THE UNIVERSITY OF THE SOUTH PACIFIC
LIBRARY

Author Statement of Accessibility- Part 2- Permission for Internet Access

Name of Candidate: ASHIKA VANDHANA SINGH · (S11000183)

Degree: MASTER OF SCIENCE INFORMATION SYSTEMS

Department/School: SCHOOL OF COMPUTING, INFORMATION AND MATHEMATICAL SCIENCES

Institution/University: UNIVERSITY OF THE SOUTH PACIFIC

Thesis Title: BIG DATA ANALYSIS FOR MOBILE COMPUTING RESEARCH :- SEMANTIC ANALYSIS FOR LOCATION AWARENESS

Date of completion of requirements for award: 17/3/15

1. I authorise the University to make this thesis available on the Internet for access by USP authorised users. Yes/No

2 I authorise the University to make this thesis available on the Internet under the International digital theses project Yes/No

Signed: Vandhana

Date. 17/3/15

Contact Address

P.O. BOX 1362, NADI

MOBILE:- +679 9212354

ashika.singh07@gmail.com

Permanent Address

RESERVOIR ROAD

NANAKA

NADI

FIJI ISLANDS

BIG DATA ANALYSIS FOR MOBILE COMPUTING RESEARCH PROJECT- SEMANTIC ANALYSIS FOR LOCATION AWARENESS

by

Ashika Vandhana Singh

A thesis submitted in fulfillment of the

requirements for the degree of

Master of Science in Information Systems

School of Computing Science, Information and

Mathematical Sciences

Faculty of Science, Technology and Environment

The University of the South Pacific

March 2015

Declaration

Statement by Author

I, Ashika Vandhana Singh, declare that this thesis is my own work and that, to the best of my knowledge, it contains no material previously published, or substantially overlapping with material submitted for the award of any other degree at any institution, except where due acknowledgement is made in the text.

Signature..... Date 17/3/15

Name ASHIKA VANDHANA SINGH

Student ID No S11000183

Statement by Supervisor

The research in this thesis was performed under my supervision and to my knowledge is the sole work of Ms. Ashika Vandhana Singh.

Signature..... Date 20/3/15

Name Prof Aruna Jayasuriya

Designation Head of School, SCIMS

Acknowledgements

I would like to express gratitude to my supervisor, Dr. Munir Naveed, for his kind support and much appreciated guidance during my supervised research project. I am also thankful to Ms. Marieata Suliana and Dr. Sushil Kumar for their guidance and assistance during the phase of my research.

Abstract

The mobile computing research phenomenon implicates optimized performance and enhanced smart information accessibility focusing on ease of anywhere, anytime information access using smart devices. Context awareness capability of smart devices is an insatiable requirement of smart device users, providing users with location awareness capabilities, using satellite communication to either broadcast their location, look for places of interest to them or finding their way using global positioning systems.

The main focus of this research work is to explore the capabilities of different machine learning algorithms in building a context-aware model. The context-awareness is based on location parameters of a mobile device. The context-aware model is evaluated using the "citywide" dataset— that is provided by CRAWDAD (an online resource community for research) for benchmarking. The dataset contains real, long-term data collected from three participants using a Place Lab Client, which collects location coordinates of the participants as they move around a specified area. The traces collected are from laptops and Personal Digital Assistants (PDAs).

In this experiment we explore the K-Means and K-Medoids algorithms to generate the clusters for each context. The stratified sampling techniques are used to generate the test sets. The test dataset is used to create data models for the semantic analysis. The results obtained in the experiment indicate that it is plausible to add semantics to mobile computing data for locations awareness. The results reveal that these machine learning algorithms are potentially the candidate solutions to identify places of significance to the user of a mobile device. These algorithms can be used to build context-aware model for the mobile-devices when a context is represented by the location of the device.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

In this paper we attempt to execute big data analysis for mobile computing research. The purpose of this research is to perform "Semantic Analysis for Location Awareness" in mobile computing data.

In this chapter we discuss our research and processes involved in performing the experiments. Primarily we are concerned with obtaining the required data from relevant sources and then describing what the "citywide" (J. H. Kang, 2006) data set is composed of, and how this data was organized to perform relevant analysis.

1.1 Background

The problem occurs as a result of pervasive computing e.g. mobile phones that create innovative research domains for mobile data or in the genre of ubiquitous computing. Such that, a mobile device user can move to several places spontaneously, depending on situations specific to them hence there is a need for these devices to possess a level of context awareness.

Semantic analysis using context awareness technologies has over the years been a challenging task due to the changeability and impulsiveness of mobile phone user movement. Being able to predict the location of a user or identify where a user is at a particular part of the day, can benefit us in terms of determining the pattern of movement or social behavioral patterns, trace criminal activity, identify popular locations or even predict the next location etc.

1.2 Objective

This focus of this research work is on exploring existing data mining techniques in analyzing the Big Data. The primary project objectives are as follows:

1. To investigate two different clustering algorithms to extract information from mobile data to identify key locations
2. To use classification algorithms to perform semantic analysis for key locations detected by the clustering algorithm

1.3 Motivation

The motivation for this research was derived from (J. K. Laurila, Jun. 2012) paper titled “The Mobile Data Challenge: Big Data for Mobile Computing Research”. The article provides motivation for using unique longitudinal smartphone datasets for interesting findings, addressing perspectives such as predictability of human behavior patterns, mobility patterns and determining relationships amongst human mobility patterns and global warming, etc. Mobile computing data is rich and diverse with context aware information which can be used to derive novelty.

Further motivation for this research was provided by the need for novelty in location aware computing as a component of context aware mobile devices. The urgency was further emphasized by the recent case of a missing airliner. While there exists, a potential of identifying a particular device on the network and obtaining data while in transit, be it an aircraft, vehicle or person, there still is potential to investigate unexplored dimensions of context aware computing for determining location. The possibilities are intriguing and immense.

In our case that motivation can be described as identifying the important locations such as home, workplace, shopping mall etc. or places of interest for the research participants, from the raw movement data of a mobile device. One of the main issues in addressing our motivation statements is performing data mining in large temporal and spatial datasets with the presence of outliers and association of several coordinates for a single location. For example, if a user moves to a location for short time or travels home via a different route, then most of location coordinates will be outlier in the movement dataset.

Therefore we use our citywide (J. H. Kang, 2006) dataset to perform an investigation of our objectives to discover the significance of relationship amongst variables.

1.4 Data collection

The citywide (J. H. Kang, 2006) dataset used in this research is provided by an online research community called CRAWDAD. The author of the citywide (J. H. Kang, 2006) dataset collects traces of his daily movement around the Seattle City limits, described as home, work, lunch, school, etc. for duration of 12hours. Most of the area around the city is covered within Place Lab AP database. Data is collected from a variety of platforms such as laptops, cell phones and personal digital assistants (PDAs).

The Place Lab AP database provides capability for a Wi-Fi enabled device to automatically determine its location by listening to radio frequency signals from known access points and radio beacons. The real, long-term data is collected from three participants using a Place Lab client that was able to record data from access points and GSM towers as the participants moved around the Seattle area.

The author exploits the fact that Access Points broadcasts its own unique MAC address periodically as part of its management beacon. Hence the Place Lab client collects data such as time of access in seconds, coordinates of the access point, access points unique MAC address, connection status, and distance from the access point. Distance is recorded in meters, where a negative distance indicates a southward movement of the participant or movement away from the access point.

Table 1-a Dataset Description

	Description
Attribute 1	Timestamp (s)
Attribute 2	Latitude
Attribute 3	Longitude
Attribute 4	AP MAC Address
Attribute 5	Access Point Name
Attribute 6	Connection Status
Attribute 7	Average Distance from AP

The table above provides a brief description of the dataset which was obtained from the source data. The original data set did not contain labels to identify each attribute.

1.5 Methodology

In this research we have attempted to adopt the data mining reference model presented by (J. Han, 2001) in their text for our model creation and data analysis process. The reference model shows the seven stages of the knowledge discovery process.

The first stage is data cleaning which involves removing duplicates and inconsistent data, followed by data integration where we look as combining multiple data sources. However, in our case we only have one. The original data set was imported into Rapid Miner 5.0. Attributes that appeared as inconsistencies were ignored and duplicates were removed.

The next stage looks at data selection where we select attributes for data transformation, for example the timestamp attribute.

In our data set the timestamp attribute represents the trace if the author per second of movement, the timestamp attribute is presented as epoch time. We convert the epoch time into human readable time to determine the actual duration of the trace using MS Office Excel 2010 application. The following formula was used to convert epoch timestamp values:

Date/ Time = (((Epoch Time/1000)-(x *y))/z) + (DATEVALUE ("1-1-1970"))

Where:
x = No. of hours of Trace
y = Total number of seconds in an hour
z = Total number of seconds in 24 hours

The DATEVALUE function was used to convert a date text to a serial number that MS Excel recognizes as a date, in this case 1/1/1970 was used as Microsoft Excel uses the 1900 date system. The dataset is then used to develop models for data mining.

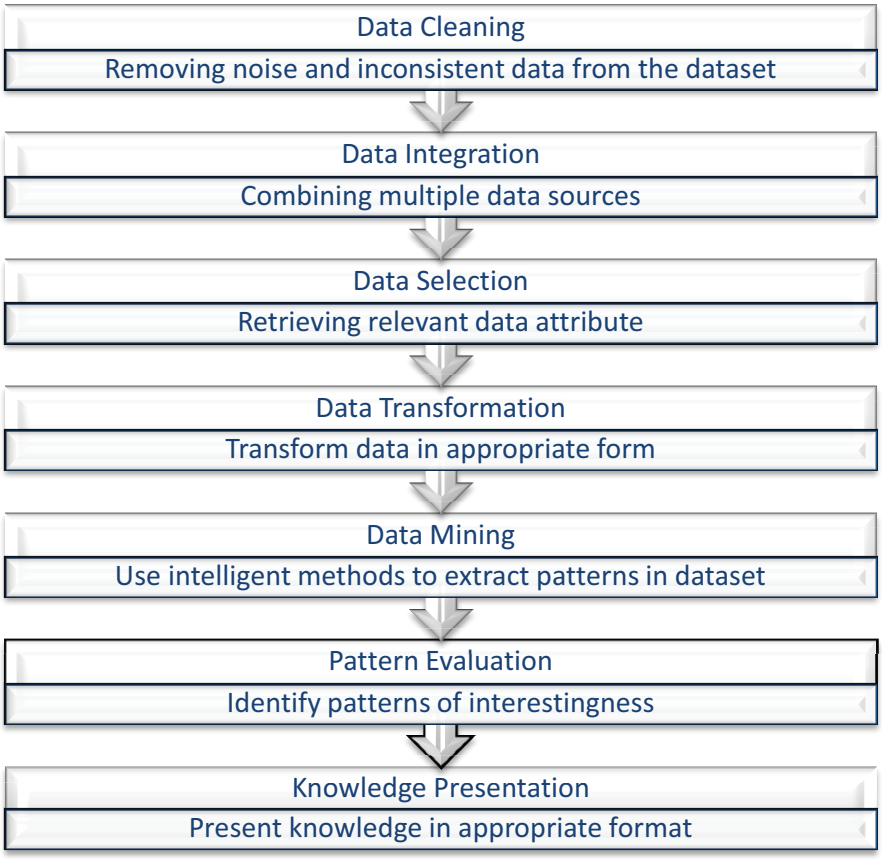

Figure 1-1 Reference Model for Data Mining Process, Adapted from (J. Han, 2001)

The process model takes as input the original citywide (J. H. Kang, 2006) data set and removes the duplicate attributes by subtracting the duplicate entries from the original dataset using the Set Minus operator. The resulting dataset is used as the training dataset. The figure below shows the model that was developed for the process.

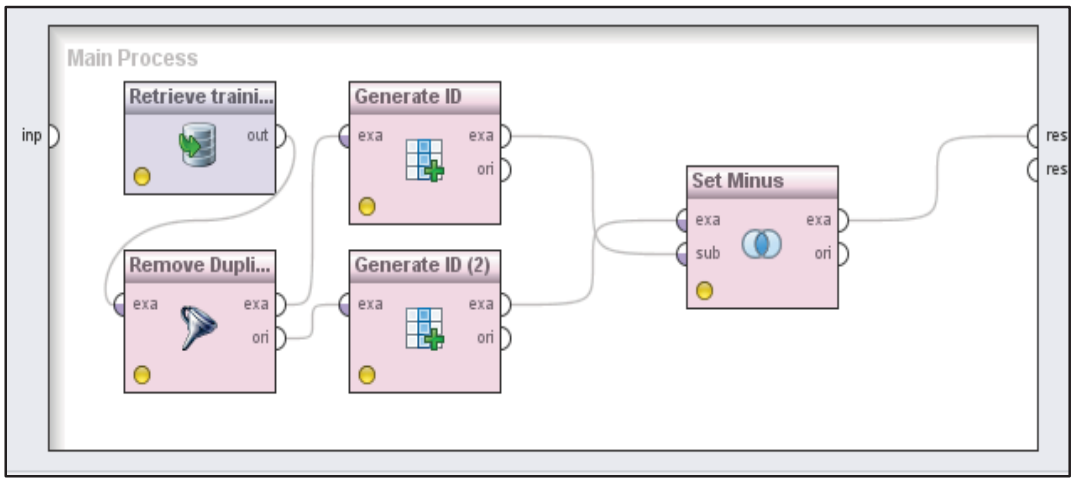

Figure 1-2 Remove duplicates - Process Model developed using Rapid Miner

The model is further expanded as shown in the next page, which uses the K-Means algorithm for clusters formation of selected attributes. The model can be applied to any data type, provided the data type is specified in the "select attribute" operator. The "select attribute" operator identifies the following data types, nominal, binomial, integers, or polynomial.

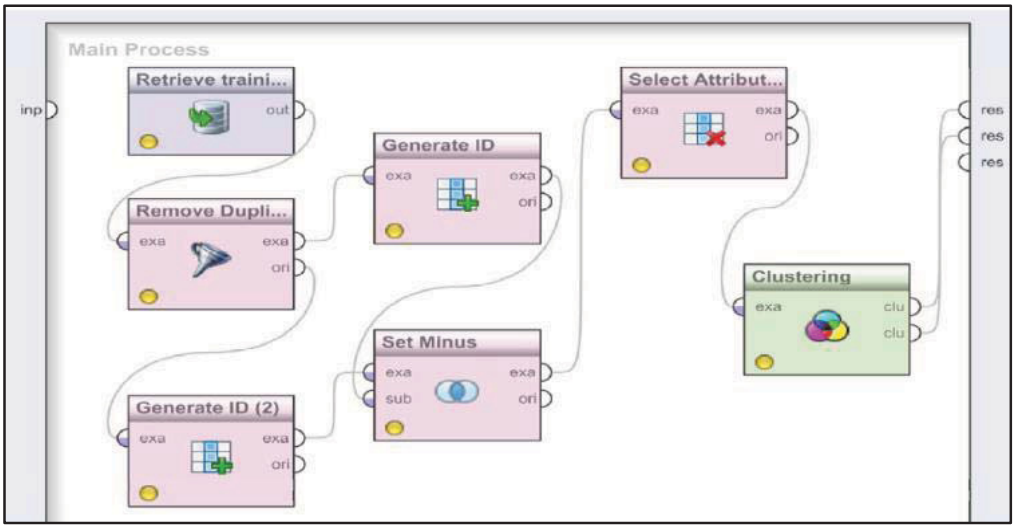

Figure 1-3 K-Means Process Model from Rapid Miner

The resulting data set is split using the 80 – 20 rule, where we randomly select 20% of the data set as the training dataset and the remaining 80% is used as the test set. The training data set was imported into the "select attribute" operator for clustering using the K-Means algorithm. A specific attribute was selected and passed into the clustering operator as parameters along with the number of clusters. The clustering operator resulted in two sets of outputs; data tables for individual clusters and a cluster distribution diagram. The process was repeated for random values of k, such as {2, 4, 6, 8, and 10}.

We further perform pattern evaluation by analyzing the recorded results for each experiment.

A similar process model was developed for clustering using the K-Medoids algorithm. In this setup we used a select attribute operator to select a specific attribute and pass into the clustering operator as parameters along with the number of clusters. The clustering operator resulted in two sets of outputs for the K-Medoids algorithm; data tables for individual clusters and a cluster distribution diagram.

