



IT4853

# Tìm kiếm và trình diễn thông tin

---

Bài 20. Vấn đề tìm kiếm trên Web

IIR.C19. Web search basics

TS. Nguyễn Bá Ngọc, *Bộ môn Hệ thống thông tin,  
Viện CNTT & TT*  
*ngocnb@soict.hust.edu.vn*

Hà Nội, 2016



# Nội dung chính

---

- Dữ liệu Web
- Ước lượng kích thước chỉ mục
- Căn bản tìm kiếm trên Web



# Sao lưu dữ liệu Web

---

- <http://www.archive.org>
  - Được ví như “cỗ máy thời gian” với khả năng hiển thị trang web như trong quá khứ
  - Thu gom bởi Alexa và Compaq
  - Năm 2001 quy mô 4 tỉ trang (40 TB)
  - Năm 2002: 100TB



# Khó khăn đối với tìm kiếm trên Web

---

- Phân tán;
- Thay đổi thường xuyên;
- Rất lớn;
- Phi cấu trúc;
- Nhiều trùng lặp;
- Chất lượng không đồng nhất;
- Đa ngôn ngữ.



# Đặc điểm đồ thị Web

---

- Cõi mỗi trang web (được xác định bởi một url duy nhất) là một đỉnh của đồ thị, các siêu liên kết là các cạnh có hướng của đồ thị.
- Broder et al (2000), WWW9
  - Công trình nghiên cứu tính chất đồ thị web quy mô lớn
- Dữ liệu được thu thập hai lần từ AltaVista
  - Tháng 5 năm 99: 203M trang, 1.5 tỉ liên kết;
  - Tháng 10 năm 99: 271M trang, 2.1 tỉ liên kết.



# Những khái niệm cơ bản của đồ thị

---

- Bậc-vào của một đỉnh là số cạnh đi tới đỉnh đó
- Bậc-ra: số cạnh đi ra từ đỉnh
- Đường kính của đồ thị:
  - Giá trị cực đại của các độ dài ngắn nhất giữa tất cả cặp đỉnh  $(u, v)$ .
- Thành phần liên thông:
  - Thành phần liên thông yếu (WCC – Weakly connected component) là tập đỉnh trong đồ thị vô hướng, trong đó luôn tồn tại đường đi giữa hai nút bất kỳ;
  - Thành phần liên thông mạnh (SCC – Strongly connected component) là thành phần liên thông trong đồ thị có hướng.

