Proceedings of the International Conference on Computer and Information Science and Technology
Ottawa, Ontario, Canada, May 11 – 12, 2015
Paper No. 113

Trip Pattern Mining Using Large Scale Geo-tagged Photos

Zahra Farzanyar, Nick Cercone
York University, Department of Computer Science and Engineering
4700 Keele Street, Toronto, Ontario, Canada M3J 1P3
zfarzan@cse.yorku.ca; ncercone@yorku.ca

Abstract- Photo-sharing websites allows people to display their experiences on the Web through rich media data such as photos and videos. These photos contain spatial context in terms of latitude and longitude where the photo was taken. The geotagged photos disclose much information about people travel behavior and tourist density. As web-based and mobile-based technologies advance, geo-tagged photos are increasingly collected beyond the capability of human analysis; so, have provided new research opportunities and challenges. In this work, we study on analyzing the travel behaviors based on the geo-tagged photos. This research proposes a framework that is used to organize a large collection of geotagged photos collected from Flickr to extract trip patterns from geo-tagged photos in Europe. This framework includes two data mining techniques based on a MapReduce framework: clustering, and frequent pattern mining. We explain using these techniques to analysis a large photo collection, while also revealing various interesting relations about popular cities and landmarks at Europe scale. We report interesting experimental results.

Keywords**:** Clustering; Frequent pattern mining; Geo-tagged photos; Photographer behavior model; Trip planning

1. Introduction

In recent years, the popularity of photo capture devices and the advent of photo sharing communities on Internet have led to huge digital photos with geographical references on the Internet. These Photo sharing sites allows people to display their experiences on the Web through rich media data such as photos and videos. These photos contain textual information such as title, notes and also are tagged with temporal context (i.e., time at which the photo was taken) and spatial context (i.e., in terms of latitude and longitude) where the photo was taken.

Flickr (Web-1) is one of the most popular photos sharing websites which is a great resource for travelers. Flickr has recently launched its own service for adding latitude and longitude information to a picture and provides the tool that admits a user to show pictures on online maps like OpenStreet-Map (Web-2). The geotagged photos disclose much information about people travel behavior and tourist density. Since, the photographed geolocations are good recommendations in terms of finding interesting locations; therefore, the tourists' photo-taking patterns are very important to identify the favorite places and trip patterns for tourism related organizations such as tour guide and recommendation systems.

As web-based and mobile-based technologies advance, geo-tagged photos are increasingly collected beyond the capability of human analysis as shown by Lee and Torpelund-Bruin (2011). So, Photos that form a huge ratio of information available on the Web, and are added every second, have provided new research opportunities and challenges.

In this work, we study on analyzing the travel behaviors based on the geo-tagged photos. To analysis the travel behaviors, several challenges need to be considered. First, with the exponential growth of online photos volume towards a terabyte or more, it has been more difficult to analyse them on a single sequential machine. To overcome this problem, we use a cloud based framework based on hadoop to develop a scheme to analyze the tourist travel patterns. Second, online photos are noisy, therefore, it is necessary to preprocess database of geotagged photos. Third, tourist behavior is based on visiting

different places; and a suitable model is needed to analyze such data. To analyze the tourist behavior, we use an Apriori based algorithm. To investigate the places, we perform mean-shift clustering on geotagged photos, in which latitude-longitude values are adopted as metric for locations.

Overall, this work focuses structural and practical aspects of the tourist travel patterns mining within Europe. Experiments show that the proposed approach can deliver promising results. The remainder of the paper is organized as follows. We briefly review related work in Section 2 and the preliminaries are investigated in Section 3. Then, we elaborate on the details of proposed approach and Experiments in Section 4. Finally, we conclude this work and give discussion of future work in Section 6.

2. Related Works

The last few years, the use of Flickr as a source of information has been a hot research topic to explore common wisdom in photo sharing community to find popular landmarks and to realize perception in the behavior of tourists. The research areas related to our work are geolocation recommender systems and geotag-based applications. We consider the latest reports that address these areas.

