Prospective Study on the Potential of Big Data

Mikio TANAKA
Signalling and Transport Information Technology Division (Former)

The era of collection, analysis and use of vast volumes of data also known as “Big Data” has dawned and has already made contributions to marketing, demand forecasting and creation of new business. Today there is a reflection on how Big Data can play a role in planning transport and administering and maintaining railways, in the hope that it will improve railway management and convenience to rail users. This report looks at the prospects of using big data in the railways and discusses studies on handling large quantities of data.

Keywords: data processing, transportation plan, demand prediction, passenger flow, facility maintenance, reliability

1. Introduction

Over recent years, “Big Data” has become a familiar term. It refers to vast volumes of data which cannot be handled easily using conventional methods; and yet, it is expected that analyzing and using Big Data will improve and engender new business. Though Big Data literally means huge amounts of information, there is no clear definition of Big Data as such and, it often simply refers to a volume of data measured in tens of petabytes that cannot be easily processed with conventional methods (database, processing algorithms, etc.) or hardware (storage). It is said that the upper limit of data size that a fairly large system today can handle, is about several hundred terabytes. As such, research and development has been actively pursued to find methods to handle data the region of 100 terabytes or more, or specifically a method uniquely applied to Big Data. This R&D focuses on big data size, but has already led to better functionality through distribution and allocation of data all over the world allowing parallel data processing and searching. Supercomputers capable of high-speed processing are already making significant contributions to this area of R&D. People may think that Big Data is just buzzword created to put a name on a new fad in another IT boom. However, in reality, this is a well-entrenched trend for the analysis and use of vast volumes of data, which can be converted into an asset through informatization. Recently there have been calls for a "Paradigm Shift: Toward the Fourth Paradigm” [1] (Table 1). As suggested there, what reflects a future “Era of Data Science” is the era of Big Data. This paper sets out to confirm this and give a broader perspective on analysis and use of a large amounts of data untethered by data size, as described above.

2. Background to the active use of big data

Big Data has drawn significant attention, and at the same time is utilized widely and actively, producing remarkable results. The background to this movement is as follows:

- Information systems and communications networks have expanded and are now widely used, resulting in an accumulation and storage of a huge amount of information assets.
- It is now possible to record and store large amounts of information about the activities (movements) of human beings and goods.
- Hardware (processing equipment, memory devices, communications networks, sensors, etc.) has more advanced functions and larger capacity.
- Technology to process and analyze data is making remarkable progress.

Against this background, there is also the change which alters the object of these information systems, converting human beings and goods “from a mass (group) to individuals”. In the past, forecasting consumer demand involved the common practice of focusing on the properties of a group of human beings. For goods, the general approach was to control and monitor goods on the basis of the average state of each object. Over recent years however the growing trend is to use methods where the target is more sophisticated: businesses for example measure (for marketing, etc,) properties and preferences of individuals, optimizing the tracking and control of a specific good bases on its individual characteristics.

As a results, society as a whole today enjoys greater efficiency and better service, and manages to save resources, among other advantages.

Whether dealing with human beings or goods, what is needed is information about the properties of each. Previously it was difficult to capture and comprehend this kind of information in a precise manner and in large quantities, so that the best practice in this regard was to consider the object “en masse”. Today however we have entered an era where a huge amount of information can be obtained in the form of “Big Data”, through collection of information from the Internet and devices (such as location information of cellular phones and information of automatic ticket gates).

Another important factor making up the background is the spread of electronic or computer-based monitoring and control. For example, railway switches have been maintained manually and uniformly based on common knowledge, such as a degree of switch fixing, a state of locking (lock), a current value of a motor. However if individual properties and conditions of each facility were available,
Empirical Data Science
Scientific research on the premise that there are a
Computing Science
Theoretical Remarks

needs simultaneous pursuit of practical level. In order to address the difficult challenge of
tem), etc., though they are still a long way from reaching a
users systems to this end, including into
weakness, compared to individual transport systems, such
changes in those requirement, which for the moment are a
convert its system into one which is better suited to ca
transportation. Therefore the railway industry needs to
shrink over the future lessening the need for capacity in
mobility. However, it appears that the population of Japan
transit

3. Study on the potential development of Big Data in railways

Various studies using and applying Big Data can be
conducted in a range of railway transportation fields. Here
are some examples.

3.1 Modeling of user behavior and improvement of transportation plan

The main merit of the railways is that it allows “mass transit”, which made an invaluable contribution to prog-
ress during Industrial Revolution when society lacked mobility. However, it appears that the population of Japan
shrink over the future lessening the need for capacity in transportation. Therefore the railway industry needs to
convert its system into one which is better suited to ca-
ter to “individual needs”, and flexible enough to follow changes in those requirement, which for the moment are a
weakness, compared to individual transport systems, such
as the road.

