



Nội dung môn học

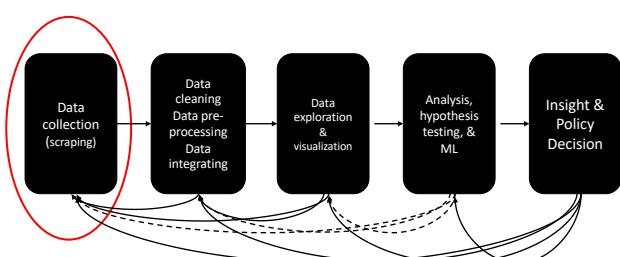
- Lecture 1: Tổng quan về Khoa học dữ liệu
- Lecture 2: Thu thập và tiền xử lý dữ liệu
- Lecture 3: Làm sạch và tích hợp dữ liệu
- Lecture 4: Phân tích và khám phá dữ liệu
- Lecture 5: Trực quan hóa dữ liệu
- Lecture 6: Trực quan hóa dữ liệu đa biến
- Lecture 7: Học máy
- Lecture 8: Phân tích dữ liệu lớn
- Lecture 9: Báo cáo tiền độ bài tập lớn và hướng dẫn
- Lecture 10+11: Phân tích một số kiểu dữ liệu
- Lecture 12: Đánh giá kết quả phân tích

Nội dung

- Trình thu thập dữ liệu
- Nền tảng Scrapy

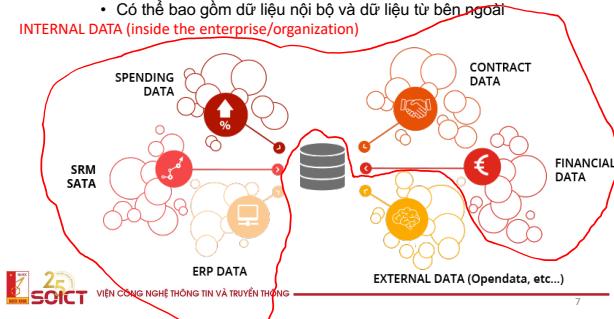
Giới thiệu

Chu trình khai thác dữ liệu



Thu thập dữ liệu là gì?

- Bước đầu tiên của làm khoa học dữ liệu
 - Mục tiêu là lấy về đủ thông tin cần thiết để phân tích trong bước sau
 - Có thể bao gồm dữ liệu nội bộ và dữ liệu từ bên ngoài



Các ví dụ về nguồn dữ liệu mở

- Global Health Facts (www.globalhealthfacts.org/)
 - UNdata (<http://data.un.org/>)
 - World Health Organization(www.who.int/research/en/) – Dữ liệu sức khoẻ
 - OECD Statistics (<http://stats.oecd.org/>) – Các chỉ số kinh tế
 - World Bank (<http://data.worldbank.org/>)
 - Census Bureau (www.census.gov/) – Dữ liệu nhân khẩu học
 - Data.gov (<http://data.gov/>) – Dữ liệu cung cấp bởi cơ quan chính phủ USA
 - Data.gov.uk (<http://data.gov.uk/>)
 - data.gouv.fr (<http://data.gouv.fr>)
 - DataSF (<http://datasf.org/>) – Dữ liệu của SanFrancisco
 - NYC DataMine (<http://nyc.gov/datalab/>) – Dữ liệu của New York.
 - ParisData (<http://opendata.paris.fr>) – Dữ liệu của Paris
 - OpenData La Rochelle (<https://opendata.larochelle.fr/>) – Dữ liệu của La Rochelle

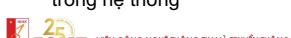


Truy cập dữ liệu trong các log files (tệp nhật ký)

- Log files chứa chuỗi nhật ký của các sự kiện đã xảy ra
 - API, network activity...
 - Các mục nhật ký có nhân thời gian và theo thứ tự thời gian
 - Cho phép phân tích các hoạt động và tương tác đã diễn ra trong hệ thống

