A
Dissertation On (Major Project-II)

"Health Prediction based on Activity Patterns by using Machine Learning Techniques

Submitted in Partial Fulfillment of the Requirement
For the Award of Degree of

Master of Technology

In

Software Technology

By

Ashish Jain
University Roll No. 2K15/SWT/506

Under the Esteemed Guidance of

Dr. Ruchika Malhotra
Associate Head & Associate Professor, Discipline of Software Engineering

2015-2018

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
DELHI – 110042, INDIA

STUDENT UNDERTAKING

Delhi Technological University
(Government of Delhi NCR)
Bawana Road, Delhi- 110042

This is to certify that the thesis entitled **"Health Prediction based on Activity Patterns by using Machine Learning Techniques"** done for the Major project-II for the achievement of **Master of Technology** Degree in **Software Technology** in the **Department of Computer Science & Engineering**, Delhi Technological University, Delhi is an authentic work carried out by me under the guidance of Dr. Ruchika Malhotra.

Signature:
Student Name
Ashish Jain
2K15/SWT/506

Above Statement given by Student is Correct.

Project Guide:
Dr. Ruchika Malhotra
Associate Head & Associate Professor,
Discipline of Software Engineering, Department of Computer Science & Engineering, DTU

ACKNOWLEDGEMENT

I would like to express sincere thanks and respect towards my guide **Dr. Ruchika Malhotra, Associate head & Associate Professor, Discipline of Software Engineering, Department of Computer Science & Engineering, Delhi Technological University Delhi.**

I consider myself very fortunate to get the opportunity for work with her and for the guidance I have received from her, while working on this project. Without her support and timely guidance, the completion of the project would have seemed a far. Special thanks for not only providing me necessary project information but also teaching the proper style and techniques of documentation and presentation.

ASHISH JAIN
M.Tech (Software Technology)
2K15/SWT/506

ABSTRACT

Nowadays Healthcare industry has become a big business. Science is advance and is able to provide treatment of any disease using advance machines used in pathology center. But the treatment would be more helpful if the disease of patient will be known in advance stage.

There are lots of online system available, which can help patients to know there disease based on taking few input from patients. This system supports an end user and online consultation. These input are related to the information which they feel while having problem. The system will treat this input/information as a symptom of the disease. The system will match this symptom with the previously available data on the database, which they collected from the doctors/patient from the previous records. Finally the system will predict the disease of the patient.

But there should be one such system available which will take care health of the person. Which will help them to make them healthier by giving them feedback base on the daily activity they perform. If such system will be available then the chances of having disease will be reduces. Since peoples are very much aware of his/her health. To become healthier they do regular exercise and consult with doctor for yearly medical checkup. But they don't know when their regular activity will spoil their health.

In this thesis, we have presented a model to predict health based on person's lifestyle, which is formed by using daily activities performed by the person. To implement this health prediction system, we have recognize human activities by taking data from sensor-rich smartphones. After this we have used these human activities to create user life document, which represents user's life style. By measuring lifestyles of users is we can predict the health and suggest them to change in the lifestyle.

First, we collected sensor data of user using mobile application and then we performed activity recognition. After finding activities of user we have created a life document of user, from which the algorithm Latent Dirichlet Allocation (LDA) [1] is used to select the life style. We further propose a similar metric algorithm in order to compute the health of users to measure the life style data available in training data, and measure the impact the life style of users by graph.

When user want to know their health, then our algorithm will return list sorted by activity score, from which user can know about his health and take prevention to not having any

disease. We have implemented this system for Android based smartphone and its performance has been evaluated on with large-scale experimental data.

Finally, the prediction results manifest that the prediction properly reflecting the preferences of activity perform by the user in day, week or month. This approach exploits gradient boosting algorithm, Auto-regression Model, Signal Magnitude Area (SMA), tilt angle, standard deviation mean & median.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

1.1 Problem Statement

Presently, health prediction is by matching information/symptoms shared by user and previously available data/symptoms stored on database, which they collected from the doctors/patient from the previous records. The prediction of health using the symptom graph may not be appropriate prediction, because as per the research conducted sometime user provide wrong information/symptoms or sometimes they were confused about the symptoms.

Hence, by the proposed health prediction model, prediction of health is dependent on life styles instead of the symptoms. The methodology behind is the implementation for discovery of life styles of users after tracking daily activity of user by using their smartphone wearable's sensor data, and prediction of health, having measuring of lifestyle of user.

In our daily lives, hundreds of activities form a significant series that define a user's life. In our approach, in the time frame of second, we used 'activity' word to indicate to the action taken, like "walking", "sitting" and, "typing", meanwhile using word 'daily life style' for actions such as "shopping" or "office work". For example, "daily exercise" life style commonly contains of the "walking", but might also consists the "sitting" or "standing". For model, daily lives design a similarity between daily lives and life documents (Figure 1.1). Probabilistic topic models treats a document as combination of topics and collection of words for a topic [1]. In the same way, combination of topic of life style is document of our daily life and collection of activity word are life style.

Figure 1.1 Relation between documents and Person's daily lives

1.2 Prediction Systems

To predicting a system we generally study the past historical data obtained and study different patterns of results of the market to analyze [3]. The Fig 2.1 shows the fuzzy logic based prediction system architecture. Past performance knowledge is required to predict any event. To learn the existed pattern we generally use the past historical data. Information on learning data of the specific pattern is taken from historical data. Future prediction is the learning from the past provided data knowledge.

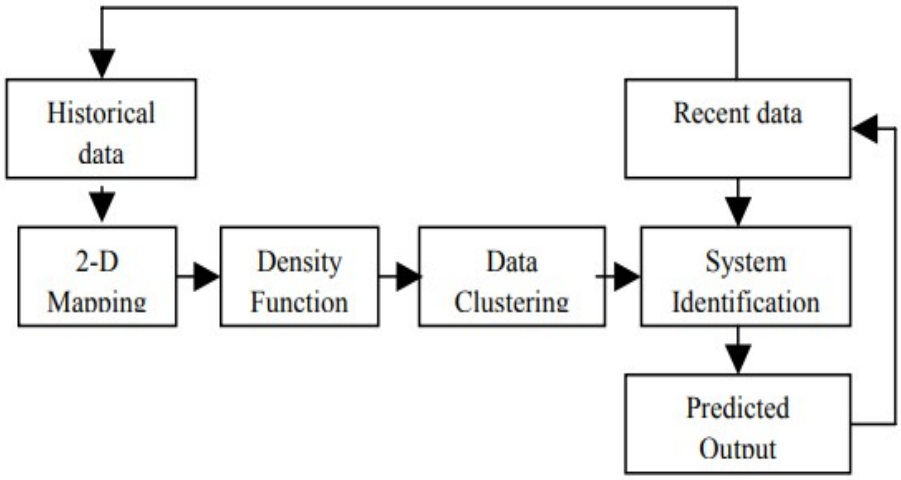

Figure 1.2 Fuzzy Based Prediction System Block Diagram

In prediction system approach, we defined the following phases of its process:-

1. **Information collecting phase**: In this phase, gather the relevant information of users for creating a model.
2. **Analyze collecting information**: In the second phase, analyze/learn the gathered information to define the model for predict the preference information to user.
3. **Prediction:** In the third phase, predict the information which user prefer. The preference information is the filtering information from gathered data for user interest.
4. **Feedback:** In the fourth phase, needed feedback from user to rate for given preference information, on basis of given feedback, we relearn the model for better use rating.

