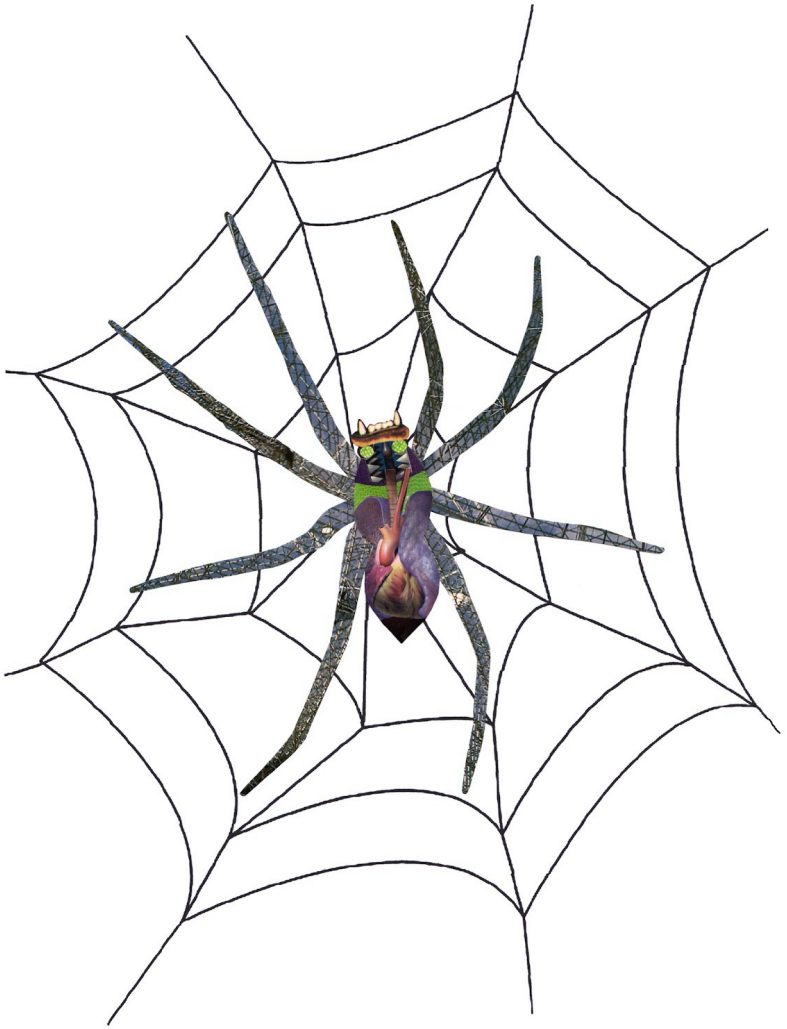
Web Crawler



Web Crawling

crawler, wanderer, robot, spider, fish, worm

Web Crawling

- Zadanie
- Motywacja
- Zasada działania
- Problemy

Problemy

- Autoryzacja, weryfikacja
- TLDR
- Strony dynamiczne
- Spider Trap
  - `http://example.com/foo/bar/baz/foo/bar/baz/...`