Jul 30 01:44:00 mailsev syslogd[45]: AOL Sender Statistics
Jul 30 01:44:00 mailsev acproxyd[45]: Function: getInterfaceInternal File:
.../.../rep/Com/Utility/MailServer/unix/cpp Line: 1715 missing PPP destination address for interface "utun". Check profile PPFErule (set to Automatic) or contact your administrator.
Jul 30 01:44:00 mailsev acproxyd[45]: A network interface has gone down.
Jul 30 01:44:00 mailsev acproxyd[45]: Function: logInterfaceDown File:
.../.../rep/Com/Utility/MailServer/unix/cpp Line: 1715 missing PPP destination address for interface "utun". Check profile PPFErule (set to Automatic) or contact your administrator.
List: PSR0:0:0:0:1672:A219:192.168.1.109 PSR0:0:0:0:1672:0181:40FC:129E
Jul 30 01:44:00 mailsev acproxyd[45]: Function: getInterfaceInternal File:
.../.../rep/Com/Utility/MailServer/unix/cpp Line: 1715 missing PPP destination address for interface "utun". Check profile PPFErule (set to Automatic) or contact your administrator.
Jul 30 01:44:00 mailsev acproxyd[45]: Function: getPrimaryInterfaceNames File:
.../.../rep/Com/Utility/MailServer/unix/cpp Line: 501 Unable to get global IPV information from system configuration [error 1009].
Jul 30 01:44:00 mailsev acproxyd[45]: Function: updateDefaultPublicIPAddresses File:
.../.../rep/Com/Utility/MailServer/unix/cpp Line: 124 Invoked Function:
CMsgSetDescription: determine the public IP candidate and store Return Code: -28395823 (0x2F48001)
Description: HOSTONLINE_ERROR_SUPPORTED_PUBLIC_ADDRESS_UNAVAILABLE IPV6
Jul 30 01:44:00 mailsev acproxyd[45]: Function: getInterfaceInternal File:
.../.../rep/Com/Utility/MailServer/unix/cpp Line: 1715 missing PPP destination address for interface "utun". Check profile PPFErule (set to Automatic) or contact your administrator.
Jul 30 01:44:00 mailsev com:avast:proxy[598]: Error connecting to 216.58.204.110:443 connect()
Network is down
Jul 30 01:44:00 ... last message repeated 4 times ...
Jul 30 01:44:00 mailsev acproxyd[45]: Function: connectTransPort File:
.../.../rep/Com/Utility/SocketTransPort/unix/cpp Line: 1045 Invoked Function: _iconnect Return Code: 11
transports) reconnection attempt

[View Details](#)



Làm thế nào để truy cập tới dữ liệu?

- Tuỳ thuộc vào nguồn dữ liệu
 - Nguồn dữ liệu nội bộ:
 - CSDL, kho dữ liệu
 - Các tệp tin
 - Có cấu trúc (Excel, log files..)
 - Phi cấu trúc
 - Viết bằng ngôn ngữ tự nhiên
 - Văn bản
 - Các báo cáo
 - Nguồn dữ liệu từ bên ngoài:
 - API (SOAP or REST): <https://youtu.be/bPNfu0lZhoE>
 - Các tệp tin có thể tải về trên Internet
 - Các dữ liệu dưới dạng HTML trên các website



Truy cập dữ liệu thông qua APIs

- Nhiều công dịch vụ cung cấp qua giao thức http
 - Định dạng dữ liệu trao đổi, cung cấp được thống nhất, có cấu trúc để dễ dàng bóc tách và xử lý
 - Định dạng dữ liệu có thể không dễ dàng để hiểu nếu không rõ mô tả
 - Các giao thức thường sử dụng
 - SOAP: sử dụng XML
 - REST: sử dụng JSON
 - Ví dụ: thông tin về số chỗ đỗ xe đang trống ở La Rochelle:
 - <https://opendata.larochelle.fr/dataset/stationnement-parking-tarifs-synthetiques/>



Truy cập dữ liệu trong các log files (tệp nhật ký)

- ```
• Định dạng chuẩn hóa
phổ biến: syslog
 • Date of emission
 • Name of the device
 • Name of the service
 • Process that triggered emission
 • Priority Level
 • Message contents
 • Message category
 • Seriousness level

• ELK là nền tảng mã nguồn mở phổ biến để xử lý dữ liệu log, cho phép trực quan hoá dễ dàng

Jul 30 01:44:00 maltese cycload[51]: ASN.1eder: Statistics
Jul 30 01:44:00 maltese cycload[51]: ASN.1der: getStatisticsInternal File:
.../.../com/Utility/MaltegoInterface/unix/obj:Line 1751 missing PPP destination address for interface "utun". Check profile PPPXclusion (set to Automatic) or contact your administrator.
Jul 30 01:44:00 maltese cycload[51]: A network interface has gone down.
Jul 30 01:44:00 maltese cycload[51]: ASN.1der: getInterfaceFile:
.../.../com/Utility/MaltegoInterface/unix/obj:Line 477 IP Address Interface List:
PFB#0:0|0:0:6:37:FA:2E|21B_16.216.1.109 PFBD#0:0|0:0:0311:4C78:3E
Jul 30 01:44:00 maltese cycload[51]: ASN.1der: getInterfaceInternal File:
.../.../com/Utility/MaltegoInterface/unix/obj:Line 1751 missing PPP destination address for interface "utun". Check profile PPPXclusion (set to Automatic) or contact your administrator.
Jul 30 01:44:00 ... last message repeated 4 times ...
Jul 30 01:44:00 maltese cycload[51]: ASN.1der: getInterfaceInternal File:
.../.../com/Utility/MaltegoInterface/unix/obj:Line 501 Unable to get global IPv4 information from system configuration [error 1004].
Jul 30 01:44:00 maltese cycload[51]: ASN.1der: updateSystemPublicAddresses File:
.../.../com/Utility/MaltegoInterface/unix/obj:Line 2245 Invoked Function:
CmostConfig@:determinePublicAddressAndUpdateRoute Return Code: -28835823 (0x840001)
Description: NOROUTCERCONX_NOM_SUPPORTED_PUBLIC_ADDRESS_UNAVAILABLE!IPv4
Jul 30 01:44:00 maltese cycload[51]: ASN.1der: getInterfaceFile:
.../.../com/Utility/MaltegoInterface/unix/obj:Line 1751 missing PPP destination address for interface "utun". Check profile PPPXclusion (set to Automatic) or contact your administrator.
Jul 30 01:44:00 maltese cycload[51]: ASN.1der: connectRequest[4]: ...
Jul 30 01:44:00 maltese cycload[51]: ASN.1der: getInterfaceInternal File:
.../.../com/Utility/MaltegoInterface/unix/obj:Line 1023 Invoked Function: :connectReturn Code: 11
.../.../com/Utility/SocketTransport/transmit[1023]: Invoked Function: :connectReturn Code: 11
```

