# Determining the Locations Visited by GPS Users: A Clustering Approach

Mohamad Saraee
School of CSE
University of Salford
Manchester, UK

Sertan Yamaner
School of CSE
University of Salford
Manchester, UK

Ming Dai
School of CSE
University of Salford
Manchester, UK

Dawei Long
School of CSE
University of Salford
Manchester, UK

## Abstract

*The aim of our research is to use the GPS log captured over 2 days from a PDA and try to extract locations the user have visited. For this research, we have logged GPS data over two days when the user was moving at least 2 miles per hour. To achieve this we cluster the information using data mining clustering software and then analyse the results to see which locations represent a place the user has spent his time.*

**Keywords:** PDA, GPS, Data Mining, Time Series, Clustering.

## 1. Introduction

Data mining is understood by us as a particular implementation which is normally implied to retrieve a certain kind of information from a considerably large scale of data. In this paper, we approach the problem by analysing time series by mining a group of GPS data which is received from a PDA that has been carried around with a user for about two days. We will cluster all the locations where the user has been detected by the GPS and try to find out places where the user visits frequently. The data contains roughly 100,000 records each of which consists of latitude, longitude and time. Every record is given a unique ID by the time order.

The system consists of a PDA that is connected to the Server through Internet. PDA application is written in Microsoft C# and it uses SQL Server CE edition as a temporary database. The server application is also written in C# and exposes it functionality through Web Service architecture. Mobile application captures data from attached GPS card if the user is moving faster than a threshold value. At any time, if the mobile device is connected to the Internet, then it will submit the day's log to server and then receive the locations visited by its user. This method was selected instead of processing locally because of performance issues. The server application will analyse the data log and then to find the time difference between any two sequential GPS reading. If the time difference is greater than the threshold value, then this reading value will be treated as a location visited by the user and then copied to locations list. It is possible to visualise reduced data set via Microsoft Mappoint application to verify the results.

## 2. GPS Technology

GPS, the Global Positioning System, is the only system today able to show you your exact position on Earth at any time, any where, and in any weather. GPS satellite orbit 11,000 nautical miles above Earth. They are monitored continuously at ground stations located around the world. The satellites transmit signals that can be detected by anyone with a GPS receiver. GPS receivers can be carried in your hand or be installed on an aircraft, ships, tanks, submarines, cars, and trucks. These receivers detect, decode and process GPS satellite signals. The typical hand-held receiver is about the size of a

cellular telephone, and the latest models are even smaller. The commercial hand-held units distributed to U.S armed forces personnel during the Persian Gulf War weighted only 28 ounces(less than two pounds). Since then, basic receiver functions have been miniaturized onto intergraded circuits that weigh about one ounce.

The principle behind GPS is the measurement of distance (or "range") between the satellites and the receiver. The satellites tell us exactly where they are in their orbits. It works something like this: if we know our exact distance from a satellite in space, we know we are somewhere on the surface of an imaginary sphere with a radius equal to the distance to the satellite radius. If we know our exact distance from two satellites, we know that we are located somewhere on the line where the two spheres intersect. And, if we take a third and a fourth measurement from two more satellites, we can find our location. The GPS receiver processes the satellite range measurements and produces its position. Simply the GPS system consists of satellites whose paths are monitored by ground stations. Each satellite generates radio signals that allow a receiver to estimate the satellite location and distance between the satellite and the receiver. The receiver uses the measurements to calculate where on or above Earth the user is located.

## 3. Implementation

Like Ashbrook and Starner's work [1][2], some of the data is discarded to make processing more effective. Therefore, if the GPS data does not change or if the user is not moving, then the data is discarded. This also makes it easier to determine the locations visited by the user as the time difference between any 2 locations will be significant. In order to consider a location significant, the user should spend some time as opposed to moving continuously. This may also introduce some false locations where GPS signal is lost. However, considering the frequency of locations, it will be possible to eliminate these false locations. Whenever a point is found that has more than a certain time between it and the previous point, we conclude that the point marks a significant location.

## 4. Determining Places

In order to infer locations visited by the user from GPS log, application in the server gets connected to the database and reads all the log entries. At any point if time difference between current entry and previous entry is greater than our pre determined threshold value, then this location is a candidate for being a significant location. When analyzing our two days of data, we have noticed that time threshold and the number of places determined followed a fairly linear relationship as shown in Figure 1. As the threshold approaches zero, the number of places grows ever more rapidly (at t = 1 seconds 2126 places were found), but there are few indication to what is a good value for threshold. In the end, we decided on 15mins 30sec. As the aim of this implementation is to analyze data in regular intervals as opposed to analyzing a few months of data, this primary data reduction strategy seems to be adequate.

Due to limitations of GPS system[3], it should be expected that there will be signal losses during data capture and these points would be considered as significant locations as well. However, as learned locations will be updated in the mobile device daily, it will then be possible to eliminate the locations that have not been visited regularly. Therefore a secondary data processing strategy is assigned to mobile device for reducing errors in location determination.

