P-DBSCAN: A density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos

Slava Kisilevich
University of Konstanz
slaks@dbvis.inf.uni-konstanz.de

Florian Mansmann
University of Konstanz
florian.mansmann@uni-konstanz.de

Daniel Keim
University of Konstanz
keim@dbvis.inf.uni-konstanz.de

ABSTRACT

The rapid spread of location-based devices and cheap storage mechanisms, as well as fast development of Internet technology, allowed collection and distribution of huge amounts of user-generated data, such as people's movement or geo-tagged photos. These types of data produce new challenges for research in different application domains. In many cases, new algorithms should be devised to better portray the phenomena under investigation. In this paper, we present P-DBSCAN, a new density-based clustering algorithm based on DBSCAN for analysis of places and events using a collection of geo-tagged photos. We thereby introduce two new concepts: (1) *density threshold*, which is defined according to the number of people in the neighborhood, and (2) *adaptive density*, which is used for fast convergence towards high density regions. Our approach is demonstrated on the area of Washington, D.C.

Categories and Subject Descriptors

H.2.8 [**Database Mangement**]: Database Applications—*Data mining, Spatial databases and GIS*; I.5.3 [**Computing Methodologies**]: Clustering

Keywords

Density based clustering, geo-tagged photos, attractive places

1. INTRODUCTION

In recent years, we witness the emergence of new types of spatio-temporal data such as GPS, mobile networks, sensors, geo-tagged images. Location-acquisition technology is becoming very popular among ordinary people, allowing them to record their positions, while the Internet allows them to share their data with others [9]. As a consequence, spatio-temporal user-generated data has become available for research in large amounts. The characteristics of these data pose new challenges in different research domains and demand for new analytical approaches. For example, Flickr[1] and Panoramio[2], photo-sharing websites, contain millions of geo-tagged photos contributed by people from all over the world and annotated with different kinds of important information like time, location (where the photo was taken), title and tags. The significance of these data has been already addressed in the following papers as landmark identification [13, 4], tag representative [1] and representative images [18]. However, since most of the research concentrates on application-based perspectives, systematic approaches for analysis of spatial or spatio-temporal aspects of geo-tagged photos only begin to emerge [8, 12, 15, 14].

The goal of this work is to propose a novel clustering algorithm for analysis of places and events using collections of geo-tagged photos. In particular, we would like to find interesting places or significant events that are characterized by high photo activity in a specific area. Identification of such places and events will be beneficial, for example, for local authorities, urban planners and service providers. Finding interesting places is a clustering problem that can be solved using different clustering algorithms, but density based clustering algorithms are more advantageous for our problem because of several aspects: (1) high photo activity can be measured by density and compared to different parts of the region under investigation, (2) the number of clusters is not known in advance, (3) clusters have arbitrary shape, and (4) sparse regions are treated as noise. DBSCAN [6], DENCLUE [10] and OPTICS [2] are examples of density based clustering algorithms. However, we argue that all existing algorithms have one serious problem with respect to the geo-tagged data: these algorithms work with generic points having equal "importance", while in geo-tagged data, owners of photos are those who determine the importance of a cluster. Consider an example presented on Fig. 1. The photo labeled as 1, where the number represents the photo's owner, is located at the center of the radius and has 6 photos in its neighborhood. When a generic density based clustering algorithm is applied on the data depicted in the left part of Fig. 1 with appropriate parameters, it would create a dense cluster. However, this cluster contains only photos taken by one person. The right illustration shows the same photo configuration but every photo belongs to a different owner. Clearly, this cluster is more significant than the cluster on the left side. Since such cases are not han-

[1]http://www.flickr.com
[2]http://www.panoramio.com

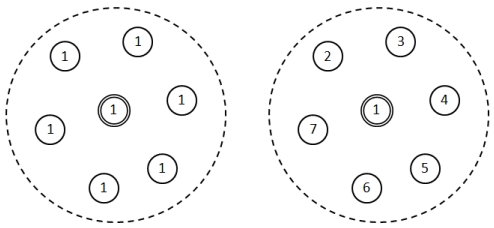

Figure 1: Problem illustration

dled by generic density based clustering algorithms, there is a need for a new specialized algorithm. We expect that clusters obtained using this new density definition should be more robust. Additionally, in many cases, the obtained clusters can have different densities in different parts of the cluster. Therefore, we introduce a notion of *adaptive density* to handle such cases. The basic idea is to split the cluster if different local areas of the cluster have large differences in density. The splitting should create small "packed" clusters in which density does not vary much.

In this paper, we present P-DBSCAN, a new density-based clustering algorithm, where the P stands for *photo*. It is a variation of DBSCAN for analysis of places and events using a collection of geo-tagged photos. Two novel concepts are introduced, namely (1) the *density threshold*, which is defined by the number of people in the neighborhood, and (2) *adaptive density* for fast convergence towards high density regions.

2. RELATED WORK

Density-based clustering methods first established a little more than a decade ago [6, 2]. The notion of density-connectivity presented in [6] served as a starting point for the number of density-based algorithms like DBSCAN [6], OPTICS [2], LDBSCAN [5] to name a few. The basic idea of DBSCAN-based algorithms is that every point in a database should contain a minimum number of *MinPts* points in its neighborhood of radius ϵ. Improvements suggested in later research aimed at generalization of clustering approaches [10], efficient selection of input parameters [2], solving the problem of local densities [5] or introducing a specialization for a particular task, such as moving clusters [11], trajectory clustering [16], spatio-temporal analysis of seawater characteristics [3], seismic activity [17], etc. Visualization of concentration of tourists using grid-based clustering was performed in [7]. Representative landmark images were found on the city and country scales in [4] combining coordinates of geo-tagged photos with content based and textual analysis using MeanShift algorithm based on kernel-density estimation for clustering. [15] presented a framework for visualization of attractive places using DBSCAN for clustering and attaching weight to every photo using kernel density estimation for mapping the weight to a color.

3. PROBLEM FORMULATION

Following the terminology of the original work on DBSCAN [6], we provide our basic definitions of P-DBSCAN with respect to the new definition of density based on the number of people (owners of photos).

Definition 1. The **neighborhood** of a photo point p, denoted as $N_\epsilon(p)$, is defined by

$$N_\epsilon(p) = (q \in D, Owner(q) \neq Owner(p) | Dist(p,q) \leq \epsilon)$$

where $(o_{i_p} \in O) = Owner(p)$ is an ownership function and $Dist(p,q)$ is distance between points p and q. We require that for a photo point p, which belongs to the owner o_i, we find at least one point q whose owner is not o_i in a neighborhood of radius ϵ.

Definition 2. A **core photo** is a photo point where at least a minimum number of owners *MinOwners* not including the owner of the photo p took photos in the neighborhood of the photo p.

Definition 3. A photo point q is **directly ownership-reachable** from a point p when $q \in N_\epsilon(p)$.

Definition 4. A photo point p is **ownership-reachable** if there is a chain of photo points p_1, $p_2 \ldots, p_n = p$ such that p_{i+1} is directly ownership-reachable from p_i.

Definition 5. A photo is a **border photo** when it is not a core, but ownership-reachable from a core photo point.

Definition 6. Adaptive density (drop) threshold is defined as the ratio of the current density of a photo point p according to the Def. 3 and the previous density. The neighbors of the photo are assigned to the current cluster until the density ratio is greater than 1 (density increase) or greater than density threshold (density drop).

4. METHOD

In this section we describe P-DBSCAN algorithm. Fig. 1 and Fig. 2 shows the pseudocode of P-DBSCAN.

Input: D - dataset of points with coordinates and ownership attributes, ϵ - neighborhood radius, Ad - adaptive density flag, Addt - adaptive density drop threshold
Output: Set of clusters

```
1  cluster-id = 0
2  while ((p = getUnprocessedPhoto(D)) ∉ ∅) do
3  |   CurrentDensity = Addt
4  |   if (|Neighborhood(p)| < MinOwners) then
5  |   |   MarkPhotoAsNoise(p)
6  |   else
7  |   |   cluster-id = cluster-id + 1
8  |   |   AssignPhotoToCluster(p,cluster-id)
9  |   |   UniqueQueue(Q,GetNeighborhoodPhotos(p))
10 |   |   while (Q is not empty) do
11 |   |   |   p = DeQueue(Q)
12 |   |   |   AssignPhotoToCluster(p,cluster-id)
13 |   |   |   if (|Neighborhood(p)| >= MinOwners) then
14 |   |   |   |   if (Ad == true) then
15 |   |   |   |   |   AdaptiveDensity(...)
16 |   |   |   |   else
17 |   |   |   |   |   UniqueQueue(Q,GetNeighborhoodPhotos(p))
18 |   |   |   |   end
19 |   |   |   end
20 |   |   end
21 |   end
22 end
```

Algorithm 1: P-DBSCAN

The algorithm starts with arbitrary photo that is not yet

assigned to any cluster and not defined as noise. If the photo is not core according to Def. 3, it is marked as noise (line 1.5). If the photo is a core, it is assigned to the current cluster and all the neighbors of the photo are queued for further processing (line 1.9), skipping the photos that were already processed or that are already in Q. The processing and assignment of photos to the current cluster continues until the queue is empty (line 1.10). The next photo is retrieved from the queue and assigned to the current cluster. Then, its neighborhood is checked to determine if the photo is core. If it is a core photo and the adaptive density version of the algorithm is running, the function AdaptiveDensity is invoked (line 1.15), otherwise, neighbors of a photo p are queued (line 1.17). When the adaptive density version of the algorithm is running (Alg. 2), several additional conditions are checked. The number of owners in the neighborhood of the point p is checked against the current density. If the number of owners in the neighborhood is equal or greater than the current density, the neighbor photos of the photo p are queued and the current density is updated (line 2.8). If the number of owners in the neighborhood is less than the current density, the adaptive density drop threshold is checked. If the number of owners in the neighborhood drops below the threshold, the neighbors of the point p are not processed, otherwise the neighbor photos of the photo p are queued and current density is updated (line 2.5).

```
   Input: p
 1 DensityDrop = |Neighborhood(p)|/CurrentDensity
 2 if |(Neighborhood(p)| < CurrentDensity ) then
 3 |   if (DensityDrop >= Addt) then
 4 |   |   CurrentDensity = |(Neighborhood(p)|
 5 |   |   UniqueQueue(Q,GetNeighborhoodPhotos(p))
 6 |   end
 7 else
 8 |   CurrentDensity = |(Neighborhood(p)|
 9 |   UniqueQueue(Q,GetNeighborhoodPhotos(p))
10 end
```

Algorithm 2: P-DBSCAN with adaptive density

5. DATASET

We collected metadata of geo-tagged photos from Flickr using its publicly available API. Since the API does not allow downloding of metadata for a particular region, the downloading was performed as follows: an initial user id was used to download his photo metadata. Then, we downloaded all the user's contacts. To speed up the process of retrieving heterogeneous users, we retrieved all groups to which the user belongs, and using group information we were able to retrieve all the people who belong to these groups. This process was applied again on other users. We collected $86,314,466$ entries from $4,137,248$ users to the time of writing this paper. Among different attributes available in the metadata, location, photo id and owner id are the only attributes relevant in this work. During the data collection process, we also converted coordinates of photos expressed in degrees into UTM (Universal Transverse Mercator) coordinate system to save computation time on applying distance measures.

6. EVALUATION

For the experimental evaluation, we concentrated on the area of Washington, D.C. spanning 306 km^2. We retrieved $151,564$ photos that were taken in year 2007. From this set, we removed photos having the same coordinate (regardless of the owner), leaving only one photo. Finally, we were left with $28,707$ photos from $4,160$ owners.

