Linköping University | Department of Computer and Information Science
Master thesis, 30 ECTS | Computer Science
202018 | LIU-IDA/LITH-EX-A--2018/XXX--SE

BEAUTIFUL TITLE, ENGLISH

VÄLDIGT FIN TITEL

Linus Kortesalmi

Supervisor : Mattias Tiger
Examiner : Fredrik Heintz

External supervisor : Simon Johansson

LINKÖPING UNIVERSITY

Linköpings universitet
SE–581 83 Linköping
+46 13 28 10 00 , www.liu.se

Upphovsrätt

Detta dokument hålls tillgängligt på Internet – eller dess framtida ersättare – under 25 år från publiceringsdatum under förutsättning att inga extraordinära omständigheter uppstår. Tillgång till dokumentet innebär tillstånd för var och en att läsa, ladda ner, skriva ut enstaka kopior för enskilt bruk och att använda det oförändrat för ickekommersiell forskning och för undervisning. Överföring av upphovsrätten vid en senare tidpunkt kan inte upphäva detta tillstånd. All annan användning av dokumentet kräver upphovsmannens medgivande. För att garantera äktheten, säkerheten och tillgängligheten finns lösningar av teknisk och administrativ art. Upphovsmannens ideella rätt innefattar rätt att bli nämnd som upphovsman i den omfattning som god sed kräver vid användning av dokumentet på ovan beskrivna sätt samt skydd mot att dokumentet ändras eller presenteras i sådan form eller i sådant sammanhang som är kränkande för upphovsmannens litterära eller konstnärliga anseende eller egenart. För ytterligare information om Linköping University Electronic Press se förlagets hemsida http://www.ep.liu.se/.

Copyright

The publishers will keep this document online on the Internet – or its possible replacement – for a period of 25 years starting from the date of publication barring exceptional circumstances. The online availability of the document implies permanent permission for anyone to read, to download, or to print out single copies for his/hers own use and to use it unchanged for non-commercial research and educational purpose. Subsequent transfers of copyright cannot revoke this permission. All other uses of the document are conditional upon the consent of the copyright owner. The publisher has taken technical and administrative measures to assure authenticity, security and accessibility. According to intellectual property law the author has the right to be mentioned when his/her work is accessed as described above and to be protected against infringement. For additional information about the Linköping University Electronic Press and its procedures for publication and for assurance of document integrity, please refer to its www home page: http://www.ep.liu.se/.

© Linus Kortesalmi

Abstract

Abstract.tex

Acknowledgments

Acknowledgments.tex

Contents

Abstract iii

Acknowledgments iv

Contents v

List of Figures vi

1 Introduction 1
- 1.1 Motivation . . . 1
- 1.2 Aim . . . 1
- 1.3 Research Questions . . . 1
- 1.4 Delimitations . . . 1

2 Theory 2

3 Data 3
- 3.1 Background . . . 3
- 3.2 Structure . . . 4

4 Experiments 7

5 Discussion 8

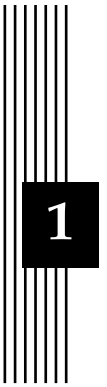
6 Conclusion 9

List of Figures

3.1 Simplified graph illustrating the data gathering process. Each bus is equipped with a GPS sensor and transmits its position to a central server/database. The dataset used in this thesis project is the log, which is a collection of documents. Each document contains the GPS data sent from all buses during a single day, together with data from the "Internal Analysis" component of the server. 4
3.2 Example of a raw `ObservedPositionEvent` entry. The header and the body is separated by `|||`. Each parameter in the header and body is separated by a single `|`. Key parameters for the `ObservedPositionEvent` event type is highlighted. . 4
3.3 The distribution of event types for a random day in the dataset. 5

Todo list

300? 3
2.5? 3
Is it interesting to plot this exact number and its variance? 4
Should we do a boxplot instead of the whole dataset? Will take a lot of time to create! . 5



1 Introduction

1.1 Motivation

1.2 Aim

1.3 Research Questions

This work explicitly answers the following questions:

1.
2. What methods can be used to detect anomalies in a dynamic real-time system?
3. How can the need for maintenance be detected from real-time processing of domain-specific data?
4. How can Predictive Maintenance be achieved by processing real-time domain-specific data?
5. How many days in advance can maintenance be predicted?
6. How can Predictive Maintenance be evaluated and compared to Corrective Maintenance?
7. How can faults be identified and isolated using Predictive Maintenance?
8. What data is required to realise fault identification and fault isolation in Predictive Maintenance?

