ĐẠI HỌC QUỐC GIA TPHCM
TRƯỜNG ĐẠI HỌC
CÔNG NGHỆ THÔNG TIN

TRUY XUẤT THÔNG TIN

CHƯƠNG IV – XÂY DỰNG SEARCH ENGINE

Nguyễn Trọng Chỉnh
chinhnt@uit.edu.vn

NỘI DUNG TRÌNH BÀY

- GIỚI THIỆU
- WEB CRAWLER
- TÁCH NỘI DUNG
- LẬP CHỈ MỤC VÀ TRUY XUẤT

GIỚI THIỆU

❖SEARCH ENGINE

- Search Engine cho nội dung văn bản.
- Có chức năng như một thư viện số.
- Không cần phân loại tài liệu trước khi tìm kiếm.
- Tìm theo từ khóa do người sử dụng chọn.

GIỚI THIỆU

❖CÁC DẠNG SEARCH ENGINE

- **Personal computer search:**
  - Tập tài liệu trên một máy tính cá nhân (file).
  - Khối lượng dữ liệu tương đối nhỏ.
- **Domain-specific search**
  - Tập tài liệu trên hệ thống mạng của một tổ chức (web page, file)
  - Nội dung xác định trước (có thể xử lý ngữ nghĩa)
  - Khối lượng dữ liệu không quá lớn