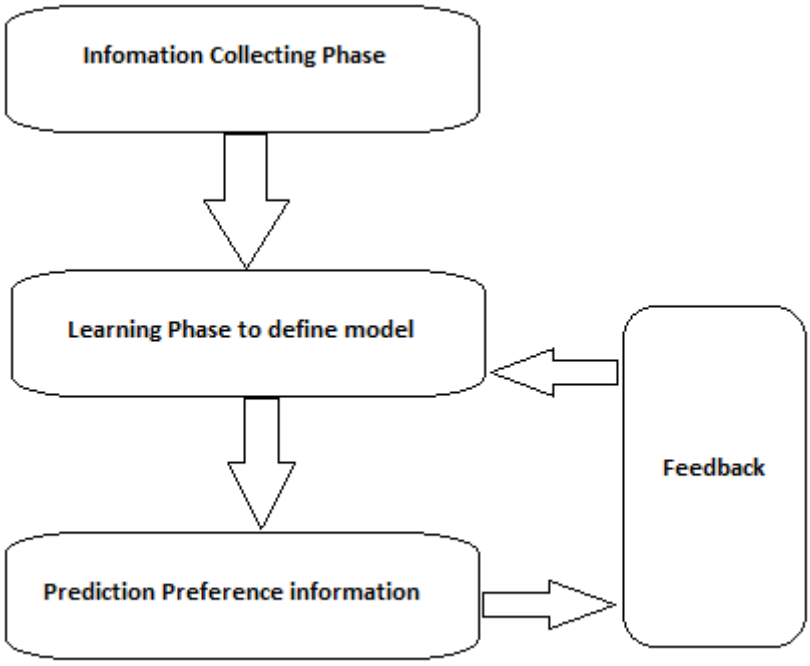

Figure 1.3 Recommendation System Model

1.3 Thesis Motivation and Goal

Mostly prediction system is designed to predict things based on data available for the particular things for which system wants to predict. This system is widely used in prediction in the field of news articles, user's old records available of disease, symptom related information taken from user etc. Regarding health prediction many prediction system are based on user' health graph, which is not appropriate in real life scenario.

Nowadays, the health prediction apps predict the health condition based on users given information. They compare user shared information with old record available and then based on data the predict health. This is not appropriate health predication for user, the reason behind is that each user has their own opinion and don't have much knowledge. So sometime user give wrong information.

To overcome this problem, we have been motivate to give a proper solution to recommend appropriate health prediction on the basis of matching activities performed by user on day to day life style. Based on the activity we can predict the health condition of the user.

1.4 Thesis Organization

We have classified thesis into six different chapters.

Chapter 1 deals with the problem statement for the thesis, the problem statement is the prediction of health of user based on user information by comparing old data available. Such prediction is not the appropriate health prediction Technique. To overcome this problem, we have been motivate to give a proper solution to recommend appropriate health prediction on the basis of matching activities performed by user on day to day life style.

Chapter 2 is describing the related work information.

Chapter 3 is containing the research details. In our research details, we explain the terms, which are being used in the thesis like “Gradient Boosting Algorithm”, “Working of Boosting Algorithm”, “Algorithm”, “Signal Magnitude Area”, “Tit Angle”, “Sensors” etc. Here, we have used Naïve Bayes Algorithms to predict health based on activity performed by user. The “cosine similarity” is employed to compute the likelihood of two vectors (non-zero). In case, the vectors are the part of inner product space, it helps to computes the cosine of the angle between them. The cosine value for 0° is 1, and in another case it is smaller than 1. Thus it is not magnitude judgment, but just orientation. ‘SMA’ is total sum of the magnitude of the all three vectors of acceleration. ‘Tilt Angle’ is the Postural orientation infers to comparative data of the body in the space.

Chapter 4 is containing the proposed approach for Health Prediction based on the basis of user activity. It also explains the system architecture of health prediction in details. Also give information about classifier and the different parameters used to predict the health condition.

Chapter 5 illustrates health prediction results. Using the representation of gray-scale image of 100 users and their 12 activity. Then prediction of user health based on top 6 activities performed by user in 30 days.

Chapter 6 is the conclusion of the thesis. It describes the benefits of that approach for prediction of health on the basis of user’s activity and their life style.

2 Related Work

In Human activity prediction for health care application using smart meter [2] proposes a system that deploy smart home big data as a meaning of learning model and for health care application, predicting human activities pattern. It also include proposed concept as cluster analysis for measuring prediction and analyzing energy usage by occupant's behavior using association rules. Since habits of people are generally identified by day to day lifestyle, to recognize activities which are regular with association of the appliances usage allows discovering such routines.

Disease Prediction Using Patient Treatment History and Health Data [3], the quick reception of electronic Health Records has made an abundance of new information about patients, which is a wealth for improving the comprehension of human Health. The above method is used to predict diseases using patient treatment history and health data.

In 2016 Hossain [8] presented a paper for representing a patient's state recognition system for healthcare using facial expressing and speech recognition. This paper discloses a model to address a general system on health care. It basically deals with the idea of identifying a patient state for giving great acknowledgment accuracy to provide minimum cost modeling. This paper basically depends on two kinds of inputs like audio and video which are caught in a multi-sensory environment which represented an average detection efficiency around 98 percent.

In 2014 Heckerman [11] proposed work on Data Mining using Bayesian Network. This paper provide information on Bayesian network which utilize to learn relationships, also provide help to increase best knowledge about the problem and to forecast the meaning. This provide best representation for combining earlier knowledge and data. This paper explains the making of Bayesian Field networks using various types of methods with the existing knowledge and given data.

