CS856 (Fall/2002) Presentation:
Crawling the Hidden Web

Discussant: Xuhui Li

Contents:
- 1. Strengths of the paper
- 2. Insufficiencies
- 3. Conclusion
Strengths

- Feedback from response analysis could be used to tune the match function.
  
  Such as, adjust the weight $M_v(v)$ for a value $v$ correspond to an form element $E$.

Strengths (Continued)

- Crawler contributes new entries to LVS
  - Finite domain form elements are good candidates as value assignments for future matching.
  
- Plan to study the dependencies among elements within the same form. (City and state, manufacturer and brand)
Insufficiencies

- Only can be used in some specific information queries, or as search agent.
  - Not support general searching
  - Not respond immediately
- Should not be compared with the crawlers used by search engines.
  - Predefined search category
  - Task-specific LVS database
  - Starting URL list

Insufficiencies (Continued)

- Not mentioned the stop condition of the crawling
- Enumeration of all values from finite domain element
  - Could be expensive
  - Not necessary
Insufficiencies (Continued)

- LVS maintenance
  - Weights of initial values should not be fixed
  - Deletion of values

- No introduction for response analyzer

Insufficiencies (Continued)

- No caching
  - No caching at any level
  - Each query needs a separate crawling
  - Fresh but expensive
Insufficiencies (Continued)

- Cannot reach all parts of the hidden Web
  - Form elements relating to some graphs
  - Forms need to fill personal information (How many years have been in your current job position?)
  - Form with only one or two elements
  - Dynamically generated pages (CGI, Server applet)

Conclusion

- This is a good trial solution in specifically well predefined interest crawling.
- To provide “Hidden Web Crawling” for general purpose, such as search engines, there is still a long way to go.