



Web Crawling

Carlos Castillo

Center for Web Research
Computer Science Department
University of Chile
www.cwr.cl

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Motivation

Behavior of a crawler

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Scheduling

- Short-term scheduling
- Long-term scheduling
- When to stop crawling

Architecture

- History
- Classification
- Implementation

Practical issues

Summary

References

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Scheduling

- Short-term scheduling
- Long-term scheduling
- When to stop crawling

Architecture

- History
- Classification
- Implementation

Practical issues

Summary

References

An astronomer watching the sky

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

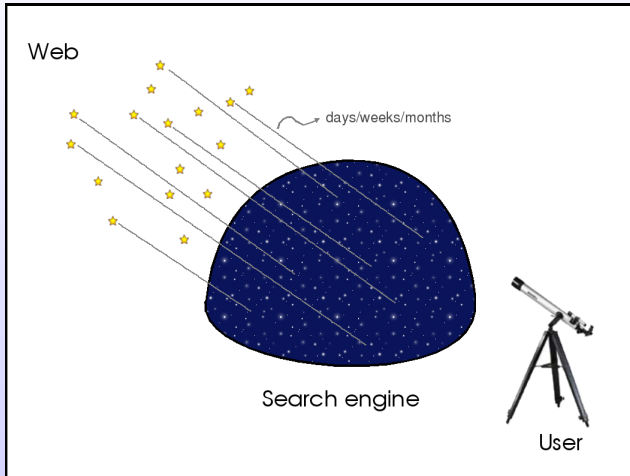
Classification

Implementation

Practical issues

Summary

References



The problem of abundance

- 5 exabytes of new information a year [Lyman and Varian, 2003] (1 exabyte = 10^{18} bytes)
- Most directories no longer encourage administrators to submit their Web sites: they have to find the page on their own
- Adversarial information retrieval

Outline

Motivation

Behavior of a crawler

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Scheduling

- Short-term scheduling
- Long-term scheduling
- When to stop crawling

Architecture

- History
- Classification
- Implementation

Practical issues

Summary

References

The problem of abundance

- 5 exabytes of new information a year [Lyman and Varian, 2003] (1 exabyte = 10^{18} bytes)
- Most directories no longer encourage administrators to submit their Web sites: they have to find the page on their own
- Adversarial information retrieval

Outline

Motivation

Behavior of a crawler

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Scheduling

- Short-term scheduling
- Long-term scheduling
- When to stop crawling

Architecture

- History
- Classification
- Implementation

Practical issues

Summary

References

The problem of abundance

- 5 exabytes of new information a year [Lyman and Varian, 2003] (1 exabyte = 10^{18} bytes)
- Most directories no longer encourage administrators to submit their Web sites: they have to find the page on their own
- Adversarial information retrieval

Outline

Motivation

Behavior of a crawler

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Scheduling

- Short-term scheduling
- Long-term scheduling
- When to stop crawling

Architecture

- History
- Classification
- Implementation

Practical issues

Summary

References

The bandwidth is expensive

“Given that the bandwidth for conducting crawls is neither infinite nor free it is becoming essential to crawl the Web in a not only scalable, but efficient way if some reasonable measure of quality or freshness is to be maintained” [Edwards et al., 2001]

The cost of a “complete” Web crawl is estimated in \$1.5 million USD [Craswell et al., 2004], only considering network usage

Outline

Motivation

Behavior of a crawler

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Scheduling

- Short-term scheduling
- Long-term scheduling
- When to stop crawling

Architecture

- History
- Classification
- Implementation

Practical issues

Summary

References

The bandwidth is expensive

“Given that the bandwidth for conducting crawls is neither infinite nor free it is becoming essential to crawl the Web in a not only scalable, but efficient way if some reasonable measure of quality or freshness is to be maintained” [Edwards et al., 2001]

The cost of a “complete” Web crawl is estimated in \$1.5 million USD [Craswell et al., 2004], only considering network usage

Outline

Motivation

Behavior of a crawler

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Scheduling

- Short-term scheduling
- Long-term scheduling
- When to stop crawling

Architecture

- History
- Classification
- Implementation

Practical issues

Summary

References

Combination of policies

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Combination of policies

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Combination of policies

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Combination of policies

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

It is necessary to prioritize

- No search engine indexes more than 16% of the Web [Lawrence and Giles, 2000]
- Download only the “important” pages
- Restrict to only a sub-domain
- Avoid spamming