1.4 Delimitations

The data available is from an existing IoT application mainly used to control units. The available sensors in the application can not be changed and additional sensors can not be added.

2 Theory

3 Data

This chapter describes the spatiotemporal dataset used in the thesis project. The data provider is briefly mentioned, followed by the structure of the data. This chapter also mentions issues encountered in the data set.

3.1 Background

The dataset was provided by Östgötatrafiken AB and contains GB of data. Östgötatrafiken AB is owned by Östergötland County and is responsible for the public transportation in the county. This thesis project only analysed the bus data available in the given data set. The dataset is a collection of documents, where one document represents a full day of data. A typical day has a document size of around GB.

300?

2.5?

Data Gathering

The process of gathering the data used in this thesis project can be generally described by the following simplified procedure:

1. Each Östgötatrafiken AB bus is running a system collecting data from sensors installed inside the bus.
2. The system collects the sensor data and transmits it to a central server or database.
3. A log containing all events for a full day is created and stored as a document in a collection.
4. The central server processes and analyses the data. The results from the data analysis is stored in the log.

Figure 3.1 illustrates the procedure. The collection of logs is the dataset used in this thesis project. The logs contain the GPS data from the buses and also events created by the "Internal Analysis" component in the system. The "Internal Analysis" component is simplified and is beyond the scope of this thesis project.

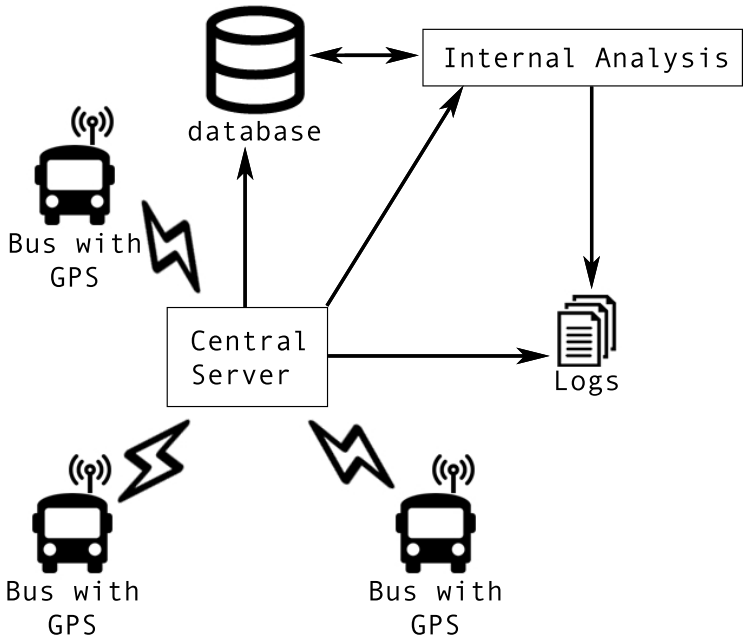


Figure 3.1: Simplified graph illustrating the data gathering process. Each bus is equipped with a GPS sensor and transmits its position to a central server/database. The dataset used in this thesis project is the log, which is a collection of documents. Each document contains the GPS data sent from all buses during a single day, together with data from the "Internal Analysis" component of the server.

3.2 Structure

A document in the dataset is made up of a large number of events representing a single day. A single day typically contains roughly 21 million events. Each event is represented by a single line of text. An event can be split into two groups: a header and a body. There are different types of events reported during the span of a single day. Each type has its own header and body structure.

Is it interesting to plot this exact number and its variance?

Event Example

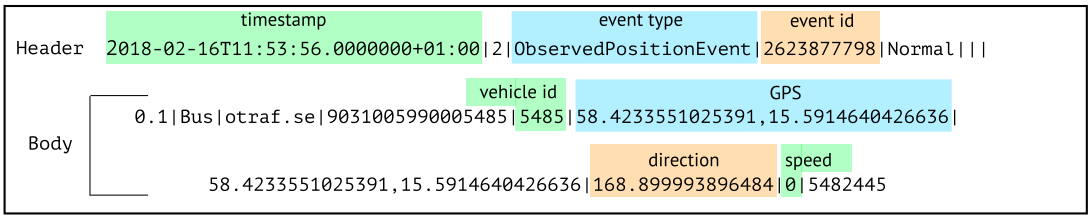


Figure 3.2: Example of a raw `ObservedPositionEvent` entry. The header and the body is separated by `|||`. Each parameter in the header and body is separated by a single `|`. Key parameters for the `ObservedPositionEvent` event type is highlighted.

