Computers, Environment and Urban Systems 54 (2015) 144–153

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Detecting tourism destinations using scalable geospatial analysis based on cloud computing platform

Xiaolu Zhou [a,*], Chen Xu [b], Brandon Kimmons [c]

[a] Department of Geology and Geography, Georgia Southern University, P.O. Box 8149, Statesboro, GA 30460, United States
[b] Department of Geography, University of Wyoming, 1000 E. University Ave., Laramie, WY 82071, United States
[c] Georgia Southern University, P.O. Box: 8136, United States

ARTICLE INFO

Article history:
Received 22 February 2015
Received in revised form 11 June 2015
Accepted 24 July 2015
Available online 26 August 2015

Keywords:
Flickr
Geographic Information Systems
RHadoop
Spatial analytics

ABSTRACT

The number of geo-tagged digital photos has grown exponentially in the past decades. Increasing numbers of digital photos with geo-tags are available on many photo-sharing websites such as Flickr and Instagram. The proliferation of online photos offers great opportunities to study people's travel experiences and preferences. Mining tourists' behavior and city preferences has become popular in recent geographic information system (GIS) research. However, the huge amount of data also poses challenges in spatial analytics. In this study, we automate the detection of places of interest in multiple cities based on spatial and temporal features of Flickr images from 2007 on. We also speed up the process by running jobs on top of the RHadoop platform. This project provides fast and accurate tourist destination detection by mining large amounts of geo-tagged Flickr images. In addition, this study provides insight in applying the RHadoop platform to strengthen large geospatial data analytics. Our methods can be applied to many other cities, and results are valuable for tourism management.

1. Introduction

Tourists are unique in that their activities are highly constrained by the space–time budget. Therefore, gleaning information about a distant place before traveling becomes an important task. In the era of Web 2.0, this task is not as daunting because many tourism websites such as homeandabroad.com and tripadvisor.com provide information about popular places. Unfortunately, although travel sites produce an enormous amount of information for potential visitors, digesting the different types of information from many contributors requires significant time and effort. Moreover, tourism websites likely apply different criteria when ranking popular destinations. Making choices among the most popular destinations is challenging. Hence, knowledge gained directly from actual travelers rather than lists of popular locations from websites can be more useful for tourists. The habits and patterns of travelers can be found through massive social media datasets, such as publically available photos, which have been the subject of an increasing number of studies (Hollenstein & Purves, 2014; Kádár, 2014; Papadopoulos, Troncy, Mezaris, Huet & Kompatsiaris, 2011). These data sources allow geographers, urban planners, and authorities to gather tourists' travel behavior (Gavric, Culibrk, Lugonja, Mirkovic & Crnojevic, 2011).

Tourists, empowered by the development of Web 2.0 technologies, can create travel stories of their own using all kinds of new social media platforms and photo services. Some of the more popular photo hosting and sharing websites include Instagram, Imgur, and Flickr. Flickr, the most popular, hosts over 5 billion photos, as of September, 2010 (Sheppard, 2010). The openly available photos not only offer possibilities for virtual tourism, but also contain an immense amount of geographic information from people of vastly different demographic backgrounds. This information has broad social and practical significance for social science studies (Kennedy, Naaman, Ahern, Nair & Rattenbury, 2007). The rich content associated with each photo provides opportunities to transform information to knowledge. Tags of images not only provide meaningful descriptions, but also convey people's perspectives and emotions (Li & Zhang, 2011). Geocoordinates and timestamps strengthen people's abilities to pinpoint photos in time–space dimensions. By studying the structured and unstructured information, studies have attempted to extract knowledge about events, city cores, and preference (Kennedy et al., 2007; Kisilevich, Krstajic, Keim, Andrienko & Andrienko, 2010a; Kisilevich, Mansmann & Keim, 2010b; Kisilevich, Mansmann, Bak, Keim & Tchaikin, 2010c; Li, Cai, Huang, Yang & Zhou, 2014).

Nevertheless, it is difficult to extract useful information from the vast amount of online photos. Due to the magnitude of uploads, it is getting more and more difficult to organize and manage the huge amount of data that can be downloaded from service providers (Kennedy et al., 2007). In addition to the huge amount of photos, the textual tags and

[*] Corresponding author.
E-mail address: xzhou@georgiasouthern.edu (X. Zhou).

http://dx.doi.org/10.1016/j.compenvurbsys.2015.07.006
0198-9715/© 2015 Elsevier Ltd. All rights reserved.

descriptions are unstructured and "noisy," making information extraction difficult. Unlike news stories, where each piece of news is related to a certain event, not every Flickr photo presents consistent or useful information (Chen & Roy, 2009). The challenges have been succinctly defined as the big data challenge. According to IBM, big data poses difficulties because of its huge data volume, rapid data velocity, broad variety of content, and problematic veracity of information (Madden, 2012; Gao, Li, Li, Janowicz & Zhang, 2014).

We seek to provide solutions to these big data challenges by constructing a prototype system for processing the large volume of Flickr photo data. The first aim of this paper is to detect tourism destinations based on analyzing the spatial and temporal features of images. The knowledge generated by the system can indicate popular tourist destinations, show the boundaries of these destinations, and extract descriptions of these places from the Flickr content. The second aim of this study is to speed up the data extraction process using the "divide and conquer" methods on top of a RHadoop platform. Although Hadoop-based approaches have been applied by computer scientists for more general big data-related studies, their implementation in the geographic domain still deserves further exploration. RHadoop seamlessly integrates R and Hadoop frameworks, enhancing the analysis capability of the pure Hadoop environment. However, RHadoop is not well documented in previous studies, especially in geospatial big data research. The following section introduces previous methods used to detect place of interests from social media data.

