DIGITAL LIBRARIES & CYBERSCHOLARSHIP COLLOQUIUM SERIES

Featuring international experts in the Digital Libraries Field

http://www.ischool.pitt.edu/colloquia/digital-libraries-series.php

HOW TO READ 15 MILLION BOOKS IN ONE SITTING

(or mining a hypertext of quotations and ideas from very large digital libraries)

Bill Schilit
Google Research

Wednesday, Feb. 3
4-5 pm
1305 Newell Simon Hall
Carnegie Mellon University

Scanning books, magazines, and newspapers is widespread because people believe a great deal of the world's information still resides off-line. In general after works are scanned they are OCR'ed, indexed for search and processed to add links.

In this talk I will describe a new approach to automatically add links by mining repeated passages. This technique connects elements that are semantically rich, so strong relations are made. Moreover, link targets point within rather than to the entire work, facilitating navigation. Our system has been run on a digital library of many millions of books (Google Book Search), has been used by thousands of people, and has generated the world's largest collection of quotations.

I will also present a follow-on project based on the theory that authors copy passages from book to book because these quotations capture an idea particularly well: Jefferson on liberty; Stanton on women's rights; and Gibson on cyberpunk. These projects suggest that mining quotations for links and ideas is an important mechanism for understanding the knowledge contained in books. (This work is in collaboration with Okan Kolak, Google Research and Google Book Search.)

Bill Schilit is part of Google Research and an adopted member of the Book Search group. Before joining Google, Bill was co-director of the Intel Research lab in Seattle, managed digital library and mobile computing research at Fuji-Xerox (FXPAL), worked on distributed computing at AT&T's Bell Labs, and was part of the team that developed Ubiquitous Computing at PARC. He is a Fellow of the IEEE, Associate Editor-in-Chief of Computer Magazine and a past member of the Board of Governors of the IEEE Computer Society. Bill received a Ph.D. from Columbia University in 1995.

This colloquium series is sponsored by

UNIVERSITY OF PITTSBURGH www.ischool.pitt.edu
School of Information Sciences

Carnegie Library of Pittsburgh

UNIVERSITY OF PITTSBURGH ULS UNIVERSITY LIBRARY SYSTEM

Carnegie Mellon SCHOOL OF COMPUTER SCIENCE

Carnegie Mellon UNIVERSITY LIBRARIES