Exploring Traversal Strategy for Web Forum Crawling

Yida Wang, Jiang-Ming Yang, Wei Lai, Rui Cai, Lei Zhang and Wei-Ying Ma

Chinese Academy of Sciences
Microsoft Research, Asia
Outline

• Motivation & Challenge

• Our Solution
  – System Overview
  – Traversal Strategy
    • Skeleton link identification
    • Page-flipping link detection

• Evaluation
Outline

• Motivation & Challenge

• Our Solution
  – System Overview
  – Traversal Strategy
    • Skeleton link identification
    • Page-flipping link detection

• Evaluation
Why Web Forum

• Web forum is a huge resource of human knowledge
  – Over 20% search results are from web forums
  – Leverage the power of users and communities

• Forum sites have complex link structures
  – Many shortcut links
  – Links with permission control
  – Page-flipping links
The Limitation of Generic Crawlers

• In general crawling, each page is treated independently, and each link is treated indiscriminately
  – Lead to more than 50% useless pages
  – Ignore the relationships between pages from a same thread

• Forum crawling needs a site-level perspective and a careful selection of links
Outline

• Motivation & Challenge

• Our Solution
  – System Overview
  – Traversal Strategy
    • Skeleton link identification
    • Page-flipping link detection

• Evaluation
What is Site-Level Perspective?

- Understand the organization structure
- Find our an optimal Traversal strategy

The site-level perspective of "forums.asp.net"
• Adopted a combined strategy of breadth-first and depth-first using a double-ended queue
• Try to cover as many as possible unseen URL Patterns
Random Sampling

• Randomly sample some pages from a given site
• Adopt a combined strategy of breadth-first and depth-first using a double-ended queue
• Try to cover as many as possible unseen URL patterns
• 1,000 pages are enough
Utilized the repetitive regions to characterize the content layout of each page

Represent links with their location and URL patterns
Sitemap Construction

• A sitemap is a directed graph consisting of a set of *vertices* and the corresponding *links*

• Cluster pages into vertices with the same page layout

• Link = its URL pattern + its location

More details about the first two parts, please refer to our previous work: iRobot: An Intelligent Crawler for Web Forums, in WWW’08
Random Sampling

Sitemap Construction

Traversal Strategy Exploring

Crawling

- Skeleton Link Identification
- Page-Flipping Link Detection
Why Skeleton Links

• Crawlers crawl as many as possible unique pages in a given forum site by following skeleton links

• Skeleton links are the most important links supporting the structure of a forum site

• Skeleton links point to all valuable pages without introducing redundant and valueless
Example of skeleton links from forums.asp.net
How to Identify Skeleton Links

• Aim at all unique pages without duplicates

• An optimal set of skeleton links leads to most unique pages and few duplicates

• Search skeleton links for each valuable vertex
  – Level by level: Inspired by user browsing behavior
  – Find an optimal combination of links
    • Optimal result comes out after exhausting all!
• Pruning while searching for optimism
  – Selected but introduce many duplicate pages
  – Rejected but cause coverage drop significantly

An illustration of the search process of skeleton links
Why Page-Flipping Links

• Crawlers can completely download a long discussion thread divided into several pages by following page-flipping links

• Page-flipping links are a kind of *loop-back* links in the sitemap. However, not all loop-back links are page-flipping ones
Example of page-flipping links from forums.asp.net
How to Detect Page-Flipping Links

• For page-flipping links, if there is a path from page A to B, there must be a path follow the same type of links from B to A

• Page-flipping links have larger connectivity score
An illustration of the characteristics of page-flipping links

Connectivity = \frac{\sum_{(A,B)} Path(A,B) \cdot Path(B,A)}{\sum_{(A,B)} Path(A,B)}

Connectivity = \frac{722}{890} = 0.81

Connectivity = \frac{108}{1153} = 0.09
• Mapping a new page to an existing layout vertex

• Follow the traversal strategy for out-links
Crawling

• From the given entry page

• Map a new page to an existing layout vertex

• Follow the explored traversal strategy for out-links from that page
Outline

• Motivation & Challenge

• Our Solution
  – System Overview
  – Traversal Strategy
    • Skeleton link identification
    • Page-flipping link detection

• Evaluation
Experimental Setup

• Contract experiments in eight forums from diverse categories
  – Mirror pages: Crawled by a real commerce crawler
  – Structure-driven: Crawled by structure-driven crawler proposed in SIGIR’06
  – Our method: Crawled by crawler using our traversal strategy
Evaluation Criteria

Informativeness

\[ \text{Info} = -\frac{1}{\log(N)} \sum_{i=1}^{K} \frac{\|D_i\|}{N} \log \left( \frac{\|D_i\|}{N} \right) \]

Coverage

\[ \text{Cov} = \frac{K'}{K} \times 100\% \]
Effectiveness and Efficiency

• Effectiveness

(a) A Generic Crawler
(b) Structure-driven
(c) Our Method
Effectiveness and Efficiency

• Efficiency

(a) A Generic Crawler

(b) Structure-driven

(c) Our Method
Evaluation of Page-Flipping Detection
Conclusions

• A complete solution to automatically explore an appropriate traversal strategy to a given target forum site is proposed
  – Skeleton link identification
  – Page-flipping link detection

• More future work directions
  – Incremental crawling
  – Forum page segmentation
Thanks!