Microsoft Research

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

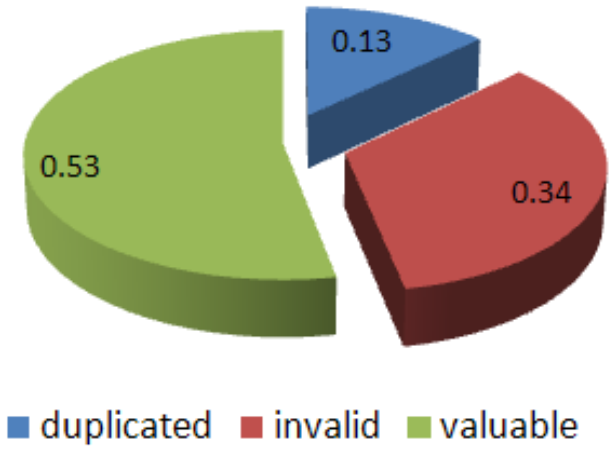
- Duplicate Pages
  - Forum is with complex in-site structure
  - Many shortcuts for browsing
- Invalid Pages
  - Most forums are with access control
  - Some pages can only be visited after registration
- Page-flipping
  - Long thread is shown in multiple pages
  - Deep navigation levels

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

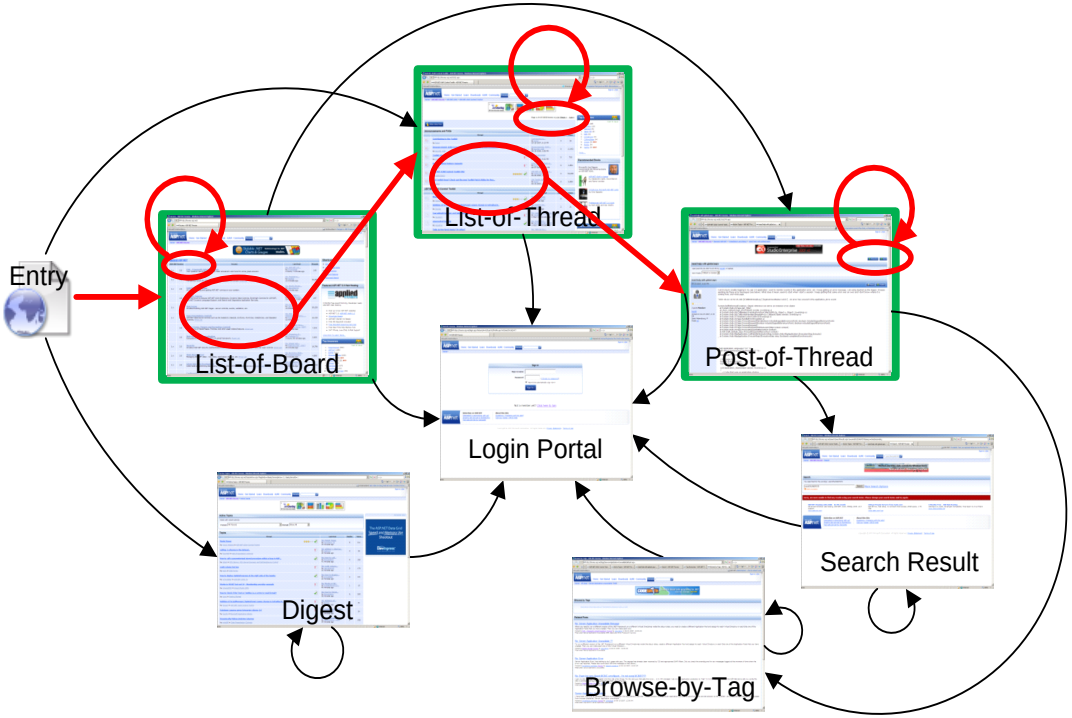


Outline

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

