Expert Systems with Applications 41 (2014) 397–405

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Exploration of geo-tagged photos through data mining approaches

Ickjai Lee *, Guochen Cai, Kyungmi Lee

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

ARTICLE INFO

Keywords:
Clustering
Association rules mining
Geo-tagged photo
Points-of-interest

ABSTRACT

With the development of web technique and social network sites human now can produce information, share with others online easily. Photo-sharing website, Flickr, stores huge number of photos where people upload and share their pictures. This research proposes a framework that is used to extract associative points-of-interest patterns from geo-tagged photos in Queensland, Australia, a popular tourist destination hosting the great Barrier Reef and tropical rain forest. This framework combines two popular data mining techniques: clustering for points-of-interest detection, and association rules mining for associative points-of-interest patterns. We report interesting experimental results and discuss findings.

1. Introduction

Web 2.0 and Web 3.0 technologies serve as a Web-as-participation-architecture with which users are encouraged to add values. With this Web structure, users are able to share user-generated social medias anytime and anywhere, and interactively communicate others. As web-based and mobile-based technologies advance, social medias are increasingly collected beyond the capability of human analysis (Lee & Torpelund-Bruin, 2011). Social networks combine the traditional blog, BBS, e-mail, instant messaging and other forms, and also add a variety of supporting applications. A key element of the technology is that it allows people to create, share, collaborate and communicate. The nature of this technology makes it easy for people to create and publish or communicate their work to either a selected group of people or to a much wider audience, or to the world.

Photo sharing platform is one kind of social network platforms. Benefitting from the development of Web album, people can easily share their photos on websites. Flickr (http://www.flickr.com) is one of the most popular photo sharing websites which is a great resource for photography enthusiasts and increasingly for travelers. Flickr has recently launched its own service for adding latitude and longitude information to a picture and provides the tool that allows a user to display pictures on online maps like OpenStreetMap (http://www.openstreetmap.org). In addition, many photos are geo-tagged automatically using GPS logs or location aware devices. Therefore, the location and time data associated with photos and other related text tags can be considered as useful geographically annotated materials on the Web. Eventually, they generate huge amount of tourist trails and detailed trajectories of what sites and in what order tourists visit. To extract tourists' photo-taking pattern is of significant importance to obtain the place the tourists visit and take photos for tourism related organizations.

Some research has been conducted with Flick datasets (Crandall, Backstrom, Huttenlocher, & Kleinberg, 2009; Kennedy, Naaman, Ahern, Nair, & Rattenbury, 2007; Kisilevich, Mansmann, & Keim, 2010; Rattenbury, Good, & Naaman, 2007a, Rattenbury, Good, & Naaman, 2007b; Yang, Gong, & Hou, 2011; Zheng, Zha, & Chua, 2012), however most work focuses on finding attractive areas points-of-interest (PoI) (Crandall et al., 2009; Kisilevich et al., 2010; Zheng et al., 2012) and folksonomy based social tagging (Kennedy et al., 2007; Rattenbury et al., 2007a, 2007b). Zheng et al. (2012) investigate regions of attractions that are similar to PoI, and use them for route analysis. Kisilevich et al. (2010) modify DBSCAN to find out PoI from Flickr photos. It is adaptive and flexible but only limited to PoI mining. Crandall et al. (2009) use classification methods to analyze Flickr geo-tagged photos. Kennedy et al. (2007) use the concept of representative tags and tag-driven approach to extract place and event semantics. Rattenbury et al. (2007a, 2007b) produce similar research and investigate ways to extract place and event semantics from folksonomy.

On the other hand, some recent studies pay more attentions to route recommendation systems (Kurashima, Iwata, Irie, & Fujimura, 2010; Lu, Wang, Yang, Pang, & Zhang, 2010; Okuyama & Yanai, 2011; Shi, Serdyukov, Hanjalic, & Larson, 2011). In these studies, geo-tagged photos are modeled as a sequence of location points, and travel sequences are then found. Shi et al. (2011) combine user-landmark preference and category-based landmark similarity to provide personalized landmark recommendation. Lu et al. (2010) study one similar work where users are able to specify personal preference in the travel route planning. Kurashima et al. (2010) integrate topic models into Markov models to provide travel route recommendations whilst Okuyama et al. (2011) extract a travel plan using trip models represented by the order sequences

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

0957-4174/$ - see front matter Crown Copyright © 2013 Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.07.065

of tourist places. These approaches do not reveal associative PoI patterns, but mainly focus on travel route recommendations. None of the previous work reveals associative PoI patterns exposing positive PoI relations. This paper concentrates on the identification of PoI and associative relationship mining among PoI.

In this paper, we focus more on structural and practical aspects of Flickr mining rather than technical and algorithmic aspects of it. Main contributions of the paper are in two folds. First, this paper proposes a mining framework for PoI associations. It first finds PoI using clustering and applies association rules mining to detect associative PoI patterns. Second, we analyze geo-tagged photos from Flickr for Queensland Australia, the second largest state hosting the Great Barrier Reef and world heritage rainforest. We report interesting experimental results and discuss findings.