Various studies and simulations have been conducted to this end, including into “demand-centered transportation systems” (a dynamically flexible transportation plan following users’ needs) (Fig. 1), “vague reservation systems” (alloca-
tion of passengers to trains using a flexible reservation sys-
tem), etc., though they are still a long way from reaching a
practical level. In order to address the difficult challenge of
simultaneous pursuit of “ability to correspond to individual
needs” and “transportation efficiency,” the railway sector

needs to work on more sophisticated and detailed capture
and forecasting of individual travel needs, and consolidate
these findings skillfully to continue operating in a manner
which is as close to mass routine transportation as possible.
For this purpose, railway must develop elaborate analysis
and prediction methods to forecast user trends.

In most cases, user trends are obtained through selected
action models based on microeconomic utility theory. This
model assumes that a share of probability of each option
can be obtained from a utility value that is calculated using
various attribute values related to a subject, an object, a situ-
ation, an environment, etc. when the subject (a purchaser,
a user, etc.) selects the object (goods purchased, a transport
mode, etc.). Now, suppose this is applied to an example for
predicting what transportation mode and what train a user
will use. Here, a utility value and a selection probability
for each option are calculated using attributes of a subject
(age, etc.), an object (price, time, etc.), and an environment
(purpose of travel, weather, etc.). Based on the results of
this calculation, it is possible to forecast passenger behavior.
But as the number of combinations of the said parameters
are almost endless, it is extremely difficult to identify pa-
rameters which offer the necessary level of reliability. If a
significant amount of Big Data is obtained however, a data
size effective for combinations of each type of attributes can
be obtained. Hence, it is expected that an elaborate pre-
diction can be made by attaching a value to each attribute
which is then introduced into a formula for a behavior model
and accumulating forecasts of individual behavior.

The above description covers action prediction and
transportation plans for ordinary operating conditions,
however, it is important to extend the method so that it can
be applied to disrupted situations, or to cases of large scale
disasters, similar to those witnessed over recent years, and
cases where there is a failure in train operations. If select-
ed actions can be analyzed on the basis of Big Data, (can-
cellation of travel, waiting for resumption, taking a detour,
etc.) collected from users during periods when services are
disrupted, it would be possible to improve remedial mea-
sures, to help stranded passengers, or offering support in
the case of exceptional situations.

3.2 Detection of change in system balance

Our environment is made up of a mosaic of factors.
This can be rephrased as a system with N variables. Fol-

| Paradigm      | Remarks                                                                 |
|---------------|-------------------------------------------------------------------------|
| 1 Empirical Science | Processing empirically data based on information obtained through experiences and observations. |
| 2 Theoretical Science | Exploring logic and laws existent behind experiences by analyzing them. |
| 3 Computing Science | Addressing a problem for which an analytical solution cannot be obtained by applying computer power to a numerical simulation method. |
| 4 Data Science  | Scientific research on the premise that there are a huge amount of data and strong computing power that is ubiquitous and available. |

change in transportation needs (Big Data)

Fig. 1 Concept of dynamic transportation plan (demand-oriented)
lowing on, it can be imagined that the state at each point can be located in an N-dimension space, which moves on in space with the flow of time.

Communications traffic, computer access, and situations related to human movement and location could be interpreted as an indication of system’s condition seen in multiple dimension. Recently, more and more studies have been aimed at detecting changes in these conditions. For example, researchers are trying to obtain stochastically the possibility of a cyber or terrorist attack from the volume and content of communications traffic, or detect or forecast a change in economic conditions or buying behavior from the state of Internet use and changes in retrieved keywords. One noteworthy effort in the railway sector is to detect passenger flows. This is useful for the preparation of long-term transportation plans or facilities plans, demand development and marketing, etc. However, it is difficult to decide how to detect a change from a wider perspective. One of the possible methods is to analyze data collected at automatic ticket gates and mutual relationships between many variables such as users at each station/OD (arrival and departure) to model results of the analyses based on multiple regression. The concept of this modeling is shown in Fig. 2. In normal circumstances a difference between regression and actual status can be within a certain range, but when the difference exceeds the range, this indicates a change in a condition. This method is used to estimate probabilities and quantitative differences related to the change. There can be a huge number of explanatory variables which influence user trends, but here simplification of multiple regression is done through use of variables concerning each station/OD as secondary explanatory variables. Figure 3 indicates an example of using this method for quantitatively evaluating an increase in the number of passengers as the result of an event.

A similar approach can be applied to equipment monitoring. Track circuits are important since they are the basis for train detection and signal control. Monitoring track circuit properties is therefore important to ensure the safety and reliability of transportation. But such properties change regularly according to various circumstances and installation requirements, making it difficult to judge soundness. So, one possible method is to grasp what the status is normally by analyzing and comparing the characteristic values of each track circuit based on past data and then estimating a degree of soundness with reference to a magnitude of deviation from the normal status. For this purpose, it is necessary to implement grouping and correlation analysis of track circuit properties based on data.

