Temporal association rule mining for the preventive diagnosis of onboard subsystems within floating train data framework

Wissam Sammouri

Université Paris-Est, IFSTTAR, GRETTIA

Instructor: Etienne Côme

Supervisor: Latifa Oukhellou

12th of June, 2012

IFSTTAR

Outline

Outline

Introduction

The recent advancement in Information and Communication Technologies (ICT) have brought important innovations that were essential in turning railways into an intelligent transportation system:

- Improvement in safety measures
- Support to operations (monitoring and control systems)
- Customer services (passenger information, electronic ticketing)
- Rolling-stock maintenance and condition monitoring

Maintenance of railway subsystems

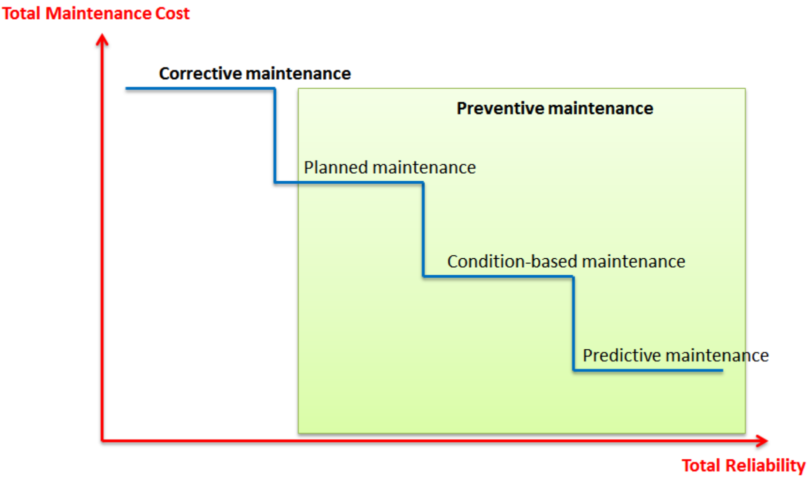
Trains today are complex, real-time, distributed and reconfigurable systems, incorporating many embedded subsystems, which concur together in performing a high quality transportation service.

Failure of any of such subsystems can have heavy impact on the service itself:

- deterioration of performance and perhaps mandatory stop
- reduction of perceived quality
- increment of costs

$\Longrightarrow$ evolution of maintenance strategies and processes towards more optimized and cost-effective solutions.

Possible maintenance strategies



Floating train data systems

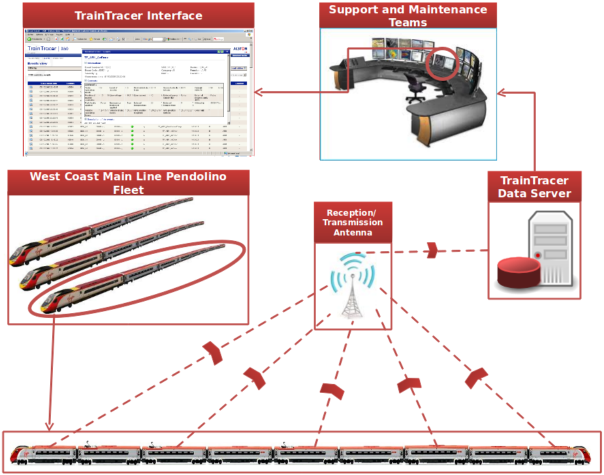
Floating train data systems (FTD):

- Commercial trains are equipped with positioning (GPS) and communications systems
- Onboard intelligent sensors monitoring various subsystems on the train

$\Longrightarrow$ each train can be seen as a mobile sensor that operates in a distributed network to collect a large amount of data transferred back to the ground automatically via wireless technology.

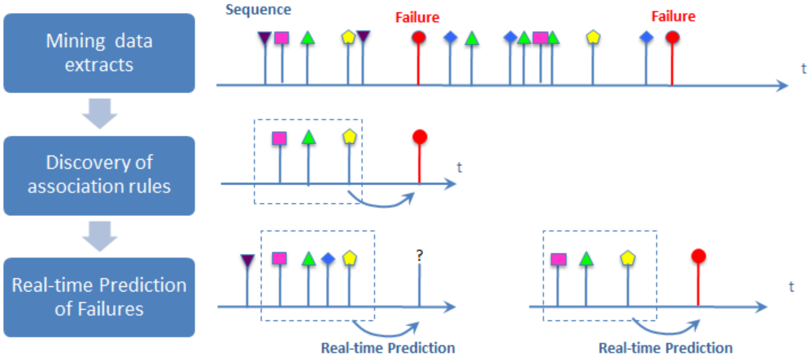
The floating train data system provides a real-time flow of information consisting of georeferenced alarms, called events, along with their spatial and temporal coordinates.

