

Các công nghệ mới trong phát triển phần mềm

# Rút trích thông tin web



KHOA CÔNG NGHỆ THÔNG TIN  
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN



- ☐ Hiểu được nhu cầu của các hệ thống rút trích thông tin web
- ☐ Hiểu được cơ chế hoạt động của các hệ thống rút trích thông tin web
- ☐ Biết cách rút trích thông tin từ một trang web
- ☐ Biết cách thực hiện các hành động tự động hóa web cơ bản

ĐĂNG NHẬP

Webmail | Trang môn học | Liên hệ  
Trang chủ | Góp ý | English

TÌM KIẾM

GIỚI THIỆU

NGHIÊN CỨU

ĐÀO TẠO

## KHOA CÔNG NGHỆ THÔNG TIN

KHOA CÔNG NGHỆ THÔNG TIN TP.HCM

Tôi muốn  
feed

Khoa Công nghệ Thông tin (CNTT) của Trường Đại học Khoa học Tự nhiên Tp. HCM được thành lập theo quyết định số 8/GĐ-ĐT ngày 13/12/1994 của Bộ Trưởng Bộ GD&ĐT, dựa trên Bộ môn tin học của Khoa Toán Trường Đại học Tổng hợp HCM. Trải qua gần 16 năm hoạt động, Khoa đã phát triển vững chắc và được chính phủ bảo trợ để trở thành một trong những khoa CNTT đầu ngành trong hệ thống giáo dục đại học của Việt Nam.

Hệ thống Webmail

Diễn đàn thảo luận

Website khảo sát

Thư viện điện tử

Bảng vàng thành tích

Lý lịch khoa học

MSDNAA (SV)

MSDNAA (GV)


### SỰ KIỆN

### LIÊN KẾT

### TIN TỨC

- 2** 2010 Hình Lễ Tốt nghiệp ngày 16/10/2010 **NEW**
- 28** 2010 Thông báo v/v chuyển sang hệ tự túc của học viên CH khóa 17 ngành KHMT
- 23** 2010 Chương trình CNTT hợp tác với Đại học Claude Bernard Lyon 1 (Pháp)
- 20** 2010 V/v triển khai email và MSDNAA cho SV khóa 2010
- 20** 2010 Syllabus và lịch giảng dạy môn PPNCKH Học Kỳ 1 (2010-2011)
- 5** 2010 Thời hạn bảo vệ Cao học ngành HTTT K17
- 7** 2010 Lịch và chương trình học của lớp cao học khóa 20

Xem tất cả các tin




Your power & your friend

[SO SÁNH GIÁ SẢN PHẨM](#)
[CHUYÊN GIA](#)
[CỬA HÀNG](#)
[RAO VẶT](#)

Hỗ trợ [Not online](#) [Not online](#)  
**Hotline** (04) 3514.7504

[Đăng nhập tài khoản](#)  
[Đăng ký mới](#)

Đang xem: [So sánh giá >>](#) [Mobile & Phụ kiện >>](#) [Điện thoại di động](#)




**Samsung P1000 Galaxy Tab**  
Khoảng giá : 15.850.000 Đ -- 17.990.000 Đ

Bạn có sản phẩm **Samsung P1000 Galaxy Tab** muốn đăng so sánh giá tại đây ? Vui lòng liên hệ hỗ trợ phía trên để được tư vấn.

[So sánh giá cửa hàng](#)
[Mô tả sản phẩm](#)
[Ý kiến đánh giá](#)

**Tìm thấy 3 giá bán Samsung P1000 Galaxy Tab**

**CHÚ Ý :** AHA chỉ cung cấp thông tin về sản phẩm **Samsung P1000 Galaxy Tab** cho quý khách mang tính tham khảo. Vui lòng liên hệ cửa hàng hỏi thông tin về sản phẩm một cách chính xác và an toàn nhất.

