Indiana University Bloomington

CSCI B 565

Data Mining

IU Bus Route Optimization - Group Name : Decoders

Author:
Dwipam Katariya
Neelam Tikone
Rudrani Angira
Sai Kumar

Supervisor:
Dr. Dalkilic

December 18, 2015

Abstract

Indiana University has a bus service which runs the buses on the Indiana University Bloomington Campus to make the commutation of the students easier around the campus. The buses are named as A, B, E and X which runs through different routes around the campus throughout the day from morning till midnight. The data for Spring 2015 and Fall 2015 Semester is considered and used to achieve aim of optimize the bus routes by identifying the variance of the scheduled and observed times and dynamically allocating the timings for each bus and analyzing how various factors like weather, stop timings play an important role. Throughout the project, our main focus has been to reduce the average time the bus needs to travel between each stop by performing various Mathematical Computations on the given data and training the model using the given data in for the dynamic prediction of the bus timings.

1 Introduction

The bus routes for the University Bus service have been the same for years since when there has been a huge change in the number of students getting admitted to the Indiana University due to various reasons like many new courses being offered each year, increase in the student count for each course etc. The weather also plays an important role in the amount of time the bus takes to travel, hence to make the transportation more efficient, there has been a need to optimize the bus routes according to the current conditions.

About IU Bus System : Following are the architecture details of database system that was used , along with the details of the data :

We had been provided with IU Bus data for Spring 2015 and Fall 2015 semesters starting from the month of January. Weather data for each hour of the day was downloaded using Forcastio

Tables present : Interval data 2014-15, Route ID, Schedule Data, Stop ID, Weather Data, Work Record.

Databases used : MongoDB, SQL Server

Hosted on:Burrows.iu.edu

Languages used : R, Python

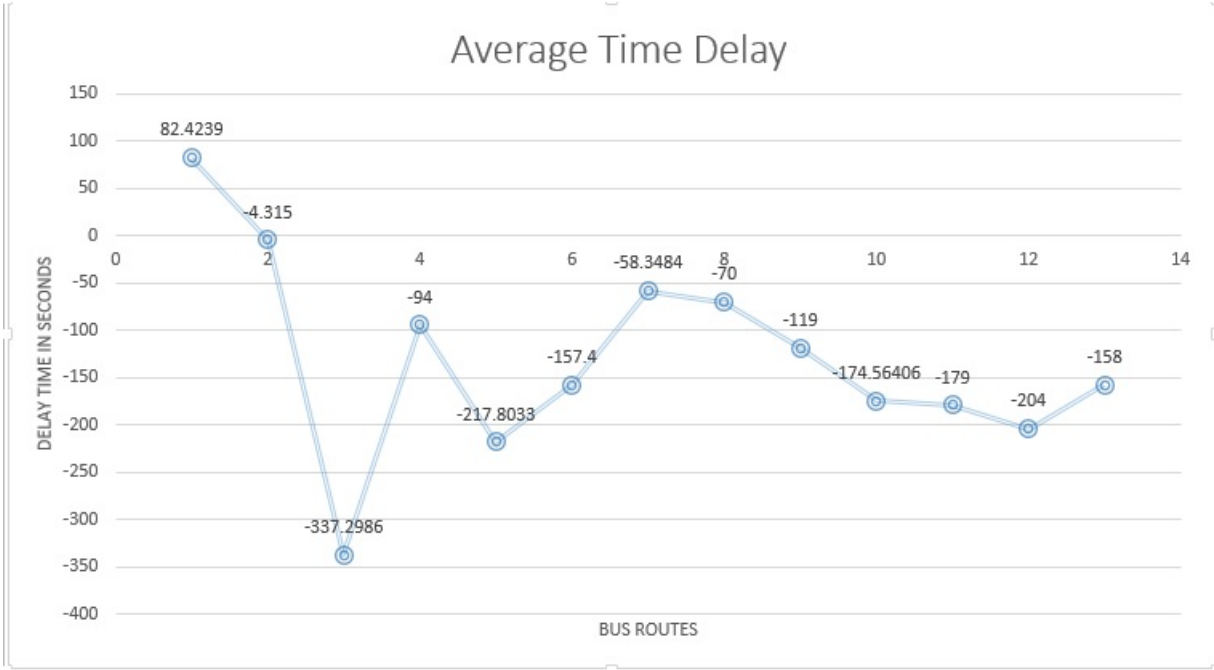
We have total 66 stops in total, 26 buses running and 4 routes.

