

World Wide Web Site Summarization

Yongzheng Zhang, Nur Zincir-Heywood, Evangelos Milios

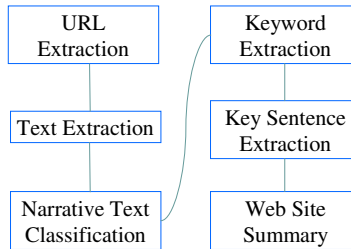
Available at: <http://www.cs.dal.ca/research/techreports/2002/CS-2002-08.html>

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

1

Overview



2

URL Extraction

- A Web crawler
- Breadth first search of Web site
- Crawling depth: 4 (four)
- Number of pages to visit: 1000
- Collect URLs of Web pages

3

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

4

Narrative Text Classification

- Define rules to extract narrative text
 - Filtering – remove *short* paragraphs
 - Use C5.0 to determine threshold
 - Classify *long* paragraphs into *narrative* and *non-narrative*
 - Part of speech tagging
 - C5.0 on frequencies of tags

5

Keyword Extraction

- Keyword – word with very high total frequency
- Extract keywords from narrative text, anchor text, and special text (italicized, bold text, etc.)
- Objective: measure how important each category of keywords is

6

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

7

W3SS Summary Example

Part 1. Top 25 Keywords				
sei	system	software	cmu	product
component	information	process	architecture	organization
course	program	report	practice	project
method	design	institute	development	research
document	management	defense	technology	team
Part 2. Top 10 Key-terms				
software engineering	carnegie mellon	development center	software process	software architecture
maturity model	risk management	software development	process improvement	software system
Part 3. Top 5 Key Sentences				
1. The Software Engineering Information Repository (SEIR) is a Web-based repository of information on software engineering practices that lead to improved organizational performance.				
2. Because of its mission to improve the state of the practice of software engineering, the SEI encourages and otherwise facilitates collaboration activities between members of the software engineering community.				
3. The SEI mission is to provide leadership in advancing the state of the practice of software engineering to improve the quality of systems that depend on software.				
4. The Software Engineering Institute is operated by Carnegie Mellon University for the Department of Defense.				
5. The Software Engineering Institute (SEI) sponsors, co-sponsors, and is otherwise involved in many events throughout the year.				

8

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
 - Time-limited site browsing

9

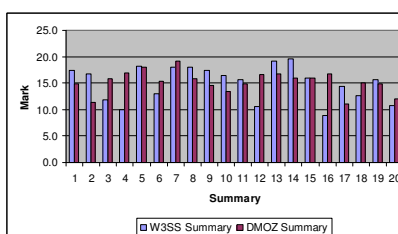
Performance Analysis

- Answers were marked out of 20
- Average scores over subjects across all summaries
- W3SS & DMOZ are better than HPB & TLSB

W3SS	DMOZ	HPB	TLSB
15.0	15.3	12.7	13.4

10

W3SS vs. DMOZ



11

Conclusions

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

12