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Selection based on links

- Order by Pagerank [Cho et al., 1998]
- Depth-first search [Najork and Wiener, 2001]
- Focused crawling [Chakrabarti et al., 1999], attempting to infer similarity to pages before downloading them

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Events

- **Creation**, which requires a link
- **Update**, can be either minor or major. Most of the changes are minor, but this is not easy to exploit
- **Deletion**, which is more damaging to the search engine's reputation

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Cost functions

- Freshness:

$$F_p(t) = \begin{cases} 1 & \text{if } p \text{ is not modified at time } t \\ 0 & \text{otherwise} \end{cases}$$

- Age:

$$A_p(t) = \begin{cases} 0 & \text{if } p \text{ is not modified} \\ t - \text{lastmod}(p) & \text{otherwise} \end{cases}$$

- Depending on the cost function used, the behavior can be different

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Cost functions

- Freshness:

$$F_p(t) = \begin{cases} 1 & \text{if } p \text{ is not modified at time } t \\ 0 & \text{otherwise} \end{cases}$$

- Age:

$$A_p(t) = \begin{cases} 0 & \text{if } p \text{ is not modified} \\ t - \text{lastmod}(p) & \text{otherwise} \end{cases}$$

- Depending on the cost function used, the behavior can be different

Evolution of freshness and age

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

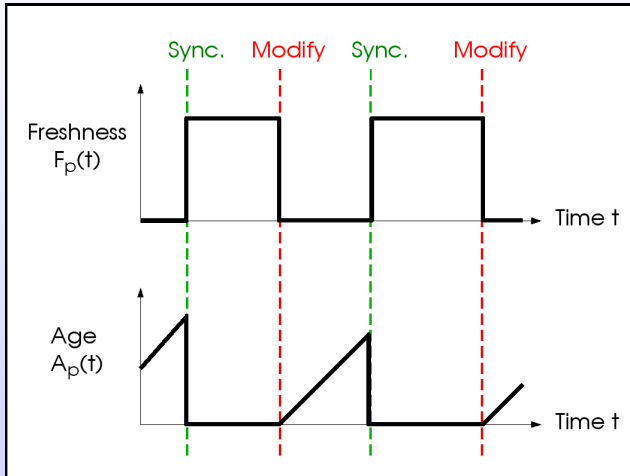
Classification

Implementation

Practical issues

Summary

References



Estimating freshness and age

- Page changes can be modeled as a Poisson process [Brewington et al., 2000]
- Probability of a page being updated at time t is

$$P(F_p(t) = 1) = e^{-\lambda_p t}$$

- λ_p can be estimated using historical data, specially if last-modification date is provided by the server [Cho and Garcia-Molina, 2003]

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Web robots can be a threat

- They consume network resources
- They can cause server overload
- The robot exclusion protocol should be honored [Koster, 1996]
- The re-visiting period should be reasonable (what is reasonable?)

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Robot exclusion

Server exclusions

Disallow: /cgi-bin

Page exclusions

```
<meta name="robots"
  content="noindex,nofollow,nocache">
```

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Robot exclusion

Server exclusions

Disallow: /cgi-bin

Page exclusions

```
<meta name="robots"  
  content="noindex,nofollow,nocache">
```

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Objectives

- Distribute the Web crawling
- Ideally, no central control point
- Reduce overhead due to communications
- Reduce overlap, ideally zero

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Types of policies

- Static assignment: typically a hash function on site names
- Dynamic assignment: more complicated to handle, usually requires central control

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Problem separation

- Indexing, downloading, and distributed crawling are done in batches – this can be exploited to separate the problem
- **Short-term scheduling:** using the network resources efficiently
- **Long-term scheduling:** ordering the crawling process to download important pages first

Outline

Motivation

Behavior of a crawler

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Scheduling

- Short-term scheduling
- Long-term scheduling
- When to stop crawling

Architecture

- History
- Classification
- Implementation

Practical issues

Summary

References

Problem separation

- Indexing, downloading, and distributed crawling are done in batches – this can be exploited to separate the problem
- **Short-term scheduling:** using the network resources efficiently
- **Long-term scheduling:** ordering the crawling process to download important pages first

