6. CONCLUSION

In this paper, we introduced MB-LDA for topic mining in microblogs and proposed distributed MB-LDA to handle the situation of large scale microblogs. Experimental results show that MB-LDA exhibits good performance and distributed MB-LDA scales well on actual dataset.

In future work, the automatic learning of hyperparameters in MB-LDA can be conducted to achieve a better performance in various application scenarios. Besides, we plan to use distributed cache to reduce the time spent on I/O transfer and speed up the execution time of distributed MB-LDA. Other strategies or implementations to reduce network communication should be also considered.

7. ACKNOWLEDGMENTS

The work is supported by Ministry of Industry and Information Technology of China (No. 2010ZX01042-002-003-001)

8. REFERENCES

[1] J. H. Kang, K. Lerman, A. Plangprasopchok. Analyzing Microblogs with Affinity Propagation. In *Proceedings of the 1st KDD workshop on Social Media Analytic*, 2010: 67-70

[2] A. Java, X. Song, T. Finin, et al. Why we Twitter: Understanding Microblogging Usage and Communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis (WebKDD/SNA-KDD)* 2007:56-65

[3] B. Krishnamurthy, P. Gill, M. Arlitt. A few chirps about Twitter. In *Proceedings of the first workshop on online social networks (WOSP)*, 2008:19-24

[4] D. Ramage, S. Dumais, D. Liebling. Characterizing microblogs with topic models. In *Proceedings of International AAAI Conference on Weblogs and Social Media*, 2010: 130-137

[5] J. Dean, S. Ghemawat. MapReduce: simplified data processing on large clusters. *Commun*. 2008, 51(1): 107-113

[6] R. Xu, D. Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 2005, 16(3): 645–678

[7] S. Deerwester, S. Dumais, T. Landauer, et al. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 1990, 41(6): 391–407

[8] G. Salton, M. McGill. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983

[9] T. K. Landauer, P. W. Foltz, D. Laham. Introduction to Latent Semantic Analysis. *Discourse Processes*, 1998, 25: 259-284

[10] D. M. Blei, A. Y. Ng, M. I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 2003, 3: 993–1022

[11] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual International ACM SIGIR Conference on Research and development in information retrieval*, 1999: 50-57

[12] T. Griffiths, M. Steyvers. Probabilistic topic models. *Latent Semantic Analysis: A Road to Meaning.* Hillsdale, NJ: Laurence Erlbaum, 2006

[13] X. Wei and W. B. Croft. LDA-based document models for ad hoc retrieval. In *Proceedings of the 29th annual International ACM SIGIR Conference on Research and development in information retrieval*, 2006: 178-185

[14] L. Dietz, S. Bickel, T. Scheffer. Unsupervised prediction of citation influences. In *Proceedings of the 24th International Conference on Machine learning*, 2007: 233-240

[15] QiaoZhu Mei, Deng Cai, Duo Zhang, et al. Topic Modeling with Network Regularization. In *Proceedings of the 17th International Conference on World Wide Web*. 2008

[16] D. M. Blei, J. Lafferty. Topic models. *Text Mining: Classification, Clustering, and Applications*. New York: Chapman & Hall/CRC, 2009

[17] R. Nallapati, W. Cohen. Link-pLSA-LDA: A new unsupervised model for topics and influence of blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*, 2008

[18] Congkai Sun, Bin Gao, Zhenfu Cao, et al. HTM: A topic model for hypertexts. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 2008: 514-522

[19] D. Newman, A. Asuncion, P. Smyth, et al. Distributed Algorithms for Topic Models. *Journal of Machine Learning Research*. 2009, 1801-1828.

[20] A. Asuncion, P. Smyth, M. Welling. Asynchronous distributed learning of topic models. In *Proceedings of the 20th Neural Information Processing Systems (NIPS)*. 2008

[21] Yi Wang, Hongjie Bai, M. Stanton, et al. PLDA: Parallel Latent Dirichlet Allocation for Large-Scale Applications. In *Proceedings of the 5th International Conference on Algorithmic Aspects in Information and Management (AAIM '09)*, 2009, 301-314.

[22] A. Smola, S. Narayanamurthy. An architecture for parallel topic models. *Proceedings of VLDB Endow*.2010, 3, 1-2, 703-710.

[23] T. L. Griffiths, M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 2004, 101:5228–5235,

[24] T. P. Minka, J. Lafferty. Expectation-propagation for the generative aspect model. In *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, 2002: 352-359

[25] Xiao Yang, Jianling Sun. An Analytical Performance Model of MapReduce. In *Proceedings of CCIS*: 306 – 310, 2011.