In 2011 Han, Pei, and Kamber [15], work in Concepts and Techniques of Data Mining, in General Concepts and Methods of Cluster Analysis. Bayesian network delivers a graphical network of underlying relations because of which we can achieved learning. For the classification Bayesian belief networks are used. These Classifications can be done based on frequent patterns. These frequent patterns brings back the relation among attribute-value pairs. In Associative classification the classification is based on rules produces from quick patterns whereas, semi-supervised categorization is useful for enormous dimensions of unsupervised data.

3 Research Background

3.1 AdaBoost the First Boosting Algorithm

The main acknowledgment of boosting that saw extraordinary accomplishment in application was Adaptive Boosting or AdaBoost for short.

AdaBoost works by weighting the perceptions, putting more weight on hard to classify instances and less on which already handled better way. New learners those are not starong, are added consecutively that give attention on their training on more hard patterns.

Forecasting are done by mostly vote of the weak learners predictions, weighted by their individual accuracy. For binary classification problems, the best form of the AdaBoost algorithm were used and were called AdaBoost.M1.

3.2 Gradient Boosting Algorithm

This framework was implemented by Friedman and known as Gradient Boosting Machines. After that is just called Gradient Boosting or Gradient Tree Boosting.

The statistical framework cast boosting as a numerical optimization problem where the purpose is to decrease the loss of the model by adding weak learners using a gradient descent like procedure.

This class of algorithm was explain as a stage wise linear model. This is because at a time 1 new weak learner is added and existing one weak learner in the model are frozen and left unchanged.

Arbitrary differentiable loss functions to be used allowed by the generalization, expanding the technique beyond binary classification problems to support regression, multi-class classification and more.

3.3 How Gradient Boosting Works

Gradient boosting involves three elements:

1. Optimize a loss function.
2. To make forecasting use a weak learner.

3. To decrease the loss function, an additive model to append weak learners.

3.4 Algorithm

Input training set (i is from 1 to n), a differentiable lossless function L(y, F(x)), M is the number of iteration

Steps of Algorithm:

1. Fist initializing the model with a constant value

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma).$$

2. For m = 1 to M:

    1. Pseudo-residuals should be compute

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n.$$

    2. Choose a base learner (e.g. tree) $h_m(x)$ to pseudo-residuals, set (i is from 1 to n).

    3. To solving below one dimensional optimization problem. Calculate multiplier $_m$.

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right)$$

    4. Update the model:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x).$$

3. Output $F_M(x)$.

3.5 Similarity of Cosine

To compute the likelihood of two vectors (non-zero), cosine similarity is used. These vector are part of inner product space and it computes the cosine of the angle between them. The cosine value for 0° is 1 or is smaller than 1 in other cases. Thus it is not magnitude judgment, it is just orientation. If two vectors same orientation then cosine will be 1, it vectors are at 90° then value will be 0, and if two vectors opposed to each other, then value will be -1, irrespective of magnitude. The name orients from the word "direction cosine".

For instance, in information gathering and text mining, each word is notionally aligned a different dimension and a doc is defined by a vector where the value of every dimension reflects to how many times that word showing in the document. Cosine similarity measure, how likely two documents are to be in terms of their subject information. The technique is also used to compute the cohesion within clusters in the field of data mining.

3.6 Auto Regression Model (AR model)

In this approach, to model the time series signal of different activities, the Auto Regression model are used [18], [19], [20], [22]. "The Auto Regression model in discrete time *t* of a random process *y(t)* is defined by "as given in equation (i) [22]" below

---------- (i)

Where

$a_1, a_2, \ldots, a_P$: the co-efficient of the model

p : Order of the model

ε(t) : Output uncorrelated error

The order of an Auto Regression model derive to the number of previous values of *y(t)* used to calculate the present value of *y(t)*.

The order of Auto Regression model can be determined by running a complete analysis of the extent to which present value of signal *y(t)* is dependent on its previous values. In this model fix order is three.

3.7 Signal Magnitude Area

The SMA (Signal Magnitude Area) [13], [14], [15], [16], [22] is defined as the total of the magnitudes of the acceleration vectors in all three direction. The unit of this measurement is

g. where g is the acceleration by gravity. It is directly proportional to the utilized metabolic energy.

"The classification of the engaged activity is defined by setting of threshold value of SMA. "As given in equation (ii) [22]" below:-

---------- (ii)

3.8 Tilt Angle

In this approach our aim is to provide differentiate between the bodily properties of walking and sitting, as well as the different sub-postures associated with standing. Tilt angle (Φ) [13], [14], [17], [16], [22] is said to be an angles between positive z axis and g. "As given in equation (iii) [22]" below: -

---------- (iii)

It is categorized as upright, if the Tilt Angle is from 0 to 60.

3.9 Sensors

A sensor is an electronic component, whose purpose is to detect changes in its surrounding and transfer the information to other electronic component. Sensors have innumerable applications in wide variety of objects, from a touch sensitive buttons (tactile sensor) to lamps which dark or lighten by touching the basal. The use of Sensors area has been enlarging from fields like pressure, temperature or flow measurement because of enriching the micromachinery. For instance, In MARG (Magnetic, Angular Rate, and Gravity) sensors. Applications of uses sensor in our day-to-day life like cars, smartphone and medicine.

Sensors that are used in this project are ACCELEROMETER and GYROSCOPE. These sensors are already present in android devices. ACCELEROMETER sensor is used to measures the acceleration ($m/s^{2)}$ force all directions including the gravity. GYROSCOPE sensor is used to measures rate of rotation (rad/s) of all directions.

3.10 Activity Recognition

The goal and action of agent are captured through a series of observation by a process of Activity recognition. Given elderly assistance scenario need to be considered for understanding activity recognition better. An elderly man living alone in his house wakes up at dawn. He ignites the stove to prepare tea, turn on oven, and take bread & jam from the cupboard. When he is done with morning medication, an automated voice prompt him to switch-off the oven. Later in the other day, his son surfs on net where he scans a lists, which were prepared by her father's house sensors network. She gets that his father is eating properly, eating drug on time, and managing his life by himself. Because of its multiple feature, different areas may point to activity identification, these activity identification treated as identification of plan, identification of goal, identification on intent, identification of behavior and location based services.

There are many kinds of activity recognition:

a) Single-user, Sensor-based,
b) Levels of sensor-based
c) Multi-user ,Sensor-based
d) Group activity Sensor-based.

4 Proposed Approach

4.1 Activity Recognition

Derivation of $p(w_i|d_k)$ is facilitated by classification or recognition of the user activities. A combination of motion activities, which have distinct occurrences probability that usually reflects Life styles. Therefore, users' motion activities are inferred using two motion sensors, accelerometer and gyroscope. In this model we have adopted one mainstream approach "Gradient Boosting Classifier" for activities recognition.