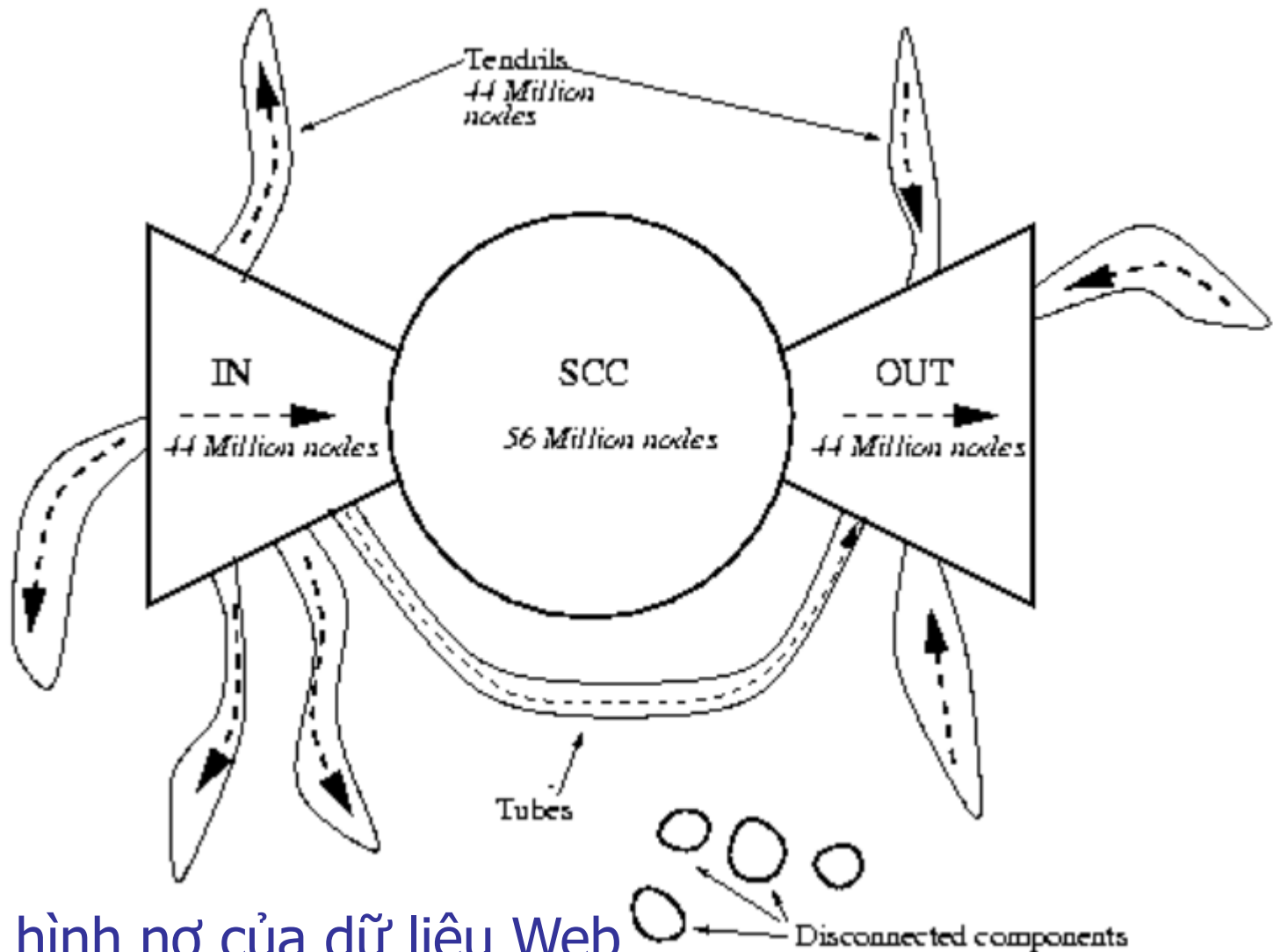


# Kết quả nghiên cứu

---

- Broder et al (2000), WWW9
- Số lượng trang với bậc vào  $i \propto 1/i^{2.1}$ 
  - Thống nhất với những nghiên cứu trên quy mô nhỏ hơn
- Kích thước của thành phần liên kết cũng tuân theo quy luật lũy thừa
  - WCC lớn nhất 91%, SCC lớn nhất 26%

## Kết quả nghiên cứu (2)



Cấu trúc hình nơ của dữ liệu Web





## Kết quả nghiên cứu (3)

---

- Đường kính tối thiểu của SCC là 28
  - Đường kính của toàn bộ Web là trên 500
- Không phải tất cả cặp đỉnh đều liên thông
  - Cho cặp  $(u, v)$  ngẫu nhiên,  $P(\text{path}(u, v)) = 0,24$ 
    - Xác suất tồn tại đường đi từ  $u$  đến  $v$  là 0,24
  - Độ dài trung bình của đường dẫn có hướng là 16
    - Đường dẫn vô hướng là 6
- Tuy nhiên trong trường hợp tổng quát, Web có mức liên thông cao
  - Nếu loại bỏ đỉnh với bậc vào  $> 5$ , trên Web vẫn tồn tại thành phần liên thông yếu  $\sim 59\text{M}$  nút



# Nội dung chính

---

- Dữ liệu Web
- Ước lượng kích thước chỉ mục
- Căn bản tìm kiếm trên Web



# Quy mô dữ liệu web

---

- Dữ liệu web có thể coi là vô hạn
  - Nội dung động
  - Soft 404: [www.yahoo.com/<anything>](http://www.yahoo.com/<anything>)
- Web tĩnh chứa nhiều trùng lặp (~30%)

