# AUTOMATED PREDICTION OF CUSTOMER HOTSPOTS TO TAXI DRIVERS USING CLUSTERING TECHNIQUES AND WEB SCRAPING

SPDT Saranatha[1] and PPNV Kumara
Department of Computer Science, General Sir John Kotelawala Defence University, Sri Lanka
[1]saranatha@yahoo.com

**Abstract**- Taxi service is one of the most important service in our society. There are many mobile application to customer to book a taxi. But the problem is that applications are not utilized properly to find taxis to customers in a city area or a busy environment because the demand for taxi exceeds the supply. Purpose of this research is to predict the taxi travel demand in a city area. So that mobile applications can utilize properly to guide taxi drivers to give a proper service to customers. The prediction is done clustering the historical data such as time, weather, location using clustering techniques like k-means and DBScan we can cluster the data and cluster hotspots of customers can be found. There are websites that shows details about the upcoming events. In that websites we can find event location, time and other details about the event. So using web scraping techniques we can scrape those data to get that event data. Using those data we can notify the taxi drivers about the nearby events. So they can easily find more customers who are attending to those events quickly. By this method the time and money of both taxi drivers can be saved. So the profit of taxi drivers will be increased.

**Keywords**- clustering, web scraping, taxi

## I. INTRODUCTION

Studying the pattern of people movements gives lot of information about their daily activities and location is one of the most commonly used forms of context. It is generally simple to gather location information, and other pieces of context may be inferred from location, for example, the nearness of other people (Daniel Ashbrook, 2003). People can be provided with many services using this information. So that people can increase the productivity of their day. Food companies, cloth companies, taxi companies need this type of data to find the hotspots of customers. This paper describes more about how taxi companies can improve their profit using these types of data. To study this moving patterns and converting this into statistical format and to see that is there any technical format by which we can analyse his moving behaviour. The study of consumers helps taxi companies to improve their strategies. One reason for driving an empty vehicle is that taxi drivers do not know where potential clients are, abandoning them with no decision however to meander around the city. The goal of this research is to predict the areas with potential demand from contexts and past history.

Global Positioning System (GPS) hardware to collect location data in a simple and reliable manner (Daniel Ashbrook, 2003). Therefore GPS can be used to get the locations of people. Analysing the data on past history, including the time and location passengers got on taxis, provides clues to the demand distribution. Given the contexts of time, location, and weather, relevant records are filtered for further computation (Neema Davis, 2016). And also the other main important thing to concern is extraction of events data from web because events are the places that we can find many customers. And normally customers are not taking their own vehicles to events because of parking problems, some customers drink alcohol in the events and vehicle safety so with these problem some are not taking their vehicle to event so taxi drivers can provide them transporting service if they could find the location and time of the event so extraction of event will give many opportunities to taxi drivers to find customers. Hence the past data can be collected from taxi service providers and using proper data mining tools and extracting events from web  the hotspot of the customers can be calculated.

## II .LITERATURE REVIEW

Detecting hotspot and studying the behaviour pattern of people is a highly interested area. Due to availability of fast computations there is more access to the technology of hotspot detection. In the paper Hotspot detection and clustering: ways and means (B.Lawson, 2010) it defines what a hotspot is and what is clustering. The most general meaning of clustering and hence clusters is where an intensity threshold or level threshold is used and any area of a map above the threshold counts as a cluster (B.Lawson, 2010). This term level thresholding furthermore, it is generally fluffy definition as there is no necessities for bunches to have neighbour integrity, be a sure shape or measure or to be particular as for this situation no limit prerequisites are characterized. This is fundamentally a type of hotspots clustering where a guide is checked for regions of 'excess level. In general, the distinguishing proof of clusters requires the meaning of cluster location, cluster size, cluster shape and furthermore potentially some measure of least force. For instance, in Figure 1 board A shows a substantial bunch though Figure 1 board B may recommend the presence of little clusters (or no groups by any means). Size is obviously a distinctive factor for this situation.

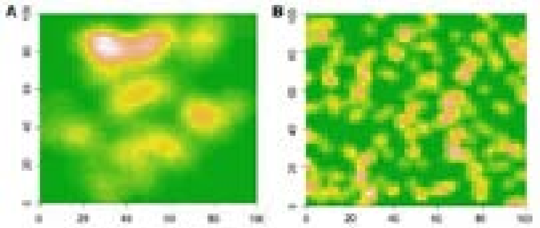


*Figure 1.   Types of hotspots*

In Taiwan each taxi driver operated the business 9.9h a day averagely. But they were wasting about 3.2 h in road without taking any passengers. So (Han-Wen Chang, 2010 january)in there research they have done this to reduce the time of taxi drivers who are roaming for fairs without a proper guide. In their research they analyse the data in past history, including the time, location, and weather. And they have used the clustering methods to find the locations of the customers with a high density or the location of customers who are requesting taxies highly. GPS signals were used to calculate the location of the customer and since the signal isn't precise the records were not indistinguishable but rather spatially near each other so they gathered the adjacent areas into clusters, and that groups additionally mapped to streets or landmarks which covers the greater part of the focuses in the cluster which is showed in figure 1. In the below illustration they show that request records (the plus-sign points) undoubtedly form three main clusters(with few outliers). In the intersection of road A and road F the upper right most cluster is formed. And in road E between road B and C the lower cluster is formed. So that is how they illustrate the formation of clusters in road. They have used three clustering algorithms k-means, x-means, hierarchical clustering and Density-Based Spatial Clustering of Applications with Noise (DBSCAN)(Martin Ester, 1996)).  Also, they expressed distinctive clustering strategies had diverse performances on various sort of data distributions with the goal that it was difficult to choose one as the best.

