Trường Đại học Khoa học Tự nhiên
Khoa Công nghệ Thông tin

TÀI LIỆU LÝ THUYẾT KHAI THÁC WEB

GIỚI THIỆU MÔN HỌC

Giảng viên: ThS. Nguyễn Ngọc Thảo
Email: nnthao@fit.hcmus.edu.vn

HKII/2011-2012

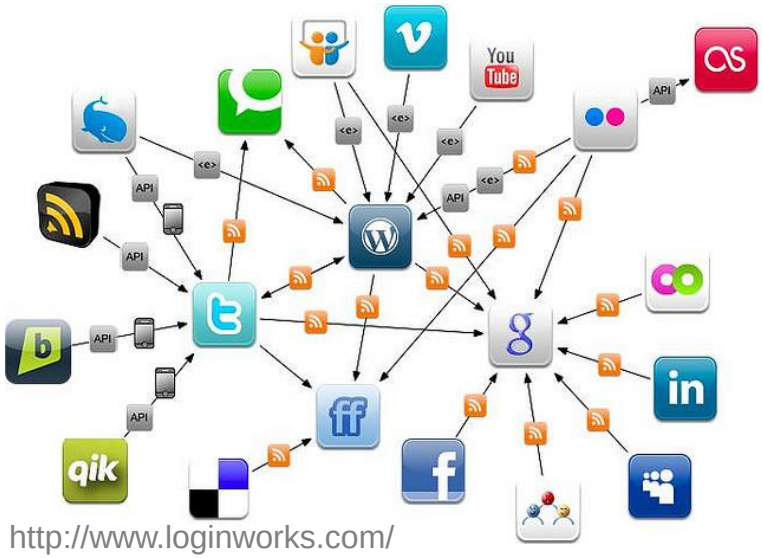
NỘI DUNG

- Giới thiệu Khai thác Web
  - Nội dung môn học
  - Mục tiêu môn học
- Kế hoạch học tập
  - Thang điểm và hình thức đánh giá
  - Nội dung học theo tuần
- Tài liệu tham khảo

GIỚI THIỆU KHAI THÁC WEB

WEB LÀ GÌ?

- Web là tập hợp các tài liệu (trang Web) liên kết với nhau thông qua Internet.
- Các trang tài liệu này được lưu trữ ở một hay nhiều máy chủ.



