

Tìm kiếm thông tin

Chương 9. Căn bản tìm kiếm thông tin trên Web


Soạn bởi: TS. Nguyễn Bá Ngọc

2021

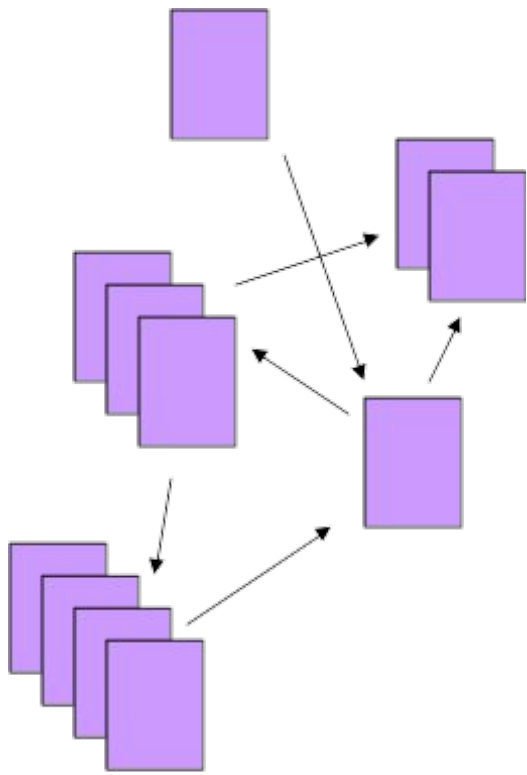
Nội dung

1. Một số đặc điểm cơ bản của Web
2. Biểu diễn đồ thị của Web
3. Gian lận nội dung trong môi trường Web
4. Quảng cáo như mô hình kinh tế
5. Giao diện tìm kiếm
6. Đặc điểm nhu cầu thông tin
7. Ước lượng kích thước chỉ mục
8. Phát hiện nội dung trùng lặp

Nội dung

- 
1. Một số đặc điểm cơ bản của Web
 2. Biểu diễn đồ thị của Web
 3. Gian lận nội dung trong môi trường Web
 4. Quảng cáo như mô hình kinh tế
 5. Giao diện tìm kiếm
 6. Đặc điểm nhu cầu thông tin
 7. Ước lượng kích thước chỉ mục
 8. Phát hiện nội dung trùng lặp

Một số đặc điểm của Web



*Các trang Web và
các liên kết*

- Không có trung tâm điều hành
- Nội dung phân tán và được liên kết
- Các nội dung bao gồm cả thật, giả, mới, cũ, mâu thuẫn, ...
- Thông tin tồn tại ở các định dạng phi cấu trúc, bán cấu trúc, có cấu trúc
- Quy mô vô cùng lớn
- Nội dung mới được công bố liên tục
- Nội dung có thể được *sinh tự động*
- v.v..

Máy tìm kiếm Web thu thập dữ liệu Web về các trung tâm dữ liệu và cung cấp dịch vụ tìm kiếm trên dữ liệu thu thập được

Bản sao lưu của 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
- Quy mô năm 2020: 486 tỉ trang

Có bao nhiêu trang Web?

- Các vấn đề

- Nếu tính đến các nội dung được sinh tự động thì số lượng trang Web là vô hạn, ví dụ:
 - Lịch vạn niên,
 - Trang 404, www.yahoo.com/<bất kỳ từ khóa nào>
 - Hệ thống Web có thể sinh vô hạn nội dung.
- Web tĩnh có nhiều trùng lặp (~30%)
- Một số máy chủ có ít liên kết

- Ảnh hưởng đến

- Giải thuật thu thập dữ liệu
- Giải thuật tìm kiếm

Vai trò của máy tìm kiếm đối với Web

- Là động lực thúc đẩy người dùng đưa nội dung lên Web
 - Có nên đăng nội dung nếu không ai đọc nó?
 - Có nên đăng nội dung nếu nó không đem lại lợi ích?
 - Máy tìm kiếm như "cửa ngõ" đưa người dùng tới các trang Web
- Tìm kiếm giải quyết vấn đề kinh phí vận hành Web
 - Máy chủ, thiết bị mạng, chi phí biên soạn nội dung v.v.
 - Kinh phí có thể được bù từ quảng cáo