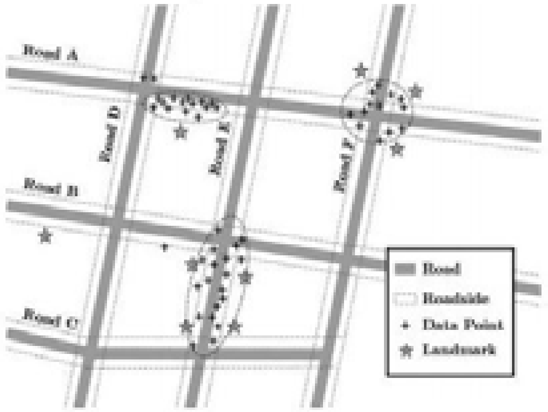


*Figure 2. Mapping clusters to road*

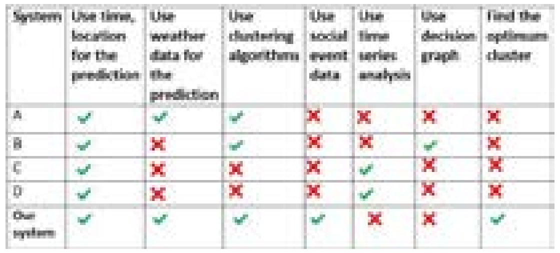
In the research a context aware  taxi demand hotspots prediction is done using data mining techniques they only consider about finding  clusters not the optimum clusters, clusters which are more profitable to taxi drivers. And they have included in their future works to consider the events in that area (social event) that a taxi driver can find more customers.

(P. X. Zhao, 2015) has done a research for detecting hotspot from taxi trajectory data using spatial cluster analysis. A method of trajectory clustering based on decision graph and data field. Contrasted and normal clustering techniques, it can naturally find out parameter as opposed to doing that by encounter and is reasonable to direction clustering. Furthermore, they apply it to direction based urban hotspot disclosure in Wuhan City, China. .Circulation and progression of the hotspots are investigated by utilizing taxi direction information regarding occasion, weekday and weekend. However, like the majority of the current clustering algorithms, the proposed strategy just thinks about spatial data, estimating similarity of points with distance between them. A trajectory data point is a trace generated by a moving object in geographical space usually represented as by series of chronological ordered points. In their paper they have mentioned that existing clustering algorithms like k-means and DBSCAN for hotspot detection discovery have some difficulties in meeting requirement of trajectory data for heterogeneous spatial distribution. Though they predict hotspot as well as traffic hotspot they didn't consider about finding the optimum clusters which are profitable to taxi drivers and finding the clusters considering the traffic data. And also they didn't include about considering event which are highly affects to the prediction of hotspots.

In the paper Predicting Taxi-Passenger Demand using Streaming Data(Luis Moreira-Matias, n.d.) they exhibited a novel use of time arrangement of time series forecasting techniques to enhance the cab driver versatility insight. They took signal emitted by 441 taxi from an organization working Porto, Portugal by transforming the GPS and occasion signals into time series to use firstly as learning base to their model and secondly as a streaming test framework. Accordingly, their model could foresee the taxi-traveller request at every single one of the 63 taxi remains at 30 minute time span interims. Furthermore, they said that the model made by them showed a more than satisfactory performance, accurately forecasting 506873 tested services with an aggregated error measure lower than 26%. That system ready to mine both the periodicity and seasonality of the traveller request, refreshing itself frequently, It at the same time utilizes long-term, midterm and short term authentic information as a learning base and it takes focal points of the pervasive attributes of a taxi network, amassing the experience and the learning everything being equal/drivers while they for the most part utilize only their own. In this research they didn't consider the weather data to their prediction. Weather is a highly effectible matter for the hotspots prediction and also they didn't consider about finding optimum clusters and also consider social event data.

In some countries the traffic changes from area to area very rapidly so a one model so unified model which can apply to countries with the same traffic in all areas cannot be apply to countries with different traffic conditions in different areas. So Neema Davis, Gaurav Raina and Krishna Jagannathan has done a research A Multi-Level Clustering Approach for Forecasting Taxi Travel Demand(Neema Davis.,2016) to make different models to different areas with different traffic conditions. They have used GPS data to find the location of the customer. Also, a time series model is based on the suspicion that the present and the future request would have some relationship with the past request, and represented as the function past information. So these are the models that they have used, moving average, exponential smoothing, i.e, Holt-Winters (HW) model, Seasonal Naive, FFT, STL decomposition, ARIMA, TBATS, linear regression and state space ETS model. So they have calculated which model fits to which area best. To do that they have used baseline model which is an averaging model so that a model is retained if it performs better than the baseline, else it is discarded. And they used a multi-level clustering technique to improve the performance of the model by 20%. In this research the importance is that they have made different models to different areas in the country but they didn't consider the weather data and also the social evet data. In here though they predict the hotspot they didn't consider about finding the optimum hotspot.

Mining hotspot are more usable to the new taxi drivers who are doing taxi service as a start-up. Because like typical taxi driver a new taxi driver doesn't know about the customer he don't have good experience about customer behaviour patterns for him prediction of cluster centres is very useful. There are taxi drivers who do taxi service as a part time job. So for taxi drivers like that it takes time to learn about customers. For by using a taxi services application those taxi drivers can find hotspots and can go that places and can provide good service rather than roaming around the city uselessly. And for the typical taxi drivers mining and predicting hotpot is not very useful but if they can know the optimum hotspots then that is useful for them.



| System | Use time, location for the prediction | Use weather data for the prediction | Use clustering algorithms | Use social event data | Use time series analysis | Use decision graph | Find the optimum cluster |
|---|---|---|---|---|---|---|---|
| A | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ |
| B | ✓ | ✗ | ✓ | ✗ | ✗ | ✓ | ✗ |
| C | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ |
| D | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ |
| Our system | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ |

*Figure 3. summary of review*

A - A context aware taxi demand hotspots prediction.

B - Detecting hotspot from taxi trajectory data using spatial cluster analysis.

C - Predicting Taxi-Passenger Demand using Streaming Data.

D - A Multi-Level Clustering Approach for Forecasting Taxi Travel Demand.