2 Problem Description

2.1 Presentation of case study

We have analyzed the given Bus Data and prepared The following Analysis:

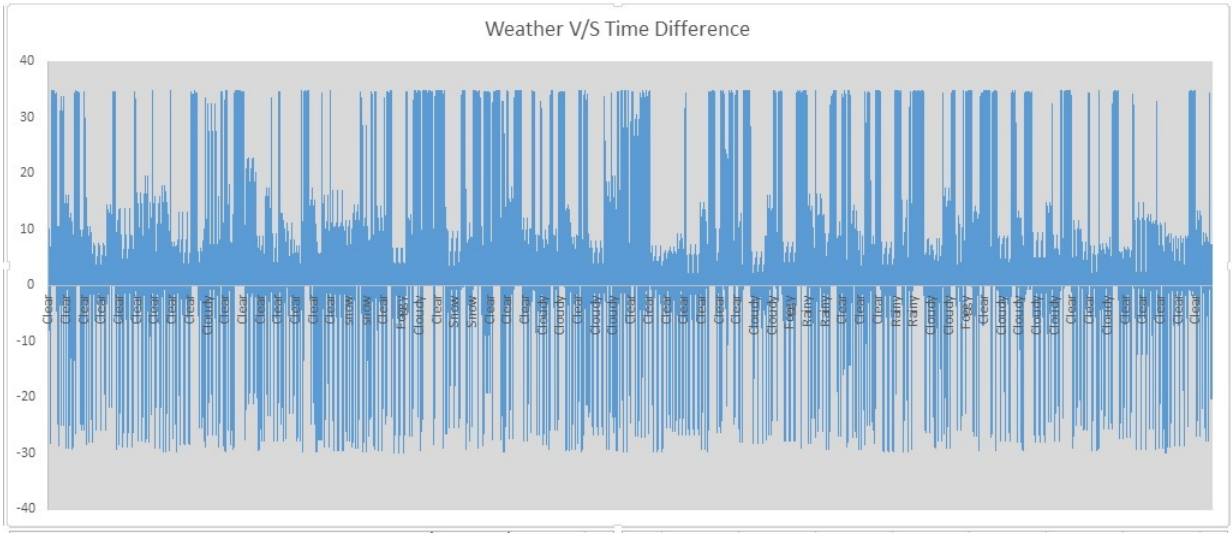
1. Average time to reach Each Stop:



The above is the graph which has Average Delay time in Seconds against the Bus routes. The X Axis contains the id's of the Routes which you can refer from the table Below. The Y Axis contains the Average Delay in Schedule Time and the Observed Time.
Following the table for the Average time for each of the the bus Routes and the Time Difference is calucated in Seconds.
It has been observed that on an Average, only the bus route A on an average is delayed when it has to reach Well's Library from Stadium.
Else, for rest of the routes on an average, the buses have reached before the Schedule time.

1	Route	From	To	Observed Average Time	Scheduled Average Time	Difference
2	A	Stadium	Well's Library	502.4329	420	82.4239
3	A	Well's Library	3rd & Jordan	115.685	120	-4.315
4	A	3rd & Jordan	IMU	226.4918	600	-337.2986
5	A	IMU	Stadium	386	480	-94
6	B	Fishercourt ()	10th & Jordan	142.1967	360	-217.8033
7	B	10th & Jordan	3rd & Jordan	82.60263	240	-157.4
8	E	Evermann	Union & 10th	181.6516	240	-58.3484
9	E	Union & 10th	Wilkie	50	120	-70
10	E	Wilkie	3rd & Jordan	61	180	-119
11	E	3rd & Jordan	IMU	65.43594	240	-174.56406
12	E	IMU	Well's Library	61	240	-179
13	X	Stadium	IMU	216	420	-204
14	X	IMU	Stadium()	262	420	-158

2. The effect of Weather on Time Difference:



X Axis: Weather.
Y Axis: Time Difference.
We calculated the Time Difference for the Observed time and the Scheduled time for the "A" bus to reach the stop. The Negative difference indicates that the bus arrived at the stop earlier than the Scheduled Time, whereas the Positive time indicates that the bus arrived at the stop later than the time it is scheduled to arrive.
We then Plotted this Time difference with the Weather Data of each day which we downloaded using Forcastio.
After plotting both the data against each other, we observed that the

weather does not affect the Timings of the bus reaching the stops.We observed and compared the time difference during the various weather condtions like "Cloudy", "Snowy", "Rainy", "Clear Day" and observed that there has not been any significant time difference due to the occurrence of any of the Weather Conditions.