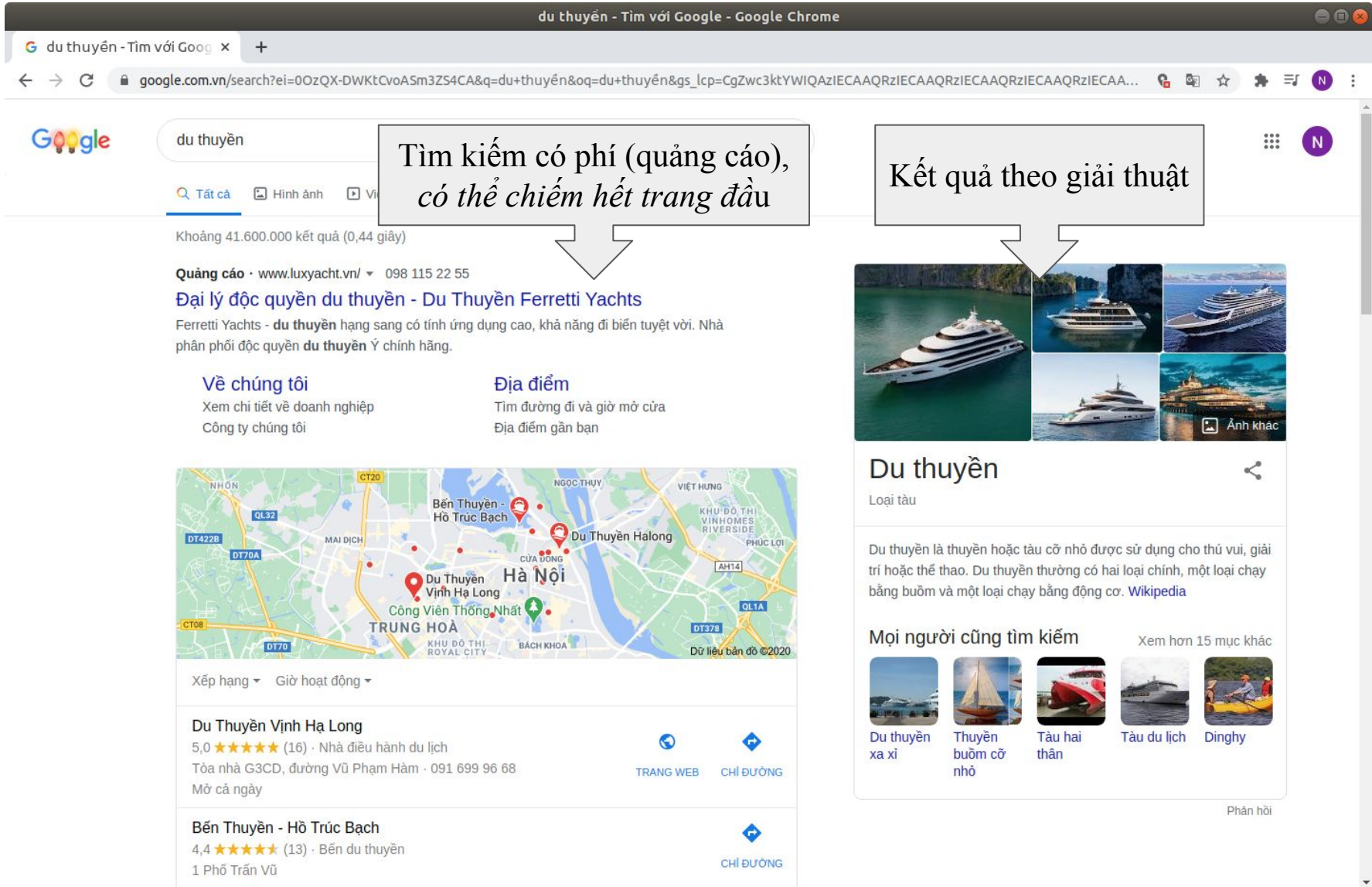
Máy tìm kiếm Web theo dòng thời gian

- Một số máy tìm kiếm đầu tiên dựa trên từ khóa, 1994-1996
 - Infoseek, Lycos, Altavista, Excite, Inktomi
- Tìm kiếm có phí Goto (1997, □ đổi tên thành Overture, 2001 □ bị thôn tóm bởi Yahoo!, 2003)
 - Người quảng cáo tự đưa ra giá trả cho quảng cáo
 - Đấu giá từ khóa: Các nội dung được đặt giá cao nhất được lựa chọn để hiển thị

