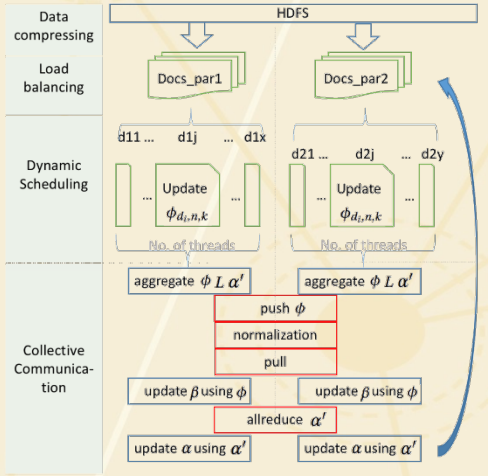
Distributed LDA on Harp

Ethan Li, Rohit Patil

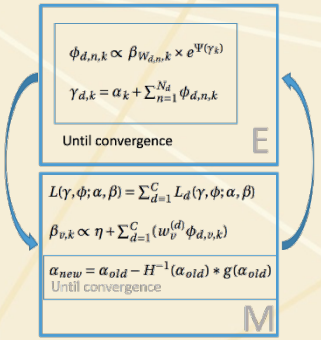
ABSTRACT

Harp LDA is a distributed variational bayes inference (VB) algorithm for LDA model which would be able to model a large and continuously expanding dataset using Harp collective communication library. We demonstrate how variational bayes inference converges within Map-Collective jobs provided by Harp. We provide results of the experiments conducted on a corpus of Wikipedia Dataset.

WORK FLOW



ALGORITHM



MATERIALS AND METHODS

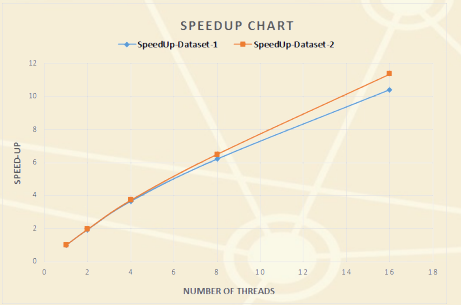
Harp[2] is a collective communication library plugged in Hadoop plug-in to accelerate machine learning algorithms.

LDA[3] is a popular topic modeling algorithm. We follow the Mr.LDA[4] to implement distributed variational inference LDA on Harp with it's dynamic schedueler, allreduce and push-pull communication models.

DATASETS

	Dataset-1	Dataset-2
docsize	744	7698
numOfWords	89907	435840
numOfTokenS	535156	5308848

EXPERIMENTS



CONCLUSIONS

Harp-LDA is proposed to provide high scalability achieve better performance with shorter time and memory requirements. A clear evidence of convergence of likelihood after a certain number of iterations is depicted.

The results from the speed up chart illustrate high scalability.

REFERENCE

[1] B. Zhang, Y. Ruan, J. Qiu, Harp: Collective Communication on Hadoop, in the proceedings of IEEE International Conference on Cloud Engineering (IC2E2015), March 9-13, 2015.

[2] Harp project https://github.iu.edu/IU-Big-Data-Lab/Harp

[3] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.

[4] Zhai, Ke, et al. "Mr. LDA: A flexible large scale topic modeling package using variational inference in mapreduce." Proceedings of the 21st international conference on World Wide Web. ACM, 2012.

INDIANA UNIVERSITY BLOOMINGTON
SCHOOL OF INFORMATICS AND COMPUTING