Tên sản phẩm	Cửa hàng	Chọn tỉnh	Sắp xếp giá bán
<b>Samsung Galaxy Tab</b> <a href="#">Not online</a> <a href="#">Not online</a> 21 Chùa Bộc, Đống Đa Tel: (04) 3572.8866 <a href="#">Xem thêm địa chỉ &gt;&gt;</a>	 <b>Anh Vũ Mobile</b> 4 nhân xét	Hà Nội	<b>15.850.000 Đ</b> Cập nhật: 03-11-2010, 11:19 am <a href="#">Xem trên website cửa hàng</a>
<b>Samsung P1000 Chic White (Galaxy Tab)</b> <a href="#">Not online</a> 45 Thái Hà Tel: 04.3.537 8899	<a href="#">FTPShop</a> 0 nhân xét	Hà Nội	<b>15.990.000 Đ</b> Cập nhật: 03-11-2010, 12:48 pm <a href="#">Xem trên website cửa hàng</a>
<b>Samsung P1000 Galaxy Tab</b> <a href="#">Not online</a> 136 Trần Phú, Phường 4, Q.5 Tel: 08.3835 3291	<a href="#">Bach Long Mobile</a> 13 nhân xét	TP HCM	<b>17.990.000 Đ</b> Cập nhật: 03-11-2010, 11:44 am <a href="#">Xem trên website cửa hàng</a>

[RSS](#) **So sánh giá bán Điện thoại di động Samsung P1000 Galaxy Tab**

Tham gia thảo luận về Samsung P1000 Galaxy Tab với các thành viên khác tại: [Thảo luận Điện thoại di động Samsung P1000 Galaxy Tab](#)

Tìm trên Denthan.com : [Samsung](#), [P1000](#), [Galaxy](#), [Tab](#).

© 2010 AHA.vn - Website so sánh giá cả hàng đầu Việt nam  
[Đăng ký gian hàng](#) - [Giới thiệu AHA](#) - [Liên hệ](#) - [Diễn đàn](#) - [Sản phẩm mới](#)

- Giá sản phẩm liên tục thay đổi
- Có nhiều cửa hàng
- Làm sao cập nhật?

# Vấn đề

Trang Vàng Gian Hàng Trực tuyến e-Catalogue Danh Ba Website Thông Tin Thị Trường

**Yellow Pages**  
Những Trang Vàng Tìm là thấy

**THIÊN HÒA**

**BẢNG TÀI HÀNH LÝ**

**BẢNG TÀI**

**DOANH NGHIỆP** CƠ QUAN HÀNH CHÍNH - QUẢN LÝ NHÀ NƯỚC NHÀ F

**Ngành nghề: VI TÍNH & TIN HỌC** trong tỉnh, TP TP. HỒ CHÍ MINH

Liệt kê kết quả theo: **Thứ tự quảng cáo** **Thứ tự Alphabet** (Thời gian tìm: 0.968 giây)

Tổng số: **1125** kết quả - Hiện thị: kết quả từ **1** đến **10** 1 2 3 4 5 6 7 8 9 10 Phần sau

**ANH QUÂN - CTY TNHH TIN HỌC ANH QUÂN (QUANTIC)**  
Ngành nghề: VI TÍNH & TIN HỌC  
Địa chỉ: 104 ĐIỆN BIẾN PHÚ, P.ĐK, Q.1, TP. HCM  
[Xem chi tiết](#) | [Bản đồ ...](#) 1

**ĐẠI TƯỜNG PHÁT - CTY TNHH TM - DV ĐẠI TƯỜNG PHÁT**  
Ngành nghề: VI TÍNH & TIN HỌC  
Địa chỉ: 113 TÂN SƠN NHÌ, P.TÂN SƠN NHÌ, Q.TP, TP. HCM  
[Xem chi tiết](#)

**NGUYỄN HOÀNG - CTY CP ĐẦU TƯ PHÁT TRIỂN CÔNG NGHỆ NGUYỄN HOÀNG**  
Ngành nghề: VI TÍNH & TIN HỌC  
Địa chỉ: 207/3 NGUYỄN VĂN THỦ, P.ĐK, Q.1, TP. HCM  
[Xem chi tiết](#) | [Bản đồ ...](#) 2

**QTC - CTY TNHH KỸ THUẬT QTC**  
Ngành nghề: VI TÍNH & TIN HỌC  
Địa chỉ: 17/33 KP7 LINH ĐÔNG, P.LINH ĐÔNG, Q.TĐ, TP. HCM  
[Xem chi tiết](#)

Tôi muốn có  
CSDL để dễ  
dàng truy vấn!

