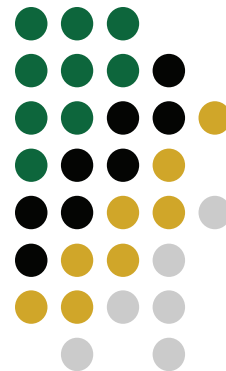


Computer Science Department

Special Colloquium



Dr. Richard L. Sites
Google, Inc.



“Statistical Language Detection in Web Pages”

Date: Friday, May 22, 2009

Time: 10:00 am

Location: Computer Science Building, Room 130

Abstract

Search engine companies prefer to show users pages that they can read. To do this, it is useful to identify the language(s) used on each web page. With billions of web pages in hundreds of languages, an automated statistical approach is needed. In contrast to previous work using short words or groups of three letters (trigrams) to identify perhaps a dozen different languages in single-language well-written text corpora, we look at the more general problem of detecting ~180 languages in ~57 Unicode scripts, in the sometimes mixed-language wild-west text of Web pages. We discuss statistical detection using quadgrams, building statistics offline from the Web itself as corpus, and dealing with unusual pages – e.g., what goes wrong.

Speaker Biography

Dick Sites is a Senior Staff Engineer at Google, where he has worked for 5 years. He previously worked at Adobe Systems, Digital Equipment Corporation, Hewlett-Packard, Burroughs, and IBM. His accomplishments include co-architecting the DEC Alpha computers, advancing the art of binary translation for computer executables, adding electronic book encryption to Adobe Acrobat, decoding image metadata for Photoshop, and building various computer performance monitoring and tracing tools at the above companies. He also taught Computer Science for four years at UC/San Diego. Most recently he has been working on Unicode text processing. Dr. Sites holds a PhD degree in Computer Science from Stanford and a BS degree in Mathematics from MIT. He also attended the Master's program in Computer Science at UNC 1969-70. He holds 46 patents and was recently elected to the National Academy of Engineering.