Figure 3.2 illustrates an event with the event type `ObservedPositionEvent`. The header is defined as all the parameters before the `|||` separator. All the parameters after

the separator is defined as the body of the event. In this example the header and body contain seven key parameters:

- *Timestamp:* A timestamp (`2018-02-16T11:53:56.0000000+01:00`), which is the timestamp from the system running on the bus.
- *Event Type:* The event type (`ObservedPositionEvent`).
- *Event ID:* The event id (`2623877798`). This is a number set by the system responsible for collecting the data from all buses. It is incremented for every event added to the log by either the database system or the "Internal Analysis" component in Figure 3.1.
- *Vehicle ID:* Unique ID for the bus transmitting its position.
- *GPS:* The GPS position of the bus in latitude and longitude.
- *Direction:* The direction of the bus.
- *Speed:* The current speed of the bus.

Event Types

The dataset contains 20 unique event types. Figure 3.3 visualises the distribution of event types for a random day in the dataset. The figure only gives an indication of what the true distribution could be, as it is computationally expensive to calculate the true distribution for the given dataset, due to its size. Knowledge about the true distribution is not required in order to reason about the event types. As the figure shows, the majority of events that occur are of the type `ObservedPositionEvent`, which is the event containing an updated GPS position for a vehicle.

Should we do a boxplot instead of the whole dataset? Will take a lot of time to create!

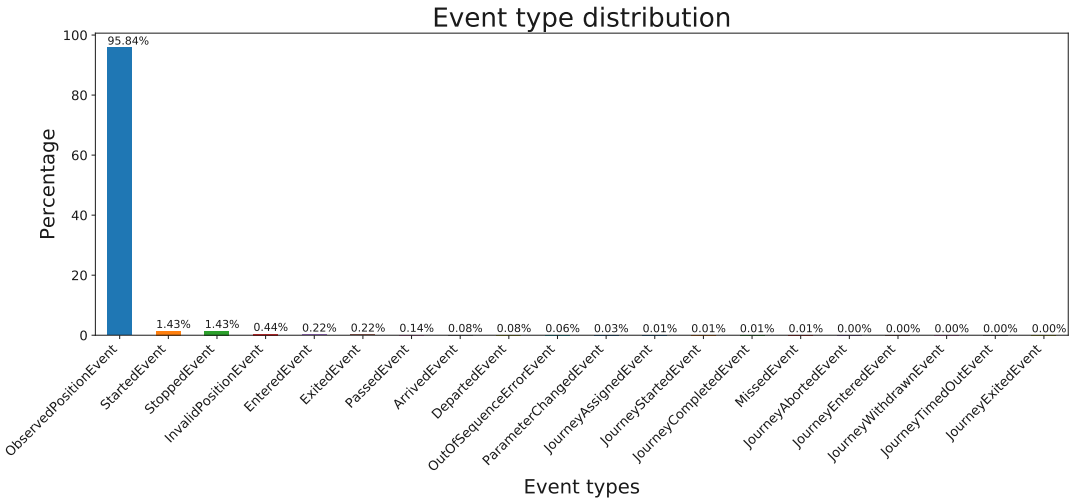


Figure 3.3: The distribution of event types for a random day in the dataset.

Of the 20 event types available in the dataset, this thesis project only used 11 of them. The events to use were chosen by analysing the log for a single day in great detail. Event types which occurred rarely and seemingly random were discarded, as no pattern could be determined for them. The `InvalidPositionEvent` type only contains the GPS position of the vehicle. The valid `ObservedPositionEvent` type also contains the *Speed* and *Direction* parameters. The GPS position of the `InvalidPositionEvent` events were always the same coordinates. In this thesis project the `InvalidPositionEvent` type was discarded, due to the missing parameters and static GPS position.

The 11 event types used in this thesis project were:

- *ObservedPositionEvent:*
- *StartedEvent:*
- *StoppedEvent:*
- *InvalidPositionEvent:*
- *ExitedEvent:*
- *EnteredEvent:*
- *PassedEvent:*
- *DepartedEvent:*
- *ArrivedEvent:*
- *ParameterChangedEvent:*
- *JourneyStartedEvent:*
- *JourneyAssignedEvent:*

4 Experiments

5 Discussion

6 Conclusion