Conclusion

Semi-supervised clustering exploits labeled data to enhance clustering results on unlabeled data. There are many methods and algorithms to exploit labeled data. This paper will choose to analysis the word distribution, applying it in the LDA clustering process. The reason choosing LDA is that documents can be associated with multiple topics under this model. We embed LDA with the weights coefficient to form WWDLDA. After adding the weights coefficient, the word distribution in every iteration is adjusted to decrease the change of the probability of word over each cluster in labeled and unlabeled documents. The adjustment could get good clustering results.

Semi-supervised clustering can be used to discover new classes in unlabeled data in addition to assigning appropriate unlabeled data instances to existing categories. In this paper, our labeled data are sampled from all document classes. Next we will consider the situation that some classes are not available in labeled documents.

Acknowledgement

The authors would like to thank the anonymous reviewers for their comments and kindly suggestions. This paper is supported by the National Natural Science Foundation of China (NSFC, NO. 60863005; No.61262006), the Science and Technology Foundation of Guizhou Province (NO.20122125), the Scientific Research Project of Introduce Talents of Guizhou University (NO.201114), the Foundation of Informatization of Manufacturing Industry of Guizhou Province (NO. GY (2011)3074) and Graduate Innovated Foundation of Guizhou University.

References

[1] XH. Hu, XD. Zhang, CM. Lu, EK. Park, and XH. Zhou, “Exploiting wikipedia as external knowledge for document clustering, ”Proc. the 15th ACM SIGKDD Int’l Conf. on Knowledge Discovery and Data Mining, ACM Press, Jun, 2009. 389-396.

[2] HT. Zheng, BY. Kang, and HG. Kim. Exploiting noun phrases and semantic relationships for text document clustering [J]. Information Sciences, 2009 179(13) 2249-2262,.

[3] M. Mahdavi and H. Abolhassani. Harmony *K*-means algorithm for document clustering [J]. Data Mining and Knowledge Discovery, 2009 18(3) 370-391.

[4] WZ. Zhao, HF. Ma, ZQ. Li, and ZZ. Shi. Efficiently active learning for semi-supervised document clustering [J]. Journal of Software, 2012 23(6) 1486−1499,.

[5] D. Blei and M. Jordan. Modeling annotated data. Proc. SIGIR, ACM Press, Jul, 2003. 127–134.

[6] Shi Zhong. Semi-supervised Model-based Document Clustering: A Comparative Study [J]. Machine Learning, 2006 (65) 3-29.

[7] S. Zhong and J. Ghosh, “A comparative study of generative models for document clustering,”Proc. SDM Workshop on Clustering High Dimensional Data and Its Applications, SDM Press, May. 2003.

[8] A. Strehl, J. Ghosh, and R. Mooney, “Impact of similarity measures on Web-page clustering,” Proc. the Workshop on Artificial Intelligence for Web Search, AAAI Press, Jul. 2000. 58-64.