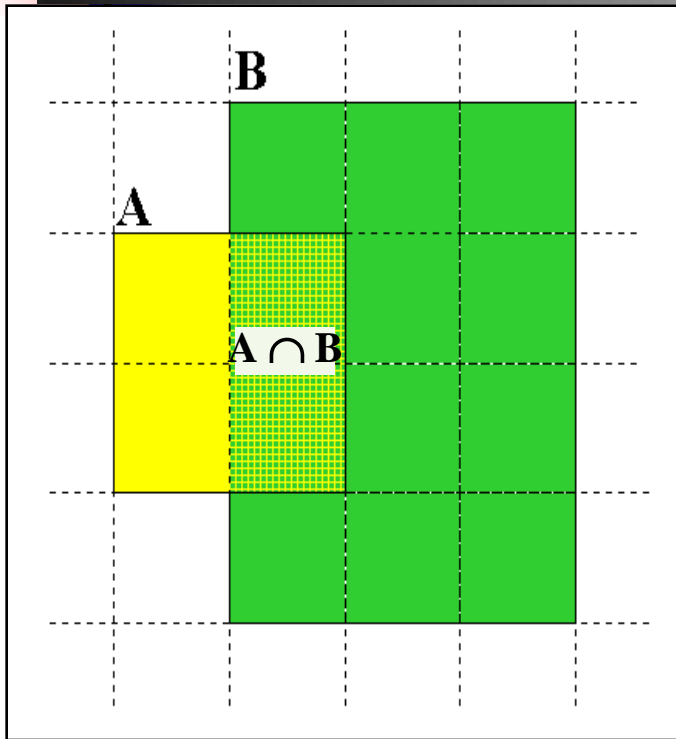


# Kích thước chỉ mục của công cụ tìm kiếm

---

- Công cụ tìm kiếm đánh chỉ mục web tĩnh và web động;
- Các công cụ tìm kiếm khác nhau có dữ liệu chỉ mục khác nhau:
  - Độ sâu url, luật phát hiện spam, độ ưu tiên v.v.
- ... thu thập các nội dung khác nhau từ một URL

# Tỉ lệ chỉ mục



**Lấy mẫu** ngẫu nhiên URLs từ A, **kiểm tra** nếu có trong B;

và ngược lại

$$A \cap B = (1/2) * \text{Size } A$$

$$A \cap B = (1/6) * \text{Size } B$$

$$(1/2) * \text{Size } A = (1/6) * \text{Size } B$$

$$\therefore \text{Size } A / \text{Size } B =$$

$$(1/6) / (1/2) = 1/3$$

**Phép thử:** (i) Lấy mẫu (ii) Kiểm tra



# Lấy mẫu URLs

---

- Mục tiêu: Sinh ngẫu nhiên URL và kiểm tra tồn tại trong chỉ mục.
  - Khó xây dựng giải thuật sinh ngẫu nhiên URL trong toàn bộ Web.
  - Có thể sinh ngẫu nhiên URL có trong chỉ mục của công cụ tìm kiếm.
- Giải pháp 1: Sinh ngẫu nhiên URL trong chỉ mục của công cụ tìm kiếm.
  - Xác định tỉ lệ chỉ mục
- Giải pháp 2: Random walks / địa chỉ IP
  - Trên lý thuyết có thể ước lượng kích thước Web.

# Ước lượng kích thước của Web

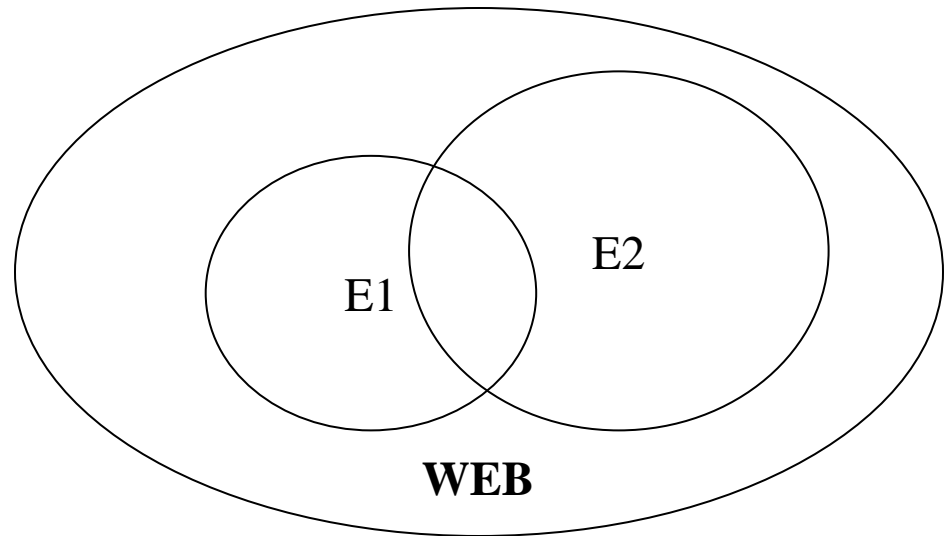
[Lawr98, Bhar98a]