trình hiện đại

# Vấn đề

- ☐ Web 2.0 cung cấp những web API làm nguồn dữ liệu cho các ứng dụng khác
- ☐ Không phải website nào cũng có API
- ☐ Web API không thể nào đáp ứng đủ nhu cầu của người dùng

# Rút trích dữ liệu (Web Data Extraction)

- ☐ Tự động thu thập dữ liệu trên web
- ☐ Lấy dữ liệu từ những nguồn xác định
- ☐ Tái cấu trúc dữ liệu
- ☐ Lưu trữ dữ liệu (đã thu thập được) vào cơ sở dữ liệu để tiện truy vấn

# Thu thập dữ liệu

- ☐ Tự viết từ đầu
- ☐ Sử dụng các mã nguồn có sẵn
  - ☒ Thư viện
    - HTMLAgilityPack, C#
    - HtmlUnit, Java
  - ☒ Ứng dụng nguồn mở
    - Watin, C#
    - Selenium, Firefox add-on



# Tái cấu trúc dữ liệu



The screenshot shows a web interface with a search bar and navigation links. The search results are displayed in a list format, showing details for four companies in the 'VI TÍNH & TIN HỌC' (Computer & IT) industry.

Company Name	Industry	Address
<b>ANH QUÂN - CTY TNHH TIN HỌC ANH QUÂN (QUANTIC)</b>	VI TÍNH & TIN HỌC	104 ĐIỆN BIÊN PHÚ, P.ĐK, Q.1, TP. HCM
<b>ĐẠI TƯỜNG PHÁT - CTY TNHH TM - DV ĐẠI TƯỜNG PHÁT</b>	VI TÍNH & TIN HỌC	113 TÂN SƠN NHÌ, P.TÂN SƠN NHÌ, Q.TP, TP. HCM
<b>NGUYỄN HOÀNG - CTY CP ĐẦU TƯ PHÁT TRIỂN CÔNG NGHỆ NGUYỄN HOÀNG</b>	VI TÍNH & TIN HỌC	207/3 NGUYỄN VĂN THỦ, P.ĐK, Q.1, TP. HCM
<b>QTC - CTY TNHH KỸ THUẬT QTC</b>	VI TÍNH & TIN HỌC	17/33 KP7 LINH ĐÔNG, P.LINH ĐÔNG, Q.TĐ, TP. HCM

- ☐ Xác định dữ liệu quan tâm
- ☐ Thiết lập quan hệ giữa các loại dữ liệu

# Vấn đề khi rút trích dữ liệu

- ☐ Server không hoạt động
- ☐ Bảo mật và hạn chế lưu lượng của web server
- ☐ AJAX
- ☐ Session và cookies
- ☐ Bản quyền

# Dữ liệu và thông tin

## Dữ liệu

- Thô
- Bán cấu trúc
- Hỗn tạp
- Chưa xử lý

## Thông tin

- Rõ ràng
- Có cấu trúc
- Tập trung
- Dựa vào ngữ cảnh để rút thông tin từ dữ liệu

# Từ dữ liệu đến thông tin

- ☐ Thu thập được nhiều dữ liệu
- ☐ Có thêm một phần cấu trúc từ mối liên hệ của trang web
- ☐ Cần thu thập đúng những gì cần

# Từ dữ liệu đến thông tin – Lọc

- ☐ Loại bỏ:
  - ☐ Banner, quảng cáo, ...
  - ☐ Header/Footer
  - ☐ Menu
  - ☐ Các thứ không liên quan đến nội dung chính
- ☐ Sử dụng XPath

# XPath

Loại nút và  
quan hệ

Chọn nút

Vị từ

Axes

Toán tử

Hàm

Tham khảo thêm:

<http://w3schools.com/xpath/default.asp>

<http://msdn.microsoft.com/en-us/library/ms256471.aspx>

# XPath – Chọn nút

Biểu thức	Mô tả
<i>tên nút</i>	Chọn tất cả nút con của <i>tên nút</i>
/	Chọn từ nút gốc
//	Chọn tất cả các nút ở bất cứ đâu miễn thỏa điều kiện
.	Chọn nút hiện tại
..	Chọn nút cha của nút hiện tại
@	Chọn các thuộc tính

# Chọn nút – Ví dụ

Biểu thức	Mô tả
/html	Chọn thẻ html của trang web
//a	Chọn tất cả thẻ a trong trang web
./div/text()	Chọn văn bản của thẻ div con của nút đang xét



