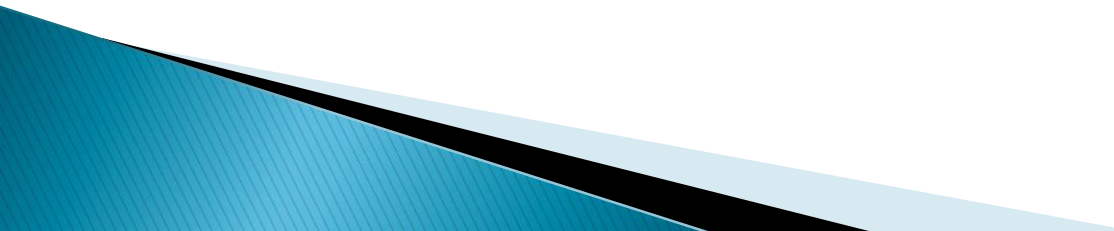


Crawling the Web

- Web pages
 - Few thousand characters long
 - Served through the internet using the hypertext transport protocol (HTTP)
 - Viewed at client end using `browsers`
- Crawler
 - To fetch the pages to the computer
 - At the computer
 - ◆ Automatic programs can analyze hypertext documents

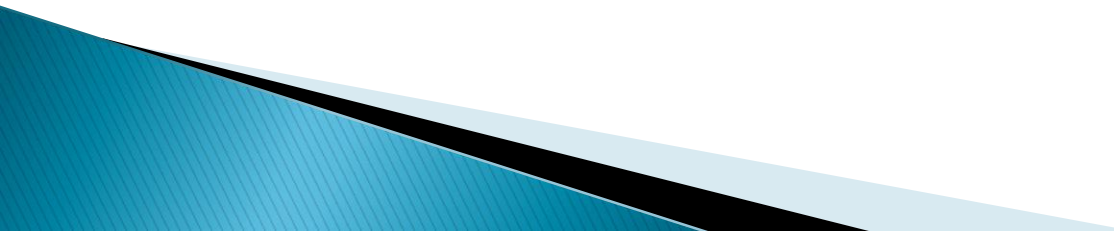
Utilities of a crawler

- ▶ Web crawler, spider.
 - ▶ Definition:
 - A Web crawler is a computer program that browses the World Wide Web in a methodical, automated manner. (Wikipedia)
 - ▶ Utilities:
 - Gather pages from the Web.
 - Support a search engine, perform data mining and so on.
 - ▶ Object:
 - Text, video, image and so on.
 - Link structure.
- 