## III. METHODOLOGY

We have used Kaggle dataset of New York taxi trip duration and also kaggle New York hourly weather data set and we combined those two datasets. So we can have both taxi trip data and weather data. And we used Scikit-learn library in python to do the clustering. To prediction of hotspot we have done three things.

1. Clustering the dataset considering date, time, weather and location.
2. Extraction of social events data from evet notifying web sites.
3. Prediction of hotspots and notify about events to taxi drivers.

### *A. K-means algorithm*

In this technique the number of cluster (k) is predefined preceding examination and afterward the choice of the underlying centroids will be made arbitrarily and it took after by iterative procedure of doling out every datum point to its closest centroid. This procedure will continue rehashing until the point when merging criteria met.

In this algorithm first we have to define the number of clusters. This algorithm work as follows: In the following illustration there are 5 data points and it takes k=2. And in the figure 4 it randomly assigns each data point to a cluster.

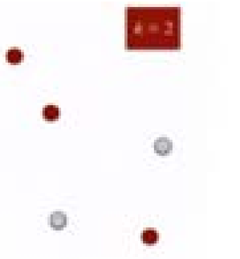


*Figure 4.*

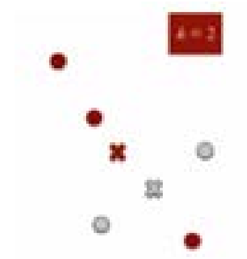
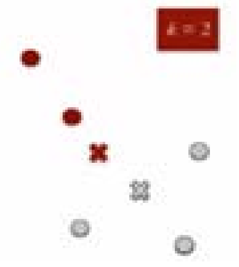


*Figure 5.*

Then the centroid of each cluster have to be compute so in figure 5 the centroid of data points in the red cluster is shown using red cross and those in grey cluster using grey cross.



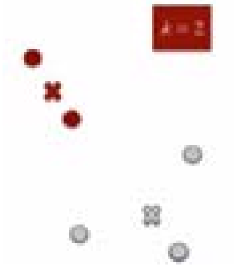


*Figure 6.* *Figure 7.*
*Figures 4,56,7 Data points(saurau,2016)*

Then in figure 6 each data point is re-assign to the closest cluster and again new cluster centroids have to be re-compute. And it is shown in figure 7

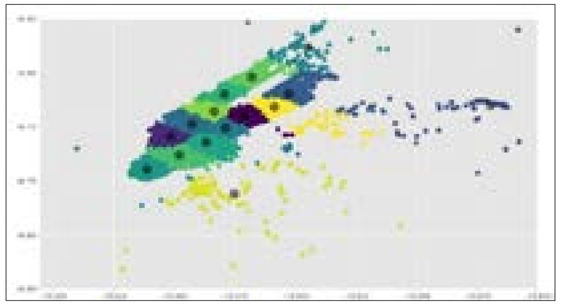


*Figure 8, output of K-means*

To our dataset when we apply K-means algorithm the output is shown above figure 8. We have used silhouette score to gauge accuracy and it is an esteem is a measure of how comparative a object is to its own cluster (cohesion) contrasted with different clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and inadequately coordinated to neighbouring clusters. The silhouette score of the above cluster data is 0.4412.

In the above algorithm the problem is that the number of clusters should be predefined and it is sensitive to noise data. And when we reducing the cluster number the silhouette score get increase to some extent but the problem is each time we have to define the number of clusters.

### B. DBSACN algorithm

In k-mean algorithm we can't discover the noise properly. DBSCAN (Density Based Spatial Clustering of Applications with Noise) which is designed to discover the clusters and the noise in a spatial database(Martin Ester, 1996)).And this algorithm work as follows:

There are two parameters:

- Eps – Maximum radius of the neighborhood
- MinPts- Minimum number of points in the Eps-neighborhood of a point.

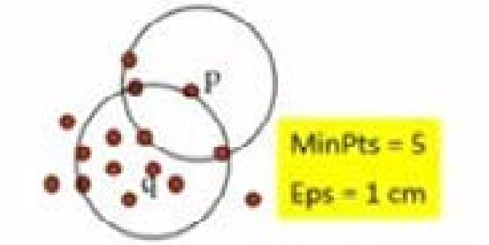


**Figure 9. Data points(saurau,2016)**

A point should be arbitrarily selected(P) and retrieve all points from p w.r.t. Eps and MinPts and if p is a core point cluster is formed and if p is border, no points are density-reachable from p, and DBSCAN visits the next point of the database. And it continues the process until all of the points processed.

When we apply DBSCAN algorithm to the dataset we get a clustering shown in the figure 8. And the silhouette score is 0.01970. When we applied as shown in the figure the noise data do not get into clusters. But problem is we need cluster centre points so in here it do not gives directly cluster centre points.

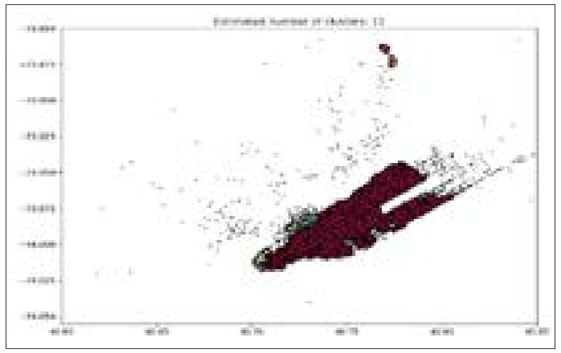


**Figure 10. Output of DBSCAN**

### C. DBbkmeans Algorithm

In the paper A Novel Density based improved k-means Clustering algorithm(K. Mumtaz, 2010) they propose a new algorithm which overcomes the draw backs of DBSCAN and K-means clustering algorithms. In this research they have combined the DBSCAN and K-means algorithms. They have first run the DBSCAN algorithm and then they have find the number of clusters and then using the cluster value they run the K-mean algorithm. So in our project we used this combined algorithm to find the clusters.

