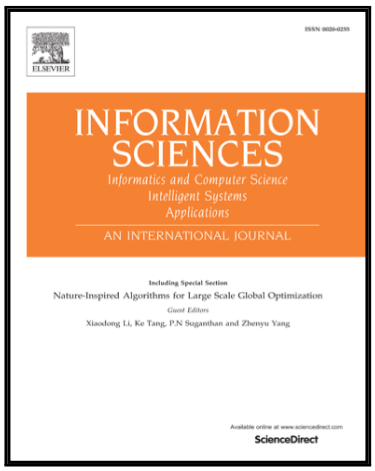

Towards Big Topic Modeling

JianFeng Yan, Jia Zeng, Zhi-Qiang Liu, Lu Yang, Yang Gao

PII: S0020-0255(16)32034-5
DOI: 10.1016/j.ins.2016.12.014
Reference: INS 12649

To appear in: *Information Sciences*

Received date: 31 December 2015
Revised date: 2 November 2016
Accepted date: 13 December 2016

Please cite this article as: JianFeng Yan, Jia Zeng, Zhi-Qiang Liu, Lu Yang, Yang Gao, Towards Big Topic Modeling, *Information Sciences* (2016), doi: 10.1016/j.ins.2016.12.014

Towards Big Topic Modeling

JianFeng Yan[a], Jia Zeng[a,c,*], Zhi-Qiang Liu[b], Lu Yang[a], Yang Gao[a]

[a]*School of Computer Science and Technology, Soochow University, Suzhou 215006, China*
[b]*School of Creative Media, City University of Hong Kong, Tat Chee Ave. 83, Kowloon Tong, Hong Kong, China*
[c]*Huawei Noah's Ark Lab, Hong Kong, China*

Abstract

To solve the big topic modeling problem, we need to reduce both the time and space complexities of batch latent Dirichlet allocation (LDA) algorithms. Although parallel LDA algorithms on multi-processor architectures have low time and space complexities, their communication costs among processors often scale linearly with the vocabulary size and the number of topics, leading to a serious scalability problem. To reduce the communication complexity among processors to achieve improved scalability, we propose a novel communication-efficient parallel topic modeling architecture based on a power law, which consumes orders of magnitude less communication time when the number of topics is large. We combine the proposed communication-efficient parallel architecture with the online belief propagation (OBP) algorithm, referred to as POBP, for big topic modeling tasks. Extensive empirical results confirm that POBP has the following advantages for solving the big topic modeling problem when compared with recent state-of-the-art parallel LDA algorithms on multi-processor architectures: 1) high accuracy, 2) high communication efficiency, 3) high speed, and 4) constant memory usage.

Keywords: Big topic modeling, latent Dirichlet allocation, communication complexity, multi-processor architecture, online belief propagation, power law

*Corresponding author
Email addresses: yanjf@suda.edu.cn (JianFeng Yan), j.zeng@ieee.org (Jia Zeng), ZQ.LIU@cityu.edu.hk (Zhi-Qiang Liu), yanglu@suda.edu.cn (Lu Yang), gaoyang.suda@gmail.com (Yang Gao)