Identifying the cause of delays in urban railways using datamining technique

Ami Kono¹, Hideyuki Yabuki¹², Norio Tomii¹

¹ 2-17-1 Tsudanuma Narashino Chiba 275-0016 Japan, {kono@tomii., tomii@}cs.it-chiba.ac.jp, Chiba Institute of Technology
² 3-19-6 Higashi-Ueno, Taito-ku, Tokyo 110-8614 Japan, yabuki@tomii.cs.it-chiba.ac.jp, Tokyo Metro Co., Ltd.

ABSTRACT

It is desired to make timetables more punctual and robust. In order to improve timetables, it is useful to analyze historical train operation records. But because the volume of the data is huge, it is strongly required to establish an approach which can deal with the huge amount of data. In addition, it is preferable that the approach does not count only on physical conditions. This is because although trains run following physical laws in principle, there exist many other factors which give influence to train operation. In addition, we should be aware that it is almost impossible to prepare all the necessary data which might influence train operation in advance. In this paper, we introduce an approach to apply a technique called the association rules to historical train traffic records to identify the causes of delays. Using the association rules make it possible to detect a relationship between items. We consider events such as a delay, an increase of a dwell time and so on as an item and try to find a relationship between them. Based on this approach, we can identify the causes of delays which very often happen and are widely influential. We show the details of the algorithm together with the results of the numerical experiments which we conducted using actual data.

INTRODUCTION

In railway companies, it is strongly desired to make timetables more punctual and robust. One useful approach to make timetables more robust is to detect primary delays which frequently occur and cause large secondary delays. If we succeed to identify such primary delays, we may be able to improve timetables by taking an appropriate delay reduction measures to prevent the primary delays from happening. But it is not so easy to find such primary delays. This is because propagation patterns are very complicated and all the propagations do not occur following physical laws. For example, in railway lines which contain single tracks, if a delay happens when two trains meet, the delay propagates to both directions. In addition, delays occur or propagate by unexpected reasons such as a delay of shunting, a delay caused by congestion on particular days and so on.

In this paper, we propose an approach to apply a data mining technique to historical train operation records in order to detect primary delays which frequently occur and cause large secondary delays (Yabuki 2015). The association rule is a very popular technique in the data mining community and we can find a relationship between items (Berry1997). We consider events such as a delay, an increase of a dwell time and so on as an item and try to find a relationship between them. As the outputs of our algorithm, we
can get association rules such as “if there is a delay larger than one minute at Station X, it is quite probable that the delay propagates to other trains and causes a delay larger than three minutes at Station Y.” We can also get some statistical indices called the certainty and the support and from these indices, we can know how frequently and how certainly the rule holds. Based on this approach, it becomes possible to identify the cause of delays which very often happen and the results suggest what kind of countermeasures to reduce delays are most effective.

Secondary delays usually happen following physical laws such as so called knock on delays. But we should be aware that not all the secondary delays happen following physical laws. In principle, trains run according to physical laws but there exist many other factors which give influence to train operation, such as a lack of human resources. We should also be aware that it is almost impossible to prepare all the necessary data for analysis in advance, partly because we do not know what factors are influential to delays and partly because some data are quite difficult to get. For example, a delay of shunting may cause a delay of other trains but the records of shunting may not be recorded. The approach based on the association rules has a merit that it could detect the cause of such secondary delays or at least it can show “doubtful” situations and the outputs can be used for an intensive analysis by human experts.

Although a lot of papers about increasing robustness using the stored data about train operation are reported (Conte 2007, Cule 2011, Flier 2009, Keckman 2015, Yamamura 2012), research using the association rules is not found in the literature.

When we apply the association rules, however, there exist a couple of things we have to notice. One is that association rules do not assure the existence of causality. The other is a greater number of rules than we can manage tend to be produced. So, we designed our algorithm so that it sorts out only the rules which explain causality and the rules which are really useful.

