High-Speed Probe Information Collection/Analysis Using Data Stream Processing Platform

NAKAMURA Nobutatsu, KIDA Koji, FUJIYAMA Kenichiro, IMAI Teruyuki

Abstract

Expectations are increasing for the advanced Telematics services that collect the massive data uploaded from automobiles (mobile objects) to a center in order to utilize the results of analyses. The implementation of these services is dependent on an information communication technology capable of collecting massive data and analyzing massive data sources at high speeds. This paper is intended to introduce the data stream technology that is used in such cases for the collection and analysis of massive data. In addition, the results of the prototyping and technological evaluation of a system prototype that can process the position and velocity data from 50,000 vehicles and provide traffic jam information in real time are also discussed.

Keywords

data stream, large-scale data collection/analysis, Telematics, traffic jam

1. Introduction

The advancement of information communication technology is bringing about a ubiquitous information society in which people can enjoy advanced information services at anytime and anywhere. In the field of automobiles, Telematics services making use of various vehicle sensor data and driver data are expanding. For example, uploading data on the positions and velocities of vehicles to a center and analyzing them at the center makes it possible to monitor the status of the traffic congestion of roads. Services such as route guidance according to the traffic congestion status may not only make driving comfortable but can also offer significant social advantages, including energy saving and environmental load reductions.

Information communication processing for the collection and analysis of data from each vehicle is indispensable for the implementation of such services. However, when the number of vehicles is large and the data to be processed is greatly increased, very high costs are necessary in order to build an effective system. At NEC, we are conducting R&D into a platform for the efficient collection of massive data and analysis of massive data sources. This platform will make it possible to construct large-scale Telematics services at low costs.

This paper introduces the technology designed to support a data stream processing platform for use in large-scale data collection/analysis and a traffic jam information monitoring system that we have prototyped using the platform.

2. Data Stream Processing Platform

Even if the size of each item of data such as log data or presence information is small (a few hundreds of bytes), the center needs to process hundreds of thousands of items or megabytes of data per second if the data is collected from millions of locations. In the case of ordinary processing, a large amount of data is stored in a database and is processed in batches. The processing of such a large amount of data requires a high-performance computer and the construction of a suitable central processing system, which entails high costs. With regard to this constraint, if the data to be processed is subsequently increased, the batch processing may not be completed within the expected time or data overflow may cause loss of data.

To deal with this issue, NEC is conducting R&D into a data stream processing technology that sequentially executes pipeline type distributed processing of data without storing it in a database (**Fig. 1**). This technology executes sampling, filtering, cleansing and statistic computations of data during the collection process in order to reduce the load of the center server.

Fig. 2 shows the architecture of the data stream processing platform. The unit of processing is referred to as a node and several nodes are combined to form a single application. Each node is a thread that acts asynchronously and independently but a module called the node manager controls all of the

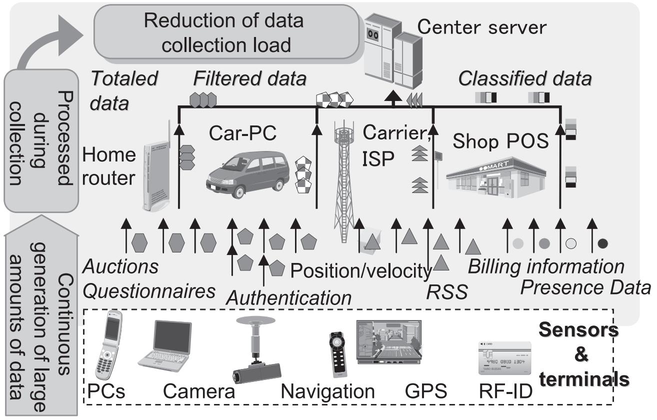

Fig. 1 Data stream processing.

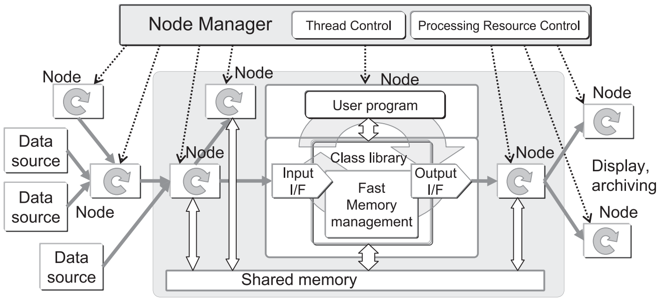

Fig. 2 Data stream processing platform architecture.

nodes by starting/stopping processing or controlling the processing speed.

The nodes can be located either in a single machine or in multiple machines. Communications between nodes in different machines are performed by exchanging the processing data via the transmission/reception nodes. Whereas nodes in the same machine exchange data by memory management, using the shared memory, which is a function unique to this platform. The memory management is optimized for sequential processing of small-sized data items and is capable of high-speed data exchange and cue management.