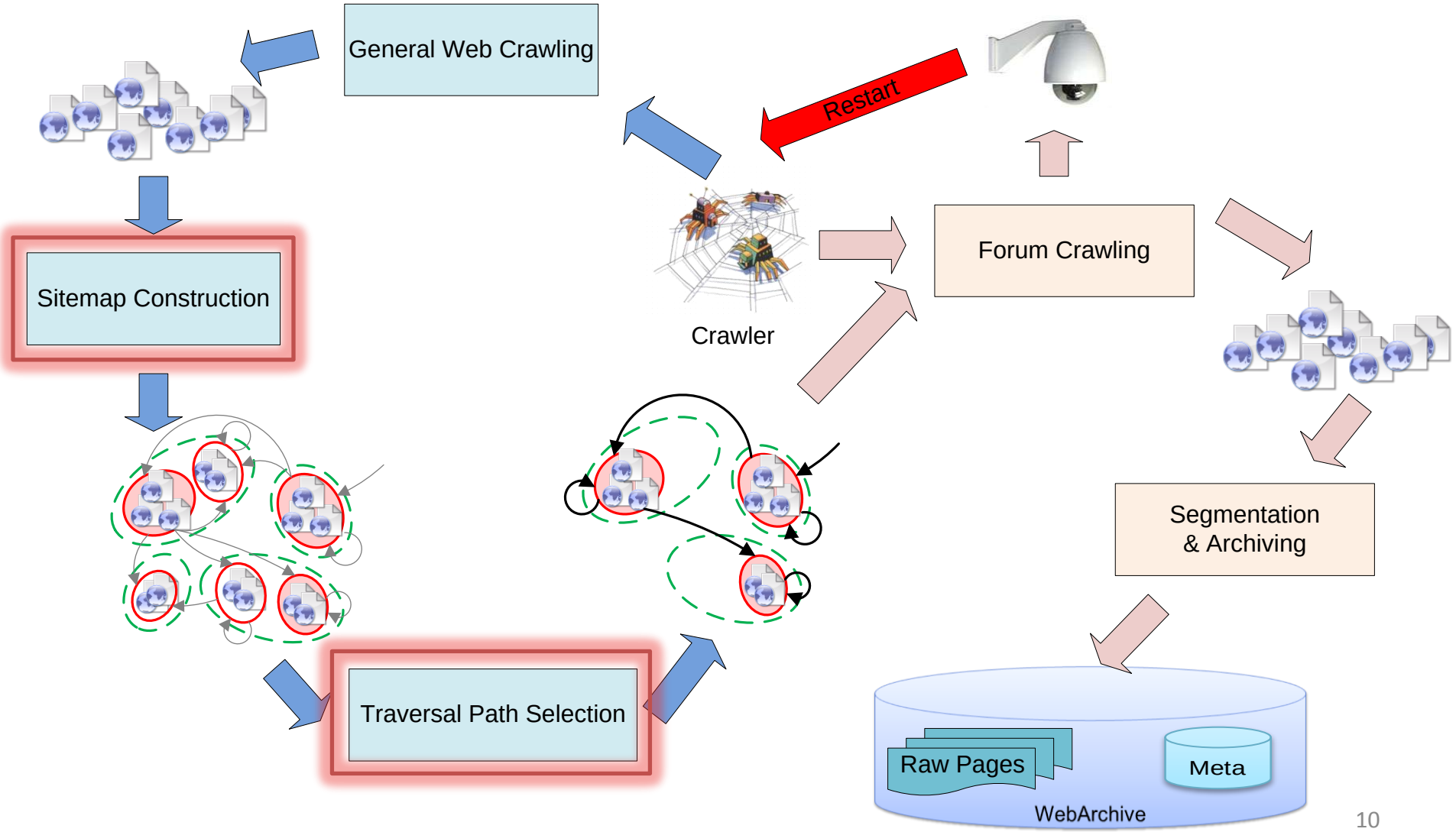
What is Site-Level Perspective?

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



The site-level perspective of "forums.asp.net"

iRobot: An Intelligent Forum Crawler



Outline

- 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

Sitemap Construction

Traversal Path Selection

Page Clustering

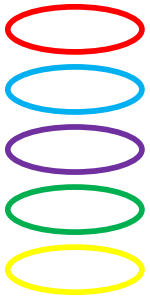
- Forum pages are based on **database** & **template**
- Layout is robust to describe template
  - Repetitive regions are everywhere on forum pages
  - Layout can be characterized by repetitive regions


(a)

