Using Metadata to Improve Search Results

Jeremy Ey
March 12th, 2007
CSC 6800-001
Background Information

• Hard to create machine understand of html.

• Use of common structures to provide information about what is contained on the page.

• Microformats provide a common set of html class names to denote this information
Client Side Use

Examples

• Example of an hCard within a webpage:

  http://michael-mccracken.net/wp/
GRDDL

- Gleaning Resource Descriptions from Dialects of Languages
- Provide references to xslt stylesheets to transform content from one format to another.
Setup

- QEMU Virtual Machine with xubuntu
- run on Windows XP hosts
- WIRE
- Local DNS Cache
WIRE Configuration

• looking at .com .net .edu .org and .us domains
• maximum of 10K urls per site
• maximum depth of 5 levels
• maximum of 50K documents per harvester run
Number of sites ok, 137
Number of sites with valid page age, 48
Average pages per site, 7.402920e+01
Average static pages per site, 6.359854e+01
Average dynamic pages per site, 1.043066e+01
Average of age of oldest page in months, 7.833333e+00
Average of age of newest page in months, 2.666667e+00
Average of age of average page in months, 4.229167e+00
Average in-degree, 0.0000000e+00
Average out-degree, 0.0000000e+00
Average internal links, 0.0000000e+00
Average site size in MB, 0.0000000e+00
Average site max depth, 1.941606e+00
Starting URLs

• http://www.csc.tntech.edu/
• http://microformats.org/wiki/hcard-examples-in-wild/
• http://michael-mccracken.net/wp/
Crawling Results

- 6 rounds of harvesting
- 800 sites
- 13436 urls
Additional Results

- I’m starting to question the determinism of computers
Next Steps

• Count and index hCard data
• Interface to access hCard Index
• Expand metadata types reviewed (GRDDL)
Future Work

• Adjust crawling depth to further explore sites where data has previously been found

• Support plugins from other systems to expand data collection ability