The process was repeated for random values of k, such as {2, 4, 6, 8, 10}. The results were recorded for each experiment for further analysis.

Our method of analysis included analyzing the recorded results for the optimal number of clusters for the algorithms by comparing the Davies Bouldin Index, performance vectors and run time complexity. Results obtained were further examined for novelty where derived k knowledge was presented in an appropriate format.

1.6 Contributions

This research explores existing data mining tools; in this case Rapid Miner 5.0 is used, to solve the given problem, i.e. analyzing a given dataset for semantic analysis. The novelty is in the application of existing tools on new dataset. The design of new algorithms is beyond possibility due to the time limitation and complexity. The main contributions of this research paper are given in the following list.

i) Data mining techniques are explored first-time for semantics analysis of mobility dataset where we try to analyses data for some form of meaning
ii) The evaluation of clustering methods - K-Means, for the semantic analysis
iii) The evaluation of clustering methods - K-Medoids, for the semantic analysis

We begin with reviewing prior works of literature relevant to our problem in the next chapter. The studied literature describes the problem from different perspectives of pervasive computing, such as context awareness, ubiquitous computing, Mobile Ad Hoc networks and location aware computing.

Chapter 3 builds up on the model design that we have used for the experiment. We further discuss the K-Means and K-Medoids algorithm, both being partition based clustering algorithms. We discuss the parameters and their operational efficiency expanding on to data pre-processing.

In chapter 4 we look at the experiment setup and how the training set was used to obtain the optimum value for k for both K- Means and K-Medoids. In this chapter we

further describe our test set and use the “k” clusters to perform our experiments and present our results.

Furthermore we perform analysis of results in Chapter 5. In this section we perform centroid mapping to determine inter and intra cluster similarity to determine the quality of the clusters obtained from our experiments.

Chapter 6 extends to performance evaluation of K-Means and K-Medoids algorithm using parameters like time complexity analysis and utilizing the correlational coefficient to determine a better relationship among the variables for the two algorithms prior to presenting a discussion on the results.

Finally in Chapter 7 we present a recommendation and suggested further works based on our results and discussions. The overall objective of this being able to utilize data mining tools, techniques and processes to determine if our research problem can be reasoned as a conceivable phenomenon.

1.7 Summary

The chapter gives a brief introduction of the basic terminologies relevant to the research work. The problem that this research work addresses is defined in this chapter. The main contribution of this research work has been highlighted. The description of dataset used for the semantics analysis is given along with brief on methodology and data mining processes.

Chapter 2 Literature Review

In this chapter we determine supporting knowledge by examining preceding works of literature performed relative to our research problem. In order to establish a foundation for our research activities in terms of deriving meaningfulness from a set of mobile phone data, we look at the methodology the authors have used and we attempt to obtain their perspective on semantic analysis.

The foundation for this research obtained from (J. K. Laurila, Jun. 2012) the Mobile Data Challenge initiated by Nokia, where large scale mobile data is used for examining issues in behavioral and social sciences in humans suggesting that mobile phone data can be used to describe and recognize real-life phenomena with respect to unconventional mobile capabilities and innovations in technology. The research focuses on smart phone data obtained from 200 volunteer users, over a 12 month period, for the purpose of determining correlation between human mobility patterns, behavior and external environment variables. (J. K. Laurila, Jun. 2012) Highlights in their research the need for open data sharing and using a proactive and holistic approach when dealing with privacy issues in mobile phone data due to "ubiquity of mobile phones and the increasing wealth of data generated from sensors and applications", giving rise to context-aware computing. In this case the authors attempted to perform the following tasks: semantic place prediction, next place prediction and demographic attribute prediction. The literature uses classification and prediction methodologies to gain novelty.

Over recent years our societies have become increasingly tech savvy, creating high demands for greater information accessibility, mobile and portable devices**.** (Budiarto, 2002) In Fiji alone there is huge competition among mobile phone vendors to improve information accessibility visible in the form of continuous upgrade of mobile phone platforms and technology. We have seen the evolution of mobile phones, analogue phones, digital phones, IP phones, iPhones, smart devices, mobile phone platforms, and much more to facilitate our insatiable need for improved information accessibility.

2.1 Big Data

Big data refers to the exponential growth and availability of structured and unstructured forms of data; examples of such data types include data from radio signals, mobile phone call logs and traces, satellite handshakes and so on. Data from collected from such ubiquitous and pervasive sources overtime constitutes to 'Big Data', and is represented by volume, velocity and variety.

Technical developments in context computing (location–aware systems) and increased information accessibility has resulted in large amounts of data which can be used to support better decision making, reduce costs and risks and potentially result in greater operational efficiencies. Over the recent years a lot of emphasis has been given to 'Big Data' for the purpose of accurate analysis, hence the prime concern is to analyze data in the most accurate and effective way.

One of the foremost advantages of big data is that it can be used with dynamic analytics systems to model data. For example organizations like SAS and IBM develop systems and platforms to support storage and analysis of big data as it comes in large volumes; these can be from machines, applications, social media, etc.; data also changes at a very fast pace, requiring the user to react at the appropriate time to analyze it effectively. Data comes in various forms as text, numerical, emails, audio, video, images, financial transactions, etc.

Some of the ways in which we can analyze these large data set as stated by (J. Han, 2001) include association rule mining, predictive analysis, data cube aggregation, classification and clustering and segmentation. In the following section we will look the different algorithms and methods used by various authors to perform analysis on their datasets and present our own method for data analysis and modeling.

2.2 Context Aware Computing

Context-aware computing has been around since the 1990s, it deals with location sensing, social networking, mobile technology and smart devices. Context awareness was derived from pervasive computing; it deals with changes in the environment with respect to computer systems, (Attribution, K-mediods, 2014) suggests it is applied mostly appropriate to mobile devices. Mobile related location aware systems provide us with an immense amount of data that can be used to predict patterns of interestingness or identify hidden knowledge in the data. User information obtained from context aware applications and the relationship with user environment can be used to improve interaction patterns or obtain knowledge on surrounding factors. Location awareness is a specialization of context aware computing, which is applied most efficiently in most industries around the world and the Pacific. For example in Fiji, context aware computing is used by organizations such as Fiji Water Limited to detect the location of trucks, when on route a consignment delivery, the location detection system on the trucks keeps a tab on the movement of the vehicle. Other organizations like MWH Global use the GPS Tracking system on their vehicles to detect location, time and speed. (Creative Commons Attributions, 2014) In United States recent development in context-aware computing has seen the introduction of applications such as “Alliance” for use in connected car systems, connected homes, etc. Another example can be using a popular smartphone application such as ‘Citi Bike’ (Creative Commons Attributions, 2014) for ‘IPhones’ to find the nearest bike station or coffee shop from your present location. The large availability of such spatial and temporal data can be used for identifying a phenomenon.

A phenomenon on an interesting aspect such as that of the mystery of a missing airliner, the first of its kind in the history of the airline industry that has left a void for authorities to obtain information from the bits and fragments of satellite data and handshakes in order to identify the right aircraft, its flight path information and location. There is a continuous need for improved information systems technology and evolution of conventional data mining techniques, where we can utilize novel techniques for semantic analysis for location awareness.

Semantic analysis exploits mobile phone data for location awareness to obtain real life phenomena e.g. the last known location of a particular device. The last ten years has seen significant development in mobile phone and smart devices technology leading to context-aware computing. (E. Turban, 2011) Suggests semantic analysis of data obtained from e-business models that use context awareness technologies can be used to add competitive advantage to the business.

The author in (A. Ciaramella, 2010) proposes a situation–aware service recommender which could be used to locate services. Smart phone devices are capable of storing many applications, increasing the interactive ability of the device. In this literature (A. Ciaramella, 2010) develops a context-aware service recommender for mobile devices, consisting of a fuzzy engine and semantic engine in the server side, whereas the client side consists of label-based resource access. The model suggested by the author comprises of a set of rules and inferences, the rules are expressed as Semantic Web Rule Language (SWRL) and characterized by fuzzy conditions, the application also makes use of spatial and temporal databases to identify the location of a user from a particular place of interest. The most interesting aspect of this literature is that the author presents his ontology in the form of a case study.

In a recent study (E.R Cavalcanti, 2013) investigates weaknesses in mobility models such as Mobile as hoc Networks, in this case refers to moving objects such as people or vehicles, taking into account six different models and seven mobility metrics.. The author uses average node degree metric, which accounts for the number of nodes within a communication range, the relative speed between nodes, and the duration of each connection. (E.R Cavalcanti, 2013) Performs his study using the classification method where simulation and regression analysis is performed to determine the gaps in modeling real time behavior of moving objects (MANETS). The simulation results show a linear positive relation between node speed and number of link changes between nodes on a MANET, however their results report a negative correlation between node speed and duration of the link. However, these techniques are not applicable for a building the context-aware model based on raw location data, as these techniques rely heavily on classification methods.

In 2012 (S. Yazji, 2013) performed a study to address the problem of intrusion detection for mobile devices through correlation of user location and time related data. This was of prime significance as smart devices contain a lot of personal information and if stolen can incur significant loss. The author adopts the use of two statistical profiling approaches for modeling the behavior of users. (S. Yazji, 2013) Further exploits the fact that most portable devices are '*equipped with location identification tools'*, this can be used to trace users, in this case 100 users over a 9 month period. The experiment develops two models, where the first is based on Location-In-Time probability measure and the second is based on Markov transition property using Row-merge algorithm and Minimum Description Length Principle (MDLP) algorithm. The author creates profiles of users and then detects attacks by recognizing any deviancy. User profiles consist of location identification and the number of minutes spent at each location, the probability of visiting a particular location. These location-based models are suitable for a system with refined stream-data—which is not possible to generate from huge location data of a device based on the mobility of the user per day.

In another study (N. Bicocchi, 2008) suggests that mobile devices can be used to create a dairy for the user's whereabouts which can in turn be used to build a user profile for the user in many applications. The author introduces a "Whereabouts Diary" application/ service that logs places visited by the user, semantically classifying places visited using unsupervised methodology. It collects data such as longitude and latitude and time using i-mate PDA 2K smart phone or a HP IPAQ RX3700 PDA connected via a Bluetooth GPS Reader. The 'Whereabouts dairy uses the Bayesian Network Classification method consisting of four nodes to collect and identify locations. The four nodes are, 'the *weekend* node', 'the *hour* node', 'the *kind of place* node' and 'the *happens* node', these encode the daily routine of a mobile device user. In the first experiment (N. Bicocchi, 2008) verifies the accuracy of the algorithm and compares it with the A-S algorithm. Next they verify results using reverse geocoding service, further more they use the discovery algorithm to cluster locations within 10 meters of each other, in this case they look at addresses of buildings. The third experiment conducted by the authors was to evaluate the performance of their algorithm by mapping results against places on the white-pages service that resulted in 40% efficiency. However, overall the authors were able to

semantically classify locations with 64% efficiency using GPS traces for the 'Whereabouts Diary'.

In 2006 (Satoh, 2006) presented his literature on using radio frequency identification (RFID) based systems as an alternative for detecting location of objects. The author explores two scenarios, first using mobile agents that are capable of migrating between public terminals and second assuming the user is carrying a PDA. The suggested framework consists of three parts, namely, location information servers, mobile agents and agent hosts. The location information system used by (Satoh, 2006) manages location sensing systems and consists of the following functionalities the RFID-based location model, location management and location dependent deployment of agents. The author also discusses the 'TaggedAgent' algorithm for abstract class object that was implemented at the service provider end to implement the proposed system. Finally the author plans to develop a methodology to test applications based on the proposed system.

A client middleware architecture is presented by (M. Malcher, 2010) which supports dynamic deployment of context and location awareness as most such applications are mobile platform specific. The proposed architecture features development of new components and switch between existing components, context-awareness, interface uniformity and provides support for collaboration services. The components of this architecture are as follows component manager, adaptation manager, context manager, Shared data manager and MD-ECI which supports remote distribution of notifications of publications as explained by the author. The context manager searches for the required services, and activates the service when found. The adaptation manager issues basic activation and connection requests to the context manager, this process is followed by implementation of the MD-ECI client by the shared data manager and finally the context management service manages collection, processing and distribution of context data. Furthermore (M. Malcher, 2010) suggests that his architecture can be implemented in location based collaboration applications such as 'Geo-Tagging', 'Track Service' and 'Bus4Blinds'.

The works of literature examined so far presented application based cases by the authors. The primary concern so far has been the ability of mobile devices to be

location aware, collecting traces, etc. However we are given the scope of using the information that these applications, process and devices collect for the purpose of analysis (S. Isaacman, 2011) States, the density of contextual information availability can be analyzed for meaningful locations using clustering and regression algorithms to explore the data sets to identify homogenous groups of objects called clusters. These homogenous groups were data sets with similar characteristics. For example trying to identify most popular locations or predicting the mostly likely place where a user may be at a particular part of the day. The experiments in this case analyze call detail record data from cell phones. The dataset is obtained from specific geographic regions of the New York and Los Angeles metropolitan areas. The location identification algorithm presented by (S. Isaacman, 2011) has two stages, the first stage is responsible for developing spatial clusters and the second stage identifies important clusters. The author(s) further implemented the use of the Hartigan's leader algorithm to form clusters of similar objects within a one mile radius from the cluster centroid. Furthermore an algorithm was developed by the authors to determine the importance of a particular location using logistic regression. Some of the factors considered in determining the importance of a cluster were, the size of the cluster, the duration of connectivity to a particular cell tower, the highest number of days a cell tower received connection. Validation for the important places algorithm is done using the nearest cluster algorithm taking into account all identified clusters. In this composition (S. Isaacman, 2011) extends the possibility to identifying HOME and WORK from the important places identified. The authors use the Home and Work Oracle algorithm to commute distance from and person's home and work locations. In their last illustration the author(s) present an extension of their Oracle algorithm to estimate carbon footprint for New York and Los Angeles. Therefore, implying we can indeed use mobile phone data for semantic analysis for the purpose of knowledge discovery.

(L. Lao, 2007)Supports the notion by proving how a person's activities and significant places can be extracted from traces of GPS data. The author(s) research work focuses on 'learning patterns of human behavior from sensor data'. The objective of this research was to provide information on the amount of time spent by the user at locations such as work, bus stop, his home and while travelling. The approach presented has the following key features, takes in users' context into account and

detects significant places, interprets user data and uses inference and learning algorithms. The proposed model is extended by use of 'approximate inference algorithms for random fields' where the model is constructed based on the resulting inference. (L. Lao, 2007) Extends his approach using the Markov models to develop a probabilistic model capable of extracting high level activities from the GPS traces, this is further extended using Conditional Random Fields (CFRs), a discriminative model to represent distribution over hidden states. Furthermore, belief propagation (BP) algorithm is used to determine the correct distribution probability. Other techniques used by the author(s) in this experiment include map estimation, parameter learning, maximum likelihood estimation, maximum-pseudo-likelihood estimation and parameter sharing modeled using probabilistic relational models. The place detection algorithm presented by the author takes GPS trace as input, generates activity segment by grouping the GPS readings into spatial segments, it generates CRF instances and the determines map sequences. It then generates places by clustering significant places, performing map inferences until the dataset is complete. The algorithm is able to identify places of significance visited by the user over a one week period, such as Bus stop, home, work, sleep, visit, etc. The author claims 90% accuracy of their suggested method, hence contributing to the subject of semantically analyzing data for location awareness.

Context aware computing has its benefits, (J.A. Burke, 2006) in his literature introduces the concept of participatory sensing, employing ubiquitous devices as sensor nodes and location aware data collection instruments suggesting location and time synchronization data capability of GPS and other technologies. (J.A. Burke, 2006) In his work develops the new 'partisan architecture' through abductive reasoning suggests applications in areas such as public health, urban planning, cultural identity and creative expression and finally natural resource management. The author presents opportunities created by participatory sensing at community level such as heavy traffic, correlation of environment data with government, health care providers, or basic human activity data. A simple example would be the relationship between air quality and health depending on user density.