3 Data Description

3.1 Existing model

In the current IU bus service, the buses namely A,B,E, and X are observed to run from morning till midnight. The following is the analysis which we did using the given data:

1. The Bus has a time constraint only during the start of the shift. That is, only for the first stop, the bus needs to be on time. Rest of the stops, the time is not considered by the driver, hence sometimes the bus reaches the stop much earlier than the scheduled time whereas, sometimes it reaches the stop much later than the scheduled time.
   In the table below contains few examples from the merged data in which the column named time_diff shows the difference in the Scheduled and Observed Time of the "A" Bus.

ID1	id	observed_date	bus_id	to	From	time	Shift.ID	Driver	Week_Day	Shift	Bus_no	observed_time	Scheduled_time	time_diff
886456	6235311	12-01-2015	642	Neal Marshall	Well's Library	135	2784	Gallalee, Jack	M	1	A6	09:19:04	09:16:00	3.07
886476	6235331	12-01-2015	650	IMU	IMU	48	2724	Foley, Shannon	M	1	A4	09:20:32	09:21:00	-0.47
886476	6235331	12-01-2015	650	IMU	IMU	48	2725	Warren, Jeremy	M	2	A4	09:20:32	09:21:00	-0.47
886483	6235338	12-01-2015	650	10th & Woodlawn	IMU	80	2724	Foley, Shannon	M	1	A4	09:21:04	09:21:00	0.07
886483	6235338	12-01-2015	650	10th & Woodlawn	IMU	80	2725	Warren, Jeremy	M	2	A4	09:21:04	09:21:00	0.07
886501	6235356	12-01-2015	642	3rd & Jordan	3rd & Jordan	47	2784	Gallalee, Jack	M	1	A6	09:22:09	09:18:00	4.15
886502	6235357	12-01-2015	645	Stadium (A)	Stadium (A)	7	2720	LaGarde, Nancy	M	1	A2	09:22:09	09:26:00	-3.85
886510	6235365	12-01-2015	642	Jordan Hall	3rd & Jordan	94	2784	Gallalee, Jack	M	1	A6	09:22:55	09:18:00	4.92

2. The bus id for each bus changes with the changes in the shift. The first bus in each shift will have the same bus id accordingly, the second, third and the fourth bus in each shift will have the same bus id. The inconsistencies have been found in the bus id's where the bus id's changes at any random stop.

3. In the current system, The bus id's haven't been assigned properly. We have observed that sometimes, the same shift, the two "A" buses have been assigned the same bus id's which increases the difficulty in interpretation of the data. The following is one of the example:

```
Buss number: 639
1/22/2015: R-A4.1, R-A7.1
```

In the above provided example, On Thursday's first shift, the bus id: 639 has been assigned to two different buses which are A4 and A7, which is the 4th and the 7th "A" bus in the same shift.

4. It is observed that sometimes all the work records are not stored properly. Hence, the Data for extracting the bus route to track the particular bus becomes incomplete.