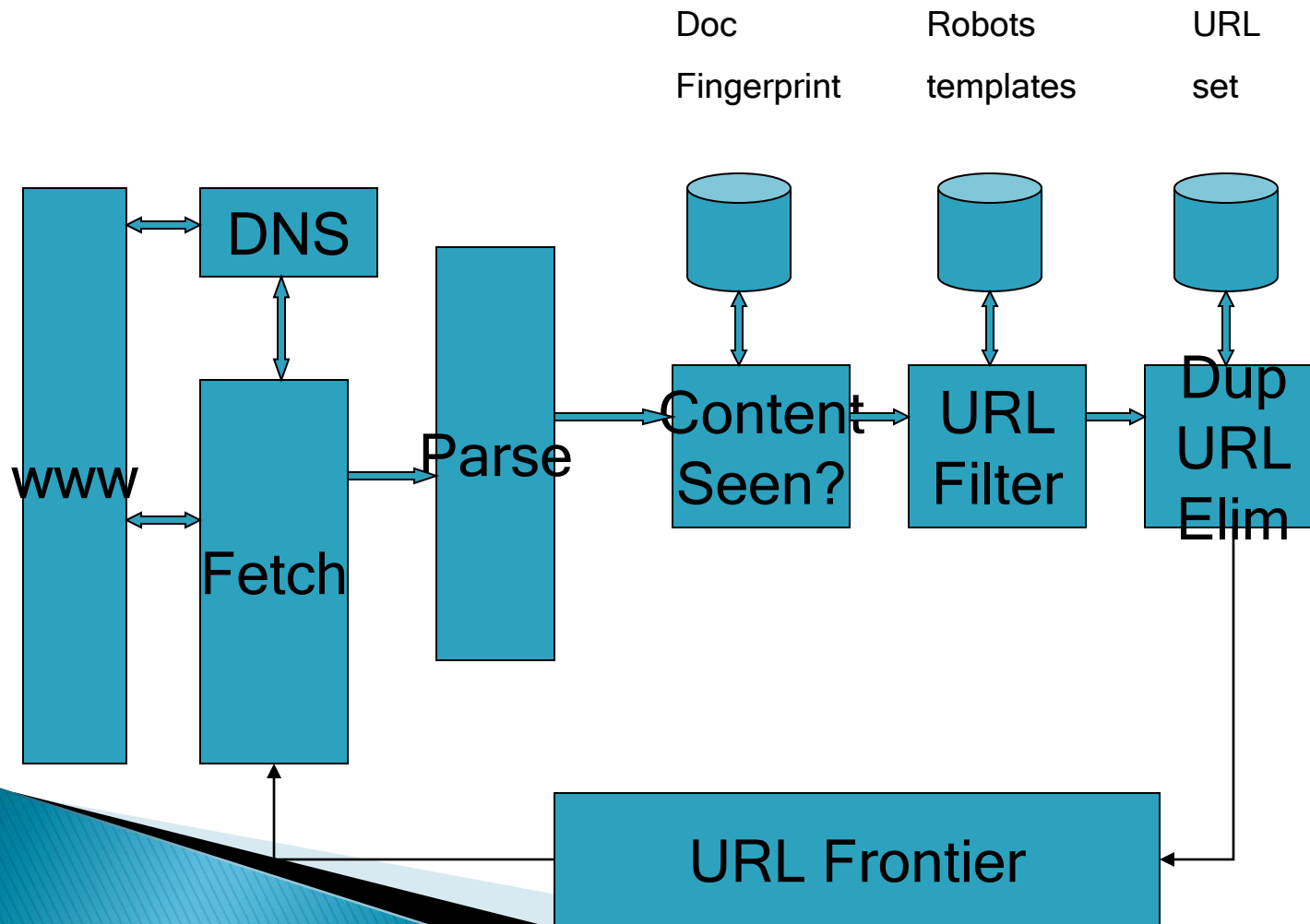
Features of a crawler

- ▶ Must provide:
 - Robustness: spider traps
 - Infinitely deep directory structures:
<http://foo.com/bar/foo/bar/foo/...>
 - Pages filled a large number of characters.
 - Politeness: which pages can be crawled, and which cannot
 - robots exclusion protocol: robots.txt
 - <http://blog.sohu.com/robots.txt>
 - User-agent: *
 - Disallow: /manage/