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Short-term scheduling

If B is the bandwidth available, then B_p , the downloading speed for page p , is

$$B_p = \frac{S_p}{T^*}$$

Where T^* is the optimal time to use all of the available bandwidth

$$T^* = \frac{\sum_p S_p}{B}$$

Outline

Motivation

Behavior of a crawler

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Scheduling

- Short-term scheduling**
- Long-term scheduling
- When to stop crawling

Architecture

- History
- Classification
- Implementation

Practical issues

Summary

References

Full parallelization

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Scheduling

- Short-term scheduling
- Long-term scheduling
- When to stop crawling

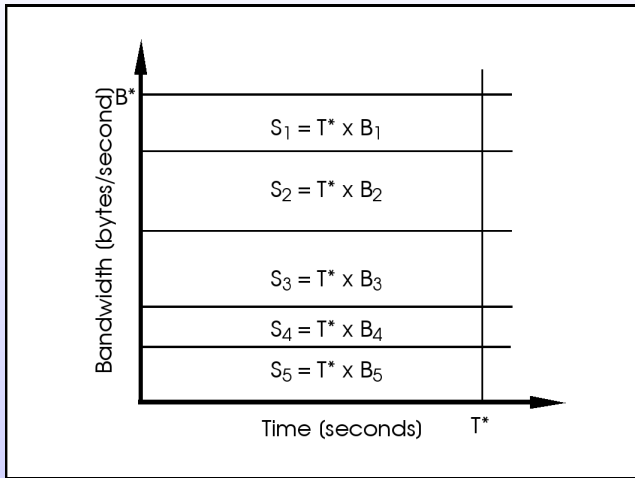
Architecture

- History
- Classification
- Implementation

Practical issues

Summary

References



Full serialization

Outline

Motivation

Behavior of a crawler

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Scheduling

- Short-term scheduling
- Long-term scheduling
- When to stop crawling

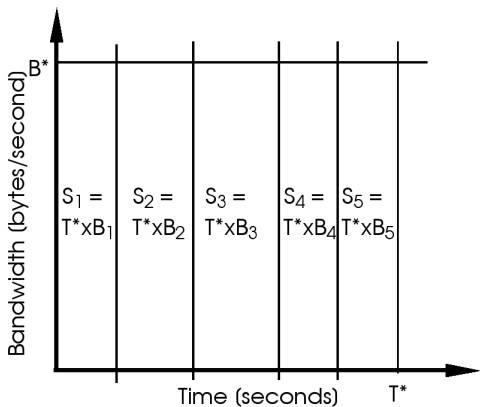
Architecture

- History
- Classification
- Implementation

Practical issues

Summary

References



Realistic scenario

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

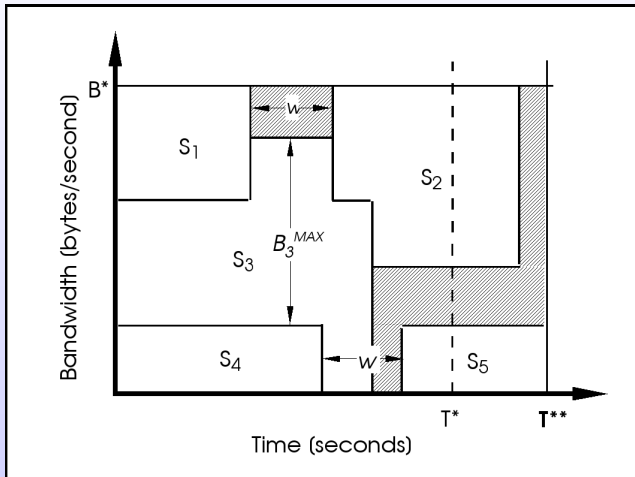
Classification

Implementation

Practical issues

Summary

References



Number of active crawlers

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

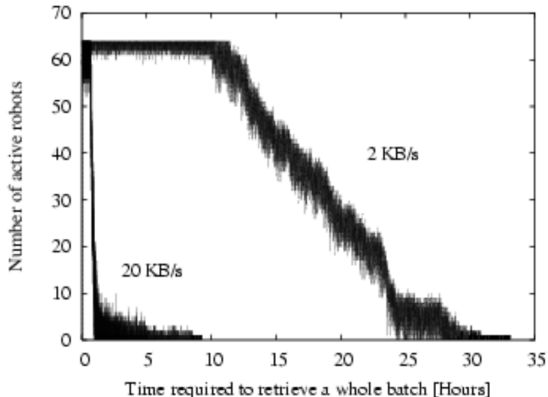
Classification

Implementation

Practical issues

Summary

References



Objective

- Download “important” pages first
- Download X% of the top Y% pages
- Cumulative Pagerank vs fraction of the Web – total Pagerank is 1, random strategy should give a straight line

