



**THE CENTRE FOR PATTERN ANALYSIS AND MACHINE INTELLIGENCE IN ASSOCIATION WITH
IEEE SIGNAL PROCESSING, COMPUTATIONAL INTELLIGENCE, AND SYSTEMS, MAN &
CYBERNETICS CHAPTERS OF KW SECTION PRESENTS:**

Modeling Term Associations for Probabilistic Information Retrieval

Speaker: Professor Jimmy Huang, Professor and Director, School of Information Technology, York University

Date: Wednesday February 26, 2014

Time: 11:00 am

Place: University of Waterloo, Centre for Environmental and Information Technology, EIT-3142

**Refreshments will be provided*

Abstract:

Traditionally, in many probabilistic retrieval models, query terms are assumed to be independent. Although such models can achieve reasonably good performance, associations can exist among terms from human beings point of view. There are some recent studies that investigate how to model term associations/dependencies by proximity measures. However, the modeling of term associations theoretically under the probabilistic retrieval framework is still largely unexplored. In this talk, I will introduce a new concept named Cross Term, to model term proximity, with the aim of boosting retrieval performance. With Cross Terms, the association of multiple query terms can be modeled in the same way as a simple unigram term. In particular, an occurrence of a query term is assumed to have an impact on its neighboring text. The degree of the query term impact gradually weakens with increasing distance from the place of occurrence. We use shape functions to characterize such impacts. Based on this assumption, we first propose a bigram Cross Term Retrieval (CRTER2) model as the basis model, and then recursively propose a generalized n-gram Cross Term Retrieval (CRTERn) model for n query terms where $n > 2$.

Specifically, a bigram Cross Term occurs when the corresponding query terms appear close to each other, and its impact can be modeled by the intersection of the respective shape functions of the query terms. For n-gram Cross Term, we develop several distance metrics with different properties and employ them in the proposed models for ranking. We also show how to extend the language model using the newly proposed cross terms. Extensive experiments on a number of TREC collections demonstrate the effectiveness of our proposed models.

Biography:

Jimmy Huang is a Professor & Director at the School of Information Technology and the founding director of Information Retrieval & Knowledge Management Research Lab at the York University. He joined York University as an Assistant Professor in July 2003. Previously, he was a Post Doctoral Fellow at the School of Computer Science, University of Waterloo. He did his PhD in Information Science at City University in London. He also worked in the financial industry in Canada, where he was awarded a CIO Achievement Award. Since 2003, he has published more than 150 refereed papers in top-tier journals (such as ACM TOIS, JASIST, IPM, IEEE TKDE, Information Sciences, IR, BMC Bioinformatics and BMC Genomics), book chapters and international conference proceedings (such as ACM SIGIR, ACM CIKM, COLING and IEEE ICDM). In the past three years, he has published 27 papers in top-tier journals and 23 papers in conferences (8 papers in ACM SIGIR and 1 in ACM CIKM). He was awarded tenure and promoted to Full Professor at York University in 2006 and 2011 respectively. He received the Dean's Award for Outstanding Research in 2006, an Early Researcher Award, formerly the Premier's Research Excellence Awards in 2007, the Petro Canada Young Innovators Award in 2008, the SHARCNET Research Fellowship Award in 2009 and the Best Paper Award at the 32nd European Conference on Information Retrieval in 2010. He was the General Conference Chair for the 19th International ACM CIKM Conference and the General Program Chair for IEEE/ACM International Joint Conferences on Web Intelligence & Intelligent Agent Technology in 2010.



IEEE KW Section