- Giả sử các công cụ tìm kiếm đánh chỉ mục một tập con ngẫu nhiên của Web

**Nếu E2 chứa x% của E1,  
thì E2 cũng chứa x%  
của Web**

**Biết kích thước E2**

**Kích thước Web =  $100 * E2 / x$**



*Bharat & Broder: 200 M (Nov 97), 275 M (Mar 98)*

*Lawrence & Giles: 320 M (Dec 97)*



# Các truy vấn trong nghiên cứu của Lawrence và Giles

---

- *adaptive access control*
- *neighborhood preservation topographic*
- *hamiltonian structures*
- *right linear grammar*
- *pulse width modulation neural*
- *unbalanced prior probabilities*
- *ranked assignment method*
- *internet explorer favourites importing*
- *karvel thornber*
- *zili liu*
- *softmax activation function*
- *bose multidimensional system theory*
- *gamma mlp*
- *dvi2pdf*
- *john oliensis*
- *rieke spikes exploring neural*
- *video watermarking*
- *counterpropagation network*
- *fat shattering dimension*
- *abelson amorphous computing*





## Tỉ lệ đánh chỉ mục Web

---

- Lawrence and Giles (1998) xác định cận dưới đối với Web: 320M trang có thể đánh chỉ mục.
- Công cụ tìm kiếm chỉ phủ một phần nhỏ của Web:
  - HotBot phủ 34%,
  - AltaVista, 28%
  - Northern Light, 20%
  - Excite, 14%
  - Infoseek, 10%
  - Lycos, 3%