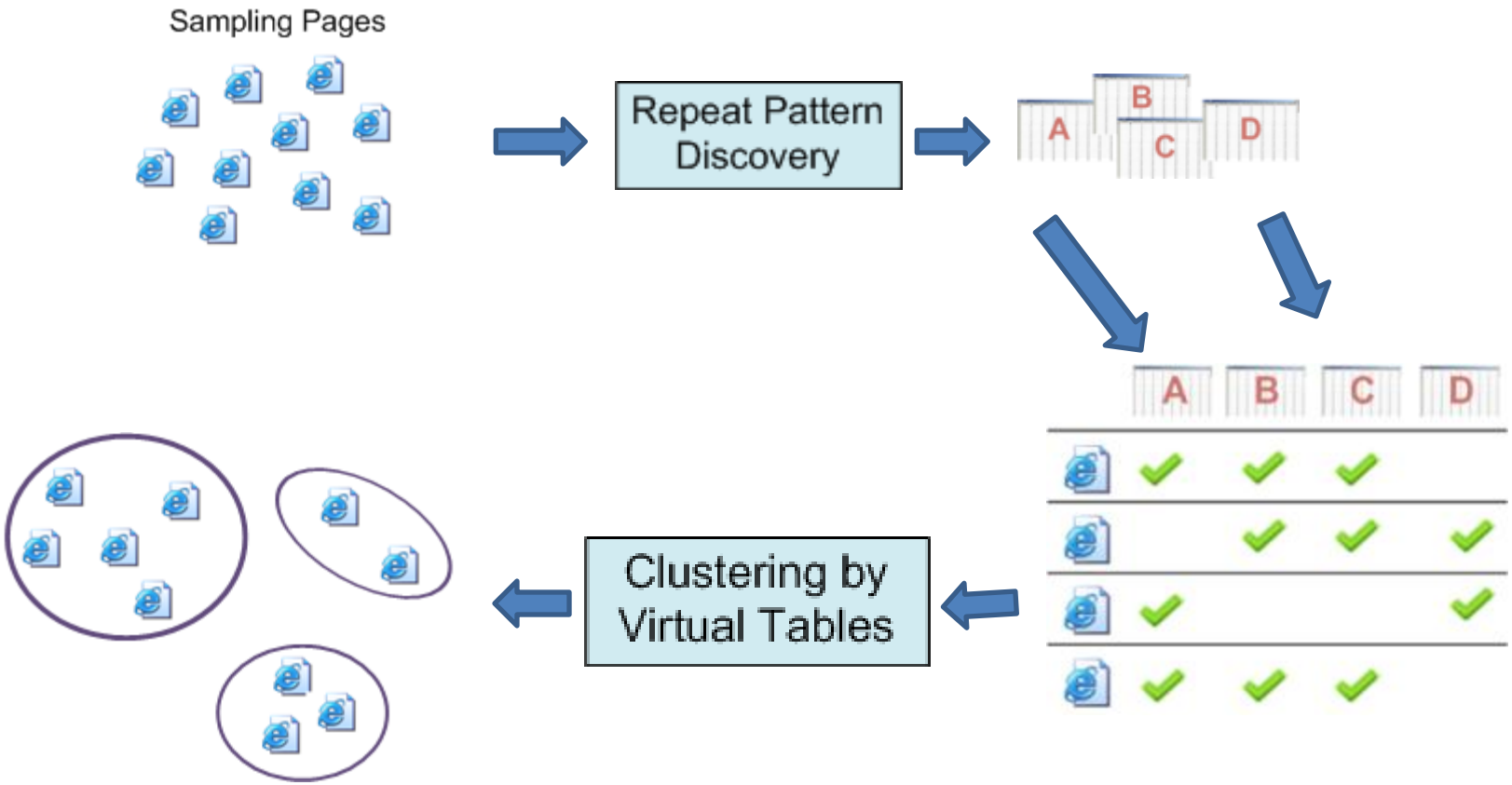

(b)

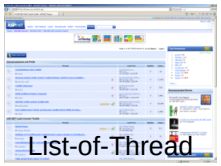

(c)


(d)