Other possible avenues of using location awareness is presented by (L. Calderoni, 2012) where the author introduces the smart city concept capable of citizen

interaction. For example, sensor networks can be used to integrate knowledge from the Internet of Things (IoT), Internet of Services (IoS) and Internet of People (IoP) where data from traffic sensors, weather stations and security cameras can be combined to provide users with 'real-time traffic information and route selection based on the current state of the urban route network' (L. Calderoni, 2012). The authors present in this paper 'an around me' application which is able to provide information based on the current location of the user, such as location of hospital, tourist landmarks, rest rooms, etc. The proposed application is a mobile client application on the Android platform consisting of the main activity, details activity and the service. The server side of the proposed application consists of JavaScript Object Notation (JSON). The application also uses MySQL Cluster and Range Query Optimization and a Java GeoLocation class to compute locations. The works of literature studied so far indicate a lot of potential research opportunities relative to location aware technology.

2.3 Summary

In this chapter we have looked at context aware computing, mobile ad hoc networks and location aware computing. From the presentation of articles in this segment we can make an assumption that it is possible to use these technologies in conjunction with mobile phone networks to determine location. We have also come across real time cases where this has indeed been put into practice, hence supporting our research notion.

Chapter 3 Model Design

In the literature examined so far, we have come across a number of techniques that can be used for performing semantic analysis for location awareness. We gain insight on using semantic web rule language (SWRL) using fuzzy logic, classification techniques using simulation and regression analysis and Correlational analysis. Other methods we have come across include use of a statistical profiling approach, such as the Markov Transition, Minimum Description Length Principle (MDLP) and the Hartigan's Leader algorithm. In our research we came across more than one author who has used the Bayesian Network classification method for semantic analysis, in combination with novel algorithms such as the TaggedAgent algorithm and the belief propagation (BP) algorithm. The novelty in our research is that we will try to solve the semantic analysis for location awareness using the K-Means and the K-Medoids algorithm.

In this chapter we present our model design for our experiments. Clustering is an example of unsupervised learning where class labels are undefined; learning is conducted by observing cluster behavior (J. Han, 2001).To explore the semantics for mobility, clustering seems an intuitive approach to extract the group of places that were part of the movement by the mobile device holder.

In the following sections we discuss the two methods, K-Means and K-Medoids, identifying their possible parameters and performance efficiency. We also look at data preprocessing techniques that we have used in this experiment.

3.1 Types of data mining techniques/algorithm

(A. Berson, 1999) In an excerpt from his book states most common data mining algorithms used today are of two types, classical techniques and next generation techniques. Classical techniques include methods such as statistics, neighborhood and clustering on the contrary next generation techniques include methods like decision trees, networks and rules. Other method of data mining includes sequential pattern and predictive analysis.

Data mining is an etymology for knowledge discovery in databases. It describes the entire range of big data analytics including collection, extraction, analysis and statistics as stated by (Rijmenam, 2013). The author presents in his article five state of the art techniques that can be used in big data analytics such as outlier detection, association rule mining, cluster analysis, classification analysis and regression analysis.

Outlier detection deals with inconsistent dataset pattern. These are also known as anomalies. Outliers contaminate the dataset whereby experiments and analysis may not result in an accurate outcome. Hence it is important to remove inconsistent data entries. (Rijmenam, 2013) Adds that 'Classification takes information, presents it and merges into defined groups. Clustering eliminates distinct groups and allows data to classify itself based on similarity. Regression analysis focuses on the function of the information and models the data based on concepts. And finally Association rule attempts to discover relationships between various data feeds.' In this research we use the K-Means and K-Medoids algorithms for clustering.

3.2 K-Means Clustering

The K-means Clustering method is a centroid based model, it is the simplest method used for unsupervised learning. It is known to have higher efficiency in large databases, which finds clusters of similar spatial (longitudinal, latitudinal) magnitude, i.e. it uses vector quantization. (Attributions, K-Means Clustering, 2013) The model returns clusters of similar size, where entities are assigned to the nearest centroid; it partitions data into a Voronoi diagram structure, i.e. partitions data points into regions based on the closeness or similarity of the data points within a subset. The centroid is the mean (μ) value for each individual cluster, implying that the distribution of items for that cluster is within the proximity of the μ value. It takes in as **parameters**, the **# of iterations** and a **value for *k***. (J. Han, 2001)

The k-means algorithm as explained by (Attributions, K-Means Clustering, 2013) is as follows:

Make initial guesses for means $m_1, m_{2,\ldots,} m_k$
Do while there are no changes in any mean

Use estimated mean to classify samples into clusters

For i from 1 to k

Replace m_i with the mean of all the sample for cluster i

End for

End While

(Attributions, K-Means Clustering, 2013; Corno, 2012) K-Means alternates between two phases, the assignment phase and the update phase. The assignment phase allocates data objects into groups with the lowest mean(μ)using the smallest squared Euclidean distance measure for objects within the cluster, the update step calculates new means as centroids of the observations for new clusters as objects are added.

3.3 K-Mediod Clustering

K-Medoids is a variation of the K-Means algorithms also known as partitioning around Medoids. This partitioning technique clusters the dataset of *n* objects into *k* clusters (J. Han, 2001; Mirkes, 2011). The K-Medoids technique deals with noise and outliers more efficiently as it reduces the 'sum of pairwise dissimilarities' (Attribution, K-mediods, 2014; Mirkes, 2011). It works best with dissimilarity matrices, such as (x,y), the algorithm takes in parameters such as *k*, number of clusters. It uses an actual point in the cluster as the center of the cluster (X. Jin, 2010). It is the most centrally located object in the cluster that has the minimum distance from all other points in the cluster. Minimum distance from other points is determined using the 'Manhattan distance' measure (Tibshirani, 2013).

The K-Medoids algorithm as explained by (Mirkes, 2011) is as follows:

Randomly select k form n data points as Medoids

Do while no change in assignment

Assign data point to nearest medoid

For i from 1 to k

Calculate average dissimilarity between a data point and all other data points in the cluster with the medoid

Select data point with the lowest dissimilarity as new cluster center

End For

End While

3.4 Methods for determining 'k'

K-means and K-Medoids algorithm require prior identification of (*k*), i.e. number of clusters for partitioning of the dataset. With reference to (Attributions, Determine the Number of Clusters, 2014) one of the proposed ways to determine k is using the rule of thumb as shown below.

$$\textbf{As a rule of thumb } k \approx \sqrt{n/2},$$

Where: n is the size of the data, k represents the number of clusters

This however, can be a drawback for k-means, as we may end up with a large value for *k*, for very large data sets if the dataset is not sufficiently transformed while data preprocessing. Another drawback is that this method is very sensitive to extreme values in the dataset, i.e. outlier data points which may affect distribution of objects within the clusters (Attributions, K-Means Clustering, 2013).Other methods for determining the value of *k* include, the **'Elbow method'** (Attributions, Determine the Number of Clusters, 2014) which looks at percentage variance where percentage variance is plotted against the number of clusters, the number of clusters at which the marginal gain drops is chosen as the '**elbow criterion**' (Attributions, Determine the Number of Clusters, 2014). Another method explained by the author includes techniques like the '**Gaussian Mixture Model**', '**Bayesian Information Criterion**' and '**Deviance Information Criterion**' (Attribution, Deviance Information Criterion, 2014).

We use the 'Davies-Bouldin Index' to evaluate the k-means and the K-Medoids algorithm. As presented by (Attribute, 2014) the Davies-Bouldin Index is 'the ratio of within cluster scatter and between cluster scatter'. The author indicates a lower value for Davies Bouldin index implies better clustering and that objects within the dataset share similar characteristics. (Attribute, 2014) Further suggests the Davies-Bouldin Index as an ideal measure for determining the number of clusters in a dataset. The author indicates that the index can be plotted against the number of clusters it has been calculated, the number of clusters for which the Davies-Bouldin Index is lowest is the optimal measure for the ideal number of clusters for the dataset.

3.5 Normalization

Database normalization refers to the process of organizing data to reduce duplication. The set of original data is mapped onto another scale (Attributions, Database Normalization, 2014; Howcast Media Inc, 2014).

The authors further expand their normalization technique where by in the first step they identify the smallest and the largest number in the original dataset, followed by identification of the smallest and largest values on the normalization scale, an example would be a scale of 1 to 10 and the final stage is to calculate the normalized value using a predefined formula. In the above steps as identified by (Howcast Media Inc, 2014), the author uses suggest the use of the following formula:

$$y = 1 + \frac{(x - A)(b - a)}{(B - A)}$$

dapted from source **(Howcast Media Inc, 2014)**

Where:

y = the normalized data

x = the actual data point from the original dataset

A = the smallest data value from the original dataset

B = the largest data value from the original dataset

a = the lowest normalization scale value

b = the highest normalization scale value

The above formula however is applied to data values of integer data type. The citywide (J. H. Kang, 2006) data set used in this research contains the following data type for each attributes:

Table 3-A Data Description with corresponding Data Type

Description	Data Type
Timestamp (s)	Epoch Date/ Time
Latitude	Real

Longitude	Real
AP MAC Address	Polynomial
Access Point Name	Polynomial
Connection Status	Binomial
Average Distance from AP	Integer

Hence for our dataset we opted to normalize the Timestamp (s) value that was represented using Epoch date and time format. Normalization for the timestamp attributed involved conversion to human readable format. The conversion formula used has been presented in the methodology section of this paper where the timestamp values have been normalized to a range from 6th June 2004 3:53:03 am to 6th June 2004 4:43:49 pm.

3.6 Data Cleaning

Data cleaning refers to the process of identifying and correcting corrupt and inaccurate data from a dataset (Attributions, Data Cleansing, 2014; J. Han, 2001). Like every dataset our citywide (J. H. Kang, 2006) dataset also contained some errors where we experienced missing values for attributes like access point names, connection status and average distance in meters from access points. The possible range of values for each attribute was identified, for example the 'Average Distance from AP' had a normal data range from 0 to -100 any other value in the range was considered inconsistent. Similarly for 'connection status' the possible values identified were 'Yes' or 'No' and so on.

The original citywide (J. H. Kang, 2006) dataset contained 208995 entries, the missing and inconsistent entries were removed by filtering the dataset using the Microsoft® Office Excel spreadsheet application, 3773 inconsistent entries were identified and removed from the data set, this constitutes to 1.8% of the original dataset.

The Table below presents a breakdown of the data cleaning process.

Table 3-B Data Preprocessing: Attributes Discrepancy as noted using Rapid Miner® 5

Missing Values for Dataset		
	Description	**# Missing Values**
Attribute 1	Timestamp	0
Attribute 2	Latitude	0
Attribute 3	Longitude	0
Attribute 4	MAC Addresses	0
Attribute 5	Radio Beacon	408
Attribute 6	Connectivity Status	3356
Attribute 7	Average of Retrieved Coordinates	3356
N – original	verified by Rapid Miner	208995
N – cleaned	verified by Rapid Miner	205222
Overall Missing Values	*** some of the missing values overlap	**3773**

3.7 Data Preprocessing

Data preprocessing techniques increase the quality of the data improving efficiency and accuracy of mining algorithms (J. Han, 2001). Noise in the dataset is detected by identifying extreme and inconsistent values. For example extreme values for longitude and latitude that are not consistent with the rest of the dataset.

The reduced dataset was imported into Rapid Miner 5.0 data mining tool for further preprocessing. In this phase duplicate objects were identified using the '**Remove Duplicate'** operator which resulted in a subset of the original dataset consisting of duplicate objects. Duplicate objects were removed from the original dataset using the '**Set Minus**' function, resulting in an optimized dataset, which was used as the training set for the experiments conducted using K-Means and K-Medoids. The process used to obtain the reduced training dataset is shown in the process diagram below.

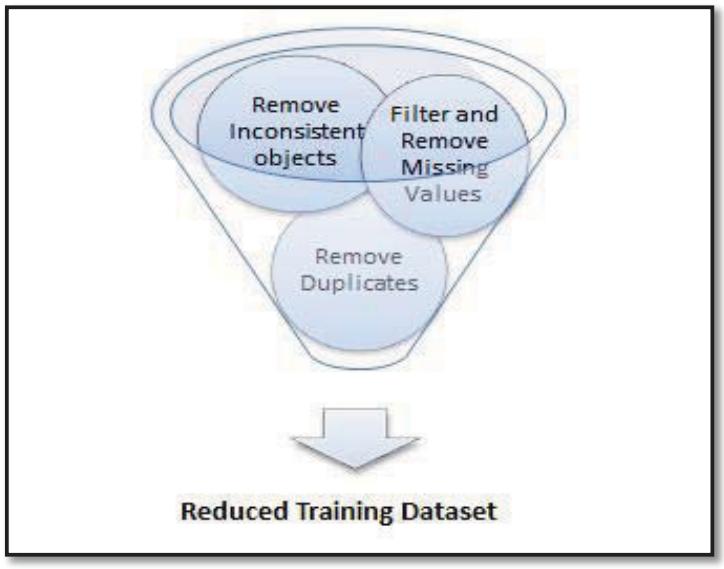

Figure 3-1 Data Preprocessing Illustration developed using SmartArt

3.8 Summary

In this chapter we present our model design by atomizing each component in the model and its behavior. We present different types of data mining techniques and expand on to the K-Means and K-Medoids algorithm, where we give a brief description of both algorithms in the form of pseudo code. Next we discuss possible methods of determining the number of clusters, k, for both the algorithms. In this section we look at the elbow method, the rule of thumb and the Davies Bouldin method and how we can use these to determine the correct number of clusters for k-means and K-Medoids. The later part of the chapter presents an illustration on data preprocessing methods used in the following chapter for our experiments. Furthermore data normalization is discussed stating procedures for normalizing the timestamp attribute from epoch time to human readable format. We conclude the chapter with data cleansing where we prepare the data for our experiment to perform semantic analysis for location awareness using k-means and K-Medoids algorithm.

Chapter 4 Experiment Setup

In this chapter we present a meticulous description of the experiments performed on our training and test dataset derived from the preprocessed data. We first expand on how we derive our training and test data sets and then present the experiment setup for k-means and K-Medoids clustering.

In the preceding chapters we state that the citywide (J. H. Kang, 2006) data file initially was in XML format and was converted to CSV format using a c# program. The resulting data file was imported into a spreadsheet application for normalization and data cleaning.

The next stage was to import the data set into Rapid Miner 5.0 for preprocessing using the data mining tool. The data set was split using the 'Split Data' operator using stratified sampling method.

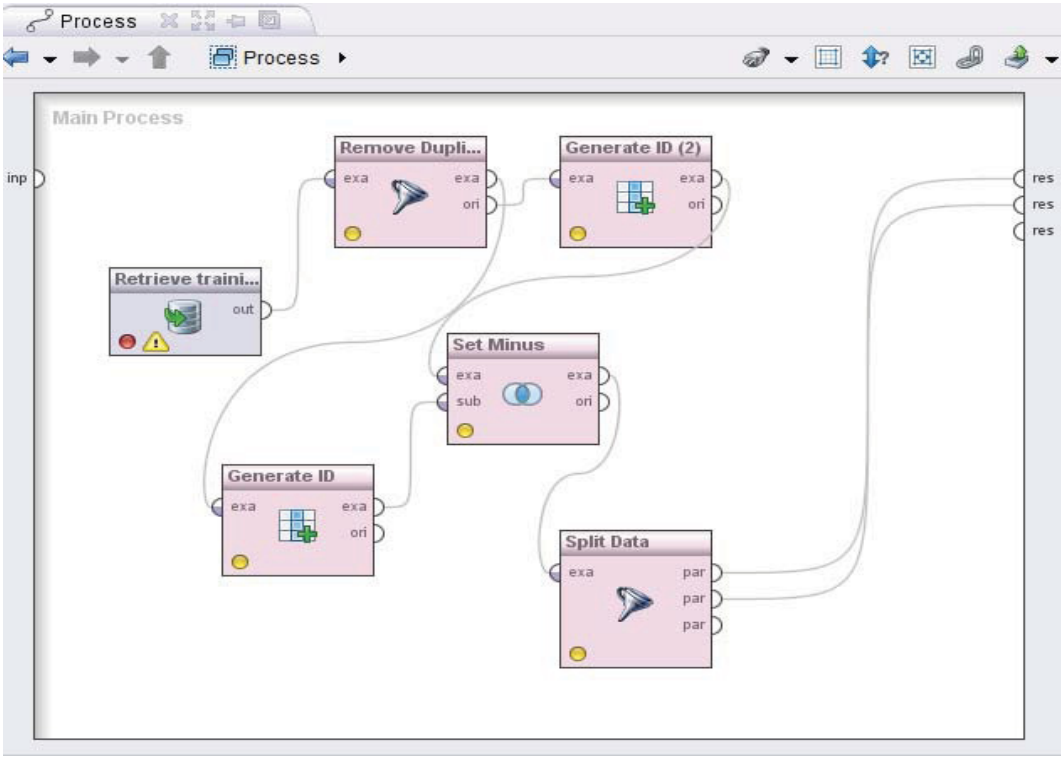

Figure 4- 1 Data Splitting Model in Rapid Miner 5.0

The diagram above shoves the process used to obtain the test and training data set; using stratified sampling we used the default random seed of 1992 for consistency.

