WIRE: an Open Source Web Information Retrieval Environment

Carlos Castillo and Ricardo Baeza-Yates

Center for Web Research
http://www.cwr.cl/
CS Dept., University of Chile

OSWIR 2005
Compiegne, France
September 19, 2005
1. WIRE Project

2. Web Crawler

3. Conclusions
Motivation

Study subsets of the Web (1-50 million pages)

- We want high performance
- We want to keep as much data as possible
- We want to study scheduling algorithms
- wget is not enough
- Large-scale crawlers were not publicly available
Motivation

Study subsets of the Web (1-50 million pages)

- We want high performance
- We want to keep as much data as possible
- We want to study scheduling algorithms
- `wget` is not enough
- Large-scale crawlers were not publicly available
Motivation

Study subsets of the Web (1-50 million pages)

- We want high performance
- We want to keep as much data as possible
- We want to study scheduling algorithms
- wget is not enough
- Large-scale crawlers were not publicly available
Motivation

Study subsets of the Web (1-50 million pages)

- We want high performance
- We want to keep as much data as possible
- We want to study scheduling algorithms

- wget is not enough
- Large-scale crawlers were not publicly available
Motivation

Study subsets of the Web (1-50 million pages)

- We want high performance
- We want to keep as much data as possible
- We want to study scheduling algorithms
- wget is not enough
- Large-scale crawlers were not publicly available
Study subsets of the Web (1-50 million pages)

- We want high performance
- We want to keep as much data as possible
- We want to study scheduling algorithms
- wget is not enough
- Large-scale crawlers were not publicly available
Characteristics

- Roughly 25,000 lines of open-source C/C++ code
- Asynchronous DNS and HTTP requests, small memory and processing requirements (except during the analysis)
- Highly configurable: rate of download, parser parameters, scheduling policy, etc.
Characteristics

- Roughly 25,000 lines of open-source C/C++ code
- Asynchronous DNS and HTTP requests, small memory and processing requirements (except during the analysis)
- Highly configurable: rate of download, parser parameters, scheduling policy, etc.
Characteristics

- Roughly 25,000 lines of open-source C/C++ code
- Asynchronous DNS and HTTP requests, small memory and processing requirements (except during the analysis)
- Highly configurable: rate of download, parser parameters, scheduling policy, etc.
Web Crawler

Manager
Page score calculations
Long-term scheduling

Seeder
Link resolving
Robots exclusions

Harvester
Short-term scheduling
Network transfers

Gatherer
Parsing
Link extraction

Collection

Carlos Castillo and Ricardo Baeza-Yates

WIRE: an Open Source Web Information Retrieval Environment

http://www.cwr.cl/
## Scheduling

<table>
<thead>
<tr>
<th>P_i</th>
<th>Quality</th>
<th>Freshness</th>
<th>Visited</th>
<th>Future Value</th>
<th>Current Value</th>
<th>Profit</th>
</tr>
</thead>
<tbody>
<tr>
<td>P_1</td>
<td>0.4</td>
<td>0.1</td>
<td>1</td>
<td>0.4</td>
<td>0.04</td>
<td>Profit: 0.36</td>
</tr>
<tr>
<td>P_2</td>
<td>0.7</td>
<td>0.9</td>
<td>1</td>
<td>0.7</td>
<td>0.63</td>
<td>Profit: 0.07</td>
</tr>
<tr>
<td>P_3</td>
<td>0.6</td>
<td>-</td>
<td>0</td>
<td>0.6</td>
<td>0</td>
<td>Profit: 0.6</td>
</tr>
</tbody>
</table>

Future Value - Current Value = Profit
Downloading pages

World Wide Web

Web sites
S1
S2
S3
S4
S5
S6
S7

Web pages
P1,1
P1,2
P1,3
P1,4
P2,1
P2,2
P2,3
P2,4
P2,5
P2,6
P3,1
P3,2
P4,1
P4,2
P4,3
P4,4
P4,5
P5,1
P5,2
P5,3
P5,4
P6,1
P6,2
P6,3
P7,1
P7,2
P7,3
P7,4
P7,5