Page Clustering



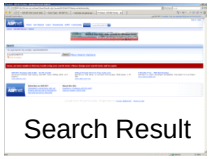


List-of-Thread

List-of-Board

Post-of-Thread

Login Portal



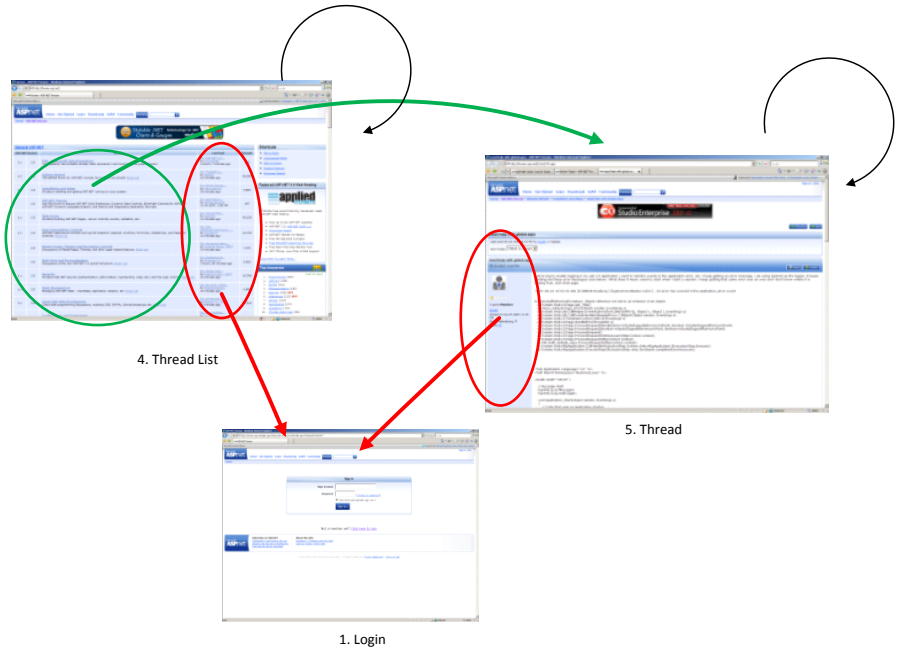
Search Result

Digest

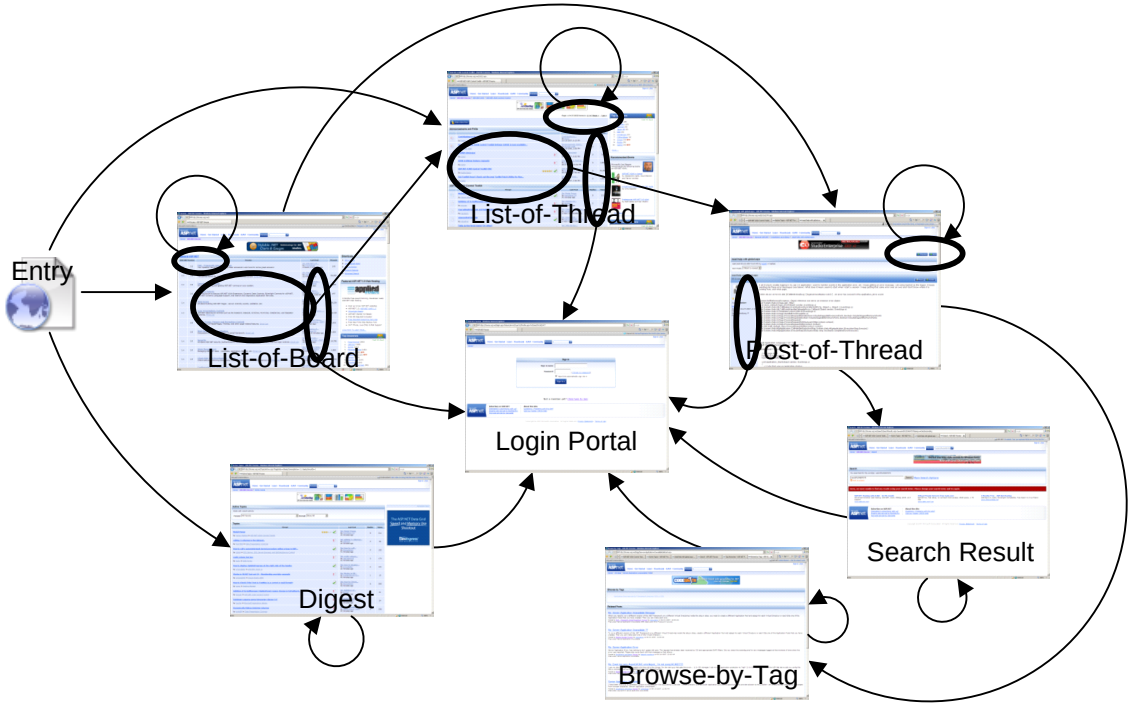
Browse-by-Tag

Link Analysis

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



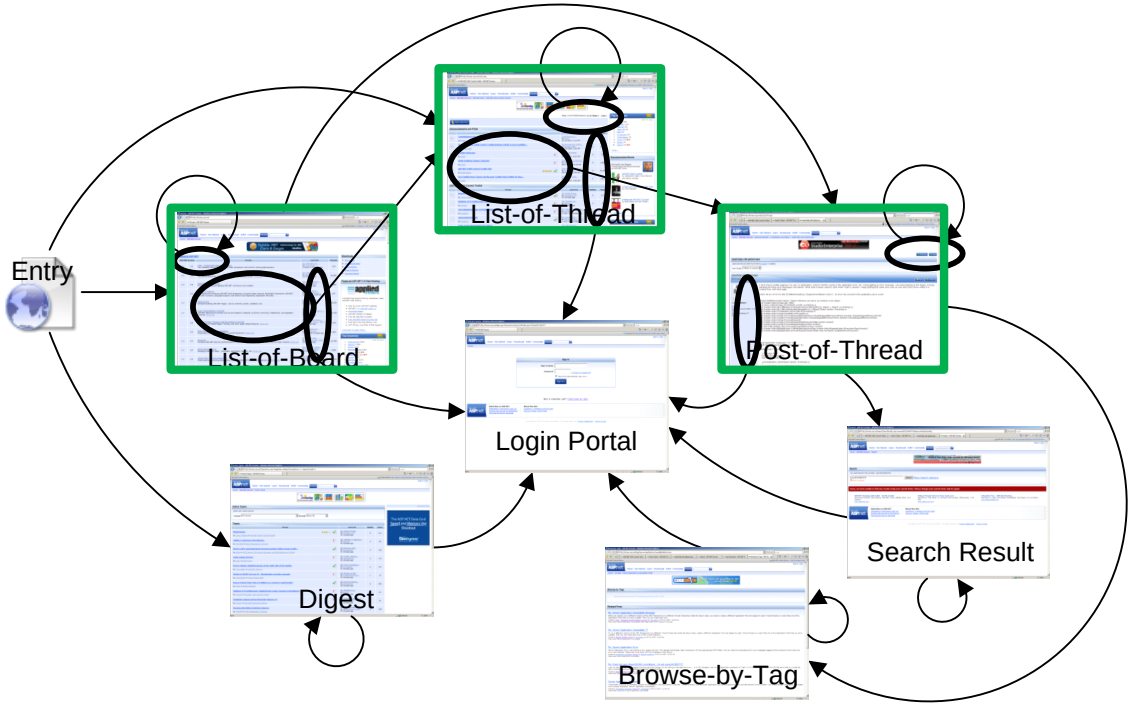
A Link = URL Pattern + Location


Entry
List-of-Board
List-of-Thread
Post-of-Thread
Login Portal