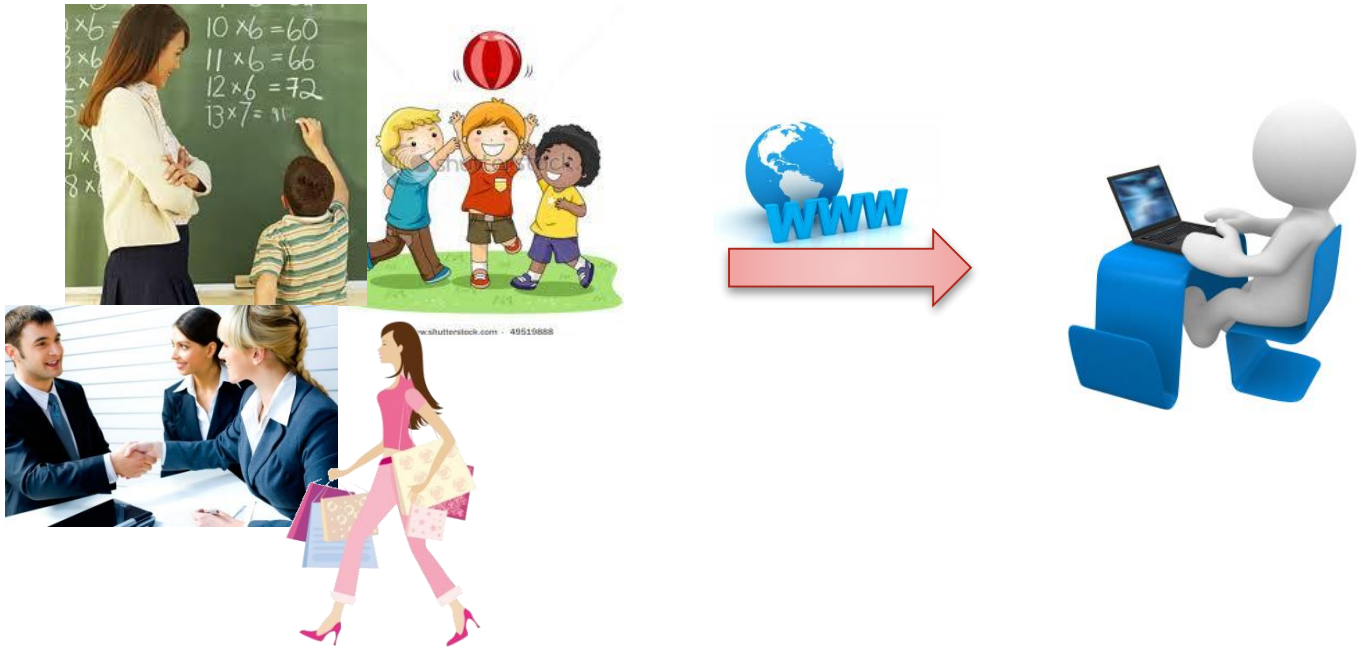
http://www.loginworks.com/

ĐỊNH DẠNG THÔNG TIN

- Trang Web có thể chứa nhiều dạng thông tin khác nhau.

VAI TRÒ CỦA WEB

- Sự ra đời của Web đã tác động sâu sắc đến nhiều khía cạnh của đời sống.
  - Ví dụ: trào lưu mua sắm trực tuyến, chương trình giáo dục từ xa



KHAI THÁC WEB LÀ GÌ?

- Khai thác Web nhằm phát hiện và rút trích tri thức hữu ích từ dữ liệu Web.
- Dữ liệu Web là
  - Nội dung Web – văn bản, hình ảnh, bản ghi,...
  - Cấu trúc Web – hyperlinks, thẻ (HTML),...
  - Hành vi sử dụng Web – http logs, app server logs,...