---



# Các định nghĩa

Data scraping, screen scraping, report mining, web scraping

## Thu thập dữ liệu thông qua cào dữ liệu (data scraping)

- Data scraping được sử dụng khi hệ thống có dữ liệu không cung cấp giao diện và API để truy cập dữ liệu
- **Data scraping** là một kỹ thuật để trích xuất dữ liệu từ nguồn dữ liệu được công khai thành dạng có cấu trúc
  - Thường là các trang web
  - Cũng có thể là các nguồn thông tin khác được hiển thị trên màn hình hoặc các giao diện khác

## Một vài giới hạn của data scraping

- Chủ sở hữu của nguồn thông tin không ưa thích data scraping
  - Gây quá tải cho hệ thống
  - Mất mát doanh thu từ quảng cáo
  - Mất kiểm soát tới dữ liệu, vi phạm bản quyền dữ liệu
- Data scraping thường là kỹ thuật được dùng khi không có phương án thay thế

## Screen mining là gì?

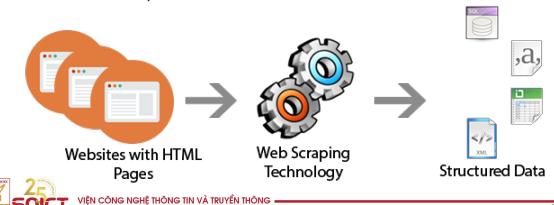
- Trích xuất dữ liệu văn bản từ màn hình hiển thị của thiết bị
- Thường sử dụng kỹ thuật chụp ảnh và OCR
- Trong một vài trường hợp, thường sử dụng kèm với chương trình giả lập thao tác của người dùng để điều khiển UI
  - Cho phép tự động chụp lại toàn bộ các màn hình

## Report mining là gì?

- Bóc tách nội dung từ báo cáo, các văn bản ở ngôn ngữ tự nhiên (PDF, text, vvv)
  - Không cần sử dụng API
- Ví dụ: Các công cụ Tabula, import in Tableau

## Web scraping là gì?

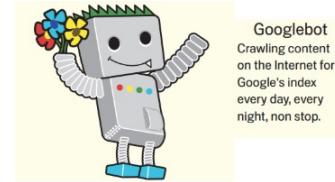
- Trang web là các tệp dữ liệu văn bản được viết theo định dạng markup-based language (HTML and XHTML)
- Tuy nhiên định dạng này phục vụ để hiển thị cho con người, không dành cho xử lý dữ động
- Để chống lại kỹ thuật cào web, một vài website sử dụng các cơ chế phòng thủ (Giới hạn số lượt truy cập theo IP, CAPTCHA...)



## Trình thu thập Web (Web crawler)

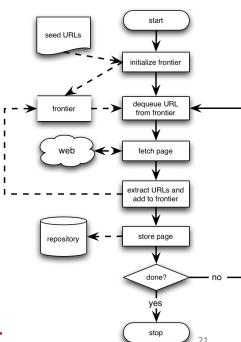
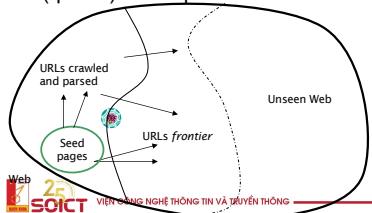
## Web crawler

- Là chương trình phần mềm, được gọi là con nhện, một robot, đi theo các links trên các websites, thu thập các thông tin và lưu vào CSDL
- Các tên gọi
  - Crawler
  - Spider
  - Robot
  - Web agent



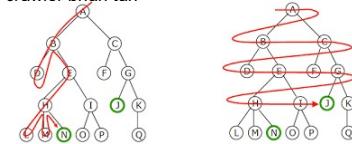
## Basic crawling operation

- BEGIN with known "seed" URLs
- Fetch and parse them
  - Extract URLs they point to
  - Place the extracted URLs on a queue
- Fetch each URL on the frontier (queue) and repeat



## Chính sách cào dữ liệu

- Hành vi của một Web crawler là kết quả của sự kết hợp các chính sách sau
  - Selection policy: Tải về trang nào
  - Re-visit policy: Khi nào cần kiểm tra lại có hay không sự thay đổi ở các trang web
  - Politeness policy: Chính sách tránh làm quá tải máy chủ
  - Parallelization policy: Chính sách phối hợp giữa các web crawler phân tán