Carlos Castillo and Ricardo Baeza-Yates

WIRE: an Open Source Web Information Retrieval Environment

http://www.cwr.cl/
Storing contents

1. Document
2. Content seen?
3. Disk Storage

Free space list

Carlos Castillo and Ricardo Baeza-Yates

WIRE: an Open Source Web Information Retrieval Environment

http://www.cwr.cl/
URL parsing

1. $h_1('host.domain.com')$
2. host.domain.com 235
3. $h_2('235 dir/file.html')$
4. 235 path/file.html 9421

SITE-ID = 235; DOC-ID = 9421
Practical problems

- The devil is in the details
  - Varying quality of service
  - Wrong DNS records, temporary DNS failures
  - HTTP responses without headers, with wrong headers, dates
  - HTML parsing has to be very tolerant
  - Duplicate pages, session-ids, etc.
Practical problems

- The devil is in the details
- Varying quality of service
  - Wrong DNS records, temporary DNS failures
  - HTTP responses without headers, with wrong headers, dates
  - HTML parsing has to be very tolerant
  - Duplicate pages, session-ids, etc.
Practical problems

- The devil is in the details
- Varying quality of service
- Wrong DNS records, temporary DNS failures
- HTTP responses without headers, with wrong headers, dates
- HTML parsing has to be very tolerant
- Duplicate pages, session-ids, etc.
Practical problems

- The devil is in the details
- Varying quality of service
- Wrong DNS records, temporary DNS failures
- HTTP responses without headers, with wrong headers, dates
- HTML parsing has to be very tolerant
- Duplicate pages, session-ids, etc.
Practical problems

- The devil is in the details
- Varying quality of service
- Wrong DNS records, temporary DNS failures
- HTTP responses without headers, with wrong headers, dates
- HTML parsing has to be very tolerant
- Duplicate pages, session-ids, etc.
Practical problems

- The devil is in the details
- Varying quality of service
- Wrong DNS records, temporary DNS failures
- HTTP responses without headers, with wrong headers, dates
- HTML parsing has to be very tolerant
- Duplicate pages, session-ids, etc.
Data analysis

- Includes link analysis and extraction of statistics (data is exported as .csv files)
- Reports are generated using \LaTeX{} and gnuplot
- Report about documents: histograms of size, in- and out-degree, link scores, page depth, HTTP responses, age, media types, etc.
- Report about sites: degree distribution in the hostgraph, maximum depth, pages per site, link structure, etc.
Data analysis

- Includes link analysis and extraction of statistics (data is exported as .csv files)
- Reports are generated using \LaTeX{} and gnuplot
- Report about documents: histograms of size, in- and out-degree, link scores, page depth, HTTP responses, age, media types, etc.
- Report about sites: degree distribution in the hostgraph, maximum depth, pages per site, link structure, etc.
Data analysis

- Includes link analysis and extraction of statistics (data is exported as `.csv` files)
- Reports are generated using LaTeX and gnuplot
- Report about documents: histograms of size, in- and out-degree, link scores, page depth, HTTP responses, age, media types, etc.
- Report about sites: degree distribution in the hostgraph, maximum depth, pages per site, link structure, etc.
Data analysis

- Includes link analysis and extraction of statistics (data is exported as .csv files)
- Reports are generated using \LaTeX{} and gnuplot
- Report about documents: histograms of size, in- and out-degree, link scores, page depth, HTTP responses, age, media types, etc.
- Report about sites: degree distribution in the hostgraph, maximum depth, pages per site, link structure, etc.
Conclusions

- A tool for Web characterization studies
- Can be extended for other purposes
- Code and documentation available at http://www.cwr.cl/projects/

Thank you.
Conclusions

- A tool for Web characterization studies
- Can be extended for other purposes
- Code and documentation available at http://www.cwr.cl/projects/

Thank you.
Conclusions

- A tool for Web characterization studies
- Can be extended for other purposes
- Code and documentation available at http://www.cwr.cl/projects/

Thank you.
Conclusions

- A tool for Web characterization studies
- Can be extended for other purposes
- Code and documentation available at http://www.cwr.cl/projects/

Thank you.
A tool for Web characterization studies
Can be extended for other purposes
Code and documentation available at http://www.cwr.cl/projects/

Thank you.