SearchResultFinder

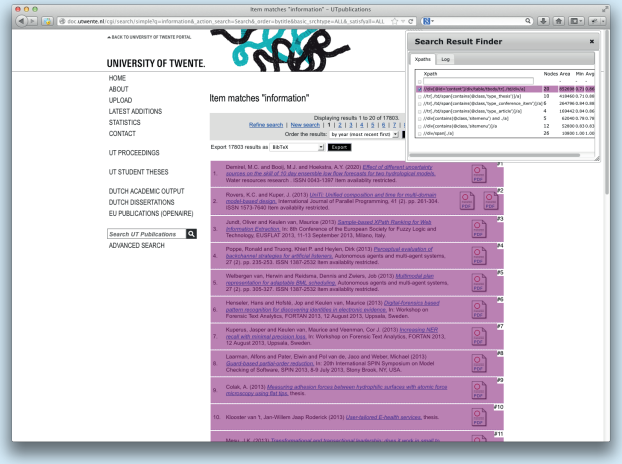
Federated Search Made Easy

Dolf Trieschnigg, Kien Tjin-Kam-Jet and Djoerd Hiemstra
{d.trieschnigg,k.t.t.e.tjin-kam-jet,d.hiemstra}@utwente.nl
UNIVERSITY OF TWENTE.

Why? Extracting search results from web search engines is a pain: Implementing an API, using existing wrapper generators or manually determining XPaths is time-consuming

What? (Semi-)Automatically determine XPaths to extract search result items

How? Based on a single search result page using a Firefox plugin



Features

- Requires only a single search result page
- Two clicks away in your browser
- Fast results (typically < 1s)
- Returns a ranked list of XPaths
- Easy inspection of the extracted nodes
- Automatable using Selenium

Steps

1. Use Firefox to render the search result page and obtain a DOM-tree
2. Find repeating nodes containing anchors <a>
3. Generate candidate XPaths, e.g. /html/body/div/div[./p/a]
4. Simplify candidate XPaths, e.g. //div[@id='results']/div
5. Filter and rank using heuristics based on size, grid-detection and node similarity

Evaluation

Evaluated search engines	148
Correct XPath at rank 1	120
Correct XPath at rank 2	8
Correct XPath at rank > 2	5
Correxct XPath not found	15
Mean reciprocal rank	0.84

Available from: snipdex.org/SRF

Interested in Federated Web Search?

You can still participate in the TREC FedWeb track!
Dataset: 2000 query samples from 157 web search engines
Topics: 200 (50 judged)

Tasks: 1) Resource selection (deadline 11 Aug 2013)
2) Resource merging (deadline 15 Sept 2013)

More info: snipdex.org/fedweb