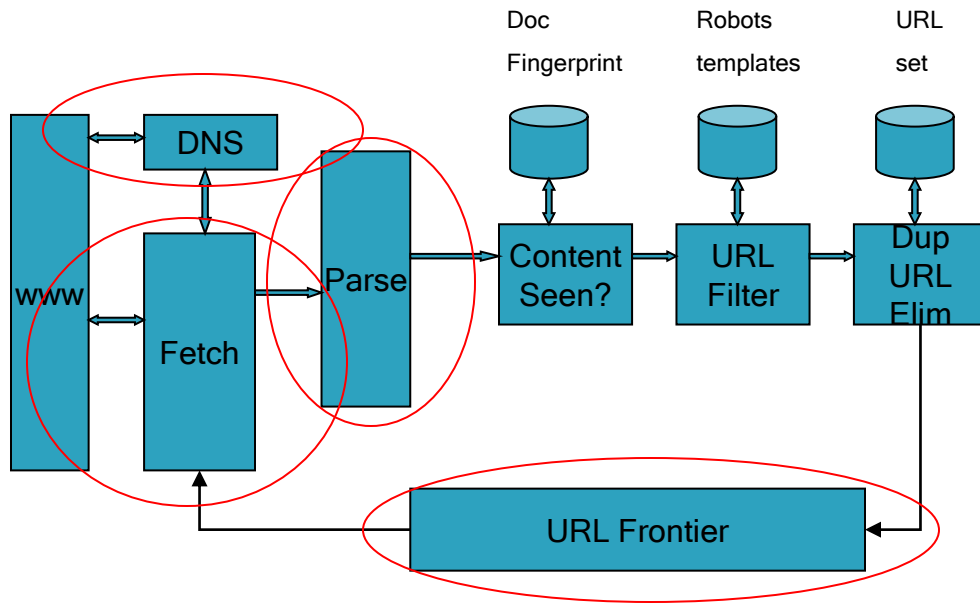
Features of a crawler (Cont'd)

- ▶ Should provide:
 - Distributed
 - Scalable
 - Performance and efficiency
 - Quality
 - Freshness
 - Extensible
- 

Architecture of a crawler

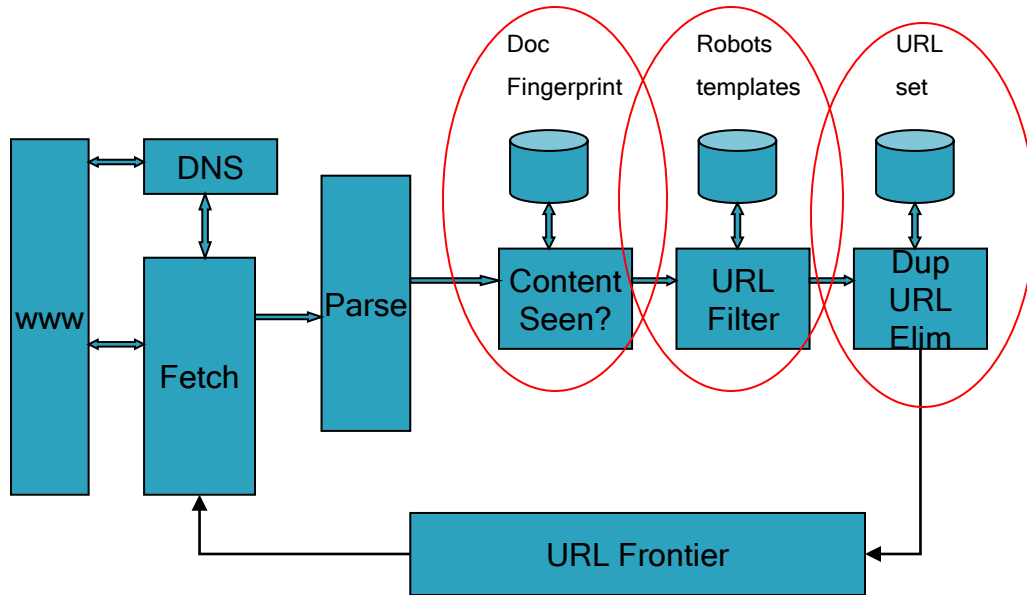


Architecture of a crawler (Cont'd)



- **URL Frontier:** containing URLs yet to be fetched in the current crawl. At first, a seed set is stored in URL Frontier, and a crawler begins by taking a URL from the seed set.
- **DNS:** domain name service resolution. Look up IP address for domain names.

Architecture of a crawler (Cont'd)



- Content Seen?: test whether a web page with the same content has already been seen at another URL. Need to develop a way to measure the fingerprint of a web page.
- URL Filter:
 - Whether the extracted URL should be excluded from the frontier (robots.txt).
 - URL should be normalized (relative encoding).
 - `en.wikipedia.org/wiki/Main_Page`
 - `Disclaimers`
- Dup URL Elim: the URL is checked for duplicate elimination.