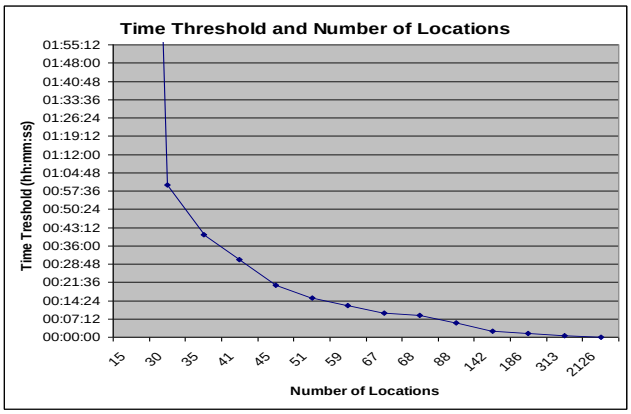


**Figure 1:** *Relationship Between Number of Locations and Time Threshold*

| time | latitude | longitude | id |
|---|---|---|---|
| 2004-4-16 | 53.48611167 | -2.25883 | 147394 |
| 2004-4-16 | 53.48611167 | -2.25883 | 147395 |
| 2004-4-16 | 53.48611167 | -2.25883 | 147396 |
| 2004-4-16 | 53.48611167 | -2.25883 | 147397 |
| 2004-4-16 | 53.48611167 | -2.25883 | 147398 |
| 2004-4-16 | 53.48611167 | -2.25883 | 147399 |
| 2004-4-16 | 53.48611167 | -2.25883 | 147400 |
| 2004-4-16 | 53.48611167 | -2.25883 | 147401 |
| 2004-4-16 | 53.48624167 | -2.25907 | 147403 |
| 2004-4-16 | 53.48624167 | -2.25907 | 147404 |
| 2004-4-16 | 53.48624167 | -2.25907 | 147405 |

**Figure 2:** *sample of the data of GPS*

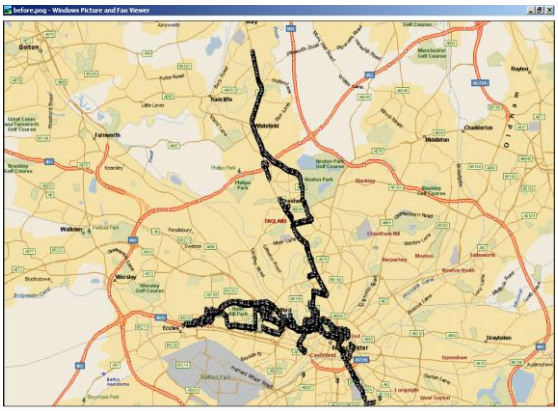

Figure 3: *the map of user's route before data mining*

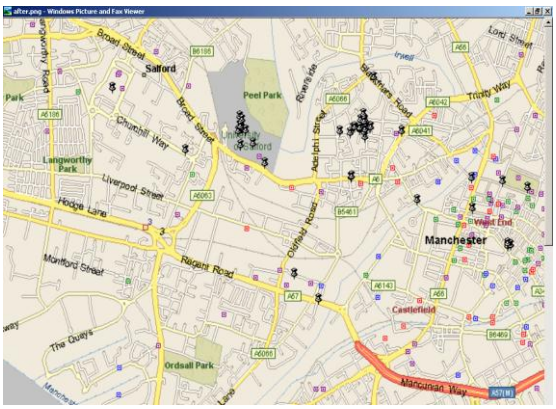


Figure 4: *the map of user's route after data mining*

## 5. Conclusion and future work

After analysing two days GPS data and clustering them with a predefined time threshold, we could determine the locations in which the GPS user has spend most of his/her time during that period. The time threshold is main attribute at this stage of the application. The linear relationship between the number of locations and the time threshold has been discovered and this relationship helps us to set a sensible threshold value in order to determine the most realistic locations. With this predefined time threshold, we can mine our time series database in which the GPS data has been recorded during the two-day experiment.

In the future, we will be adding some more sophisticated functionalities to the system. We are planning to do clustering the reduced data set into actual locations instead of GPS points. In addition, once data has been clustered, algorithms that aid predicting the destination of the users will be implemented. Currently, there are some algorithms implemented that can predict the destination which does not include the expected arrival time arrive to destination. Therefore, our aim will be taking time dimension into consideration in our prediction algorithms.

## 6. References

[1] Ashbrook, D. and T. Starner (2002). Learning Significant Locations and Predicting User Movement with GPS, IEEE Computer Society.

[2]Ashbrook, D. and T. Starner (2003). "Using GPS to learn significant locations and predict movement across multiple users." Personal and Ubiquitous Computing **7**(5): 275 - 286.

[3]Choi, E. and D. A. Cicci (2003). "Analysis of GPS static positioning problems." Applied Mathematics and Computation **140**(1): 37-51.