of knowledge in a given domain, in a way that reduces the likelihood of misunderstanding and fosters retrieval and reuse [6].

(ii) As these ontologies are formal theories, they enable reasoning services on top of them. These reasoning services assist at different stages. They ensure that the developed ontologies are consistent. They help to make implicit knowledge explicit, discover incompatibilities and, thus, prevent users from combining data, models and tools that were developed with different underlying assumptions in mind. They allow querying across different sources and the semi-automatic alignment of different ontologies to foster the reuse and integration of data, models, and services. And finally, they support the design of smart user interfaces that go beyond simple keyword search and improve accuracy in search, cross-domain discovery, and other tasks which require data and information integration.

Linked Data is the data infrastructure of the Semantic Web [7]. It has rapidly grown over the last years and has found substantial uptake in industry and academia, since it significantly lowers the barrier for publishing, sharing, and reuse of data. Linked Data is an easily adoptable and ready-to-use paradigm that enables data integration and interoperation by opening up data silos. Combining Semantic Web technologies and Linked Data with ontologies also enables the discovery of new knowledge and the testing of scientific hypotheses. Consequently, the Semantic Web allows for vertical and horizontal integration, which is of central importance for EarthCube in order to realize the required interoperability of data, models, and tools while preserving the heterogeneity that drives the motor of interdisciplinary science.

However, the use of semantic technologies and ontologies in itself does not automatically guarantee interoperability or better access to data if not supported by a clear roadmap and guiding principles. The following list reflects a minimal set of principles that should guide the community for the next years. For EarthCube to be successful and transformative, we propose the following lines of action:

1. Be driven by concrete use cases and needs of the members of the EarthCube community. Collect, at the outset, a set of use cases from each EarthCube group, and conduct a substantial study of interconnected use cases which expose requirements related to data, models, and tools interoperability. These requirements need to be thoroughly analyzed as to the requirements they impose on the EarthCube data, ontology, and semantics infrastructure.
2. The choice of methods and the degree of knowledge formalization, e.g., lightweight versus heavyweight approaches, should be chosen based on use cases and application needs. This reduces the entry barrier for domain scientists to contribute data and ensures that a semantics-driven infrastructure is available for use in early stages of EarthCube.
3. Foster semantic interoperability without restricting the semantic heterogeneity introduced by the diverse community representing EarthCube. Provide methods that

enable users to flexibly load and combine different ontologies instead of hardwiring data to particular ontologies and, thus, hinder their flexible reusability.

4. Allow for bottom-up and top-down approaches to semantics to ensure a vertical integration from the observations-based data level up to the theory-driven formalization of key domain facts.
5. Involve domain experts in ontology engineering and enable them to become active participants by providing building blocks, strategies, documentations, and workshops on how to publish, retrieve, and integrate data, models, and workflows.
6. Apply semantics and ontologies to capture the body of knowledge in various Earth science domains for the purpose of organizing and accessing data, models and tools, learning about them, and extracting information from legacy data.
7. Exploit the power of classical and non-classical reasoning services to develop user interfaces, dialog systems and service chains that assist domain scientists at different stages ranging from discovering data and integrity constraint checking to the generation of new knowledge and hypothesis testing.

A detailed, more technical argumentation why these points need to be realized and how the heterogeneity of the geosciences requires new directions of research beyond schema standardization, can be found in the report of the Semantics and Ontology Technical Committee Report [8].

Acknowledgments

Krzysztof Janowicz acknowledges support by the National Science Foundation EarthCube Charette June 2012 travel grant. Pascal Hitzler acknowledges support by the National Science Foundation under award 1017225 "III: Small: TROn – Tractable Reasoning with Ontologies," and under a travel grant to the June 2012 EarthCube Charette. Anne Thessen acknowledges funding by NSF Data Net Program 0830976 and the National Science Foundation EarthCube Charette June 2012 travel grant.(c) 2012, by the respective authors, and The MITRE Corporation (for Leo Obrst). All Rights Reserved.

References

1. Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E.A., Tao, J., Zhao, Y.: Scientific workflow management and the kepler system. Concurrency and Computation: Practice and Experience **18**(10) (2006) 1039–1065
2. Obrst, L.: Ontologies for semantically interoperable systems. In: Proceedings of the 12th international conference on Information and knowledge management. CIKM ’03, ACM (2003) 366–369
3. Jones, M., Schildhauer, M., Reichman, O., Bowers, S.: The new bioinformatics: Integrating ecological data from the gene to the biosphere. Annual Review of Ecology, Evolution, and Systematics **37**(1) (2006) 519–544
4. Janowicz, K., Hitzler, P.: The Digital Earth as knowledge engine. Semantic Web Journal **3**(3) (2012) 213–221

5. Gahegan, M., Smart, W., Masoud-Ansari, S., Whitehead, B.: A semantic web map mediation service: interactive redesign and sharing of map legends. In: Proceedings of the 1st ACM SIGSPATIAL International Workshop on Spatial Semantics and Ontologies. SSO '11, ACM (2011) 1–8
6. Kuhn, W.: Semantic Engineering. In Navratil, G., ed.: Research Trends in Geographic Information Science. Lecture Notes in Geoinformation and Cartography, Springer (2009) 63–76
7. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data – The Story So Far. International Journal on Semantic Web and Information Systems **5**(3) (2009) 1–22
8. Hitzler, P., Janowicz, K., Berg-Cross, G., Obrst, L., Sheth, A., Finin, T., Cruz, I.: Semantic Aspects of EarthCube. Technical report, Semantics and Ontology Technical Committee. Available online at: http://knoesis.wright.edu/faculty/pascal/pub/EC-SO-TC-Report-V1.0.pdf (2012)

Geocoding Billions of Addresses:

Toward a Spatial Record Linkage System with Big Data

Sen Xu[1,2], Soren Flexner[1], Vitor Carvalho[1]

[1]Data Systems, Intelius, Inc., Bellevue, WA, U.S.A.
[2]Department of Geography, Pennsylvania State University, University Park, PA, U.S.A.
Email: senxu@psu.edu, sflexner@intelius.com, vitor@cs.cmu.edu

Abstract: Address is one of the most commonly used spatial data in everyday life. Comparing two addresses (e.g., if they are referring to the same location) is a fundamental problem for address-related record linkage. In this paper, a fast, reliable, expandable address parser/standardizer/geocoder has been developed as an initial step towards spatial record linkage. First, a CASS-based geocoding test set was created and performance of on-line geocoding API providers (Google, Yahoo, Bing) was evaluated. Considering high time consumption and geocoding precision flaws, we developed an in-house TIGER/Line based hierarchical geocoder, Intelius Address Parser (IAP) that provides on-par geocoding precision compared to on-line geocoding APIs. Given over one billion addresses, on a 25-node Hadoop cluster setup on with Amazon AWS, the time consumption and cost are reported and compared with commercial solutions. Strategies for using geocoded addresses for record linkage is presented and plans on expanding the use of geocoded result are discussed.

Keywords: address geocoding, spatial record linkage, big data, TIGER/Line, vanity city name

1. Introduction

Spatial record linkage describes the process of comparing two or more records based on the spatial footprints (e.g., addresses, spatial coordinates, political region, etc.) from each record. One of, if not the most commonly used spatial footprint is postal addresses, in forms of "123 Main ST, Unit A, Springfield, MA 01108" (U.S. addresses is the focus of this paper). *Parsing*, *standardizing*, and *geocoding* postal addresses for the purpose of spatial record linkage have attracted wide research attention in inter-disciplinary fields such as health science (Churches et al., 2002, Cayo & Talbot, 2003, Christen & Churches, 2005). The motivation of address geocoding in this paper comes directly from the demand to support people search, which aims at returning individuals by query such as name and location. From a people search perspective, many records (such as bill payments) were anchored to a person by a name and an address; given that there are many person sharing the same name (e.g., there are over 500 "Jim Smith" in the State of California[1]), using a combination of name and address to differentiate records is one of the fundamental techniques for reliable record linkage (Elmagarmid et al., 2007) using people data.

Geocoding describes the process of converting text-based postal address data into digital geographic coordinate (Boscoe, 2007), specifically latitude and longitude (lat/lon). The following challenges make geocoding a quintessential pre-requisite for spatial record linkage:

1. Addresses can be written in flexible ways. *500 108th Ave NE, Bellevue, WA 98004* maybe written as *500 108 Avenue NorthEast, Bellevue, WA*. The latter address has non-standard street post-directions (*NorthEast*), street name (*108*) and type (*Avenue*), and is missing zipcode (98004), yet human can still comprehend it is referring to the same location as the former standard address. Although USPS has a very specific standard on address format (United States Postal Service, 2012), due to the flexible nature of human language, it still very common to see non-standard addresses being

[1] Based on data from Intelius.com, as of June, 2012.

published and used in everyday life. Because of this flexibility, same address could be represented in completely different written forms; As a result, comparing two addresses requires more complicated procedure than string comparison.