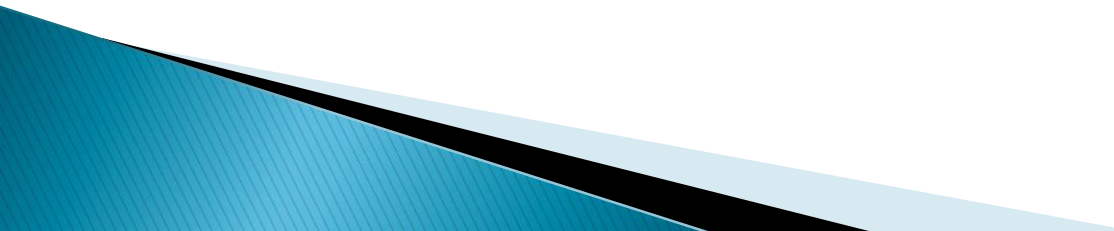
Architecture of a crawler (Cont'd)

- ▶ Other issues:
 - Housekeeping tasks:
 - Log crawl progress statistics: URLs crawled, frontier size, etc. (Every few seconds)
 - Checkpointing: a snapshot of the crawler's state (the URL frontier) is committed to disk. (Every few hours)
 - Priority of URLs in URL frontier:
 - Change rate.
 - Quality.
 - Politeness:
 - Avoid repeated fetch requests to a host within a short time span.
 - Otherwise: blocked ☹️

Crawl “all” Web pages?

- ▶ Problem: no catalog of all accessible URLs on the Web.
- ▶ Solution:
 - start from a given set of URLs
 - Progressively fetch and scan them for new outlinking URLs
 - fetch these pages in turn.....
 - Submit the text in page to a text indexing system
 - and so on.....

Utilities of a crawler

- ▶ Web crawler, spider.
 - ▶ Definition:
 - A Web crawler is a computer program that browses the World Wide Web in a methodical, automated manner. (Wikipedia)
 - ▶ Utilities:
 - Gather pages from the Web.
 - Support a search engine, perform data mining and so on.
 - ▶ Object:
 - Text, video, image and so on.
 - Link structure.
- 