  $\Rightarrow$ *Dữ liệu thô*

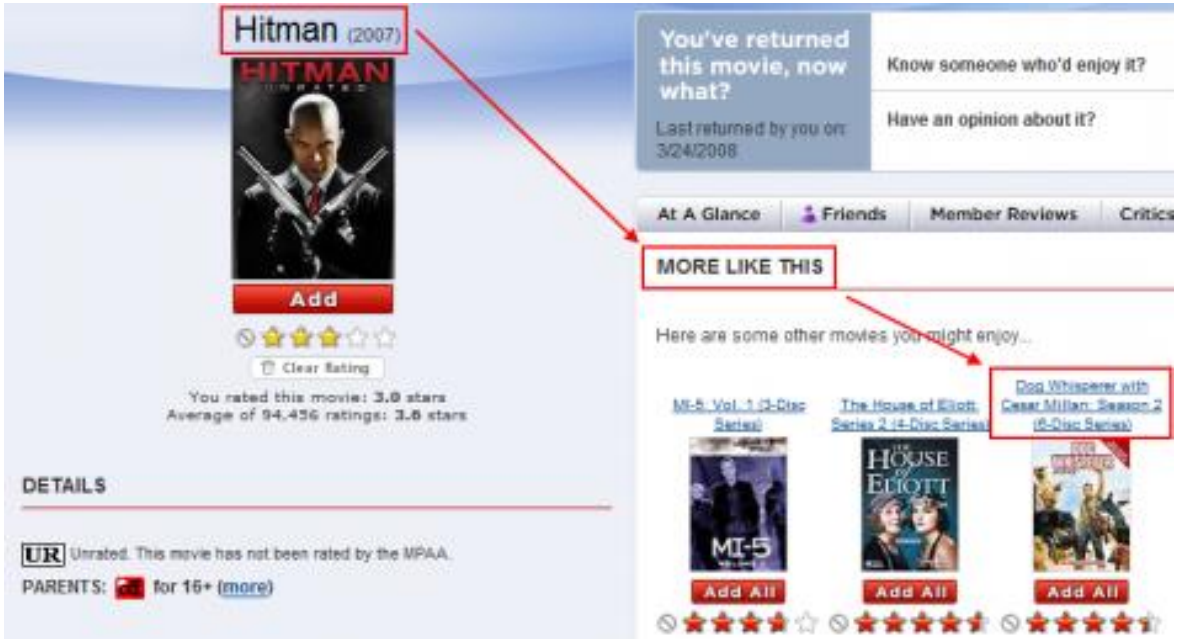
DỮ LIỆU THÔ VÀ TRI THỨC

- Dữ liệu thô và tri thức khác nhau thế nào?
  - Dữ liệu thô: là dữ liệu ở dạng nguyên thủy thu thập được từ một nguồn dữ liệu
    - Ví dụ: hóa đơn bán hàng ở siêu thi, web-log tại server,...
  - Tri thức: là mối quan hệ tiềm ẩn trong dữ liệu
    - Ví dụ: hóa đơn bán hàng $\Rightarrow$ thói quen mua sắm của khách, mua bột giặt sẽ mua nước xả vải

Khai thác

ỨNG DỤNG KHAI THÁC WEB

- Giải trí: các hệ thống tư vấn sản phẩm dựa trên sở thích người dùng.
  - Ví dụ: trang Web xem phim (IMDB, Netflix), hệ thống bán hàng trực tuyến Amazon.com



ỨNG DỤNG KHAI THÁC WEB

- Kinh doanh: thu thập ý kiến khách hàng về sản phẩm để định hướng cải tiến.
  - Ví dụ: Apple muốn biết đánh giá của người dùng về Ipad2
    - Khảo sát, lập bình chọn 👍 – chậm, ít người tham gia
    - Thu thập từ trang bán hàng, blog 👎 – nhanh, chân thật

★☆☆☆☆ **Overpriced piece of junk**
Extremely flimsy. Display is lack luster. Apps that one can receive free on an android, you have to pay for via the app store.
Read more
Published 2 days ago by :)

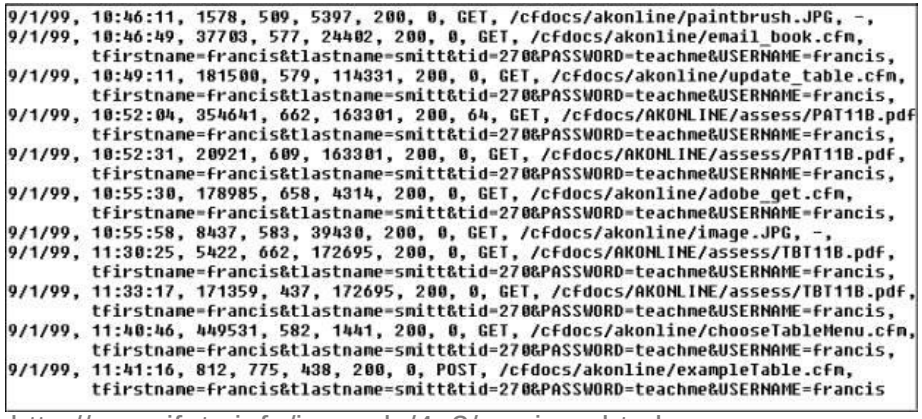
★★★★★ **Awesome**
Love the ipad as it is so easy to navigate and very handy to use anywhere around the house. It is lightweight and turns on fast.

★★★★★ **show**
The product is excellent. Reached all my expectations.
It is easy to use and each day I discover a new application. Read more
Published 1 day ago by mhendrick