2. Address standardization could be the solution to the above challenge. However, dictionary-based standardization is not sufficient; a complete street database is required to achieve reliable address standardization. Dictionary-based standardization is a common method in natural language processing for normalizing certain language flexibility. Street Type, for example, can be standardized based on a dictionary (e.g., *Avenue*, *Ave.*, *Av* should all be standardized as *Ave* when the string is recognized as a Street Type, according to the USPS address standard). This kind of irregularity, known as *Lexical Heterogeneity*, is a common problem in NLP and there is a variety of approaches to tackle it (see (Elmagarmid et al., 2007). However, address standardization requires spatial knowledge to be able to fix errors in the input. Take street type as an example, if *108TH ST* is not found for the region specified but *108TH PL* can be found, it's highly likely that the input is referring to the latter. Another common example is addresses that are missing the correct Pre-directional terms or Post-directional terms. Spatial databases are required in order for spatial record linkage system to be tolerant enough for these mistakes and make corrections based on potential possible matches to the real-world streets name.
3. Geocoding will allow more complex functions in spatial record linkage. Comparing two addresses and tell if they are referring to the same address is basic; more complex functions like "getting the distance between two addresses" (Tobler, 1970), "tell if they are located in the same zipcode/city/CBSA/State" (see Section 3.1), and "get the census statistics about the region of an address". Such derived potential "neighbour relationship", "located in the same zipcode/city/CBSA/State" and "located regions that share the same population/income level" can be very useful in conflating people data, which would require address data to be geocoded.

A variety of data source for address geocoding has been used for developing existing geocoding services, including **building centroids**, **parcel geometries**, **street segments**, and **centroids for USPS ZIP codes, cities, counties,** and **states** (Goldberg & Cockburn, 2010). Due to the fact that most fine-scale spatial databases (e.g., building centroids, parcel, and street segments) are incomplete, most geocoding services in place (Google Inc., 2012, Microsoft Corporation, 2012, Yahoo! Inc., 2012) use a combination of the data sources mentioned above. As a result, the precision of the geocoded results varies depends on which data source is available for the input address. Geocoding precision has been pointed out to be a very important component in interpreting the geocoded coordinates (Goldberg, 2011a). The *hierarchical geocoding strategy* of modern geocoders was adopted in developing our in-house address parser, standardizer and geocoder: Intelius Address Parser (IAP).

In this paper, the development of a CASS-based geocoding test set and evaluation of three commonly used geocoding API (Bing, Google and Yahoo!) is presented in Section 2. Upon realizing the high time consumption of using web API, the methods we used in developing a TIGER/Line based geocoder (IAP) aiming at geocoding Big Data was presented in Section 3. In Section 4, we present the process of setting up a Hadoop cluster on Amazon Web Service (AWS) and report the cost for geocoding over one billion addresses. Discussion and future development was presented in Section 5.

2. Evaluation of Existing Geocoders

Geocoding a few addresses can be easily achieved by calling one of the online geocoding APIs. Bing, Google, and Yahoo! all provide free and commercial geocoding services. The quality of the result from each geocoder, however, is self-defined representations without a unified standard. This means it is not straightforward to compare the geocoding results from multiple geocoders (Goldberg, 2011b). A summarization for the different geocoding precision/quality was presented in Section 2.1.

Another challenge is there has not been a publicly freely available geocoding test set at large scale. Because of the flexibility in address format, a large test set would be helpful for developing a geocoder that is tolerant enough for different kinds of irregularities. To this end, we adopted the Stage 1 test set for CASS (150,000 address pairs, in non-standard format and USPS standardized format) as our test set. Coding Accuracy Support System, known as CASS (United States Postal Service, 2008), is a system of tests for ensuring the quality of software that correct and matches postal addresses. We put the 150,000 non-standard format addresses to online geocoding APIs and record the geocoded result (returned lat/lon pairs and corresponding quality), and then compared with the geocoded result from IAP (see Figure 2).

2.1 Summary of geocoding API from Bing, Google and Yahoo

We chose three of the most popular geocoding web services, from Bing, Google and Yahoo! for this evaluation. Although they can all function as a general purpose geocoder, they vary at technical details from data source to their self-defined geocoding precisions. Table 1 provides an overview of the differences among the three online geocoding services. Note that there is a daily cap for the number of geocoding calls you can make within a day if you are using a free account (Bing) or from the same IP (Google and Yahoo!). In order to make the precision comparable, the different precision defined from each provider was summarized into three simple categories: **street level or better**, **region centroid (zip, city/state)**, or **unknown** (color coded as green, blue and red, refer to Figure 2). This generalized geocoding precision was used in comparing the three online geocoders versus IAP.

Table 1. Comparison of features of online geocoding APIs (February, 2012)

	Bing Geocode Service	Google Geocoding API	Yahoo! PlaceFinder
Example API call	http://dev.virtualearth.net/REST/v1/Locations/US/WA/98004/bellevue/500+108th+ave+ne?key=[APIkey]	http://maps.googleapis.com/maps/api/geocode/xml?address=500+108th+ave+ne,+bellevue+wa+98004&sensor=true	http://where.yahooapis.com/geocode?q=500+108th+ave+ne,+bellevue,+wa+98004
Data Source	TeleAtlas, Navteq, Map Data Sciences (Pendleton, 2008)	Internal street networks (Lookingbill, 2009)	Navteq (Yahoo! Inc., 2012)
Daily Cap	30,000 per API Key	2,500 per IP	50,000 per IP
Commercial license cost	$8000 per 1,000,000 transaction (Hansen, 2009)	$10,000 per year (Google, 2012); daily API call limit raise to 100,000	Commercial service not yet provided as of August 2012.
Key required	Yes	Optional	Optional
Return format	JSON/XML	JSON/XML	XML/JSON/Serialized PHP/
Geocoding Precision	Parcel, Interpolation, Rooftop, InterpolationOffset Null	Rooftop, Geometric_Center, Range_Interpolated, Approximate, Zero_Results	99, 90, 87, 86, 85, 84, 82, 80, 75, 74, 72, 71, 70, 64, 63, 62, 60, 59, 50, 49, 40, 39, 30, 29, 20, 19, 10, 9, 0

Addition functions	Reverse Geocoding Return bounding box Return Type	Reverse Geocoding Return Type (e.g., locality, political) Viewport Biasing, Region Biasing,	Reverse Geocoding Return time zone and telephone area code Return in other language (French) Allow POI name as input

2.2 Lesson learned from the evaluation

The above geocoding APIs have one huge caveat when it comes to Big Data processing: the time consumption is too high due to that the request and reply were sent through the internet. In other words, the communication overhead of calling web API is high. If the average duration for making geocoding calls and receiving the geocoded results is 0.1 sec, geocoding one billion addresses will take over 1,000 days on one machine. With the daily cap (for free public API usages), the time consumption will be much higher.

More importantly, we also found the existing geocoding APIs are far from perfect in dealing with messy addresses (some of the non-standard addresses can be very tricky to parse and standardize). Because before geocoding, there needs to be *address parsing* and *address standardization*, which is quite prone to errors due to flexible address format. The following types of addresses have been found frequently causing the online geocoders to fail:

1. Non-residential addresses, including PO Box (e.g., *PO Box 123, Springfield, OR 97477*), Rural Route Boxes (e.g., RR 2, Box 12, Springfield, IL), General Delivery and Military mail (e.g., John Jackson, Unit 123, Box 456, APO AE 09001). Non-residential addresses represent only a Postal Box for receiving mail (such as at a postal office), so it is important to understand that the geocoding precision for non-residential type addresses is limited (e.g., PO Box addresses can only be best geocoded to zipcode level). Specifically, they are missing street and house number, which a lot of the times are the reason for the incorrectly parsing. Take PO Box type address as an example, the PO Box number gets dropped (Google) or mistaken for House Number or POI Name (Yahoo, Bing). The online geocoders do not differentiate non-residential addresses from residential addresses, which is the main cause for the observed error. This problem is addressed in IAP.
2. Non-English addresses. Although we are focusing on U.S. addresses, there still exists street names that are in languages other than English (majority of these cases in the US is in Spanish). The challenge lies in that the sequence of Pre-directional term, Street Name, Street Type, Post-directional term would be different due to different language patterns. *Washington Ave* (i.e., *Washington Avenue*) in English would become *Av Washington* in Spanish (i.e., *Avenida Washington*, real street name in San Juan, Puerto Rico). This reverse sequence would require a different parsing mechanism; in IAP, we use a different set of RegExs for when the StreetType is recognized as non-English.
3. General parsing errors: Yahoo!'s geocoding service, at quality below 84 (Yahoo!'s quality representation, equivalent to region-centroid level precision), contains a lot of cases where the lat/lon returned was outside of U.S. Many PR (Puerto Rico) addresses was geocoded to Europe (could be the error from the language issue). Some well formatted address are also geocoded wrong[2].

Given the above concern on speed and precision, we concluded that developing an in-house address geocoder is required for processing Big Data.

[2] The errors are found as of Jun 18th, 2012 from Yahoo!'s PlaceFinder API (which was released in June 2010). Examples:
http://where.yahooapis.com/geocode?q=91+ROSELAND+AVE+#+601,+CALDWELL+NJ+07006: one valid address in New Jersey results in lat/lon pairs in Russia from the PlaceFinder API.

3. Intelius Address Parser

The address data in Intelius is at a very large data scale and varies in data quality. Intelius Address Parser (IAP) is designed to be *highly tolerant of noisy data*, *fast*, can *deliver reliable geocoding precision*, and *customizable* for complex spatial record linkage demands. IAP is designed based on an open-source project JGeocoder (JGeocoder, 2008), which was further developed and customized based on the different cases identified from the above evaluation. The design of IAP is presented in Section 3.1 and evaluation of IAP's performance compared to the three on-line geocoding API is presented in Section 3.2.