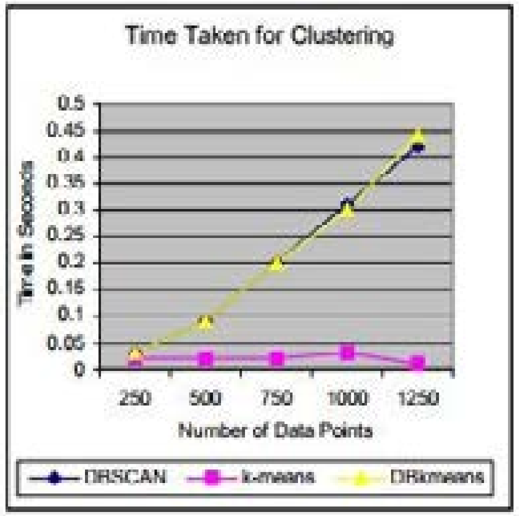


**Figure 11. Time taken for clustering(K. Mumtaz,2010)**

As shown in the above figure 11 time taken for clustering is increasing when number of records are high and that is disadvantage but in our project since we are updating the hotspot hourly it will not be big issue.

| Algorithm | Silhouette Score |
|---|---|
| K-means | 0.44012 |
| DBSCAN | 0.01970 |
| Dbkmeans | 0.48403 |

**Figure 12. silhouette scores**

As mentioned in the above table the silhouette score for each algorithm when we done the clustering using different algorithms are shown. So according to that the Dbkmeans shows the highest score. Because of that using that algorithm to our system will give more accurate results.

After calculating clusters the next task is to select the optimum clusters. To calculate the optimum cluster mean that the most profitable cluster to the taxi driver. To do that we first calculate distance between pickup and drop-off location using google maps API and then we multiply that distance by charge(cost)taken per unit distance. So we do that for all the data points in a cluster and add all the values and take the total. Then that total value is divided by the total number of taxi rides in that cluster. The final answer is the average income of taxi driver who use that cluster.

- Distance between pickup and drop-off location in data point in a cluster – distance(i,(x,y)) -> i- number
- of the cluster, (x,y) – coordinate of the data point in the cluster.
- Total Distance of the all the data points in the cluster Σ distance(I,(x,y)).
- Total income of the cluster=Total Distance × Charge per unit distance.
- Average income of a taxi driver = Total income ÷ total number of points in the cluster

So the cluster with the highest average income of the taxi driver will be the optimum cluster.

### D. Extraction of events from web

An occasion happens at a specific area, has a begin date and time, and a title or portrayal. In other words to be valuable to a client, an occasion must have the capacity to answer the inquiries: What?,When?, and Where?(John Foley, 2015). In the paper learning to extract local events from web (John Foley, 2015). So they have focused on the identification and extraction of events on the open internet and recommend events that users might want to attend. So this method is very useful to taxi drivers because they can find more customers in an event area. So the taxi service providers can use this method to notify the location and the time of the event to taxi drivers.

Web scraping is also known as web data extraction, web data extraction, web harvesting, and screen scraping. And it is great technique of extracting unstructured data to a structured format. So that structured data can be stored in databases, spread sheets, XML files. Information like online price comparison, context scraping, weather data monitoring, extract offers and discounts, and extract information from job notification web sites, collect government data and market data are the data that can scrap using web scraping.

### E. Web scraping techniques

There are several methods and ways that we can do web scraping. And each one has its own advantages and disadvantages

Computer vision web page analysers, they can analyse the web page like human using machine learning and image processing techniques and find the important unstructured data and store them in structured format. Example is Diffbot. And this technique uses high computation power because it using both machine leaning and image processing techniques.

In our research we have used web scraping using python and the BeautifulSoup library and we used it because the simplicity and easy to handle. But the problem in that method is to we have to customize the web scraping code to site to site because in that method what we do is we extract text from HTML tags so tag arrangement is differ from site to site. So by that method we can find the location, time and name of the event. And store that data

# AUTOMATED PREDICTION OF CUSTOMER HOTSPOTS TO TAXI DRIVERS USING CLUSTERING TECHNIQUES AND WEB SCRAPING

SPDT Saranatha[1] and PPNV Kumara

Department of Computer Science, General Sir John Kotelawala Defence University, Sri Lanka

[1]*saranatha@yahoo.com*

**Abstract**- Taxi service is one of the most important service in our society. There are many mobile application to customer to book a taxi. But the problem is that applications are not utilized properly to find taxis to customers in a city area or a busy environment because the demand for taxi exceeds the supply. Purpose of this research is to predict the taxi travel demand in a city area. So that mobile applications can utilize properly to guide taxi drivers to give a proper service to customers. The prediction is done clustering the historical data such as time, weather, location using clustering techniques like k-means and DBScan we can cluster the data and cluster hotspots of customers can be found. There are websites that shows details about the upcoming events. In that websites we can find event location, time and other details about the event. So using web scraping techniques we can scrape those data to get that event data. Using those data we can notify the taxi drivers about the nearby events. So they can easily find more customers who are attending to those events quickly. By this method the time and money of both taxi drivers can be saved. So the profit of taxi drivers will be increased.

**Keywords**- clustering, web scraping, taxi

## I. INTRODUCTION

Studying the pattern of people movements gives lot of information about their daily activities and location is one of the most commonly used forms of context. It is generally simple to gather location information, and other pieces of context may be inferred from location, for example, the nearness of other people (Daniel Ashbrook, 2003). People can be provided with many services using this information. So that people can increase the productivity of their day. Food companies, cloth companies, taxi companies need this type of data to find the hotspots of customers. This paper describes more about how taxi companies can improve their profit using these types of data. To study this moving patterns and converting this into statistical format and to see that is there any technical format by which we can analyse his moving behaviour. The study of consumers helps taxi companies to improve their strategies. One reason for driving an empty vehicle is that taxi drivers do not know where potential clients are, abandoning them with no decision however to meander around the city. The goal of this research is to predict the areas with potential demand from contexts and past history.