The activity recognition flowchart is shown in Figure 4 which describes the steps of recognizing activities of human. The raw information has been collected on the phone, and then features is gathered to categories the pre-processed data, thus, further improving recognition accuracy. After testing multiple features like mean, standard deviation, median, Auto regression coefficient, and the collection of them on the data, we find the standard deviation, Auto regression coefficient [18], [19], [20], [22] , SMA [13] , [14], [15], [16], [22] tilt angle [13], [14], [17], [16], [22] are the most representative feature for characterizing motion activities.

This approach using 39 features including (mean_accX, mean_accY, mean_accZ, mean_gyroX, mean_gyroY, mean_gyroZ, median_accX, median_accY, median_accZ, median_gyroX, median_gyroY, median_gyroZ, std_accX, std_accY, std_accZ, std_gyroX, std_gyroY, std_gyroZ, model_AR_acc_coefficeint (9), model_AR_gyro_coefficeint (9), SMA_acc, SMA_gyro, and tilt angle) for Gradient Boosting Classifier Activity Recognition System. 'mean_acc_#': Mean value in all three directions of accelerometer sensor. 'median_acc_#': median value in all three directions of accelerometer sensor, 'sd_acc_#': standard deviation value in all three directions of accelerometer sensor, similarly take same value for gyroscope sensor data. 'AR_acc_coefficeint'(9) : Auto regression co-efficient [34] of all three direction with order 3 , we mean AR_acc_XC1, AR_acc_XC2,and AR_acc_XC3 (three coefficient for X direction of sensor data) , similar taken for Y and Z direction .

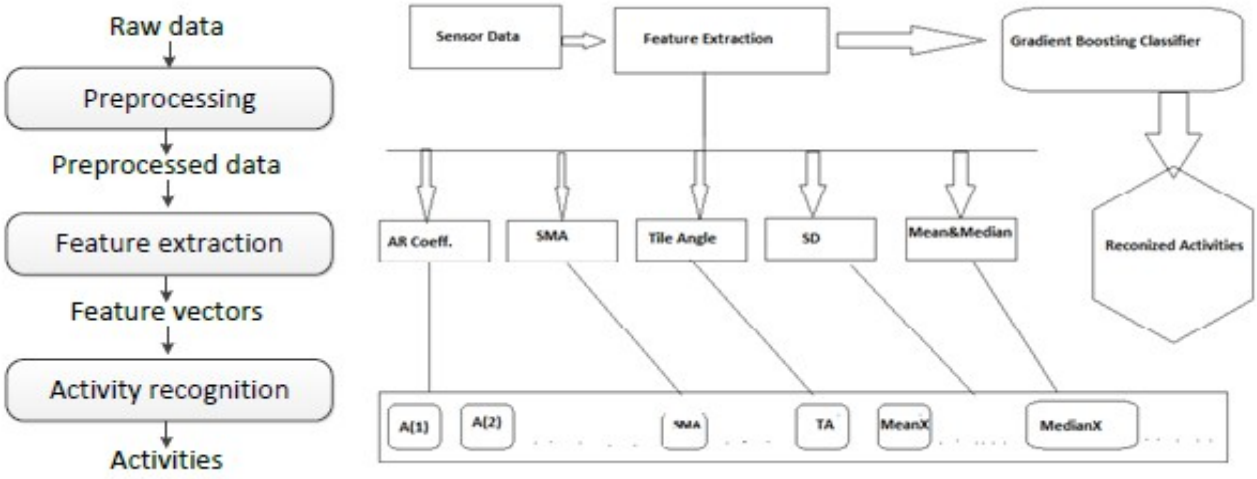

Figure 4.4 The Flowchart of Activity Recognition

In our experiment including with feature mean, median, Standard deviation. AR Coefficient, SMA [13], [14], [15], [16],[22] , tilt angle [13], [14], [17], [16],]22] we got accuracy of 87% in result of activities recognition. We have used the same Gradient Boosting Classifier for activities recognition and calculated the accuracy by using the library "sklearn.ensemble" and "sklearn.metrics" respectively. Source code as follow: -

```
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.metrics import accuracy_score

/* Gradient Boosting Classifier for activities recognition */
clf = GradientBoostingClassifier(estimators=500, learningRate=0.3, maxDepth=1,
randomState=0).fit(xTrain, yTrain)
   predictions =  clf.Predict(xTest)

/* measure Accuracy  score */
 Score  = accuracy_score(y_test, predictions)
```

Figure 4.5 Source Snippet for activities recognition and compute accuracy

The significance of this accuracy of the activity recognition will impact more to define the analogous of user life style consequence. We can recommend more appropriate friend which will be based on user life style and behavior.

4.2 System Architecture of Health Prediction

The Health Prediction system architecture follow client-server model. Where a phone used by a user is a client, while data centers are servers, as shown in the proposed system (Figure 3).

On client side, every user's data (Accelerometer and Gyroscope sensor data) is recorded by his smartphone.

In this approach, activity recognition was done by using Gradient Boosting Classifier. To develop a reasonable activity classifier, Data and training phase is required. Thirty days were spent to build a large training data set by collecting raw data of 100 volunteer.

Collecting the weekly sensor data of 100 people for predefined activities for Gradient Boosting training, we applied it on rest of data for activity recognition. We are able to observe detailed human life-styles by probabilistic-topic model [1], because user continuously accumulates even more activities in life documents.

We have also collected information about their health, after completion of thirty days data collection of mobile sensor of user mobile. We stored this information with the user activity recognized data.

On the server side, we choose only six major activity performed by user. This six major activity plays major role to keep user healthy. This 6 activities are Walking, WalkingUpstairs, WalkingDownstairs, Sittng, Standing and Laying. The activity information table consist of the particular activity performed by user in a day. Further we make Activity-Health Vector table by replacing the activity information by Better, Best, Good and Bad string. The conversion is done on the basis of below information.

Days	Activity Type	Level
0 - 9 Days	Bad	3
10 - 14 Days	Good	2

15 - 20 Days	Best	1
21 - 30 Days	Better	0

Status	Level
Healthy	0
Not Healthy	1

Finally we use this data to predict health condition of user by apply Naïve Bayes Algorithm.