The idea of using Flickr to find points-of-interest (PoIs) has already been studied in some works. As shown by Crandall et al. (2009) the PoIs are determined which tourists have found most interesting. It uses classification methods to analyze Flickr geo-tagged photos. Zheng et al (2012) investigate regions of attractions that are similar to PoI, and use them for route analysis which focuses on analyzing tourist movement patterns. Kisilevich et al. (2010) change DBSCAN to find out PoI from Flickr photos. It is adaptive and flexible but only limited to PoI mining. Cao et al. (2010) deals with the problem of determining PoIs whose visual features resemble that of a photo or textual description provided by the user. Kennedy et al. (2007) use the concept of representative tags and tag-driven approach to extract place and event semantics. Rattenbury et al. (2007a, b) produce similar research and investigate ways to extract place and event semantics from folksonomy.

By Lee et al. (2014), associative PoI patterns are detected from Queensland Australia. The difference between our work and theirs is that ours generates frequent PoI patterns on a very large scale environment like Europe continent; where produces a big dataset. Most of works find well-liked trips within a city from data in user-generated photographic collections. Consequently, these studies covered only one or two tour destinations. In contrast, the proposed approach mines the travel information from Internet photos on Europe continent by using cloud environment, which renders the data acquisition highly efficient, and thus, allows the travel analysis to easily scale up to a multitude of destinations. This paper concentrates on the identification of PoI and relationship mining among PoI on Europe continent. Our work concenters on practical aspects of Flickr mining on the cloud environment rather than technical and algorithmic aspects. The contributions of the paper are; first, proposing a MapReduce based mining framework for PoI patterns. It first finds PoI using clustering on Hadoop and applies our previous algorithm by Farzanyar and Cercone (2013 a, b) for frequent pattern mining to detect PoI patterns from a big dataset on the cloud. Second, we analyze geo-tagged photos from Flickr for Europe continent. This study uses cloud based mining algorithms.

3. Preliminaries

With the exponential growth of data volume towards a terabyte or more, it has been more difficult to mine them on a single sequential machine. Researchers try to parallelize data mining algorithms to speed up the mining of the ever-increasing sized databases. While the parallelization may improve the mining performance, it also raises several issues for solution including load balancing, jobs assignment and monitoring, data partition and distribution, parameters passing between nodes, etc. To overwhelm this problem, the *MapReduce* framework as shown by Dean and Ghemawat (2004), one of most important techniques for cloud computing, has been introduced. Hadoop is an open source implementation of Google MapReduce architecture, sponsored by Apache Software Foundation; for efficiently writing applications which process huge amount of data in-parallel on large clusters of commodity hardware in a reliable and fault-tolerance manner (Web-4).

The sequential computing algorithms need to be redesigned to *MapReduce* algorithms. Converting a serial data mining algorithm into a distributed algorithm on the MapReduce framework might not be difficult, but the mining performance might be unsatisfactory as shown by Farzanyar and Cercone (2013a). In this work, we have used mean-shift algorithm based on MapReduce framework to find locations. We also use our proposed *MapReduce based* Apriori algorithm as shown by Farzanyar and Cercone (2013 a, b) to find trip patterns.

3. 1. Mean-Shift Algorithm

Clustering is the task of grouping objects in such a way that objects in the same group is more similar to each other than to those in other clusters as shown by Comaniciu and Meer (2002).
K-means clustering as shown by Hartigan and Wong (1979) and Mean-shift clustering have different applications in the geospatial context. K-means is better suited for location optimization problem or the number of clusters is known, whereas mean-shift is better suited for finding geospatial aggregations in the presence of noise points. In this work, we are interested in geospatial information of geo-tagged photos and the number of clusters is unknown. Thus, mean-shift is used as a default clustering approach in this research. Given n data points x_i, i = 1, ..., n on a d-dimensional space R^d, the multivariate kernel density estimate acquired with kernel K(x) and window radius h is

$$f(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \quad (1)$$

For radially symmetric kernels, it serves to specify the profile of the kernel k(x) satisfying

$$K(x) = c_{k,d}\, k(\|x\|^2) \quad (2)$$

$C_{k,d}$ is a normalization constant which ensures K(x) integrates to 1. The modes of the density function are set at the zeros of the gradient function $\nabla f(x) = 0$. The gradient of the density estimator (1) is

$$\nabla_{f(x)} = \frac{2c_{k,d}}{nh^{d+2}}\sum_{i=1}^{n}(x_i - x)g\left(\left\|\frac{x-x_i}{h}\right\|^2\right) = \frac{2c_{k,d}}{nh^{d+2}}\left[\sum_{i=1}^{n} g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)\right]\left[\frac{\sum_{i=1}^{n} x_i g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)} - x\right] \quad (3)$$