Global Positioning System (GPS) hardware to collect location data in a simple and reliable manner (Daniel Ashbrook, 2003). Therefore GPS can be used to get the locations of people. Analysing the data on past history, including the time and location passengers got on taxis, provides clues to the demand distribution. Given the contexts of time, location, and weather, relevant records are filtered for further computation (Neema Davis, 2016). And also the other main important thing to concern is extraction of events data from web because events are the places that we can find many customers. And normally customers are not taking their own vehicles to events because of parking problems, some customers drink alcohol in the events and vehicle safety so with these problem some are not taking their vehicle to event so taxi drivers can provide them transporting service if they could find the location and time of the event so extraction of event will give many opportunities to taxi drivers to find customers. Hence the past data can be collected from taxi service providers and using proper data mining tools and extracting events from web  the hotspot of the customers can be calculated.

## II .LITERATURE REVIEW

Detecting hotspot and studying the behaviour pattern of people is a highly interested area. Due to availability of fast computations there is more access to the technology of hotspot detection. In the paper Hotspot detection and clustering: ways and means (B.Lawson, 2010) it defines what a hotspot is and what is clustering. The most general meaning of clustering and hence clusters is where an intensity threshold or level threshold is used and any area of a map above the threshold counts as a cluster (B.Lawson, 2010). This term level thresholding furthermore, it is generally fluffy definition as there is no necessities for bunches to have neighbour integrity, be a sure shape or measure or to be particular as for this situation no limit prerequisites are characterized. This is fundamentally a type of hotspots clustering where a guide is checked for regions of 'excess level. In general, the distinguishing proof of clusters requires the meaning of cluster location, cluster size, cluster shape and furthermore potentially some measure of least force. For instance, in Figure 1 board A shows a substantial bunch though Figure 1 board B may recommend the presence of little clusters (or no groups by any means). Size is obviously a distinctive factor for this situation.

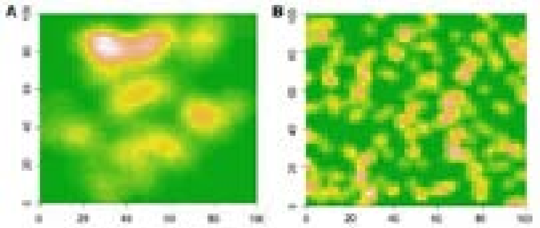


*Figure 1.   Types of hotspots*

In Taiwan each taxi driver operated the business 9.9h a day averagely. But they were wasting about 3.2 h in road without taking any passengers. So (Han-Wen Chang, 2010 january)in there research they have done this to reduce the time of taxi drivers who are roaming for fairs without a proper guide. In their research they analyse the data in past history, including the time, location, and weather. And they have used the clustering methods to find the locations of the customers with a high density or the location of customers who are requesting taxies highly. GPS signals were used to calculate the location of the customer and since the signal isn't precise the records were not indistinguishable but rather spatially near each other so they gathered the adjacent areas into clusters, and that groups additionally mapped to streets or landmarks which covers the greater part of the focuses in the cluster which is showed in figure 1. In the below illustration they show that request records (the plus-sign points) undoubtedly form three main clusters(with few outliers). In the intersection of road A and road F the upper right most cluster is formed. And in road E between road B and C the lower cluster is formed. So that is how they illustrate the formation of clusters in road. They have used three clustering algorithms k-means, x-means, hierarchical clustering and Density-Based Spatial Clustering of Applications with Noise (DBSCAN)(Martin Ester, 1996)).  Also, they expressed distinctive clustering strategies had diverse performances on various sort of data distributions with the goal that it was difficult to choose one as the best.

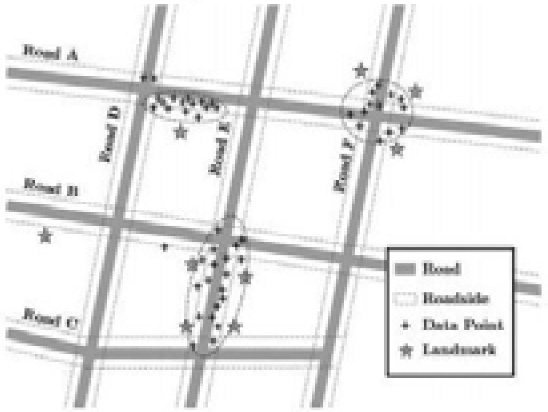


*Figure 2. Mapping clusters to road*

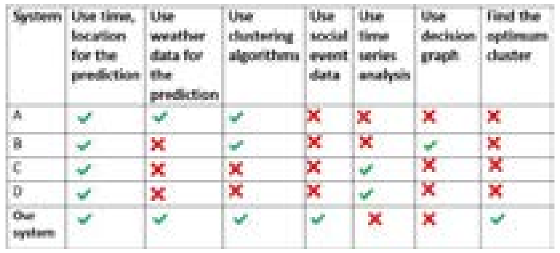
In the research a context aware  taxi demand hotspots prediction is done using data mining techniques they only consider about finding clusters not the optimum clusters, clusters which are more profitable to taxi drivers. And they have included in their future works to consider the events in that area (social event) that a taxi driver can find more customers.

(P. X. Zhao, 2015) has done a research for detecting hotspot from taxi trajectory data using spatial cluster analysis. A method of trajectory clustering based on decision graph and data field. Contrasted and normal clustering techniques, it can naturally find out parameter as opposed to doing that by encounter and is reasonable to direction clustering. Furthermore, they apply it to direction based urban hotspot disclosure in Wuhan City, China. .Circulation and progression of the hotspots are investigated by utilizing taxi direction information regarding occasion, weekday and weekend. However, like the majority of the current clustering algorithms, the proposed strategy just thinks about spatial data, estimating similarity of points with distance between them. A trajectory data point is a trace generated by a moving object in geographical space usually represented as by series of chronological ordered points. In their paper they have mentioned that existing clustering algorithms like k-means and DBSCAN for hotspot detection discovery have some difficulties in meeting requirement of trajectory data for heterogeneous spatial distribution. Though they predict hotspot as well as traffic hotspot they didn't consider about finding the optimum clusters which are profitable to taxi drivers and finding the clusters considering the traffic data. And also they didn't include about considering event which are highly affects to the prediction of hotspots.