Figure 4 presents an example (a dendrogram) of a cluster analysis of track circuit facilities in a certain area, and Fig. 5 shows an example of identifying a track circuit featured by similar characteristics by use of cluster analysis and correlation analysis.

Figure 6 shows an example of detecting property changes where multi regression values, including property...
values of other facilities and environmental variables as explanatory variables, and actual values are presented together. Correlation between multi regression values and actual values is strong. So, research is now driven to work out a method of early detection of a change in soundness through checking their degree of concordance.

3.3 Property analysis of facilities

Safety and reliability of railway transport is supported by various facilities, and railway operators confirm their soundness through maintenance inspections and monitoring. Railway companies traditionally adopt the basic idea that a maintenance method (inspection locations, a method, a threshold, etc.), an inspection period (a cycle), etc. should be specified in a single uniform way for each type of facility. However, current changes in the state of equipment depend on the environment in which the equipment is used or installed, lapse of time, past history, and other factors, so that it is impossible to judge change according to a fixed set of standards. Persons in charge of field maintenance understand this very well, but because the current situation does not allow for effective collection and analysis of data in a large scale, there is no progress in sharing and systematization of knowledge and transmission of skills and technologies.

In recent years, monitoring devices have evolved significantly with growing automatic data collection and storage on a large scale. This paper introduces an analysis approach geared to leverage such data (note that its result is from an example of using mere samples, so that it never leads to a general conclusion).

This paper presents the example an analysis of a pull-off arm, part which is often used for overhead contact lines. Table 2 shows, for each type of setup environment (a general environment/a bridge), a ratio (a ratio of measures taken) of cases where measures had to be taken, in relation to inspections conducted tens of thousands of times. This results confirmed whether or not “the ratio of measures taken depends on a type of environment” by applying statistical analysis (a chi-square test on a contingency table). Leading to the following fact: at a level of significance of 0.05, the “difference is significant” or, in other words, a "probability of occurrence of a problem is high in bridge sections.”

Next, results on switches are shown in Table 3. Like in the above case, as a result of a statistical testing, a trailing switch has a higher occurrence ratio of problems than a facing switch. In addition, a similar analyses were conducted for each type of attribute, resulting in the following findings:

- Attributes for which a difference between ratios of measures taken became statistically significant:
  - Trailing/facing, types (NS/TS, flexible/articulated), a toe/a nose
  - Attributes for which a difference between ratios of measures taken did not become statistically significant:
    - Simple/symmetrical, opening to the right/opening to the left, a main track/a side track, a switch number, a lapse of time

Moreover, it is interesting to see whether these results match real findings and intuitions of persons in charge of maintenance. In the case of the pull-off arm, they matched each almost perfectly (generally speaking severe conditions were observed on bridge due to exposure to wind and rain). However, although most people were of the opinion that a facing switch would have more problems (as wheels run...
into the toe of a point) results demonstrated the opposite. Figure 7 shows an example of a survival analysis for each type of installation environment of facilities in order to study equipment life span and replacement cycles. The purpose here was to analyze changes in reliability and survival rate (soundness) with reference to data of states of occurrence of problems and then optimize the replacement cycle according to each type of environment.

If Big Data can be employed and actively utilized for analysis by promoting studies, like the ones stated above, systematically and extensively on a large scale, it will be possible to find more suitable and effective maintenance methods (for example, method and cycles adapted to attributes) corresponding to individual properties of a facility. Moreover, valuable knowledge included in qualitative findings and experience of experienced workers in charge of maintenance will be confirmed (or negated), thus contributing to discovery of a new facts and technological progress.

4. Conclusions

Figure 8 presents possible purposes for which Big Data can be employed while preparing common databases in each field of railway. Though the railway business tends to be complicated and technical and creating a vertical division between specialized disciplines, railway operators should share precious Big Data in each field to make the best use of it.

The content of this paper relates only related to technical aspects, whereas there are many other different aspects and challenges in the railways. For example, to what extent is it acceptable to use personal information from a viewpoint of privacy protection; whether it is welcomed or not in terms of convenience to adopt a scheme where "the operational schedule dynamically evolves following a need," and whether maintenance corresponding to properties of individual facilities can be adapted to the field or be compliant with institutional maintenance standards. The author will continue to gather more views and opinions from all quarters and plan a study to see how Big Data can be leveraged in the future.

References

[1] Microsoft Research : The Fourth Paradigm : Data-Intensive Scientific Discovery, http : //research. Microsoft. com / en - us / Collaboration / Fourth-paradigm / .