Where g(s) = $-k'$(s). The first term is proportional to the density estimate at x calculated with kernel$G(x) = c_{g,d}\, g(\|x\|^2)$ and the second term is the mean shift. The mean shift vector always points toward the direction of the maximum raise in the density. The points which are in the same basin of attraction are related with the same cluster. In this work, we use Apache Mahout library for Mean-shift algorithm that is based on MapReduce framework.

$$m_h(x) = \frac{\sum_{i=1}^{n} x_i g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{x-x_i}{h}\right\|^2\right)} - x \quad (4)$$

3. 2. Apriori Algorithm

Frequent itemset mining plays an important role in mining associations, correlations, emerging patterns, and other data mining tasks. The Apriori algorithm as shown by Agrawal and Srikant (1994) is one of most well-known methods for mining frequent itemsets in a transactional database. It uses a "bottom up" approach, frequent subsets are extended one item at a time, and groups of candidates are tested against the data. The algorithm terminates when no more candidate itemsets can be created. Converting a serial Apriori-like mining algorithm into a distributed algorithm on the MapReduce framework might not be difficult, but the mining performance might be unsatisfactory as shown by Farzanyar and Cercone (2013a). To redesign a serial Apriori-like mining algorithm into the MapReduce framework, the multiple-pass feature of the Apriori algorithm needs multiple MapReduce phases. The

master node must schedule jobs to initialize each MapReduce phase. The map function of a MapReduce phase can begin after all the reduce functions of its previous phase complete. Nodes that terminate their reduce functions must wait for all the unfinished nodes to done. The scheduling and the waiting are pure overheads to the mining task. In this paper, we use our proposed algorithm, IMRApriori, as shown by Farzanyar and Cercone (2013 a,b). It finds all frequent itemsets by using two MapReduce phases in a time and communication efficient manner.

4. The Proposed Frame Work for Trip Pattern Mining

Fig. 1 shows our proposed framework for trip patterns mining from Flick photos. Since, we want to find unbiased trip patterns in Europe; we use a large sample of geotagged photos in 5 years. The framework collects datasets using Flicker API. The case study presented in this article studies geo-tagged photos in Europe, in the years of 2008–2013. To do this, we defined place_type="continent", place_type_id="29" and woe_name="Europe" in our query from Flickr. We download metadata (photo_id, photographer_id, geolocation).