Outline

Motivation

Behavior of a crawler

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Scheduling

- Short-term scheduling
- Long-term scheduling**
- When to stop crawling

Architecture

- History
- Classification
- Implementation

Practical issues

Summary

References

Objective

- Download “important” pages first
- Download X% of the top Y% pages
- Cumulative Pagerank vs fraction of the Web – total Pagerank is 1, random strategy should give a straight line

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Strategies

- Oracle with Pagerank
- Depth-first search
- Bigger sites first
- Partial pagerank calculations

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Strategies

- Oracle with Pagerank
- Depth-first search
- Bigger sites first
- Partial pagerank calculations

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Strategies

- Oracle with Pagerank
- Depth-first search
- Bigger sites first
- Partial pagerank calculations

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Strategies

- Oracle with Pagerank
- Depth-first search
- Bigger sites first
- Partial pagerank calculations

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

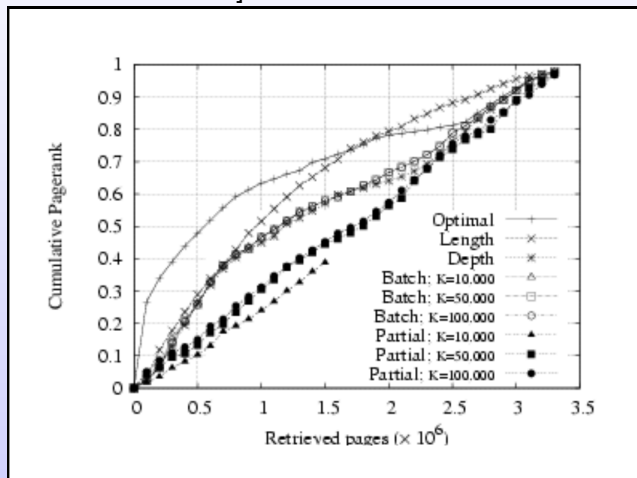
References

Comparison of strategies

Web Crawling

Carlos Castillo

[Castillo et al., 2004]



Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

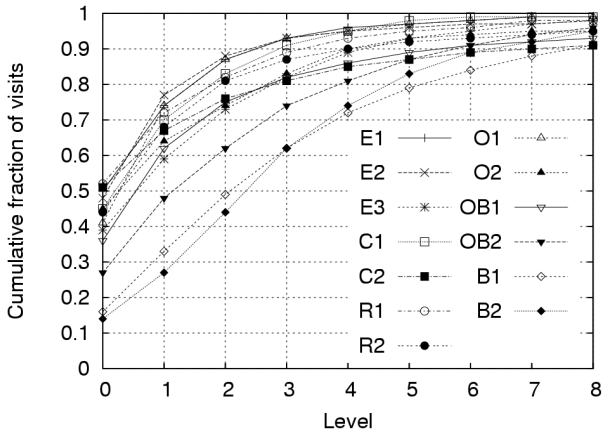
References

Distribution of visits per level

Web Crawling

Carlos Castillo

[Baeza-Yates and Castillo, 2004]



Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

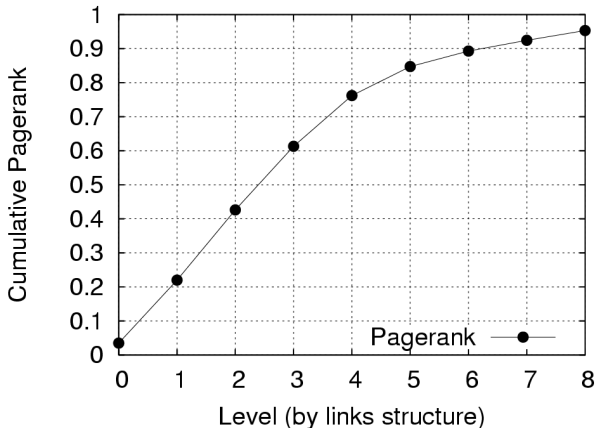
Practical issues

Summary

References

Pagerank and depth

Cumulative Pagerank by levels in the Chilean Web



Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

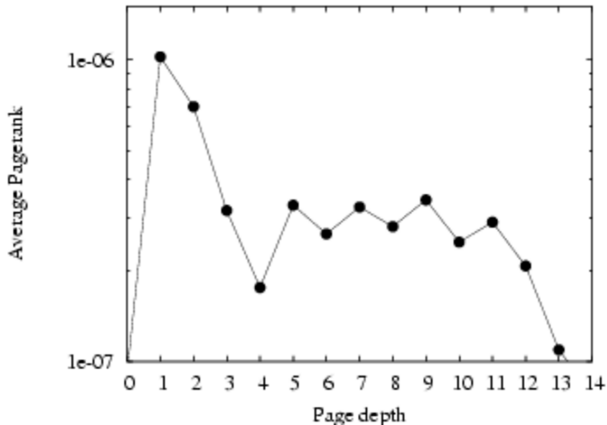
Practical issues

Summary

References

Pagerank and depth

Correlation of Pagerank and depth is low at deeper levels



Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

First crawlers

- RBSE spider - size of the Web: 100,000 pages
- Internet archive crawler - www.archive.org
- Webcrawler - first search engine powered by a Web crawler
- Pages were a scarce resource

Outline

Motivation

Behavior of a crawler

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Scheduling

- Short-term scheduling
- Long-term scheduling
- When to stop crawling

Architecture

History

- Classification
- Implementation

Practical issues

Summary

References

First crawlers

- RBSE spider - size of the Web: 100,000 pages
- Internet archive crawler - www.archive.org
- Webcrawler - first search engine powered by a Web crawler
- Pages were a scarce resource

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Second generation

- Mercator, SPHINX - focused crawling
- Lycos, Excite, Google - large-scale crawling
- Parallel crawlers
- Problem of abundance

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Second generation

- Mercator, SPHINX - focused crawling
- Lycos, Excite, Google - large-scale crawling
- Parallel crawlers
- Problem of abundance

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Standard architecture

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

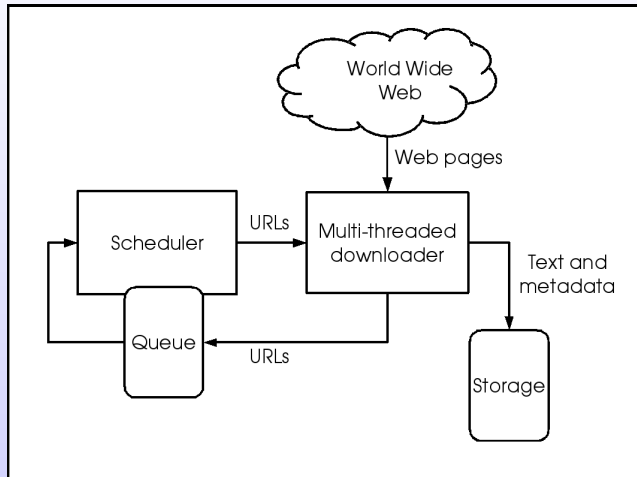
Classification

Implementation

Practical issues

Summary

References



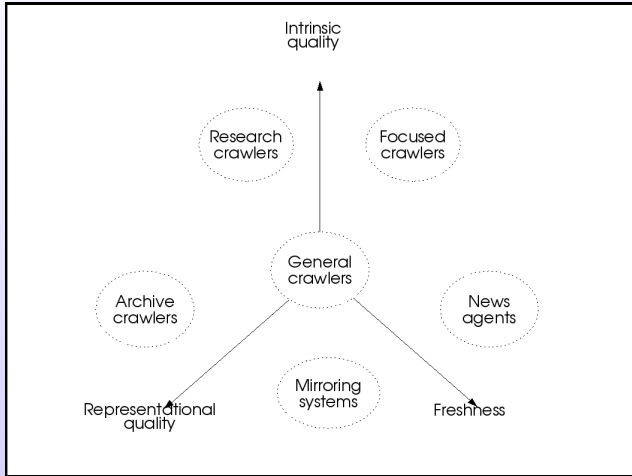
Different crawlers have different focus

- Different issues
- **Quality**: having “good resources”
- **Representation**: having complete copies
- **Freshnes**: having updated copies
- A global-scale crawler tries to balance them all

