Reading Exam

Vaibhav Kulkarni

June 6th 2016

Contents

1 Motivation

In recent years, we have witnessed a proliferation of mobile devices with global positioning (GPS) functionality and internet connectivity. This has led to a rapid emergence and a notable progress in the development of Location-based Service (LBS). The typical examples of LBS include automotive traffic monitoring, network resource allocation, location based targeted advertisements and social networking. A large amount of user data is collected by these companies which goes into training user specific learning models to predict user mobility and behavioural patterns. Although LBS offer valuable services, revealing personal location data to potentially untrustworthy service providers raises several privacy concerns.

The heart of this thesis lies in augmenting the current mobility prediction learning frameworks in order to incorporate privacy awareness as a key paradigm. Meanwhile, researchers today use cryptographic techniques and Location Privacy Preserving Mechanisms (LPPM) for privacy preservation which are computationally complex and does not facilitate collaborative learning. We will also explore distributed means of computing user independent learning models in realtime, maintaining the utility-privacy space. According to recent surveys, 55% of LBS users have shown concern towards loss of their location privacy and about 50% of U.S. residents who have a profile on social networking sites are concerned about their privacy [6]. It is clear that the success of LBS depends on the location privacy in the near future.

Since the major component of this research lies in a privacy aware mobility prediction framework, we limit the literature survey to these two critical topics, mobility prediction and privacy aware computation. Firstly, it is important to understand to what extent, human mobility is predictable. To this end, we review a classical paper which provides insights into human mobility and the prediction limits. Once the GPS data is collected, it is necessary to understand how to extract the mobility patterns of individuals. We perform a thorough research in this area and review two papers which highlight how to discover mobility patterns and user profiles from raw GPS logs. Next, we review literature related to existing privacy preserving mobility prediction models. In this area, we present an article which depicts a technique to preserve location privacy maintaining the quality of services offered by the LBS. We highlight several limitations of the method explained. Finally, we survey machine learning techniques which can be utilised to construct user mobility models. In this area, we review a paper which presents an novel architecture of privacy-preserving distributed learning approach.

2 Limits of Predictability in Human Mobility [4]

Summary

Context

Human behaviour is characterised by spontaneity, randomness and change. This paper puts forth an important contribution in stating to what extent human mobility is predictable. The authors identify a metric, "Entropy" as a means to measure the potential predictability in user mobility. Although, the paper might seem trivial, it can be seen as a base research in order to make explicit predictions on user whereabouts.

Problem

The authors take an interesting approach to model human activity which is not stochastic based. There exist several probabilistic based models, as a result it is necessary to know the degree of randomness in human behaviour and the extent to which human mobility is predictable. Furthermore, it is important to know the bounds of predictability characterised by the mobility entropy, travel distance, frequency of visits, time spent at a particular location and heterogeneity of the visitation patterns. This papers fills this void by experimentally analysing the above bounds and concluding that human mobility is characterised by high regularity and this is predictable.

Contributions

The paper dismisses many of the common assumptions associated with human mobility prediction by experimental evaluation of several mobility datasets. Here lies the major contribution of the paper as it established bounds for various aspects of prediction.

- The article presents a technique to measure the entropy associated with user movement. The entropies can be classified in to random entropy (number of distinct locations visited by the user), temporal-uncorrelated entropy (characterises the heterogeneity of visitation patterns) and the actual entropy (accounts the order in which the nodes are visited). This distinction aids to arrive at the conclusion that user movement can be predicted irrespective of the entropy thus dismissing the general assumption that only lower entropy implies higher predictability.
- The authors evaluate the Fano's inequality bound on predictability when a user with a given entropy moves between N locations. It was discovered that, despite of the apparent randomness of the individual trajectories there exists a high degree of potential predictability in user movement.
- The analysis also led to a conclusion that, the predictability across a large user base is insignificant and varies from person to person. Further, it was also found that, users covering larger distances on regular basis are just as predictable as users commuting in a small area.
- Similar results were obtained when experiments were performed on diverse demographics of varying ages and genders, i.e. only insignificant variations were found in regularity. This concludes that regularity and thus predictability is not imposed by demographic factors, but instead by intrinsic human activities.
- The combination of the empirically determined user entropy by the authors and Fano's inequality leads to a potential 93% average predictability in user mobility.