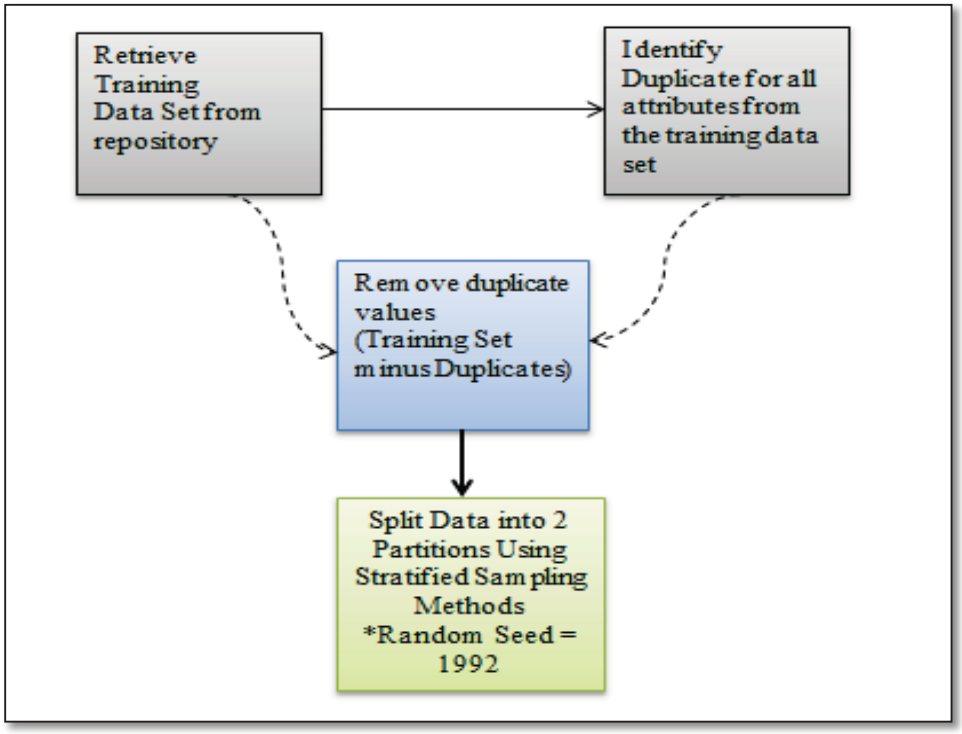

Figure 4-2 Split Data Process Chart using Stratified Sampling

Experiments on preprocessed training and test data were performed using both K-Means and K-Medoids algorithms. A set of six experiments were run using the K-Means algorithm operator in Rapid Miner 5.0 using different values for$k = \{2, 4, 6, 8, 10, 12\}$. The same was repeated using the K-Medoids algorithm operator.

4.1 Conceptual Model for K-Means Algorithm

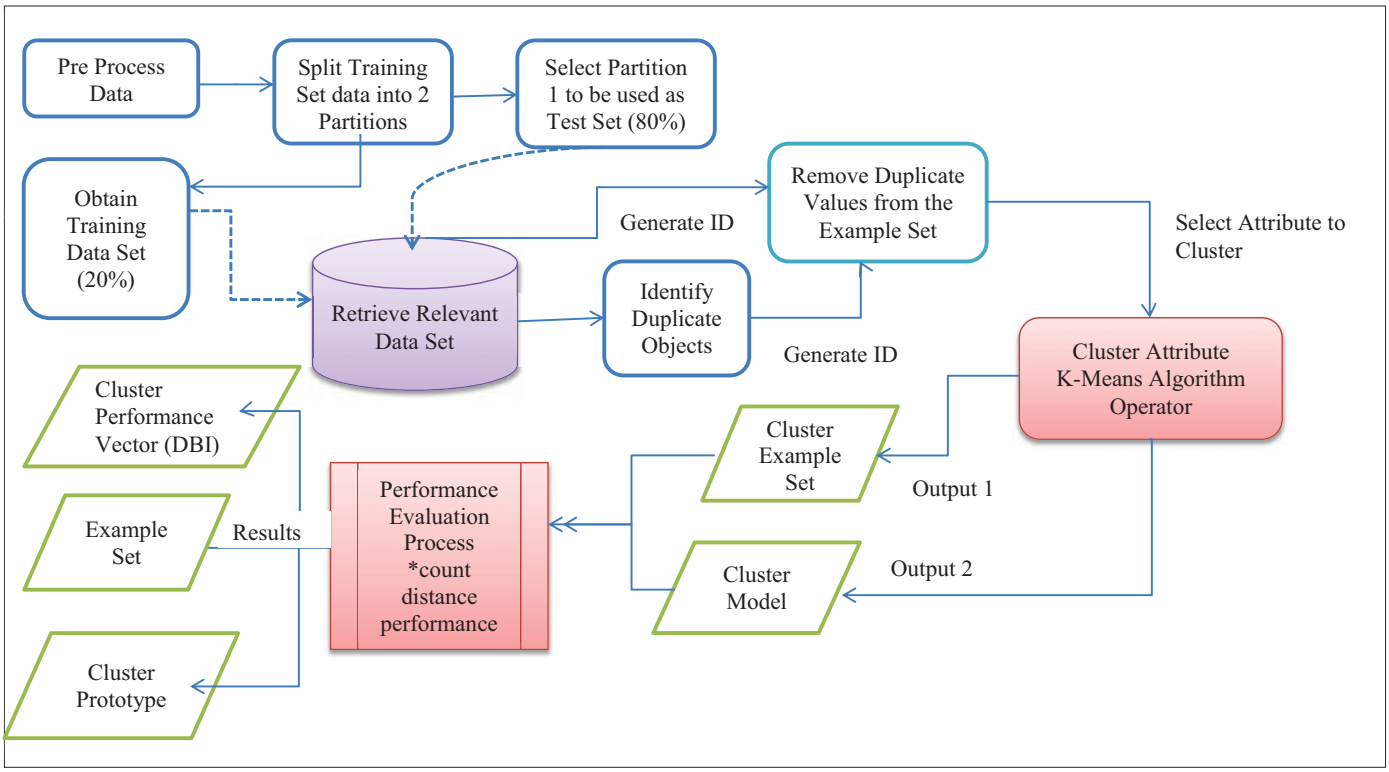

Figure 4-3 Model describing experiment setup for k-means

We present an overview of how we performed our experiment for k-means algorithm in the data flow chart above. The original preprocessed dataset was split using the 80-20 rule, where the 20% of the originally preprocessed data was used as the training set and the other 80% of the data was used as the test set. The 'Split Data' operator used in the experiment used stratified sampling method to randomly split the data.

Once the training data was obtained it was passed into the 'K-Means' operator which had the following constraints applied to get a k-means clustering model.

- Special criteria – k (#. of clusters)
- We used Maximum # Runs as 10 (default setting)
- The Determine Good Start Values criteria was selected
- Measurement type used was mixed measures as our training set consists of both nominal and integer class variables
- Numerical Measure selected was Euclidean Distance measure for integers and determining the Davies Bouldin index
- We used the default setup for optimized number of steps as 100
- The local random seed option was selected for random selection of cluster objects
- Local Random Seed value used was 1992, default value.

The resulting cluster model and example set was further input into the 'Cluster Performance Vector' operator to determine the Davies Bouldin Index (DBI) each time the experiment was repeated for a different value of k. The value for 'k' with the lowest DBI was chosen as the most optimal number of clusters.

In the final stage of the K-Means experiment we used the 'Apply Model' operator to generate clusters for our test set as illustrated in the diagram below.

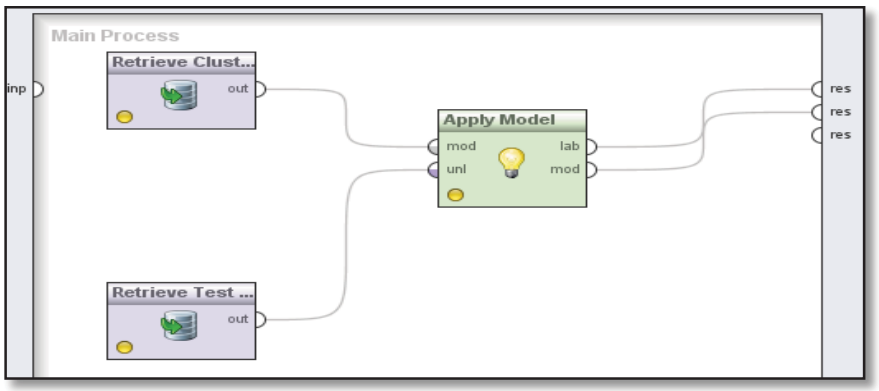

Figure 4-4 Applying k-means clustering model to test set (k=4)

4.2 Conceptual Model for K- Medoids

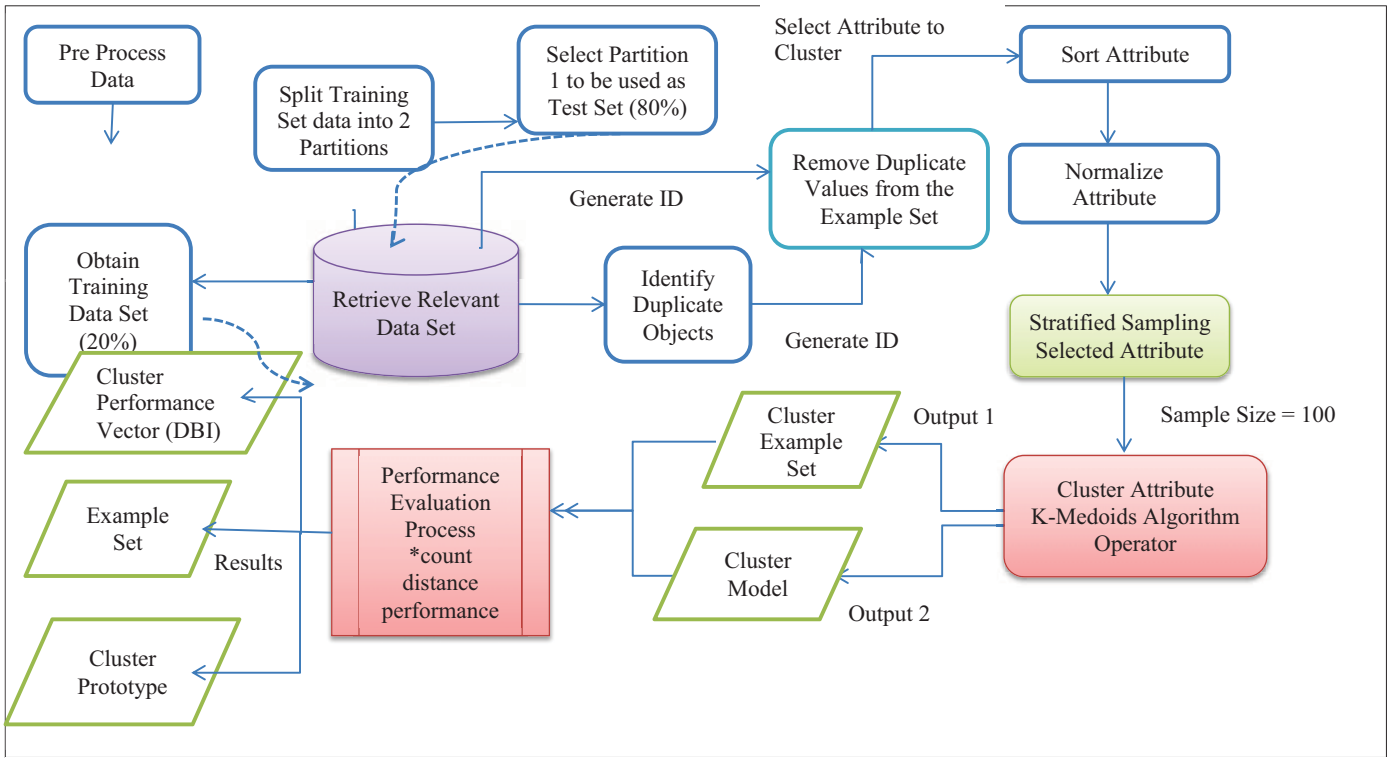

Figure 4-5 Model describing experiment setup for k-Medoids

The data flow chart shown above presents an overview of how we performed our experiment for K-Medoids algorithm. We primarily applied the model to our training data in order to determine the appropriate number of clusters for the K-Medoids algorithm. We passed the training data into the 'K-Medoids' operator to which we applied the following constraints to generate a K-Medoids clustering model.

- Special criteria – k (#. of clusters)
- The add cluster attributes criteria was selected
- We used Maximum # Runs as 10 (default setting)
- We used the default setup for optimized number of steps as 100, consistent with the K-Means setup
- Measurement type used was mixed measures as our training set consists of both nominal and integer class variables, also because K-Medoids is able to efficiently deal with nominal class values (J. Han, 2001)
- Numerical Measure selected was Euclidean Distance measure for integers and determining the Davies Bouldin index, we used Euclidean Distance measure for consistency and standardization for comparison with results obtained for k-means
- The local random seed option was selected for random selection of cluster objects
- Local Random Seed value used was 1992, default value. Consistent with k-means setup.

The resulting K-Medoids cluster model and example set was input into the 'Cluster Performance Vector' operator to determine the Davies Bouldin Index (DBI).the experiment was repeated for different values of k, similar to the k-means model. The value for 'k' with the lowest DBI was chosen as the most optimal number of clusters for K-Medoids. The final stage of the K-Medoids experiment used the 'Apply Model' operator to generate clusters for the test set.

4.3 Summary

This Chapter describes the experiment setup used in this research for the K-Means and the K-Medoids algorithm. The results obtained from the above experiments have been presented in the next chapter along with the characteristics of the different clusters and their significance. Furthermore, the results obtained were analyzed and used to identify patterns of interestingness and build relationships among attributes for location awareness.

Chapter 5 Results

In this chapter we present the results of the k-means and the K-Medoids algorithm clustering method experiment setup. In the first part of the experiment we show results obtained for obtaining the optimal value for k for both algorithms. We then compare results obtained for the training set and test set using k-means and K-Medoids. And in the later part we present performance of the two algorithms.

5.1 Determine 'k' using Training Data

5.1.1 K-Means

The following Table shows results obtained for the Davies Bouldin Index using the K-Means algorithm for clustering. The cluster with the lowest index was chosen as the value for 'k' for the algorithm. The training data set was used to run the experiments to determine k. The value obtained for 'k' was used to run experiments on the test set.

Table 5 – A Results Table for K-Means

# of. Clusters *(k)*	Davies Bouldin Index
2	-99.297
4	-37.049
6	-47.53
8	-40.706
10	-63.884
12	-117.984

The results obtained were also represented graphically for comparison as shown below.

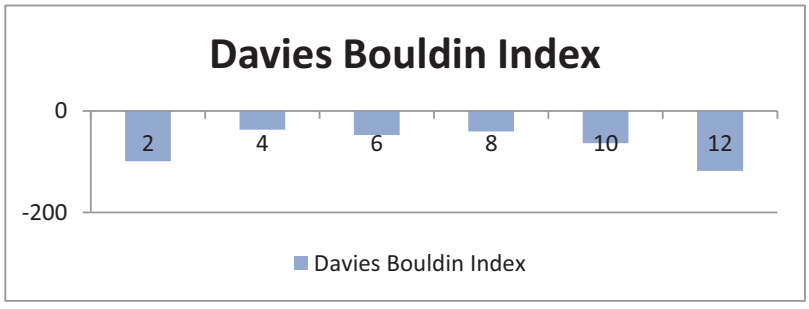

Figure 5-1 Davies Bouldin Index for K-Means

For K-Means algorithm the most appropriate value for ‘k’ that was identified was 4. Hence number of clusters used for analysis and running experiments on the test set is 4.