3.1 IAP Design

IAP segments the task of geocoding into three consecutive steps: *parsing*, *standardization* and *geocoding*. Each step produces an intermediate output, which was used for the next step. IAP uses the TIGER/Line database (The US Census Bureau, 2011) and a zipcode/city name database (federalgovernmentzipcodes.us, 2012). The workflow is shown in Figure 1:

Parsing: The input full-address string is segmented into *Address Components* using a library of Regular Expressions and rule-sets in the parsing stage. *Address Components* consist of *POIName*, *HouseNumber*, *PreDir*, *StreetName*, *StreetType*, *PostDir*, *Line2*, *City*, *State*, and *Zip* (JGeocoder, 2008). When the input address is recognized as non-residential type addresses (PO Box, Rurul Route, General Delivery, and Military), corresponding tags are put in the geocoding precision field for further processing.

Standardization: Each parsed *Address Component* goes through a standardization process where a rule set developed following USPS postal address standard are applied (United States Postal Service, 2012). For example, StreetType *Avenue* will be changed into *AVE*, PostDir *NorthWest* will be changed to *NW*. When zipcode is present, the city and state name will be fixed for spelling mistakes using Levenshtein distance to match to the zip-city name database (e.g., change *Bellevua, WA* into *Bellevue, WA*). Moreover, a vanity city name (referring to the alternative of a city name) database is used to fix non-standard city names. For example, *City Hollywood* will be changed to *Los Angeles*; *City Manhattan* will be changed to *New York City*.

Geocoding: When recognized as residential address, the standardized *Address Components* will be used for construct SQL query for TIGER/Line address range database (refer to (Goldberg & Cockburn, 2010) for details on using address range for geocoding). Here, *Zip*, *StreetType*, *PreDir* and *PostDir* will be fixed if the standardized result contains errors or missing components from the input. If the address range can be found for the given street name, find the closest address range for the input *HouseNumber* and return imputed lat/lon pairs and put Street-level in *Precision*. Otherwise, return lat/lon of region centroid for zipcode or city. Additional infomation (*County* and *CBSA*) are also returned.

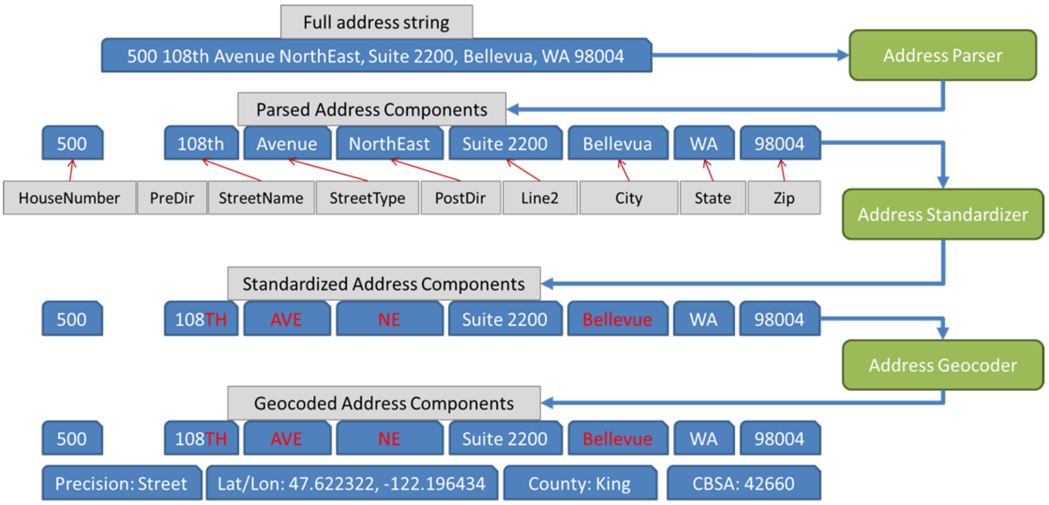

Figure 1. Workflow of the Intelius Address Parser (IAP). Red represents modification from original input.

A few design features from IAP are explained in detail below:

1. IAP differentiates non-residential address and residential address at the parsing stage. Non-residential address will not make call to TIGER/Line database (because they miss <u>*HouseNumber*</u> and <u>*Street*</u>), which saves time. Also, <u>*Precision*</u> field is used to deliver address type information. Non-residential address types will be reflected in the precision field. This feature is particularly useful in address comparison.
2. IAP allows flexible input. The input can be a simple string, which means addresses that are organized differently from different data source can be parsed and standardized into the same format. Also, most *Address Components* can be missing yet IAP can still deliver geocoded result. An address can be missing <u>*HouseNumber*</u>, *PreDir*, *PostDir*, <u>*Zip*</u> and still be parsed, standardized, and geocoded correctly. Even if the input consists only of city name and state, IAP can still deliver lat/lon of on the city centroid, thanks to the *hierarchical geocoding strategy*.
3. IAP delivers additional information other than lat/lon for addresses. County and CBSA (Core Based Statistical Areas, obtained from (United States Office of Management and Budget, 2009) are regions that inexplicitly associated with an address. They are larger than Zip and smaller than a State, which can be useful for spatial record linkage.

3.2 Performance of IAP against online geocoding APIs

We put IAP to the CASS Stage 1 test and compared the performance with online geocoding APIs. Because online geocoding APIs does not return non-residential address types, this comparison only reflects coarse granularity of returned lat/lon pairs. As shown in Figure 2, IAP provides on-par precision compared to the three online geocoding APIs.

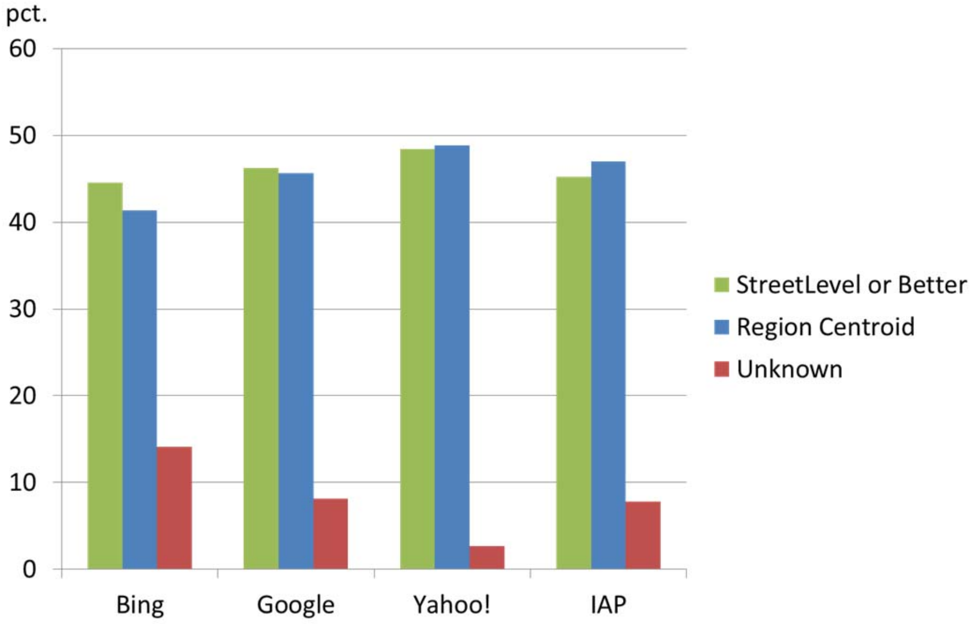

Figure 2. Summarized geocoding performance from Bing, Google, Yahoo and IAP on the CASS Stage 1 test set (150,000 addresses)

However, of the CASS test set, which consists only of US addresses (except for certain overseas military addresses), Yahoo!'s PlaceFinder returns more than 7,800 lat/lon that are located outside of U.S., presenting a potential bug in parsing. Since this evaluation relies on the reported geocoding precision, it is difficult to judge how precise the geocoded lat/lon pairs actually are. A golden set of geocoded addresses and corresponding lat/lon pairs are required for further evaluation, which is discussed in Section 5.

4. Geocoding Billions of Addresses

To experiment geocoding Big Data with IAP, we leverage some of the current Big Data processing tools and report the procedure of how over one billion addresses was geocoded. Hadoop, a free implementation of MapReduce (Dean & Ghemawat, 2004) has become the *de facto* industrial standard framework for Big Data processing. Amazon Web Service (AWS) is also becoming more and more popular as a *quick*, *inexpensive*, and *disposable* way of setting up a computing cluster. We deploy a 25-node Hadoop cluster on AWS for this experiment.

