Andrew McCallum, Professor
University of Massachusetts Amherst
Computer Science Department

Structured Topic Models for Natural Language and Social Network Analysis

Linguistics and social network analysis are being transformed by analysis of large collections of data. One approach to making sense of large data is to find a digestable number of latent components that summarize the original data. In this talk I will present several pieces of research in topic models---mixed-membership Bayesian networks that provide presentable, interpretable views of various data. I will focus on models that incorporate both text and additional meta-data, such as time stamps, relations and other attributes. For example, the "Author-Recipient-Topic" model discovers role-similarity between entities by examining not only social network connectivity, but also the words communicated on on those edges; I'll demonstrate this method on a large corpus of email data subpoenaed as part of the Enron investigation. The "Group-Topic" model discovers groups of entities and the topical conditions under which different groupings arise; I'll demonstrate this on coalition discovery from 16 years worth of voting records in the U.S. Senate and the U.N. I'll conclude with
further examples of topic models successfully applied to various large text collections, as well as discussion of their applicability to trend analysis, expert-finding and bibliometrics.

Joint work with David Mimno, Hanna Wallach, Xuerui Wang, Wei Li, Andres Corrada-Emmanuel and Natasha Mohanty.

Bio:

Andrew McCallum is a Professor and Director of the Information Extraction and Synthesis Laboratory in the Computer Science Department at University of Massachusetts Amherst. He has published over 200 papers in many areas of AI, including natural language processing, machine learning, data mining and reinforcement learning, and his work has received over 25,000 citations. He obtained his PhD from University of Rochester in 1995 with Dana Ballard and a postdoctoral fellowship from CMU with Tom Mitchell and Sebastian Thrun. In the early 2000's he was Vice President of Research and Development at at WhizBang Labs, a 170-person start-up company that used machine learning for information extraction from the Web. He is a AAAI Fellow, the recipient of the UMass NSM Distinguished Research Award, the UMass Lilly Teaching Fellowship, and research awards from IBM, Microsoft and Google. He is the General Chair for the International Conference on Machine Learning (ICML) 2012, a member of the board of the International Machine Learning Society and the editorial board of the Journal of Machine Learning Research. For the past ten years, McCallum has been active in research on statistical machine learning applied to text, especially information extraction, co-reference, semi-supervised learning, topic models, and social network analysis. Work on search and bibliometric analysis of open-access research literature can be found at http://rexa.info. McCallum's web page is http://www.cs.umass.edu/~mccallum.