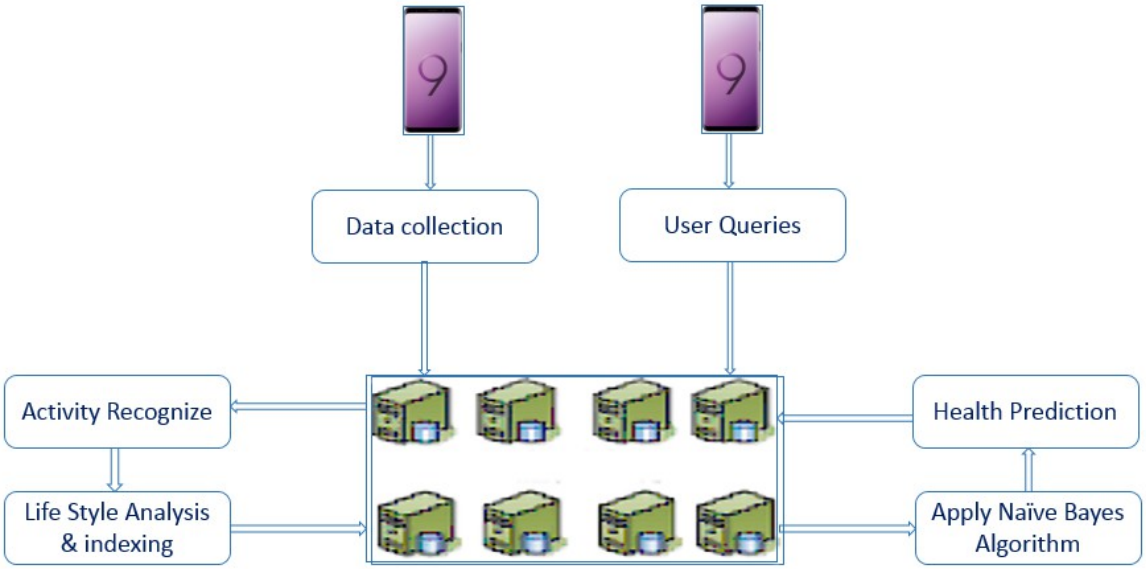

Figure 4.6 System Architecture of Friend Recommendation

4.3 Health Prediction Using Naïve Bayes Algorithm

Here we have used the Naïve Bays Algorithm to predict the health of user using predefined activity and health condition of user, which we formed using 30 days sensor data of 100 users.

Following are the mechanism to predict the health of a user.

4.3.1 The Classifier

The Bayes Naive classifier chooses the most similar classification V_{nb} on attribute value given $a_1, a_2, \ldots, a_n$.

It shows in following form:

$$V_{nb} = \text{argmax}_{vjv} \; P(v_j) \; P(a_i|v_j) \qquad \text{(i)}$$

We normally estimate $P(a_i \mid v_j)$ using m estimates:

$$P(ai|vi) = \qquad \text{(ii)}$$

where:

n = the number of training examples for which $v = v_j$

n_c = number of examples for which $v = v_j$ and $a = a_i$

p = a priori estimate for $P(a_i \mid v_j)$

m = the equivalent sample size

4.3.2 Data Set

Activity	Walking	Walking Upstairs	Walking Downstairs	Sitting	Standing	Laying	Status
User	**A1**	**A2**	**A3**	**A4**	**A5**	**A6**	
1	Better	Best	Good	Good	Bad	Good	Healthy
2	Best	Good	Bad	Bad	Good	Best	Healthy
3	Good	Good	Bad	Good	Good	Good	Not Healthy
4	Best	Good	Good	Good	Good	Bad	Not Healthy
5	Good	Good	Best	Best	Bad	Good	Not Healthy
6	Best	Good	Good	Bad	Best	Good	Healthy
7	Good	Best	Good	Good	Good	Best	Healthy
8	Good	Bad	Bad	Bad	Good	Good	Not Healthy
9	Good	Good	Best	Good	Best	Bad	Healthy
10	Good	Bad	Good	Bad	Bad	Good	Not Healthy

Table 4.1 Data set Table

4.3.3 Health Prediction using Training data

We want to classify a user data based on six activities (A1, A2, A3, A4, A5 and A6) captured from mobile sensor data.

If we can see in equation (ii), we can check computation logic. Probability can be calculate with this.

For health prediction we need to calculate following probabilities and apply this computed value in equation (3) and equation (4). The health will be predicted based on largest value received from the equation (3) and (4). If equation (3) have largest value then user is healthy and if equation (4) has largest value then user has Not-Healty.

P(BetterInA1 | Healthy), P(BestInA1 | Healthy),
P(GodoInA1 | Healthy), P(BadInA1 | Healthy),
P(BetterInA2 | Healthy), P(BestInA2 | Healthy),

P(GodoInA2 | Healthy), P(BadInA2 | Healthy),
P(BetterInA3 | Healthy), P(BestInA3 | Healthy),
P(GodoInA3 | Healthy), P(BadInA3 | Healthy),
P(BetterInA4 | Healthy), P(BestInA4 | Healthy),
P(GodoInA4 | Healthy), P(BadInA4 | Healthy),
P(BetterInA5 | Healthy), P(BestInA5 | Healthy),
P(GodoInA5 | Healthy), P(BadInA5 | Healthy),
P(BetterInA6 | Healthy), P(BestInA6 | Healthy),
P(GodoInA6 | Healthy), P(BadInA6 | Healthy),

P(BetterInA1 | NotHealthy), P(BestInA1 | NotHealthy),
P(GodoInA1 | NotHealthy), P(BadInA1 | NotHealthy),
P(BetterInA2 | NotHealthy), P(BestInA2 | NotHealthy),
P(GodoInA2 | NotHealthy), P(BadInA2 | NotHealthy),
P(BetterInA3 | NotHealthy), P(BestInA3 | NotHealthy),
P(GodoInA3 | NotHealthy), P(BadInA3 | NotHealthy),
P(BetterInA4 | NotHealthy), P(BestInA4 | NotHealthy),
P(GodoInA4 | NotHealthy), P(BadInA4 | NotHealthy),
P(BetterInA5 | NotHealthy), P(BestInA5 | NotHealthy),
P(GodoInA5 | NotHealthy), P(BadInA5 | NotHealthy),
P(BetterInA6 | NotHealthy), P(BestInA6 | NotHealthy),
P(GodoInA6 | NotHealthy), P(BadInA6 | NotHealthy),

P(Healthy), P(NotHealthy)

P(Healthy | A1…A6) = P(Healthy) * P(BetterInA1 | Healthy) *
P(BestInA1 | Healthy) * P(GodoInA1 | Healthy)
* P(BadInA1 | Healthy) *….. * P(GodoInA6 | Healthy)
* P(BadInA6 | Healthy) ---------- (3)

P(Healthy | A1…A6) = P(NotHealthy) * P(BetterInA1 | NotHealthy) *
P(BestInA1 | NotHealthy) * P(GodoInA1 | NotHealthy)
* P(BadInA1 | NotHealthy) *….. * P(GodoInA6 | NotHealthy)
* P(BadInA6 | NotHealthy) ---------- (4)

5 Results & Analysis

5.1 Data Collection and Preprocessing Details

We have collected user's real time sensor data for twelve different activities. For collection of these row data, made a android application which is deployed on user's smartphone. Our application records data of accelerometer and data of gyroscope sensor in x, y & z direction mapping with respective activities.

