In Favour of Posting

香港中文大學
The Chinese University of Hong Kong

Institute of Theoretical Computer Science and Communications

ITCSC Seminar

Efficient Indexes of Document Retrieval Problems

By
Prof. Wing-Kai Hon
Assistant Professor, Department of Computer Science
National Tsing Hua University, Taiwan

January 11, 2010 (Monday)

4:00 pm – 5:00 pm

Rm. 121, Ho Sin Hang Engineering Building, CUHK

Abstract:
Given a collection C of text documents and a pattern P, a natural problem is to report the documents in C that are most relevant with respect to P. For instance, one may want to find out which documents in C contain more than 30 occurrences of P. The above problem is somewhat more advanced than the traditional pattern matching problem, as some notion of "most relevance" is involved.

Muthukrishnan (2002) showed how one can index the collection C of total length n with O(n log n) space, so that for any query pattern P one can solve the above problem in optimal time. In this talk, we will describe a new framework so that the index space is reduced to O(n) while query time remains optimal. Our framework can also be applied for reporting the top k most relevant documents (in sorted order) in optimal time under similar or more general relevance metrics.

This is a joint work with Rahul Shah (LSU) and Jeff Vitter (TAMU).

Biography:
Wing-Kai Hon received his PhD degree in Computer Science from University of Hong Kong in 2005. He is currently an assistant professor in the Department of Computer Science at National Tsing Hua University. Prior to that, Wing-Kai has visited Purdue University as a post-doc under the supervision of Jeff Vitter. His research interests include indexing, data compression, external memory data structures, and combinatorial optimization.

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