For the present, the platform is provided as C++ class libraries. The developer uses the classes in implementing a program in each node and specifies the processing triggering conditions and then describes data processing program for each node. When specifying the node connection, it is also possible to specify more than one connection destination. This option enables various implementations, such as the provision of more than one service or parallel data archiving by means of branching as well as integrated analysis via centralization.

3. Prototyping of a Traffic Congestion Status Monitoring System

Using the data stream processing platform described above, we have prototyped a traffic congestion status monitoring system that collects the position and velocity information of massive vehicles and moving objects and is able to provide accurate and very fresh traffic congestion status information.

The most of traditional systems for identifying the traffic congestion status have installed sensors on the road and processed the information collected from them. Such a system can collect accurate information on target roads once the system has been installed. However, the method is also accompanied with some disadvantages such as its high cost, which makes it hard to install sensors on all of the roads within the targeted area. Meanwhile, as GPS (Global Positioning System) and wireless communications (cellular phones) are now widely disseminated, the infrastructures for the easy collection of the position/velocity data of individual vehicles is almost ready. The system proposed here has been designed in order to collect the position/velocity data from each vehicle or mobile object.

In order to generate the requisite traffic congestion information the position/velocity data of each vehicle is processed as follows.

(1)Collection of vehicle information

(2)Matching of each vehicle on a specific road of the map data using the mosaic matching method

(3)Data sampling using the calculation cost variable approximation method

(4)Computation of the degree of traffic congestion for each road section

With the mosaic matching method, the road network is divided into the grids formed by the latitude and longitude and the vehicle position (GPS) data is matched with the gridded road data at a high speed. It can thus identify the driving road section, crossing of crossroads and the lane (direction) of each vehicle.

With the calculation cost variable approximation method, the sampling rate of the data collected to be analyzed is varied according to the analysis speed and accuracy of the data. It is naturally desirable not to perform sampling in order to maximize the acquired information. However, since the data

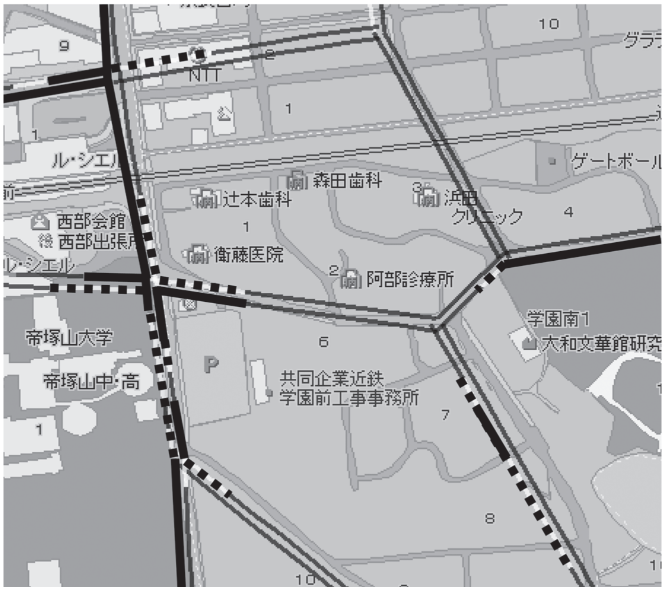

Fig. 3 Example of traffic jam status monitoring system display.

from vehicles using the same road at a specific moment are similar, the effect on the analysis is minimal even when certain data items are abandoned in sampling. Based on this assumption, the method groups roads according to their geological connection relationships and uses a lower data sampling rate for the road groups that can offer an adequate amount of data for analysis.

After prototyping a large-scale traffic jam status monitoring system using the above techniques, we simulated a virtual traffic flow with a road network composed of about 2,000 roads in Nara Prefecture, Japan, using the commercially available traffic simulation software NETSIM. Following this we simulated the transmission of the position, velocity and direction of each vehicle to an analysis server, which computed the degree of congestion of each road section based on the map matching. **Fig. 3** shows an example (enlarged view) of the display of the prototyped system. This display shows the average velocity of vehicles with values and the road color. The thick lines indicate traffic jams, the thick dotted lines indicate relatively low velocity traffic flows and the thin lines indicate smooth traffic flows.