# Xpath – Vị từ

- Được đặt trong dấu [ ] nhằm xác định các nút thỏa điều kiện xác định

Biểu thức	Mô tả
<code>//table[@id='content']</code>	Chọn thẻ table có id là content
<code>//table[@id='content']/td[2]</code>	Chọn thẻ td của table có id là content

# Regular Expression

Expresso

Ký hiệu	Ý nghĩa
<code>^ ... \$</code>	Dấu hiệu bắt đầu và kết thúc một Expression
<code>\t</code>	Có chứa Ký tự Tab
<code>\n</code>	Có chứa Ký tự xuống dòng
<code>.</code>	Có chứa Ký tự bất kỳ khác <code>\n</code>
<code>[qwerty]</code>	Có chứa Ký tự bất kỳ trong ngoặc vuông
<code>[^qwerty]</code>	Không chứa ký tự nào trong ngoặc vuông
<code>[a-z]</code>	Có chứa ký tự trong khoảng từ a đến z
<code>\w</code>	Có chứa một từ bất kỳ (word). Tương tự <code>[a-zA-Z0-9]</code>
<code>\W</code>	Có chứa một chuỗi bất kỳ không phải là một từ (nonword)
<code> </code>	Hoặc

# Regular Expression

Ký hiệu	Ý nghĩa
\s	Có chứa ký tự khoảng trắng
\S	Không chứa ký tự khoảng trắng
\d	Có chứa ký tự số
\D	Không phải ký tự số
*	Chỉ định 0 hoặc nhiều
+	Chỉ định 1 hoặc nhiều
?	Chỉ định 0 hoặc 1
{n}	Chỉ định có đúng chính xác n lần
{n,}	Chỉ định có nhiều hơn n lần
{n,m}	Chỉ định có từ n đến m lần

# Rút trích thông tin (Web Information Extraction)

- ☐ Đến vùng thông tin quan tâm (thông qua XPath trên XHTML)
- ☐ Xem dữ liệu là các đối tượng
- ☐ Tìm kiếm các đối tượng liên quan
- ☐ Gán nhãn thuộc tính cho các đối tượng

# Vấn đề khi rút trích thông tin

- ☐ Nguồn dữ liệu không có cấu trúc tốt
- ☐ Cấu trúc thông tin thay đổi liên tục
- ☐ Nhận dạng sai thông tin
- ☐ Thông tin rút trích bị lỗi