## Thách thức đối với cào web

- Internet là rất rộng lớn, bao la
  - Googlebot là trình cào web phân tán
- Lọc, phân biệt các trang quan tâm/ không quan tâm/ trang độc hại (đối với bot)
  - Spam pages
  - Spider traps – trang được sinh ra tự động dễ dụ bot
- Tính mới của dữ liệu (Content freshness)
  - Crawler cần thu thập được dữ liệu mới, có yếu tố thời điểm
- Trùng lặp dữ liệu
  - Các trang trùng lặp hoặc cả website được nhân bản

## Cân bằng giữa exploitation vs. exploration

- Khai thác - Exploitation
  - Thu thập các trang mà xác xuất cao có dữ liệu cần thu thập
- Khám phá - Exploration
  - Khám phá các nguồn dữ liệu mới mà có thể có dữ liệu cần thu thập

## Tính lịch thiệp - Politeness

- Mô tả rõ - Explicit
  - Chỉ định bởi chủ trang web mô tả phần nào của website có thể được thu thập (robots.txt)
- Ngầm định - Implicit
  - Tránh gây quá tải, thu thập quá thường xuyên dẫn tới tiêu tốn tài nguyên máy chủ, ảnh hưởng đến chất lượng cung cấp dịch vụ của máy chủ

## Robots.txt

- Là giao thức đặc tả thể hiện sự giới hạn đối với các "robots", đưa ra từ 1994
  - [www.robotstxt.org/wc/norobots.html](http://www.robotstxt.org/wc/norobots.html)
- Website chỉ định các nội dung không được thu thập**
  - Tạo tệp tin /robots.txt
  - Chỉ định các giới hạn
- Ví dụ

```
User-agent: *
Disallow: /yoursite/temp/
```

```
User-agent: searchengine
Disallow:
```

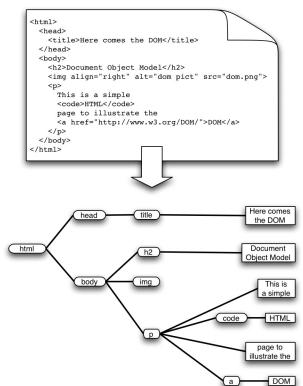
## Bóc tách thông tin khi thu thập (Web-scraping)

## Kỹ thuật bóc tách

- Để bóc tách thông tin từ trang web, thường sử dụng **XPath hoặc CSS Selector**
- XPath là chuẩn của W3C để tìm kiếm một phần tử trong văn bản XML
  - W3C là tổ chức xây dựng và quản lý các chuẩn web
- XPath sử dụng cấu trúc cây gồm các nút (và thuộc tính) trên tệp XML
- HTML là đặc tả tuân theo cấu trúc cây của XML

## Cấu trúc mã HTML

- HTML có cấu trúc của cây Document Object Model (DOM)
- DOM khác nhau tuỳ từng trang, thậm chí với cùng một dạng trang
  - Do nội dung động, chèn quảng cáo, vvv.



## Ví dụ về XPath

| Path Expression | Result                                                                                                                       |
|-----------------|------------------------------------------------------------------------------------------------------------------------------|
| bookstore       | Selects all nodes with the name "bookstore"                                                                                  |
| /bookstore      | Selects the root element bookstore                                                                                           |
|                 | <b>Note:</b> If the path starts with a slash ( / ) it always represents an absolute path to an element!                      |
| bookstore/book  | Selects all book elements that are children of bookstore                                                                     |
| //book          | Selects all book elements no matter where they are in the document                                                           |
| bookstore//book | Selects all book elements that are descendant of the bookstore element, no matter where they are under the bookstore element |
| //@lang         | Selects all attributes that are named lang                                                                                   |

Practice it yourself on [https://www.w3schools.com/xml/xpath\\_examples.asp](https://www.w3schools.com/xml/xpath_examples.asp)

## Ví dụ về XPath

| Path Expression                      | Result                                                                                                                                                                                                                                                                                                        |
|--------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| /bookstore/book[1]                   | Selects the first book element that is the child of the bookstore element.<br><b>Note:</b> In IE 5,6,7,8,9 first node is [0], but according to W3C, it is [1]. To solve this problem in IE, set the SelectionLanguage to XPath:<br>In JavaScript: <code>xml.setProperty("SelectionLanguage", "XPath");</code> |
| /bookstore/book[last()]              | Selects the last book element that is the child of the bookstore element                                                                                                                                                                                                                                      |
| /bookstore/book[last()-1]            | Selects the last but one book element that is the child of the bookstore element                                                                                                                                                                                                                              |
| /bookstore/book[position()<3]        | Selects the first two book elements that are children of the bookstore element                                                                                                                                                                                                                                |
| //title[@lang]                       | Selects all the title elements that have an attribute named lang                                                                                                                                                                                                                                              |
| //title[@lang='en']                  | Selects all the title elements that have a "lang" attribute with a value of "en"                                                                                                                                                                                                                              |
| /bookstore/book[price>35.00]         | Selects all the book elements of the bookstore element that have a price element with a value greater than 35.00                                                                                                                                                                                              |
| /bookstore/book[price > 35.00]/title | Selects all the title elements of the book elements of the bookstore element that have a price element with a value greater than 35.00                                                                                                                                                                        |