In the paper Predicting Taxi-Passenger Demand using Streaming Data(Luis Moreira-Matias, n.d.) they exhibited a novel use of time arrangement of time series forecasting techniques to enhance the cab driver versatility insight. They took signal emitted by 441 taxi from an organization working Porto, Portugal by transforming the GPS and occasion signals into time series to use firstly as learning base to their model and secondly as a streaming test framework. Accordingly, their model could foresee the taxi-traveller request at every single one of the 63 taxi remains at 30 minute time span interims. Furthermore, they said that the model made by them showed a more than satisfactory performance, accurately forecasting 506873 tested services with an aggregated error measure lower than 26%. That system ready to mine both the periodicity and seasonality of the traveller request, refreshing itself frequently, It at the same time utilizes long-term, midterm and short term authentic information as a learning base and it takes focal points of the pervasive attributes of a taxi network, amassing the experience and the learning everything being equal/drivers while they for the most part utilize only their own. In this research they didn't consider the weather data to their prediction. Weather is a highly effectible matter for the hotspots prediction and also they didn't consider about finding optimum clusters and also consider social event data.

In some countries the traffic changes from area to area very rapidly so a one model so unified model which can apply to countries with the same traffic in all areas cannot be apply to countries with different traffic conditions in different areas. So Neema Davis, Gaurav Raina and Krishna Jagannathan has done a research A Multi-Level Clustering Approach for Forecasting Taxi Travel Demand(Neema Davis.,2016) to make different models to different areas with different traffic conditions. They have used GPS data to find the location of the customer. Also, a time series model is based on the suspicion that the present and the future request would have some relationship with the past request, and represented as the function past information. So these are the models that they have used, moving average, exponential smoothing, i.e, Holt-Winters (HW) model, Seasonal Naive, FFT, STL decomposition, ARIMA, TBATS, linear regression and state space ETS model. So they have calculated which model fits to which area best. To do that they have used baseline model which is an averaging model so that a model is retained if it performs better than the baseline, else it is discarded. And they used a multi-level clustering technique to improve the performance of the model by 20%. In this research the importance is that they have made different models to different areas in the country but they didn't consider the weather data and also the social evet data. In here though they predict the hotspot they didn't consider about finding the optimum hotspot.

Mining hotspot are more usable to the new taxi drivers who are doing taxi service as a start-up. Because like typical taxi driver a new taxi driver doesn't know about the customer he don't have good experience about customer behaviour patterns for him prediction of cluster centres is very useful. There are taxi drivers who do taxi service as a part time job. So for taxi drivers like that it takes time to learn about customers. For by using a taxi services application those taxi drivers can find hotspots and can go that places and can provide good service rather than roaming around the city uselessly. And for the typical taxi drivers mining and predicting hotpot is not very useful but if they can know the optimum hotspots then that is useful for them.



| System | Use time, location for the prediction | Use weather data for the prediction | Use clustering algorithms | Use social event data | Use time series analysis | Use decision graph | Find the optimum cluster |
|---|---|---|---|---|---|---|---|
| A | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ |
| B | ✓ | ✗ | ✓ | ✗ | ✗ | ✓ | ✗ |
| C | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ |
| D | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ |
| Our system | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ |

*Figure 3. summary of review*

A - A context aware taxi demand hotspots prediction.

B - Detecting hotspot from taxi trajectory data using spatial cluster analysis.

C - Predicting Taxi-Passenger Demand using Streaming Data.

D - A Multi-Level Clustering Approach for Forecasting Taxi Travel Demand.

## III. METHODOLOGY

We have used Kaggle dataset of New York taxi trip duration and also kaggle New York hourly weather data set and we combined those two datasets. So we can have both taxi trip data and weather data. And we used Scikit-learn library in python to do the clustering. To prediction of hotspot we have done three things.

1. Clustering the dataset considering date, time, weather and location.
2. Extraction of social events data from evet notifying web sites.
3. Prediction of hotspots and notify about events to taxi drivers.

### *A. K-means algorithm*

In this technique the number of cluster (k) is predefined preceding examination and afterward the choice of the underlying centroids will be made arbitrarily and it took after by iterative procedure of doling out every datum point to its closest centroid. This procedure will continue rehashing until the point when merging criteria met.

In this algorithm first we have to define the number of clusters. This algorithm work as follows: In the following illustration there are 5 data points and it takes k=2. And in the figure 4 it randomly assigns each data point to a cluster.

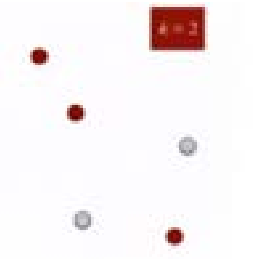


*Figure 4.*

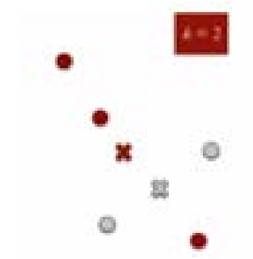
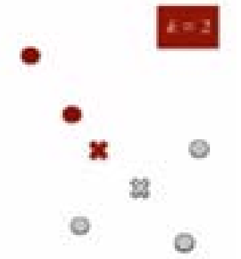


*Figure 5.*

Then the centroid of each cluster have to be compute so in figure 5 the centroid of data points in the red cluster is shown using red cross and those in grey cluster using grey cross.