1.1. Related work

Tourism research uses different methods to map different tourists' favorite destinations. Clustering is an important method used to detect where people like to visit, and K-means is a popular type of unsupervised locational clustering (Ahern, Naaman, Nair & Yang, 2007; Kennedy & Naaman, 2008). One drawback of the K-means method is the fixed number of clusters that are predefined manually. Realizing that the K-means approach might generate non-discriminating results for cities with different levels of tourist interest, several studies used a mean shift algorithm to replace the fixed-based approach when finding the most popular places (Crandall, Backstrom, Huttenlocher & Kleinberg, 2009; Yin, Cao, Han, Luo & Huang, 2011). In contrast to the K-means method, the mean shift approach is non-parametric in nature and detects clusters by relying on the probability distribution of samples. Thus, the number of clusters is not pre-determined and varies according to a given estimate of the scale of the input data (Crandall et al., 2009). The scale-based estimation thus produces comparable results for cities that attract different numbers of tourists. In addition, more sophisticated density-based approaches such as DBSCAN and P-DBSCAN have been used to detect landmarks inside a city. The benefit of applying density-based approaches is that they are flexible and therefore require minimum knowledge of the search radius and a minimum number of points in a cluster. Density-based approaches also use algorithms that can detect the realistic shapes of geographic boundaries, a persistent interest of geographers (Kisilevich et al., 2010a,b,c; Majid et al., 2013). Cluster-based methods effectively reveal patterns hidden in data and have been continually studied and improved by researchers. For example, Yang et al. (2011) proposed a self-tuning spectral clustering approach to properly assign the parameters for the clustering.

While clustering approaches can be implemented to group the spatial data from geotags on photos, they are also applicable to non-spatial data—textual tags and actual photos, for instance. When grouping similar images, global and local features are extracted as signatures of photos to be clustered (Moxley, Kleban, Xu & Manjunath, 2009; Kennedy and Naaman, 2008). One of the most prominent algorithms of image feature detection is Scale-invariant feature transform (SIFT), which detects distinctive signatures that are invariant to scale and rotation (Lowe, 2004). Clustering methods are then applied to find the most attractive destinations. For calculating clusters of tags, different algorithms have been used to process the unstructured textual information. The most popular method is term frequency–inverse document frequency (tf–idf) that locates the most important and representative descriptive words. Combined image features and tag features are also used together to detect popular destinations. Papadopoulos et al. (2011) calculated a similarity index using both Flickr tags and images and used a hybrid similarity image graph to cluster images.

Recent studies also leverage data from multiple sources to strengthen tourist spot detection. Majid et al. (2013) applied a density-based clustering algorithm on geotags to extract tourist locations and then supplied them with additional information extracted from Google Place Service. Studies have developed mash-up systems to collect information and use hybrid filtering techniques to recommend events (Kayaalp, Özyer & Özyer, 2011). Kisilevich et al. (2010a,b,c) developed a framework to identify attractive spots using density-based clustering and produced interactive visualizations using Google Earth Mashup.

Spatial data grow rapidly, partially due to the fast improvement of data acquisition technologies. To benefit from the increasingly large volume of observational data for scientific research and practical applications, it has become more critical to better manage large spatial data (Aji et al., 2013). The limitations of traditional stand-alone computing technology become more apparent with data size growing beyond the capacity of single machine. Grid computing has been proposed and applied mostly within academic fields to ensure agile growth of computational capacity in response to ever-growing data volume. Commercially, grid-based approaches have only attracted limited interest. In contrast, cloud computing, a more flexible means of proportionally expanding physical hardware according to demands, commercialized by companies like Amazon, has become immediately popular in academic and industrial domains. In geospatial domains, the combination of GIS and cloud computing provides a promising framework for spatial information storage, processing, and analytics (Weng & Liu, 2013; Yang, Raskin, Goodchild & Gahegan, 2010). According to Yang et al. (2011), this analytical framework can potentially solve four intensity problems: the data intensity challenge, the computing intensity challenge, the concurrent-access-intensity challenge, and the spatiotemporal-intensity challenge that are all relevant to geospatial analytics. Recent studies have attempted to apply cloud computing and CyberGIS infrastructure to solve geospatial problems with the availability of social media data, contributed voluntarily by a massive amount of social media users. For instance, Padmanabhan et al. (2014) presented FluMapper, a system that used massive social media data to investigate flu risks across spatial and temporal dimension, based on a data-driven framework using CyberGIS infrastructure.

With the computational paradigm shifting from single machine to cloud computing, the impact of effective and efficient data management becomes imperative. Because there are many forms of cloud computing that must be adjusted to the unique demands of tasks and customers, several successful cloud resource management architectures have been created. The Hadoop-based architecture divides large computing tasks into many individual itemized smaller tasks and then integrates all itemized results into one overall output. The two steps are accordingly termed Map and Reduce. Gao et al. (2014) applied the map-and-reduce paradigm on a cloud-computing platform for supporting gazetteer research. This study constructed a scalable geoprocessing workflow based on the Hadoop ecosystem to organize and analyze crowd-sourced gazetteer entries. For image analysis, the map-and-reduce framework has been used to efficiently extract features from large-scale images (Cheng et al., 2014; Yan & Huang, 2014).

Unfortunately, Hadoop-based cloud computing techniques have been less used to address the big online photo data challenge and to extract useful tourist knowledge. To address this gap, this paper provides timely and valuable information on leveraging emerging advanced