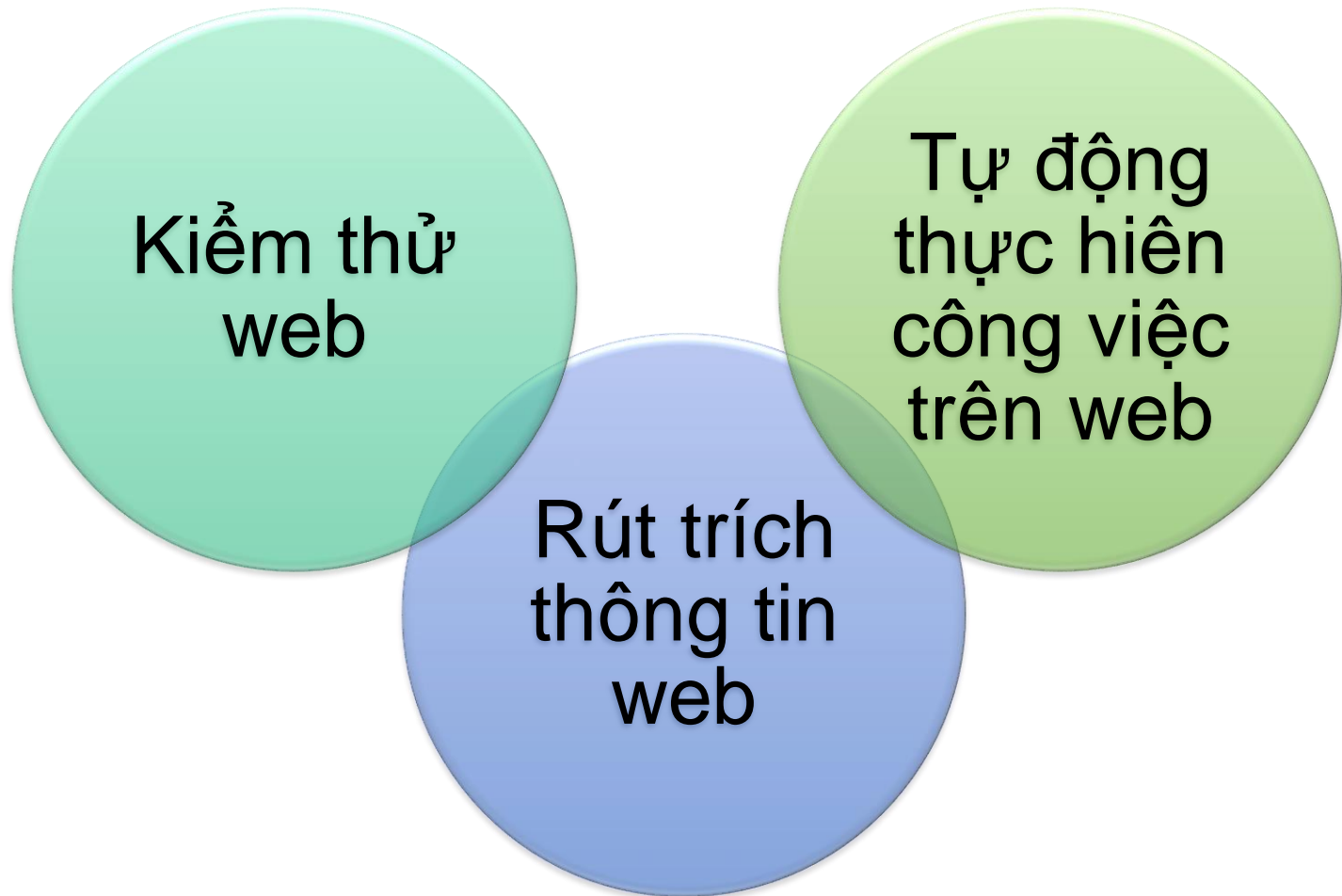
# Tự động hóa web (Web Automation)

- ☐ Giả lập hành động duyệt web của người dùng
- ☐ Các hành động bao gồm
  - ☐ Bấm liên kết
  - ☐ Điền văn bản vào TextBox
  - ☐ Bấm Button
  - ☐ Submit dữ liệu
  - ☐ Chọn ComboBox
  - ☐ ...

# Tự động hóa Web - Phân loại

Sử dụng trình duyệt	Thông qua giao thức HTTP
<ul style="list-style-type: none"><li>● Dựa trên các trình duyệt thông dụng (IE, Firefox)<ul style="list-style-type: none"><li>● Xây dựng Add-on (iMacros)</li><li>● Xây dựng ứng dụng khác điều khiển trình duyệt (Watin)</li><li>● Sử dụng WebBrowser control (csEXWB)</li></ul></li></ul>	<ul style="list-style-type: none"><li>● Sử dụng GET, POST<ul style="list-style-type: none"><li>● Gửi và nhận các gói tin tương ứng với các hành động và xử lý kết quả trả về của server</li></ul></li></ul>
<ul style="list-style-type: none"><li>● Thao tác chậm</li></ul>	<ul style="list-style-type: none"><li>● Thao tác nhanh</li></ul>
<ul style="list-style-type: none"><li>● Xử lý được hầu hết các hành động duyệt web</li></ul>	<ul style="list-style-type: none"><li>● Khó thực hiện các hành động phức tạp</li></ul>
<ul style="list-style-type: none"><li>● Hoạt động giống như người duyệt web</li></ul>	<ul style="list-style-type: none"><li>● Dễ bị phát hiện thao tác do máy thực hiện</li></ul>

# Tự động hóa Web - Ứng dụng





# Các bước thực hiện

Xác định  
mục tiêu

Theo dõi và  
ghi nhận  
hành động

Tự động  
thực hiện lại  
hành động

# Tự động hóa Web – Cách thức

## Sử dụng trình duyệt

- Xác định các phần tử HTML trên cây DOM
- Ghi nhận các sự kiện trên phần tử HTML

