Expert Systems with Applications 41 (2014) 3514–3526

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Sequential pattern mining of geo-tagged photos with an arbitrary regions-of-interest detection method

Guochen Cai, Chihiro Hio, Luke Bermingham, Kyungmi Lee, Ickjai Lee *

School of Business (IT), James Cook University, Cairns, QLD 4870, Australia

ARTICLE INFO

Keywords:
Sequential pattern mining
Geo-tagged photo
Regions-of-interest
Trajectory pattern mining

ABSTRACT

Geo-tagged photos leave trails of movement that form trajectories. Regions-of-interest detection identifies interesting hot spots where many trajectories visit and large geo-tagged photos are uploaded. Extraction of exact shapes of regions-of-interest is a key step to understanding these trajectories and mining sequential trajectory patterns. This article introduces an efficient and effective grid-based regions-of-interest detection method that is linear to the number of grid cells, and is able to detect arbitrary shapes of regions-of-interest. The proposed algorithm is combined with sequential pattern mining to reveal sequential trajectory patterns. Experimental results reveal quality regions-of-interest and promising sequential trajectory patterns that demonstrate the benefits of our algorithm.

1. Introduction

With the advance of Web 2.0/3.0 techniques, mobile technology, and social media, the Web-as-a-data-generator framework enables users to produce their own content and communicate with others on the Web. A photo sharing service is one such example of social networking through social media. Benefiting from the advancement of mobile digital camera technology, people can easily take photos when they find something interesting and upload them to a photo sharing platform to manage and share exciting moments with their friends, families and colleagues.

These photos uploaded are associated with location (spatial) information and time (temporal) information and other metadata through the geo-tagging service available on the photo sharing site. Geo-tagging is a technique that includes a geog raphic reference inside the metadata of specific types of content: photos, videos, and SMS. The effect geo-tagging has had on the amount of valuable data available for extraction is profound. Flickr (http://www.flickr.com/), one of the most popular photo sharing sites on the Web, reported in 2011 that its photo upload count was "about 4.5 million daily". Clearly, photo sharing and social media sites such as this present a potential ocean of valuable information that can be harnessed.

In particular, the spatial and temporal metadata can be used to reflect where a photo-taker was and what he/she was doing. With this in mind, a series of photos, presented chronologically, leaves a trail of sequential whereabouts, and can describe approximate spatio-temporal movements of an individual. A collection of spatio-temporal entries connected to represent movement is known as a trajectory. Analysis of multiple photo-taker trajectories can reveal valuable, previously unknown, information such as frequent travel patterns and regions-of-interest (RoI) (Lee, Cai, & Lee, 2013a; Lee, Cai, & Lee, 2014; Yin, Cao, Han, Luo, & Huang, 2011; Zheng, Zha, & Chua, 2012).

Sequential trajectory pattern mining reveals frequent trajectory patterns and it has been studied in several other domains in recent years (Kang & Yong, 2010; Kisilevich, Keim, & Rokach, 2010; Zheng, Zhang, Xie, & Ma, 2009). For example both (Kang & Yong, 2010; Zheng et al., 2009) mine people's travel sequences using GPS trajectory data. Similarly, Kisilevich et al., (2010) conducts research extracting sequential movement patterns from geo-tagged photos. However, these approaches only consider the spatial dimension and did not consider both spatial and temporal dimensions simultaneously. Clearly the exclusion of the temporal dimension greatly limits the value of constructed patterns. Specifically, the temporal dimension is an important feature of trajectory data, which can provide rich and specialized details about the underlying structural pattern in the data. Giannotti, Nanni, Pinelli, and Pedreschi, (2007) propose a well-defined trajectory pattern mining (TPM) framework to find trajectory patterns that identify frequent sequences of visited RoIs with travel time. TPM combines RoI mining and sequential mining. It first finds RoIs using a grid-based indexing approach, and then applies sequential pattern mining to find a series of frequent RoI patterns. However, the RoI mining technique used in TPM suffers from several drawbacks. Firstly, it can only generate

* Corresponding author. Tel.: +61 740421083.
E-mail addresses: Guochen.Cai@my.jcu.edu.au (G. Cai), Chihiro.Hio@my.jcu.edu.au (C. Hio), Luke.Bermingham@my.jcu.edu.au (L. Bermingham), Joanne.Lee@jcu.edu.au (K. Lee), Ickjai.Lee@jcu.edu.au (I. Lee).

0957-4174/$ - see front matter
http://dx.doi.org/10.1016/j.eswa.2013.10.057

rectangular shaped RoIs and it is not able to detect arbitrary RoI shapes. Furthermore, TPM also tends to create large RoIs that contain many sparse grid cells (false positives). Thus, the resulting RoIs produced by TPM do not truly represent interesting dense regions (true positives), and fail to distinguish nearby landmarks and popular places.

In this paper, we present two improved RoI mining algorithms, called: (1) slope and (2) hybrid RoI mining. We show that both of these techniques are able to detect finer and more accurate arbitrary RoI shapes than TPM, thus allowing them to capture true landmarks and interesting geo-places. These algorithms are combined with the sequential pattern mining segment of TPM to discover more meaningful RoIs and more insightful frequent patterns in the study area. We extract real Flickr datasets covering New York city, United States of America in 2012 in order to validate our approach. Experimental results demonstrate the robustness, usefulness and effectiveness of our proposed method. The sequential trajectory patterns our approach produces highlight frequent sequences of locations visited by photo-takers. By uncovering accurate information about photo-takers' movement behaviour, various application areas can be greatly benefited including tourism, retail, transport, and marketing.