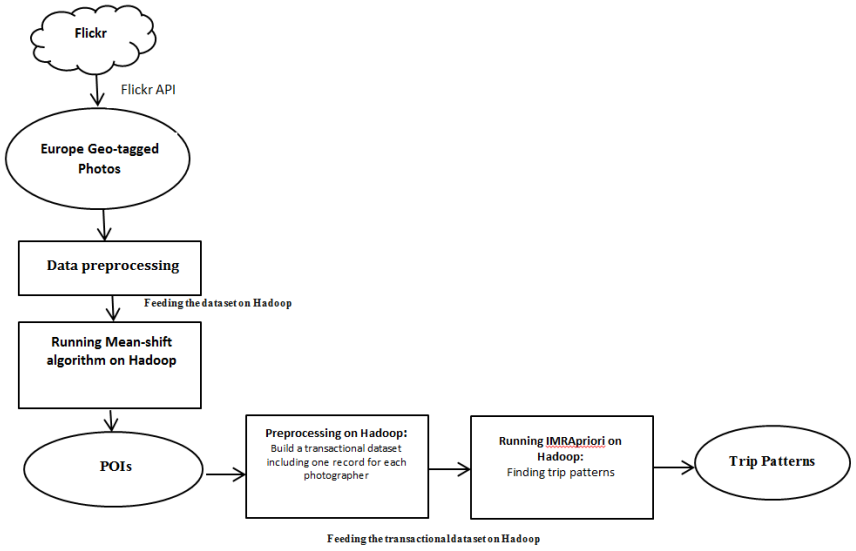

Fig.1. Trip pattern mining Framework

We want to find often-photographed places in the Europe to recommend popular places. Extracting highly-photographed places can be considered as a clustering task in a two-dimensional space. There are a huge number of photos in this case that leads to a huge dataset. The raw geo-tagged photos are repetitive. Photos should be pre-processed to consider the number of distinct photographers. Multiple photos taken by one user in the same place in the same day are viewed as the same photo. For this purpose, the dataset are fed into Hadoop to be pre-processed and preprocessed datasets are fed into *mean-shift* algorithm to identify PoIs. Frequent pattern mining algorithm for clustered PoIs is used to find frequent patterns which reveal strong relations among PoIs.

4. 1. Finding Point of Interests

In total, there are 9,205,326 public records in the study region. Fig. 2 depicts the distribution of the dataset using Google Earth 7.1.1 (Web-3). In our dataset, there is an unobservable probability distribution of where people take photographs, with modes based on interesting places to photograph. We are able to see the spots at which people take photos, from which mean shift algorithm estimates the modes of the underlying distribution. As shown by Crandall et al. (2009) the mean shift approach is well-suited to highly multimodal probability density functions with very different mode sizes and no known functional form, such as we have here.

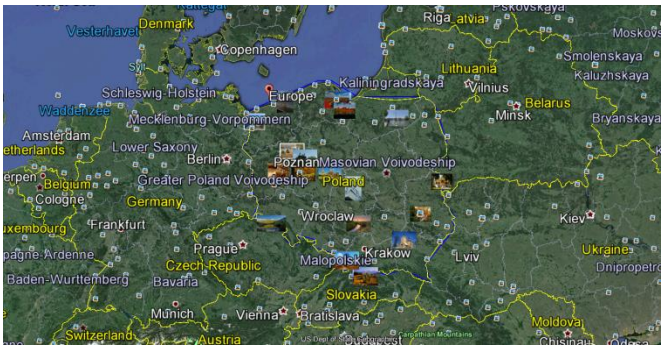

Fig. 2.Geotagged photos in Europe

In order to find the number of distinct photographers, photos are preprocessed to remove replicates and formatted for clustering. We preprocess dataset in one MapReduce phase. Fig. 3 shows the form of Geo-tagged photo data. The compound of latitude and longitude is used for spatial clustering.

photo id	owner	Latitude	Longitude

Fig. 3. Geo-tagged photo data form

Due to the hierarchical nature of geospatial data, we apply clustering in two levels to find PoIs in the continent level and city level. At first, Clustering extracts PoIs in the continent level by using the entire dataset. Then the clusters are further divided into city level PoIs by using finer mean-shift parameter values. We use the latitude-longitude values in degrees for each photo which is proposed byCrandall et al. (2009), considering them as points in the Europe continent, 1 degree for continent level (100 km) and .01 degree for city level (1 km). In the pre-processing phase, at a given scale, for each photographer we sample a single photo from each latitude-longitude value. We then perform the mean shift procedure at each scale. We identify the significance of each PoIs by simply counting the number of distinct photographers who took photos at that location. Table1 shows the 12 most photographed continent-scale PoIs on Europe found via MapReduced based mean shift algorithm, ranked according to number of distinct photographers. In this case, 1,638,312 out of 9,205,326 are clustered and 7,567,014 are assumed as replicates.

Table 1. The Clustering results at the Continent level

Continent level POI	Users	Photos
London	924567	3767265
Paris	434156	1642181
Berlin	124566	854080
Rome	85654	584705
Amsterdam	45675	496771
Vienna	29343	409987
Stockholm	23335	369072
Dublin	18544	352738
Brussels	12343	274676
Oslo	9876	186796
Zurich	5432	134043
Athens	5321	133012

In order to extract city level interesting locations in photo collections, we conducted clustering with .01 degree for city level (1km) in order to divide continent level PoIs into smaller city level PoIs. Table 2 presents the five most photographed landmarks in each of the top 12 cities.

Table 2.The five most photographed landmarks, found using MapReduced based mean-shift clustering.

	1st landmark	2nd landmark	3rd landmark	4th landmark	5th landmark
London	Tower Hill	Whitehall	Trafalgar Square	South Bank	Embankment
Paris	Saint-Germain-L'auxerrois	Gros Caillou	Les Iles	Vendôme	Saint-Merri
Berlin	Moabit	Hansaviertel	Alt-Treptow	Friedrichsberg	Südende
Rome	Campo Marzio	San Paolo	Trastevere	Pinciano	Appio Latino
Amsterdam	Jordaan	Watergraafsmeer	Buitenveldert	Bos en Lommer	Nieuwendam
Vienna	Hofburg	Innere Stadt	Staatsoper	Schonbrunn	Wieden
Stockholm	Norrmalm	Old Town	Djurgården	Riddarshomen	Södermalm
Dublin	Ringsend	Ballsbridge	Rathmines	Kilmainham	Phibsborough
Brussels	Pl. de Brouckere	Grand' Place	Sint-Joost-ten-Noode	Etterbeek	Eeuwfeestwijk
Oslo	Vika	Hammers	Vaterland	St.hanshaugen	Bygdøy
Zurich	Oberstrass	Enge	Aussersihl	Industriequartier	Fluntern
Athens	Makrygianni	Syntagma	Areopagos Hill	Pláka	Monastiraki

4. 2. Trip Pattern Mining

We use a frequent pattern mining to find significant relations among PoIs. At first, we must build a transaction dataset to find trip patterns in both continent level and city level. To do this, we first sample a photographer id randomly from the space of Flickr photographer id numbers, search the corresponding photographer on all PoIs in both levels, and build a record for this user in each level. We iterate the entire process for another randomly sampled user id number, keeping track of users who have already been considered so that their photos are not re-crawled.

Finally, there are a transactional dataset in continent level and a transactional dataset for each city. The transactional datasets are fed into Hadoop to identify frequent patterns by using *IMRApriori* as shown by Farzanyar and Cercone (2013a, b). This subsection reports some interesting frequent patterns in continent level and city level.

A user-specified minimum support is used to find frequent patterns. We use minimum support = 1%.

Frequent patterns are those k-itemsets that persuade the user-specified minimum support. Top 7 frequent 1-itemset of the continent level PoIs and their support are shown in Table 3 (a). Frequent 2-itemsets display more associative patterns. Table 3 (b) presents top 7 frequent 2-itemsets of the continent level PoIs. Top 5 frequent 1-itemset of city level PoIs in London that is top landmark are shown in Table 4 (a). Table 4(b) and Table 4(c) show frequent 2-itemsets and 3-itemsets of city level PoIs in London consecutively.

Table 3. Continent level PoIs (a) Top 7 frequent 1-itemset (b) Top 7 frequent 2-itemset (c) Frequent 3-itemset

(a)

Itemsets	Support (%)
London	41
Paris	38
Berlin	22
Rome	19
Amsterdam	14
Vienna	11
Stockholm	8

(b)

Itemsets	Support (%)
London, Paris	14
London, Berlin	11
London, Amsterdam	8
Paris, Rome	7
Paris, Amsterdam	5
Berlin, Vienna	3
Amsterdam, Stockholm	2

(c)

Itemsets	Support (%)
London, Paris, Rome	3
London, Paris, Vienna	1
Paris, Berlin, Rome	1

Table 4. City level PoI in London (a) Top 5 frequent 1-itemset (b) Top 5 Frequent 2-itemsets (c) Frequent 3-itemsets

(a)

Itemsets	Support (%)
Tower Hill	38
Whitehall	22
Trafalgar Square	16
South Bank	10
Embankment	8

(b)

Itemsets	Support (%)
Tower Hill, Whitehall	20
Tower Hill, Trafalgar Square	16
Tower Hill, Embankment	10
Whitehall, South Bank	5
South Bank, Embankment	3