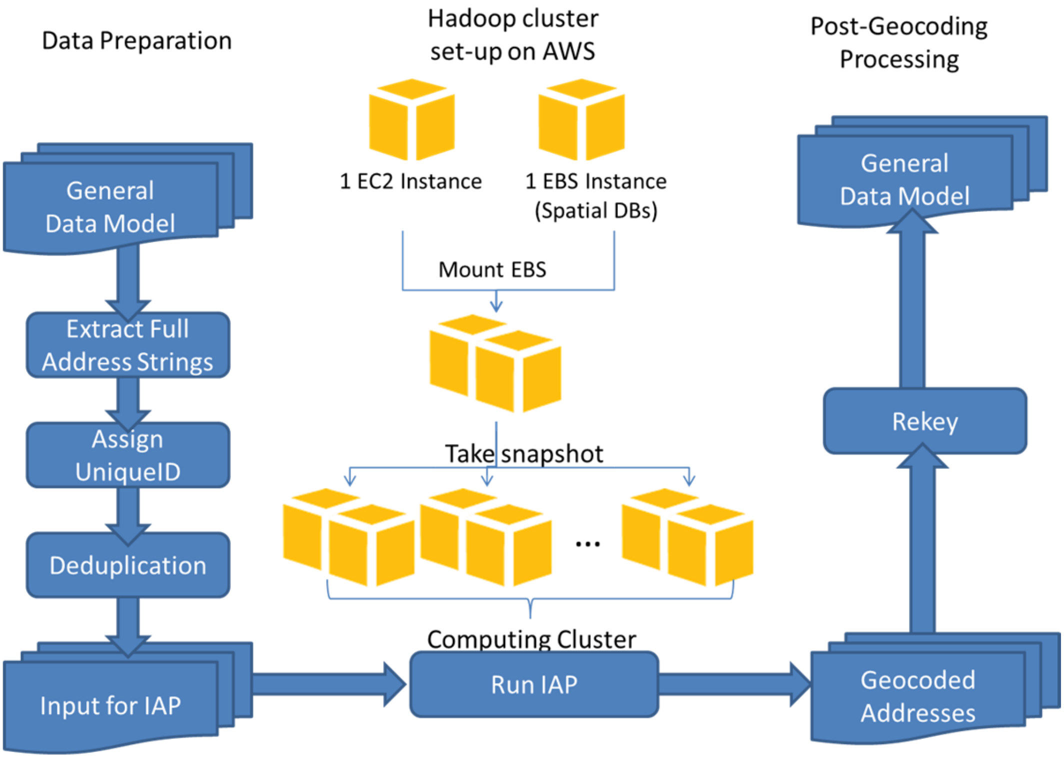

Figure 3. Workflow for running IAP on AWS.

Setting-up AWS for this IAP experiment requires two types of instances, EC2 (Elastic Computing Cloud) for high capacity computing CPUs and EBS (Elastic Block Storage) for storing the required spatial databases (TIGER/Line and zipcode database, approximately 17GB as MySQL databases). Java and MySQL comes pre-configured in AMI (Amazon Machine Image) for EC2; once the spatial DBs gets uploaded to 1 EBS instance, the EBS instance gets mounted to 1 EC2 instance; this EC2+EBS image are then snapshotted then copied into as many instances as needed. For this procedure, we setup a 25 nodes High-CPU Extra Large EC2 cluster.

Because address data is usually part of a general data model, to reduce cost of data transfer, we first prepare the input data for geocoding by extracting addresses as strings and assign unique ID to each record. Then we use string-based string comparison to deduplicate apparent redundant addresses. After this we have over one billion addresses ready for geocode. Running IAP over one billion addresses on a 25 nodes High-CPU Extra Large cluster takes about 38 hours. The cost of this geocoding process (High-CPU Extra Large instances cost \$0.66/hour/instance) is \$627; considering the time consumption of setup and data transfer, the overall cost should be under \$800[3], which would be drastically less than the cost of using any of the commercial services (see Table 1).

5. Discussion and Future Work

Admittedly, the evaluation of the four geocoders is limited. Because CASS test set only provides addresses (non-standard and standard), we can only rely on the precision (or quality) field provided by the geocoder to tell if the result is good or bad (because we do not have the true and precise lat/lon for each of the 150,000 addresses). However, based on hand-

[3] All cost estimates in this paper were obtained in May 2012.

examination of geocoding result from Yahoo!, we found the returned lat/lon to be not as precise as the precision field indicates. This calls for a golden set consisting of addresses and their corresponding verified lat/lon for evaluating geocoders. Crowdsourcing method could be leveraged to build such a golden set, where human participants manually check the lat/lon of an address and popular the golden set with the best precision.

Another drawback of the evaluation is that due to the mismatch of the precisions used in different geocoders, the summarized categories can only be general and coarse in granularity. It is possible that on-line geocoders like Google may offer finer-than-street-level precision (e.g., building centroid) on certain addresses, which are not credit for in this evaluation. Different kinds of spatial databases may complement each other, which urge IAP to include other types of spatial databases to offer finer and more comprehensive geocoding.

The limitations of making web API call (communication overhead, daily cap) compared to local database lookup shows its significance when it comes to Big Data processing. The time consumption for getting geocoded result for the 150,000 addresses ranges from four days (Yahoo! PlaceFinder) to two months (Google Geocoding API). In the case of Google, even with a $10k per year commercial API (daily limit raised to 100,000), geocoding one billion unique addresses will take 10,000 days (more than 27 years, costing over $270k).

From developing IAP through examining messy addresses, we learned that high tolerance for address parsing is crucial to achieve reliable geocoding result. Non-residential type addresses and addresses with missing components all needs to be considered to achieve a robust geocoding service.

Geocoders may not be limited to provide lat/lon pairs and precision; to better serve as a foundation for spatial record linkage, providing additional spatial data, such as county and CBSA, can be helpful for making address comparisons. U.S. Census data for different regions can be used to derive further linkages between two locations. For example, whether two addresses comes from county with similar population level and income. This additional information would particularly be useful for linking people data, which will be developed for the next version of IAP.

Finally, we recognize the benefit of using AWS for Big Data computing. First, building a high-capacity computing cluster using AWS is fast, easy and cheap. What used to take months on buying hardware, configuring OS and software now takes minutes on AWS. The cost for geocoding billions of addresses is much lower comparing to other commercial offerings. Another advantage for using AWS is data security. Sending data to commercial service providers increases the risk of data exposure, which would be a big concern for sensitive data. Computing on AWS, however, uses disposable instances where after computing, the instances can be destroyed. On the other hand, because AWS is charged by use time, the instances should be destroyed to save cost. For example, the AWS cluster used for this geocoding experiment was only alive for less than a week. This disposable setup reduces data exposure in the cloud to the minimal.

In sum, our evaluation of 150,000 addresses shows current on-line geocoding API is still lacking with regard to speed and precision. With an in-house geocoder IAP, we experimented on geocoding over one billion addresses, which take 38 hours and cost less than $800 on a 25 nodes high-CPU cluster on AWS.

Acknowledgements

This study is conduct during Sen Xu’s internship at Intelius, Inc. The test dataset was obtained from Intelius, Inc.

References

Boscoe, F. P. . (2007). *Geocoding Health Data*, chapter 5. The Science and Art of Geocoding-Tips for Improving Match Rates and Handling Unmatched Cases in Analysis, (pp. 95–109). CRC Press.

Cayo, M. & Talbot, T. (2003). Positional erro in automated geocoding of residential addresses. *International Journal of Health Geographics*, 2(10).

Christen, P. & Churches, T. (2005). A probabilistic deduplication, record linkage and geocoding system. In *Proceedings of the Australian Research Council Health Data Mining Workshop, Canberra, Australia*.

Churches, T., Christen, P., Lim, K., & Zhu, J. (2002). Preparation of name and address data for record linkage using Hidden Markov Models. *Medical Informatics and Decision Making*, 2(9).

Dean, J. & Ghemawat, S. (2004). Mapreduce: simplified data processing on large clusters. In *Proceedings of the 6th conference on Symposium on Opearting Systems Design & Implementation - Volume 6*, OSDI'04 (pp. 10–10). Berkeley, CA, USA: USENIX Association.

Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. volume 19 (pp. 1–16). Piscataway, NJ, USA: IEEE Educational Activities Department.

federalgovernmentzipcodes.us (2012). Free zipcode database. http://federalgovernmentzipcodes.us/.

Goldberg, D. W. (2011a). Advances in geocoding research and practice. *Transactions in GIS*, 15(6), 727–733.

Goldberg, D. W. (2011b). Improving geocoding match rates with spatially-varying block metrics. *Transactions in GIS*, 15(6), 829–850.

Goldberg, D. W. & Cockburn, M. G. (2010). Improving geocode accuracy with candidate selection criteria. *Transactions in GIS*, 14, 149–176.

Google, I. (2012). Google maps api for business documentation. http://www.google.com/enterprise/earthmaps/maps-faq.html.

Google Inc. (2012). The Google Geocoding API. https://developers.google.com/maps/documentation/geocoding/.

Hansen, R. (2009). Price Check on Bing Maps. http://rexdotnet.blogspot.com/2009/09/price-check-on-bing-maps.html.

JGeocoder (2008). http://jgeocoder.sourceforge.net/index.html.

Lookingbill, A. (2009). Google latlong: Your world, your map. http://google-latlong.blogspot.com/2009/10/your-world-your-map.html.

Microsoft Corporation (2012). Bing Maps Geocode Service. http://msdn.microsoft.com/en-us/library/cc966793.aspx.

Pendleton, C. (2008). New feature release of live search maps. http://blogs.msdn.com/b/virtualearth/archive/2008/09/24/new-feature-release-of-live-search-maps.aspx.

The US Census Bureau (2011). TIGER/Line databases. http://www.census.gov/geo/www/tiger/.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic Geography*, 46, 234–240.

United States Office of Management and Budget (2009). Table of united states core based statistical areas. http://www.census.gov/population/metro/files/lists/2009/List1.txt.

United States Postal Service (2008). CASS Certification Requirements - A Mailer's Guide. https://ribbs.usps.gov/cassmass/documents/tech_guides/CASS%20Cert%20Req%20MAILERS%20Guide.pdf.

United States Postal Service (2012). Publication 28 - Postal Addressing Standards. http://pe.usps.gov/text/pub28/welcome.htm.

Yahoo! Inc. (2012). Yahoo! PlaceFinder. http://developer.yahoo.com/geo/placefinder/.