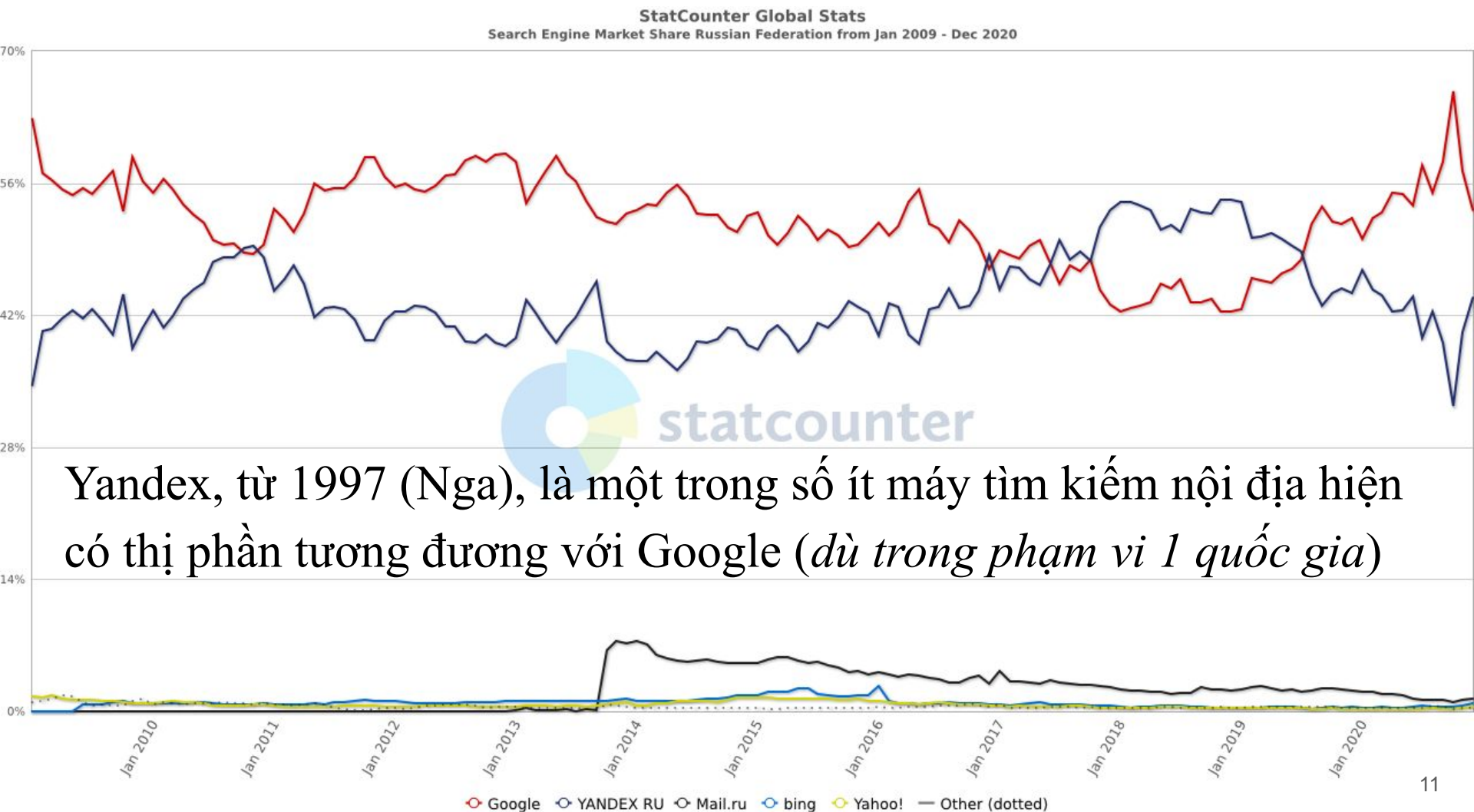
Máy tìm kiếm Web theo dòng thời gian (2)

