Abstract:

Microblogging services are widely used to share information in various domains. It is significant to seek to understand the opinions of individuals, or to gauge aggregated sentiment of mass populations. To date, Sentiment analysis has been mainly studied for product and movie reviews, which always contain sufficient and standard words, and differ substantially from microblogging. Unlike standard product and movie reviews, the texts in microblogging contain a various of slang, new abbreviations and specific web vocabulary from various domains. According to the characteristics of microblogging mentioned above, we propose a domain adaption Co-Training algorithm for sentiment analysis in microblogging based on a extended sentiment dictionary.

In this talk, I will introduce our algorithms for Sentiment Analysis in microblogging. Firstly, we extend the original LDA model to expand sentiment dictionary using some nondomain-specific sentiment words as seed words. Secondly, we remove the documents that contain no opinion words based on the generated sentiment dictionary. This process is in fact able to remove a certain objectivity documents from unlabeled training data set. Afterwards, We propose a co-training approach based on the extended sentiment dictionary to take advantage of a small number of labeled training data and a large number of unlabeled data. To utilize the orientation of sentiment words (i.e., positive and negative sentiment), we apply non-negative logistic regression as initial classifier for co-training. Extensive experiments have been conducted to verify the proposed approaches on real-world Twitter dataset in English and SINA micoblogging data set in Chinese. The experiment results show that our approaches significantly improve the accuracy of sentiment classification with a small number of labeled data in microblogging.

About the Speaker:

Yang Min is a full-time PHD Student at the Department of Computer Science, The University of Hong Kong. Her supervisor is Dr. K.P. Chow. Her research interest is Computer Forensics, Data Privacy, Machine learning, and Information retrieval.