外 国 語 要 旨

学位論文題目：Optimization Method for Data Access Processing in Real-time
氏名：　Miki Enoki

Recently many amount of data such as location, weather, social, and censor data is being produced continuously from various devices. It is essential to analyze such data in real-time. For example, monitoring censor data of a particular machine can detect or predict any fault as soon as possible. To deliver such analysis environment, real-time processing infrastructure is needed.

The infrastructure system handling big data consists of two parts. One is focusing on "Velocity" for real-time data processing. It enables real-time analysis and data calculation against continuous data. Another one if focusing on "Volume" for massive data analysis such as OLAP (online analytical processing). Large volume of data is stored in HDD based storage, so it is important to provide various query results as soon as possible.

In this paper, I propose a data processing system enables both real-time and off-line data analysis with big data. Especially, I investigate performance of data access processing and resolve some bottleneck points for faster data access. I use social data to evaluate the system. A social media service such as Twitter characterized by frequent message posting and transient topics. Over four hundred million messages are posted around the world in a day. There is a need to respond quickly on the basis of an analysis of user behaviors and to quickly identify trending messages because much of the information shared through social media quickly loses its impact. In the current Twitter system, tweets can be searched for by using keywords related to a company and monitored manually. I develop a system for analyzing the diffusion of information through retweeted tweets. The retweeting of a tweet message by many users indicates that they find the content interesting and/or entertaining.

For real-time data processing, I propose custom time window model. Streaming data is usually divided into segments called windows. However, static time window fragments diffusion data. Therefore, we retain diffusion data in the data store for only as long as it is being retweeted frequently by using retweet propagation model. When the data becomes stale, they are removed from the data store. To control capacity of incoming tweet data against bursting, I propose a filtering method using extra attribute

information of tweets. I evaluate the effectiveness of maintenance and filtering methods for data control with Twitter data. There was a trade-off between query performance and accuracy of analysis results.

For massive data analysis, I create data access layer between application and database with OpenJPA which is an implementation of the Java persistence API (JPA). It has a caching layer for databases queries to share cached objects among multiple client sessions. However the performance is limited when an application includes write transactions, because the default OpenJPA cache invalidation mechanism is course-grained and this results in a low cache hit rate. I implement two kinds of finer-grained invalidation mechanisms by using query dependency analysis and invalidation index. In experiments with TPC-W benchmark, I show the OpenJPA with the finer-grained invalidation mechanisms outperform the current OpenJPA. To access diffusion data more quickly, I also develop a matrix index containing some of the data needed for diffusion analysis in the data access layer. This would be a large and sparse matrix, so data compression techniques are applied. I evaluate the performance and overhead when compressing and reconstructing the matrix.