Từ mang cảm nghĩ

ỨNG DỤNG KHAI THÁC WEB

- Giáo dục: nâng cao chất lượng các hệ thống giáo dục từ xa.
  - Ví dụ: sử dụng Web-log để quản lý sinh viên, tái bố trí bài giảng,...



```
9/1/99, 10:46:11, 1578, 509, 5397, 200, 0, GET, /cfdocs/akonline/paintbrush.JPG, -,
9/1/99, 10:46:49, 37703, 577, 24402, 200, 0, GET, /cfdocs/akonline/email_book.cfm,
        tfirstname=francis&tlastname=smitt&tid=270&PASSWORD=teachme&USERNAME=francis,
9/1/99, 10:49:11, 181500, 579, 114331, 200, 0, GET, /cfdocs/akonline/update_table.cfm,
        tfirstname=francis&tlastname=smitt&tid=270&PASSWORD=teachme&USERNAME=francis,
9/1/99, 10:52:04, 354641, 662, 163301, 200, 64, GET, /cfdocs/AKONLINE/assess/PAT11B.pdf
        tfirstname=francis&tlastname=smitt&tid=270&PASSWORD=teachme&USERNAME=francis,
9/1/99, 10:52:31, 20921, 609, 163301, 200, 0, GET, /cfdocs/AKONLINE/assess/PAT11B.pdf,
        tfirstname=francis&tlastname=smitt&tid=270&PASSWORD=teachme&USERNAME=francis,
9/1/99, 10:55:30, 178985, 658, 4314, 200, 0, GET, /cfdocs/akonline/adobe_get.cfm,
        tfirstname=francis&tlastname=smitt&tid=270&PASSWORD=teachme&USERNAME=francis,
9/1/99, 10:55:58, 8437, 583, 39430, 200, 0, GET, /cfdocs/akonline/image.JPG, -,
9/1/99, 11:30:25, 5422, 662, 172695, 200, 0, GET, /cfdocs/AKONLINE/assess/TBT11B.pdf,
        tfirstname=francis&tlastname=smitt&tid=270&PASSWORD=teachme&USERNAME=francis,
9/1/99, 11:33:17, 171359, 437, 172695, 200, 0, GET, /cfdocs/AKONLINE/assess/TBT11B.pdf,
        tfirstname=francis&tlastname=smitt&tid=270&PASSWORD=teachme&USERNAME=francis,
9/1/99, 11:40:46, 449531, 582, 1441, 200, 0, GET, /cfdocs/akonline/chooseTableMenu.cfm,
        tfirstname=francis&tlastname=smitt&tid=270&PASSWORD=teachme&USERNAME=francis,
9/1/99, 11:41:16, 812, 775, 438, 200, 0, POST, /cfdocs/akonline/exampleTable.cfm,
        tfirstname=francis&tlastname=smitt&tid=270&PASSWORD=teachme&USERNAME=francis
```

http://www.ifets.info/journals/4_2/garrison.html

Web Server Log

Số lượt truy cập, tần số truy cập,...

Mẫu truy cập (vào những học phần nào, thứ tự truy cập,...)

ỨNG DỤNG KHAI THÁC WEB

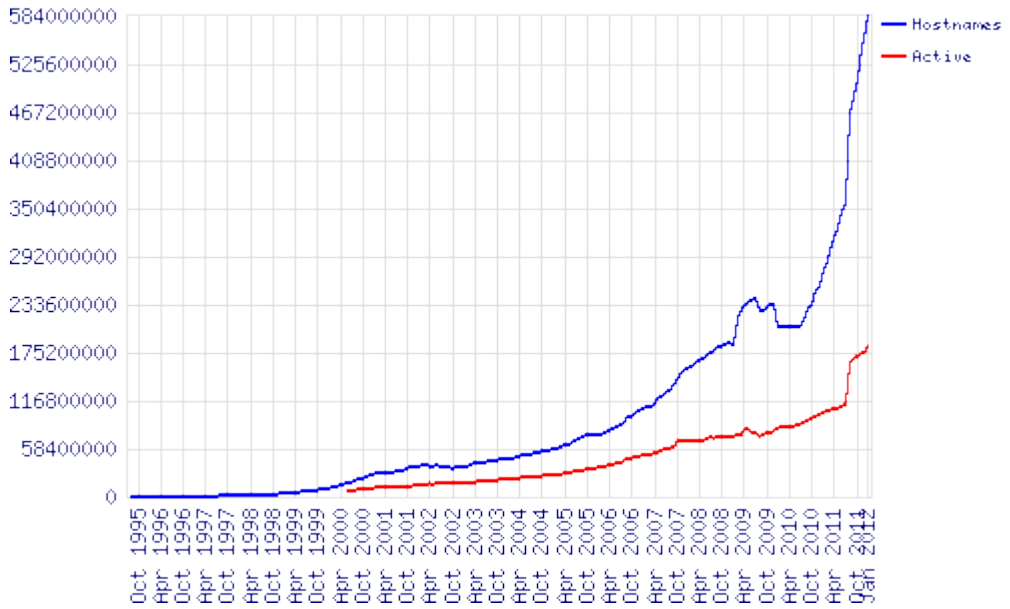
- Khoa học:
  - Cung cấp nhiều nguồn dữ liệu đa dạng về nội dung và hình thức.

http://www.zdnet.com/

  - Thúc đẩy sự phát triển các thuật toán, kỹ thuật phù hợp với đặc thù của dữ liệu Web.
    - Dữ liệu Web là dữ liệu động, kích thước lớn, nhiễu cao, chứa nhiều dạng dữ liệu và không có cấu trúc chuẩn

NHU CẦU KHAI THÁC WEB

- Web phát triển nhanh chóng trong hai thập kỷ qua.
- Dữ liệu Web là nguồn dữ liệu khổng lồ, dễ dàng truy cập và không ngừng phát triển.