TrainTracer: FTD system developed by Alstom Transport



Problematic

Apply temporal data mining techniques on an extract of the TrainTracer data to discover temporal associations between timestamped alarms, that can predict the occurrence of severe failures within a complex bursty environment.



TrainTracer data extract

- Alstom FTD system: TrainTracer
- Temporal sequence of alarms extracted from the TrainTracer database, West Coast Main Line Network (52 Pendolino Trains).
- Time period covered: 6 months
- 9,046,217 alarms
- 1113 type of alarms
- 31 subsystems

Alarms categories

5 alarm intervention categories

- Status: Cat 1
- Driver Information: Cat 2
- Driver Intervention Low: Cat 3
- Maintenance: Cat 4
- Driver Intervention High: Cat 5

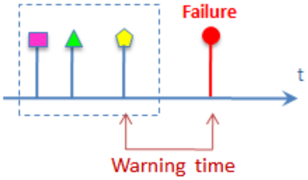
Target alarms: Tilt and Traction Cat.5

- 69 alarm types in total, 46 existing in data
- Count in all data: 40902 $\simeq$ 0.4521%

Constraints

Prediction Constraint:

- Warning time $>$ 30 minutes
- High Precision



Obstacles

Obstacles:

- Rareness of target alarms
- High noise frequency due to redundant alarms, alarms with no valuable information and alarms due to driver intervention.
- High number of bursts
- Heavy Calculation time
- Weak Apriori knowledge on the design of pendolino trains and the relevance of each alarm

Orientation:

Concentrating on extracting L2 patterns (association rules $A \longrightarrow B$) that end with target alarm (tilt and traction Cat.5)

General Methodology

General methodology applied to all algorithms:

1. Establish an algorithm
2. Test the efficiency of the algorithm on a toy dataset
3. Run the algorithm on TrainTracer data
4. Result analysis

Outline

1. Introduction summary
   - Introduction
   - Problematic
   - Constraints and Obstacles
   - General Methodology
2. T-Patterns Algorithm
   - T-Patterns - Aim and Principle
   - Methodology
3. Results and discussion
4. Conclusion and Future work

T-Patterns (Magnusson 2000, Tavenard 2007, Salah 2010)

Aim: study the dependency between couples of alarms

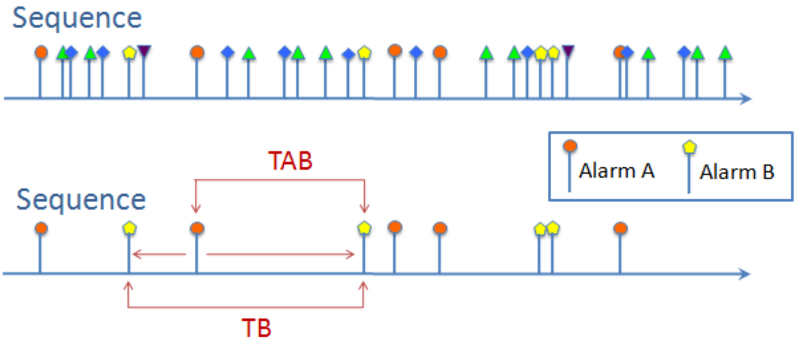
Two temporal point processes A and B are considered to be independent if the arrival of an A-alarm does not lead to an increase in the probability of occurence of a B-alarm.