Taxonomy of Web crawlers

Web Crawling

Carlos Castillo



Outline

Motivation

Behavior of a crawler

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Scheduling

- Short-term scheduling
- Long-term scheduling
- When to stop crawling

Architecture

- History
- Classification
- Implementation

Practical issues

Summary

References

Key operations

- Have I seen this URL ?
- Have I seen this page (or a very similar one) ?
- Which pages should I download next ?
- Store this page
- Download the batch of pages

Outline

Motivation

Behavior of a crawler

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Scheduling

- Short-term scheduling
- Long-term scheduling
- When to stop crawling

Architecture

- History
- Classification
- Implementation**

Practical issues

Summary

References

Key operations

- Have I seen this URL ?
- Have I seen this page (or a very similar one) ?
- Which pages should I download next ?
- Store this page
- Download this batch of pages

Outline

Motivation

Behavior of a crawler

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Scheduling

- Short-term scheduling
- Long-term scheduling
- When to stop crawling

Architecture

- History
- Classification
- Implementation**

Practical issues

Summary

References

Key operations

- Have I seen this URL ?
- Have I seen this page (or a very similar one) ?
- Which pages should I download next ?
- Store this page
- Download this batch of pages

Outline

Motivation

Behavior of a crawler

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Scheduling

- Short-term scheduling
- Long-term scheduling
- When to stop crawling

Architecture

- History
- Classification
- Implementation**

Practical issues

Summary

References

Key operations

- Have I seen this URL ?
- Have I seen this page (or a very similar one) ?
- Which pages should I download next ?
- Store this page
- Download this batch of pages

Outline

Motivation

Behavior of a crawler

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Scheduling

- Short-term scheduling
- Long-term scheduling
- When to stop crawling

Architecture

- History
- Classification
- Implementation**

Practical issues

Summary

References

Key operations

- Have I seen this URL ?
- Have I seen this page (or a very similar one) ?
- Which pages should I download next ?
- Store this page
- Download this batch of pages

Outline

Motivation

Behavior of a crawler

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Scheduling

- Short-term scheduling
- Long-term scheduling
- When to stop crawling

Architecture

- History
- Classification
- Implementation**

Practical issues

Summary

References

The architecture needs to be highly optimized

*“While it is fairly easy to build a slow crawler that downloads a few pages per second for a short period of time, building a high-performance system that can download hundreds of millions of pages over several weeks presents a number of challenges in system design, I/O and network efficiency, and robustness and manageability”
[Shkapenyuk and Suel, 2002].*

Outline

Motivation

Behavior of a crawler

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Scheduling

- Short-term scheduling
- Long-term scheduling
- When to stop crawling

Architecture

- History
- Classification
- Implementation

Practical issues

Summary

References

Problems arise in large crawls

- Network and protocol problems
- Page contents problems
- Server problems

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Network and protocol problems

- Variable quality of service
- Misconfigured firewalls
- Crashing DNS servers
- Wrong DNS servers pointing to good hosts

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Server problems

- Responses lacking headers
- Fancy “error” pages
- “Deeep Web” pages which could be accessible otherwise
- Embedded session-ids in URLs

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Page contents problems

- High prevalence of duplicates
- Browsers are very tolerant
- Malformed markup
- Physical over logical formatting

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Summary

- Web crawling is studied at multiple levels
- Long-term scheduling, page selection
- Scalability, parallelization
- Practical issues, network usage

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Summary

- Web crawling is studied at multiple levels
- Long-term scheduling, page selection
- Scalability, parallelization
- Practical issues, network usage

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Summary

- Web crawling is studied at multiple levels
- Long-term scheduling, page selection
- Scalability, parallelization
- Practical issues, network usage

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Summary

- Web crawling is studied at multiple levels
- Long-term scheduling, page selection
- Scalability, parallelization
- Practical issues, network usage

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Open problems

- Scheduling using historical information
- Exploiting the Web's structure
- Adversarial IR: Spam detection **before** downloading the pages

