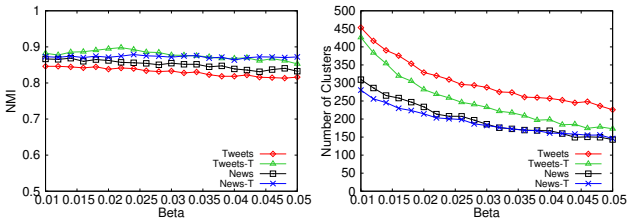

(a) Performance **(b) Number of clusters**

Figure 8: Influence of Beta.

in each cluster, and the probability of the document choosing the potential cluster grows with α.

5.8 Influence of Beta

In this part, we try to investigate the influence of β to the peformance and the number of clusters found by MStream and MStreamF. We use MStream to deal with Tweets and News datasets and MStreamF to deal with Tweets-T and News-T datasets. The range of β is from 0.01 to 0.05.

Figure 8a shows the performance of MStream and MStreamF with different values of β. From Figure 8a, we can see MStream and MStreamF can achieve stable performance with different β on these datasets. Figure 8b shows the number of clusters found by MStream and MStreamF with different values of β. An observation is that the number of clusters found drops when β gets larger. The reason is that β is the pseudo frequency of each word in each cluster, and the probability of a document choosing a cluster is less sensitive to the similarity between the documents and the clusters when β gets larger. As a result, "richer gets richer" property makes MStream and MStreamF get fewer clusters.

6 CONCLUSION

In this paper, we first propose a short text stream clustering algorithm based on the Dirichlet process multinomial mixture model, call MStream, which can deal with the concept drift problem and sparsity problem naturally. The MStream algorithm can achieve state-of-the-art performance with only one pass of the stream, and can have even better performance when we allow multiple iterations of each batch. We propose an improved algorithm of MStream with forgetting rules called MStreamF, which can efficiently delete outdated documents by deleting clusters of outdated batches. Our extensive experimental study shows that MStream and MStreamF can achieve better performance than three baselines on real datasets. As future work we intent to use the proposed methods to improve the performance of other related applications such as search result diversification, event detection and tracking, and text summarization in the context of short text streams.

ACKNOWLEDGMENTS

This work was supported in part by National Basic Research 973 Program of China under Grant No. 2015CB352502 and 2014CB340505, National Natural Science Foundation of China under Grant No. 61702190, 61532010, and 61521002.