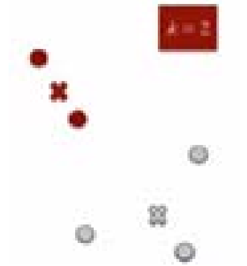


*Figure 6. Figure 7.*
*Figures 4,56,7 Data points(saurau,2016)*

Then in figure 6 each data point is re-assign to the closest cluster and again new cluster centroids have to be re-compute. And it is shown in figure 7

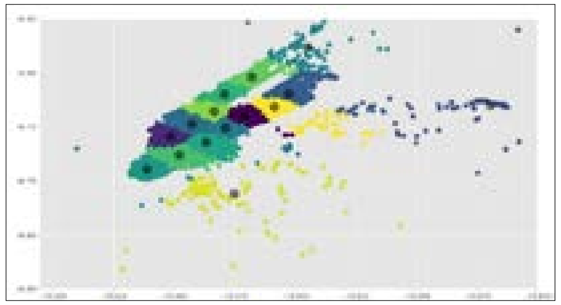


*Figure 8, output of K-means*

To our dataset when we apply K-means algorithm the output is shown above figure 8. We have used silhouette score to gauge accuracy and it is an esteem is a measure of how comparative a object is to its own cluster (cohesion) contrasted with different clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and inadequately coordinated to neighbouring clusters. The silhouette score of the above cluster data is 0.4412.

In the above algorithm the problem is that the number of clusters should be predefined and it is sensitive to noise data. And when we reducing the cluster number the silhouette score get increase to some extent but the problem is each time we have to define the number of clusters.

### B. DBSACN algorithm

In k-mean algorithm we can't discover the noise properly. DBSCAN (Density Based Spatial Clustering of Applications with Noise) which is designed to discover the clusters and the noise in a spatial database(Martin Ester, 1996)).And this algorithm work as follows:

There are two parameters:

- Eps – Maximum radius of the neighborhood
- MinPts- Minimum number of points in the Eps-neighborhood of a point.

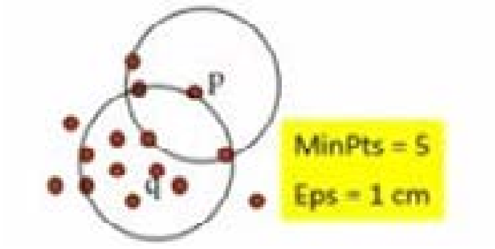


*Figure 9. Data points(saurau,2016)*

A point should be arbitrarily selected(P) and retrieve all points from p w.r.t. Eps and MinPts and if p is a core point cluster is formed and if p is border, no points are density-reachable from p, and DBSCAN visits the next point of the database. And it continues the process until all of the points processed.

When we apply DBSCAN algorithm to the dataset we get a clustering shown in the figure 8. And the silhouette score is 0.01970. When we applied as shown in the figure the noise data do not get into clusters. But problem is we need cluster centre points so in here it do not gives directly cluster centre points.

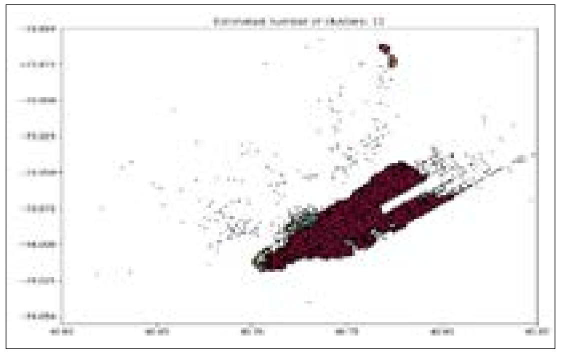


*Figure 10. Output of DBSCAN*

### C. DBbkmeans Algorithm

In the paper A Novel Density based improved k-means Clustering algorithm(K. Mumtaz, 2010) they propose a new algorithm which overcomes the draw backs of DBSCAN and K-means clustering algorithms. In this research they have combined the DBSCAN and K-means algorithms. They have first run the DBSCAN algorithm and then they have find the number of clusters and then using the cluster value they run the K-mean algorithm. So in our project we used this combined algorithm to find the clusters.

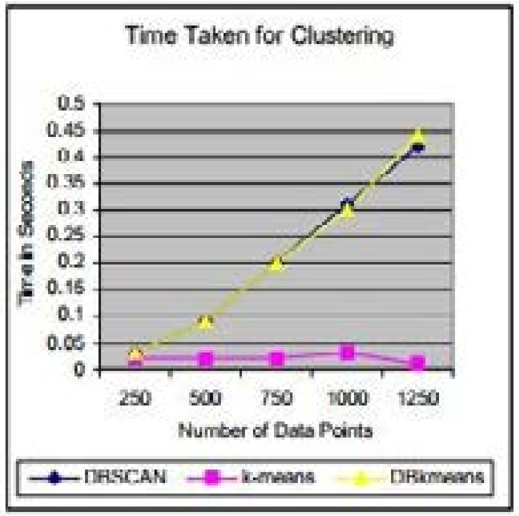


*Figure 11. Time taken for clustering(K. Mumtaz,2010)*

As shown in the above figure 11 time taken for clustering is increasing when number of records are high and that is disadvantage but in our project since we are updating the hotspot hourly it will not be big issue.

| Algorithm | Silhouette Score |
|---|---|
| K-means | 0.44012 |
| DBSCAN | 0.01970 |
| Dbkmeans | 0.48403 |

*Figure 12. silhouette scores*

As mentioned in the above table the silhouette score for each algorithm when we done the clustering using different algorithms are shown. So according to that the Dbkmeans shows the highest score. Because of that using that algorithm to our system will give more accurate results.

After calculating clusters the next task is to select the optimum clusters. To calculate the optimum cluster mean that the most profitable cluster to the taxi driver. To do that we first calculate distance between pickup and drop-off location using google maps API and then we multiply that distance by charge(cost)taken per unit distance. So we do that for all the data points in a cluster and add all the values and take the total. Then that total value is divided by the total number of taxi rides in that cluster. The final answer is the average income of taxi driver who use that cluster.