These raw data collected from 100 users for 30 days. Using this raw data we have made a feature vector having 39 features including mean, median, standard deviation, AR coefficients, SMA and Tilt Angle. This feature vector defines user activity.

To define life style of user, we have prepared a user activity document with the help of feature vector. The twelve different activities that we have used in our recommendation system are as follows:

1. Activity of walking
2. Activity of walking downstairs
3. Activity of walking upstairs
4. Activity of sitting
5. Activity of laying
6. Activity of standing

Collected raw data table format for one activity are as follows:

User ID	AccX	AccY	AccZ
1	1.420	-0.340	-0.125
2	1.002	-0.204	-0.108
3	0.683	-0.061	-0.108
4	0.733	-0.083	-0.120
5	0.956	-0.263	-0.137
6	1.050	-0.402	-0.144
7	1.013	-0.415	-0.104
8	0.950	-0.393	-0.105
9	0.950	-0.359	-0.102
10	0.952	-0.315	-0.086
11	0.913	-0.213	-0.055
12	0.912	-0.125	-0.026
13	0.950	-0.111	-0.063
14	0.969	-0.130	-0.104
15	0.652	-0.075	0.222
16	0.652	-0.075	0.222
17	0.716	-0.055	0.201
18	0.809	-0.141	0.213
19	0.809	-0.141	0.213

Table 5.2 Accelerometer Sensor of WALKING activity

In above Table 5.1 we represent the Accelerometer Sensor data in the form of Acc X, Acc Y and Acc Z for activities like walking, walking downstairs, walking upstairs, sitting, laying and standing.

User ID	GyroX	GyroY	GyroZ
1	-0.275	1.642	-0.0821
2	-0.675	0.670	-0.083
3	-1.133	-0.391	0.118
4	-1.290	-0.763	0.105
5	-1.204	-0.759	0.034
6	-0.853	-0.632	-0.087
7	-0.566	-0.653	-0.118
8	-0.351	-0.733	-0.091
9	-0.175	-0.485	0.066
10	-0.127	-0.409	0.169
11	-0.166	-0.479	0.314
12	-0.300	-0.574	0.445
13	-0.497	-0.525	0.457
14	-0.724	-0.367	0.425
15	-0.790	-0.201	0.368
16	-0.743	-0.191	0.289
17	-0.700	-0.218	0.278
18	-0.674	-0.252	0.262
19	-0.693	-0.427	0.264

Table 5.3 Gyroscope Sensor of WALKING activity

In above Table 5.2 we represent the Gyroscope Sensor data in the form of AccX, AccY and AccZ for activities like walking, walking downstairs, walking upstairs, sitting, laying and standing.

As our experiment on given training and test data, first taken only the features Mean, Median and Standard Deviation with 18 features only. With this features we got accuracy of 64% in result of activities recognition. We used Gradient Boosting Classifier for activities recognition.

Below is the list of the feature names (18 features) for data of gyroscope and data of accelerometer sensor in x, y & z directions.

- **mean_accX:** Mean of X accelerometer sensor's direction data.
- **mean_accY**: Mean of Y accelerometer sensor's direction data.
- **mean_accZ:** Mean of Z accelerometer sensor's direction data.

- **mean_gyroX** : Mean of X gyroscope sensor's direction data.
- **mean_gyroY** : Mean of Y gyroscope sensor's direction data.
- **mean_gyroZ** : Mean of Z gyroscope sensor's direction data.

- **median_accX**: Median of X accelerometer sensor's direction data.
- **median _accY**: Median of Y accelerometer sensor's direction data.
- **median _accY**: Median of Z accelerometer sensor's direction data.

- **median _gyroX:** Median of X gyroscope sensor's direction data.
- **median _gyroY:** Median of Y gyroscope sensor's direction data.
- **mean _gyroZ:** Median of Z gyroscope sensor's direction data.

- **std_accX:** SD(Standard deviation) of X accelerometer sensor's direction data.
- **std_accY:** SD of Y accelerometer sensor's direction data.
- **std_accZ:** SD of Z accelerometer sensor's direction data.

- **std_accX:** SD of X accelerometer sensor's direction data.
- **std_accY:** SD of Y accelerometer sensor's direction data.
- **std_accZ:** SD of Z accelerometer sensor's direction data.

In another experiment, we added a feature AR Coefficient [18], [19], [20] of data of accelerometer and data of gyroscope sensor in x, y & z directions. With this features, we got accuracy of 83% in result of activities recognition. We used Gradient Boosting Classifier for activities recognition.

Below is the list of AR Coefficient feature for data of accelerometer and data of gyroscope sensor data in x, y & z directions.

- **AR_AccX_C1:** first AR co-efficient C1 of X accelerometer sensor's direction data.
- **AR_AccX_C2:** first AR co-efficient C2 of X accelerometer sensor's direction data.
- **AR_AccX_C3:** first AR co-efficient C3 of X accelerometer sensor's direction data.

- **AR_AccY_C1:** first AR co-efficient C1 of Y accelerometer sensor's direction data.
- **AR_AccY_C2:** first AR co-efficient C2 of Y accelerometer sensor's direction data.
- **AR_AccY_C3:** first AR co-efficient C3 of Y accelerometer sensor's direction data.

- **AR_AccZ_C1:** first AR co-efficient C1 of Z accelerometer sensor's direction data.
- **AR_AccZ_C2:** first AR co-efficient C2 of Z accelerometer sensor's direction data.
- **AR_AccZ_C3:** first AR co-efficient C3 of Z accelerometer sensor's direction data.

- **AR_ GyroX _C1:** first AR co-efficient C1 of X gyroscope sensor's direction data.
- **AR_ GyroX _C2:** first AR co-efficient C2 of X gyroscope sensor's direction data.
- **AR_ GyroX _C3:** first AR co-efficient C3 of X gyroscope sensor's direction data.

- **AR_ GyroY _C1:** first AR co-efficient C1 of Y gyroscope sensor's direction data.
- **AR_ GyroY _C2:** first AR co-efficient C2 of Y gyroscope sensor's direction data.
- **AR_ GyroY _C3:** first AR co-efficient C3 of Y gyroscope sensor's direction data.
- **AR_ GyroZ _C1:** first AR co-efficient C1 of Z gyroscope sensor's direction data.
- **AR_ GyroZ _C2:** first AR co-efficient C2 of Z gyroscope sensor's direction data.
- **AR_ GyroZ _C3:** first AR co-efficient C3 of Z gyroscope sensor's direction data.