Independency is identified by means of a hypothesis test, where:

- H_0: Alarms A and B are independent
- H_1: Alarms A and B are dependent

H_0 is accepted/rejected with respect to a threshold α

T-Patterns - Main proposition



T_{AB}: the time distance between each A-alarm and the first succeeding B-alarm
$\tilde{T}_B$: the time interval between two successive B-alarms between which at least one A-alarms occured

T-Patterns - Key property

Proposition: If A and B are independent temporal point processes, then $T_{AB} \sim U(0, \tilde{T}_B) \Longrightarrow \dfrac{T_{AB}}{\tilde{T}_B} \sim U(0,1)$

T-patterns pseudo-code

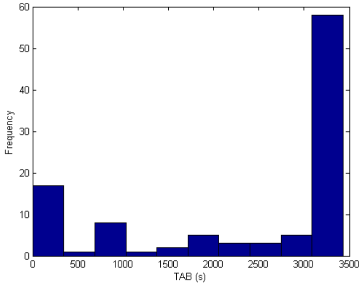
- A-List: List of all alarms occuring in data
- B-List: List of target alarms occuring in data

For every combination of A and B alarms in the A-list and B-list:

1. Extract all timestamps of A
2. Find the first B succeeding the A-alarm and Calculate T_{AB}
3. Calculate $\tilde{T}_B$
4. Calculate the ratio vector U
5. Test if the ratio vector U is uniformly distributed using a Kolmogorov Smirnov statistical test
6. If H_0 is rejected, $A \to B$ is considered to be statistically significant and added to a list of possibly dependent AB couples.

Modeling inter-event time intervals

A preliminary approach towards modeling inter-event time intervals is the T_{AB} frequency histogram. This histogram provides a visual representation of the distribution of inter-arrival times between an A-alarm and a B-alarm.



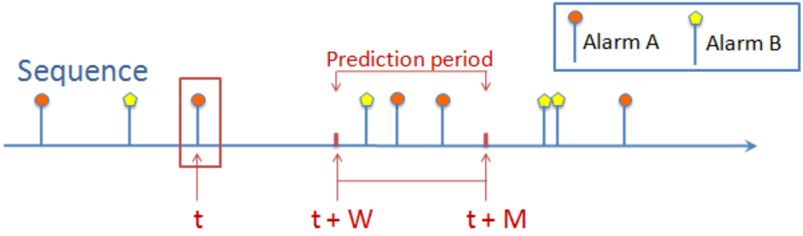
AB-couples are filtered with respect to their histogram peaks.

Warning time and Monitoring time

A prediction is correct if a target alarm occurs within its prediction period. The prediction period is defined by a warning time, W, and a monitoring time, M.

Warning time: time delay before a target alarm becomes highly probable to occur

Monitoring time: how far into the future the prediction extends



Recall and Accuracy

Other measures to examen couples: Recall and Accuracy

$$Recall = \frac{\#\ Predicted\ target\ alarms}{Total\ target\ alarms}$$

$$Precision = \frac{\# True\ predictions}{\# Total\ predictions}$$

- High recall means no target alarms were missed
- High precision reflects a high predictive capability, but might imply a low recall if a big precentage of target alarms weren't detected

$\Longrightarrow$ importance of a Recall-Precision trade-off

Outline

1. Introduction summary
   - Introduction
   - Problematic
   - Constraints and Obstacles
   - General Methodology
2. T-Patterns Algorithm
   - T-Patterns - Aim and Principle
   - Methodology
3. Results and discussion
4. Conclusion and Future work

Results

The direct application of the T-patterns algorithm on the TrainTracer data will lead to the discovery of many spurious couples.
$\Longrightarrow$ a filter was introduced prior to the evaluation of an alarm couple by the T-patterns algorithm. This filter prunes out trains where the frequency of either the A or B alarm is superior to $\overline{x} + 3\sigma$,
$\Longrightarrow$ results are more robust as the filter decreases the number of statistically dependent couples discovered by T-patterns by 20%.

Results

T-Patterns Algorithm

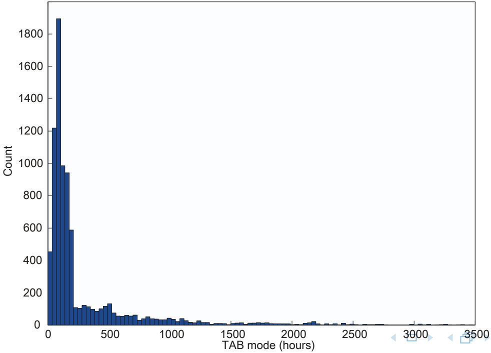
- 8281 L2 patterns discovered, $\Delta_t = 36$ hours
- L2 patterns: patterns length 2 (type $A \rightarrow B$, where B is a target alarm)
- Parameters: significance level α of the Kolmogorov-Smirnov test = 1%, Size S = 50, minimal warning time = 30 minutes

