SaberLDA:
Sparsity-Aware Learning of Topic Models on GPUs

Wenguang Chen

Tsinghua University

Wednesday, February 28, 2018 at 3:00 pm
Room: Andreas-Pfitzmann-Bau, Nöthnitzer Str. 46, APB 3105

Abstract:

Latent Dirichlet Allocation (LDA) is a popular tool for analyzing discrete count data such as text and images. Applications require LDA to handle both large datasets and a large number of topics. Though distributed CPU systems have been used, GPU-based systems have emerged as a promising alternative because of the high computational power and memory bandwidth of GPUs. However, existing GPU-based LDA systems cannot support a large number of topics because they use algorithms on dense data structures whose time and space complexity is linear to the number of topics.

In this talk, I introduce SaberLDA, a GPU-based LDA system that implements a sparsity-aware algorithm to achieve sublinear time complexity and scales well to learn a large number of topics. To address the challenges introduced by sparsity, we propose a novel data layout, a new warp-based sampling kernel, and an efficient sparse count matrix updating algorithm that improves locality, makes efficient utilization of GPU warps, and reduces memory consumption. Experiments show that SaberLDA can learn from billions-token-scale data with up to 10,000 topics, which is almost two orders of magnitude larger than that of the previous GPU-based systems. With a single GPU card, SaberLDA is able to learn 10,000 topics from a dataset of billions of tokens in a few hours, which is only achievable with clusters with tens of machines before.

Bio:

Wenguang Chen is a professor in Department of Computer Science and Technology, Tsinghua University. His research interest is in parallel and distributed systems and programming systems. He received the Bachelor's and Ph.D. degrees in computer science from Tsinghua University in 1995 and 2000 respectively. Before joining Tsinghua in 2003, he was the CTO of Opportunity International Inc. He was appointed as the associate head of Department of Computer Science and Technology from 2007 to 2014.

He has published over 50 papers in international conferences and journals like OSDI, ASPLOS, PLDI, EuroSys, USENIX ATC, OOPSLA, ICSE, PPoPP and SC. He is a distinguished member and distinguished speaker of CCF (China Computer Foundation). He is an ACM member and co-chair of ACM China Council. He serves in the program committee of many conferences, such as SOSP, PLDI, PPoPP, SC, and ASPLOS.