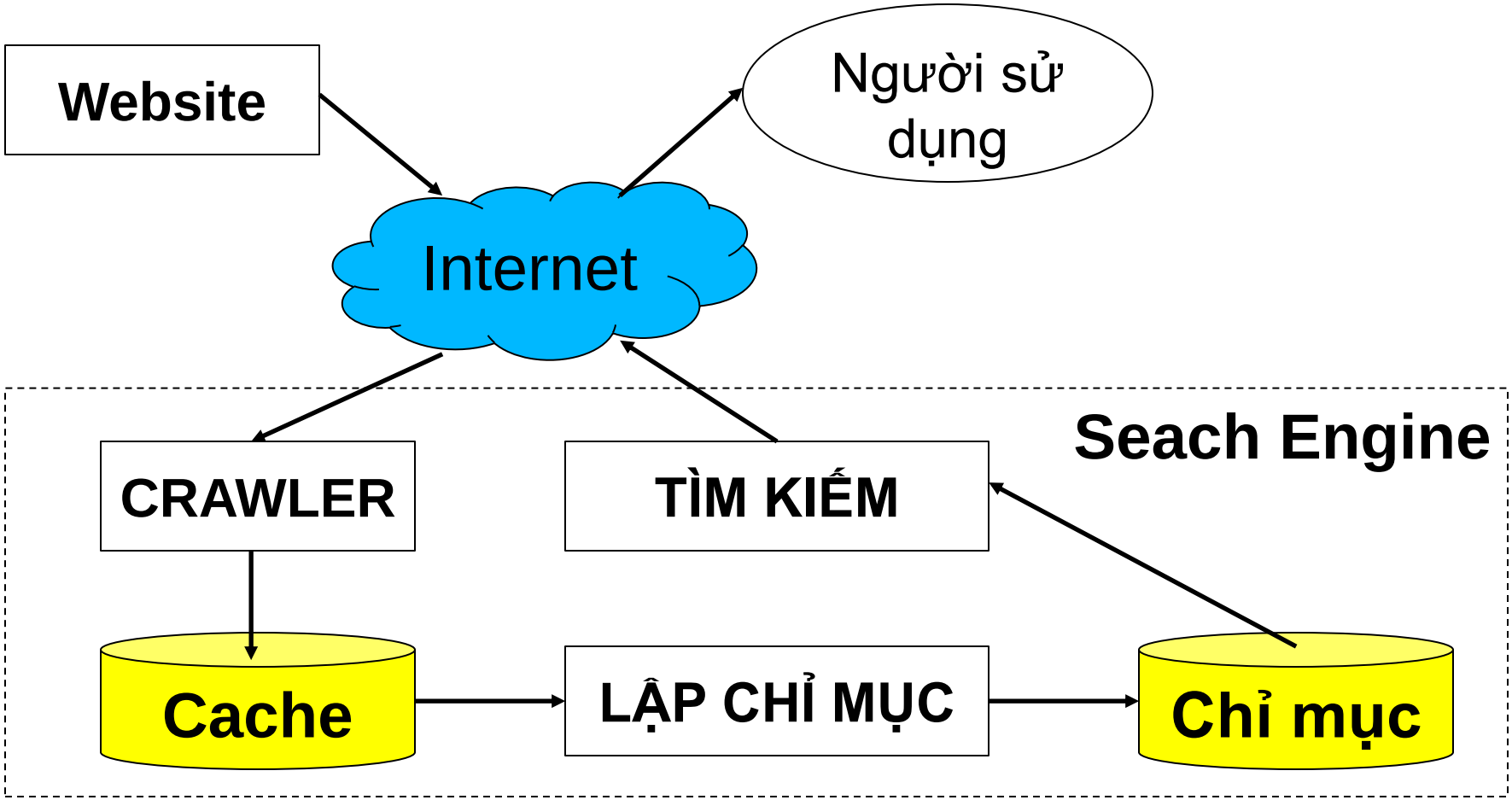
GIỚI THIỆU

❖CÁC DẠNG SEARCH ENGINE

- **Web search: tương tự Domain-specific search**
  - Tài liệu trên các website (web page, file)
  - Nội dung đa dạng (khó đảm bảo ngữ nghĩa)
  - Khối lượng tài liệu rất lớn (vấn đề hiệu quả trong việc lập chỉ mục và tìm kiếm)

