THE UNIVERSITY OF CHICAGO

Department of Statistics

DISSERTATION PROPOSAL

MINZHE WANG

Department of Statistics
The University of Chicago

From Topic Model to Information Retrieval

THURSDAY, November 15, 2018, at 2:00 PM

ABSTRACT

Topic model plays a crucial role in modern text mining. One of the classical models *pLSI* assumes the word document matrix has a low-rank structure in which the word-topic matrix is of great interest but is also difficult to estimate at the same time. We propose a new SVD-based method to learn the word-topic matrix from the corpus data which is based on a deep connection between the word-topic matrix and the low-dimensional simplex structure in the left singular vector ratios of the word-document matrix. The proposed method is the first one that has been explicitly shown to achieve the minimax l_1 error convergence rate lower bound up to a multi-logarithmic term, and it is also shown to be more computationally efficient compared to the existing methods. The applications of our method on two data sets, Associated Process (AP) and Statistics Literature Abstract (SLA), have also shown encouraging results which largely validates our discovery.

Information retrieval is another important task in text mining. But the popular existing ways of tackling this problem usually fail to take into account the intrinsic rate heterogeneity between query words generation and document words generation or enjoy any theoretical guarantee of IR under any measure of success. We propose a general Naive-Bayes based probability model framework for query generation in which this rate heterogeneity has been explicitly considered. Simple and interpretable algorithms have also been proposed under this framework with a Poisson query language model. We show through application on the SLA data set that our proposed methods outperform the popular existing methods in terms of 0-1 loss and demonstrate that this superiority comes exactly from the fact that our proposed methods have successfully taken into account this intrinsic rate heterogeneity while the other methods failed. We also

provide theoretical guarantees of the performance of our algorithms in terms of both 0-1 loss and l_2 loss when the *pLSI* model in the corpus is assumed.

Although the proposed modeling framework can be naturally extended to incorporate the word-topic heterogeneity in query generation, the improvement in real applications is tiny. One of the possible reasons is the inaccurate estimation of the topic-document matrix which is another component in the pLSI model in addition to the word-topic matrix. We discover that a pre-SVD elimination of the high-frequency anchor words can improve the accuracy of topic-document matrix estimation. We propose two LRT-based methods for this anchor words elimination step which has been shown to work well in simulations. Application on SLA data sets has also shown encouraging results that match human judgment. More theoretical work is needed to fully understand the proposed method, and a Higher-Criticism-based thresholding is also under investigation of its usefulness in determining adaptively the cutting value for the testing statistics in the screening step.

For information about building access for persons with disabilities, please contact Kirsten Wellman at 773.834.5169 or send an email to kwellman@galton.uchicago.edu. If you wish to subscribe to our email list, please visit the following web site: https://lists.uchicago.edu/web/subscribe/statseminars.