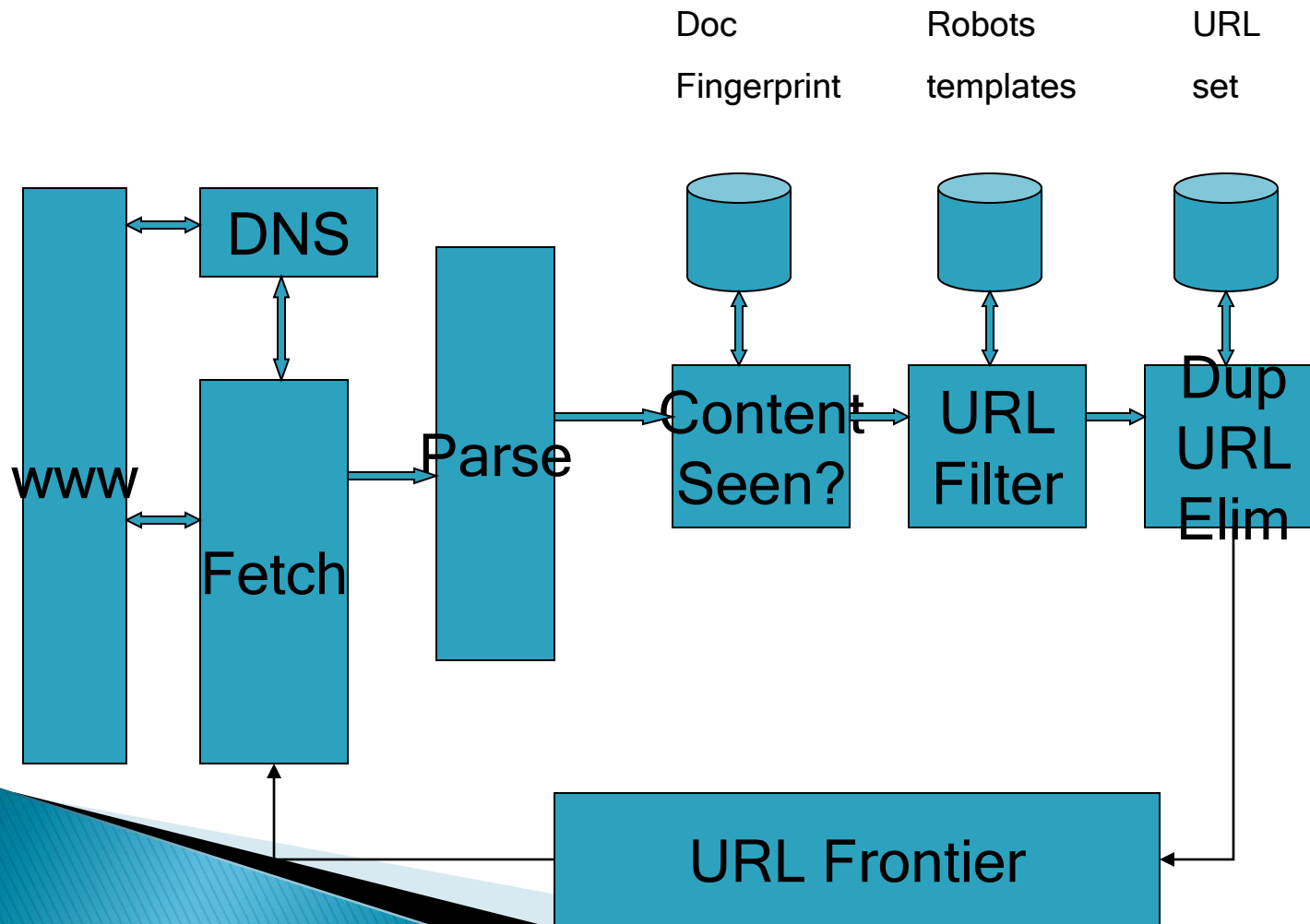
Features of a crawler

- ▶ Must provide:
 - Robustness: spider traps
 - Infinitely deep directory structures:
<http://foo.com/bar/foo/bar/foo/...>
 - Pages filled a large number of characters.
 - Politeness: which pages can be crawled, and which cannot
 - robots exclusion protocol: robots.txt
 - <http://blog.sohu.com/robots.txt>
 - User-agent: *
 - Disallow: /manage/

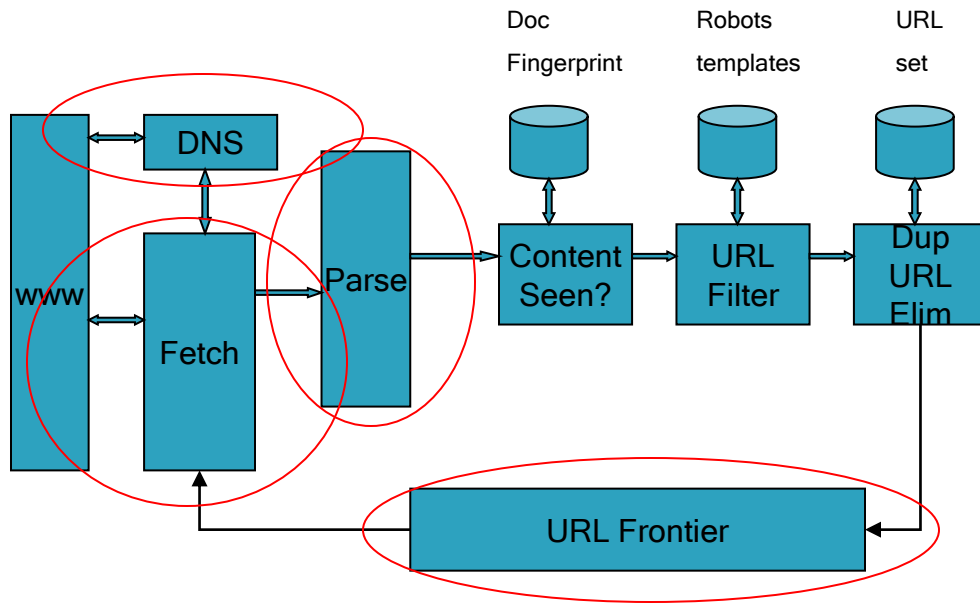
Features of a crawler (Cont'd)