5.1.2 K-Medoids

The Table below shows results obtained for the Davies Bouldin Index obtained using the K-Medoids algorithm. In this case also the cluster with the lowest index was chosen as the value for ‘k’ for the algorithm.

Table 5 - B Results Table for K-Medoids

# of. Clusters *(k)*	Davies Bouldin Index
2	-0.929
4	
6	-0.362
8	-0.654
10	-0.457
12	-1.287

For the K-Medoids algorithm the most appropriate number of clusters chosen was 6, as the experiments showed the lowest Davies Bouldin Index for k=6. It was also noted that for number of cluster = 4 showed a negative infinity value, implying it was the worst configuration of clusters using the K-Medoids algorithm.

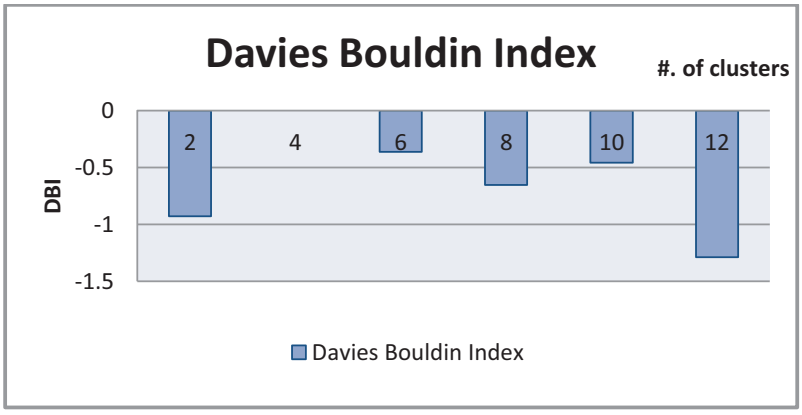

Figure 5 -2 Davies Bouldin Index for K-Medoids

5.2 Cluster Characteristics Training Set

In this section we present the characteristics of clusters produced by the K-Means and K-Medoids algorithm respectively using the training dataset. The K-Means clustering method produced four clusters as shown in the table below, containing the number of items per cluster, the cluster center and average distance within the cluster objects.

Table 5 -C Results Table- Training Set Cluster Characteristics (K-Means)

Average -Inter Cluster Centroid Distance			-12.379		
Davies Bouldin Index			-0.475		
Cluster	**Number of Objects**	**AVG. Distance of objects from Cluster Centers**	**Cluster Centers**	**Timestamp** 6/16/2004	**Access Points where Connection was established**
Cluster 0	13663	-11.091	-72.659	16:20 – 16:36	1011, IR, University of WashingtonCSE
Cluster 1	5993	-17.929	-49.724	12:36 – 12:42	MSHOME, Eric's Network, GTwirelessAHoc
Cluster 2	7730	-10.022	-61.336	11:57 – 12:00	1011, IR, EE
Cluster 3	13658	-12.567	-88.923	6:28 – 6:59	University of Washington, SpeedStream, Default

The table above shows the cluster characteristics produced for the K-Means test set. The table shows attributes such as the peak time of activity, range of access points connected to and the number of data point per cluster. The table also lists the average distance of each object from the cluster centers. The results above were obtained using the Rapid Miner data mining tool.

Table 5 -D Results Table – Training Set Cluster Characteristics (K-Medoids)

Average -Inter Cluster Centroid Distance		-11.069			
Davies Bouldin Index		-0.362			
Cluster	**Number of Objects**	**AVG. Distance of objects from Cluster Centers**	**Cluster Centers**	**Timestamp** 6/16/2004	**Access Points where Connection was established**
Cluster 0	4978	-12.246	-56.715	05:59 – 06:23	TeaHouse_Public, UniversityofWashington
Cluster 1	10130	-10.515	-88.923	12:15-12:35	1100,Sam Langford Network
Cluster 2	5009	-9.407	-62.316	06:23– 06:59	BLUMECO, Netgear,linksys
Cluster 3	2370	-11.404	-93.441	14:29-16:59	Sahara, Kyle Home, Seattle Best Coffee
Cluster 4	11958	-13.095	-77.689	11:57 – 13:45	1100, IR, University Of WashingtonCSE, GTwirelessAdHoc
Cluster 5	6599	-9.751	-49.724	None	None

The K-Medoids clustering method produced six clusters as shown in the table below the same criteria as K-Means was used to derive the results. The table further describes the characteristics of clusters obtained using K-Medoids, such as the peak connection times, and range of access points with which the mobile user established connection.

5.3 Cluster Characteristics Test Data

In this section of the paper we present the results obtained by relating the k-means and the K-Medoids clustering model to the test data representing 80% of our preprocessed data. The table below shows the results obtained when the k-means clustering model was applied to the test data using k = 4 as the number of clusters.

Table 5 -E Results Table- Test Set Cluster Characteristics (K-Means)

Average -Inter Cluster Centroid Distance		-12.469			
Davies Bouldin Index		-0.477			
Cluster	**Number of Objects**	**AVG. Distance of objects from Cluster Centers**	**Cluster Centers**	**Timestamp Most Common** 6/16/2004	**Connected Access Points Names**
Cluster 0	54359	-12.632	-88.937	6:28 – 6:59	101, 206-726-8032
Cluster 1	23808	-18.444	-49.702	12:30 – 12:42	HCK344, Eric's Network, GTwirelessAHoc
Cluster 2	31411	-10.103	-61.313	11:40 – 12:10	101, 206-726-8032
Cluster 3	54600	-11.063	-72.63	16:20 – 16:36	None

The results table above shows cluster information using the test set. We again look at the peak timestamp of user activity and possible access point that the user established connections. The tables also lists the number of items per cluster obtained using the k-means algorithm on the test data. Advanced chart method from the Rapid Miner data mining tool was used to develop an illustrated using the clustered example set, where we mapped most common timestamp values against the different clusters, Cluster 0, Cluster 1, Cluster 2, and Cluster 3.

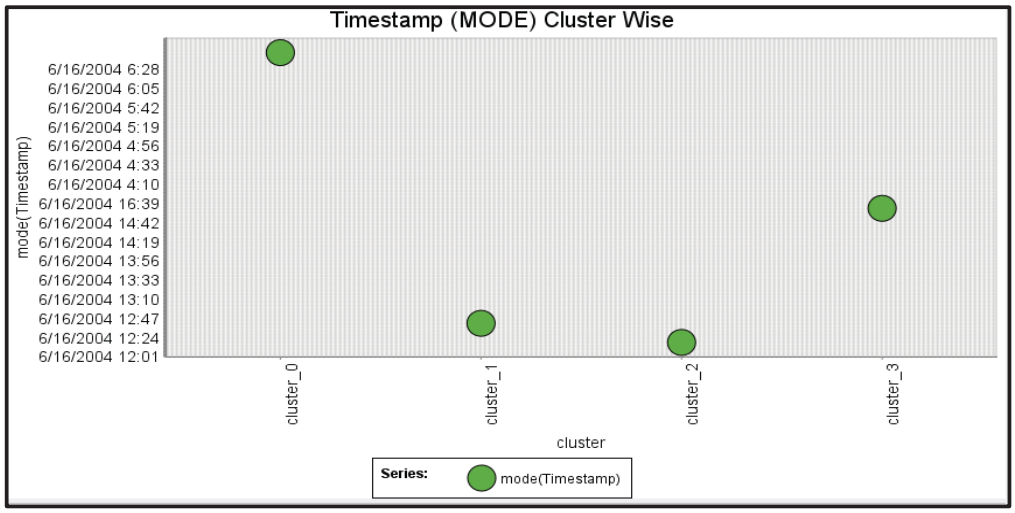

Figure 5 -3 Results for Peak Time per Cluster for Test Data (K-Means)

The graph above shows the results obtained for determining the peak time for each cluster. Mode timestamp values for each cluster has been plotted, however note that this does not state that the clusters were subject to only these timestamp values.

Table 5 -F Results Table- Test Set Cluster Characteristics (K-Medoids)

Average -Inter Cluster Centroid Distance			-12.469		
Davies Bouldin Index			-0.477		
Cluster	**Number of Objects**	**AVG. Distance of objects from Cluster Centers**	**Cluster Centers**	**Timestamp**	**Most Common Location**
Cluster 0	19898	-12.301	-55.625	05:59 – 06:23	TeaHouse_Public, UniversityofWashington, Home Office Jills Airport
Cluster 1	40519	-10.535	-88.913	05:47 – 06:23	Seattle Mac Store, University of Washington,BLUMECO
Cluster 2	20029	-8.917	-63.406	06:23– 06:59	21558, @Nelle, Apple Network, University of Washington, Green lake,
Cluster 3	47842	-14.195	-77.689	14:29-16:59	1100, IR, University Of WashingtonCSE
Cluster 4	9490	-11.784	-93.441	None	Apple Network, Trabant Chai Lounge, University of Washington
Cluster 5	26400	-9.751	-49.684	None	21558, Home Office, University of Washinton, Linksys

The table above shows the results obtained when the K-Medoids clustering model was applied to the test data using k = 6 as the number of clusters. We developed an illustration for cluster wise peak connectivity period using the timestamp attribute.

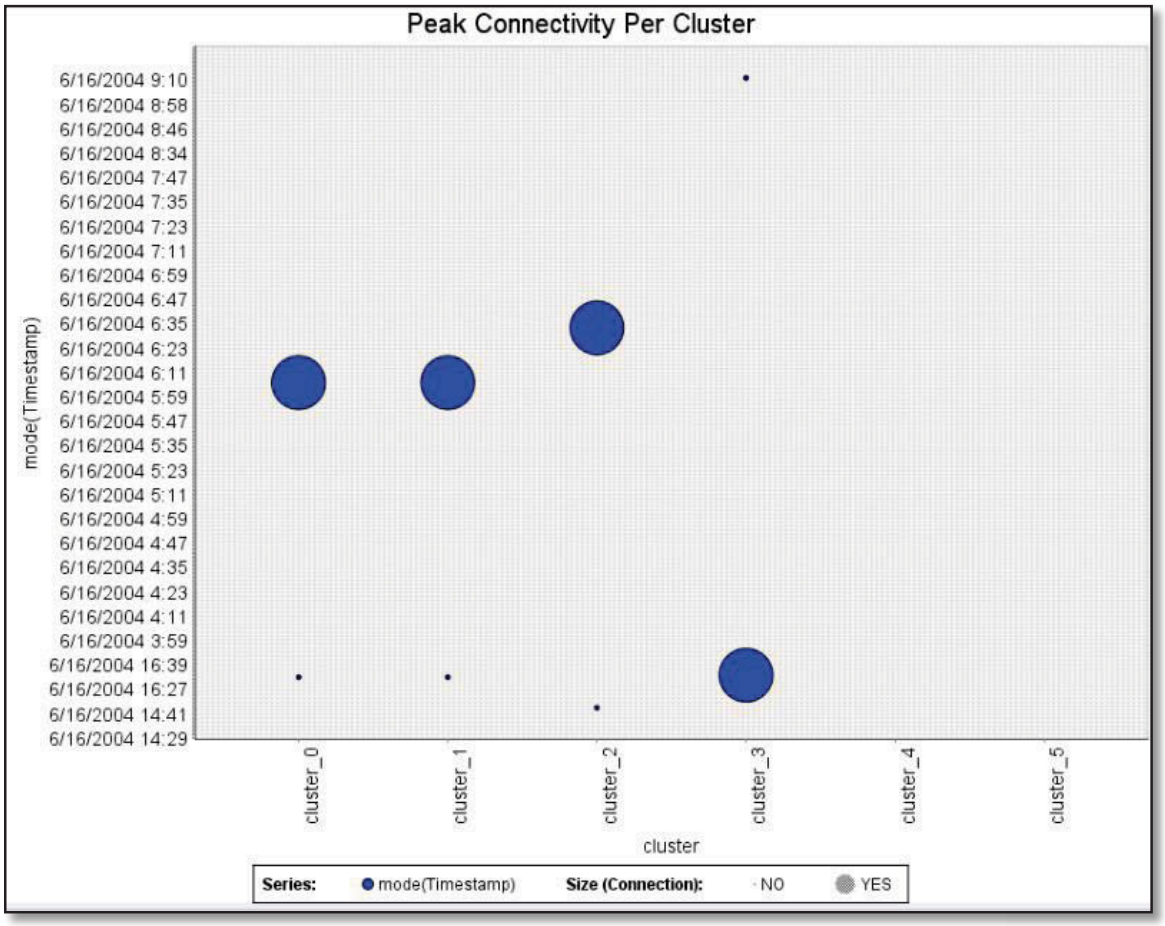

Figure 5-4 Results for Peak Time per Cluster for Test data (K-Medoids)

The graph obtained shows there was not peak connectivity period observed for cluster 4 and cluster 5 for the test data, cluster behavior which was consistent with observations made for cluster 5 of the training data.

5.4 Summary

In Chapter 5 we presented results obtained from our experiments using the training data set and the test data set using both the k-means and the K-Medoids algorithm respectively. We observe a level of consistency in the results that were obtained, that provide validity to our experiment model. We next analyze our results for the purpose of location awareness, where we attempt to derive knowledge from the movement pattern of our user within the 12 hour trace period.

Chapter 6 Analysis

In this chapter we analyze results obtained from our experiments, we perform centroid mapping to compare our clusters obtained for the training data using k-means and K-Medoids and clusters obtained using test data using k-means and K-Medoids. In the initial stages of the experiments we used Rapid Miner to derive the Davies Bouldin Index for each model for cluster validity, based on which the number of clusters for k-means and K-Medoids was obtained.

6.1 Centroid Mapping

We gain an impression on the quality of clusters produced by the k-means and K-Medoids algorithm. The quality of clusters was determined by observing intra-cluster similarity and inter-cluster similarity. Intra-cluster similarity looked at the resemblance between objects within the same cluster whereas inter-cluster similarity looked at the comparison between the different clusters.

6.1.1 K-Means

The k-means algorithm produced four clusters. Training set and test set results were plotted to identify the intra cluster similarity. In this instance the average distance of an object within a cluster from the cluster centroid was used.

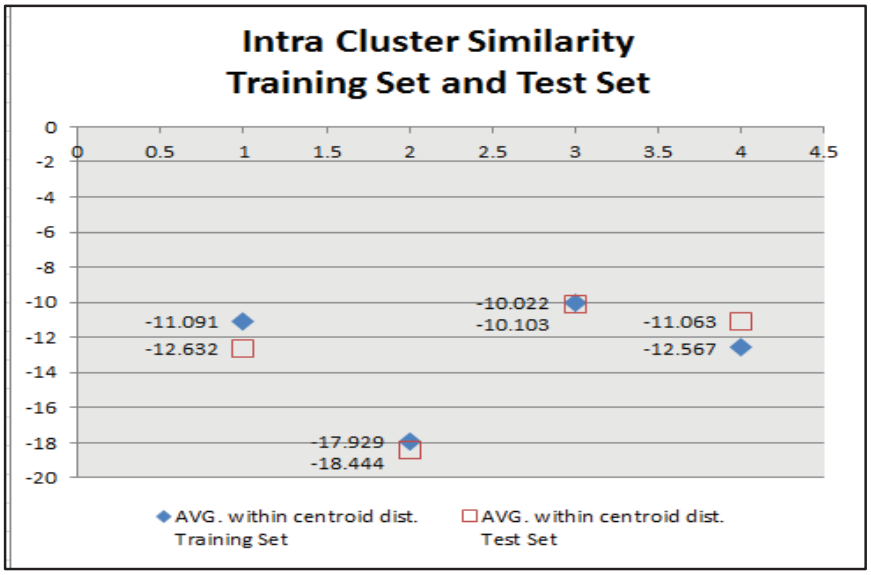

Figure 6-1 Mapping Average Distance from Cluster Centers (K-Means)

It was observed that there was some similarity between the cluster objects produced by the training data set and the test data set. Specifically for Cluster 1 and Cluster 2 there was an overlap on the plots, implying that objects within these two clusters from the training data set shared similar characteristics with those from the test data set. We note that the distance of objects from the centroid in Cluster 0 of the training data set was notably similar to distance of objects from the centroid in Cluster 4 of the test data set, we also observe the same for Cluster 4 of the training data set and Cluster 0 of the test data set.