Approach

A dataset representing the call patterns of 10 million mobile phone users was used containing the routing tower location. The data was filtered to have only the users with a sufficient calling frequency and high movement which was further characterised by individual call/motion activity. This data was processed to construct a time series for each user to determine their entropy, movement regularity and dependence on the demographic and population density.

Discussion

- The used dataset was collected in a high income country where having a 93% potential predictability in user mobility is acceptable despite very large differences in travel distances due to the regularity and availability of transportation modes. However, it will be interesting to characterise similar parameters in the low income and densely populated countries which face much more extreme conditions so as to generalise the findings.
- Although the authors calculated the bounds on predictability by defining entropy and movement regularity, the paper did not show how close to the maximum potential predictability, the accuracy of actual algorithms can come in practice.

3 Understanding Mobility Based on GPS Data [5]

Summary

Context

A considerable amount of research studies have focussed on detecting significant locations in the user trajectory data, predicting future locations of a user or recognising user specific activity at a particular location. On the contrary, classifying user GPS trajectories based on transportation modes has not received substantial attention. Knowledge regarding the transportation modes is significant in order to provide pervasive computing systems with meaningful context information.

Problem

To date, the transportation mode classification, relies on manual labelling by the users or utilised GSM radio signals, which could only discriminate between simple motions such as moving and being stationary. On the other hand, identification methods based on trivial approaches such as velocity based classification leads to inherent errors due to the frequency at which users switch the travel modes. The velocity of the travel modes is also influenced by traffic conditions and weather. In this paper the user GPS trajectories are used to understand and distinguish between users transportation mode, such as walking, driving or taking a bus based on supervised learning.

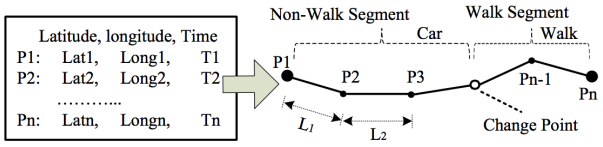

Figure 1: GPS logs and transportation mode prediction (From [5])

Contributions

The main contributions of the paper can be summarised as below:

- The papers presents a change point based segmentation technique to extract mobility trajectories from raw GPS logs.
- The technique presented in the paper to infer the mobility mode is independent of any additional database of road networks or Point of Interests (POI's). Therefore it can be deployed in a range of web applications.

Results

In order to infer the transportation mode, the authors identify a set of novel features beyond simple velocity and acceleration. These features along with a post processing algorithm are robust to the traffic condition and contain significant information of users motion. The technique was evaluated using GPS logs collected by 65 users over a time period of 10 months. The authors evaluated the proposed approach by considering various combinations of the suggested features and achieved a prediction accuracy of 72.8 %.

Approach

The framework to infer the transportation mode consists of an offline learning stage and a online processing stage. In the offline learning part, the GPS trajectories are partitioned into segments depending on the direction change points. The individual segments are than utilised to extract several features which are used to train a classification model for the online inference stage.

When the GPS trajectory is extracted, it is first partitioned into segments from which features such as direction change rate, velocity change rate, stop rate and heading direction are computed. Using these features, the inference model is utilised to predict the transportation mode for a given segment in a probabilistic manner. As a classification model, decision tree is used based on a change point based segmentation method. The change points are grouped into clusters using a density-based clustering algorithm. The derived clusters are used to construct an undirected graph with nodes as an individual cluster and edges being the transportation between nodes. A spatial index is built over the graph to improve the efficiency of accessing information of each node and edge. Finally, the probability distribution is calculated of different transportation modes on each edge.

Discussion

The contribution of the above paper is a methodology to infer the mobility mode, given raw GPS data logs. The non-user specific model presented in the paper is independent of external sensors such as GSM signals, heart rate and map information like road networks and bus stops. The change point based segmentation method devised by the authors is proved to be robust to variable traffic conditions. Thus, it can be deployed in a range of web applications.