The rest of paper is organized as follows. Section 2 briefly outlines preliminaries on trajectory data mining and the original TPM framework. Section 3 proposes our new RoI mining algorithms, slope and hybrid algorithm. Section 4 introduces our framework for mining frequent trajectory patterns. Section 5 compares the results of original TPM and our hybrid TPM, whilst also presenting some interesting dataset specific patterns discovered by our hybrid TPM. Finally, Section 6 concludes with final remarks and opens a road to possible future work.

2. Preliminaries

2.1. Trajectory data mining

Trajectory data mining is the process of producing interesting previously unknown knowledge from a moving entity (trajectory) dataset. Trajectory data mining approaches can be divided into two clear categories: entity-based and field-based. Entity based methodologies calculate results using the intra-object characteristics (i.e. spatial similarity) whereas field-based approaches partition the study region and compare neighbour characteristics (i.e. cell density) to discover knowledge. Our research is focused on extracting RoIs from trajectory data using a field-based approach. However in order to define where our research integrates into the field we provide a broad overview of trajectory data mining approaches: clustering, RoI mining, and pattern mining.

2.1.1. Trajectory clustering

Han, Kamber, & Pei, (2011) define clustering as; "the process of grouping a set of physical or abstract objects into classes of similar objects". Trajectory clustering is simply an extension of this concept where the objects analyzed are either whole trajectories or sets of trajectory line segments (sub-trajectories). Trajectory clustering is useful in variety of data mining applications, but notably in relation to this research it has even been applied to the problem space of discovering RoIs in trajectory data. In Palma, Bogorny, Kuijpers, and Alvares (2008) apply the novel idea that interesting places occur in a trajectory when entity movement resembles a "stop", using this theory they cluster around stop points to form cluster RoIs. It is clear to see that by its definition trajectory clustering, for whatever application, is a classic example of an entity based mining method. However, similar to the geoinformatics axiom that, "raster is faster but vector is corrector" (Lee & Phillips, 2008; Estivill-Castro and Lee, 2004), entity based methods often trade higher quality results for a longer computational running time, as is the case in Palma, Bogorny, Kuijpers, and Alvares (2008).

2.1.2. Trajectory RoI mining

RoI mining is the extraction of interesting, previously unknown regions from a trajectory dataset. Most RoI mining techniques work using field-based approaches. Generally, the study region is partitioned into some space-filling structure for fast indexing and region-based computations. For example in Giannotti et al. (2007) partition a trajectory dataset into grid cells on 2D spatial plane (i.e latitude, longitude). Following this the density of each grid cell is determined by the number of trajectories that pass through it. Finally using some intelligence grid cells above a certain threshold are expanded to form rectangular shaped RoIs in the study region. As expected due to its field-based nature in practise this method turns out to be computationally efficient and scalable. However, the rectangular shape of the resulting RoI poses a problem to the quality of the results in some datasets, meaning some RoIs exhibit an undesirable false positive characteristic.

2.1.3. Trajectory pattern mining

Trajectory pattern mining is the process of discovering frequent entity sequences between visited locations in a trajectory dataset. For example in a tourist dataset a frequent trajectory sequence might be {Airport → Train Station → Shopping Centre}. The simplified process of trajectory pattern mining is to intelligently convert a trajectory dataset into a set of distinct visitation sites (i.e RoIs or clusters) and then translate each trajectory into a series of visitation sites (i.e A → B → A → C). The final step is then to use these sequences of visitation sites as input for a suitable sequential pattern mining algorithm, such as PrefixSpan (Han et al., 2011). In Giannotti et al. (2007) this process is formally defined in a trajectory pattern mining framework using RoIs as the visitation sites and PrefixSpan as the sequential pattern mining algorithm. It should be clear to see that any trajectory pattern mining algorithm is highly reliant on the quality and type of visitation site used to discover frequent patterns. Therefore, this research aims to improve the false positive characteristic of the RoI algorithm used in Giannotti et al. (2007)) and in turn lay the groundwork to produce both better quality RoIs and frequent trajectory patterns.

2.2. Flickr data mining

Massive amounts of photos are uploaded to Flickr each day. These photos contain additional metadata that can lead to mining interesting patterns about photo-takers' behaviours and movements. Unsurprisingly, this has lead to Flickr data mining to become a hot area of research. In general traditional Flickr data mining research can be categorized into the clustering and Association Rules Mining (ARM).

2.2.1. Flickr clustering

The basic concept of Flickr clustering is to partition geo-tagged photos into similar groups in order to maximize inter-group difference and maximizes intra-group similarity. There have been several clustering methods proposed in the literature located geo-referenced photos indicate points-of-interest (PoI) where relatively a large amount of photos is taken. Zheng et al. (2012) investigate regions of attractions that are similar to PoI, and use them for route analysis. Lee, Cai, and Lee (2013b) propose a framework to discover geographical hierarchical PoI and PoI in several temporal categories. Based on the extracted PoI, Zheng et al. (2012) find popular travel routes presented as a set of visited PoI. Kisilevich et al. (2010) use a density-based clustering variant in their study for POI identification.