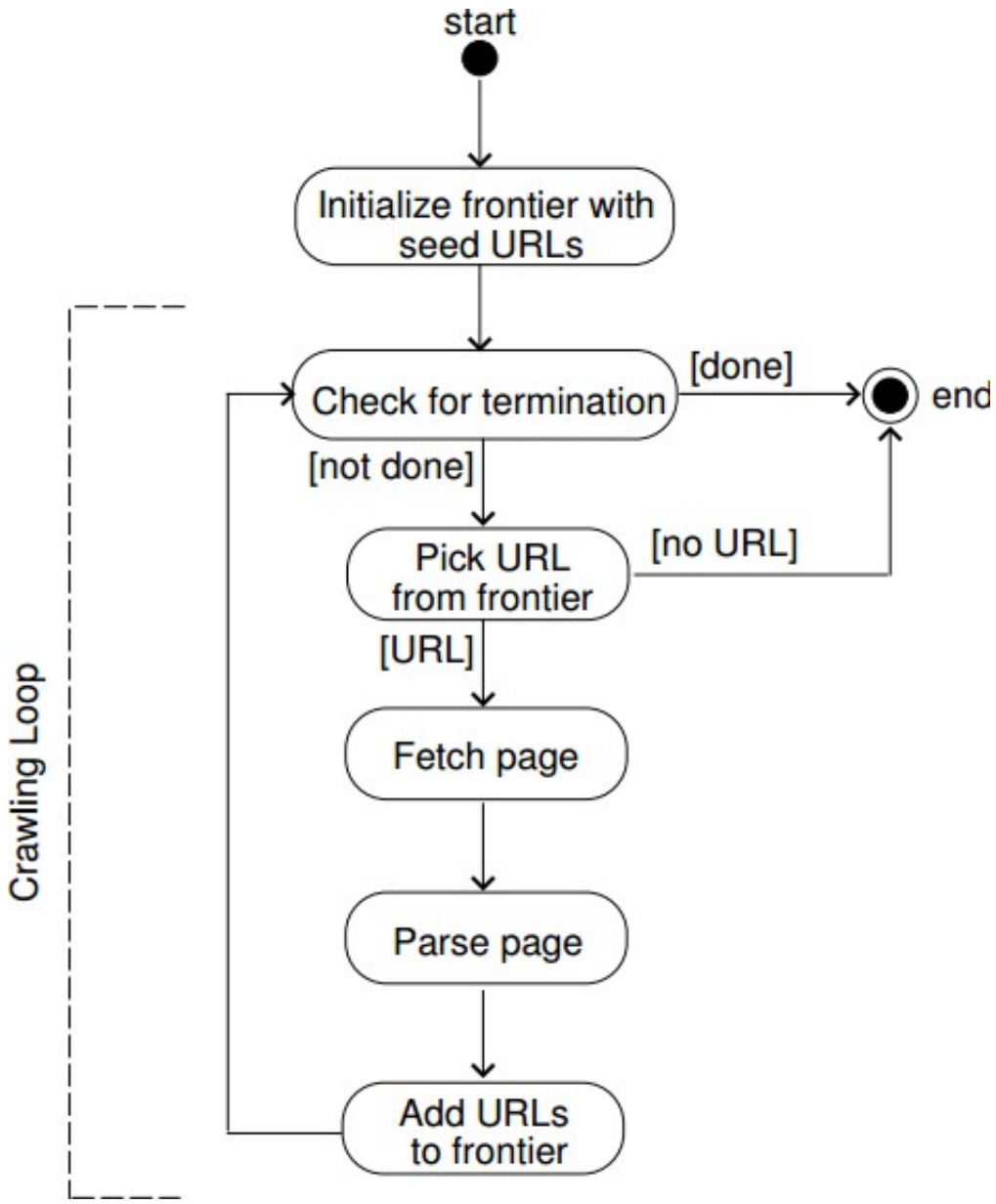
Problemy

- Robot Exclution Protocol
  - „polite” crawler

```
User-agent: *
Crawl-delay: 10
Allow: ../myfile.html
Disallow: /folder1/
Sitemap:
../sitemap.xml
Host: example.com
```

Crawler sekwencyjny

- Pula adresów do odwiedzenia (*frontier*)
  - Każda pętla dodaje kolejne adresy do frontier'u z parsowanej odpowiedzi
  - Sprawdzanie czy wcześniej odwiedzona
  - Przeskoczenie w chwili przekroczenia głębokości etc.


start
Initialize frontier with seed URLs
Check for termination
[done]
end
[not done]
Pick URL from frontier
[no URL]
[URL]
Fetch page
Parse page
Add URLs to frontier
Crawling Loop

Historia, repozytorium

- Data odwiedzenia
- Postać kanoniczna URL'a
- Zdarzenie
  - Niezwykle ciekawe źródło
- Sprawdzanie kiedy strona była odwiedzona
  - Odświerzenie wiadomości