- The technique presented, involves an offline training phase which uses the extracted features to generate the online transportation mode inference model. It would have been interesting to quantify the data required and computational time/complexity involved at the offline stage to arrive at a sufficient online prediction accuracy.
- The paper states that the devised approach is a non-user specific model. However, sufficient evaluation of the experimental results regarding generalisation of user specific models is not presented in the paper.
- The authors derive features from the raw GPS logs to train the inference model. Although, it is evident that a combination of certain features results in a sufficient prediction accuracy, a combination of a particular features reduces the accuracy. The reasoning behind such a behaviour is lacking in the evaluation section.
- Considering majority of the population owning a smart phone have Internet connection, it will be interesting to map the extracted segments to road maps and points of interests in the training phase to increase the prediction accuracy.
- The segmentation technique described, extracts segments from consecutive GPS points. However, there will exist some segments which do not offer any valuable information for feature extraction. It will also be interesting to extract the only segments providing valuable information based on other attributes.

4 Discovering SpatioTemporal Mobility Profiles of Cellphone Users [1]

Summary

Context

In order to develop applications, related to context-based search and advertising, inherent in Location Based Services, it is trivial to understand the mobility patterns and profiles of users. These patterns need to be extracted from the raw GPS traces which are available from the cellphones. Different users have distinct behaviours which influence their mobility patterns, as a result it is crucial to discover these factors which affect the mobility path information retrieval. This paper focuses on discovering the spatiotemporal mobility patterns and mobility profiles from cell phone location logs.

Problem

Today, the LBS providers continuously monitor their users and log their location information, in the form of GPS coordinates. However, extracting meaningful information out of these raw traces is a challenge due to the anomalies caused while tracking and user specific behaviour. On the contrary to the existing work which are restricted to small scale environments in order to study human mobility, this paper puts forth a framework "Mobility Profiler", for discovering user mobility patterns and user profiles at a city wide level using cellular networks.

Contributions

The main contributions of the paper can be summarised as below:

- The paper introduces formal definitions for the concepts of mobility path, mobility pattern and mobility profile and the factors influencing each.
- The authors design and implement the complete framework of Mobility Profiler to discover mobility profiles from raw data.
- The paper also presents several experiments conducted using the Reality Mining data set to determine realistic thresholds for when to consider location end times, interim location on a mobility path and others alike.
- The previous studies have concluded that typical users spend 85% of their time in 3 to 4 locations. The paper sheds light on user behaviour and patterns during the remaining 15% of the time.

Approach

The Mobility Profiler Framework consists of the following phases:

- Path Construction: Here, an ordered set of cell tower IDs corresponding to user's travel path is constructed. Cell clustering is further employed to eliminate oscillating cell towers and replace them with their corresponding clusters.
- Topology Construction: The extracted travel paths of cell clusters are used to construct topology of user movements.

- Pattern Discovery: Here, the frequent mobility patterns of each user are discovered. This step is carried out by employing the topology information and a string matching support criteria, i.e. for every subsequent pair of cell clusters in a sequence, the former one should be a neighbour of the latter one in the cell-cluster topology graph.
- Post Processing: The extracted personal mobility patterns are then used to generate cellphone user profiles.

Results

The paper presents several interesting results summarised as below:

- The authors find that the average duration spent in a cell is 10 min where it is defined as the duration between the cell end time and the cell start time. The average cell transition time was also found to be 10 min where it is defined as the time duration between the subsequent cell start time and the current cell end time. These values were experimentally found by analysing the ratio of cell span duration and cell span transitions, smaller than predefined time values in the experiment phase.
- In order to determine the right cluster in the case of oscillations, it is important to determine the minimum switching count. It was analytically found out that in order to distinguish between oscillations due to user mobility and cell tower oscillations the minimum switching threshold should be 3.
- Further, experiments were performed to discover patterns for generating both global and personal frequent patterns. It was found out that frequency of mobility paths is inversely correlated with the path-length.

Discussion

A single cell may encompass several points of interests which are masked by the resolution offered by logging only the network tower ID which covers a large area. As a result, the thresholds presented in the paper may not hold in practicality which consists of user mobility characterised by short movements and stay times. It would be interesting to implement and analyse, how these parameters change when it is applied to a dataset consisting of GPS logs mapped at a high frequency.