31

## Các giới hạn về kỹ thuật

### • Sự không nhất quán / hỗn loạn trong tổ chức thông tin

- Thông tin trong 1 website có bố cục HTML không giống nhau, ngay cả khi biểu diễn cùng một loại thông tin

### • Sự thay đổi trong cấu trúc

- Các trình thu thập dữ liệu chạy liên tục
- Cấu trúc website cũng thay đổi theo thời gian, do cập nhật, chỉnh sửa
  - Đặc biệt với các trang có trình quản lý themes, Content Management Systems (e.g. Wordpress)

## Các giới hạn về kỹ thuật

### • Giới hạn truy cập

- Nội dung bị giới hạn cho người dùng được xác thực
- Vẫn có thể tự động hóa được việc xác thực nhưng cần tài khoản

### • Nội dung được sinh ra tự động

- Công nghệ web tiên tiến không tải về toàn bộ nội dung trong một lượt truy cập (request)
  - Tùy theo tương tác người dùng mà tải về nội dung tương ứng
- Một vài công cụ có thể giả lập tương tác và trình duyệt, nhưng quá trình thu thập sẽ khó khăn hơn

33

## Các giới hạn về kỹ thuật

### • Chủ sở hữu của nguồn thông tin không ưa thích data scraping

- Gây quá tải cho hệ thống
- Mất mát doanh thu từ quảng cáo
- Mất kiểm soát tới dữ liệu, vi phạm bản quyền dữ liệu

### • Chủ sở hữu website có thể sử dụng các kỹ thuật, công cụ để chặn truy cập từ trình thu thập

- Giới hạn lượt truy cập theo thời gian
  - Con người thường không di chuyển giữa các trang quá nhanh

## Các giới hạn về kỹ thuật

### • Chủ sở hữu website có thể sử dụng các kỹ thuật, công cụ để chặn truy cập từ trình thu thập

- Chặn IP**
  - Tự động chặn khi quá nhiều request đến từ 1 IP
  - DoS là một dạng của cyber-attacks
- Giới hạn băng thông**
- Xác thực yêu cầu truy cập đến từ người**
  - Vd. CAPTCHA, vvv.

35

## Khía cạnh đạo đức / pháp lý

### • Pháp luật đối xử với việc thu thập thông tin

- Không rõ ràng với hầu hết các quốc gia
  - European regulation « General Data Protection Regulation » (GDPR)
- Khác nhau tùy từng quốc gia

### • Có OK không khi chúng ta đăng lại thông tin và vẫn chỉ rõ nguồn dữ liệu?

- Câu hỏi khó!
- Một mặt, quyền sở hữu trí tuệ được tôn trọng
- Mặt khác, nguồn dữ liệu gốc bị mất lượt truy cập (traffic), dẫn tới mất doanh thu

## Khía cạnh đạo đức / pháp lý

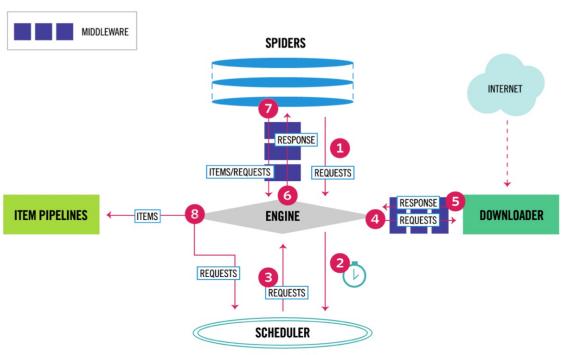
- Chúng ta nên thông báo cho chủ sở hữu website về việc chúng ta có mong muốn sử dụng dữ liệu của họ
  - Có thể nhận được sự cho phép từ đó tránh được các vấn đề về pháp lý
  - Có thể nhận được API truy cập nếu có
- Chủ sở hữu website có thể yêu cầu đưa URL về trang của họ từ đó giúp cải thiện thứ hạng của họ trong máy tìm kiếm
  - Lợi cả đôi bên, Win-win!

## Các công cụ thu thập dữ liệu

- Thư viện lập trình
  - Scrapy, BeautifulSoup (Python), PhantomJS
- Cloud: ScrapingHub, Dexi.io...
- Phần mềm: ParseHub, OctoParse...
- Web-browser plugins:
  - Data Scraper - Easy Web Scraping, Instant Data Scraper, Web Scraper