Tổng số site tính trên mọi domain (08/1995 – 01/2012)

http://news.netcraft.com/archives/2012/page/2/

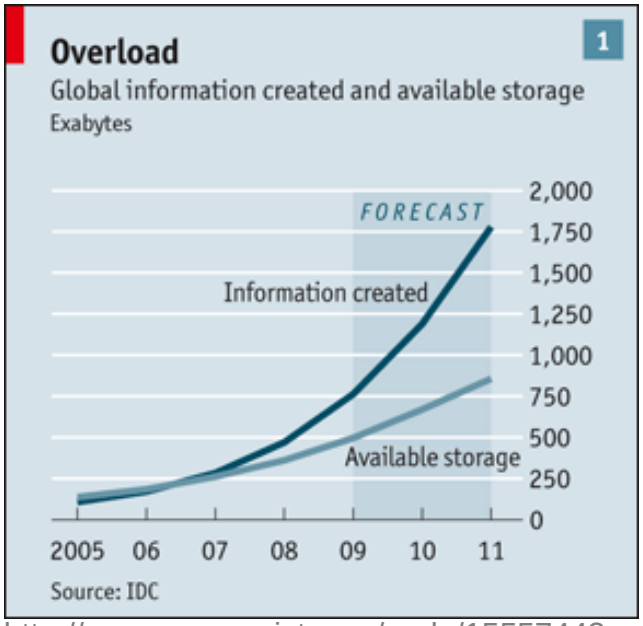
NHU CẦU KHAI THÁC WEB

- Ta không thể phân tích dữ liệu bằng tay.
  - Con người cần hàng tuần để khám phá tri thức có ích từ dữ liệu.
    - Ví dụ: tìm các trang web có chứa từ “bóng đá”
      - Google trả về 113,000,000 kết quả trong 0.23 giây
      - Giả sử mỗi trang web chỉ chứa 10 từ, tốc độ xử lý của người là 200 từ/phút $\Rightarrow$ cần ~ 11 năm

“*We are drowning in information, but we are starved for knowledge*”.
(John Naisbitt, 1982)

NHU CẦU KHAI THÁC WEB

- Ta không thể phân tích dữ liệu bằng tay.
  - Phần lớn dữ liệu chưa bao giờ được phân tích