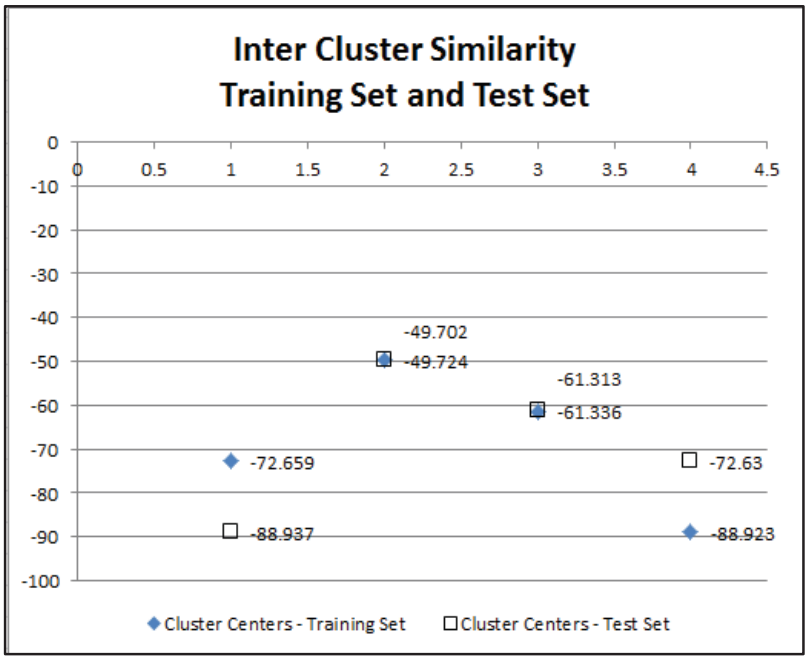

Figure 6-2 Mapping K-Means Cluster Centers

The centroid values for the four clusters were plotted for both the training set and the test set to determine inter cluster similarity and to validate our clusters produced for the training set and the test set. From observations made, it was noted that Cluster 1 and Cluster 2 for the training set had significantly similar cluster centroids in comparison to Cluster 1 and Cluster 2 of the test set. On the same note there were substantial likeness between Cluster 0 of the training set and Cluster 4 of the test set and vice versa. Hence we were able to ascertain that our training set was a true exemplification of our test set for the k-means algorithm.

6.1.2K-Medoids

The same technique was used to determine the similarity and likeliness of clusters produced using the K-Medoids algorithm for the training and the test set, we again use cluster centroids and average distance of cluster objects from cluster centroids as the measure.

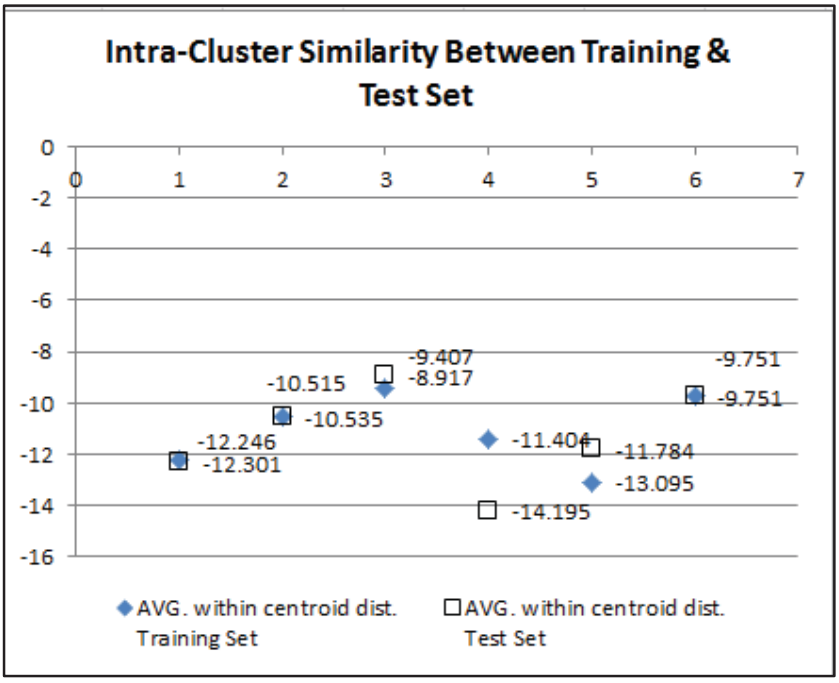

Figure 6-3 Mapping Average distance from Cluster Centers (K-Medoids)

There was greater similarity observed between objects within a cluster produced by the training data set and the test data set using K-Medoids algorithm in comparison to k-means algorithm. A very close likeness was observed for average distance of objects within the cluster from cluster centers for Clusters 0, Cluster 1 and Cluster 5.

Differences were observed in Cluster 2, Cluster 3 and Cluster 4, where the distance of objects in the clusters seemed significantly unrelated and did not correlate with each other.

However, for Cluster 3 and Cluster 4, we do note that the distance of objects from the centroid in Cluster 3 of the training data set was notably similar to distance of objects from the centroid in Cluster 4 of the test data set, we also observe the same for Cluster 4 of the training data set and Cluster 3 of the test data set.

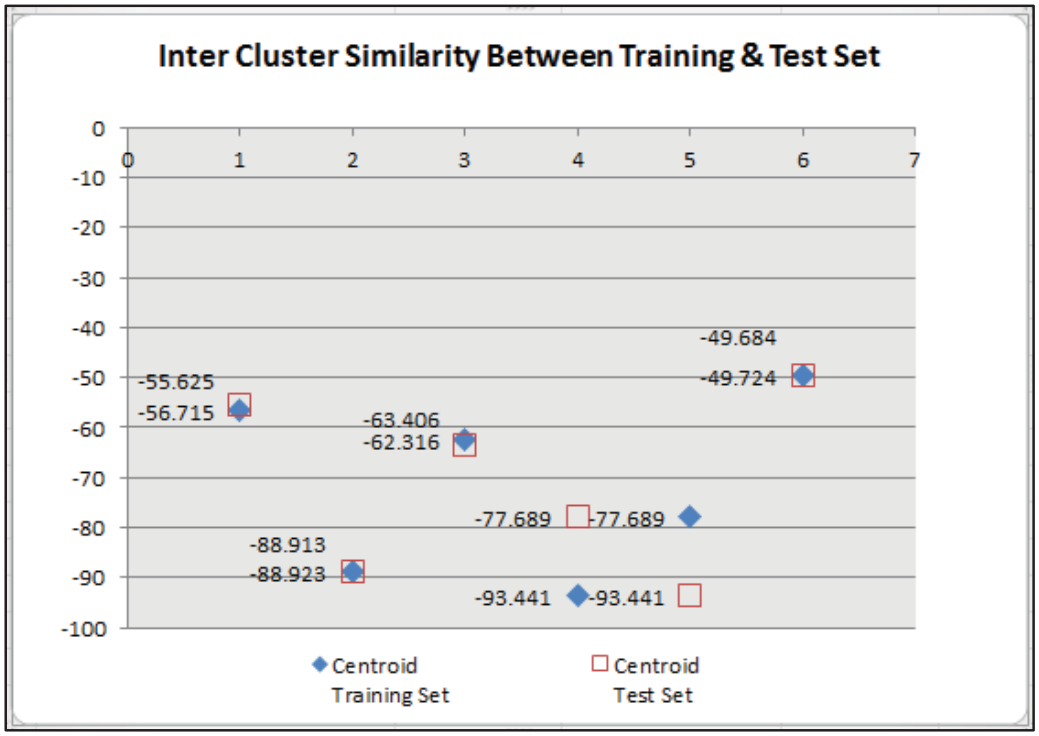

Figure 6-4 Mapping K-Means Cluster Centers

The centroid values for the six clusters were plotted for both the training set and the test set to determine inter cluster similarity of clusters produced and to validate our clusters for K-Medoids. From observations made, it was noted that cluster centroids were considerably alike for Cluster 0, Cluster 1, Cluster 2 and Cluster 5.

We observed significant dissimilarity between Cluster Centers obtain for Cluster 3 and Cluster 4 of the training and test data set. However we do note that cluster centers for Cluster 3 of the training set and Cluster 4 of the test set are exactly alike, also Cluster 4 of the training set and Cluster 3 of the test set have same centroids.

Hence for k–Medoids our results indicated that our training set was a true exemplification of our test set for the algorithm

6.2 Comparing K-Means and K-Medoids

In this section of the paper we compare the results obtained for K-Means and K-Medoids algorithm. The comparison is made using consistent variables for both algorithms such as sample size used, number of cluster and the numerical measure used.

Table 6 – A Results comparison K-Means and K-Medoids

	K-Medoids	K-Means
Sample Size	164178	164178
Optimal #. Clusters	6	4
Numerical Measure	Mixed Euclidean Distance	Mixed Euclidean Distance
AVG. within Centroid Distance	-11.247	-12.469
Davies Bouldin Index	-0.326	-0.477

Our foremost concern in this experiment was deciding upon the better algorithm. Hence comparing the Davies Bouldin Index for the two algorithms we note K-Medoids resulted in a lower Davies Bouldin Index of -0.326 compared to -0.475 for K-Means.

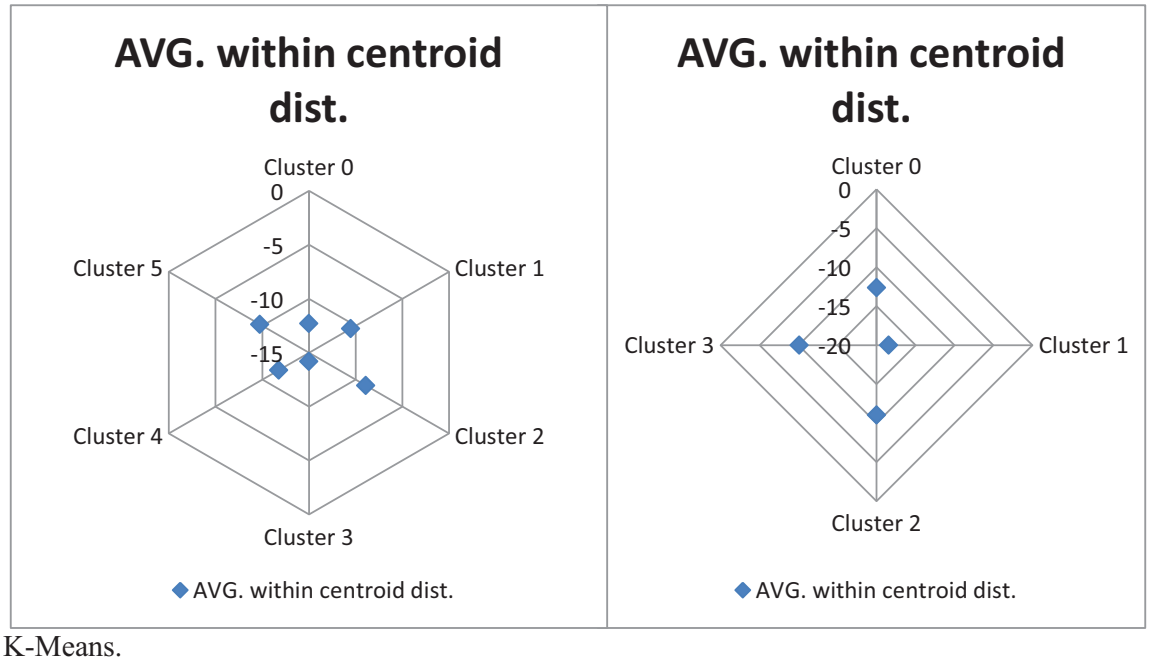

Figure 6-5 Cluster Distribution Graph for K-Means and K-Medoids

The average distance of objects from the cluster centroid for within cluster objects was notably smaller for K-Medoids, in other words the objects were closer together whilst for those in the K-Means objects were distributed further apart.

From the results obtained in our experiments we were able to observe that the different ways in which the two methods of clustering behaved under different conditions, such as for different values of k, sample size used and the numerical measure used.

6.2.1 Time Complexity K-Means and K-Medoids Method

The runtime efficiency of both algorithms was compared in this experiment for each value of k = {2, 4, 6, 8, 10, 12}. The process used to obtain the results has been presented in the conceptual models included as a supplement of this critique. The same sample size of 164178 is used for both K-Medoids and K-Means algorithm. Time values measured are in seconds, these were also recorded after the respective models were run each time using Rapid Miner 5.0 in the process of identifying the optimal value for k.

Table 6-B Run Time for K-Medoids and K-Means

Clusters (k)	K-Medoids Time (s)	K-Means Time (s)
2	372	6
4	983	20
6	1647	33
8	2069	42
10	3278	61
12	5221	88

The table above shows the runtime measured in seconds for the increasing number of *k* clusters for both K-Means and K-Medoids algorithm. Initially to measure the runtime for K-Medoids on the test data set (*n= 164178)* resulted in an unreasonable amount of time, we used the training dataset of size n = 41044 to measure time complexity. We then created a graph for the record time data using a spreadsheet application to derive the time complexity graph as shown below.

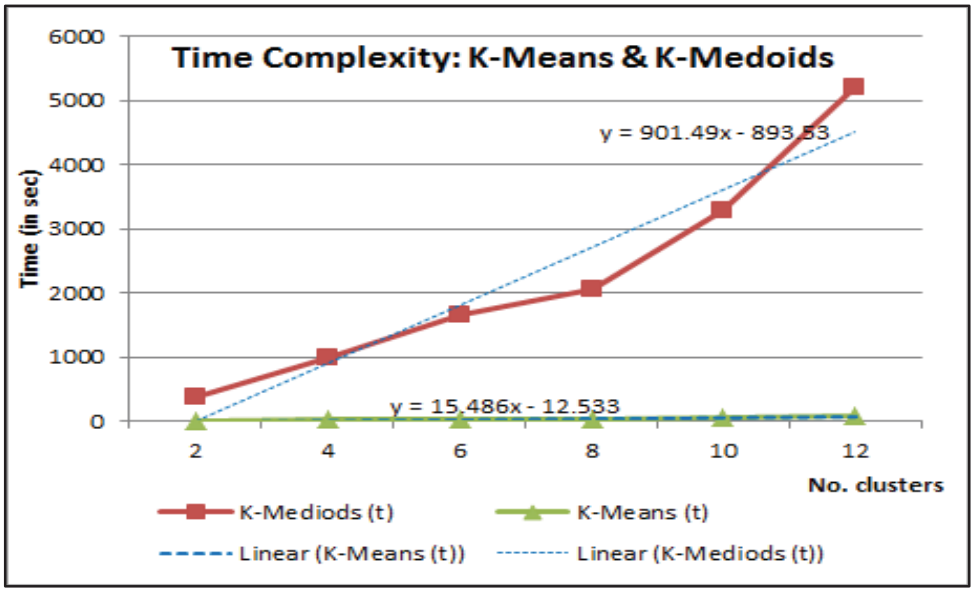

Figure 6-6 Time Complexity Graph for K-Medoids and K-Means

It was observed from the results obtained that K-Medoids algorithms took a greater amount of time to form clusters compared to K-Means algorithm. K-Medoids had an exponential time growth compared to K-Means. Hence, it can be assumed that although-Medoids highly robust and efficient, it has a greater (J. Han, 2001)'time complexity of $O\ (k\ (n-k)^2)$ whereas K-Means has a time complexity of ***O (nk).'***

6.2.2 Correlational Coefficient

The results obtained for time complexity were used to determine the linear equations for the two sets of time values for k. The two linear equations we used to obtain the correlational coefficients were:

i. f(x)= 901.49x - 893.53

ii. f(x)= 15.48x- 12.53

The correlational coefficient was used to determine the strength of association between x and f(x), where x represents the number of clusters and f(x) represents time in seconds.

Table 6-C Comparison between K-Medoids & K-Means

	K-Medoids (PAM)	K-Means
Equation	f(x) = 901.49x - 893.53	f(x) = 15.48x - 12.53
Correlational Coefficient	0.999	0.948

A strong linear relationship was indicated in the case of both K-Medoids and K-Means as both algorithms have a correlational coefficient closer to 1, however in the case of K-Medoids the linear association between the number of clusters and time taken to form the clusters was stronger. Also a positive correlation indicates that as the number of clusters increases, time taken to form the clusters also increases.

6.3 Summary

This chapter primarily focuses on determining the quality of clusters and discussing performance of the K-Means and K-Medoids algorithm. We perform centroid mapping for K-Means and K-Medoids algorithm, followed by comparison of results obtained from the experiments using measures like the average within centroid distance, correlational coefficient and time complexity to present our results logically and to support our work.