| General Features Comparison |                          |                         |                          |                          |                            |
|-----------------------------|--------------------------|-------------------------|--------------------------|--------------------------|----------------------------|
|                             | Octoparse                | Parsehub                | Mozenda                  | Descripto                | Importio                   |
| Usability                   | ★★★★★                    | ★★★★★                   | ★★★★★                    | ★★★★★                    | ★★★★★                      |
| Functionality               | ★★★★★                    | ★★★★★                   | ★★★★★                    | ★★★★★                    | ★★★★★                      |
| Easy to learn               | ★★★★★                    | ★★★★★                   | ★★★★★                    | ★★★★★                    | ★★★★★                      |
| Customer support            | Email, phone, community  | Email, live chat, forum | Phone, email, video chat | Email, phone, community  | Email, chat bot, community |
| Price                       | \$0 - \$249              | \$149 - \$499           | \$100/5000 page credits  | \$119 - \$699            | \$299 - \$999              |
| Trial/Free version          | Free Version             | Free Version            | 30 days trial            | Trial                    | 7 days trial               |
| OS (Specification)          | Win                      | Win, Mac, Linux         | Win                      | Win, Mac, Linux          | Win, Mac, Linux            |
| Data Export Formats         | TXT, CSV, XLS, Databases | CSV, XLS                | CSV, TSV, XLS, XLS, JSON | CSV, XLS, XLS, JSON, Zip | CSV, JSON, Google sheets   |
| Multi-thread                | ✓                        | ✓                       | ✓                        | ✓                        | ✓                          |
| API                         | ✓                        | ✓                       | ✓                        | ✓                        | ✓                          |
| Scheduling                  | ✓                        | ✓                       | ✓                        | ✓                        | ✓                          |

## Các thành phần của Scrapy



## Thực hành

Thu thập dữ liệu với Scrapy

## Giới thiệu về Scrapy

- Scrapy là một thư viện lập trình mã nguồn mở mạnh mẽ, viết bằng Python
- Scrapy có thể cài đặt sẵn mã chương trình cho các vấn đề của trình thu thập dữ liệu như
  - Throttling
  - Concurrency
  - XML sitemaps
  - Filtering duplicated URLs
  - Retry on Error

## Các thành phần của Scrapy

- 
- The diagram shows a detailed view of the Scrapy components. At the top, there is a box labeled "SPIDERS" with a red square inside. Below it is a box labeled "ITEM PIPELINES" with a green square inside. To the right, there is a box labeled "DOWNLOADER" with a blue square inside. Further right is a box labeled "SCHEDULER" with an orange square inside. At the bottom, there is a box labeled "ENGINE" with a yellow square inside. Arrows show the flow of data between these components. A legend on the left identifies the colors: blue for MIDDLEWARE, green for ITEM PIPELINES, red for SPIDERS, orange for SCHEDULER, and yellow for ENGINE. A legend on the right identifies the symbols: a red square for SPIDERS, a blue square for DOWNLOADER, an orange square for SCHEDULER, and a green square for ITEM PIPELINES. The numbers 1 through 8 are placed near specific arrows in the main diagram to indicate the flow of data.
- Scrapy Engine
    - Điều khiển luồng giữa tất cả các thành phần
  - Scheduler
    - Nhận yêu cầu từ engine và đưa vào hàng đợi để xử lý
  - Downloader
    - Tải trang web và đưa về cho engine
  - Spiders
    - Bóc tách các trả lời gồm dữ liệu và các đường dẫn yêu cầu mới cần phải truy cập tới
  - Item pipeline
    - Xử lý dữ liệu sau khi được bóc bởi spider
  - Downloader middlewares
    - Xử lý các yêu cầu khi chúng đi từ engine tới downloader và ngược lại
  - Spider middlewares
    - Nằm giữa Engine và Spiders
    - Xử lý các trả lời (responses) và các mục dữ liệu (items) và yêu cầu (requests)

## Bài tập

### 1. XPath

- Học trên [https://www.w3schools.com/xml/xpath\\_examples.asp](https://www.w3schools.com/xml/xpath_examples.asp)

### 2. WebScraping tutorial

- <https://www.webscraper.io/tutorials>

### 3. Scrapy

- <https://doc.scrapy.org/en/latest/intro/tutorial.html>
- <https://doc.scrapy.org/en/latest/topics/media-pipeline.html>
- « Web Scraping in Python using Scrapy \_ Codementor »: PDF in the Google Teams

## PageRank

## Giới thiệu

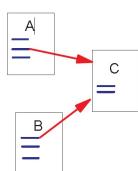
- Thách thức thực trong tìm kiếm thông tin trên World Wide Web.
  - Số lượng các website khổng lồ: 150 million by 1998, 1000 billion by 2008
  - Thông tin trên các website rất đa dạng: đa dạng chủ đề và chất lượng, vv.
- PageRank là gì?
  - Một phương pháp để đánh trọng số độ quan trọng của một trang web sử dụng cấu trúc liên kết giữa chúng.

## Lịch sử của PageRank

- PageRank được phát triển bởi Larry Page và Sergey Brin
- Là một phần của dự án nghiên cứu về một máy tìm kiếm Internet. Dự án bắt đầu vào 1995 và có bản prototype nguyên mẫu vào 1998.
- Page và Brin sau đó sáng lập của Google.
- Ngày nay
  - Có nhiều kỹ thuật dựng cấu trúc website và nội dung sao cho tối ưu hoá cho công cụ tìm kiếm, tăng thứ hạng (SEO).

## Cấu trúc liên kết của Web

- 150 triệu trang web → 1.7 tỷ liên kết (link)



Backlinks and Forward links:  
➤ A and B are C's backlinks  
➤ C is A and B's forward link

Một trang web là quan trọng nếu nó có nhiều backlinks.