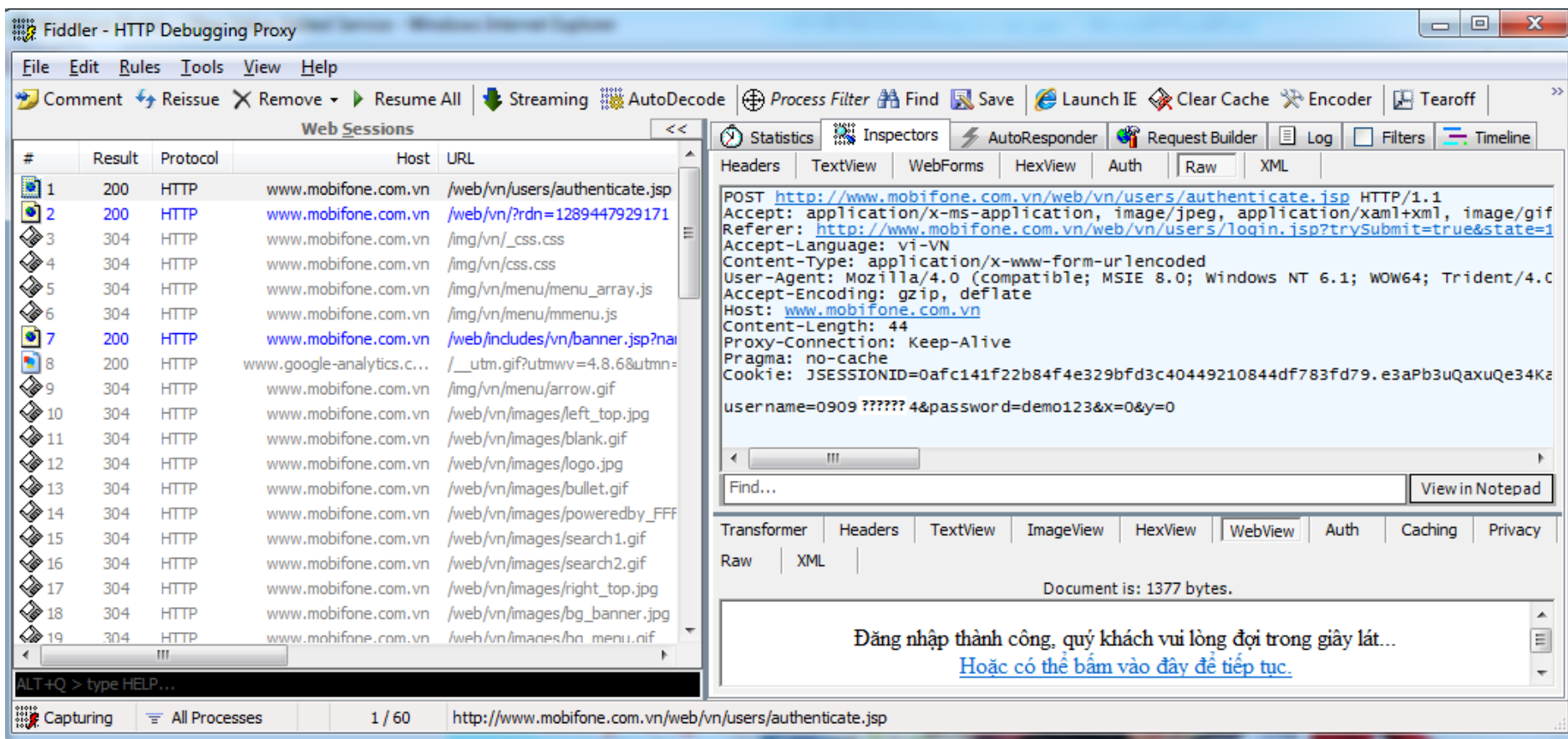
## Thông qua giao thức HTTP

- Ghi nhận các yêu cầu GET, POST
- Quản lý session và cookie

- Firefox Add-on: Firebug
- IE/Chrome Development tools

- Fiddler

# Đọc POST bằng Fiddler



**Fiddler - HTTP Debugging Proxy**

File Edit Rules Tools View Help

Comment Reissue Remove Resume All Streaming AutoDecode Process Filter Find Save Launch IE Clear Cache Encoder Tearoff

