MPI/OpenMP hybrid parallel inference for Latent Dirichlet Allocation

摘要

用于 LDA 参数估计的变分贝叶斯推断和 collaopsed Gibbs sampling 在处理大规模数据时计算开销非常大。本文利用并行计算技术提高 Gibbs sampling 推断的效率。我们使用一种近年来被广泛使用的共享内存集群 (SMP 集群)。在 LDA 并行推理的前期工作中，MPI 和 OpenMP 都以被使用过。另一方面，对于 SMP 集群而言，更适合采用在 SMP 节点和循环指令之间通过消息传递来实现通信的混合并行化，以此在每个 SMP 节点之间实现并行化。本文设计了一种用于 LDA 的 MPI/OpenMP 混合并行推断方法。

1 Introduction

针对大量文档集的分析中，主题模型方法是一种极为成功的学习算法。主题模型基于的思想是，每篇文档都是从单词分布的“混合”中生成的，每一个这样的“混合”称为“主题”。...

2 Related Work

2.1 LDA

2.2 Fast inference methods for LDA

collapsed Gibbs sampling 的计算复杂性来自于文档集中的主题数量和词汇表大小的乘积。针对此问题的改进工作有很多：

- ...
- ...