- Distance between pickup and drop-off location in data point in a cluster – distance(i,(x,y)) -> i- number
- of the cluster, (x,y) – coordinate of the data point in the cluster.
- Total Distance of the all the data points in the cluster Σ distance(I,(x,y)).
- Total income of the cluster=Total Distance × Charge per unit distance.
- Average income of a taxi driver = Total income ÷ total number of points in the cluster

So the cluster with the highest average income of the taxi driver will be the optimum cluster.

### D. Extraction of events from web

An occasion happens at a specific area, has a begin date and time, and a title or portrayal. In other words to be valuable to a client, an occasion must have the capacity to answer the inquiries: What?,When?, and Where?(John Foley, 2015). In the paper learning to extract local events from web (John Foley, 2015). So they have focused on the identification and extraction of events on the open internet and recommend events that users might want to attend. So this method is very useful to taxi drivers because they can find more customers in an event area. So the taxi service providers can use this method to notify the location and the time of the event to taxi drivers.

Web scraping is also known as web data extraction, web data extraction, web harvesting, and screen scraping. And it is great technique of extracting unstructured data to a structured format. So that structured data can be stored in databases, spread sheets, XML files. Information like online price comparison, context scraping, weather data monitoring, extract offers and discounts, and extract information from job notification web sites, collect government data and market data are the data that can scrap using web scraping.

### E. Web scraping techniques

There are several methods and ways that we can do web scraping. And each one has its own advantages and disadvantages

Computer vision web page analysers, they can analyse the web page like human using machine learning and image processing techniques and find the important unstructured data and store them in structured format. Example is Diffbot. And this technique uses high computation power because it using both machine leaning and image processing techniques.

In our research we have used web scraping using python and the BeautifulSoup library and we used it because the simplicity and easy to handle. But the problem in that method is to we have to customize the web scraping code to site to site because in that method what we do is we extract text from HTML tags so tag arrangement is differ from site to site. So by that method we can find the location, time and name of the event. And store that data in the database so that the taxi service provider can use that data to notify the taxi drivers.

### *F. Client and server side*

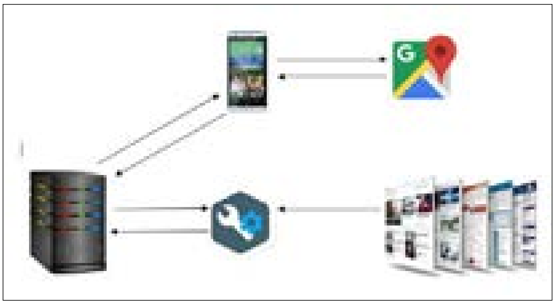


*Figure 13. Data flow*

The entire system is designed to be implemented in a client server architectural design. The clustering model is stored in the server. And when the taxi driver request for hotspot the server run the clustering model considering location, time, and weather condition. And also considering the location it finds the event details of the nearby locations using the web scraped data. And it process all data together and send it the taxi driver. As shown in the above data flow diagram first the taxi mobile application should pass the location, time, weather data to the service providers server. Then the server process the data and start clustering considering location, time, weather condition and it output the hotspot data to taxi application. And the service providers server run the web scraping application by passing location and time data to the web scraping application so the application has stored unstructured scraped data in a structured format and that application use that data and pass the event details considering location and time given by the server. And the server pass the that event details data to the taxi application and taxi application uses google map API to find the location send by the service providers server and then the hotspot location and event details are shown in the taxi application.

## IV. CONCLUSION & FUTURE WORK

Automated prediction of customer hotspot system for taxi drivers system is introduced to overcome from the problem in a city area the demand for taxi exceeds the supply. This system cluster the historical taxi booking data like date, time location and the weather condition at time of the booking to calculate customer hotspots. And it collect the data from web site to find about events and notify the taxi driver about the customer hotspots. And it gives the most optimum cluster that taxi drivers should use it order to have higher income. As future work traffic data also should be considered when we finding the optimum hotspots to taxi drivers and finding most accurate and fast clustering techniques to cluster the historical data.

## ACKNOWLEDGMENT

We would like to thank all the people who helped us in doing a lot of Research and we came to know about so many new things and we are really thankful to them.

## REFERENCES

Anja Struyf, M. H. ,. J. R., 1997. Clustering in an Object-OrientedEnvironment. Journal of statical sofware, Issue 4.

B.Lawson, A., 2010. Hotspot detection and clustering: ways and means. p. 15.

Chang, H.-W., 2010. Context-aware taxi demand hotspots. International Journal of Business Intelligence and Data Mining , p. 7.

Daniel Ashbrook, T. S., 2003. Using GPS to Learn Significant Locations and Predict Movement Across. Personal and Ubiquitous Computing , 7(5), pp. 275-286.

Han-Wen Chang, J. Y.-j. H., 2010 january. Context-aware taxi demad hotspots prediction. Internation Journal of Business Intelligence and Data Mining , p. 17.

Hassan Sayyadi, M. H. M., 2009. Event Detection and Tracking in Social Streams. Proceedings of the Third International ICWSM Conference .

John Foley, M. B. J., 2015. Learning to Extract Local Events from the Web.

K. Mumtaz, D. K. D., 2010. A Novel Density based improved k-means. (IJCSE) International Journal on Computer Science and Engineering, Volume 02, pp. 213-218.

Luis Moreira-Matias, J. G. F. D., n.d. Predicting Taxi-Passenger Demand using Streaming Data.

Martin Ester, H.-P. K. ,. J. S. ,. X. X., 1996). A density-based algorithm for discovering clusters in large spatial databases with noise.

Neema Davis, G. R. a. K. J., 2016. A Multi-Level Clustering Approach for Forecasting Taxi Travel Demand. 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC).

P. X. Zhao, K., 2015. DETECTING HOTSPOTS FROM TAXI TRAJECTORY DATA USING SPATIAL. International workshop for spatiotemporal computing , p. 5.

Wang S, G. W. L. D. e. a., 2011. Data field for hierarchical. Journal of Data Warehousing and Mining (IJDWM).

SaURAV KAUSHIK , NOVEMBER 3, 2016(https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering)