Computers, Environment and Urban Systems xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/compenvurbsys

Constructing gazetteers from volunteered Big Geo-Data based on Hadoop

Song Gao [a,*], Linna Li [c], Wenwen Li [b], Krzysztof Janowicz [a], Yue Zhang [b]

[a] STKO Lab, Department of Geography, University of California, Santa Barbara, CA, USA
[b] GeoDa Center for Geospatial Analysis and Computation, School of Geographical Sciences and Urban Planning, Arizona State University, Tempe, AZ, USA
[c] Department of Geography, California State University, Long Beach, CA, USA

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Gazetteers
Volunteered geographic information
Hadoop
Scalable geoprocessing workflow
Big Geo-Data
CyberGIS

ABSTRACT

Traditional gazetteers are built and maintained by authoritative mapping agencies. In the age of Big Data, it is possible to construct gazetteers in a data-driven approach by mining rich volunteered geographic information (VGI) from the Web. In this research, we build a scalable distributed platform and a high-performance geoprocessing workflow based on the Hadoop ecosystem to harvest crowd-sourced gazetteer entries. Using experiments based on geotagged datasets in Flickr, we find that the MapReduce-based workflow running on the spatially enabled Hadoop cluster can reduce the processing time compared with traditional desktop-based operations by an order of magnitude. We demonstrate how to use such a novel spatial-computing infrastructure to facilitate gazetteer research. In addition, we introduce a provenance-based trust model for quality assurance. This work offers new insights on enriching future gazetteers with the use of Hadoop clusters, and makes contributions in connecting GIS to the cloud computing environment for the next frontier of Big Geo-Data analytics.

1. Introduction

Place is a fundamental concept in daily life and reflects the way humans perceive, experience and understand their environment (Tuan, 1977). Place names are pervasive in human discourse, documents, and social media when location needs to be specified and referred to. Digital gazetteers are dictionaries of georeferenced place names, and play an important role in geographic information retrieval (GIR), in digital library services, and in systems for spatio-temporal knowledge organization (Hill, 2006; Goodchild & Hill, 2008; Li, Yang, & Zhou, 2008; Li, Raskin, & Goodchild, 2012). Several well-known authoritative digital gazetteers have been developed such as the Alexandria digital library (ADL) gazetteer at the University of California Santa Barbara (Goodchild, 2004; Hill, Frew, & Zheng, 1999), the Getty Thesaurus of Geographical Names (TGN) at the Getty Research Institute, the gazetteer maintained by the US Board on Geographic Names (BGN), and a Chinese gazetteer, KIDGS, at Peking University (Liu, Li, et al., 2009). Such authoritative projects require expert teams to make lengthy efforts and the maintenance costs are high, thus often leading to lengthy delays in updating the databases.

With the emergence of the *social Web*, new forms of crowd-sourced gazetteers have become possible. They can be categorized in two types. One is collaborative mapping platforms, such as Wikimapia[1] and OpenStreetMap (OSM),[2] in which volunteers create and contribute geographic features and detailed descriptions to websites where the entries are synthesized into databases. The other way is socially constructed place, that is, gazetteer entries constructed from the Web documents and diverse social-media sources (such as Facebook, Twitter, Foursquare, Yelp, and Flickr) where the general public uses place names, describes sense of place, and makes diverse comments according to their experiences (Goldberg, Wilson, & Knoblock, 2009; Jones, Purves, Clough, & Joho, 2008; Li, Goodchild, & Xu, 2013; Uryupina, 2003). Note that the term *gazetteer* in this paper also includes *point of interest* (POI) databases such that the ***P*** stands for place not point. By mining such rich resources, it is possible to construct or enrich gazetteers in a bottom-up approach instead of in a traditional top-down approach (Adams & Janowicz, 2012; Adams & McKenzie, 2013). However, the data mining and harvesting processes are computationally intensive. Especially in the age of Big Data, the volume, the updating velocity, and the variety of data are too big, too fast and too (semantically and syntactically) diverse for existing tools to process (Madden, 2012).

* Corresponding author. Address: Department of Geography, University of California, 1832 Ellison Hall, Santa Barbara, CA 93106, USA. Tel.: +1 805 980 8424.
E-mail address: sgao@geog.ucsb.edu (S. Gao).

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

http://dx.doi.org/10.1016/j.compenvurbsys.2014.02.004
0198-9715/

In the GIScience/GIS community, researchers may not be willing to wait for weeks or longer to process the terabyte or petabyte-scale geotagged data streams. Fortunately the emerging cloud-computing technologies offer scalable solutions for some of the processing problems in Big Data Analytics.

In this research, we present a novel approach to harvest crowd-sourced gazetteer entries from social media and to conduct high-performance spatial analysis in a cloud-computing environment. The main contribution of this paper is two-folds: First, it introduces the design and implementation of a scalable distributed-platform based on Hadoop for processing Big Geo-Data and facilitating the development of crowd-sourced gazetteers. Second, it provides valuable demonstrations about how to efficiently extract multiple feature types of gazetteer entries at multiple scales and how to integrate emerging data and technologies to improve GIScience research.

The rest of the paper is organized as follows. In Section 2, we introduce some relevant work about space and place, gazetteers, VGI, and Big Data, as well as cloud-computing infrastructures, to help understand the challenges involved in the presented research. In Section 3, we design and implement a novel Hadoop-based geoprocessing platform for mining, storing, analyzing, and visualizing crowd-sourced gazetteer entries; this is followed by experiments and results, as well as a trust evaluation in Section 4. We conclude the paper with discussions and directions for future research (Section 5).

2. Related work

In this section we briefly point to related work and background material.

2.1. Space and place

Space and place are two fundamental concepts in geography, and more broadly in the social sciences, the humanities, and information science (Agnew, 2011; Goodchild, 2011; Goodchild & Janelle, 2004; Harrison & Dourish, 1996; Hubbard, Kitchin, & Valentine, 2004; Tuan, 1977). The spatial perspective is studied based on geometric reference systems that include coordinates, distances, topology, and directions; while the alternative "platial" (based on place) perspective is usually defined by textual place names, linguistic descriptions, and the semantic relationships between places (Gao, Janowicz, McKenzie, & Li, 2013; Goodchild & Li, 2012a; Janowicz, 2009). There would not be any places without people's perception and cognition. As argued by Tuan (1977), it is humans' interactions and experiences that turn space into place. Place is not just a thing in the world but a social and cultural way of understanding the world. Giving names and descriptions to locations is a process to make space meaningful as place. Social-tagging, tweets, photo sharing, and geo-social check-in behaviors have created a large volume of place descriptions on the Web.

Researchers have made significant efforts toward georeferencing place descriptions and processing spatial queries, such as using ontologies of place (Jones, Alani, & Tudhope, 2001), using a qualitative spatial reasoning framework (Yao & Thill, 2006), using fuzzy objects (Montello, Goodchild, Gottsegen, & Fohl, 2003), using probability models in combination with uncertainty (Guo, Liu, & Wieczorek, 2008; Liu, Guo, Wieczorek, & Goodchild, 2009), using kernel-density estimation (Jones et al., 2008), using description logics (Bernad, Bobed, Mena, & Ilarri, 2013), as well as knowledge discovery from data techniques for platial search (Adams & McKenzie, 2012). Recently, a review by Vasardani, Winter, and Richter (2013) has suggested that a synthesis approach would provide improvements in locating place descriptions, and that new opportunities exist in identifying places from public media and volunteered sources by using Web-harvesting techniques.

2.2. Gazetteers

Existing GIS and spatial databases are mature in representing space, but limited in representing place. In order to locate place names on a map with precise coordinates and to support GIR, efforts have been taken to convert place to space. One major mechanism is the use of gazetteers, which conventionally contain three core elements: place names (N), feature types (T), and footprints (F) (Hill, 2000). A place name is what people search for if they intend to learn about a place, especially its location, in a gazetteer. A place type is a category picked from a feature-type thesaurus for classifying similar places into groups according to explicit or implicit criteria. Janowicz and Keßler (2008) argued that an ontological approach to defining type classifications will better support gazetteer services, semantic interoperability (Harvey, Kuhn, Pundt, Bishr, & Riedemann, 1999; Scheider, 2012), and semi-automated feature annotation. A footprint is the location of a place, and is almost always stored as a single point which represents an extended object as an estimated center, or the mouth in the case of a river. Recent work is providing additional spatial footprints including polygons and part-of relations.

One major role of a gazetteer is thus to link place names to location coordinates. For example, the ADL model which links places to spatially defined digital library resources requires a comprehensive gazetteer as part of its spatial query function to provide access to web services, including collections of georeferenced photographs, reports relating to specific areas, news and stories about places, remote sensing images, or even music (Goodchild, 2004). The minimum required elements of a place in ADL model are represented by the triples ⟨N,T,F⟩. As a start, ADL combines two databases: the Geographic Names Information System (GNIS) and the Geographic Names Processing System (GNPS), both from US federal-government agencies. Frequently, it is necessary to consult and combine results from multiple gazetteer sources, which is generally described as (feature) conflation (Saalfeld, 1988). Hastings (2008) has proposed a computational framework for automated conflation of digital gazetteers based on three types of similarity metrics: geospatial, geotaxial, and geonomial. In addition, efforts have been made in mining gazetteers semi-automatically from the Web (e.g., Goldberg et al., 2009; Uryupina, 2003). Challenges such as interoperability and quality control need to be investigated in such crowd-sourced gazetteers. The conflation of POI databases is widely considered an important next research step to combine the different attributes stored by various systems to more powerful joint database.

2.3. Big Data and VGI

Big Data is used to describe the phenomenon that large volumes of data (including structured, semi-structured, and unstructured data) on various aspects of the environment and society are being created by millions of people constantly, in a variety of formats such as maps, blogs, videos, audios, and photos. Big Data is "big" not only because it involves a huge amount of data, but also because of the high dimensionality and inter-linkage of a multitude of (small) datasets that cover multiple perspectives, topics, and scales (Janowicz, Scheider, Pehle, & Hart, 2012). The Web has lowered previous barriers to the production, sharing, and retrieval of varied information linked to places. VGI (Goodchild, 2007), a type of user-generated content (UGC) with a geospatial component, has gradually been taking the lead as the most voluminous source of geographic data. For example, there were over 20 million

Please cite this article in press as: Gao, S., et al. Constructing gazetteers from volunteered Big Geo-Data based on Hadoop. *Computers, Environment and Urban Systems* (2014), http://dx.doi.org/10.1016/j.compenvurbsys.2014.02.004