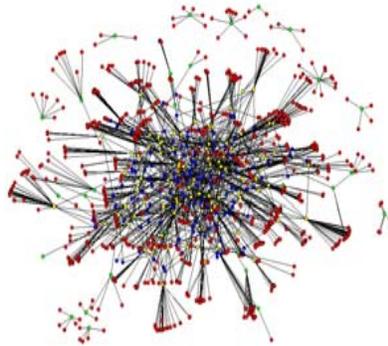


Four-Campus Video-Conference Series, 2006-7 f-w-s

Human Sciences and Complexity



May 18, 2007: Friday, 1:30-3:00 telecast from UCI 3030 Anteater Center

Padhraic Smyth

Department of Computer Science, UCI

“Automated Analysis of Relations between Words, Entities, Topics, and Documents using Statistical Topic Models”

<http://eclectic.ss.uci.edu/~drwhite/center/cac.html#Smyth>

Abstract: The availability of very large online corpora of text in digital form has led in recent years to the development of algorithms that try to automatically extract useful information and relationships from such text. In this talk I will describe a recent statistical approach that has proven to be very useful in this general context. Specifically I will discuss a representation for documents as mixtures of topics, where a topic is a probability distribution over words. The topics can be learned in a completely automated and unsupervised manner using a statistical estimation method called Gibbs sampling. I will illustrate the results of applying this approach to a diverse set of large corpora, including 250,000 emails from the Enron investigation, 300,000 news articles from the New York Times, 12,000 technical papers from UCI and UCSD faculty, and 80,000 articles from the Pennsylvania Gazette (from the 18th century). Once the statistical topic model is estimated for a specific corpus, a wide variety of interesting questions can be posed and answered: for example, how have topics changed over time in a particular corpus? which authors write on a particular topic? and so on. I will conclude with a discussion of how these statistical topic models can provide an interesting basis for automatically constructing large and complex networks from text and how these networks can support interesting inferences and insights that would be difficult (or impossible) to obtain by purely manual means.

Video Conference Locations for Participants

UCLA: 285 Powell Library

UCI: 3030 Teaching, Learning & Technology Center, corner E. Peltason and Anteater

UCSD: 260 Galbraith Hall

UCR: A139 Olmsted Hall