The rest of paper is organized as follows. Section 2 briefly outlines preliminaries on clustering and association rules mining. Section 3 introduces our framework for associative PoI mining. Section 4 reports PoI clustering results and Section 5 further explores PoI association mining. Section 6 concludes with final remarks.

2. Preliminaries

2.1. Clustering

Clustering is a process of grouping objects into classes by measuring similarities among objects. Objects within one cluster have high similarity compared with objects in other clusters. Partitioning clustering and density-based clustering are two popular clustering approaches. Given a set of n objects, and a cluster number k, a partitioning method assigns the n objects into k clusters. One typical algorithm is k-means (MacQueen, 1967). The k-means algorithm sets the mean value of objects in a cluster as a cluster centre. It is a simple and efficient method for most data sets because its computational complexity is $O(nkt)$, where n is the number of objects, k is the number of clusters, and t is the number of iterations. For most data sets, $k \ll n$ and $t \ll n$, k-means becomes $O(n)$. However, k-means method is not effective in the presence of noise outliers. One drawback of k-medoid method, a popular variant of k-means, is that it requires an exhaustive search in order to find the best candidate of a cluster centre, which is inefficient in data-rich environments.

However, density-based clustering is based on data intensity. This technique defines one cluster as a set of points within an area with high density. DBSCAN (Sander, Ester, Kriegel, Wimmer, & Xu, 1998) is the most representative density-based clustering approach. It searches for areas of high density with two parameters *Eps* and *MinPts*. Areas of given neighborhood (*Eps*) contain at least a minimum number of points (*MinPts*) are reported as a part of a cluster. A point p_1 is directly density-reachable from a point p_2 if p_1 is within a given neighborhood distance *Eps* and also if p's *Eps* contain at least *MinPts*. A point p is density-reachable from a point q if there is a sequence $p_1, p_2, \ldots, p_n$ of points with $p_1 = p$ and $p_n = q$ where each $p_i + 1$ is directly density-reachable from p_i. A set of density-reachable points forms a cluster. DBSCAN first picks a core point and then expands it by searching for all density-reachable points from the core point. The set of points that do not belong to any cluster (not density-reachable from any other clusters) is classified as noise points. It continues searching and expanding until no point is left.

k-means clustering and DBSCAN have different applications in the geospatial context. k-means is better suited for location optimization problem or the number of clusters is known beforehand, whereas DBSCAN is better suited for finding geospatial aggregations in the presence of noise points. In this research, we are interested in geospatial concentrations of geo-tagged photos and the number of clusters is unknown beforehand. Thus, DBSCAN is used as a default clustering approach in this work.

2.2. Association rules mining

Association rules mining is an important technique in data mining which aims to extract frequent associative patterns from transaction databases.

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of items. Let $D = \{t_1, t_2, \ldots, t_n\}$ be a set of transactions. Each transaction in D contains a subset of items in I. An association rule is defined as $X \Rightarrow Y$ which is interpreted as X implies Y, where $X, Y \in I$ and $X \cap Y = \Phi$. X is called antecedent while Y is consequence. The rule $X \Rightarrow Y$ holds with support s and confidence c, where s is the percentage of transactions in D that contains X and Y (the union of transactions that contains X and Y) and c is the percentage of transactions in D containing X that also contain Y.

$$s = \text{probability } (X \cup Y),$$
$$c = \text{probability } (X \cup Y)/\text{probability } (X).$$

Two user-provided constraints minimum support and minimum confidence are used to focus on frequent patterns and association rules. Frequent patterns are those k-itemsets that satisfy the user-supplied minimum support constraint whilst strong rules are those association rules satisfying minimum support threshold and minimum confidence threshold. Generally, we are interested in rules with high support and strong confidence. Association rues mining is to find all strong rules.

Apriori algorithm (Agrawal & Srikant, 1994) and FP-growth algorithm (Han, Pei, Yin, & Mao, 2004) are two prevalent methods. The former is based on candidate generation that is based on the downward closure property which guarantees that all subsets of one frequent k-itemsets are frequent. This property prunes away unnecessary generations of candidate frequent itemsets. Frequent itemsets will be selected by meeting the minimum support, and the association rules are then generated from the frequent itemsets that meet the minimum confidence. While, FP-growth algorithm is based on an extended prefix-tree structure which is called frequent patterns tree that expresses and stores frequencies of items. The frequent itemsets are generated from the tree by traversing in a bottom-up fashion, and then association rules are produced based on frequent itemsets. The apriori algorithm is used by default in this study, and all parameter values presented in this paper are for the apriori algorithm.

3. Framework of mining PoI association rules

3.1. Framework overview

Fig. 1 displays a proposed framework for mining associative PoI patterns from Flick photos. The framework collects datasets using Flicker API. They are preprocessed and cleaned to remove duplicates and formatted for clustering, and remove the dominance of heavy-photo takers (those who take many photos in one place). Preprocessed datasets are fed into a clustering approach (DBSCAN in this study) to identify PoI. Association rules mining is in place for clustered PoI to find association rules which reveal strong associations between/among PoI. Detected rules are displayed with 3D association visualization (Wong, Whitney, & Thomas, 1999) and also along with Google Earth (http://www.earth.google.com).

3.2. Study region

Tourism is one of biggest industries in Australia, and Queensland (the second largest state) is a high profile tourist