# Nội dung chính

---

- Dữ liệu Web
- Ước lượng kích thước chỉ mục
- Căn bản tìm kiếm trên Web

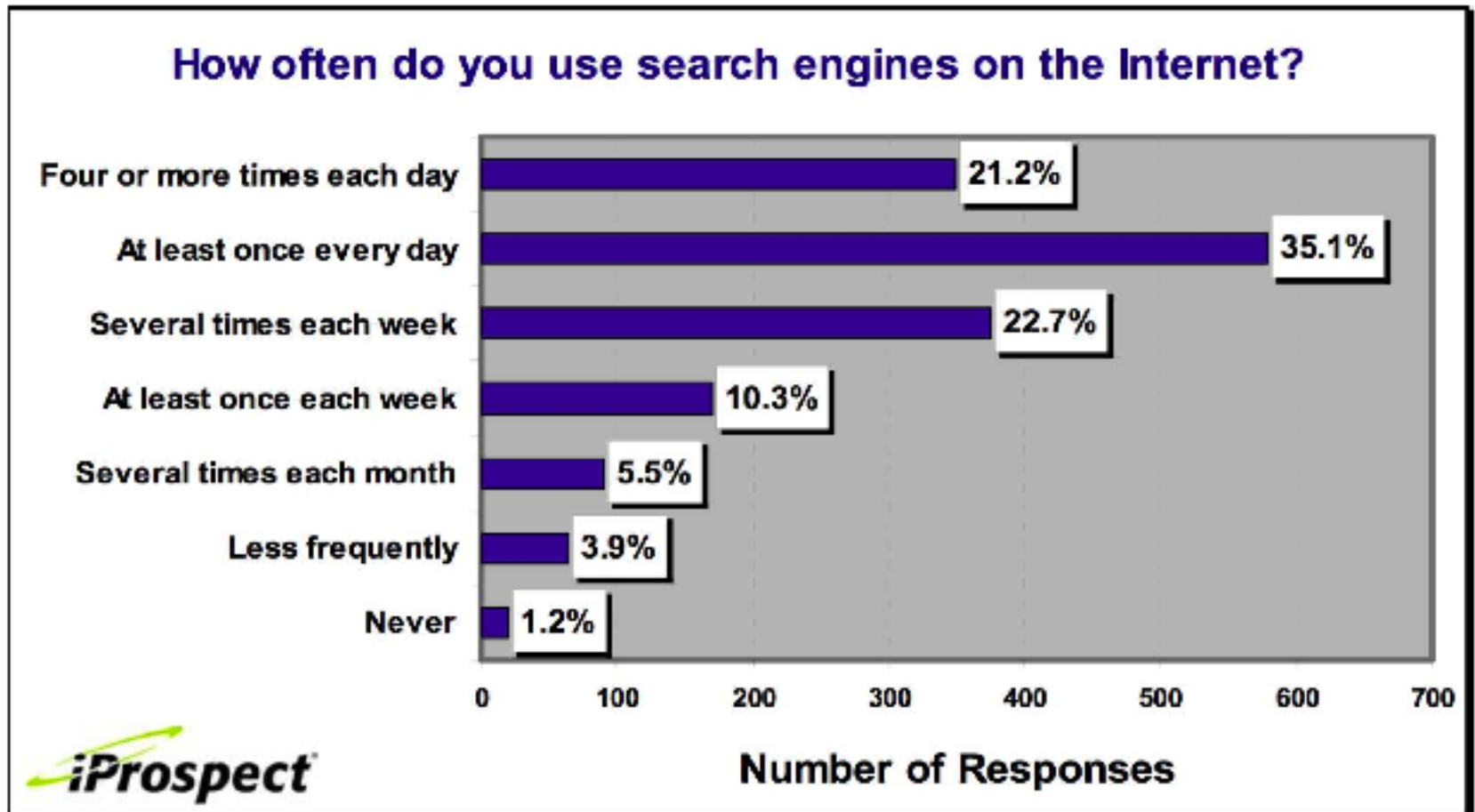


# Vai trò của công cụ tìm kiếm web

---

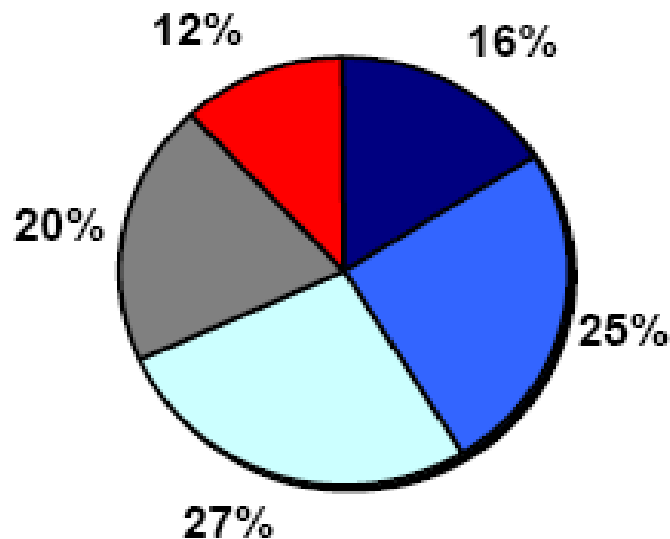
- Là động lực thúc đẩy người dùng công bố nội dung trên web
  - Có nên công bố thông tin nếu không ai đọc nó?
  - Có nên công bố nội dung nếu không thu được lợi nhuận?
- Tìm kiếm giải quyết vấn đề kinh phí vận hành web
  - Máy chủ, thiết bị mạng, việc biên soạn nội dung v.v.
  - Ngày nay phần lớn chi phí được trả nhờ quảng cáo trong tìm kiếm;