**Web Sessions**

#	Result	Protocol	Host	URL
1	200	HTTP	www.mobifone.com.vn	/web/vn/users/authenticate.jsp
2	200	HTTP	www.mobifone.com.vn	/web/vn/?rdn=1289447929171
3	304	HTTP	www.mobifone.com.vn	/img/vn/_css.css
4	304	HTTP	www.mobifone.com.vn	/img/vn/css.css
5	304	HTTP	www.mobifone.com.vn	/img/vn/menu/menu_array.js
6	304	HTTP	www.mobifone.com.vn	/img/vn/menu/mmenu.js
7	200	HTTP	www.mobifone.com.vn	/web/includes/vn/banner.jsp?nai
8	200	HTTP	www.google-analytics.c...	/_utm.gif?utmwv=4.8.6&utmn=
9	304	HTTP	www.mobifone.com.vn	/img/vn/menu/arrow.gif
10	304	HTTP	www.mobifone.com.vn	/web/vn/images/left_top.jpg
11	304	HTTP	www.mobifone.com.vn	/web/vn/images/blank.gif
12	304	HTTP	www.mobifone.com.vn	/web/vn/images/logo.jpg
13	304	HTTP	www.mobifone.com.vn	/web/vn/images/bullet.gif
14	304	HTTP	www.mobifone.com.vn	/web/vn/images/poweredby_FFF
15	304	HTTP	www.mobifone.com.vn	/web/vn/images/search1.gif
16	304	HTTP	www.mobifone.com.vn	/web/vn/images/search2.gif
17	304	HTTP	www.mobifone.com.vn	/web/vn/images/right_top.jpg
18	304	HTTP	www.mobifone.com.vn	/web/vn/images/bg_banner.jpg
19	304	HTTP	www.mobifone.com.vn	/web/vn/images/bg_menu.gif

**Inspectors**

Headers TextView WebForms HexView Auth Raw XML

POST <http://www.mobifone.com.vn/web/vn/users/authenticate.jsp> HTTP/1.1  
 Accept: application/x-ms-application, image/jpeg, application/xaml+xml, image/gif  
 Referer: <http://www.mobifone.com.vn/web/vn/users/login.jsp?trySubmit=true&state=1>  
 Accept-Language: vi-VN  
 Content-Type: application/x-www-form-urlencoded  
 User-Agent: Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 6.1; WOW64; Trident/4.0  
 Accept-Encoding: gzip, deflate  
 Host: [www.mobifone.com.vn](http://www.mobifone.com.vn)  
 Content-Length: 44  
 Proxy-Connection: Keep-Alive  
 Pragma: no-cache  
 Cookie: JSESSIONID=0afcf141f22b84f4e329bfd3c40449210844df783fd79.e3aPb3uQaxuQe34Ka  
 username=0909 ?????? 4&password=demo123&x=0&y=0

Find... View in Notepad

Transformer Headers TextView ImageView HexView WebView Auth Caching Privacy

Raw XML

Document is: 1377 bytes.

Đăng nhập thành công, quý khách vui lòng đợi trong giây lát...  
[Hoặc có thể bấm vào đây để tiếp tục.](#)

ALT+Q > type HELP...

Capturing All Processes 1 / 60 http://www.mobifone.com.vn/web/vn/users/authenticate.jsp

# Tự động hóa Web – Vấn đề

## ☐ Thay đổi

- ☐ Cấu trúc trang web thay đổi
- ☐ Website ngăn cấm các hành động lặp đi lặp lại

## ☐ Trạng thái

- ☐ Session/Cookie
- ☐ Xác định kết quả thực hiện của các hành động

# Bài tập



- ☐ Xây dựng feed tin tức cho trang chủ của khoa Công nghệ thông tin, trường Đại học Khoa học Tự nhiên.
- ☐ Thang điểm:
  - ☐ Rút trích được danh sách tin từ web khoa: 4 điểm. Bao gồm:
    - Tựa đề của tin
    - Ngày đăng.
  - ☐ Xuất kết quả dạng RSS bằng RESTful web service: 4 điểm.
  - ☐ Đăng tải lên host thực tế để có thể đọc được bằng Feed reader (Feedly, ...): 2 điểm.



# Tham khảo

□ <http://www.strathweb.com/2012/04/rss-atom-mediatypeformatter-for-asp-net-webapi/>