Trending on Foursquare:

Examining the Location and Categories of Venues that Trend in Three Cities

Jessica Benner and Cristina Robles

{jgb14, cmr93}@pitt.edu

Abstract. This study examines categorical and spatial patterns for trending venues in three cities. The three cities in our experiment are New York City, NY, Pittsburgh, PA and Erie, PA. We examine common categories of venues that trend in our test cities and see a similar theme in each city favoring food-based venues, nightlife venues and other venues. We map the venues in each city and perform a nearest neighbor analysis to identify any spatial clustering in the datasets. We find that trending venues in New York City and Pittsburgh exhibit clustering while the trending venues in Erie are more dispersed.

1 Introduction

This study is an extension of a previous work done on Foursquare and trending (Robles and Benner, 2012) that focused on comparing the temporal trends in three distinct cities. The previous work briefly examined categories and trending, but only in terms of what categories of venues trended the most often. This work focuses on the categories and spatial pattern of trending venues in three cities. The broader goal of our work is to begin to understand the trending feature on Foursquare in general and its use.

Foursquare is a popular location-based social network (LBSN) boasting 20 million members and nearly one million businesses using the service plus the collection of 2.5 billion check-ins since its release in 2009 (Foursquare, 2012a). Some researchers try to discern patterns of use in Foursquare from the individual user's perspective to make social link predictions (Scellato et al. 2011) or understand the location-sharing patterns of users (Cramer et al. 2011). On the contrary, the trending feature on Foursquare offers global insights into the most popular venues in a city and when they are most visited. This kind of information is useful to both researchers interested in urban analytics and Foursquare venue owners. For example, an urban analyst could create an application to collect real-time trending data from Foursquare and share visualizations of currently trending venues in a city and other interpretations of the data. Then, a Foursquare venue owner in the same city could use this resource to monitor their own venue and similar venues with the aim to improve their business.

In order for this kind of collection, sharing and use of trending data to be realized methods for retrieving, filtering, storing, integrating, and sharing large datasets of check-in data are needed. A first step in this process is to gain a basic understanding of the phenomenon of trending and the structure of trending data. The aim of this paper is to investigate the spatial distribution of trending venues in general and specific categories of venues.

2 Background and Related Work

2.1 Location Based Social Networks

Location Based Social Networks, or LBSNs, are social networks similar to Facebook or Twitter, but with a location component. The LBSN we chose for our study is Foursquare. The main purpose of Foursquare is to share where you are in the form of 'check-ins' via a free downloaded mobile phone application available from Foursquare. Users are able to 'check-in' to any venue currently listed on Foursquare, or they can make their own.

2.2 Trending

In addition to 'checking in' at venues, and seeing the locations of friends, users are able to explore Foursquare to see which venues around their location are 'trending now.' Venues that are 'trending now' are venues near the user that currently have several other people checked in at them. To date, little is known about this new feature yet 'trending' may provide a filter to extract the most important venues in a city at certain times from large check-in datasets.

2.3 Related Work

In our previous paper (Robles and Benner, 2012), we show commonly trending venues in all three cities are food-based venues, nightlife venues, and venues not well represented by the common categories used to study Foursquare (referred to as 'other'). Examples of 'other' venues are doctor's offices, work offices, hotels and hospitals. Overall, food based venues dominated in New York City while nightlife venues (i.e., bars) dominated both Pittsburgh and Erie.

Lindqvist et al. (2011) conduct a survey-based study with users of Foursquare to discover motivations for checking-in on Foursquare. Participants of the study share a variety of reasons for deciding to check-in or not including: letting others know where you are, finding out where others are, game playing to collect badges or points, becoming the 'mayor' of a venue for all friends to see, and presenting a 'self' to others through the locations they choose to check-in. Understanding why people choose to check-in or not has value for attaching meaning to venues on Foursquare and determining why certain categories of venues trend. For example, does the meaning of a venue to certain groups of users impact the number of check-ins and its likelihood of trending?

Finally, a new project from Carnegie Mellon University called the Livehoods project (Cranshaw, 2012) uses machine learning clustering techniques to re-define city neighborhoods by pattern of use on Foursquare as opposed to the traditional neighborhood that are politically defined. Livehoods clusters represent groups of check-ins by similar users. The Livehoods project offers a new expression of check-in data and combines the characteristics of users to find undiscovered pockets of activity in cities.

3 Experimental Design

This study obtained data for venues that are 'trending now' using the publicly available Foursquare API. Specifically, we retrieved data from the Trending Venues endpoint and when necessary the Venues endpoint over 25 consecutive days. The Trending Venues endpoint gave us a list of the venues that were trending during each hour we sent the request coupled with the category of the venue and the time stamp. We then plotted the trending venues on a map using their latitude and longitude and symbolized the venues using the number of times trended (sum over 25 days) and the category. Table 1 shows the three different city 'types,' extra large, mid-sized, or small, we selected for the same region in the USA. We also took into account the total population of the city when assigning the city 'type' city populations as reported in the 2010 US Census, are shown in Table 1 below.

Table 1. City description with population for cities in study

City Name, State	City Type	Population
New York, NY	Extra Large	8,175,133
Pittsburgh, PA	Mid-Sized	305,704
Erie, PA	Small	101,786

After choosing which cities we would use for our experiment, we then had to choose which latitude and longitude coordinates to give the API. The Trending Venues endpoint of Foursquare API needs an exact set of latitude and longitude coordinates to know where to get the list of trending venues. To choose a standardized set for each city, we decided to use the latitude and longitude coordinates from each city's Wikipedia page. This is different from a user getting the list of trending venues from their phone because their physical location is what provides the latitude and longitude coordinates to Foursquare. After collecting trending venues hourly for the 25 day period, we had over 10,000 trending events for the three cities. Table 2 shows the number of trending events (i.e., records in the database) and the unique venues for each city. Finally, we used the spatial software, ArcGIS (version 9.2), to conduct a nearest neighbor analysis in each city. The nearest neighbor calculation assumes the data is randomly distributed and results in a nearest neighbor ratio value and a corresponding z value for each set of features.

Table 2. Data collection

City	Trending Events	Unique Venues
New York City	9,447	843
Pittsburgh	1,178	149
Erie	275	32

4 Results

First we calculated a set of summary statistics for the number of times unique venues in each city trended over the 25 days of our study. Table 3 shows the summary including the average number of times a venue trended, the percent above and below the average value and the standard deviation. New York City has the highest average value, 11.21 times trending and Pittsburgh has the lowest average value, 7.91. This is expected for New York since it is the largest of the three test cities in terms of people and the number of venues. All cities show a similar pattern for the percent of data above and below the average in which most venues trended below the average number of times. Finally, the measures of spread for each city show that the distribution for New York City is more dispersed than the other two cities while the distribution of trending in Erie forms a curve that is closer to the average.

Table 3. Summary Statistics for Times Trending in three cities

Summary Statistics, timesTrending				
	Avg timesT	% Above Avg	% Below Avg	SD
NYC	11.21	0.21	0.79	23.06
PGH	7.91	0.26	0.74	15.26
ERIE	8.59	0.34	0.66	9.01

Next, we mapped the trending data for each city and symbolized the data using shape, color and size. The shape and color of an icon in the map depicts the category of a venue while the size of the icon corresponds to the number of times a unique venue trended during our study period and the larger the icon, the more times a venue trended. Figure 1 illustrates the data collected for New York City, Figure 2 Pittsburgh and Figure 3, Erie.