In our last experiment, we added two more features SMA and tile angle got 84% accuracy in result of activities recognition.

Feature vector table having format as follows:

XMeanAcc	YMeanAcc	ZMeanAcc	XStdAcc	YStdAcc	ZStdAcc	XMedianAcc
1.021	-2.370	-7.280	1.980	1.551	1.421	9.822
9.810	-2.343	-1.648	2.533	1.740	1.362	9.442
1.014	-2.532	-3.826	2.771	2.002	1.584	9.863
1.026	-2.327	-2.713	2.583	2.005	1.742	9.592
9.875	-2.172	-3.622	2.304	1.803	1.401	9.671

Table 5.4 Feature vector table 1

ZMedian Acc	XMeanGyro	YMeanGyro	ZMeanGyro	XStd Gyro	YStd Gyro	ZStd Gyro	XMedian Gyro
-1.101	-4.989	1.466	2.878	5.354	6.327	2.580	-5.320
-3.892	-8.251	-1.506	3.775	5.132	7.495	3.152	5.731
-4.518	-5.023	2.404	-1.272	5.704	9.011	3.436	2.753
-6.183	-9.143	8.293	-1.591	5.020	8.332	3.553	1.292
-4.586	-3.190	-1.991	-2.020	4.631	7.521	3.458	4.411

Table 5.5 Feature vector table 2

Z_median_gyro	XARCoefAacc_1	XARCoefAcc_2	XARCoefAcc_3	YARCoefAcc_1	YARCoefAcc_2	YARCoefAcc_3	ZARCoefAcc_1
3.631	3.511	1.351	-6.911	-6.567	1.511	-7.878	-8.090
5.025	2.453	1.286	-5.276	-8.155	1.295	-6.367	-3.290
3.514	2.765	1.343	-6.148	-8.741	1.397	-7.316	-7.381
3.515	2.232	1.268	-4.802	-7.304	1.338	-6.383	-5.264
4.280	2.429	1.374	-6.139	-5.080	1.311	-5.588	-5.234

Table 5.6 Feature vector table 3

ZARCoefAcc_2	ZARCoefAcc_3	XARCoefGyro_1	XARCoefGyro_2	XARCoefGyro_3	YARCoefGyro_1	YARCoefGyro_2	YARCoefGyro_3
1.241	-3.482	-8.166	1.639	-8.045	8.664	1.443	-6.169
1.196	-4.214	6.182	1.566	-7.582	-1.401	1.181	-4.456
1.222	-4.253	-1.936	1.483	-7.744	1.148	1.166	-5.194
1.260	-4.511	-6.079	1.521	-8.311	2.586	1.209	-5.272
1.031	-1.966	-9.951	1.500	-7.919	3.523	1.232	-4.784

Table 5.7 Feature vector table 4

ZARCoefGyro_1	ZARCoefGyro_2	ZARCoefGyro_3	smaAcc	smaGyro	tiltAngle
7.554	1.442	-6.218	7.071	-4.551	1.641
9.291	1.158	-5.032	7.318	2.277	1.593

Table 5.8 Feature vector table 5

We have used 70% user data as a training dataset and 30% data used to verify model. Accuracy of final model is 84% on testing data.

Activities	WALKING	WALKING_UPSTAIRS	WALKING_DOWNSTAIRS	SITTING	STANDING	LAYING	Status
User	A1	A2	A3	A4	A5	A6	
1	27	19	11	13	8	13	Healthy
2	19	12	7	8	14	15	Healthy
3	12	11	5	12	14	11	Not Healthy
4	15	14	13	10	11	6	Not Healthy
5	10	13	15	19	8	12	Not Healthy
6	18	13	13	7	19	10	Healthy
7	14	17	14	13	11	15	Healthy
8	13	9	6	8	14	13	Not Healthy
9	12	10	16	11	15	9	Healthy
10	12	9	10	7	8	13	Not Healthy

Table 5.9 Activity Summary for model

Activity	Walking	Walking Upstairs	Walking Downstairs	Sitting	Standing	Laying	Status
User	A1	A2	A3	A4	A5	A6	
1	Better	Best	Good	Good	Bad	Good	Healthy
2	Best	Good	Bad	Bad	Good	Best	Healthy
3	Good	Good	Bad	Good	Good	Good	Not Healthy
4	Best	Good	Good	Good	Good	Bad	Not Healthy
5	Good	Good	Best	Best	Bad	Good	Not Healthy
6	Best	Good	Good	Bad	Best	Good	Healthy
7	Good	Best	Good	Good	Good	Best	Healthy
8	Good	Bad	Bad	Bad	Good	Good	Not Healthy
9	Good	Good	Best	Good	Best	Bad	Healthy
10	Good	Bad	Good	Bad	Bad	Good	Not Healthy

Table 5.10 Activity Conversion Table

Days	Activity Type	Level
0 - 9 Days	Bad	3
10 - 14 Days	Good	2

Status	Level
Healthy	0
Not Healthy	1

15 - 20 Days	Best	1
21 - 30 Days	Better	0

Table 5.11 Data Conversion Type Table

User ID	A1	A2	A3	A4	A5	A6	Health Prediction
1	0	1	2	2	3	2	0
2	1	2	3	3	2	1	0
3	2	2	3	2	2	2	1
4	1	2	2	2	2	3	1
5	2	2	1	1	3	2	1
6	1	2	2	3	1	2	0
7	2	1	2	2	2	1	0
8	2	3	3	3	2	2	1
9	2	2	1	2	1	3	0
10	2	3	2	3	3	2	1

Table 5.12 Activity-Health Vector Table

5.2 DATA SUMMARY FOR OUR MODEL:

Total User	Total Activities	Total Features	Total Data Records For Model	Training Data Size (67%)	Testing Data Size (33%)
100	**6**	**39**	**20000**	**13400**	**6600**

Table 5.13 Data Summary for model

In our model we collected data of 100 users. The data contains 6 different type of activity perform by users in there day to day life. The activities are like Walking, WalkingDownstairs, WalkingUpstairs, Sitting, Laying and Standing. Based on activity perform by a user our model predict his/her health condition. We divided our data in 67% of training data and 33% of testing data to measure accuracy.