# Tìm kiếm là hoạt động thường xuyên nhất trên Web



# Phạm vi tìm kiếm kết quả

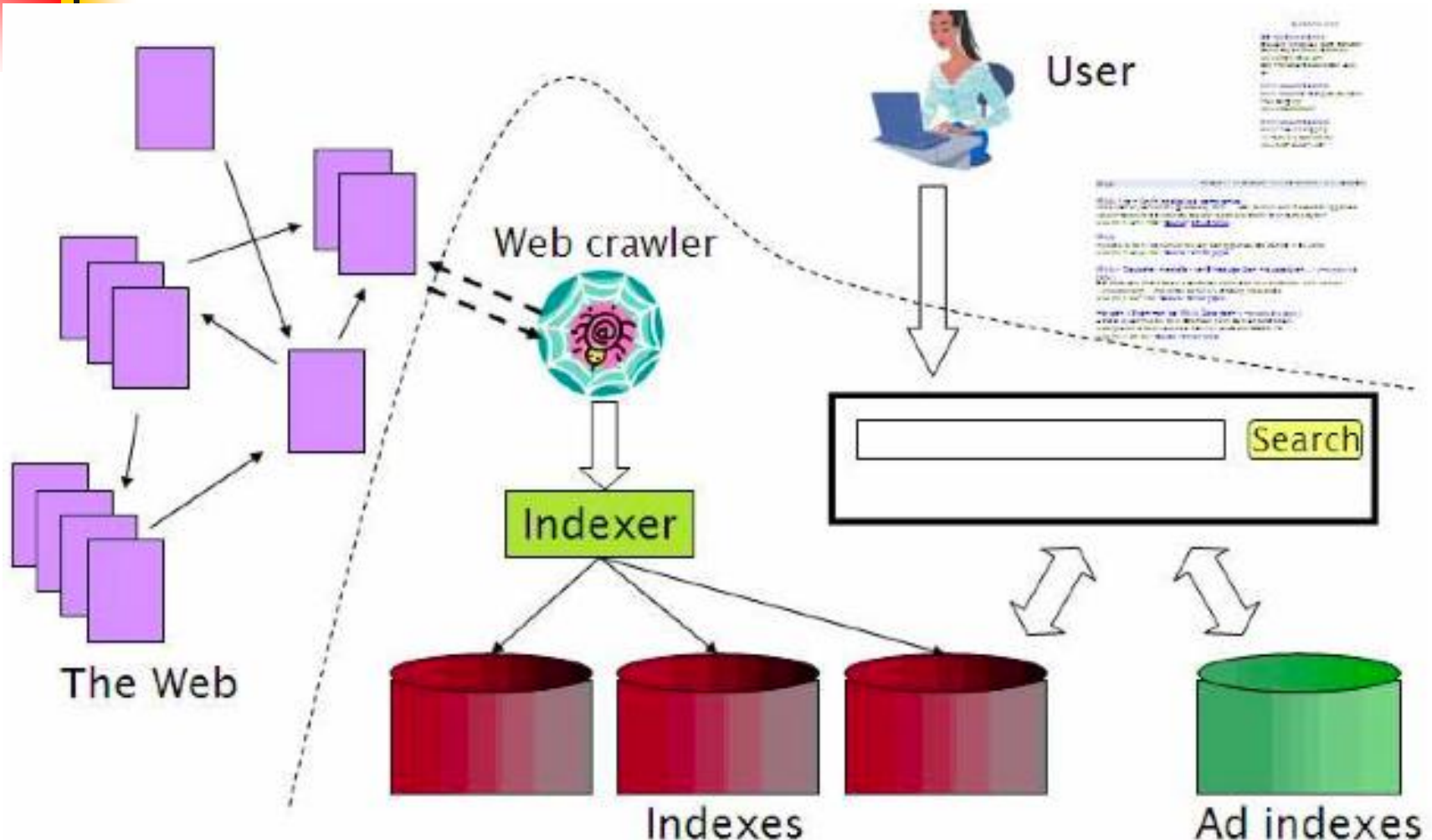
“When you perform a search on a search engine and don't find what you are looking for, at what point do you typically either revise your search, or move on to another search engine? (Select one)”



- After reviewing the first few entries
- After reviewing the first page
- After reviewing the first 2 pages
- After reviewing the first 3 pages
- After reviewing more than 3 pages

(Source: [iprospect.com](http://iprospect.com) WhitePaper\_2006\_SearchEngineUserBehavior.pdf)

# Tổng quan công cụ tìm kiếm trên Web





# Nhu cầu thông tin

---

- Need [Brod02, RL04]

- Thông tin (Informational): Học về một vấn đề nào đó (~40%/65%)
- Định vị (Navigational): Địa chỉ một trang cụ thể (~25%/15%)
- Giao dịch (transactional): Dịch vụ, tải dữ liệu, mua sắm, v.v., (~35%)
- Trung gian (Gray areas)



## Bài tập 20.1

---

Nếu số trang có bậc vào  $i$  tỉ lệ thuận với  $1/i^{2.1}$ , bậc vào trung bình của một trang web bằng bao nhiêu?  
Khi  $i$  tăng đến vô cùng điều gì sẽ xảy ra với số trang có bậc vào  $i$ : tăng, không đổi, tiến đến 0?





## Bài tập 20.2

---

Giá trị trung bình bậc vào của tất cả các nút bằng 9. Giá trị trung bình bậc ra bằng bao nhiêu?



## Bài tập 20.3

---

Hai công cụ tìm kiếm A và B sinh ngẫu nhiên một số lượng lớn trang web từ chỉ mục. 30% trang của A có trong chỉ mục của B, 50% trang của B có trong chỉ mục của A. Hãy tính tỉ lệ chỉ mục  $|A|/|B|$