http://www.economist.com/node/15557443

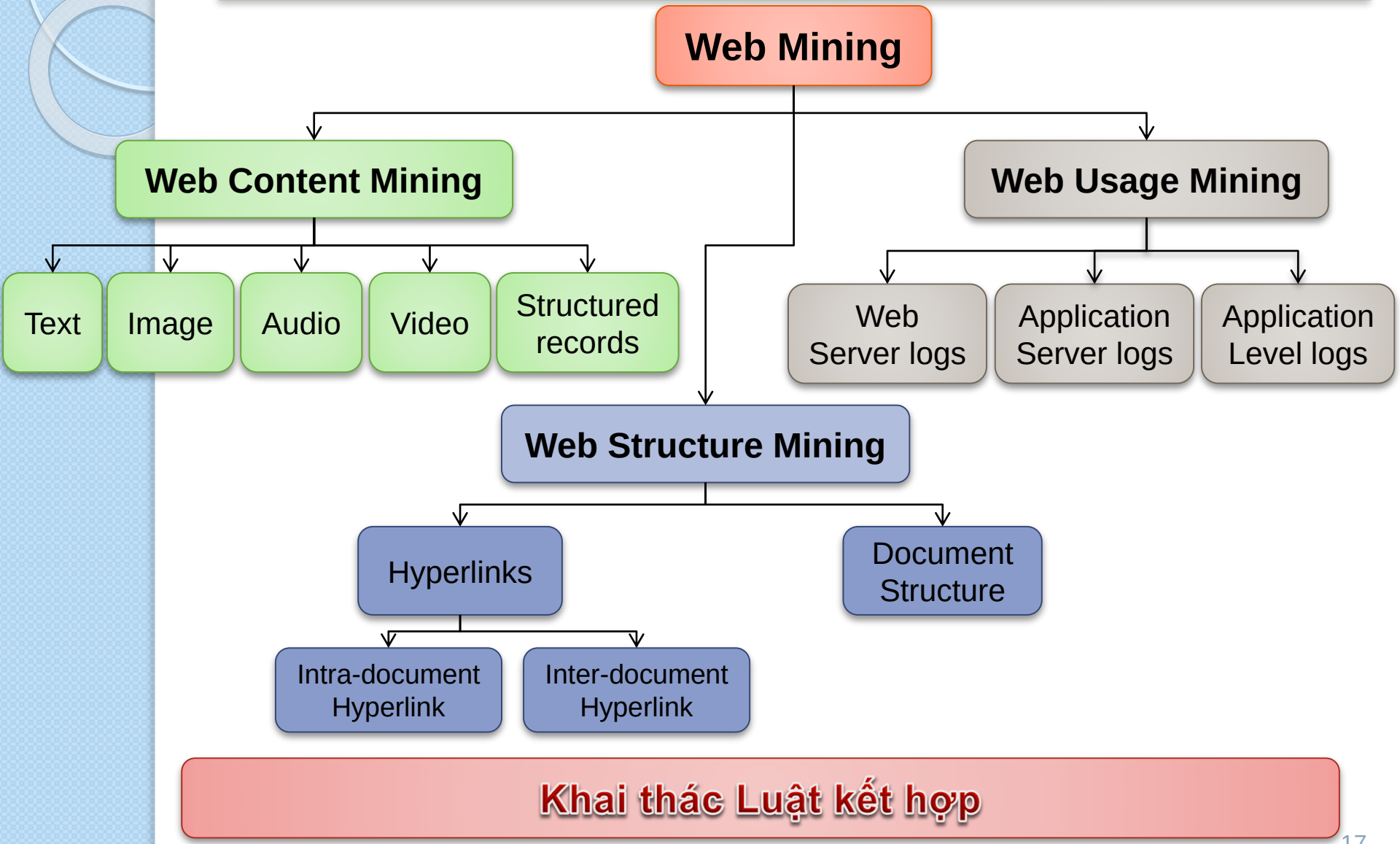
Biểu đồ so sánh giữa dữ liệu phát sinh và khả năng lưu trữ

"*A rapidly growing gap between our ability to generate data, and our ability to make use of it*". (Usama Fayyad)

LĨNH VỰC KHAI THÁC WEB

- Phần lớn các kỹ thuật khai thác Web dựa trên phương pháp khai thác dữ liệu, học máy và truy tìm thông tin.
  - Các kỹ thuật xử lý ngôn ngữ tự nhiên ngày càng chiếm vị trí quan trọng.
  - Thống kê xác suất, quản lý cơ sở dữ liệu, công nghệ đa phương tiện,...

BÀI TOÁN KHAI THÁC WEB



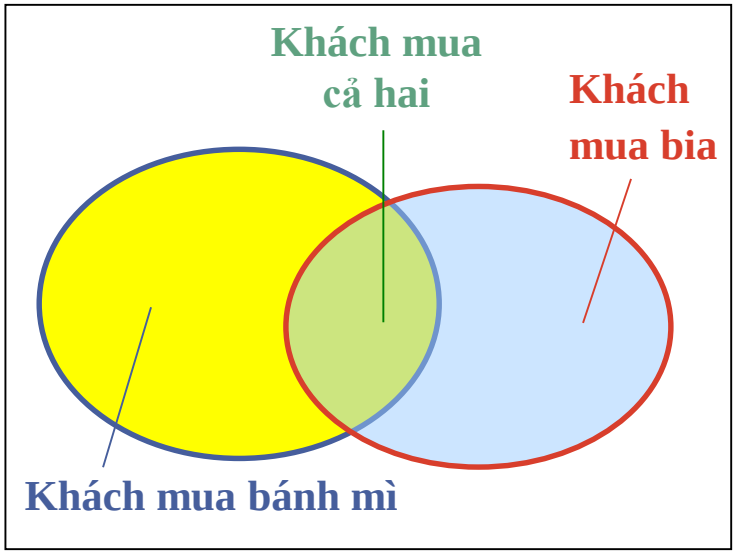
KHAI THÁC LUẬT KẾT HỢP

- Mục tiêu: Cho trước tập dữ liệu, tìm luật dự đoán sự xuất hiện của một hạng mục dựa trên sự xuất hiện của các hạng mục khác.