6 Conclusion

6.1 CONCLUSIVE RESULTS

In our experiment, we collected 100 user's sensor data mapping with respectively 12 activities as Activity of Walking, , Activity of laying, Activity of Walking Downstairs, Activity of Walking Upstairs, Activity of Sitting, Activity of Standing, Activity of position of sit to position of stand, Activity of position of stand to position of sit, Activity of position of stand to position of lie, Activity of position of sit to position of lie, Activity of position of lie to position of stand and Activity of lie to sit. While considering 39 features in our model we got 84% accuracy in activities recognition. Data Summary for our activities recognition model as following:

Total User	Total Activities	Total Features	Accuracy Measure for model of Activity Recognition
100	12	39	84%

Table 6.14 Data summary for Activities Recognition for model

We predict the health on the basis of user's activities by applying Naïve Bayes Algorithm.

Total User	Total Activities	Accuracy Measure for model of Health Prediction
100	6	93.9%

Table 6.15 Data summary for Health Prediction for model

6.2 SUMMARY OF RESULT

The proposal that we are representing contains design and development of Health Prediction system. Here we collected sensor data of user and recognize his 12 different activities using Gradient Boosting Algorithm. Further based on the top 6 activities we predict the health condition of user by applying Naïve Bayes Algorithm.

This approach is different from the other health prediction system here we predict the health of user on the basing of activity performed by user, recognize using 39 different feature.

Implemented Health Prediction System can run on the Android-based smartphones, and can learn things on runtime. We have measured its performance on real small series of data. The results show that the prediction accurately identify health condition of user. The future task would be to determine our proposed system on bigger- series field experiments. We would add more field in our system and predict the diseases of user like headache, fever etc. We would be adding more feedback module to evaluate the impact of health. Also, we would add more devices for discovering more users' activities.

7 REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3:993-1022, 2003.

[2] S.Bakiyalakshmi1, K.Malini2, H.Fametha3 HUMAN ACTIVITY PREDICTION FOR HEALTH CARE APPLICATIONS USING SMART METER

[3] Disease Prediction Using Patient Treatment History and Health Data

[4] L. Bian and H. Holtzman. Online friend recommendation through personality matching and collaborative filtering. Proc. of UBI-COMM,pages 230-235, 2011.

[5] J. Kwon and S. Kim. Friend recommendation method using physical and social context. International Journal of Computer Science and Network Security, 10(11):116-120, 2010.

[6] X. Yu, A. Pan, L.-A. Tang, Z. Li, and J. Han. Geo-friends recommendation in gps-based cyber-physical social network. Proc. Of ASONAM, pages 361-368, 2011.

[7] W. H. Hsu, A. King, M. Paradesi, T. Pydimarri, and T. Weninger. Collaborative and structural recommendation of friends using weblog-based social network analysis. Proc. of AAAI Spring symposium Series, 2006.

[8] L. Gou, F. You, J. Guo, L.Wu, and X. L. Zhang. Sfviz: Interestbased. friends exploration and recommendation in social networks. Proc. of VINCI, page 15, 2011.

[9] Y. Zheng, Y. Chen, Q. Li, X. Xie, and W.-Y. Ma. Understanding Transportation Modes Based on GPS Data for Web Applications. ACM Transactions on the Web (TWEB), 4(1):1-36, 2010.

[10] J. Lester, T. Choudhury, N. Kern, G. Borriello, and B. Hannaford. A Hybrid Discriminative/Generative Approach for Modeling Human Activities. Proc. of IJCAI, pages 766-772, 2005.

[11] Li, J. A. Stankovic, M. A. Hanson, A. T. Barth, J. Lach, and G. Zhou. Accurate, Fast Fall Detection Using Gyroscopes and Accelerometer-Derived Posture Information. Proc. of BSN, pages 138-143, 2009.

[12] E. Miluzzo, N. D. Lane, S. B. Eisenman, and A. T. Campbell. Cenceme-Injecting Sensing Presence into Social Networking Applications. Proc. of EuroSSC, pages 1-28, October 2007.

[13] J. Biagioni, T. Gerlich, T. Merrifield, and J. Eriksson. EasyTracker: Automatic Transit Tracking, Mapping, and Arrival Time Prediction Using Smartphones. Proc. of SenSys, pages 68-81, 2011.

[14] M. J. Mathie, A. C. F. Coster, N. H. Lovell, and B. G. Celler, “A pilot study of long term monitoring of human movements in the home using accelerometry,” J. Telemed. Telecare, vol. 10, pp. 144–151, 2004.

[15] D. M. Karantonis, M. R. Narayanan, M. Mathie, N. H. Lovell, and B. G. Celler, “Implementation of a real-time human movement classified using a tri-axial accelerometer for ambulatory monitoring,” IEEE Trans. Inf. Technol. Biomed., vol. 10, no. 1, pp. 156–67, Jan. 2006.

[16] M. Mathie, B. Celler, N. Lovell, and A. Coster, “Classification of basic daily movements using a triaxial accelerometer,” Med. Biol. Eng. Comput., vol. 42, pp. 679–687, 2004.

[17] C.V. Bouten,K. T.Koekkoek, M.Verduin, R.Kodde, and J. D. Janssen, "A triaxial accelerometer and portable data processing unit for the assessment of daily physical activity," IEEE Trans. Biomed. Eng., vol. 44, no. 3, pp. 136–147, Mar. 1997.

[18] P. H. Veltink, H. B. J. Bussmann, W. de Vries, W. L. J. Martens, and R. C. van Lummel, "Detection of static and dynamic activities using uniaxial accelerometers," IEEE Trans. Rehabil. Eng., vol. 4, no. 4, pp. 375–385, Dec. 1996.

[19] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in Pervasive Computing, Berlin/Heidelberg, Germany: Springer-Verlag, 2004, pp. 158–175.

[20] N. Ravi,N.Dandekar, P. Mysore, and M. L. Littman, "Activity recognition from accelerometer data," in Proc. 20th Nat. Conf. Artif. Intell., 2005, pp. 1541–1546.

[21] U. Maurer, A. Smailagic, D. Siewiorek, and M. Deisher, "Activity recognition and monitoring using multiple sensors on different body positions," and monitoring using multiple sensors on different body positions," pp. 113–116.

[22] Zhibo Wang, Student Member, IEEE, Jilong Liao, Qing Cao, Member, IEEE, Hairong Qi,Senior Member, IEEE, and Zhi Wang, Member, IEEE, "Friendbook: A Semantic-based Friend Recommendation System for Social Networks" – 2013

[23] A. M. Khan,Y.K. Lee, andT.-S. Kim, "Accelerometer signal-based human activity recognition using augmented autoregressive model coefficients and artificial neural nets," in Proc. 30th Annu. IEEE Int. Conf. Eng. Med. Biol. Soc., 2008, pp. 5172–5175.

[24] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava. Using Mobile Phones to Determine Transportation Modes. ACM Transactions on Sensor Networks (TOSN), 6(2):13, 2010.