- ▶ Should provide:
 - Distributed
 - Scalable
 - Performance and efficiency
 - Quality
 - Freshness
 - Extensible

Architecture of a crawler



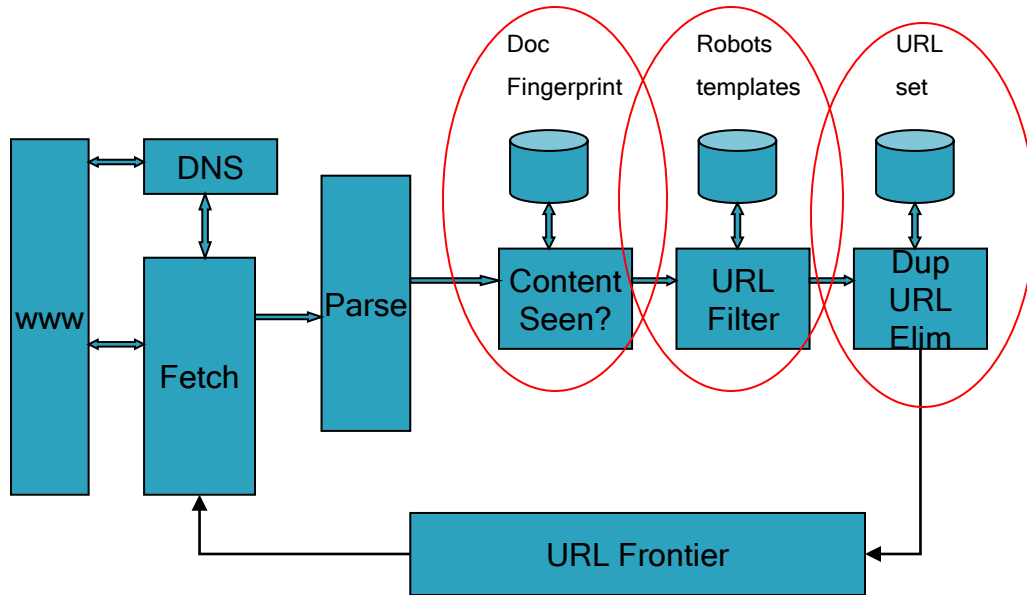
Architecture of a crawler (Cont'd)



- **URL Frontier:** containing URLs yet to be fetched in the current crawl. At first, a seed set is stored in URL Frontier, and a crawler begins by taking a URL from the seed set.
- **DNS:** domain name service resolution. Look up IP address for domain names.

Fetch the URL from the URL Frontier

Architecture of a crawler (Cont'd)



- Content Seen?: test whether a web page with the same content has already been seen at another URL. Need to develop a way to measure the fingerprint of a web page.
- URL Filter:
 - Whether the extracted URL should be excluded from the frontier (robots.txt).
 - URL should be normalized (relative encoding).
 - `en.wikipedia.org/wiki/Main_Page`
 - `Disclaimers`
- Dup URL Elim: the URL is checked for duplicate elimination.