Market Basket

Mã giao dịch	Món hàng đã mua
10	Bánh mì, Sữa, Bia
20	Bánh mì, Bia
30	Bánh mì, Nước ngọt
40	Sữa, Cà phê, Trà

Bánh mì ⇒ Bia (supp = 50%)



KHAI THÁC LUẬT KẾT HỢP

- Được áp dụng để giải quyết các bài toán khai thác Web.
  - Phân loại nội dung trang Web theo một số chủ đề cho trước.
  - Tìm mẫu hành vi của người dùng từ chuỗi trang viếng thăm.
    - Ví dụ: 60% người dùng truy cập đường dẫn /home/products/file1.html, sẽ đi theo chuỗi /home ==> /home/whatsnew ==> /home/products ==> /home/products/file1.html

KHAI THÁC NỘI DUNG WEB

- Là quá trình rút trích thông tin hữu ích từ nội dung của các tài liệu Web.
  - Dữ liệu Web có thể bao gồm văn bản, hình ảnh, âm thanh, video hoặc bản tin cấu trúc (danh sách, bảng biểu).
- Bài toán cần sự hỗ trợ của các kỹ thuật truy tìm thông tin và xử lý ngôn ngữ tự nhiên.

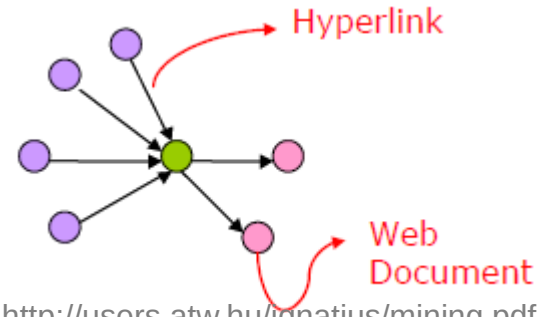
KHAI THÁC NỘI DUNG WEB

- Ứng dụng khai thác nội dung Web
  - Nhận diện các chủ đề trong tài liệu Web
  - Phân loại các tài liệu Web
  - Tìm các trang Web giống nhau trên nhiều server
  - Các ứng dụng liên quan truy vấn nội dung

http://blog.alchemyapi.com/?p=62

KHAI THÁC CẤU TRÚC WEB

- Cấu trúc của một đồ thị Web cơ bản gồm:
  - Trang Web là nút
  - Siêu liên kết là cạnh nối giữa hai trang liên quan



http://users.atw.hu/ignatius/mining.pdf

Cấu trúc đồ thị Web

- Khai thác cấu trúc Web là quá trình phát hiện thông tin cấu trúc từ Web.
  - Có hai mức độ: Mức tài liệu (intra-page) hoặc mức liên kết (inter-page)

KHAI THÁC CẤU TRÚC WEB

- Ứng dụng khai thác cấu trúc Web
  - Xác định chất lượng của trang Web: uy tín của trang Web về một chủ đề nào đó, xếp hạng trang Web
  - Phát hiện các cấu trúc Web thú vị: co-citation, đồ thị hai phía đầy đủ
  - Phân loại trang Web theo nhiều chủ đề khác nhau
  - Chọn lựa trang Web để thu thập dữ liệu
  - Tìm kiếm trang Web liên quan với một trang cho trước
  - Phát hiện trang Web trùng

KHAI THÁC HÀNH VI WEB

- Phát hiện các mẫu có ý nghĩa từ dữ liệu giao dịch client-server trên một hay nhiều vị trí trên Web.
- Các nguồn dữ liệu cơ bản
  - Dữ liệu tự động phát sinh lưu ở server: access logs, referrer logs, agent logs, và cookies phía client
  - Hồ sơ người dùng
  - Siêu dữ liệu: thuộc tính trang, thuộc tính nội dung, dữ liệu hành vi

KHAI THÁC HÀNH VI WEB

- Ứng dụng khai thác hành vi Web
  - Hỗ trợ sự phát triển của thương mại điện tử: các hệ thống tư vấn sản phẩm cho người dùng (Amazon.com – online "Wal-Mart")
  - Phát hiện các thông tin hữu ích tiềm ẩn trong dữ liệu
    - Ví dụ: ý kiến của người dùng về một sản phẩm, hành vi phổ biến của người dùng để nâng cao chất lượng dịch vụ.