- 1998+: Google tiên phong áp dụng xếp hạng dựa trên phân tích liên kết
 - Vượt qua những máy tìm kiếm Web đã có trước đó
 - Trải nghiệm người dùng tốt hơn
- 2002: Google bổ xung kết quả tìm kiếm tính phí
 - Sau đó Yahoo thuê tóm Overture (mảng quảng cáo) và Inktomi (mảng tìm kiếm), đệ đơn kiện Google.
- 2005+: Thị phần tìm kiếm của Google liên tục tăng
- 2020: Google chiếm hơn 90% thị trường tìm kiếm trên phạm vi toàn thế giới
 - Trung bình ~6 tỷ yêu cầu tìm kiếm mỗi ngày

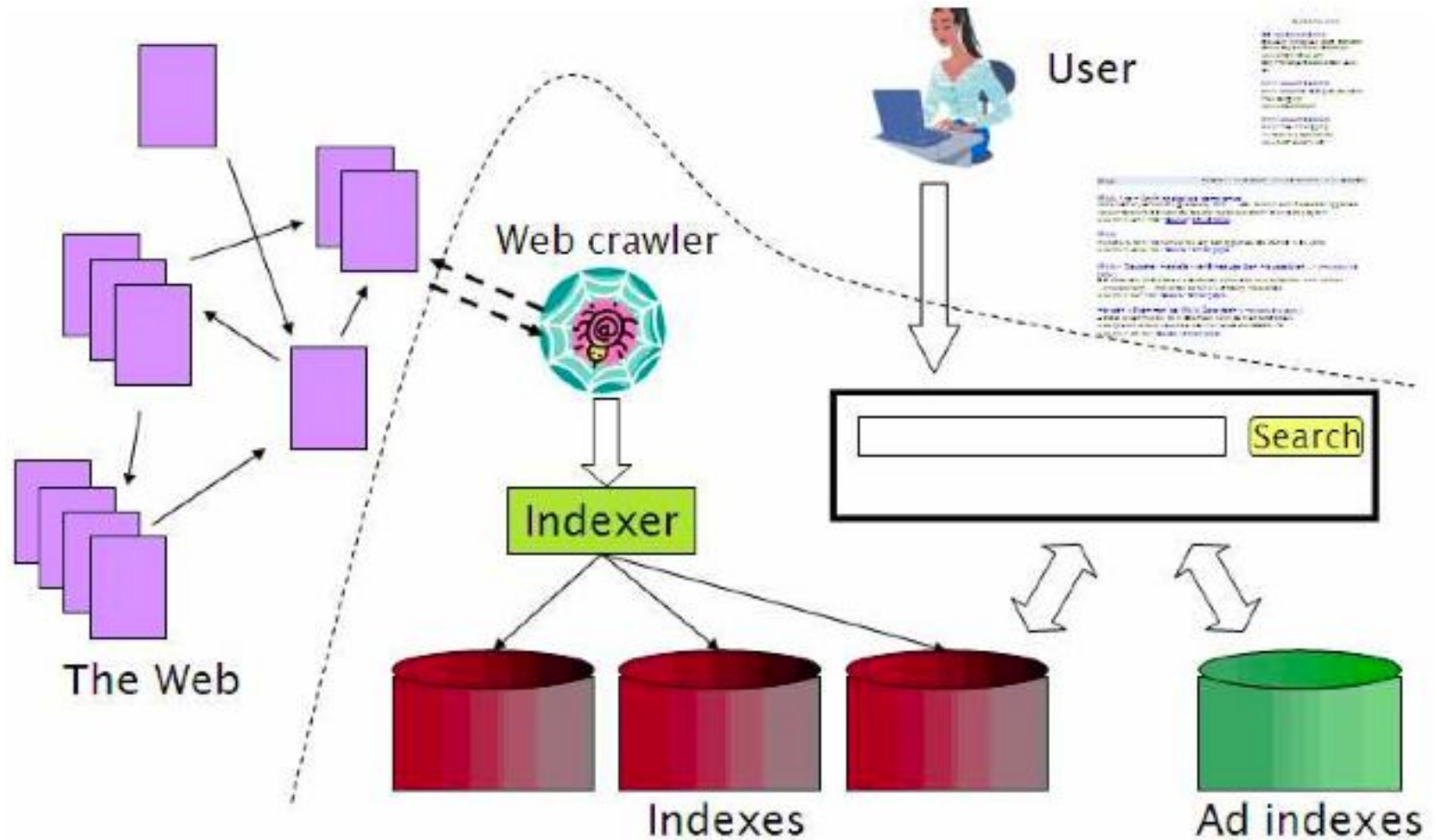
Google tiếng việt



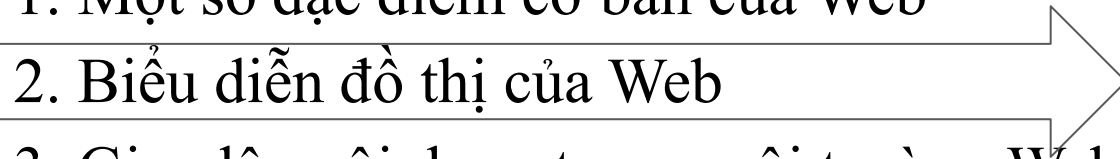
Google vs. máy tìm kiếm nội địa



Bức tranh toàn cảnh



Nội dung

1. Một số đặc điểm cơ bản của Web
 2. Biểu diễn đồ thị của Web
 3. Gian lận nội dung trong môi trường Web
 4. Quảng cáo như mô hình kinh tế
 5. Giao diện tìm kiếm
 6. Đặc điểm nhu cầu thông tin
 7. Ước lượng kích thước chỉ mục
 8. Phát hiện nội dung trùng lặp
- 

Biểu diễn 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, nghiên cứu tính chất đồ thị của Web quy mô lớn
 - Dữ liệu được thu thập hai lần từ AltaVista
 - Tháng 5 năm 1999: 203M trang, 1.5 tỷ liên kết;
 - Tháng 10 năm 1999: 271M trang, 2.1 tỷ liên kết.