REFERENCES

[1] Charu C Aggarwal. 2013. A Survey of Stream Clustering Algorithms. (2013).
[2] Charu C Aggarwal and S Yu Philip. 2010. On clustering massive text and categorical data streams. *Knowledge and information systems* 24, 2 (2010), 171–196.
[3] Charu C Aggarwal, S Yu Philip, Jiawei Han, and Jianyong Wang. 2003. -A Framework for Clustering Evolving Data Streams. In *VLDB*. Elsevier, 81–92.
[4] Amr Ahmed and Eric Xing. 2008. Dynamic non-parametric mixture models and the recurrent chinese restaurant process: with applications to evolutionary clustering. In *SDM*. SIAM, 219–230.
[5] Hesam Amoualian, Marianne Clausel, Eric Gaussier, and Massih-Reza Amini. 2016. Streaming-lda: A copula-based approach to modeling topic dependencies in document streams. In *SIGKDD*. ACM, 695–704.
[6] Charles E Antoniak. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The annals of statistics* (1974), 1152–1174.
[7] David Blackwell and James B MacQueen. 1973. Ferguson distributions via Pólya urn schemes. *The annals of statistics* (1973), 353–355.
[8] David M Blei and John D Lafferty. 2006. Dynamic topic models. In *ICML*. ACM, 113–120.
[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* (2003). http://dl.acm.org/citation.cfm?id=944919.944937
[10] Feng Cao, Martin Estert, Weining Qian, and Aoying Zhou. 2006. Density-based clustering over an evolving data stream with noise. In *SDM*. SIAM, 328–339.
[11] Arnaud Doucet, Nando De Freitas, and Neil Gordon. 2001. An introduction to sequential Monte Carlo methods. In *Sequential Monte Carlo methods in practice*. Springer, 3–14.
[12] Nan Du, Mehrdad Farajtabar, Amr Ahmed, Alexander J Smola, and Le Song. 2015. Dirichlet-hawkes processes with applications to clustering continuous-time document streams. In *SIGKDD*. ACM, 219–228.
[13] Thomas S Ferguson. 1973. A Bayesian analysis of some nonparametric problems. *The annals of statistics* (1973), 209–230.
[14] Hemant Ishwaran and Lancelot F James. 2001. Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* 96, 453 (2001), 161–173.
[15] Tomoharu Iwata, Shinji Watanabe, Takeshi Yamada, and Naonori Ueda. 2009. Topic Tracking Model for Analyzing Consumer Purchase Behavior.. In *IJCAI*, Vol. 9. 1427–1432.
[16] Anil K. Jain. 2010. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* 31, 8 (2010), 651–666.
[17] Argyris Kalogeratos, Panagiotis Zagorisios, and Aristidis Likas. 2016. Improving text stream clustering using term burstiness and co-burstiness. In *SETN*. ACM, 16.
[18] Shangsong Liang, Emine Yilmaz, and Evangelos Kanoulas. 2016. Dynamic clustering of streaming short documents. In *SIGKDD*. ACM, 995–1004.
[19] Alireza Rezaei Mahdiraji. 2009. Clustering data stream: A survey of algorithms. *International Journal of Knowledge-based and Intelligent Engineering Systems* 13, 2 (2009), 39–44.
[20] Hai-Long Nguyen, Yew-Kwong Woon, and Wee-Keong Ng. 2015. A survey on data stream clustering and classification. *Knowledge and information systems* 45, 3 (2015), 535–569.
[21] Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom M. Mitchell. 2000. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning* 39, 2/3 (2000), 103–134.
[22] Gerard Salton, A. Wong, and C. S. Yang. 1975. A Vector Space Model for Automatic Indexing. *Commun. ACM* 18, 11 (1975), 613–620.
[23] Lidan Shou, Zhenhua Wang, Ke Chen, and Gang Chen. 2013. Sumblr: continuous summarization of evolving tweet streams. In *SIGIR*. ACM, 533–542.
[24] Jonathan A Silva, Elaine R Faria, Rodrigo C Barros, Eduardo R Hruschka, Andre CPLF De Carvalho, and João Gama. 2013. Data stream clustering: A survey. *ACM Computing Surveys (CSUR)* 46, 1 (2013), 13.
[25] Alexander Strehl and Joydeep Ghosh. 2003. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research* 3 (2003), 583–617.
[26] Yee Whye Teh. 2010. Dirichlet process. In *Encyclopedia of machine learning*. Springer, 280–287.
[27] Xuerui Wang and Andrew McCallum. 2006. Topics over time: a non-Markov continuous-time model of topical trends. In *SIGKDD*. ACM, 424–433.
[28] Yu Wang, Eugene Agichtein, and Michele Benzi. 2012. TM-LDA: efficient online modeling of latent topic transitions in social media. In *SIGKDD*. ACM, 123–131.
[29] Xing Wei, Jimeng Sun, and Xuerui Wang. 2007. Dynamic Mixture Models for Multiple Time-Series.. In *IJCAI*, Vol. 7. 2909–2914.
[30] Jianhua Yin and Jianyong Wang. 2014. A dirichlet multinomial mixture model-based approach for short text clustering. In *SIGKDD*. ACM, 233–242.
[31] Jianhua Yin and Jianyong Wang. 2016. A model-based approach for text clustering with outlier detection. In *ICDE*. IEEE, 625–636.
[32] Shinjae Yoo, Hao Huang, and Shiva Prasad Kasiviswanathan. 2016. Streaming spectral clustering. In *ICDE*. IEEE, 637–648.
[33] Shi Zhong. 2005. Efficient streaming text clustering. *Neural Networks* 18, 5-6 (2005), 790–798.