MỤC TIÊU MÔN HỌC

- Môn học cung cấp cho sinh viên kiến thức cơ bản về khai thác Web.
  - Các bài toán và kĩ thuật khai thác Web
  - Ứng dụng khai thác Web thực tế
- Sinh viên sau khi hoàn thành khóa học sẽ
  - Nắm vững kiến thức cơ bản về khai thác Web
  - Có khả năng chọn lựa phương pháp xử lý dữ liệu Web phù hợp với ứng dụng cụ thể

KẾ HOẠCH HỌC TẬP

THANG ĐIỂM

- Lý thuyết: 7 điểm
  - Bài tập lý thuyết giữa kỳ 2 điểm
  - Thi viết cuối kỳ 5 điểm
- Thực hành: 3 điểm
- Điểm cộng: 2 điểm
- **Tổng cộng: 12 điểm**

Sinh viên phải tham gia thực hành và thi lý thuyết cuối kỳ.

HÌNH THỨC ĐÁNH GIÁ

- Bài tập lý thuyết giữa kỳ: nhận đề bài và nộp bài làm qua Moodle.
- Thi cuối kỳ:
  - Thi viết, thời điểm theo qui định của Khoa.
  - Thời gian: 90 phút.
  - Được sử dụng tài liệu giấy và máy tính cầm tay.

HÌNH THỨC ĐÁNH GIÁ

- Thực hành:
  - Sử dụng công cụ khai thác dữ liệu WEKA
  - Cài đặt chương trình đơn giản bằng C#/C++
  - Thực nghiệm và viết báo cáo
- Bài tập cộng điểm:
  - Tìm hiểu chủ đề mở rộng do giáo viên gợi ý
  - Cài đặt (nếu có) và viết báo cáo
  - Chỉ áp dụng cho một số nhóm đăng ký sớm

NỘI DUNG HỌC THEO TUẦN

Tuần	Nội dung
1	Giới thiệu sơ nét về các nội dung trong môn học Phổ biến nội quy môn học: hình thức đánh giá, thông tin tài liệu tham khảo
2	Chương 1 – Tổng quan về Khai thác Web
3	Chương 2 – Khai thác luật kết hợp và mẫu tuần tự
4	Chương 2 – Khai thác luật kết hợp và mẫu tuần tự
5	Chương 2 – Khai thác luật kết hợp và mẫu tuần tự Bài tập lý thuyết giữa kỳ
6	Chương 3 – Khai thác nội dung Web
7	Chương 3 – Khai thác nội dung Web
8	Chương 3 – Khai thác nội dung Web

NỘI DUNG HỌC THEO TUẦN

Tuần	Nội dung
9	Chương 4 – Khai thác cấu trúc Web
10	Chương 4 – Khai thác cấu trúc Web
11	Chương 4 – Khai thác cấu trúc Web
12	Chương 5 – Khai thác hành vi sử dụng Web
13	Chương 5 – Khai thác hành vi sử dụng Web
14	Chương 6 – Các xu hướng phát triển trong tương lai
15	Tổng kết và ôn tập

NỘI QUI MÔN HỌC

- Sinh viên chuẩn bị trước nội dung lý thuyết và thực hành trước mỗi buổi học.
- Các bài làm giống nhau (≥ 50%) đều bị điểm 0. Sinh viên có trách nhiệm bảo quản bài làm của mình.

TÀI LIỆU THAM KHẢO

- B. Liu, *Web Data Mining-Exploring Hyperlinks, Contents, and Usage Data*, Springer Series on Data-Centric Systems and Applications, 2007.
- S. Chakrabarti, *Mining the Web: Discovering Knowledge from Hypertext Data*, Morgan Kauffman, 2003.
- J.Han, M.Kamber, Data Mining: Concepts & Technique, 2nd edition, Morgan Kauffman, 2006.
- Phần mềm WEKA http://www.cs.waikato.ac.nz/ml/weka/

KẾT THÚC