Search Result
Digest
Browse-by-Tag

Informativeness Evaluation

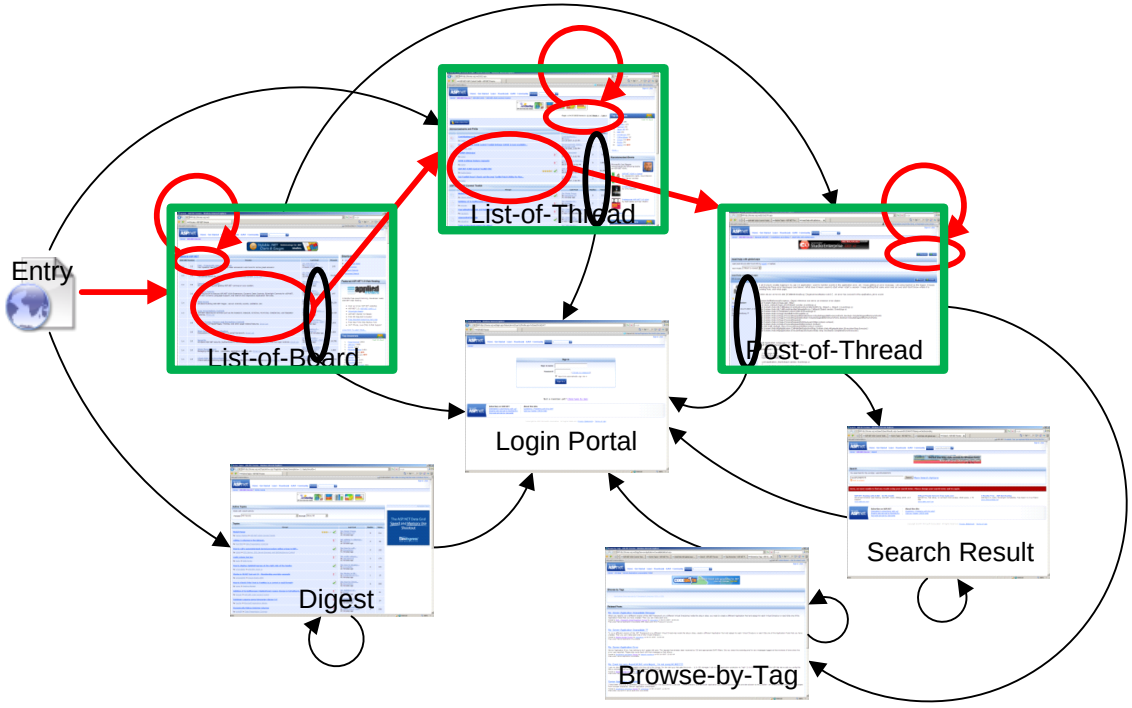
- Which kind of pages (nodes) are valuable?
- Some heuristic criteria
  - A larger node is more like to be valuable
  - Page with large size are more like to be valuable
  - A diverse node is more like to be valuable
    - Based on content de-dup

$$Infor(V_i) = \frac{N_i}{N} \times \frac{S_i^{avg}}{S^{avg}} \times \left(1 - \frac{N_i^{dup}}{N_i}\right)$$