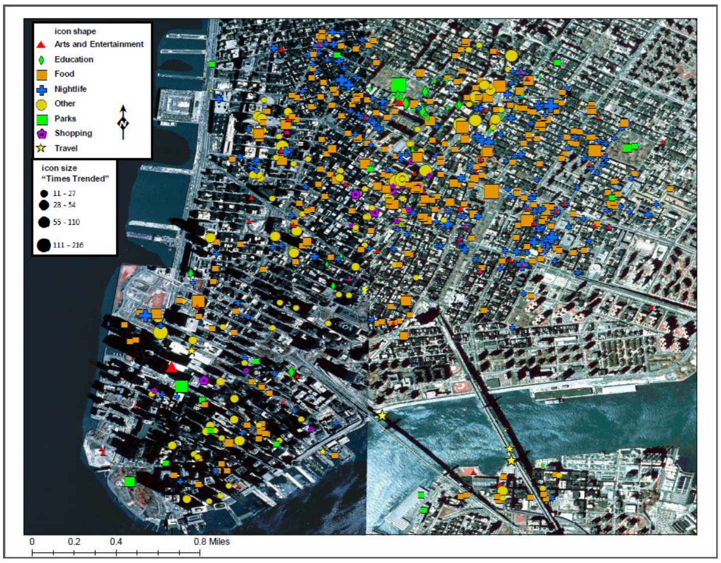

Fig. 1. Trending venues with categories in New York City, NY

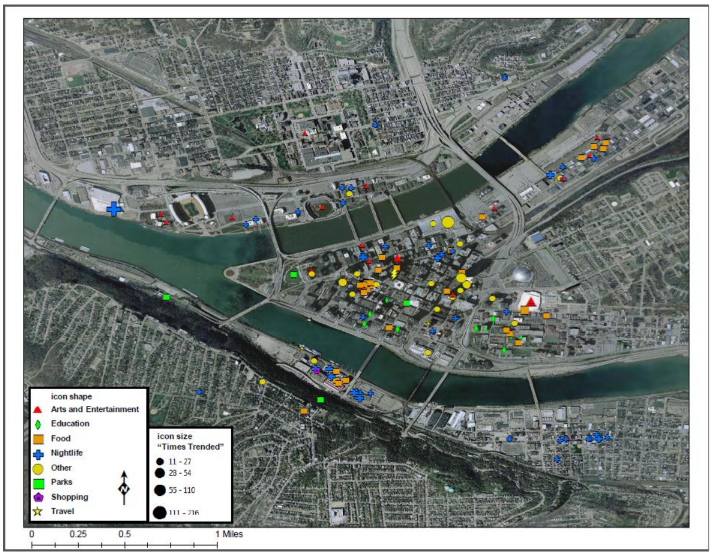

Fig. 2. Trending venues with categories in Pittsburgh, PA

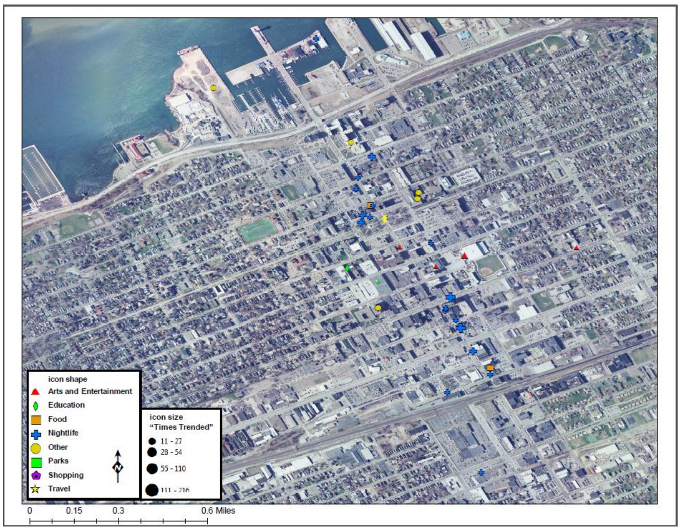

Fig. 3. Trending venues with categories in Erie, PA

Finally, a nearest neighbor analysis was performed using the three datasets. The results are presented in Table 4. The observed mean distance is the average distance between the points in the dataset. The expected mean distance is an expected value for the average distance between points in a dataset based on a random distribution. We observe from Table 4 that the observed and expected mean distances for both New York City and Erie are fairly close to one another but the observed mean distance for Pittsburgh is much lower than the expected distance.

Table 4. Nearest Neighbor Analysis for Three Cities

City	Observed Mean Dist (m)	Expected Mean Dist (m)	NN Ratio	Z Score
NYC	35.6	57.9	0.615	-21.25
PGH	76.4	132.8	0.575	-9.925
ERIE	122.1	128.8	0.948	-0.57

The nearest neighbor ratio is the ratio of observed mean distance divided by expected mean distance. This ratio is interpreted by values lower than 1 considered as clustered datasets and values above 1 considered as dispersed datasets. The nearest neighbor ratios for the three cities all fall below 1 with Pittsburgh having the most clustered result, 0.575, and Erie a more dispersed value, 0.948, fairly close to 1. A Z-score is a score of statistical significance indicated in measures of standard deviation. For this measurement, we begin with the null hypothesis that the datasets are randomly distributed through space. We select a 95% confidence interval in which we would reject the null hypothesis if the z-score lies between -1.96 and 1.96. Given the results for both New York City and Pittsburgh we reject the null hypothesis in favor of the presence of an underlying pattern. The high Z-scores and the nearest neighbor ratio results indicate that the New York City and Pittsburgh datasets are clustered not randomly dispersed. For Erie, we cannot reject the null hypothesis due to the value of z = -0.57 which falls within our 95 % confidence interval. As a result, we cannot say the Erie dataset displays clustering.

5 Discussion

Our analysis shows that most unique venues trend less than the average number of times trending for the three test cities over 25 days. The trending data in Erie shows a tighter fit around the average value and a more even distribution above and below the average values than the other cities. From the map, we observe the majority of the trending venues in Erie are on one specific street. This indicates a concentration of the trending venues in Erie; however, we were unable to confirm the presence of clustering in our nearest neighborhood analysis. Pittsburgh displayed trending patterns that were fairly close to those of the other three cities. The map of the Pittsburgh venues showed a lot of different places trending in the heart of the city and smaller areas trending outside of the city. In the nearest neighbor analysis, Pittsburgh showed the highest degree of clustering among the three cities. We believe the presence of the bridges in Pittsburgh also make this city unique in terms of the spatial patterns of trending and contribute to the higher degree of clustering we detected. In a large city like New York, there are trending venues everywhere in a variety of categories. This variety of venues explains why New York City has the highest standard deviation from the average times unique venues trend. We found that New York City displayed clustering of the trending venues and is the most densely packed city with shorter distances between neighboring venues. We can now add to the findings of our previous study that trending venues are located in a concentrated area in Erie, PA, but in New York City they are located everywhere, with Pittsburgh, PA showing a little bit of both of these patterns.

6 Limitations

Although our experimental setup was simple, we have a few limitations. To begin, our conclusions were dependent on the latitude and longitude coordinates given to the Foursquare API. Different coordinates could have yielded different results regarding the trending venues we received. For example, latitude and longitude coordinates in the middle of downtown are not representative of the entire city and since most cities have a lot of possible latitude and longitude pairs, the pairs we chose for each city may not be in similar areas (e.g., downtown, shopping district).

Secondly, our study period of 25 days is not long enough to get a complete picture of the categorical and spatial patterns of trending events in the three cities. Although we establish preliminary patterns, a period of time covering more holidays and weekends would provide a more complete picture. Finally, in order to confirm our results for Erie, we believe that a larger dataset is required since the number of unique venues in Erie during our study period was 32.

7 Conclusions and Future Work

In conclusion, this study begins to help us understand the categorical and spatial patterns of trending on Foursquare. The purpose of which is to support our long term goal to understand why certain venues trend. A solid understanding of the trending phenomenon on Foursquare can help researchers in urban analytics understand the meaning of trending and make use of the data in creative ways and help venue owners maximize their business's exposure on Foursquare. We selected three cities of varying sizes to obtain the strongest case for our analysis, New York City, NY, and Pittsburgh and Erie PA. We find that for all three cities, the majority of venues trend less than the average number of times trended for the city as a whole and that both New York City and Pittsburgh exhibit clustering of unique trending venues over a 25 day study. This paper representes a work in progress thus, more studies are necessary to confirm the spatial patterns we find and to complete our understanding of the trending phenomenon.

8 References

1. AboutFoursquare Blog, "Foursquare's "Trending Now" makes popular venues easier to find", http://aboutfoursquare.com/foursquares-trending-now-makes-popular-venues-easier-to-find/
2. BitsBot Labs, Foursquare Trends Blog, Available at: http://bitsybot.com/foursquare/trends.html
3. Cramer, H.; Rost, M.; and Holmquist, L. E. 2011. Performing a check-in: emerging practices, norms and 'conflicts' in location sharing using foursquare. In MobileHCI '11. ACM.
4. Cranshaw, J., Schwartz, R., Hong, J., and Sadeh, N., "The livehoods project: Utilizing social media to understand the dynamics of a city" Association for the Advancement of Artificial Intelligence, 2012.
5. Foursquare. Available at: http://www.foursquare.com
6. Foursquare (2012a) About Foursquare Page, https://foursquare.com/about/
7. Lindqvist J., Cranshaw, J., Wiese, J., Hong, J., and Zimmerman, J., "I'm the mayor of my house: examining why people use foursquare - a social-driven location sharing application" ACM CHI Vancouver, BC, Canada, 2011.
8. Robles, C. and Benner, J., "A Tale of Three Cities: Looking at the Trending Feature on Foursquare" IEEE SocialCom Amsterdam, The Netherlands, 2012.
9. Scellato, S. Noulas, A. and Mascolo, C. "Exploiting place features in link prediction on location-based social networks," in Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '11. New York, NY, USA: ACM, 2011, pp. 1046–1054.

MoveBank Track Annotation Project:

Linking Animal Movement Data with the Environment to Discover the Impact of Environmental Change in Animal migration

Somayeh Dodge[1], Gil Bohrer[1], Rolf Weinzierl[2]

[1]Department of Civil, Environmental & Geodetic Engineering, The Ohio State University
dodge.66@osu.edu, bohrer.17@osu.edu

[2]Max Planck Institute Radolfzell, MoveBank Project
rolf@strd.de

Abstract

The behavior of animals is very much influenced by their surrounding environment. With the advances in positioning and sensor technologies, it is now possible to capture data of animal locations as well as their surrounding environmental information, at previously unseen spatial and temporal granularities. As a consequence, research interest in developing computational methods for the analysis of movement has increased significantly over the past few years. Yet, the link between movement data and the environmental variables has been largely ignored in existing exploratory tools, as well as in previous studies of movement behavior of animals. The MoveBank environmental data annotation project expands an open portal of animal tracking data and enriches it with automated access to environmental variables, as well as effective computational methods to study and process movement and environment data. The aim is to facilitate the investigation and develop a new understanding of spatiotemporal patterns of animal movement in response to a changing environment. The outcomes will contribute to a better modeling, understanding, and ultimately prediction of the behavioral changes of animals in response to global change.

1. Introduction

Today, with the advances in sensing technologies and satellite observations, we have access to a large array of remote sensing datasets capturing the past and current states, and informing models that calculate future forecasts of our dynamic environment. As a consequence, researchers developed a great interest in exploiting these valuable sources of information to gain a better understanding of the interaction between environment and spatiotemporal processes in various disciplines, including animal movement.