Frontier - Problemy

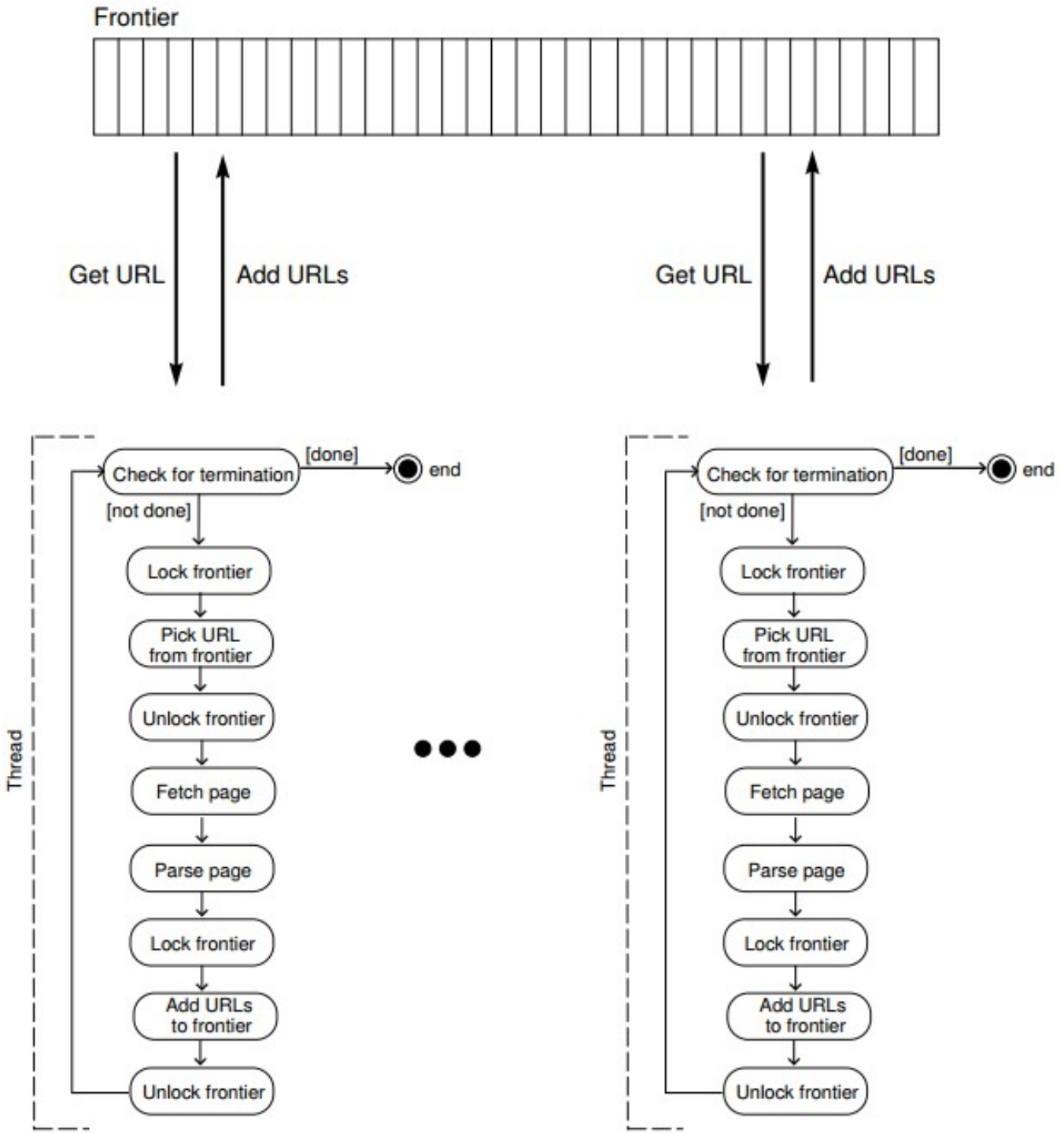
- Pojemność
  - Unikanie duplikacji
- Implementacja ?
  - FIFO

Parsowanie

- Wyrzucenie wszystkich stop-words'ów
  - Słownik
  - Statystyka
  - Hybryda
- Rozszerzenie i kanonizacja adresów
  - EXAMPLE.com
  - www.fizyka.umk.pl/~jacek %/E
  - http://www.example.com/../index.htm(l)

Crawler wielowątkowy

Osobny wątek – osobna pętla


Frontier
Get URL
Add URLs
Get URL
Add URLs
Thread
Check for termination
[done]
end
[not done]
Lock frontier
Pick URL from frontier
Unlock frontier
Fetch page
Parse page
Lock frontier
Add URLs to frontier
Unlock frontier
Thread
Check for termination
[done]
end
[not done]
Lock frontier
Pick URL from frontier
Unlock frontier
Fetch page
Parse page
Lock frontier
Add URLs to frontier
Unlock frontier

Problemy

- Blokowanie frontier'u
- Blokowanie historii, repozytorium
- Pusty frontier

Docelowo

- Wielowątkowość
- Proxy

Bibliografia

- http://www.wikipedia.com/
- http://webkitdotnet.sourceforge.net/
- http://htmlagilitypack.codeplex.com/
- http://msdn.microsoft.com/
- Pant G., Srinivasan P., Menczer F., Crawling the Web [online : dostęp 3 czerwca 2013] Dostępne w Internecie: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.4776&rep=rep1&type=pdf