(c)

Itemsets	Support (%)
Tower Hill, Whitehall, Embankment	5
Tower Hill, Trafalgar Square, South Bank	2

5. Conclusion

Understanding Photographer behaviour could be useful in related organizations such as tour guide and recommendation systems. In this paper we present a framework to automatically recognize places that people find attractive to photograph at both city and landmark scales. In this framework we introduce techniques for analyzing a global collection of geo-tagged photos from Flickr. The scalability of our methods based on Hadoop platform enables for automatically mining the trip patterns in very large sets of images. This study specifically, focuses on Europe, one of the hottest tourist destinations in the world. The findings demonstrate the value and applicability of the proposed framework. Development of the study area to the all of world is a future study. Sequence analysis and trajectory analysis of travelers is also the next step to study.

Acknowledgements

Thanks to CIV-DDD and NSERC Discovery Grant and NSERC Strategic Project Grant 430214.

References

Agrawal R. and Srikant R. (1994) Fast Algorithms for Mining Association Rules in Large Databases. In: Proceedings of the Twentieth International Conference on Very Large Databases (VLDB), pp. 487-499.

Comaniciu, D. & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(5), 603–619.

Crandall D. J., Backstrom L., Huttenlocher D., and Kleinberg J. (2009) Mapping the world's photos. In Proceedings of the 18th International Conference on World Wide Web, pages 761–770.

Cao L., Luo J., Gallagher A., Jin X, Han J, and Huang T. S.. (2010) A worldwide tourism recommendation system based on geotagged web photo. Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, pages 2274–2277.

Dean J. and Ghemawat S., (2004) Mapreduce: Simplified Data Processing on Large Clusters. In: Proceedings of the Sixth Symposium on Operating System Design and Implementation (OSDI), pp. 137-150.

Farzanyar, Z., and Cercone N. (2013) "Efficient mining of frequent itemsets in social network data based on MapReduce framework." Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. ACM.

Farzanyar Z, and Cercone N. (2013)"Accelerating Frequent Itemsets Mining on the Cloud: A MapReduce-Based Approach." Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on. IEEE.

Hartigan, J. A. & Wong, M. A. (1979). "Algorithm AS 136: A K-Means Clustering Algorithm". Journal of the Royal Statistical Society, Series C **28** (1): 100–108. JSTOR 2346830.

Kisilevich, S., Mansmann, F., & Keim, D. A. (2010). P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. In Proceedings of the 1 st International Conference and Exhibition on Computing for Geospatial Research & Application, New York, NY: ACM.

Kennedy, L., Naaman, M., Ahern, S., Nair, R., & Rattenbury, T. (2007). How Flickr helps us make sense of the world: Context and content in communitycontributed media collections. In Proceedings of the conference on multimedia (pp. 631–640). New York, NY: ACM.

Lee, I., & Torpelund-Bruin, C. (2011). Geographic knowledge discovery from Web 2.0 technologies for advance collective intelligence. Informatics, 35, 453–461.

Lee, Ickjai, Guochen Cai, and Kyungmi Lee. (2014) "Exploration of geo-tagged photos through data mining approaches." Expert Systems with Applications 41.2: 397-405.

Rattenbury, T., Good, N., & Naaman, M. (2007). Towards extracting Flickr tag semantics. In Proceedings of the 16th international conference on World Wide Web (pp. 1287–1288).

Rattenbury, T., Good, N., & Naaman, M. (2007a). Towards automatic extraction of event and place semantics from Flickr tags. SIGIR, 103–110.

Zheng, Y-T., Zha, Z-J., & Chua, T-S. (2012). Mining travel patterns from geotagged photos. ACM Transactions on Intelligent Systems and Technology, 3(3), 56.

Web sites:

Web-1: http://www.flickr.com consulted 1 Dec. 2014.

Web-2: http://www.openstreetmap.org consulted 20 Sep. 2014.

Web-3: http://www.google.com/earth/ consulted 14 Nov. 2014.

Web-4: http://hadoop.apache.org/ consulted 18 Aug. 2014.