Makoto Yui

Senior Researcher at AIST

yuin405@gmail.com

Experience

Senior Researcher at AIST

April 2010 - Present (4 years 10 months)

Working on distributed and parallel data processing and large-scale machine learning at the data science research group.

Visiting Researcher at University of Edinburgh

September 2011 - November 2011 (3 months)

Worked with Paolo Basala and Prof. Malcolm Atkinson at Data Intensive Research (DIR) group. Designed a distributed streaming data processing system on EDIM1 data-intensive machine (an energy-efficient PC cluster) for scientific workflows.

Visiting Postdoc at CWI

October 2009 - March 2010 (6 months)

Worked with Peter Boncz and Prof. Martin Kersten at INS1 database research group. Designed and implemented a parallel database system on the top of shared-nothing MonetDB servers.

Visiting Researcher at Waseda University

April 2009 - March 2010 (1 year)

JSPS Research Fellow (PD) at Japan Society for the Promotion of Science

April 2009 - March 2010 (1 year)

JSPS Research Fellow (DC2) at Japan Society for the Promotion of Science

April 2008 - March 2009 (1 year)

System Engineer at NEC Infomatic Systems, Ltd

April 2004 - March 2006 (2 years)

Designed and implemented RM4GS (Reliable Messaging for Grid Services) which provides reliable messaging facilities for Web Services.

Honors and Awards

Bossie Awards 2014: The best open source big data tools (for Hivemall)

InfoWorld, Inc.

September 2014

The InfoWorld Bossie Awards are selected by InfoWorld Test Center editors and reviewers, and judging is based upon software products and companies that demonstrate innovation, functionality, ease of use, implementation, and a proven track record in serving the needs of today's businesses. A full list of winners is available here: http://www.infoworld.com/article/2688074/big-data/big-data-164727-bossie-awards-2014-the-best-open-source-big-data-tools.html

Projects

Hivemall: Hive scalable machine learning library

September 2013 to Present

Members:Makoto Yui

Hivemall is a scalable machine learning library running on Hive/Hadoop, licensed under the LGPL 2.1. Hivemall is designed to be scalable to the number of training instances as well as the number of training features. Hivemall got the InfoWorld's Bossie (The Best of Open Source Software) Awards 2014: The best open source big data tools.

Languages

English (Professional working proficiency)

Japanese (Native or bilingual proficiency)

Publications

Hivemall: Scalable Machine Learning Library for Apache Hive

Hadoop summit 2014, San Jose June 3, 2014

Authors: Makoto Yui

This talk will introduce Hivemall, a new open-source machine learning library for Apache Hive. Hivemall provides a number of machine learning functionalities across classification, regression, ensemble learning, and feature engineering through UDFs/UDAFs/UDTFs of Hive and is very easy to use as every machine learning step is done within HiveQL. We present a series of experiments using the KDD cup 2012, click-through rate prediction task and compare Hivemall to state-of-the-art scalable machine learning frameworks, including Vowpal Wabbit and Mahout. We consider that this talk is particularly interesting and relevant to people already familiar with Hive and working on big data analytics.

Hivemall: Hive scalable machine learning library

NIPS 2013 Workshop on Machine Learning Open Source Software: Towards Open Workflows December 10, 2013

Authors: Makoto Yui

We gave a Hivemall demo that shows provisioning a machine learning service (i.e., Hivemall) on Amazon Elastic MapReduce.

A Database-Hadoop Hybrid Approach to Scalable Machine Learning

Proc. IEEE 2nd International Congress on Big Data June 30, 2013

Authors: Makoto Yui, Isao KOJIMA

There are two popular schools of thought for performing large-scale machine learning that does not fit into memory. One is to run machine learning within a relational database management system, and the other is to push analytical functions into MapReduce. As each approach has its own set of pros and cons, we propose a database-Hadoop hybrid approach to scalable machine learning where batch-learning is performed on the Hadoop platform, while incremental-learning is performed on PostgreSQL. We propose a purely relational approach that removes the scalability limitation of previous approaches based on user-defined aggregates and also discuss issues and resolutions in applying the proposed approach to Hadoop/ Hive. Experimental evaluations of classification performance and training speed were conducted using a commercial advertisement dataset provided in the KDD Cup 2012, Track 2. The experimental results show that our scheme has competitive classification performance and superior training speed compared with state-of-the-art scalable machine learning frameworks; 5 and 7.65 times faster than Vowpal Wabbit and Bismarck, respectively, for a regression task.

Nb-GCLOCK: A Non-blocking Buffer Management based on the Generalized CLOCK

Proc. ICDE March 2010

Authors: Makoto Yui, Jun Miyazaki, Shunsuke Uemura, Hayato Yamana

In this paper, we propose a non-blocking buffer management scheme based on a lock-free variant of the GCLOCK page replacement algorithm. Concurrent access to the buffer management module is a major factor that prevents database scalability to processors. Therefore, we propose a non-blocking scheme for buffer#x operations that #x buffer frames for requested pages without locks by combining Nb-GCLOCK and a non-blocking hash table. Our experimental results revealed that our scheme can obtain nearly linear scalability to processors up to 64 processors, although the existing locking-based schemes do not scale beyond 16 processors.

Skills & Expertise

Databases
Distributed Systems
Hadoop
Machine Learning
Data Mining
Research
Computer Science
Artificial Intelligence
Data Warehousing
Parallel Computing
Database Development
Hive
Data Analytics
Adtech

Education

Nara Institute of Science and Technology

Ph.D, Computer Science, 2006 - 2009

Nara Institute of Science and Technology

M.E. Computer Science, Computer Science, 2005 - 2006

Shibaura Institute of Technology, Japan

Bachelor's Degree, Computer Science, 1999 - 2003

Honors and Awards

Best Student Award, Nara Institute of Science and Technology, Japan, 2009/3.

Interests

Parallel/Distributed Database Systems, Distributed Systems, Large-Scale Machine Learning, MapReduce and Hadoop.

Makoto Yui

Senior Researcher at AIST

yuin405@gmail.com

Contact Makoto on LinkedIn