Movement is essential to almost all organisms and spatiotemporal processes. Recent years have witnessed an explosion of research activities on movement datasets, triggered by the advent of inexpensive and ubiquitous positioning technologies (e.g. GPS, geo-sensors, RFID tags), in many disciplines such as biology, Geographic Information Science (GIScience), computer science, environmental science, movement ecology, and cognitive science. As a consequence, the study of movement has gained a great momentum in science and technology, as evidenced by the vast amount of literature published on the subject during the past decade (MOVE, 2009).

Currently, knowledge discovery and data mining techniques for analyzing movement data are mostly based on the geometric properties of the trajectories (i.e. the path of an object through space and time) and embedding environmental variables has been ignored (Miller and Han, 2009, Buchin et al., 2011, Dodge et al., 2012). However, in real world applications the movement of an organism is very much influenced not only by its internal state (i.e. the focal individual) but also external factors (i.e. the environment and underlying context) (Nathan et al., 2008). That is, environmental conditions may cause certain movement patterns, and thus can potentially be considered as important indicators for the identification of patterns in the movement of animals. For instance, an animal may move faster or slower depending on how high the ambient temperature is, or stop altogether when there is precipitation. Likewise, abnormal changes in environment temperature may influence the behavior of organisms (Gordon, 1991). For instance, behavioral changes may occur in the migration of birds when cold weather arrives very rapidly rather than when the weather gradually turns cold. Animal can also optimize their energy expenditure during flight by selecting for locations and time when the conditions are supportive for movement. For example, vultures and eagles in their southward fall migration select for a preferential mode of uplift that best fits their flight capacity (Mandel et al., 2011; Bohrer et al., 2012). Therefore, it is essential to gain knowledge about how movement and decisions in moving animals are induced and interact with the physical environment that the animal is exposed to. In order to better understand the behavior of migratory animals and to answer questions such as “when do animals start migrating", “which strategies should they adopt while migrating" and “how if at all do movement rules change in a changing environment”, it is necessary to take a closer look into the interaction of the organisms with their environment, in particular in the continental scale migration.

Animals, particularly birds, travel long distances in their migration courses, and thus, their trajectories cross broad areas of the globe and a diverse environment. Moreover, in order to investigate changes in their migratory behaviors, the animals need to be tracked for entire migratory routes, and preferably with some replication of migration events, over several years. Thereby, very large datasets of remote sensing observations are required to extract the environmental information embedding the animals’ migration paths in space and time. Here, scientific and technical challenges rise in developing the link between the growing collections of animals’ movement data and the *big data* repositories of remote sensing observations containing environmental variables, obtained from satellite remote sensing products such as the MODIS ecological, ocean, land cover and land use data sets, the NCEP-NCAR weather reanalysis datasets, as well as high resolution Digital Elevation Models (DEMs). Namely, efficient storage, indexing, retrieval, and analytical techniques are required for handling and analysis of these data. Also such vast datasets demand for sophisticated context aware data mining and pattern recognition techniques, in order to discover patterns of movement in response to changes in the environment. Hence, an integrated system capable of managing and analyzing movement tracks of animals linked to large climatic and land use datasets is widely needed in the movement ecology community.

2. Research Objectives

The main objective of this study is to develop an open portal that will streamline the co-registration of animal tracking data with a variety of environmental variables such as weather and land surface data. The aim is to provide efficient knowledge discovery methods to examine relationships between observed animal movements (spatiotemporal data of biological observations) and a breadth of information about environmental conditions. The methods will enable discovering unique information about weather dependencies of habitat, migration and landscape connectivity of migratory birds and other threatened and endangered species. These kinds of information are crucial for planning and management of areas allocated as refuges and for forecasting the population status and habitats needs in future conditions of climate and land use changes.

3. Research Questions

The study will help to investigate biological research questions about movement behavior of animals, including migratory birds that are of concerns to the impact of climate change and environmental changes. In order to achieve the objectives, we will, particularly, investigate the following two research questions:

1. How does the movement behavior of animals change in response to a change in environment?
2. Do animals optimize their migration paths according to the climatic conditions?

4. Research Plans

The project is based on extending the capabilities of the existing migration data portal: MoveBank[1] (Kranstauber et al. 2012; Wikelski and Kays, 2012). MoveBank is a free, online database of animal tracking data, which provides biologists and animal tracking researchers with a secure on-line archive to store, manage, process, and share animal movement data. We are currently developing new capabilities within MoveBank, which include:

- Generating an automated system with a graphic user interface to annotate animals' movement trajectories with environmental information. Path annotation attributes environmental data to each reported tracking location (in space and time) along the migration paths.
- Generating and attributing virtual tracks, such as those based on temporal offsets or random walk algorithms, allowing statistical comparisons between observed data and other hypotheses.
- Developing knowledge discovery and visualization techniques to be applied to explore patterns in the linked movement data (e.g. segmentation, movement pattern recognition techniques).

To ensure its relevance and effectiveness, the portal and its toolboxes are designed in collaboration with our wildlife research partners from the US Fish and Wildlife Service (FWS), the US National Park Service (NPS), the US Geological Survey (USGS), who are active participants in this project and are contributing the bird migration data, to guarantee the portal's applicability and relevance to contemporary conservation and wildlife management.

5. MoveBank Data Archive

The proposed system requires a large database containing environmental variables from remote sensing products, as well as spatio-temporal movement trajectories of migratory birds. Both environmental and movement data are growing in time. Moreover, the project aims at handling such data at the global scales. Hence, the system requires a large volume of memory and data space for data storage and retrieval. To alleviate this problem and effective management of our large database we have secured a 25 TB storage space and 100,000 CPU time unites at the Ohio Supercomputer Center (OSC) in Columbus, USA, and a 50 TB storage system at the Max Plank Institute supercomputer at Garching.

As of January 2012, MoveBank currently holds 429 studies of animal movement data published by 159 contributors. These studies include tracking worldwide data of 185 species, 19,414 tracks, which so far contains about 51,000,000 data points that expands by day[1].

[1] www.Movebank.org

Table 1 Available environmental datasets for the trajectory annotation service

Data	Data Source	Projection system / Grid	Temporal Coverage	Geographic coverage	Temporal resolution	Spatial resolution	Data Format
Tropical Rainfall Measuring Mission (TRMM)	NASA (http://trmm.gsfc.nasa.gov/)	Regular lat/lon grid	1998-present	Latitude: 50°N - 50°S Longitude:180°E - 180°W	3-hour	0.25°	unformatted binary
AVHRR land NDVI	USGS (http://phenology.cr.usgs.gov/ndvi_avhrr.php)	Regular lat/lon grid	1989-present	Latitude: 90°N - 90°S Longitude:180°E - 180°W	1-week, 2-week	1 Km	unformatted binary
NCEP Global Reanalysis	NOAA (http://www.cpc.ncep.noaa.gov/)	regular (non-Gaussian) grid	1948-present	Latitude: 90°N - 90°S Longitude:180°E - 180°W	6-hour 8-day	2.5° (208 Km)	NetCDF
North American Regional Reanalysis (NARR)	NOAA (http://www.emc.ncep.noaa.gov/mmb/rreanl/)	Lambert Conformal, Conic Grids	1979-present	Latitude: 90°N - 1°N Longitude: 0° - 170W°	3-hour	32 Km (at 40°N)	GRIB
ECMWF Reanalysis	ECMWF (http://www.ecmwf.int/)	Regular grid	1979-present	Latitude: 89.463°N - 89.463°S Longitude:180°E - 180°W	6-hour	0.7°	GRIB
MODIS Land	NASA (https://lpdaac.usgs.gov/)	Geographic/ Sinusoidal grid	2002 - 2012	Latitude: 90°N - 90°S Longitude:180°E - 180°W	Daily, 8-day, 16-day, monthly	5.6 Km (0.05°)	HDF-EOS
MODIS Ocean	NASA (http://oceancolor.gsfc.nasa.gov/)	Cylindrical Equidistant				4 Km, 9 Km	HDF-EOS
MODIS Snow	NASA (http://modis-snow-ice.gsfc.nasa.gov/	Cylindrical Equidistant				1 Km, 4 Km	HDF-EOS
Ocean productivity	http://www.science.oregonstate.edu/ocean.productivity/	Regular lat/lon grid	1997- 2009	Latitude: 90°N - 90°S Longitude:180°E - 180°W	8-day, monthly	grid sizes 1080x2160 2160x4320	HDF
ASTER GDEM	USGS (https://lpdaac.usgs.gov/content/view/full/11033)	Regular grid, (WGS84 ellipsoid)		Latitude: 83°N - 83°S Longitude:180°E - 180°W		1 arc-second	GeoTIFF
GlobCover	ESA (http://dup.esrin.esa.it/prjs/prjs68.php)	Plate-Carrée projection (WGS84 ellipsoid)	2009	Latitude: 90°N - 65°S Longitude:180°E - 180°W		20 acres	HDF
Socioeconomic data (Population Density Grid)	http://sedac.ciesin.columbia.edu/gpw/global.jsp	Regular grid (WGS84 ellipsoid)	1990-2010	Latitude: 85°N - 58°S Longitude: 180°E - 180°W	5 years	30 arc-second (1km)	ASCII

On the other hand, up to now a large amount of remote sensing data has been obtained from various sources (e.g. NASA, NOAA, USGS, ECMWF), and archived at OSC to support the MoveBank project. Table 1 summarizes the available data sources that are already added to system. As seen on the table, the archive contains large resources of global climatic data, population densities, topography, and land use data over the last several decades. The data are obtained in different formats, such as NetCDF, GRIB, HDF, GeoTIFF, ASCII, and are used to link to animal tracking data. In order to maintain the link between these different datasets we have developed a trajectory annotation service, described in the next section.

