香港浸會大學理學院
HKBU Faculty of Science

for favour of posting

DEPARTMENT OF COMPUTER

PhD Degree Oral Presentation

PhD Candidate:	Mr Chi Wa CHENG
Supervisor:	Dr Chun Hung LI
External Examiner:	Dr Hau Sang WONG Dr Hon Wah TAM (Proxy for Prof Yan ZHANG)
Time:	31 March 2011 (Thursday) 3:00 pm – 5:00 pm (35 mins presentation and 15 mins Q & A)
Venue:	T716, Cha Chi Ming Science Tower, HSH Campus

"Probabilistic Topic Modeling and Classification Probabilistic PCA for Text Corpora"

Abstract

Topic modeling is one of the most common tools to analyze a large volume of unlabeled documents, which are usually represented with bag-of-words. This thesis firstly discusses the connections between the exchangeability property of bag-of-words, popular topic modeling algorithms, and de Finetti-Hewitt-Savage theorem. We showed that these algorithms are special cases of this theorem and the exchangeability of words, rather than independence of words, is the sufficient condition for applying them. Works are then focused on the latent Dirichlet allocation (LDA) because of its higher modeling capability. The investigation of asymmetric prior for LDA and derivation of per-document topic distribution for unseen documents are also presented. Since topics are often denoted by multinomial distributions of words, the semantic meaning cannot be easily understood especially when people are not familiar with the background of the studying corpus. To address this problem, automatic topic labeling is applied to propose understandable topic labels to users.

Apart from the text of a corpus, there are usually some meta information accompanied for analyses, e.g. author name, date, category, etc. Integrating them with text documents during topic modeling not only enable better topic analysis but also more information can be revealed. For instance, an author's interests can be identified if his/her name occurrences are tightly coupled to some words. Based on LDA and the author-topic model, we propose a Bayesian model with Dirichlet priors for combining text and author information to identify topics and interests associated with the corpus and the authors, respectively. With both the topics and the interests, generalization of the model is significantly improved. We also propose a composite model to combine identified topics and interests so that the overall composite topics of a corpus can be derived. These composite topics have a desirable property that the correlation between them is lower and hence they can represent more different aspects of the corpus.

Text corpus analyses are often performed in low dimensional spaces rather than high dimensional spaces formed by bag-of-words. Dimensionality reduction can be done with PCA or some other algorithms. Nevertheless, most of them are unsupervised and complementary information, if available, such as labels of documents is ignored. Even there are some supervised dimensionality reduction algorithms such as supervised probabilistic PCA, they treat labels as real numbers but not nominal categories. We propose the classification probabilistic PCA (CPPCA) to incorporate label information of documents, in which labels are treated as categories. Documents can be projected into a lower dimensional space where variances and labels are considered simultaneously. Semi-supervised version of this algorithm was applied to domain adaptation problems and experimental results showed that CPPCA performs significantly better than unsupervised and supervised probabilistic PCA.

***** ALL INTERESTED ARE WELCOME *****