Chapter 7 Knowledge Presentation

In this section of the paper we present knowledge derived from exploring our clusters. As stated previously our data set is a 12 hour trace of a single user movement taken on the 16th June 2004, between 03:32am and 16:46pm. From the dataset we attempt to build cluster models using K-Means and K-Medoids and based on characteristics of these clusters we identify places common to each cluster or the time of day the user had visited a particular place. In this part of the paper we will present knowledge extracted using K-Medoids clusters as both k-means and K-Medoids produced similar results, cluster quality of K-Medoids clusters was better in comparison to k-means.

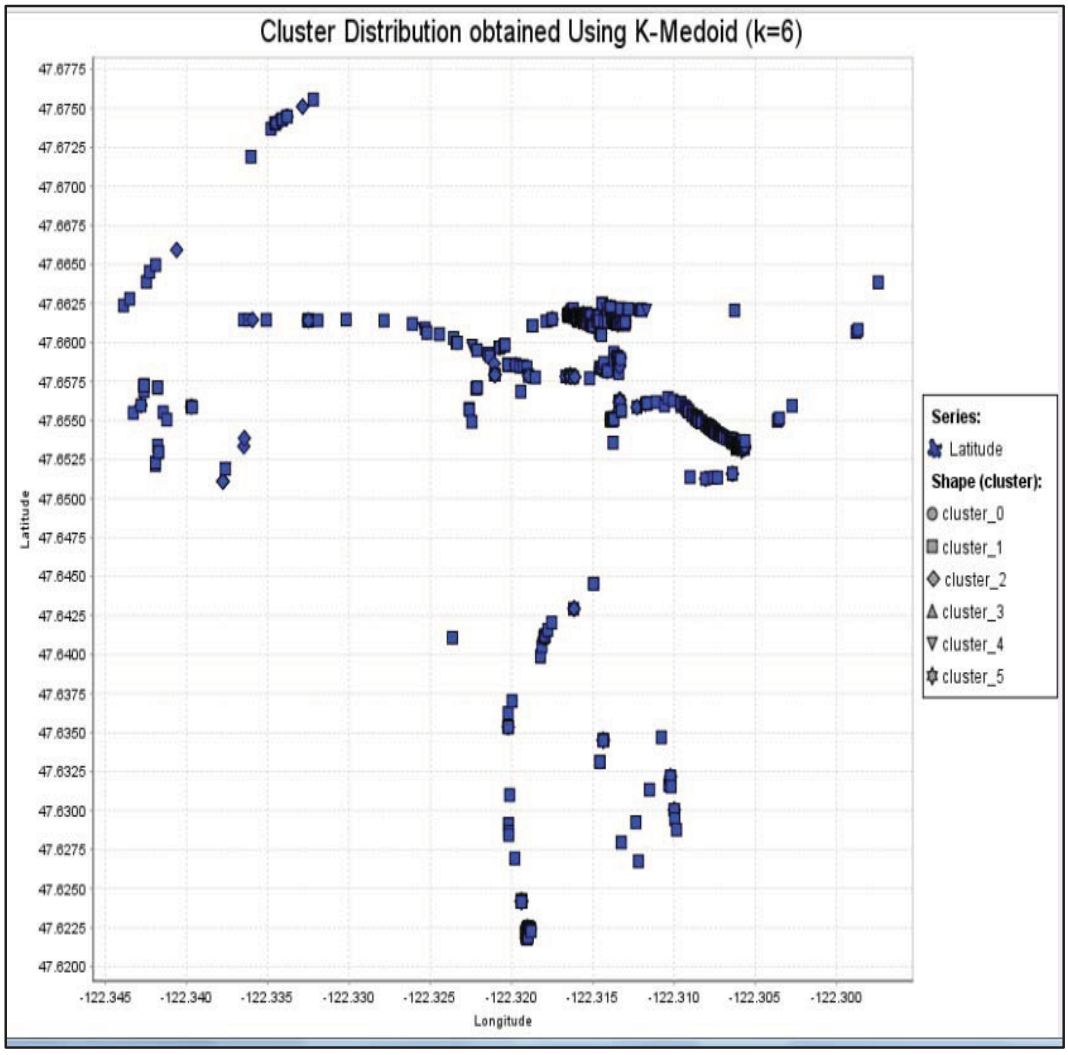

Figure 7-1 K-Medoids Cluster Distribution

The illustration above shows the distribution of clusters along dissimilarity matrix of latitude and longitude, extracted using the K-Medoids algorithm. The clusters are represented by different shapes showing user movement and because we are using coordinates, we can map these coordinates to specific places visited by the user on 16th June 2004. We derive a list of seven highly dense coordinates that indicate user trace:

1. Place 1 Coordinate (47.6550, -122,345) - 3929 Woodland Park Avenue, Seattle
2. Place 2 Coordinate (47.6600, -122.320)- 4311 8th Avenue NE
3. Place 3 Coordinate (47.6625, -122.315)-4600 – 4698 12th Avenue NE
4. Place 4 Coordinate (47.6575, -122.315) - 4116 12th Avenue NE
5. Place 5 Coordinate (47.6550,-122.310) - 1535 NE Grant Ln. University of Washington, Seattle
6. Place 6 Coordinate (47.6525, -122.305) - More Hall (MOR), University of Washington, Seattle
7. Place 7 Coordinate (47.6225, -122.320) - 415 10th Avenue East

Furthermore, we use Google Maps® to find out the places visited by the user according to the coordinates. In our investigation we discover that:

1. Place 1 as 3929 Woodland Park Avenue, Seattle
2. Place 2 as 4311 8th Avenue NE
3. Place 3 as 4600 – 4698 12th Avenue NE
4. Place 4 as 4116 12th Avenue NE
5. Place 5 as 1535 NE Grant Ln. University of Washington, Seattle
6. Place 6 as More Hall (MOR), University of Washington, Seattle
7. Place 7 as 415 10th Avenue East

Hence, we can conclude the user had been to the above locations at some part of the day during the 12 hour trace period.

7.1 Popular Locations

To identify the location where the user spent most of his time, out of the 12 hour trace period, we use a simple analogy to generalize i.e. we create a pivot table with timestamp count sorted from the largest to the smallest value for access points names with established connection. As shown in the table below.

Table 7-A Access Point Connectivity Duration

Location	Access Point Names	Timestamp Count (milli-seconds)	Time (mins)
Place 1	1100	28383	7.88
Place 2	IR	28027	7.78
Place 3	UniversityOfWashingtonCSE	15751	4.37
Place 4	University of Washington	5416	1.50
Place 5	linksys	1845	0.512
Place 6	NETGEAR	1403	0.389
Place 7	EE	1053	0.292
Place 8	geronimo	911	0.253
Place 9	tesseract01	818	0.227
Place 10	Default	743	0.206

(**Please note as we do not know the locations denoted by the Access Point names, we have assigned our own location to the table as place 1, place 2 and so on…)

The table above shows 10 most popular access points from the dataset the user connected with. At Place 1, access point name 1100 the user spent 28383 milli-seconds which approximates to almost 8 minutes of activity, similarly for Place 2 there is an activity period of 7.78 minutes almost 8 minutes. However from the data table above, we can only deduce that the user spent maximum amount of time at Place 1 also shown in the illustration below.

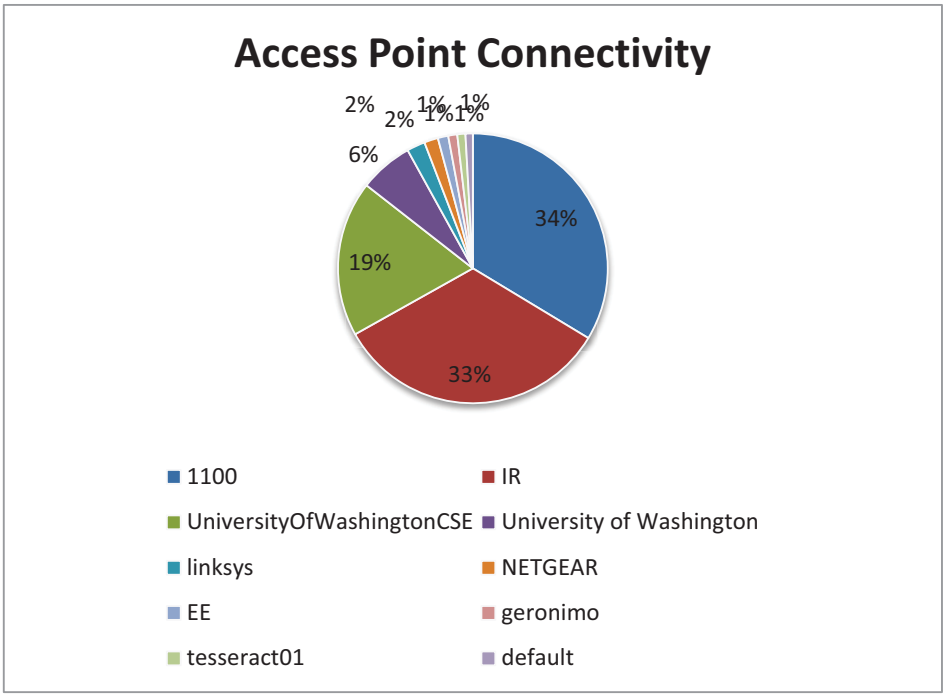

Figure 7-2 Timestamp distribution along access point connectivity

The user seems to have spent a considerable amount of time at Place 1, Place 2 and Place 3. By further scrutinizing the data we can assume that the user is either a student or a professor at University of Washington. In other words there exists some relationship between the user and University of Washington.

We further obtain a view of the clusters by grouping clusters based on most common access points the user connected to. In the bubble graph below the size of the bubble indicates connectivity and the bubble itself represents the access point name.

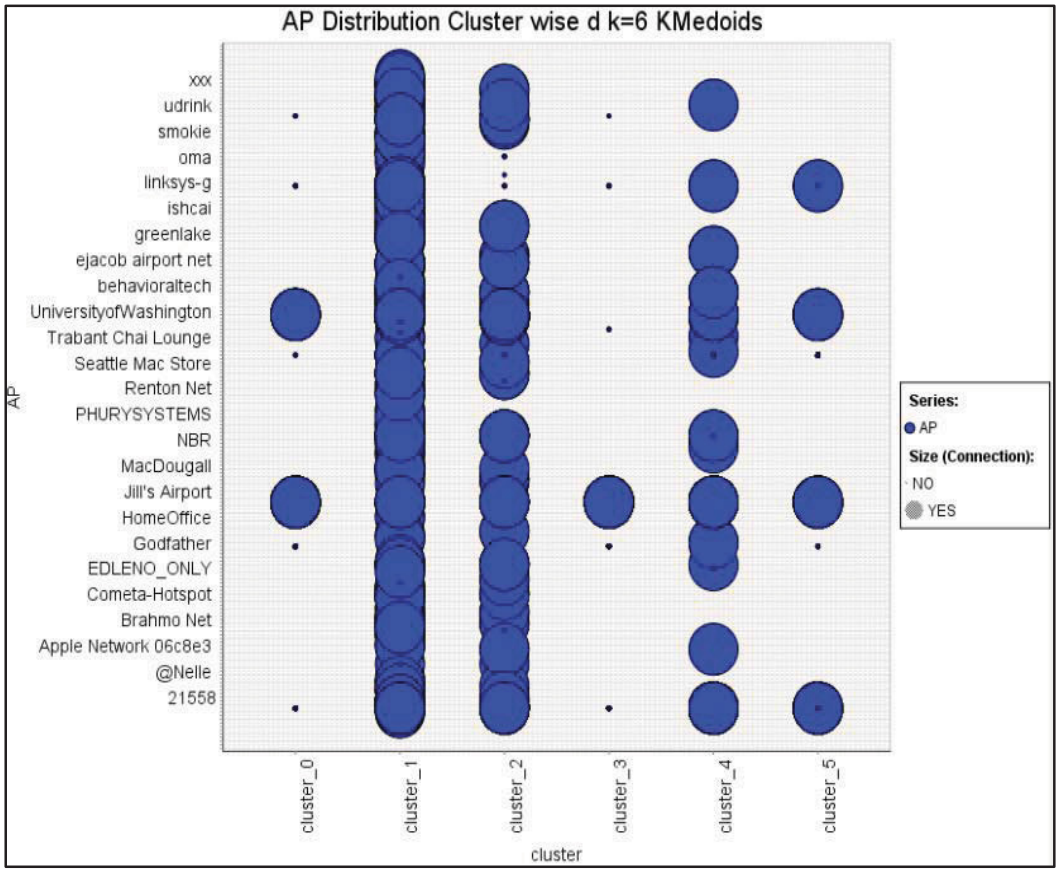

Figure 7-3 Cluster based AP Connectivity

From the above illustration we can indicate the user spent some time at access points where the user established a connection. We can further assume that since the user spent some time at these locations, we can classify them as important locations or places of interest. Observations made show the following cluster based grouping of access points:

1. **Cluster 0** – TeaHouse_Public, University of Washington, Home Office and Jills Airport
2. **Cluster 1** – Seattle Mac Store, 21558, Brahmo Net, Home Office, University of Washington and Green Lake
3. **Cluster 2** – 21558, @Nelle, Apple Network, NBR, University of Washington, Green Lake and Smokie
4. **Cluster 3** – Home Office
5. **Cluster 4** - 21558, Apple Network, Home Office, NBR, Trabant Chai Lounge, University of Washinton, Linksys andudrink

6. **Cluster 5** - 21558, Home Office, University of Washington, and Linksys

Most of the Cluster 0, Cluster 1, Cluster 2, Cluster 4 and Cluster 5 indicate that the most common location of the user was around University of Washington. We can also assume that other access point names used in our datasets belong to locations around the University of Washington area, either on campus locations or off campus locations.

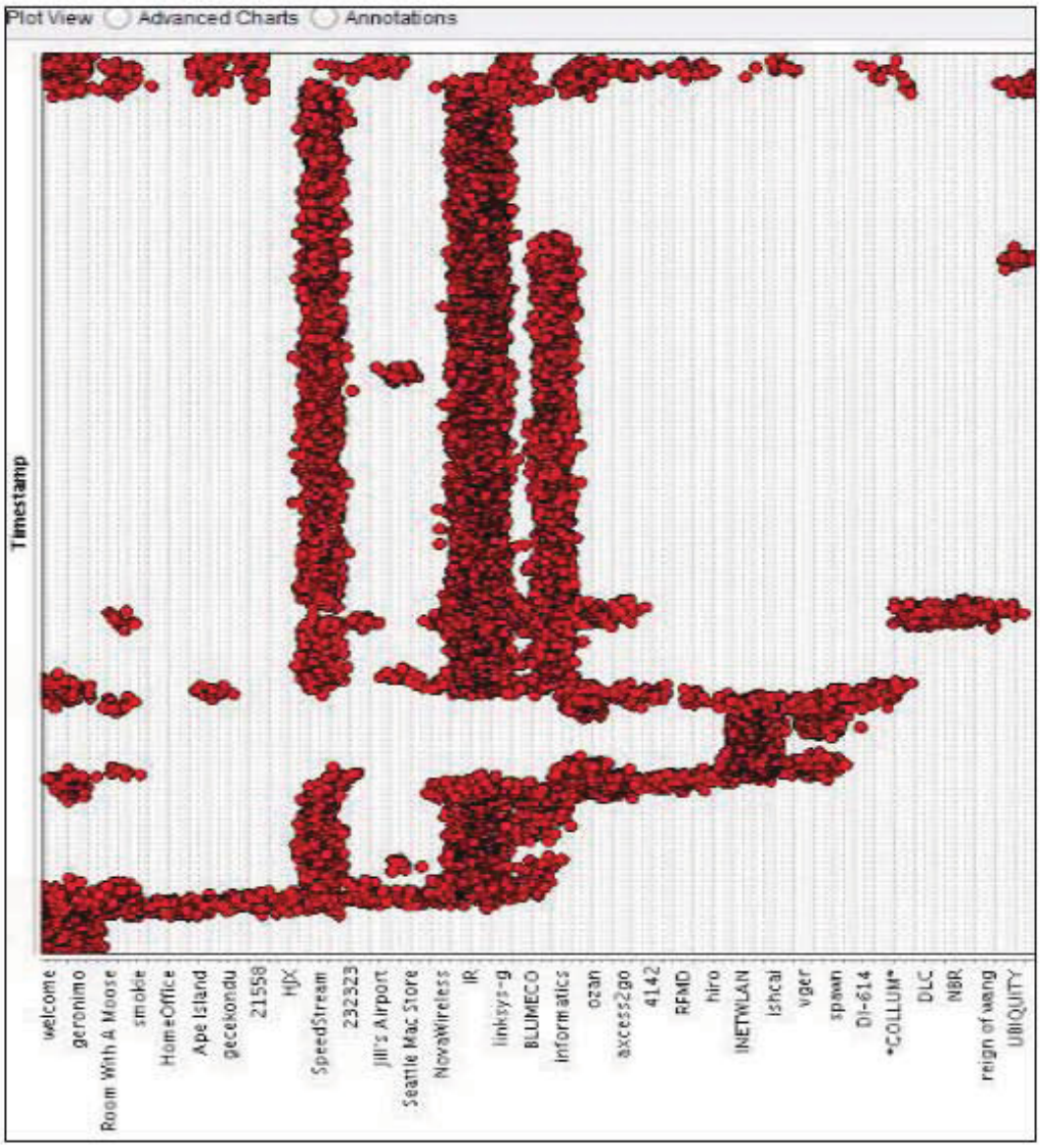

Figure 7-4 Access Point Distribution against Timestamp

Another approach we used to model the clustered data was to treat it as a single entity irrespective of cluster distribution. We used the clustered example set produced using the K-Medoids algorithm to build a plot view of the timestamp and access point attribute. Our major objective here was to identify the access point with the longest period of connectivity. The results obtained showed that the user was connected to the following access points for a significant amount of time:

1. Speedstream
2. IR
3. Linksys-g
4. BLUMECO
5. Informatics

The results also indicate that the user was either able to connect to or had access to the above five access points instantaneously, as these five access points share the same timestamp. We also note that for timestamp values where the user did not connect to any one of the access points listed above, connectivity was also missing for the remaining four access points. For instance if the user did not connect to Speedstream, connection was also not established with IR, Linksys-g, BLUMECO, Informatics. Therefore, we can conclude that the access points Speedstream, IR, Linksys-g, BLUMECO and informatics are either within the same building or very close to each other. We can also propose with reference to our cluster based information that the access points are in a building on the university campus.