In the first evaluation, we applied $MinPts = 150$ (DBSCAN), $MinOwners = 150$ (P-DBSCAN) and $\epsilon = 30$. We used 10% for the density drop threshold. The goal of the evaluation is to compare DBSCAN and P-DBSCAN with respect to the clustering result. Fig. 2 shows the results of the clustering. The boundaries of the clusters were obtained using the PostgreSQL's Convex Hull spatial query. It can be seen that DBSCAN creates 3 regions: the region around Lincoln Memorial, Tidal Basin and around the center of the city. P-DBSCAN has created four regions: the first is around Lincoln Memorial but smaller than the region created by DBSCAN, the second region (Tidal Basin) spans almost the same territory as in DBSCAN, but having less photos and owners in a cluster. DBSCAN included 502 owners and 1232 photos, while the cluster created by P-DBSCAN included 476 owners and 1151 photos respectively. The third region is around National World War II Memorial, while the fourth region is around the remaining part of the city center. P-DBSCAN with adaptive density has created 5 "packed" regions around Lincoln Memorial, Vietnam Veterans Memorial, National World War II Memorial, Jefferson Memorial and Washington Monument. These places are the most visited among tourists. For example, among 502 owners that were assigned by DBSCAN to the cluster covering Jefferson Memorial, 203 people took photos in close vicinity of the Jefferson Monument.

In the second evaluation, we applied $MinOwners = 50$ (P-DBSCAN), $\epsilon = 50$ and 10% for the density drop threshold. The goal of this evaluation is to compare P-DBSCAN with respect to different parameters. Since, the neighborhood radius increased by 20 meters and the number of owners decreased by 100, 8 clusters were produced by P-DBSCAN and 40 clusters were created by P-DBSCAN with adaptive density (Fig. 3). Among the new discovered places by P-DBSCAN there are Marine Corps War Memorial (107 owners), Arlington House (181 owners), Tomb of the Unknowns (96 owners), Washington National Cathedral (109 owners), Dupont Circle (390 owners) and Union Station (179 owners).

The observations of the results support our assumption, that density based on ownership tends to create smaller clusters, while adaptive density leads to creation of small "packed" clusters with high density.

7. CONCLUSIONS

In this paper we presented a new clustering algorithm, based on DBSCAN, specialized for the problem of analysis of places and events using large collections of geo-tagged photos. We introduced two improvements to the original definition of DBSCAN. (1) We defined neighborhood density as the number of people who take photos in the area. (2) We proposed a notion of *adaptive density* for optimizing search for dense areas and faster convergence of the algorithm towards clusters with high density.

Figure 2: Washington, D.C. Clusters produced by DBSCAN (red), P-DBSCAN (yellow), P-DBSCAN with adaptive density (blue) using 150 photos as *MinPts* (DBSCAN) or *MinOwners* (P-DBSCAN) and $\epsilon = 30$ meters minimum radius and a density drop threshold of 10%

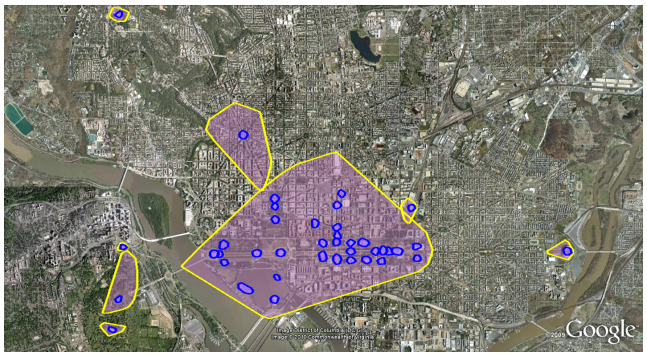

Figure 3: Washington, D.C. Clusters produced by P-DBSCAN (yellow), P-DBSCAN with adaptive density (blue) using 50 owners as *MinOwners* and $\epsilon = 50$ meters minimum radius and a density drop threshold of 10%

Different aspects of the proposed approaches were not mentioned in this paper as it is an ongoing research. In our future work we will concentrate on the evaluation approaches, runtime optimization, database integration and different analytical tasks.