6. Showcase: Trajectory Annotation Service

Trajectory annotation service is one of the main components of the system. Annotation is a data integration approach that meets the needs of understanding movement in its environmental context. Borrowed from computer science, where it is used in web browsing, the term “path annotation” is used when additional data about important variables, encountered through a particular path, are added to the object trajectory. In the context of animal movement, an annotated trajectory would include the values of environmental and physiological variables, co-located in time and space with the moving organism (Mandel et al. 2011). Figure 1 shows an example of Galapagos Albatross trajectory, tracked from June to September 2008, annotated with air temperature (a), wind speed (b), and (c) movement speed of the bird.

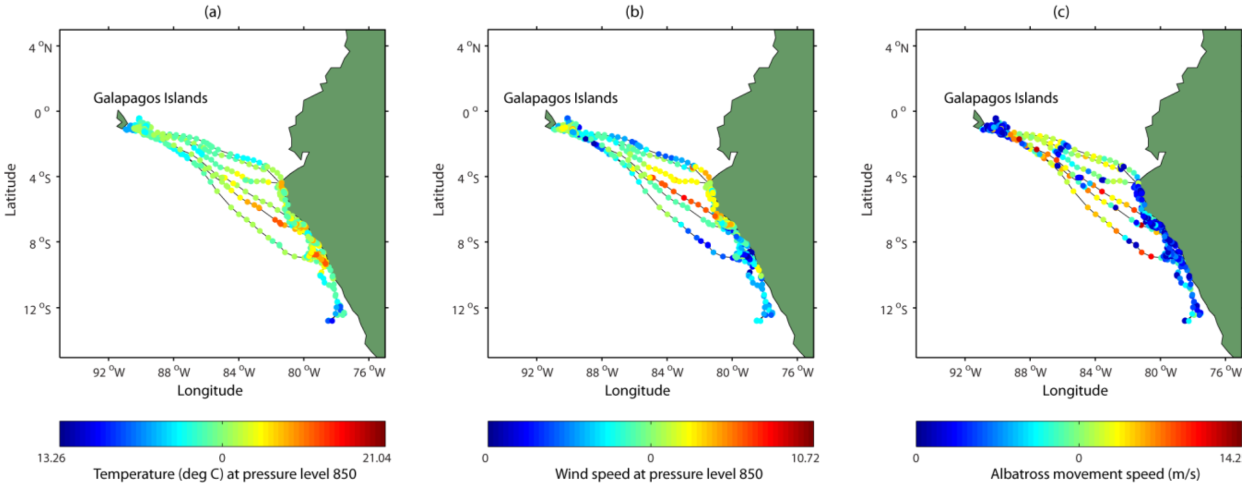

Figure 1 Annotated trajectories of an albatross

Environmental and animal movement data are usually collected in different spatial and temporal scales. It is therefore crucial to choose appropriate spatial and temporal scales for the annotation process, and accordingly a relevant interpolation technique needs to be applied. The developed annotation service allows a number of interpolation techniques in space and time, such as *k*-Nearest Neighbors, Bilinear, Linear, and Inverse Weighted Distance interpolations. Because of the large volume of raster data, efficient interpolations, and indexing strategies have to be undertaken to provide an effective annotation service for the users. The bottleneck of the service is data download from different providers such as (see Table 1).

The annotation service includes the following processes:

a. **Annotation Request:** The annotation service is prompted by a user request through the system interface. The users may request for two types of annotation: annotation of a gridded geographic area, or annotation of a set of trajectories. An annotation request starts with the selection of the requested sub-set of environmental variables, spatial and temporal resolutions, as well as interpolation method. In case of trajectory annotation request, the spatial and temporal resolution and coverage of the environmental data are determined by the system according the spatial and temporal information of the input trajectories.

b. **Data Acquisition:** As the combined volume of globally available environmental data is at the order of petabytes and keeps increasing, it is unfeasible to mirror all the source environmental data locally. Instead, we have developed a caching strategy using mySQL database according to which we keep the most accessed data, and download any other data upon request if it is not locally stored. New data requests (provided as a list of locations, times and variables) are translated to lists of needed data sources, sorted according to data service, variable, timestamp and scene (i.e. a raster tile). Multiple data sources are listed when the location in the movement path requires interpolation between scenes and/or in time, or when derived variables, such as available thermal uplift demand the combination of several input environmental variables. The data-sources list is compared with the stored metadata table and data that are not stored locally are being requested from their provider using an ftp or OPeNDAP interface. We rank each scene according to the frequency it was accessed since download. The least accessed scenes are deleted when space is needed for new data download.

c. **Data Interpolation:** Once all needed data sources are locally available, the environmental data from each scene is interpolated for all trajectory points that are within that scene's domain. The interpolation strategy differs according to the type of data. For instance, for categorical data, the Nearest Neighbor Interpolation is applied, whereas for continuous-numeric data types either a bilinear (in regular grids) or Inverse Weighted Distance interpolation is chosen based on the resolution of data.

The annotated trajectories and gridded geographic areas are delivered to the user via ftp download. The user receives an e-mail when the download is ready. Data are used for the investigations of the interaction between the migratory behaviors of animals and their environment, using data mining and visualization approaches. We provide codes for suggested knowledge discovery and data mining methods and a user support-group site where users can post new analysis and visualization tools (typically in R) and comment on existing tools.

7. Concluding Remarks and Open Questions

This project has to overcome the following technical and methodological challenges in order to achieve the objectives:

- Optimizing storage and retrieval times for a very large dataset of environmental variables from multiple data provides
- Applying effective interpolation techniques in order to maintain the link between animal tracks and their embedding environment in space and time.
- Applying suitable spatiotemporal indexing strategies for data retrieval
- Maintaining a large database of remote sensing data

In addition to the technical challenges, the research has to address scientific problems regarding the development of exploratory methods to investigate the impact of climate change on the migratory behavior of animals. For this purpose, this study will exploit deterministic GIScience, and spatiotemporal data mining techniques, as well as well-known statistical approaches to discover patterns and structures in the migration of animals. There are several methodological questions that have to be taken into consideration in the development of such knowledge discovery approaches. We would like to share these questions at the workshop of *GIScience in the Big Data Ages*, as we think they could be relevant to most of GIScience researches that are dealing with the study of movement and spatiotemporal phenomenon:

- How to integrate context variables (e.g. environmental data) in movement pattern analysis effectively?
- To what extent are deterministic knowledge discovery approaches used in GIScience applicable for finding structures in the movement of animals?
- How generic are the proposed pattern recognition methods for different animal species or various movement datasets?

8. Acknowledgment

This research is funded by NASA under grant number NNX11AP61G. We would like to thank Roland Kays at NCSU, and Martin Wikelski and Kamran Safi at the Max Planck Institute of Ornithology in Germany, for supporting this project through MoveBank portal (www.movebank.org).

9. References

Bohrer, G., D. Brandes, J. T. Mandel, K. L. Bildstein, T. A. Miller, M. Lanzone, T. Katzner, C. Maisonneuve, and J. A. Trembley. (2012). Estimating updraft velocity components over large spatial scales: contrasting migration strategies of golden eagles and turkey vultures. *Ecology Letters* **15**:96–103.

Buchin, K., Buchin, M., van Kreveld, M.J., Luo, J. (2011): Finding long and similar parts of trajectories. *Computational Geometry: Theory and Applications* **44**(9), 465–476.

Dodge, S., Laube, P., and Weibel, R. (2012). Movement Similarity Assessment Using Symbolic Representation of Trajectories. *International Journal of Geographic Information Science*. 26 p., DOI:10.1080/13658816.2011.630003

Gordon, D. M. (1991). Variation and Change in Behavioral Ecology. *Ecology*, **72**(4):1196 –1203.

Kranstauber, B., A. Cameron, R. Weinzerl, T. Fountain, S. Tilak, M. Wikelski, and R. Kays. (2011). The Movebank data model for animal tracking. *Environmental Modelling & Software* **26**:834-835.

Mandel, J. T., G. Bohrer, D. W. Winkler, D. R. Barber, C. S. Houston, and K. L. Bildstein. 2011. Migration path annotation: cross-continental study of migration-flight response to environmental conditions, *Ecological Applications* **21**:2258–2268.

Miller, H.J., Han, J. (2009): Geographic Data Mining and Knowledge Discovery, 2nd edition. Taylor & Francis Group.

MOVE (2009). Knowledge discovery from moving objects (move). Memorandum of understanding for the implementation of a European concerted research action designated as cost action ic0903. http://w3.cost.eu/typo3conf/ext/bzb_securelink/pushFile.php?cuid=253\&file=fileadmin/domain_files/ICT/Action_IC0903/mou/IC0903-e.pdf.

Nathan, R., Getz, W. M., Revilla, E., Holyoak, M., Kadmon, R., Saltz, D., and Smouse, P. E. (2008). A movement ecology paradigm for unifying organismal movement research. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(49):19052–9.

Wikelski, M., and Kays, R. (2012). Movebank: archive, analysis and sharing of animal movement data. World Wide Web electronic publication. http://www.movebank.org, accessed on 2012.