7.2 Peak Timestamp

The results were also used to investigate the data in terms of the peak periods of mobile user activity. In this case we plot peak periods of activity per cluster. The K-Medoids example set was used to obtain a cluster based overview with respect to timestamp as shown in the graph below.

The graph below shows both the peak time of activity and inactivity per cluster.For Cluster 0 we note the peak time of activity as 6:29am and time of inactivity as 12:47pm. Cluster 1 shows peak activity as 6:49am and time of inactivity as 16:31pm

and so on. Peak activity time suggests the user was most actively connected to either one or more access points during this period.

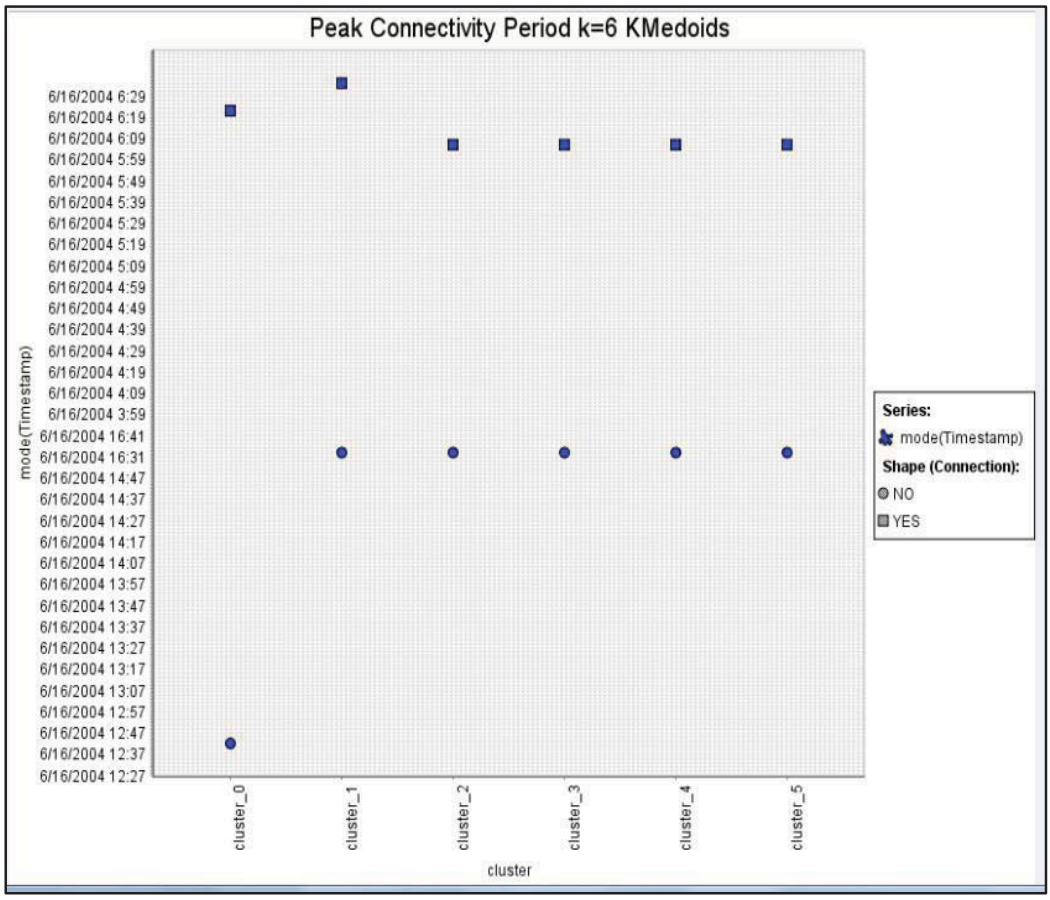

Figure 7-5 Popular Timestamps Per Cluster (K-Medoids Example Set)

7.3 Summary

In this chapter we presented knowledge derived by performing cluster analysis for location awareness using the K-Medoids algorithm. We identify popular locations visited by the user and extract cluster based information regarding peak time of user activity. We present graphical illustrations and cluster attributes to support our discussions and prepositions. In the following section we make conclusions and recommendations on using cluster based algorithms for location awareness.

Chapter 8 Conclusion

In our experiment we used the partition based clustering methods, namely K-Means and K-Medoids algorithms to semantically analyze our data. Semantic analysis involved describing our data sets in a logical way to derive an understanding of how objects behave in our datasets or correlated with each other. We used data mining reference model defined by (J. Han, 2001) to design our experiments. Furthermore, we used the Davies Bouldin index derived for each value for k to determine the appropriate number of clusters for our data set. Once the value for k for the two algorithms was identified using the training set we applied it to our test data set.

In determining the efficiency of K-Means and K-Medoids, it was noted that the K-Medoids algorithm was more efficient with large data sets in comparison to K-Means; however it had greater time complexity, where we experienced a longer time period for our K-Medoids model runs. We also noted a linear correlational relationship between number of clusters and time threshold. Where, as the number of clusters for the algorithm increases the time taken for the algorithm to run increases as well.

Using our clustering method we were able to obtain the following information:

1. The user had visited the following places at some part of the day.
    a. Place 1 as 3929 Woodland Park Avenue, Seattle
    b. Place 2 as 4311 8th Avenue NE
    c. Place 3 as 4600 – 4698 12th Avenue NE
    d. Place 4 as 4116 12th Avenue NE
    e. Place 5 as 1535 NE Grant Ln. University of Washington, Seattle
    f. Place 6 as More Hall (MOR), University of Washington, Seattle
    g. Place 7 as 415 10th Avenue East
2. The user had some relationship with the University of Washington i.e. He was either a student or a professor at the university.
3. The user was most commonly connected to access points 1100, IR, BLUMECO and Speedstream.

4. Access Points 1100, IR, BLUMECO and Speedstream were very close to each other as they shared common timestamp values and exhibit a direct dependency relationship, possibly located within the same building.
5. Most common time the user was active was around 5:59am to 6:29am.

Hence we derive knowledge that:

1. The user is a student or a professor at the University of Washington, where he/ she spent a significant part of the 12 hour trace period in a building on the University Campus.
2. The place or building visited by the user is within range of the following access points, 1100, IR, BLUMECO, and Speedstream.
3. The peak time of user activity is between 5:59am to 6:29am

The methods that we have used in this experiment can be further improved where we can look at mapping the latitude and longitude values data set with specific locations as derived using Google Maps®, hence reducing our dependency on access point names for clustering and analysis. Classifying our data based on specific locations can help us determine the actual amount of time spent at each location by the user.

Hence, I would like to recommend further research work on this topic to better analyze the data and acquire meaning from the dataset using other data mining techniques such as Classification, Prediction analysis, Association rule mining or through use of decision trees. The dataset holds high potential in terms of analysis using these techniques, where we can obtain more specific knowledge regarding user behavior.

The quintessence of this research paper was to derive meaning out of mobile user traces; hence we can finally state that it is possible to perform semantic analysis on mobile data for location awareness as shown by our research.

Bibliography

1. A. Berson, S. S. (1999). Building Data Mining Applications for CRM . In S. S. A. Berson, & S. Yates (Ed.), *An Overview of Data Mining Techniques.* New York, USA: McGraw-Hill Companies.

2. A. Ciaramella, M. G. (2010, January). Situation Aware Mobile Service Recommendation with fuzzy Logic and Semantic Web. *International Journal of Uncertainty, Fuzziness & Knowledge based Systems, 18*, 411-430.

3. Attribute, C. C. (2014, October 27). *Davies–Bouldin index*. (Wikimedia Foundation Inc.) Retrieved December 01, 2014, from Wikipedia - The free encyclopedia: http://en.wikipedia.org/wiki/Davies%E2%80%93Bouldin_index

4. Attribution, C. C. (2014, November 14). *Deviance Information Criterion*. (Wikimedia Foundation Inc.) Retrieved December 01, 2014, from Wikipedia - The Free Encyclopedia: http://en.wikipedia.org/wiki/Deviance_information_criterion

5. Attribution, C. C. (2014, May 9). *K-mediods*. (Wikimedia Foundation Inc.) Retrieved May 15, 2014, from Wikipedia - The Free Encyclopedia: en.wikipedia.org/wiki/K-mediods

6. Attributions, C. C. (2013, November 27). *K-Means Clustering*. (Wikimedia Foundations Inc.) Retrieved May 16, 2014, from Wikipedia - The Free Encyclopedia: http://en.wikipedia.org/wiki/K-means_clustering

7. Attributions, C. C. (2014, December 8). *Data Cleansing*. (Wikimedia Foundations Inc.) Retrieved December 11, 2014, from Wikipedia - The Free Encyclopedia: http://en.wikipedia.org/wiki/Data_cleansing

8. Attributions, C. C. (2014, December 3). *Database Normalization*. (Wikimedia Foundation Inc.) Retrieved December 11, 2014, from Wikipedia - The Free Encyclopedia: http://en.wikipedia.org/wiki/Database_normalization

9. Attributions, C. C. (2014, May 15). *Determine the Number of Clusters*. (Wikimedia Foundation Inc.) Retrieved May 23, 2014, from Wikipedia - The Free Encyclopedia: http://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set

10. Budiarto, S. N. (2002). Data Management Issues in Mobile and Peer-to-Peer Environments. *Data & Knowledge Engineering, 41*, 183-204.

11. Corno, M. (2012, February). *K-Means Clustering*. (Dipartimento di Elettronica e Infomazione, Politecnico di Milano) Retrieved November 02, 2014, from A

Tutorial on Clustering Algorithms: http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

12. Creative Commons Attributions. (2014, November 27). *Citi Bike*. (W. F. Inc, Producer) Retrieved December 02, 2014, from Wikipedia - The Free Encyclopedia: http://en.wikipedia.org/wiki/Citi_Bike

13. Creative Commons Attributions. (2014, July 23). *Context Awareness*. (C. C. Attributions, Editor, & W. Foundation, Producer) Retrieved December 02, 2014, from Wikipedia - The Free Encyclopedia: http://en.wikipedia.org/wiki/Context_awareness

14. E. Turban, L. V. (2011). *Information Technology for Management - Improving Strategic and Operational Performance* (8th ed.). New Jersey, United States: John Wiley & Sons.

15. E.R Cavalcanti, M. S. (2013). On the Interactions between Mobility Models and Metrics in Mobile ad hoc Networks. *International Journal of Networks and Communications, 3*(3), 63-80.

16. Howcast Media Inc. (2014, April). *How to Normalize Data*. (J. &. Liebman, Producer, & Howcast Media, Inc.) Retrieved December 11, 2014, from Howcast: http://www.howcast.com/videos/359111-How-to-Normalize-Data

17. J. H. Kang, W. W. (2006, may). *uw-places-placelab-city-wide-2006-05-02.* Retrieved March 11, 2014, from CRAWDAD: http://crawdad.cs.dartmouth.edu/uw/places/placelab/city-wide

18. J. Han, M. K. (2001). *Data Mining - Concepts and Techniques* (1st. ed.). San Diego, United Stated: Morgan Kaufmann Publishers.

19. J. K. Laurila, D. G.-P. (Jun. 2012). The Mobile Data Challenge: Big Data for Mobile Computing Research. *in Proc. Mobile Data Challenge Workshop (MDC) in conjunction with Int. Conf. on Pervasive Computing.* Newcastle.

20. J.A. Burke, D. E. (2006). Participatory Sensing. *Workshop on World-Sensor-Web (WSW'06): Mobile Device Centric Sensor Networks and Applications*, (p. n/a). Colorado, USA.

21. L. Calderoni, D. M. (2012, December). Location-aware Mobile Services for a Smart City: Design, Implementation and Deployment. *Journal of Theoretical and Applied Electronic Commerce Research, 7*(3), 74-87.

22. L. Lao, D. F. (2007, January). Extracting Places and Activities from GPS Traces Using Hierarchical Conditional Random Fields. *International Journal of Robotics Research, 26*(1), 119-134.

23. Lawton, G. (2010, August). *Is Context-Aware Computing Ready for the Limelight?* (C. Metra, Ed.) Retrieved March 15, 2014, from Computing Now: http://www.computer.org/portal/web/computingnow/archive/news068

24. M. Malcher, J. A. (2010). A Middleware Supporting Adaptive and Location-aware Mobile Collaboration. *Mobile Context Workshop: Capabilities, Challenges and Applications, Adjunct Proceedings of UbiComp 2010.* Copenhagen.

25. META Group. (2001, February). *Big Data: Why it matters?* Retrieved March 17, 2015, from www.sas.com: http://www.sas.com/en_us/insights/big-data/what-is-big-data.html

26. Mirkes, E. M. (2011, September). *K-Means and K-Medoids Applet*. (E. M. Mirkes, Editor) Retrieved June 16, 2014, from University of Leicester: http://en.wikipedia.org/wiki/K-means_clustering

27. N. Bicocchi, G. C. (2008, February). Supporting Location Aware Services for Mobile Users with the Whereabouts Diary. *1st International Conference on Mobile Wireless MiddleWARE, Operating Systems and Applications, MobileWARE.*

28. Rijmenam, M. V. (2013, April). *Five Data Mining Techniques That Help Create Business Value*. (BIG DATA) Retrieved December 2, 2014, from Datafloq: Connecting Data and People: https://datafloq.com/read/data-mining-techniques-create-business-value/121

29. S. Isaacman, R. B. (2011, June). Identifying Important Places in People's Lives from Cellular Network Data. *Pervasive and Mobile Computing, 6696*, 133-151.

30. S. Yazji, P. S. (2013, January 18). Efficient Location Aware Intrusion Detection to Protect Mobile Devices. *Pervasive and Ubiquitous Computing*(Open Access).

31. Satoh, I. (2006). Location-based services in ubiquitous computing environment. *International Journal on Digital Libraries*.

32. Tibshirani, R. (2013, January 24). Clustering 1: K-means, K-medoids. In R. Tibshirani, *Data Mining* (pp. 36-463/ 36-662).

33. X. Jin, J. H. (2010). K-Medoids Clustering. (G. I. C. Sammut, Ed.) *Encyclopedia of Machine Learning*, 564-565.