Acknowledgements

This work was partially funded by the Priority Program (SPP) 1335 ("Visual Spatio-temporal Pattern Analysis of Movement and Event Data").

8. REFERENCES

[1] S. Ahern, M. Naaman, R. Nair, and J. Yang. World explorer: Visualizing aggregate data from unstructured text in geo-referenced collections. In *Proceedings of the 7th ACM/IEEE joint conference on Digital libraries*, pages 1–10, 2007.

[2] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. Optics: ordering points to identify the clustering structure. *SIGMOD Rec.*, 28(2):49–60, 1999.

[3] D. Birant and A. Kut. St-dbscan: An algorithm for clustering spatial-temporal data. *Data Knowl. Eng.*, 60(1):208–221, 2007.

[4] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *WWW*, 2009.

[5] L. Duan, L. Xu, F. Guo, J. Lee, and B. Yan. A local-density based spatial clustering algorithm with noise. *Inf. Syst.*, 32(7):978–986, 2007.

[6] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *Data Mining and Knowledge Discovery*, pages 226–231, 1996.

[7] F. Girardin, F. Calabrese, F. D. Fiore, C. Ratti, and J. Blat. Digital footprinting: Uncovering tourists with user-generated content. *Pervasive Computing, IEEE*, 7(4):36–43, 2008.

[8] F. Girardin, A. Vaccari, A. Gerber, A. Biderman, and C. Ratti. Quantifying urban attractiveness from the distribution and density of digital footprints. *International Journal of Spatial Data Infrastructures Research*, 4:175–200, 2009.

[9] M. Goodchild. Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0. *International Journal*, 2:24–32, 2007.

[10] A. Hinneburg and D. A. Keim. An efficient approach to clustering in large multimedia databases with noise. pages 58–65. AAAI Press, 1998.

[11] P. Kalnis, N. Mamoulis, and S. Bakiras. On discovering moving clusters in spatio-temporal data. *Advances in Spatial and Temporal Databases*, pages 364–381, 2005.

[12] D. Keim, P. Bak, S. Kisilevich, N. Andrienko, and G. Andrienko. Analysis of community-contributed space- and time referenced data by example of panoramio photos. *Vision, Modeling, and Visualization Workshop (VMV)*, 2009.

[13] L. Kennedy, M. Naaman, S. Ahern, R. Nair, and T. Rattenbury. How flickr helps us make sense of the world: Context and content in community-contributed media collections. *Proceedings of the 15th international conference on Multimedia*, pages 631–640, 2007.

[14] S. Kisilevich, D. Keim, and L. Rokach. A novel approach to mining travel sequences using collections of geotagged photos. *13th AGILE International Conference on Geographic Information Science*, 2010.

[15] S. Kisilevich, F. Mansmann, P. Bak, and D. Keim. Where would you go on your next vacation? - a framework for visual exploration of attractive places. *Proceedings of the GeoProcessing 2010*, 2010.

[16] A. T. Palma, V. Bogorny, B. Kuijpers, and L. O. Alvares. A clustering-based approach for discovering interesting places in trajectories. In *SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*, pages 863–868, 2008.

[17] M. Wang, A. Wang, and A. Li. Mining Spatial-temporal Clusters from Geo-databases. *Lecture Notes in Computer Science*, 4093:263, 2006.

[18] Y.-T. Zheng, M. Zhao, Y. Song, H. Adam, U. Buddemeier, A. Bissacco, F. Brucher, T.-S. Chua, and H. Neven. Tour the world: building a web-scale landmark recognition engine. *Proceedings of ICCV, Miami, Florida, USA*, 2009.