Entry
List-of-Board
List-of-Thread
Post-of-Thread
Login Portal
Search Result
Digest
Browse-by-Tag

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

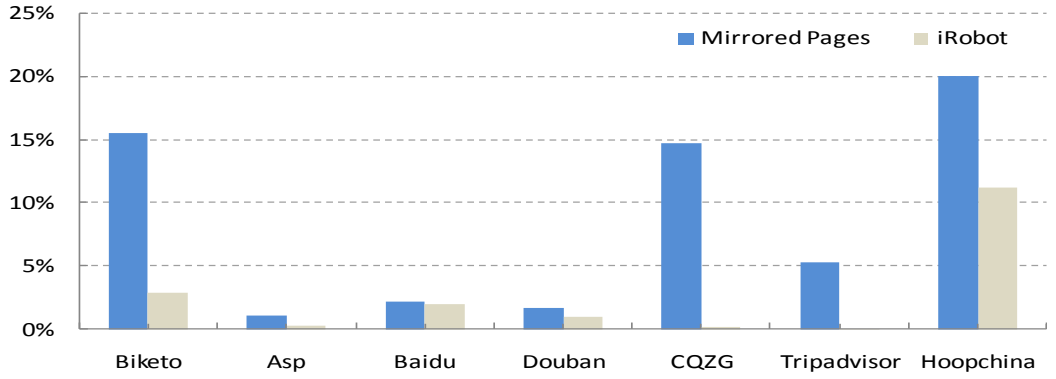

Entry
List-of-Board
List-of-Thread
Post-of-Thread
Login Portal
Digest
Search Result
Browse-by-Tag

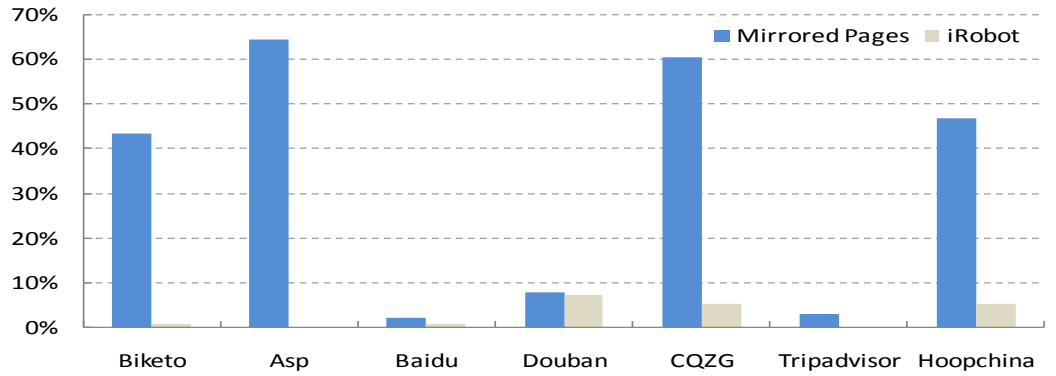
Outline

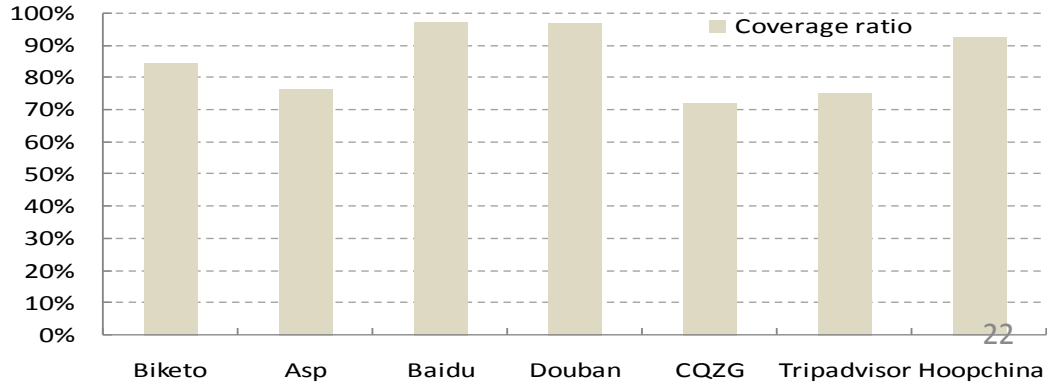
- Motivation & Challenge
- iRobot – Our Solution
  - System Overview
  - Module Details
- Evaluation