**AN ALGORITHM TO IDENTIFY THE CAUSE OF DELAYS USING ASSOCIATION RULES**

**Association rules** Let \( I = \{i_1, i_2, \ldots, i_n\} \) be a set of \( n \) binary attributes called items. Let \( T = \{t_1, t_2, \ldots, t_m\} \) be a set of transactions. Each transaction in \( T \) contains a subset of the items in \( I \). An association rule is defined as an implication of the form \( X \Rightarrow Y \) where \( X, Y \subseteq I \) and \( X \cap Y = \emptyset \). The set of items \( X \) and \( Y \) are called the left hand side (LHS) and the right hand side (RHS) of the association rule respectively.

The support \( \text{supp}(X) \) of an item (or a set of items) \( X \) is defined as the proportion of transactions which contain \( X \) in the whole transactions. The confidence of an association rule is defined as \( \text{conf} (X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} \).

**Algorithm to produce association rules**

(1) Basic ideas

We consider the following events as an item: an arrival delay of a train at a station, a departure delay of a train from a station, an increase of a dwell time of a train at a station, an increase of a running time of a train from a station to the next station, an increase of an interval between two consecutive trains at a station.

We try to obtain association rules from these data. We expect to find an association rule such as “if a delay of train X at Station A is larger than \( \alpha \) seconds, then the delay propagates to trains at Station B and cause delays larger than \( \beta \) seconds.” (We assume that Station B is a big and an important station in this railway line).
(2) Details of the algorithm

We show the overall structure of our algorithm in Fig. 1.

**Input:** The input of the algorithm is train traffic records which contain actual and planned arrival / departure times of all the trains for a certain period of time.

**Extract Delays etc.:** From these data, we extract delays larger than threshold, etc. For example, if Train X is delayed larger than the threshold and if Train Y and Train Z are also delayed larger than the threshold on the same day, the delay of Train X, that of Train Y and that of Train Z form one transaction.

**Obtain Association Rules:** From this binary matrix, we extract association rules using the well known “a priori” algorithm (Agrawal 1994) implemented in R (R2015).

**Delete inappropriate Association Rules:** Sometimes too many association rules are output. We select meaningful ones focusing on the value of the confidence and the support. We also eliminate the rules which are against causality. We check the times of occurrence of LHS and RHS and if the time of occurrence of LHS is larger than that of RHS, we delete the rule judging there is no causality between LHS and RHS.

**NUMERICAL EXPERIMENTS**

We have applied our algorithm for the actual historical train traffic records for the railway lines in a suburban area. The target railway lines are shown in Fig. 2.

- We wanted to obtain association rules which have a delay at Station A in LHS and a delay at Station B in RHS.
- The thresholds for the delays in Station A and Station B are both 60 seconds.
- Station B is a big station and we are very much interested in reducing delays of trains which arrive at this station.
- Station A is located in the junction where three lines meet and we are suspicious if delays at this station might cause a delay at Station B.
- It takes around one hour and a half from Station B to Station A.
Tracks in some parts in Fig. 2 are single and tracks in other parts are double.

The number of trains is approximately one hundred.

We used the historical train operation records for a whole day (24 hours) of two months.

We set 0.1 as the threshold of the support, which means association rules whose support is larger than 0.1 are output. This value suggests that the delay mentioned in the rule happens approximately once ten days.

We set 0.8 as the threshold of the confidence, which means association rules whose confidence is larger than 0.8 are output. This value suggests that the rule holds more than eight times out of ten.

Before we eliminate unnecessary rules, 78 association rules were output. Then as the result, we got 37 rules after the elimination of unnecessary rules.

We checked the association rules one by one and learned that many of them suggest an existence of causality and we are still continuing analysis.

CONCLUSION

We have proposed to apply association rules to historical train operation records in order to detect primary delays which frequently occur and cause large secondary delays.

We have implemented the algorithm and applied it to actual data. We proved our approach is promising and we will continue to evaluate the results of our algorithm.

ACKNOWLEDGEMENTS

This research is partly supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C) 15K01199.

REFERENCES

Flier, H., Gelashvili, R. Graffagnino, T., Nunkesser, M. (2009) Mining Railway Delay Dependencies in Large-Scale Real-World Delay Data, Robust and Online Large-Scale Optimization, Lecture Notes in Computer Science Volume 5868