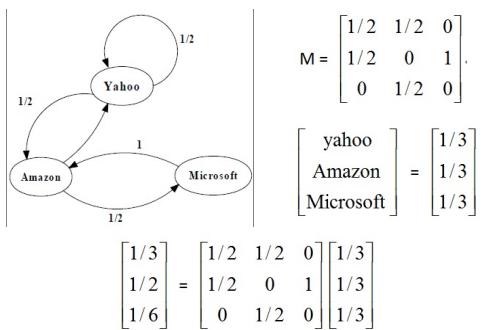
Một trang web có quan trọng không nếu nó không có link từ [www.cnn.com](http://www.cnn.com)?

## Giải thuật PageRank đơn giản hóa

- u: một trang web
- B<sub>u</sub>: tập hợp các backlinks của u
- N<sub>v</sub>: Số lượng các forward links của trang v
- c: Hệ số chuẩn hóa để đạt được
  - $\|R\|_{L1} = 1 (\|R\|_{L1} = |R_1| + \dots + |R_n|)$

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

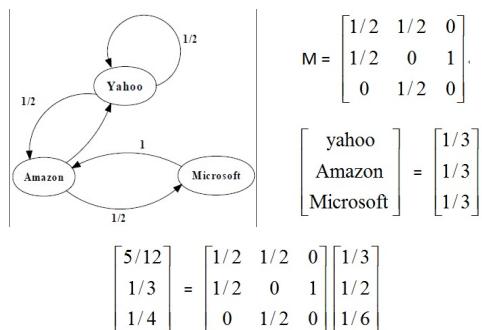
## Giải thuật PageRank đơn giản hóa



 PageRank Calculation: first iteration  
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

49

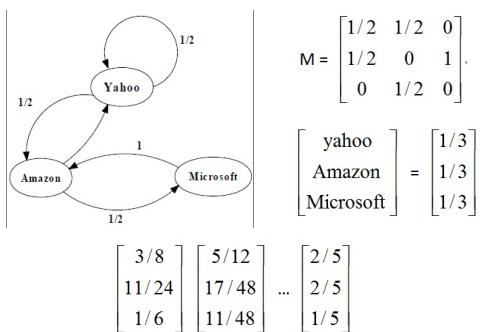
## Giải thuật PageRank đơn giản hoá



PageRank Calculation: second iteration

50

## Giải thuật PageRank đơn giản hóa

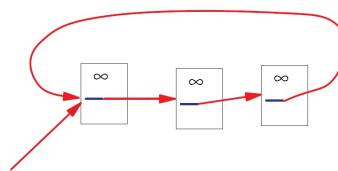


Convergence after some iterations

51

## Vấn đề với PageRank đơn giản hóa

### Vòng lặp:



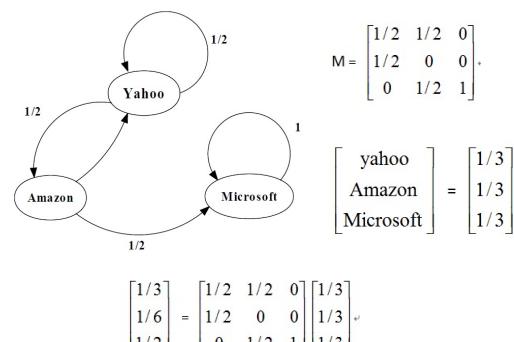
Sau mỗi bước lắp, các trang nhận được thứ hạng nhưng không phân phối cho các trang khác!

**25**  
**SQICT**

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

52

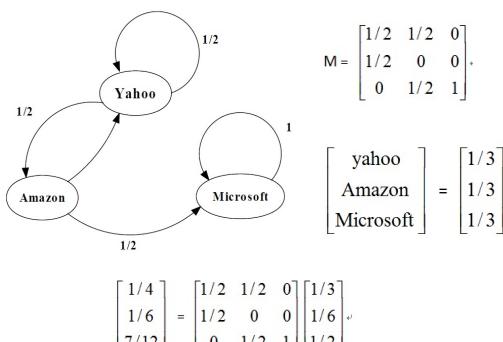
Ví du



[1/2] [0]

三

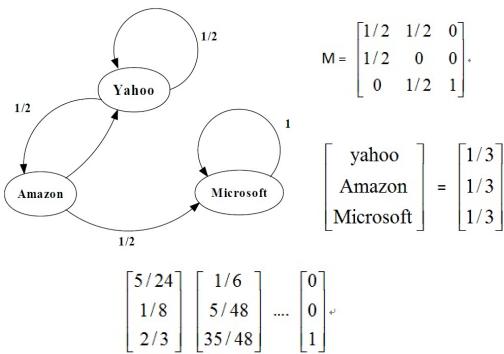
Ví du



[ 112 ] [ 8 ]

