NATIONAL UNIVERSITY OF SINGAPORE

School of Computing

PH.D DEFENCE - PUBLIC SEMINAR

**Title:** **Studies on Machine Learning for Data Analytics in Business Application**

Speaker: Mr Fang Fang

Date/Time: 15 August 2014, Friday, 02:00 PM to 03:30 PM

Venue: Executive Classroom, COM2-04-02

Supervisor : Dr Datta Anindya, Associate Professor, School of Computing

Abstract:

The volume of data is growing at an unprecedented rate now. Data are being produced everywhere, from Facebook, Twitter, YouTube to Google search records, and more recently, mobile apps. The tremendous amount of data embodies incredible valuable information. Analysis of data, both structured and unstructured such as text, are quite valuable and useful to a number of groups of people such as marketers, retailers, investors, consumers and so on. In this thesis, we focus on predictive analytics problems in the context of business application and utilize machine learning methods to solve it.

Study I
focuses on cross-domain sentimental classification. Sentiment analysis is quite useful to consumers, marketers, organizations, etc. One of the tasks of sentiment analysis is to determine the overall sentiment orientation of a piece of text and supervised learning methods, which require labeled data for training, have been proven quite effective to solve this problem. One assumption of supervised methods is that the training domain and the data domain share exactly the same distribution, otherwise, accuracy drops dramatically. However, in some circumstances, labeled data is quite expensive to acquire. For instance, Tweets and comments in Facebook. Study I addresses this problem and proposes an approach to determine the sentiment orientation of a piece of text when in-domain labeled data is not available.

Study II
focuses on Industry Classification. Industry analysis, which studies a specific branch of manufacturing, service, or trade, is quite useful for various groups of people. Before industry analysis, we need to define industry boundaries effectively and accurately. Existing schemes like SIC, GICS or NAICS have two major limitations. Firstly, they are all static and assume that the industry structure is stable. Secondly, these schemes assume binary relationship and do not measure the degree of similarity. Study II aims to contribute the literature by proposing an industry classification methodology that can overcome these limitations. Our

method is on the basis of business commonalities using the topic features learned by the Latent Dirichlet Allocation (LDA) from firms? business descriptions.

Study III
focuses on mobile app download estimation. Mobile apps represent the fastest growing consumer product segment of all times. To be successful, an app needs to be popular. The most commonly used measure of app popularity is the number of times it has been downloaded. For a paid app, the downloads will determine the revenue the app generates; for an ad-driven app, the downloads will determine the price of advertising on this app. In addition, researches in the app market necessities download numbers to measure the success of an app. Even though the app downloads are quite valuable, it turns out that number of download is one of the most closely guarded secrets in the mobile industry ? only the native store knows the download number of an app. Study III intends to propose a model of daily free app downloads estimation.