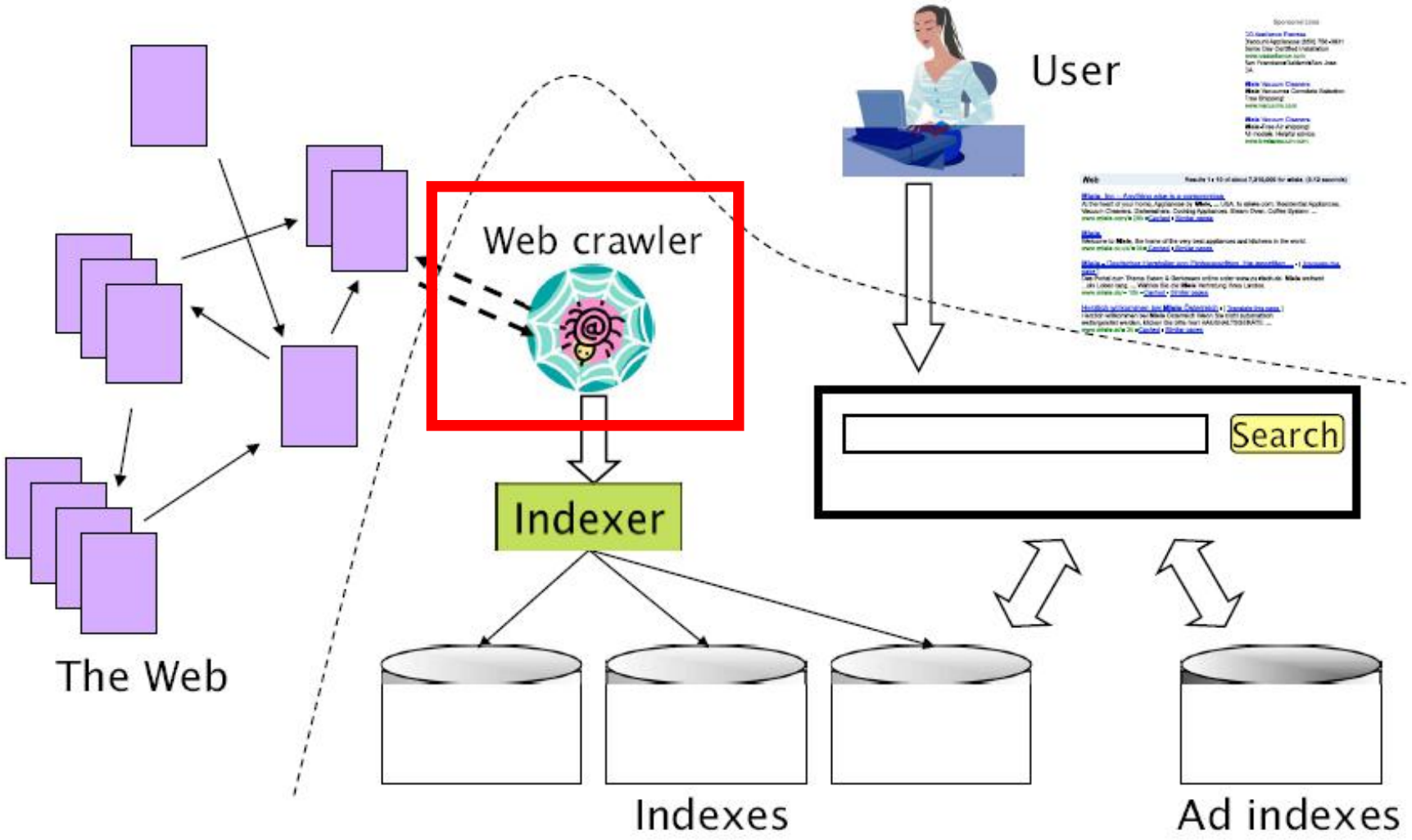
Knowledge Discovery and Data Mining

Lab 2 Introduction to Python Data Crawler

Xuan Song
Songx@sustech.edu.cn

Web Crawler

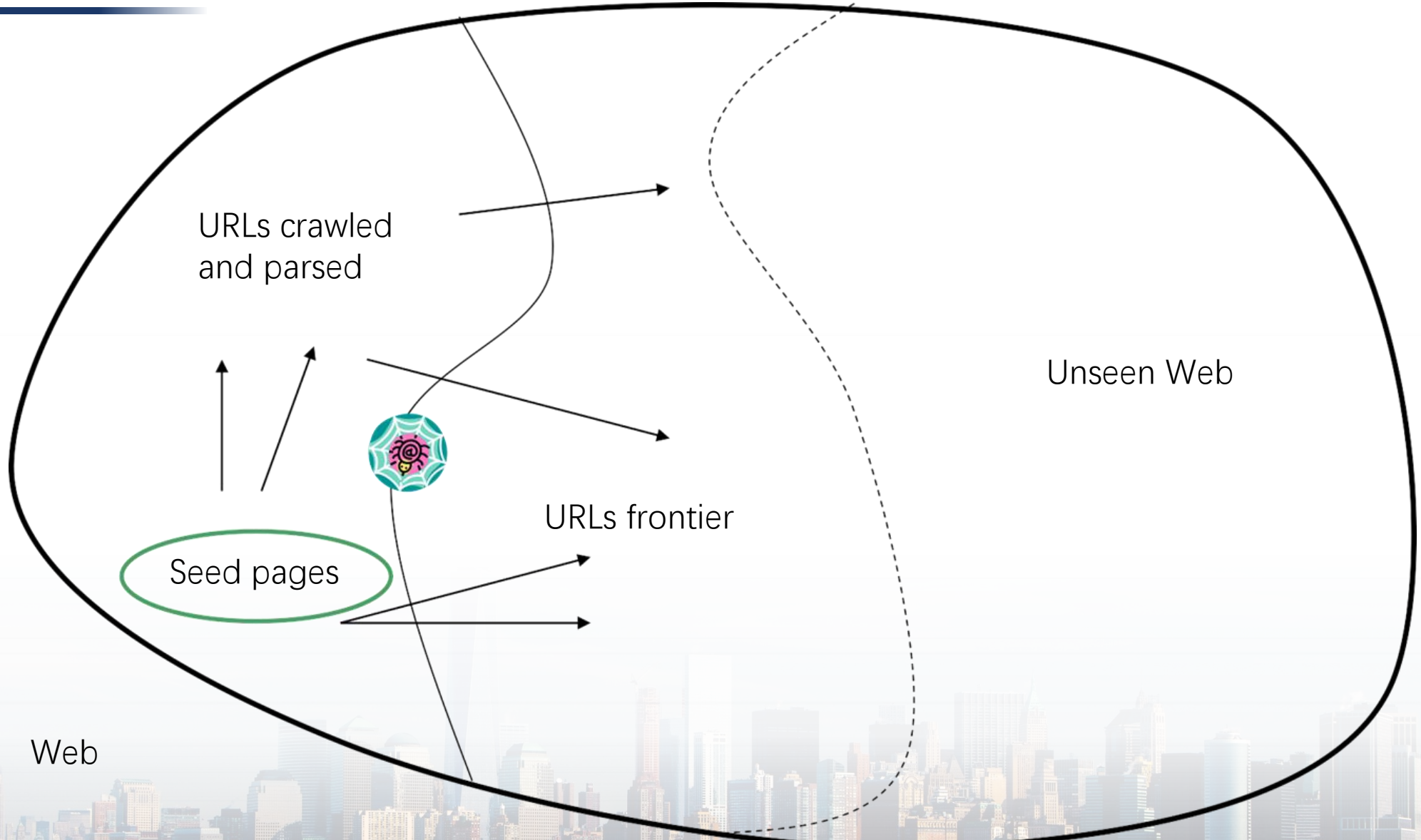
The Crawler



▶ Figure 19.7 The various components of a web search engine.

How hard can crawling be?

Basic Crawler Operations



Retrieved from: https://www.slideshare.net/muhammadbilalyousaf9/lecture16-crawling

Simple Crawler

```
urlqueue := (some carefully selected set of seed urls)

while urlqueue is not empty:
    myurl := urlqueue.getlastanddelete()
    mypage  := myurl.fetch()
    fetchedurls.add(myurl)
    newurls := mypage.extracturls()

    for myurl in newurls:
        if myurl not in fetchedurls and not in urlqueue:
            urlqueue.add(myurl)

    indexer.index(mypage)
```

What's wrong with this crawler?