GIỚI THIỆU

❖MÔ HÌNH CỦA WEB SEARCH ENGINE



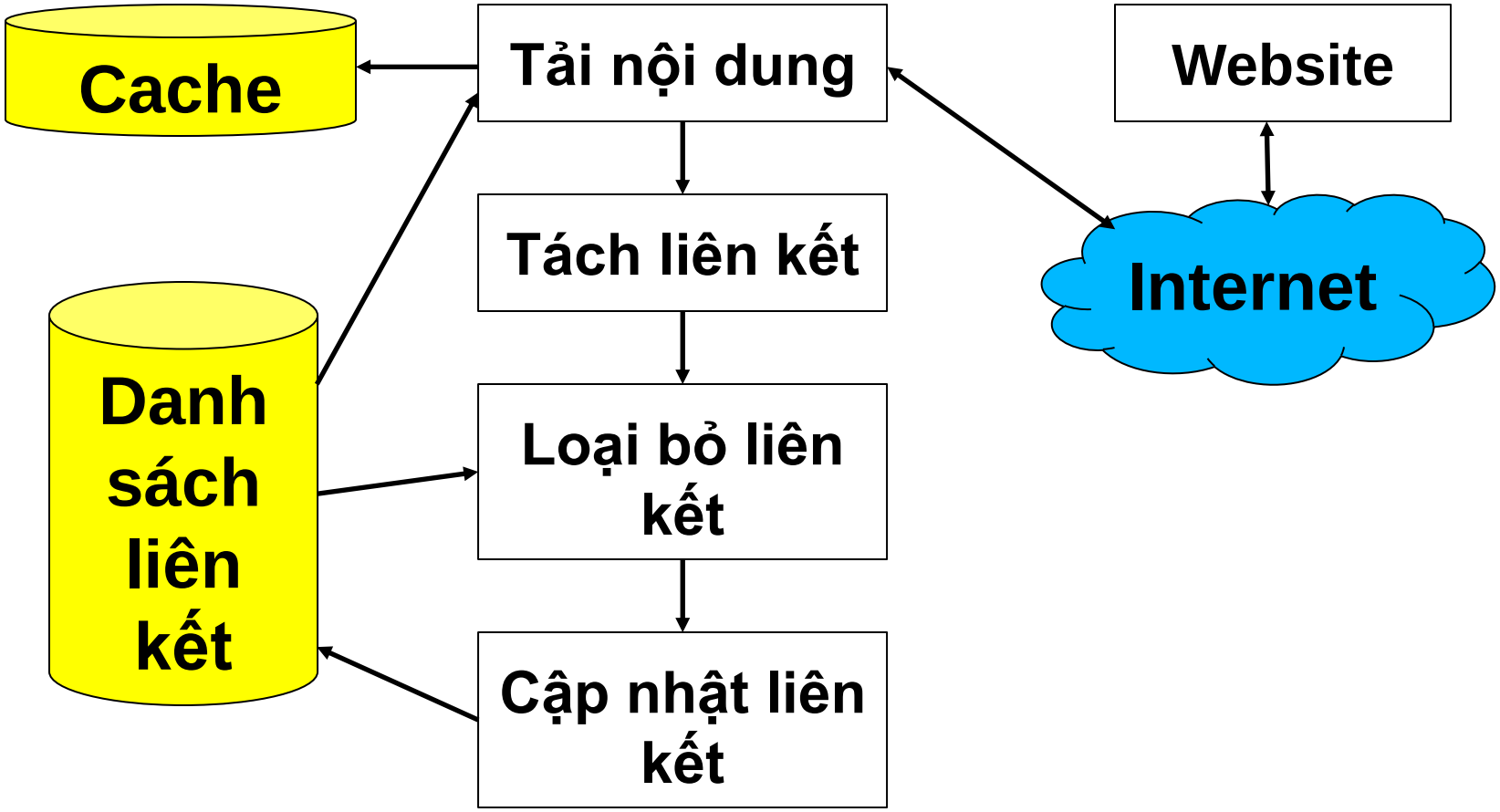
WEB CRAWLER

❖GIỚI THIỆU

- Còn gọi là web spider hoặc bot.
- Duyệt các trang web trên internet để thu thập nội dung.
- Tự động duyệt theo các liên kết (hyper-link)
- Cần có một trang web ban đầu, gọi là seed để duyệt đến những trang khác. Thông thường seed được chọn là sitemap

WEB CRAWLER

❖MÔ HÌNH CỦA CRAWLER



WEB CRAWLER

❖MỘT SỐ VẤN ĐỀ CỦA WEB CRAWLER

- Trùng trang web.
- Bẫy crawler (crawler trap – không thể kết thúc duyệt).
- Có thể bị xác định là DDoS (Distributed Denial of Service) → kiểm tra file robots.txt của website (Robots Exclusion Protocol).
- Website sử dụng công nghệ AJAX.

WEB CRAWLER

❖MỘT SỐ CRAWLER

- Crawler4j.
- Wget.
- Apache Nutch.

TÁCH NỘI DUNG

❖MỤC ĐÍCH

- Lấy nội dung văn bản của các file (doc, xls, pdf, …)
- Lấy nội dung chính của trang web. Loại bỏ:
  - Quảng cáo
  - Web Navigator (hệ thống menu của website)
  - Những liên kết chứa nội dung không liên quan đến nội dung chính của trang web.

Thứ năm, 6/4/2017 | 21:35 GMT+7

Putin tiết lộ thói quen ăn uống

Tổng thống Nga Vladimir Putin tiết lộ với báo giới sở thích và thói quen trong cuộc sống hàng ngày của ông.

- **Putin trích dẫn nhầm lời cựu tổng thống Mỹ**

Tổng thống Nga Vladimir Putin. Ảnh: *Sputnik*.

Trong buổi họp báo tại Diễn đàn Truyền thông ONF ngày 5/4, Tổng thống Nga Vladimir Putin tiết lộ món ăn ưa thích nhất của ông là cháo Nga truyền thống. Ông có thể ăn bất cứ loại cháo nào, trừ ngũ cốc, theo *Sputnik*.

Tư liệu

Xung đột của ông Trump và ông Tập

Toàn cầu hóa, EU, Trung Đông và Triều Tiên là những vấn đề toàn cầu được dự báo tạo ra đối chọi lớn giữa ông Trump và ông Tập trong ...

- **Người dàn xếp cuộc gặp đầu tiên Trump - Tập**

TÁCH NỘI DUNG

Thứ Năm, 06/04/2017 - 09:48

Vợ nhạc sĩ Châu Kỳ bất ngờ lên tiếng về việc "Con đường xưa em đi" bị cấm

Liên quan đến việc 5 ca khúc sáng tác trước năm 1975, trong đó có "Con đường xưa em đi" của nhạc sĩ Châu Kỳ bị Cục NTBD cấm lưu hành vô thời hạn do cho rằng đã sửa lời và có nội dung không đúng với bản gốc, vợ của nhạc sĩ Châu Kỳ đã lần đầu tiên lên tiếng chia sẻ về việc này.

>> Sao không công bố bản gốc "Con đường xưa em đi"?
>> Sửa lời khác bản gốc, "Con đường xưa em đi" bị cấm vĩnh viễn
>> "Con đường xưa em đi" bị tạm dừng lưu hành vì ca từ không đúng?

