National Security Education Center

A CONSORTIUM OF LANL INSTITUTES

UNCLASSIFIED

Statistical Sciences Seminar Series

Peter Chew, Principal Investigator
Galisteo Consulting Group

"Solving Big Data Problems: From Heuristics to Theory"

Wednesday, April 18, 2012
10:00 - 11:00 AM
TA-3, Bldg. 40, N125, Moon Room

Abstract: In a 2011 Gartner survey, 47% of respondents identified data growth in their top 3 challenges, and this survey indicates that data capacity at enterprises is growing 40%-60% a year due to factors including an explosion in unstructured data such as e-mails and documents. As data grows, the need for reliable and generalizable approaches to data analytics is becoming more critical.

In the first part of this talk I shall review some major techniques and concepts in Information Retrieval (a field which deals with analysis of unstructured data), including Latent Semantic Analysis, Latent Dirichlet Allocation, term weighting, stemming, and stoplists. I will show how insights from information theory, similar to those which have produced advances such as Statistical Machine Translation in the field of computational linguistics, can be applied to reformulate common techniques in Information Retrieval more elegantly, and in a way which has the double benefit of significant (and sometimes dramatic) empirical improvements to the analytical output.

I shall then review two highly diverse big-data mining problems where I have been able to apply and test these techniques, one involving clustering multilingual documents by topic, and the other involving financial account reconciliation. In both cases we find that an information-theoretic approach to IR is highly effective as a method of solving non-trivial data mining problems, with accuracy between 90% and 100%.

Biography: Dr. Peter Chew is a Principal Investigator with Galisteo Consulting Group, Inc., a consulting firm in Albuquerque, NM. Through his work with Galisteo, he provides services in data and text analytics, national security, and business process improvement to clients in the public and private sector. Dr. Chew's current work includes projects focused on making sense of large datasets, including cutting-edge and patent-pending applications of statistics and informatics to improve analytic financial processes such as account reconciliation. His work also includes data analytics to support fraud investigations, and research into the cultural and linguistic aspects of national security (e.g. sentiment analysis of multilingual text).

Prior to Galisteo, Dr. Chew served as a Senior Manager at Moss Adams LLP (formerly Neff + Ricci LLP), the largest public accounting office in New Mexico, and at Price Waterhouse in the United Kingdom and Russia, where he participated in financial and technology audits of a variety of local and international organizations. Dr. Chew also served for over 5 years on the technical staff at Sandia National Laboratories, leading LDRD research into the application of linear and multilinear algebraic analytic techniques to natural language, in particular multilingual text.

Dr. Chew has a B.A. in Russian Language and Literature with subsidiary Polish from the School of Slavonic and East European Studies at the University of London, and an M.St. in Slavonic Studies and a D.Phil. in Computational Linguistics from the University of Oxford. He is also a Certified Public Accountant and Certified Fraud Examiner.

Los Alamos NATIONAL LABORATORY EST.1943

UNCLASSIFIED

For more information contact Kary Myers, kary@lanl.gov, 606-1455.
Hosted by the Information Science and Technology Center (ISTC)