iRobot: An Intelligent Crawler for Web Forums

Rui Cai, Jiang-Ming Yang, Wei Lai, Yida Wang, and Lei Zhang
Microsoft Research, Asia
Outline

• Motivation & Challenge
• iRobot – Our Solution
  – System Overview
  – Module Details
• Evaluation
Outline

• Motivation & Challenge
• iRobot – Our Solution
  – System Overview
  – Module Details
• Evaluation
Why Web Forum is Important

• Forum is a huge resource of human knowledge
  – Popular all over the world
  – Contain any conceivable topics and issues

• Forum data can benefit many applications
  – Improve quality of search result
  – Various data mining on forum data

• Collecting forum data
  – Is the basis of all forum related research
  – Is not a trivial task
Why Forum Crawling is Difficult

• Duplicate Pages
  – Forum is with complex in-site structure
  – Many shortcuts for browsing

• Invalid Pages
  – Most forums are with access control
  – Some pages can only be visited after registration

• Page-flipping
  – Long thread is shown in multiple pages
  – Deep navigation levels
The Limitation of Generic Crawlers

• In general crawling, each page is treated independently
  – Fixed crawling depth
  – Cannot avoid duplicates before downloading
  – Fetch lots of invalid pages, such as login prompt
  – Ignore the relationships between pages from a same thread

• Forum crawling needs a site-level perspective!
Statistics on Some Forums

- Around 50% crawled pages are useless
- Waste of both bandwidth and storage
Outline

• Motivation & Challenge

• Our Solution – iRobot
  – System Overview
  – Module Details

• Evaluation
What is Site-Level Perspective?

- Understand the **organization structure**
- Find our an optimal **crawling strategy**

The site-level perspective of "forums.asp.net"
iRobot: An Intelligent Forum Crawler

General Web Crawling

Sitemap Construction

Traversal Path Selection

Forum Crawling

Segmentation & Archiving

Raw Pages

Meta

WebArchive

Restart
• Motivation & Challenge
• Our Solution – iRobot
  – System Overview
  – Module Details
    • How many kinds of pages?
    • How do these pages link with each other?
    • Which pages are valuable?
    • Which links should be followed?
• Evaluation
Page Clustering

- Forum pages are based on **database & template**
- Layout is robust to describe template
  - Repetitive regions are everywhere on forum pages
  - Layout can be characterized by repetitive regions
Page Clustering

Sampling Pages

Repeat Pattern Discovery

Clustering by Virtual Tables
Link Analysis

• URL Pattern can distinguish links, but not reliable on all the sites
• Location can also distinguish links

A Link = URL Pattern + Location
Informativeness Evaluation

• Which kind of pages (nodes) are valuable?
• Some heuristic criteria
  – A larger node is more likely to be valuable
  – Page with large size are more likely to be valuable
  – A diverse node is more likely to be valuable
  • Based on content de-dup

\[ \text{Inf}_{\text{or}}(V_i) = \frac{N_i}{N} \times \frac{S^\text{avg}}{S^\text{avg}} \times \left( 1 - \frac{\hat{N}_i^\text{dup}}{N_i} \right) \]
Traversal Path Selection

- Clean sitemap
  - Remove valueless nodes
  - Remove duplicate nodes
  - Remove links to valueless / duplicate nodes

- Find an optimal path
  - Construct a spanning tree
  - Use depth as cost
    - User browsing behaviors
  - Identify page-flipping links
    - Number, Pre/Next
Outline

• Motivation & Challenge
• iRobot – Our Solution
  – System Overview
  – Module Details
• Evaluation
Evaluation Criteria

- Duplicate ratio
- Invalid ratio
- Coverage ratio
Effectiveness and Efficiency

- **Effectiveness**

- **Efficiency**
Performance vs. Sampled Page#
# Preserved Discussion Threads

<table>
<thead>
<tr>
<th>Forums</th>
<th>Mirrored</th>
<th>Crawled by iRobot</th>
<th>Correctly Recovered</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biketo</td>
<td>1584</td>
<td>1313</td>
<td>1293</td>
</tr>
<tr>
<td>Asp</td>
<td>600</td>
<td>536</td>
<td>536</td>
</tr>
<tr>
<td>Baidu</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Douban</td>
<td>62</td>
<td>60</td>
<td>37</td>
</tr>
<tr>
<td>CQZG</td>
<td>1393</td>
<td>1384</td>
<td>1311</td>
</tr>
<tr>
<td>Tripadvisor</td>
<td>326</td>
<td>272</td>
<td>272</td>
</tr>
<tr>
<td>Hoopchina</td>
<td>2935</td>
<td>2829</td>
<td>2593</td>
</tr>
</tbody>
</table>

- **94.5%** for Mirrored
- **87.6%** for Crawled by iRobot
Conclusions

• An intelligent forum crawler based on site-level structure analysis
  – Identify page templates / valuable pages / link analysis / traversal path selection

• Some modules can still be improved
  – More automated & mature algorithms in SIGIR’08

• More future work directions
  – Queue management
  – Refresh strategies
Thanks!