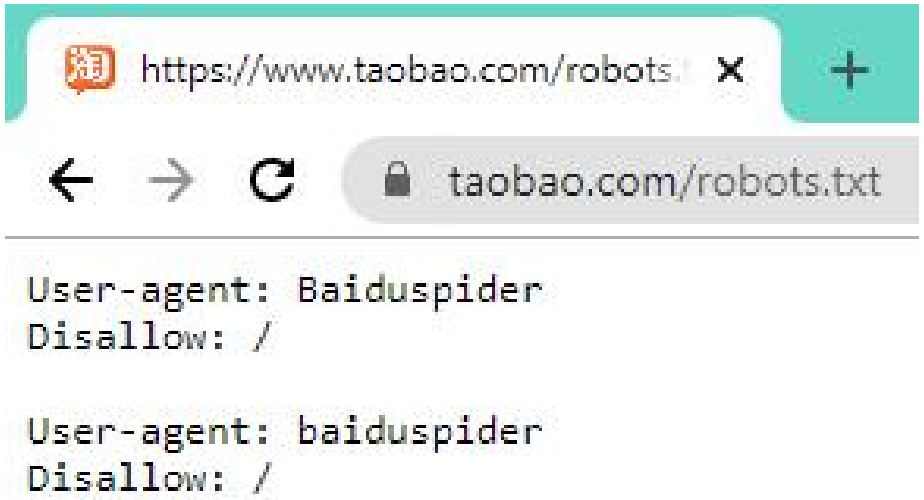
Features a crawler MUST provide

Robustness

Politeness

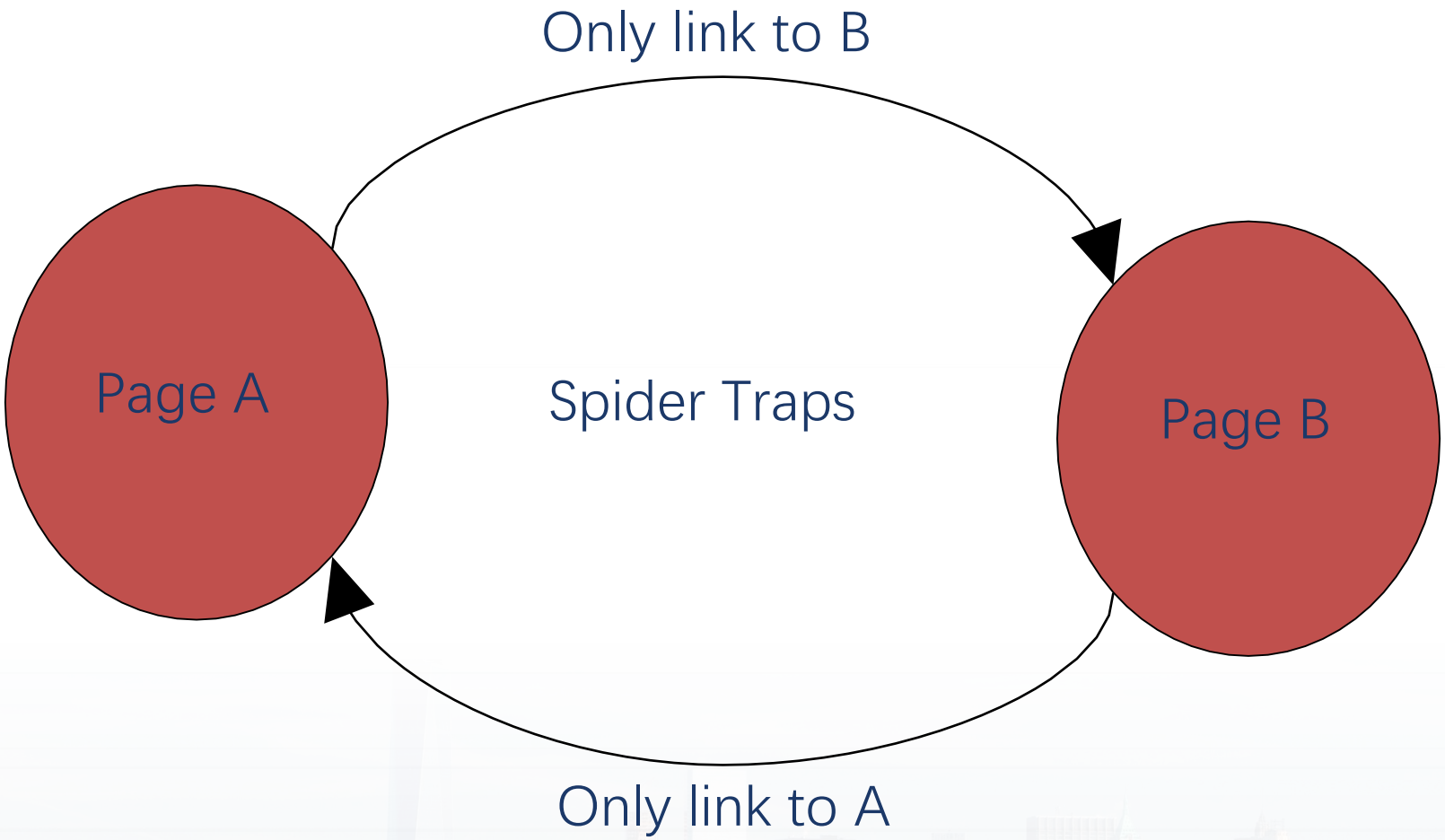
Politeness

1) Robots.txt



2) Do not frequently request the same site

Robustness



Features a crawler SHOULD provide

Distributed

Scalable

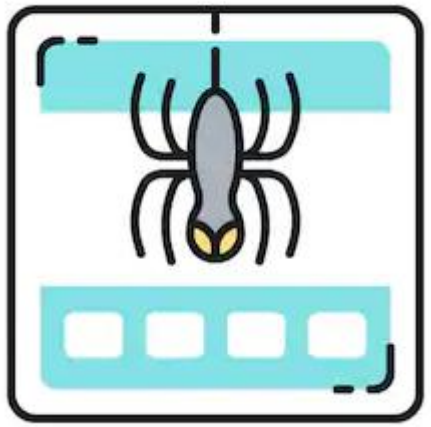
Performance and efficiency

Quality

Freshness

Extensible

What we do today



THE SCRAPINGHUB BLOG

Turn Web Content Into Useful Data

LARGE SCALE WEB SCRAPING

January 14, 2021 Attila Tóth 0 Comment

From inconsistent website layouts that break our extraction logic to badly written HTML, web scraping comes with its share of difficulties. Over the last few years, the single most

KEEP UP TO DATE WITH WEB SCRAPING AND DATA TIPS...

Email*

SIGN ME UP

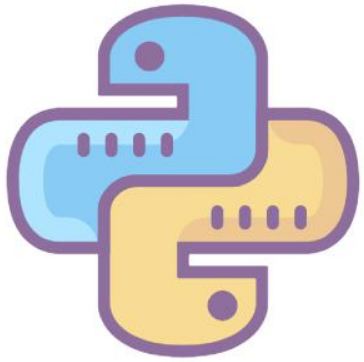
Story of the Month

A PRACTICAL GUIDE TO WEB DATA QA: BROAD CRAWLS

In this article, we will show you some of our favorite web

What we want today

Tools



Beautiful Soup 4

https://www.crummy.com/software/BeautifulSoup/bs4/doc.zh/#

Setup

```
from urllib.request import urlopen
from bs4 import BeautifulSoup
url = "https://blog.scrapinghub.com"
html = urlopen(url)
soup = BeautifulSoup(html, 'html.parser')
type(soup)
# Print out the text
text = soup.get_text()
#print(soup.text)
```

Exercise

Get the title, date and author from every post on the Scrapinghub blog.

```
{'title': 'Large scale web scraping', 'date': 'January 14, 2021 ', 'author': 'Attila Tóth'}