These couples were subject to two major evaluation processes before scrutiniqing their physical significance:

- modeling inter-event times
- calculation of interestingness measures (recall and precision)

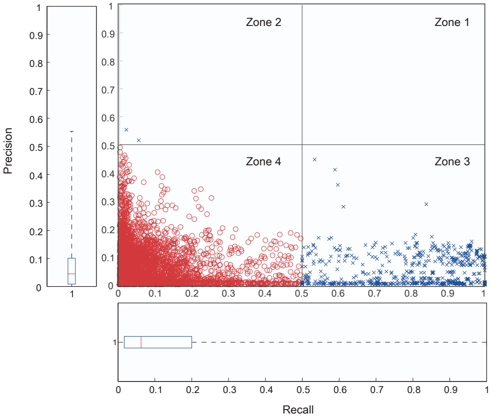
Results - Modeling inter-event times

- mining focused on couples with inter-event times at least equal to 30 minutes
- 4796 discovered couples with a mode value superior to the 30-minute threshold are accepted



Results - Calculation of interestingness measures

The interestingness measures for the 4796 couples were calculated:



Results - physical analysis of rules

- The analysis of the obtained rules has to be both statistical and physical
- All discovered association rules were submitted to railway maintenance experts for futher analysis in order to identify those having a real physical meaning
- Spurious association rules with no technical significance were omitted

Results - example

Consider the following association rule:

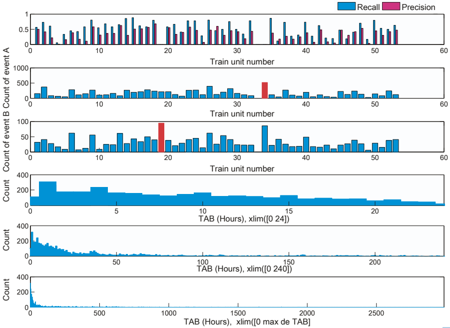
Tilt Authorization and Speed Supervision Not Available (TNA)
$\Longrightarrow$ Train Speed Exceeds 113mph with Tilt Not Available (TOS)

Recall: 59%
Precision: 41%

- Recall value indicates that 59% of the "Train Speed Exceeds 113mph with Tilt Not Available" alarms have been predicted by "Tilt Authorization and Speed Supervision not available" alarms.
- Precision value indicates only 41% of the TNA alarms have lead to a TOS alarms within a time window of [30min , 24h].

Results - example

- Recall and precision values of the association rule per train as well as the distribution of the two alarms of the couple amongst trains are considered.
- The observation of unusual distributions may decrease the chances of a rule to be significant.



Outline

Conclusion, Current and future work

Conclusion

- Few potentially interesting rules were found that can be extended to L3
- The quality of the data mining process is heavily influenced by the rareness of target alarms in addition to the frequent existence of data bursts and flows.

Current and future work

- Cleaning of data bursts and redundancies
- Application of new algorithms efficient in mining rare patterns within bursty environement (ex: Randomization methods)
- Extension of pattern length towards L3

References

S.Pal and P.Mitra
Pattern Recognition Algorithms for Data Mining.
Champman and Hall/CRC, 2004.

Mannila, H. and Tiovonen, H. and Vermako, A.
Discovery of Frequent Episodes in Event Sequence
Data Mining and Knowledge Discovery 1, 259-289, 1997.

Magnusson, M. S.
Discovering hidden time patterns in behavior: T-patterns and their detection
Behav. Res. Methods. Instrum. Comput. 2000, 32, 93-110.

References

AGRAWAL, R. and SRIKANT, R.
Fast Algorithms for Mining Association Rules,
Proceedings of the 20th VLDB Conference Santiago, Chile, 1994. .

AGRAWAL, R. and SRIKANT, R.
Mining Sequential Patterns,
In ICDE, pages 3-14, October 1995.

J.ZAKI, M. and LESH, N. and OGIHARA, M.
Sequence Mining for Plan Failures,
4th International Conf. Knowledge discovery and data mining, 1998.

SALAH, A. A. and PAUWELS, E. and TAVENARD, R. and GEVERS, T.
T-Patterns Revisited: Mining for Temporal Patterns in Sensor Data,
In Sensors, number 8, volume 10, pages 7496–7513, 2010.