三

## Ví dụ



## Random Walks in Graphs

- Mô hình duyệt ngẫu nhiên - Random Surfer Model

• Duyệt web liên tục đi theo các link với xác suất ngẫu nhiên

- Mô hình duyệt có chỉnh sửa - The Modified Model

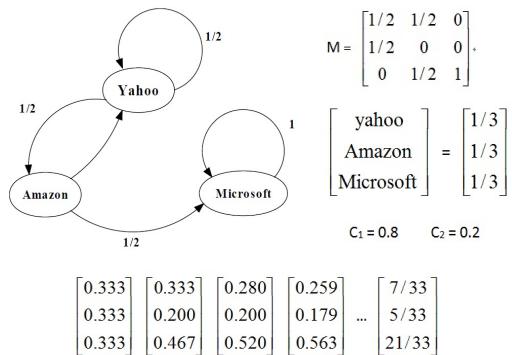
• Duyệt web liên tục theo các link với xác suất ngẫu nhiên, nhưng định kỳ nhảy vào một trang theo phân phối  $E$

## Một phiên bản cải tiến của PageRank

$$R'(u) = c_1 \sum_{v \in B_u} \frac{R'(v)}{N_v} + c_2 E(u)$$

$E(u)$ : a distribution of ranks of web pages that “users” jump to when they “gets bored” after successive links at random.  
For uniform random jump:  $E(i) = 1/n$

## Ví dụ Modified PageRank



## Dangling Links

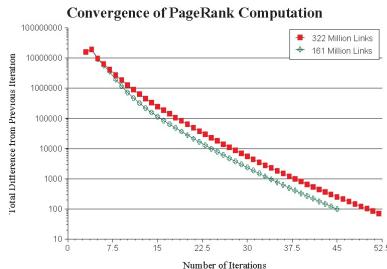
- Link trả tới bất kỳ page nào mà không có link đi ra
- Phần lớn là các trang chưa được tải về xong
- Ảnh hưởng tới mô hình vì nó gây ra sự không rõ ràng về trọng số của nó sẽ được phân phối đi đâu
- Không ảnh hưởng tới tính thứ hạng của các trang khác
- Có thể đơn giản là bỏ đi trước khi tính pagerank và thêm vào sau đó

## Cài đặt PageRank

- Chuyển đổi mỗi URL thành một số nguyên ID duy nhất, và lưu các hyperlink trong 1 cơ sở dữ liệu sử dụng ID để định danh các trang
- Sắp xếp cấu trúc liên kết bởi ID
- Xóa bỏ các dangling links
- Khởi tạo thứ hạng và bắt đầu các vòng lặp
  - Giá trị khởi tạo tốt có thể tăng tốc hội tụ cho pagerank
- Thêm lại các dangling links.

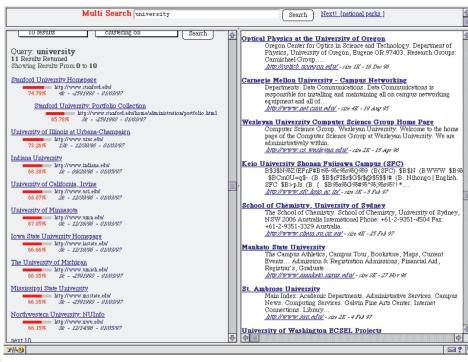
## Tính hội tụ Convergence Property

- PR (322 Million Links): 52 iterations
- PR (161 Million Links): 45 iterations
- Scaling factor is roughly linear in logn



61

## Tìm kiếm với PageRank



62



## Tìm kiếm với PageRank

| Web Page                                              | PageRank (average is 1.0) |
|-------------------------------------------------------|---------------------------|
| Download Netscape Software                            | 11589.00                  |
| http://www.w3.org/                                    | 10717.70                  |
| Welcome to Netscape                                   | 8673.51                   |
| Point: It's What You're Searching For                 | 7930.92                   |
| Web-Counter Home Page                                 | 7254.97                   |
| The Blue Ribbon Campaign for Online Free Speech       | 7010.39                   |
| CERN Welcome                                          | 6562.49                   |
| Yahoo!                                                | 6561.80                   |
| Welcome to Netscape                                   | 6203.47                   |
| Wusage 4.1: A Usage Statistics System For Web Servers | 5963.27                   |
| The World Wide Web Consortium (W3C)                   | 5672.21                   |
| Lyves, Inc. Home Page                                 | 4683.31                   |
| Starting Point                                        | 4501.98                   |
| Welcome to Magellan!                                  | 3866.82                   |
| Oracle Corporation                                    | 3587.63                   |

Top 15 Page Ranks: July 1996

63



VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

## Cá nhân hoá PageRank

- Một thành phần quan trọng trong tính toán PageRank là E
- E vector thể hiện phân phối của các trang web mà việc duyệt web liên tục theo các link với xác xuất ngẫu nhiên, nhưng định kỳ nhảy vào một trang theo phân phối E
- Có thể sử dụng E vector để cá nhân hoá



VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

64

## Kết luận

- PageRank là giải thuật xếp hạng toàn cục cho tất cả các trang web dựa trên vị trí của trang web trên cấu trúc đồ thị web
- PageRank sử dụng thông tin để xếp hạng là các backlinks

Chân thành  
cảm ơn!!!



VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

65