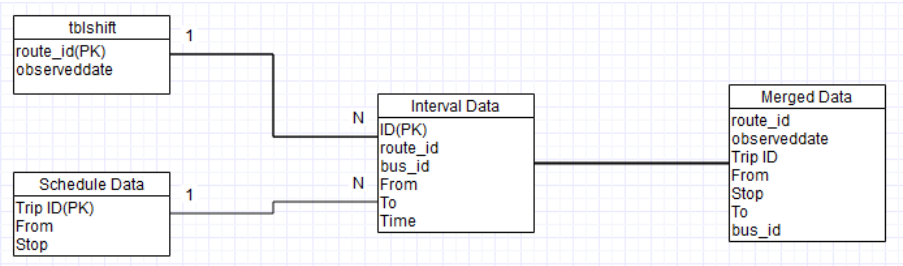
3.2 Proposed model

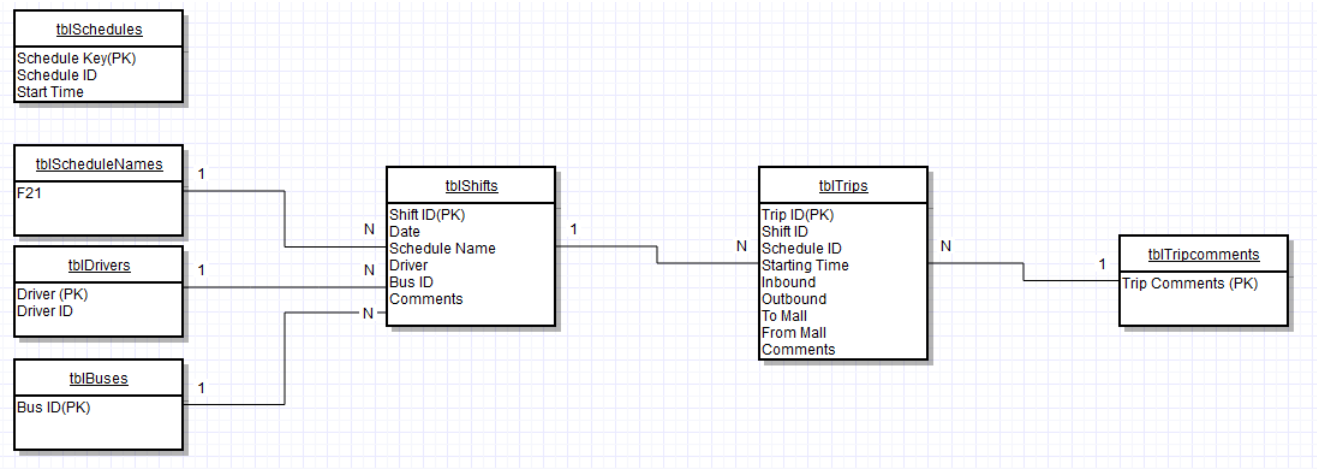
After Observing the bus data given to us of the Speing and Fall Semester, we came up with the following Proposals:

1. One of the major problem we observed while Analyzing the bus data was figuring out the Bus id's. So, we propose that unique bus id's should be given for each shift. i.e for shift A, the bus id's should be 1,2,3,4. and for shift B the bus id should be 5,6,7,8. and so on. Following this pattern will help us know which buses run in which shift.

2. A single table should be maintained to calculate the time difference of the observed time and the Scheduled time of the bus simultaneously in the table.

3. After working on the Bus data, the following are the UML diagrams of the models which we want to propose:





The data set which we propose consists of the route model for every route.The following is the description about the data based on the columns present:

id: Consists of the unique id that represents the entire tuple.
observed_date : Contains information of the date the bus runs.
bus_id : It is the unique identification number assigned to each bus.
to: the stop from which the bus started.
From: the destination stop of the bus
time: Consists of the time taken to travel between the stops in seconds.
Shift ID : the unique ID given to a particular shift
Driver: the driver assigned to the particular shift
Week_Day : M represents Monday, T represent Tuesday, W represents Wednesday,R represent Thursday,F for Friday,S for Saturday and U for Sunday.
Shift : Each route runs in three shifts whre 1 represents 1st shift,2- 2nd shift and 3-3rd shift
Bus_no: Contains information regarding the bus. Here A1 represents a route 'A' bus in shift 1.
observed_time:It is the actual time observed upon reaching a particular stop.
Scheduled_time:It is the time that the bus is actually scheduled or expected to arrive the stop.
time_diff:It is the difference (observed-scheduled) represented in seconds.

4 Computational Implementation

For this project, we have followed two approaches:

1. Mapping the Data in Java using Hash Maps and using the data as a key value pair for every bus id in a following way:

   ```
   Map<Date_BusID , Route>schedule
   ```

   Above, we have created a Key-Value Pair where key is Bud$_i$*datthepartic*ular*dateandRouteisth*

   As well as:

   ```
   Map< Shift , ArrayList Stop<ArrayList arrivalTime <ArrayList DepartureTime
   <ArrayList<TimeDifference>>>>> Route
   ```

   Above we are using shift as a key and arraylist structure as a Value which includes the following data about the bus: Stop, Arrival Time, Departure Time and Time Difference to generate the Route of each Bus.