4. Performance Evaluation

We evaluated the performance of the prototyped system by

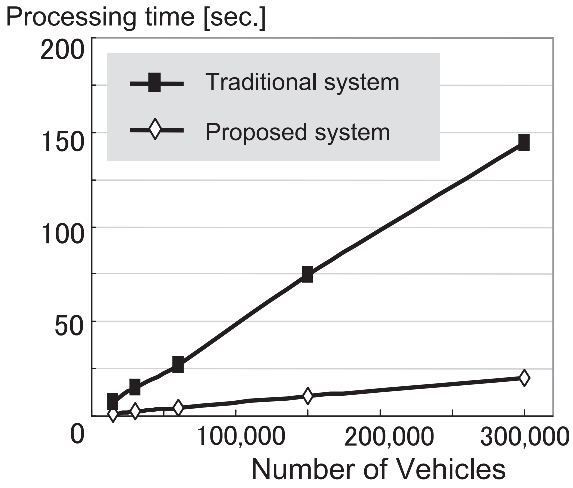

Fig. 4 Change in processing time when the number of vehicles is varied.

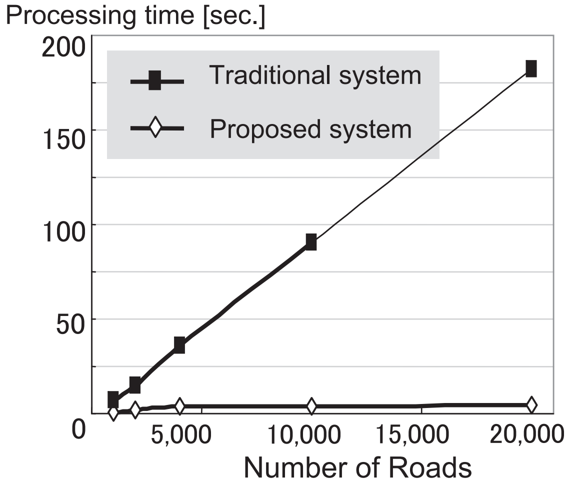

Fig. 5 Change in processing time when the number of road sections is varied.

using a single PC (CPU: Pentium 4, 3.0GHz, RAM: 512M Bytes. OS: Windows 2003 Server) for the following two cases; 1) fixing the number of roads to 2,000 and varying the number of vehicles; 2) fixing the number of vehicles to 30,000 and varying the number of roads. The results are shown in **Fig. 4** and **Fig. 5** .

Since the traditional system is based on simple collation processing without using the mosaic matching method, the processing time increases linearly when the number of either vehicles or roads increases. Since expansion of the processing target area generally increases the numbers of both vehicles and roads linearly, the processing time of the traditional system with respect to the target area is increased in the order of $O(n^2)$. On the other hand, the processing time of the proposed system increases linearly with respect to the increase in the number of vehicles, but it varies little with respect to that of the number of roads. Additionally, the change in the processing time with respect to the number of vehicles is only about 15% of the traditional system. Therefore, the order of increase in the processing time of the proposed system with respect to the target area can be regarded to be O(n) when the grid size is reduced to a level at which the number of roads is almost irrelevant. This means higher scalability compared to the traditional system.

5. Conclusion

The most well known existing research related to the data stream processing technology dealt with in this paper includes the Aurora Project by MIT and Brown University and the STREAM Project of the Stanford University. These projects are basically theoretical and are focused on research into the query language, processing architecture and analysis algorithm. In the products field, Apama is well known for its application in the financial world but its processing performance in large-scale applications is not yet fully understood. The products associated with log collection are active in relation to the internal governance issue. However, these products specialize in collection and do not consider the analysis and real-time utilization of the collected data.

We are conducting R&D into the data stream processing platform aiming at a technology for collecting and analyzing continuously generated massive data that enables advantages in computer resource efficiency and high throughput. Up to the present, we have simulated and demonstrated the performance of this platform by performing a map matching of 50,000 items of vehicle data per second. In the future, we aim to improve the functions and performance of the platform and also to ensure its practicality by applying relevant demonstration experiments. We intend that our efforts will contribute the implementation of advanced ITS services with a capability for performing the collection and analysis of massive data.

*Pentium is a trademark or registered trademark of the Intel Corporation.

*Windows is a registered trademark of the Microsoft Corporation in the United States and other countries.

*Apama is a registered trademark of the Sonic Software Corporation.

*Netsim is a product of the Phoenix Research Co., Ltd.

*The map data has been compiled using MapDK, a product of the Increment P. Corporation.

*As the products introduced in this paper are mainly sold for the domestic market, some figures feature explanations by the Japanese Language.

Authors' Profiles

NAKAMURA Nobutatsu
Principal Researcher,
Service Platforms Research Laboratories,
NEC Corporation

KIDA Koji
Assistant Manager,
Service Platforms Research Laboratories,
NEC Corporation

FUJIYAMA Kenichiro
Researcher,
Service Platforms Research Laboratories,
NEC Corporation

IMAI Teruyuki
Researchers,
Service Platforms Research Laboratories,
NEC Corporation