World Wide Web Site Summarization
Yongzheng Zhang, Nur Zincir-Heywood, Evangelos Milios
Available at: http://www.cs.dal.ca/research/techreports/2002/CS-2002-08.html

Motivation
- Web information overload problem
- Manually constructed summaries available at DMOZ, but expensive to build
- Automatic web site summarization derives from:
  - Text summarization
  - Web page summarization

Text Extraction
- Extract plain text from Web pages
- Tools available:
  - HTML2TXT v2.0
  - Text browser Lynx
  - Our own module HTML Parser
- Evaluations show Lynx is the best

Narrative Text Classification
- Define rules to extract narrative text
  - Filtering – remove short paragraphs
  - Use C.0 to determine threshold
  - Classify long paragraphs into narrative and non-narrative
  - Part of speech tagging
  - C.0 on frequencies of tags

Key Sentence Extraction
- Measure the significance of sentence by maximum cluster weight
- Cluster: a sequence of consecutive words:
  - The sequence starts and ends with a keyword
  - Less than 2 non-keywords separate any two neighboring keywords
- Compute cluster weight:
  - Add weights of keywords
  - Divide the sum by number of keywords

W3SS Summary Example

Experiments
- Collect 20 Web sites from DMOZ (10 academic, 10 commercial)
- Extrinsic evaluation: 4 groups of subjects, 5 in each group
- Answer pre-defined questions based on W3SS summaries, DMOZ summaries, home page browsing, and time-limited site browsing

Conclusions
- W3SS summaries are as informative as DMOZ summaries
- Significantly better than home page browsing and time-limited site browsing