5 Hiding Stars with Fireworks: Location Privacy through Camouflage [2]

Summary

Context

There has been a rapid proliferation of Location Based Services (LBS) in recent years due to ubiquitous wireless connectivity and GPS modules integrated with smart phones. These LBS rely on accurate, continuous and realtime streaming of location data. However, revealing this information to service providers poses a significant privacy risk. In this paper, the authors device a method to preserve user privacy without trading off the quality of services offered by the LBS.

Problem

Existing research on user privacy protection in LBS takes the approach of obscuring user's path, compromising the accuracy of services offered by the LBS. Hiding parts of user's paths can lead to degrading the the spatial accuracy, increased delay in reporting user location or temporarily preventing the user from reporting locations completely. This leads to user data being less useful after enabling privacy protection. As a result a framework is needed which can protect the user against location tracking by the service providers at the same time offer high quality services.

Contributions

1. The paper proposes CacheCloak, which acts as an intermediary server between the users' and the LBS. The framework utilises mobility prediction in order to camouflage the user by requesting information for a series of predicted interweaving paths instead of a single GPS coordinate.

2. CacheCloak extends "Path Confusion" technique and fixes its trivial flaw which is the inherent delay caused while answering queries. Path prediction and prospective caching retains the benefits of path confusion without incurring the delay of path confusion.

3. The iterated Markov prediction model is robust to high number of mis-predictions as the user will always see only up-to-date cached data. Requesting new data for mis-predictions comes at a low cost compared to necessity of privacy.

4. The authors, evaluate the proposed framework considering a realistic attacker model. The diffusive method, models an attacker trying to follow every possible way the user might go considering different speeds and directions.

5. The paper proposes a quantitative measure of privacy based on the attackers ability or inability to track the user over time.

6. Finally the authors discuss a practical implementation of a distributed form of CacheCloak under the assumption that the intermediary server is untrusted.

Approach

1. The authors carry out a trace-based simulation in order to have realistic operating conditions. A city map was loaded into a simulator with virtual drivers following physical laws and defined speed limits, with random placement of vehicles on the map as a bootstrapping criteria. The user location was written in a file system as the simulation progressed. These traces were loaded into CacheCloak chronologically, simulating a real time stream.

2. Two cases can arise while operating CacheCloak, the submitted coordinates can already have the information associated with them cached which is termed as cache hit. On the contrary, if the information for a coordinate is not cached, a path prediction needs to be performed. The predicted path is extrapolated until it is connected on both ends to other predicted paths present in the cache.

3. Next, the entire generated path is sent to the LBS and all responses for all locations are retrieved and cached. If the user deviates from the predicted path, new requests to the LBS will be triggered and corresponding results will be cached.

4. Based on a formulated attacker model, the authors evaluate and quantify the location privacy based on entropy which provides a measure of attacker's uncertainty.

Results

The results show that CacheCloak can provide users with multiple bits of entropy within a maximum of 10 minutes. The evaluations also show that such entropies can be achieved even in sparse populations.

Discussion

1. The paper also discusses a distributed implementation of CacheCloak in which the devices will have to perform their own mobility predictions. It will be interesting to investigate the feasibility of the diffusion schemes and the iterative Markov model and quantify the cost of incorrect predictions. As incorrect predictions only lead to communication costs while running it on the CacheCloak Server, now it will also result in computational costs.

2. In the paper, the the authors primarily investigate the application of CacheCloak to vehicular mobility, more specifically, cars. However, a considerable proportion of the population use mixed modes of transportation. Their mobility behaviour involves bus/train/ walking or all of them, in which case the predictions are not so straightforward. As a result some alterations in the proposed framework needs to be made to have a practical viability.

6 Privacy-Preserving Deep Learning [3]

Summary

Context

Today, commercial companies such as Facebook, Google and Apple collect a large amount of user data to learn about user preferences and suggest recommendations. This requires a large amount of data related to the users, a considerable part of which is highly sensitive personal information. This data is used to formulate user specific models, generation of which is aided by techniques such as deep learning. On the other hand, biomedical and clinical researchers cannot attain benefits from these techniques as they are not permitted to share the data. As a result privacy and confidentiality restrictions reduce the utility.