Evaluation Criteria

- Duplicate ratio
- Invalid ratio
- Coverage ratio

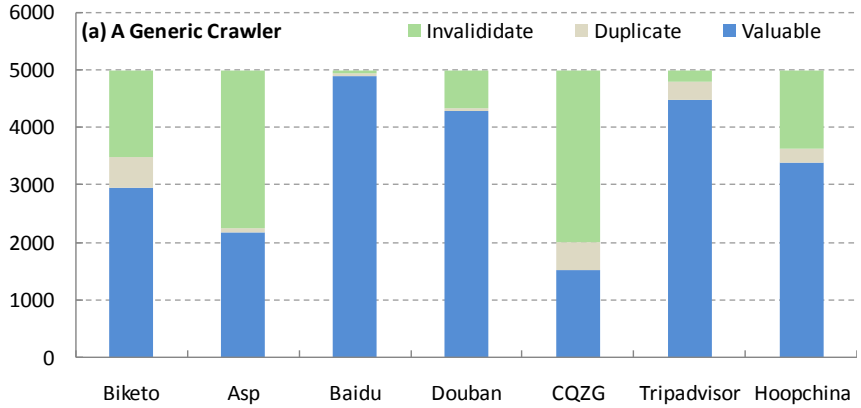


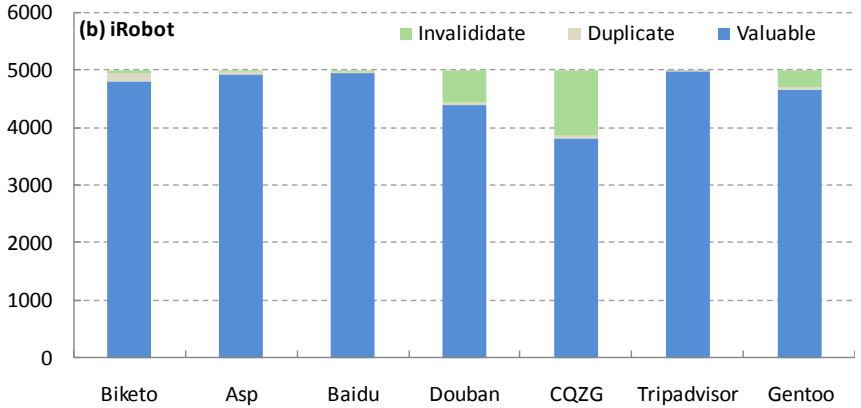




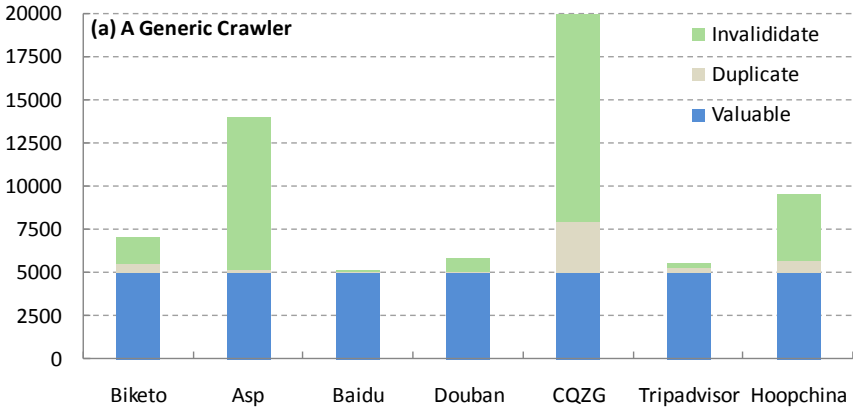
Effectiveness and Efficiency

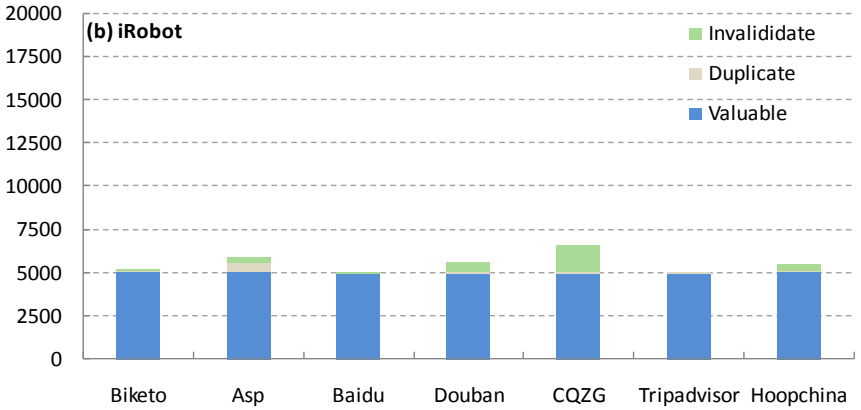
- Effectiveness



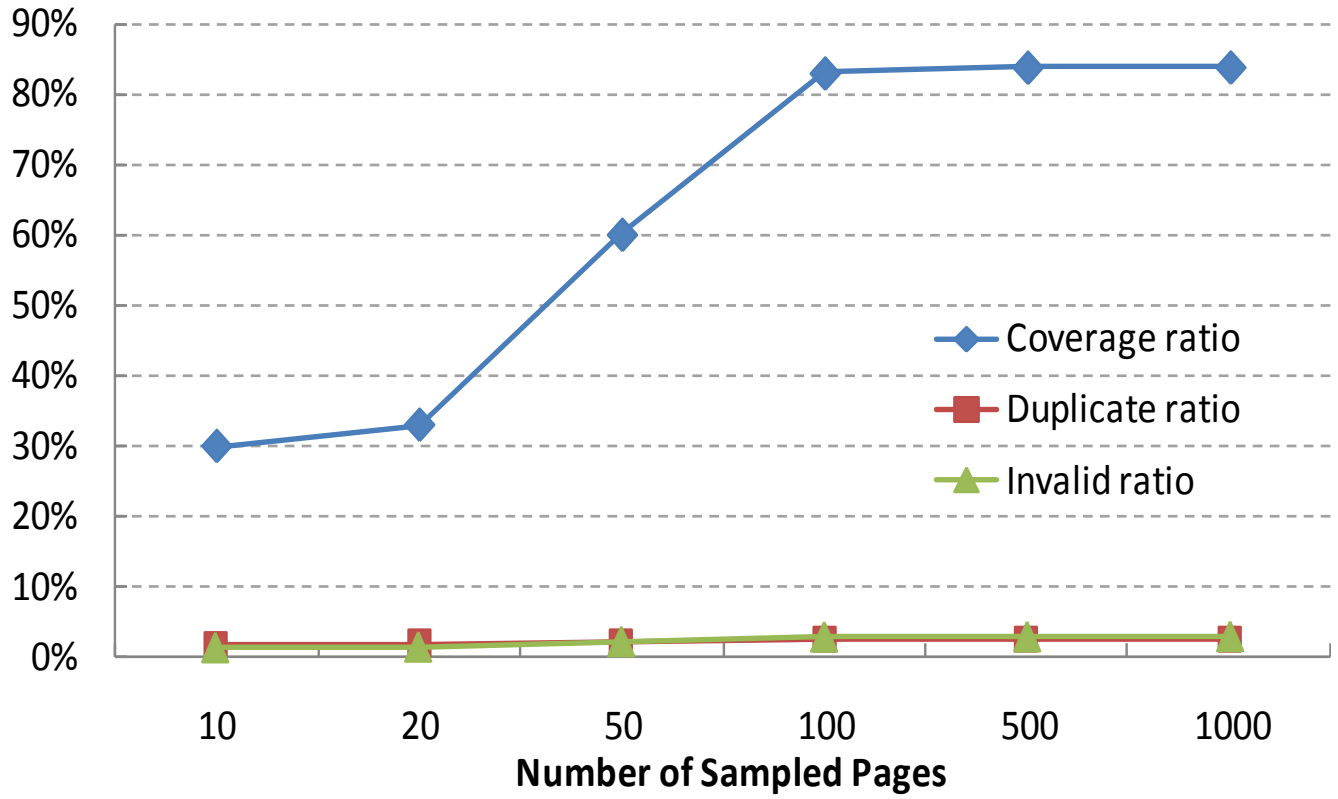


- Efficiency



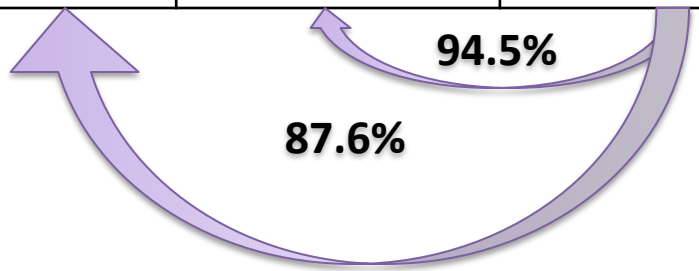


Performance vs. Sampled Page#



Preserved Discussion Threads

Forums	Mirrored	Crawled by iRobot	Correctly Recovered
Biketo	1584	1313	1293
Asp	600	536	536
Baidu	–	–	–
Douban	62	60	37
CQZG	1393	1384	1311
Tripadvisor	326	272	272
Hoopchina	2935	2829	2593



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!