{'title': 'A Practical Guide to Web Data QA Part V: Broad Crawls', 'date': 'September 30, 2020 ', 'author': 'Ivan Ivanov'}
{'title': 'Announcing The Web Data Extraction Summit 2020', 'date': 'September 24, 2020 ', 'author': 'Himanshi Bhatt'}
{'title': 'News & Article Data Extraction: Open Source vs Closed Source Solutions', 'date': 'September 10, 2020 ', 'author': 'Attila Tóth'}
{'title': 'A PRACTICAL GUIDE TO WEB DATA QA PART IV: COMPLEMENTING SEMI-AUTOMATED TECHNIQUES', 'date': 'September 03, 2020 ', 'author': 'Ivan Iva
nov and Warley Ferreira Lopes'}
{'title': 'Real Estate: Use Web Data Extraction to Make Smarter Decisions', 'date': 'August 27, 2020 ', 'author': 'Attila Tóth'}
{'title': 'Scrapy Cloud Secrets: Hub Crawl Frontier and How To Use It', 'date': 'August 06, 2020 ', 'author': 'Júlio César Batista'}
{'title': 'Blog Comments API (BETA):\xa0Extract Blog Comment DATA At Scale', 'date': 'July 30, 2020 ', 'author': 'John Campbell'}
{'title': 'Your Price Intelligence Questions Answered', 'date': 'July 28, 2020 ', 'author': 'Himanshi Bhatt'}
{'title': 'Data Center Proxies vs. Residential Proxies', 'date': 'July 21, 2020 ', 'author': 'Attila Tóth'}
```

Hint

- Hint 1: Understand the html structure of the page can be very helpful!
- Hint 2: You can use BeautifulSoup to find css element to pin-point what you need.
- Hint 3: You can also grab the url for next page to recursively scrape the whole site.

End of Lab 2