Computer Science
Seminar

(Re-) Discovering Lost Web Pages

Michael Nelson
Old Dominion University

Abstract: Missing web pages (pages that return the "404 Page Not Found" error) are part of the browsing experience. So too are pages whose owners failed to renew their domain and whose old URls now have unexpected content. Users that encounter a missing page or unexpected page may try to use search engines to discover either the same page at a new location or a similar, "good enough" page to satisfy their information needs, but this can be laborious. To address this need, we are developing a semi-automated framework to assist users to first discover the topic of the missing page, and then locate the same or similar page at a new URl.

We have been investigating a number of techniques to discover the "aboutness" of an unknown web page. If the page is in the internet Archive's Wayback Machine or in a search engine cache, the user may be satisfied with the old copy. If an old copy is insufficient, we can use either the page's title or generate a lexical signature to serve as a queiry to a search engine to find the resource. A lexcial signature is a 5-7 word "abstract" of a document that is suitable for using as a queiry to a search engine. The performance of titles and lexcial signatures are comparable, with both achieving over 60 percent success. The combination of titles and lexical signatures from link neighborhoods as well as using tags from del.icio.us, but at this point neither method performs well.

Speaker Bio:
Michael L. Nelson is an associate professor of computer science at Old Dominion University. Prior to joining ODU, he worked at NASA Langley Research Center from 1991-2002. He is co-editor of the OAI-PMH and OAI-ORE specifications and is a 2007 recipient of an NSF CAREER award. In 2008, Dr. Nelson was named a "Digital Preservation Pioneer" by the Library of Congress. He has developed many digital libraries, including the NASA technical Report Server. His research interests include repository-object interaction and alternative approaches to digital preservation. More information about Dr. Nelson can be found at http://www.cs.odu.edu/ min/

Friday, October 2, 2009, 3:00 pm
Mathematics and Science Center: W301

Mathematics and Computer Science
Emory University