Theo đó, bà Kha Thị Đàng là vợ của nhạc sĩ Châu Kỳ hiện đang sinh sống cùng gia đình ở TP.HCM. Bà Đằng cho biết, từ khi ca khúc "Con đường xưa em đi" bị cơ quan quản lý văn hóa tạm dừng lưu hành cho đến khi có thông tin bài hát này sẽ bị cấm vĩnh viễn, bà và mọi người trong nhà đều biết nhưng không hiểu rõ lý do của việc này.

TÁCH NỘI DUNG

❖PHƯƠNG PHÁP

- Phân tích cấu trúc tài liệu HTML thành dạng cây. Công cụ phân tích là các HTML Parser, chẳng hạn jSoup.
- Dựa vào quy ước hoặc đặc điểm trình bày trang web để xác định node chứa nội dung chính.
- Lấy phần text của node chứa nội dung chính.

TÁCH NỘI DUNG

❖PHƯƠNG PHÁP

- **Dựa vào quy ước**: Các thẻ <DIV>, <P>, <TABLE>, … có thuộc tính (thường là **id**) đánh dấu nội dung chính.

```
496         </div>
497     </div>
498 <div class="clearfix adm-mainsection">
499     <div class="ads-sponsor type-2 adm-hidden"><div id="admsection2"></div></d
500     <div class="ads-sponsor type-2 adm-hidden"><div id="admsection3"></div></d
501     <div class="clearfix wrapper">
502         <div class="container">
503     <div class="fl wid470 adm-leftsection">
504         <div id="ctl00_IDContent_Tin_Chi_Tiet">
505             <div id="ctl00_IDContent_ctl00_divContent" class="clearfix">
506
507 <div class="box26 clearfix">
508
509 <!--VuLV - edit 11/06-->
510
511         <ol class="breadcrumb clearfix inline" itemscope itemtype="http://sche
512
513         <li class="fl" itemprop="itemListElement" itemscope itemtype="http://s
514             <a class="breadcrumbitem1" itemprop="item" href="/">
```

TÁCH NỘI DUNG

❖PHƯƠNG PHÁP

- **Dựa vào quy ước**:
  - Ưu điểm:
    - ❖ Phương pháp đơn giản.
    - ❖ Dễ thực hiện.
  - Nhược điểm:
    - ❖ Website khác nhau có quy ước khác nhau.
    - ❖ Một website cũng thay đổi theo thời gian.

TÁCH NỘI DUNG

❖PHƯƠNG PHÁP

- Dựa vào hình thức trình bày vùng nội dung chính của trang web:
  - Chiều ngang lớn nhất.
  - Tỉ lệ liên kết so với văn bản thấp.
  - Cấu trúc không lặp với phần khác trong trang web.
  - ….

TÁCH NỘI DUNG

❖PHƯƠNG PHÁP

- Dựa vào hình thức trình bày vùng nội dung chính của trang web:
  - Ưu điểm.
    - ❖ Có thể áp dụng cho nhiều website khác nhau.
    - ❖ Có thể áp dụng khi một website thay đổi không đáng kể.
  - Nhược điểm:
    - ❖ Phức tạp
    - ❖ Không đảm bảo việc áp dụng được cho tất cả Website

LẬP CHỈ MỤC VÀ TÌM KIẾM

❖TIỀN XỬ LÝ

- Chuyển bộ mã.
- Loại bỏ các ký tự không sử dụng.
- Sửa lỗi chính tả.

LẬP CHỈ MỤC VÀ TÌM KIẾM

❖PHÂN TÍCH TERM

- Tách từ.
- Xử lý ngữ nghĩa:
  - Stemming hoặc Lemmatizing (Vd: tiếng Anh).
  - Xác định ranh giới từ (Vd: tiếng Việt, tiếng Hoa)
  - Xác định khái niệm hoặc thực thể.
  - Xác định ý nghĩa của câu (cần nhiều nghiên cứu)
- Kết quả phân tích term ảnh hưởng đến độ phủ của hệ thống.

LẬP CHỈ MỤC VÀ TÌM KIẾM

❖LẬP CHỈ MỤC VÀ TÌM KIẾM

- Chọn mô hình: Boolean, Vector Space, Xác suất,..
- Chọn phương pháp tính trọng sô cho term.
- Chọn công thức tính độ tương đồng giữa truy vấn và tài liệu.

LẬP CHỈ MỤC VÀ TÌM KIẾM

❖ MỘT SỐ THƯ VIỆN HỖ TRỢ LẬP CHỈ MỤC VÀ TÌM KIẾM

- Apache Lucene
- Apache Solr