Problem

Deep learning presents a interesting avenue to extract highly accurate models by deriving complex features from high dimensional data. Although, such accurate models can give rise to high utility applications they present serious privacy issues due to user data stored in the servers, which is also used for monetary gains by these companies. As a result, there is a need to alter these model generation techniques to offer a satisfactory point in the utility/ privacy tradeoff space.

Contributions

The authors device a distributed deep learning technique that collaborates with multiple participants to learn a neural-network model on their own inputs, without explicitly sharing these inputs. The key contributions can be summarised as follows:

- A selective parameter update model: During training iterations, some attributes contribute largely towards a neural networks objective function as compared to others. This model selects parameters whose current value is far away from the local optima. Only these parameters are updated collaboratively to undergo bigger changes in subsequent iterations.
- Distributed collaborative learning: After every iteration of local training, participants asynchronously share the computed gradients with each other. Therefore benefiting from each others training data without actually sharing it. Thus preserving the privacy of the participants.

Approach

The system architecture consists of a local training database for performing the training locally and parameter server which can be an actual server or a distributed system for running the parameter and gradient exchange protocol.

- Each participant maintains a local vector of neural network parameters. A training iteration is performed over the local training data. Next, the participant downloads the parameters uploaded at the server and computes the gradient of each parameter.

- The parameter server initialises the parameter vector and handles the participants upload and download requests.
- Distributed Stochastic Gradient Descent: While training alone, every participant is more likely to converge to a local optima. However using a distributed approach in learning, by using parameters trained on different datasets, can help to escape local optima resulting in more accurate models.
- Parameter exchange protocol: The authors assume round robin to run the gradient decent sequentially. Every participant downloads the most updated parameters from the server, runs local training and uploads selected gradients and the next participant follows in the fixed order.

Results and Discussion

The authors implemented distributed gradient decent with round robin, random order, and asynchronous parameter exchange protocols. The results were compared with two scenarios, running a gradient descent on the entire dataset and the other is standalone gradient descent where participants train only on their own training data without collaboration. The evaluation was performed on two major datasets used as benchmarks in deep-learning. The key results can be summarised as below:

- The authors achieved the same accuracy as simple stochastic gradient decent as compared to gradient descent ran by sharing only a small fraction of gradients at each iteration step.
- The results obtained show that even at sharing only 1 percent of parameters, results in higher accuracy than standalone or centralised learning.
- Distributed gradient descent using round robin parameter resulted in the highest accuracy and was discovered that round robin protocol is suitable for scenarios where all participants have similar computational capacity.
- It was also observed that number of participants has a lower impact on accuracy than the percentage of shared parameters. Assuming each participant shares his largest gradients with other participants.
- The system computes the differential privacy of each parameter and than decides which one to share with other participants resulting in lower privacy leakage. Further, the authors find the sensitivity index for each parameter, quantifying the amount of random noise which needs to be added to achieve a certain level of differential privacy.

The experiments conducted in the paper are based on supervised learning, it will be interesting to evaluate the accuracy of distributed stochastic gradient descent on unsupervised learning approaches.

References

[1] M. A. Bayir, M. Demirbas, and N. Eagle. Discovering spatiotemporal mobility profiles of cellphone users. In *WOWMOM 2009, Greece, June*.

[2] J. Petzold, F. Bagci, W. Trumler, and T. Ungerer. In *Euro-Par*, volume 4128, 2006.

[3] R. Shokri and V. Shmatikov. Privacy-preserving deep learning. CCS '15, 2015.

[4] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 2010.

[5] Y. Zheng, Q. Li, Y. Chen, X. Xie, and W.-Y. Ma. Understanding mobility based on gps data. UbiComp '08. ACM, 2008.

[6] Y. Zheng, L. Wang, R. Zhang, X. Xie, and W. Y. Ma. Geolife: Managing and understanding your past life over maps. In *Mobile Data Management, 2008. MDM '08. 9th International Conference on.*