Scalable Machine Learning for the Web: Foundations and Advances  
(Innovative Claims and Major Impacts)  

Yixuan Li

Machine learning is a method that can learn from and make predictions on data, and has been placed at the core of many computing tasks when devising explicit algorithms is unfeasible (e.g., search engines, spam filtering, and computer vision). Despite its success, however, many machine learning algorithms can be computationally inefficient or even intractable when dealing with massive data generated on the web. The goal of my research is to develop principled foundations and practical advances for scaling machine learning methods to web-scale data. In my thesis work, I aim to bridge the gap between theoretical understanding and practical usefulness, and I have applied such methodology to many different real-world problems. I will demonstrate using a few examples in the following.

**Scalable Graph Clustering.** Graph clustering is a primitive research problem with a wealth of research interest. Conventional approaches relying on the global graph structure is no longer feasible as graph scales to billions of nodes. I formulate rigorous principles for scalable graph clustering -- finding small clusters in large graphs in a localized and parallelizable way; and then deployed the proposed method within Google’s system for detecting spammer groups on YouTube. As a result, the deployed method greatly expanded the daily detection volume of spamming activities on YouTube, and achieved an accuracy of 98%, with 10 times faster running time compared to the state-of-the-art proposal within Facebook for detecting fake Likes. Part of the work has been integrated into Google’s production, search for daily deceptive practices on YouTube spanning over a billion users. The research results have been published in the top-tier conference on World Wide Web (WWW) in 2015 and 2016 (acceptance ratio < 15%) [2, 8,10,11,12], and has been covered in media by Cornell Chronicle.

**Modeling and Predicting Group Dynamics.** An integral part of my thesis is to develop machine learning model that reveals the dynamics of groups: the processes by which online groups grow new members and evolve over time. In a research collaboration with Tencent Corporation, we initiated and led the first research studying the real-world data from WeChat group messaging platform -- the largest social messaging service in China, with more than 600 million monthly active users. One compelling feature in WeChat is “chat group”, which allows users to chat with several others all at once. We provided analysis on dynamics of millions of chat groups by keeping track of their emergence, growth and demise over time. The developed prediction model takes into account user’s historical behavior as well as the social network structure, and tries to predict which user is likely to join the chat group in the next. We demonstrated the effectiveness by achieving a prediction accuracy of 98.66%. The research results have been published in WWW 2016 [9].

**Interpretable and Scalable Learning Neural Networks.** Deep learning, a subfield of machine learning, has recently been placed at the core of many web-driven tasks such as image classification. My thesis research takes on two important challenges in deep learning: the lack of principled understanding of what neural networks learn, and the cost of expensive computation and long training time (which can last for days or weeks). My thesis work has made unique contributions by addressing these issues. It sheds light on the inner workings and properties of such complex learning systems, and further leads to new techniques of how we can scale the expensive training of deep networks with limited computation resources. The value and broad impact of my work has been acknowledged by the research community, where it was selected for oral presentations in both NIPS workshop (6.7%) and top-tier deep learning conference ICLR (5.7%) in 2016 [5,6]. A recent work submitted to ICLR 2017 [3] has also established a new world record for image classification accuracy on widely used datasets.

---

1 Citation numbers reference the numbered publication list in my CV.