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Open problems

- Scheduling using historical information
- Exploiting the Web's structure
- Adversarial IR: Spam detection **before** downloading the pages

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Open problems

- Scheduling using historical information
- Exploiting the Web's structure
- Adversarial IR: Spam detection **before** downloading the pages

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Outline

Motivation

Behavior of a crawler

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Scheduling

- Short-term scheduling
- Long-term scheduling
- When to stop crawling


Architecture


- History
- Classification
- Implementation


Practical issues

Summary

References

 Baeza-Yates, R. and Castillo, C. (2004).
Crawling the infinite Web: five levels are enough.
In *Proceedings of the third Workshop on Web
Graphs (WAW)*, volume 3243 of *Lecture Notes in
Computer Science*, pages 156–167, Rome, Italy.
Springer.

 Brewington, B., Cybenko, G., Stata, R., Bharat,
K., and Maghoul, F. (2000).
How dynamic is the web?
In *Proceedings of the Ninth Conference on World
Wide Web*, pages 257 – 276, Amsterdam,
Netherlands.

 Castillo, C., Marin, M., Rodriguez, A., and
Baeza-Yates, R. (2004).
Scheduling algorithms for Web crawling.

Outline

Motivation

Behavior of a crawler

- Selection policy
- Re-visit policy
- Politeness policy
- Parallelization policy

Scheduling

- Short-term scheduling
- Long-term scheduling
- When to stop crawling

Architecture

- History
- Classification
- Implementation

Practical issues

Summary

References

In *Latin American Web Conference (WebMedia/LA-WEB)*, Riberao Preto, Brazil.


IEEE CS Press.


(To appear).

 Chakrabarti, S., van den Berg, M., and Dom, B. (1999).

Focused crawling: a new approach to topic-specific web resource discovery.

Computer Networks, 31(11–16):1623–1640.

 Cho, J. and Garcia-Molina, H. (2003).
Estimating frequency of change.
ACM Transactions on Internet Technology, 3(3).

 Cho, J., García-Molina, H., and Page, L. (1998).
Efficient crawling through URL ordering.
In *Proceedings of the seventh conference on World Wide Web*, Brisbane, Australia.

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

 Craswell, N., Crimmins, F., Hawking, D., and Moffat, A. (2004).

Performance and cost tradeoffs in web search.

In *Proceedings of the 15th Australasian Database Conference*, pages 161–169, Dunedin, New Zealand.

 Edwards, J., McCurley, K. S., and Tomlin, J. A. (2001).

An adaptive model for optimizing performance of an incremental web crawler.

In *Proceedings of the Tenth Conference on World Wide Web*, pages 106–113, Hong Kong. Elsevier Science.

 Koster, M. (1996).

A standard for robot exclusion.

<http://www.robotstxt.org/wc/exclusion.html>.

 Lawrence, S. and Giles, C. L. (2000).

Accessibility of information on the web.

Intelligence, 11(1):32–39.



Lyman, P. and Varian, H. R. (2003).

How much information.

<http://www.sims.berkeley.edu/how-much-info-2003>.



Najork, M. and Wiener, J. L. (2001).

Breadth-first crawling yields high-quality pages.

In *Proceedings of the Tenth Conference on World Wide Web*, pages 114–118, Hong Kong. Elsevier Science.



Shkapenyuk, V. and Suel, T. (2002).

Design and implementation of a high-performance distributed web crawler.

In *Proceedings of the 18th International Conference on Data Engineering (ICDE)*, pages 357 – 368, San Jose, California. IEEE CS Press.

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

Selection policy

Re-visit policy

Politeness policy

Parallelization policy

Scheduling

Short-term scheduling

Long-term scheduling

When to stop crawling

Architecture

History

Classification

Implementation

Practical issues

Summary

References

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

- Selection policy

- Re-visit policy

- Politeness policy

- Parallelization policy

Scheduling

- Short-term scheduling

- Long-term scheduling

- When to stop crawling

Architecture

- History

- Classification

- Implementation

Practical issues

Summary

References

Web Crawling

Carlos Castillo

Outline

Motivation

Behavior of a crawler

- Selection policy

- Re-visit policy

- Politeness policy

- Parallelization policy

Scheduling

- Short-term scheduling

- Long-term scheduling

- When to stop crawling

Architecture

- History

- Classification

- Implementation

Practical issues

Summary

References