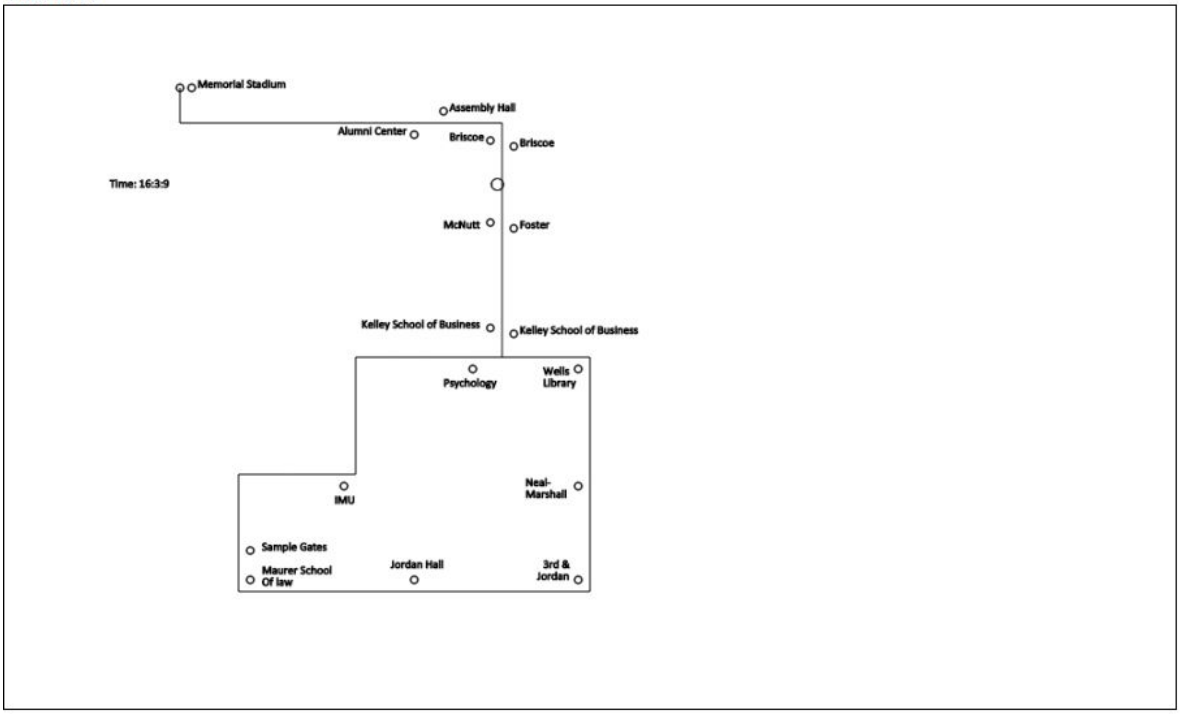
2. Using R to merge the table using the Primary key and unique column ids: As per the Schema Definition model, we merged the data with respect to the primary keys thereby doing Inner and Outer Join.

```
[Interval_data \Schedule_data]^{I_I} y_{\Interval_data \schedule_data}^{I_J} z_{\Interval_data.trip_id, Schedule_data.trip_id,Interval_data.from, schedule_data.from}->
entire_data[Interval_data[i]][schedule_data[j]
frac{x_{\Interval_data \tblschedule^{I_I} y_{\Interval_data \tblschedule}^{I_J} z_{\Interval_data."bus_id","observed_date","route_id", tblschedule."bus_id","observed_date","route_id",Interval
entire_data[Interval_data[i]][schedule_data[j]
Deviate[Actual_time\observed_time]^{I_I} -> Interval_data.actual_time^[I_I]-Interval_data.schedule_time^[I_I]
```

5 Visualizations

We created a Visualization using HTML 5 and JQuery for the Route "A" to provide the real time bus location at a given time.

Time: 16:3:9



6 Proposed Changes

1. We propose to change the schema definition of the current architecture model to our new architecture, where all the data are consolidated into one table. This approach gives us ease to mine the data using data mining algorithms for classification and regression. Every sample of data can be associated with the time deviation as label and can help us take an accurate decision on the same.

2. We propose to install a pedestrian signal at Kelly, as we can analyse that , major times bus gets late at Kelly due to student crossing and inconsistency between previous signal and next. Let this pedestrian signal by sync with the main signal of wells library, this will Reduce the delay and for reduction for other trips, and buses behind it.

3. We also propose for dynamic scheduling of drivers and maintaining accurate data of the the employees, as this can hep for dynamic allotment to the bus.

4. We propose to maintain the bus_id unique to every bus type, as this would allow for easy mapping of the data.

7 Concluding remarks

We can conclude that given the dynamic model for Naive Bayes, we can actually predict if the bus would be late for the next trip and thus take an absolute decision for the number of buses, speed limit, passenger intake and scheduled time for each and every stop.
We have satisfied following goals for this project:

1. Calculating the Time Difference for the Observed and Scheduled time for each bus and its respective stop.
2. Identified the routes for all the buses.
3. Analyzed the effect of weather on the Time Deviation.
4. Calculated the Average time of commutation between the major stops of each route.
5. Visualization of the Proposed Model.

8 References

1. www.forcastio.com - used to retrieve information regarding the actual weather data.
2. https://cran.r-project.org/doc/manuals/r-release/R-intro.html - referred for the help on R modules.
3. https://bloomington.doublemap.com/map/ -used for receiving the information about the bus location details.