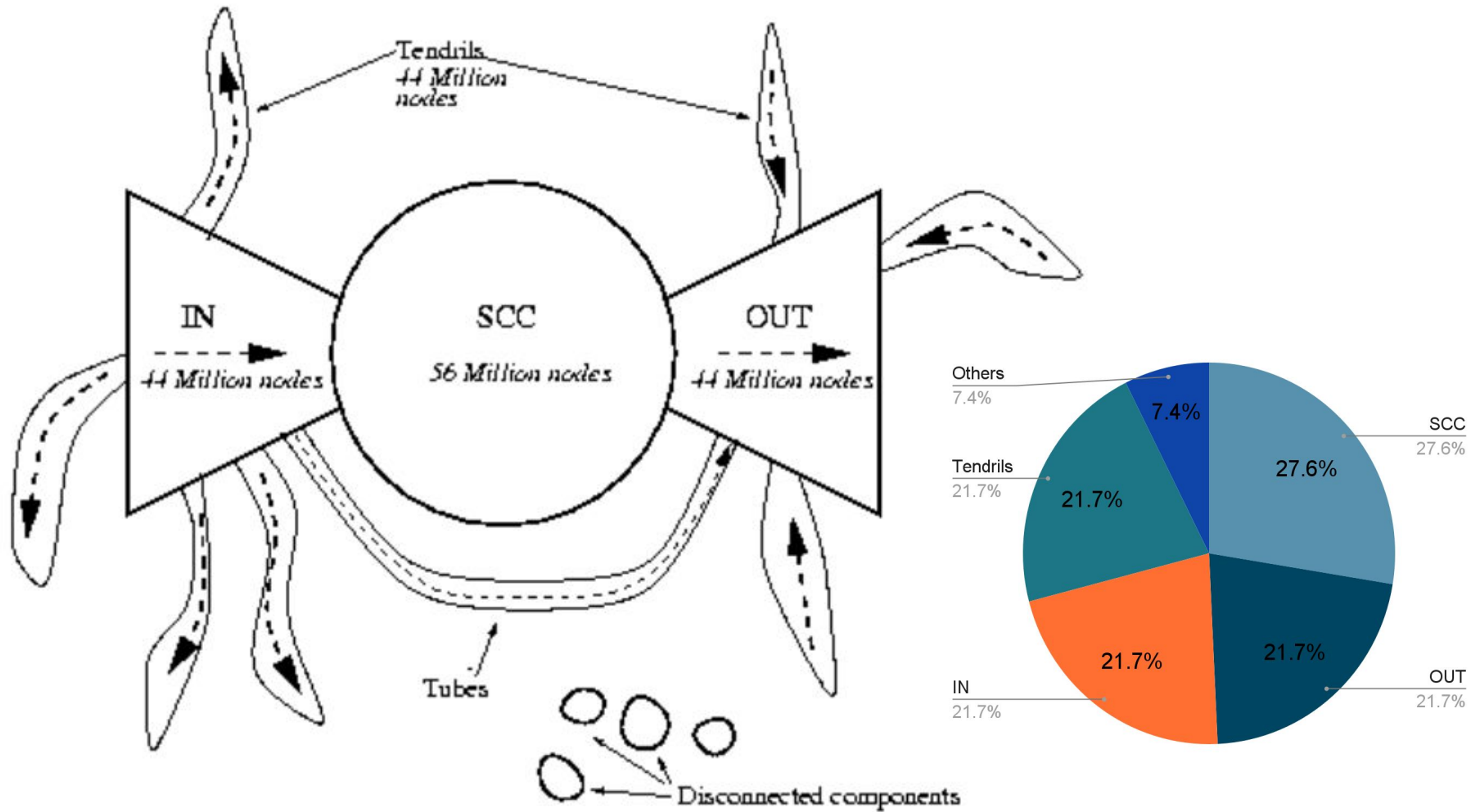
Broder et al (2000), WWW9

- Số lượng trang với bậc vào $i \propto 1/i^{2.1}$
 - Giống với những nghiên cứu trên quy mô nhỏ hơn
- Đồ thị Web không phải là đồ thị liên thông nhưng có tính liên kết cao
 - WCC lớn nhất 91%, SCC lớn nhất ~28%

WCC/Weak Connected Component/Thành phần liên kết yếu - Được xác định dựa trên các cạnh vô hướng;

SCC/Strong Connected Component/Thành phần liên kết mạnh - Được xác định dựa trên các cạnh có hướng.

Broder et al (2000), WWW9₍₂₎



Cấu trúc hình nơ của Web

Broder et al (2000), WWW9₍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 tỉ lệ 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

1. Một số đặc điểm cơ bản của Web
2. Biểu diễn đồ thị của Web
3. Gian lận nội dung trong môi trường Web
4. Quảng cáo như mô hình kinh tế
5. Giao diện tìm kiếm
6. Đặc điểm nhu cầu thông tin
7. Ước lượng kích thước chỉ mục
8. Phát hiện nội dung trùng lặp

Tìm kiếm có trả phí và lựa chọn thay thế

Quảng cáo có thể tiêu tốn ngân quỹ, vậy lựa chọn thay thế là gì?

- *SEO - Search Engine Optimization*
 - Biên soạn nội dung trang Web để được xếp hạng cao trong trang kết quả tìm kiếm cho những từ khóa được lựa chọn (*trong máy tìm kiếm được lựa chọn*).
 - Có thể là một công cụ Marketing
 - Có thể thay thế quảng cáo
- Được thực hiện bởi công ty, cá nhân, cung cấp dịch vụ cho khách hàng
- Có một số hành vi là hợp lệ (SEO), một số khác không hợp lệ (SPAM - gian lận)

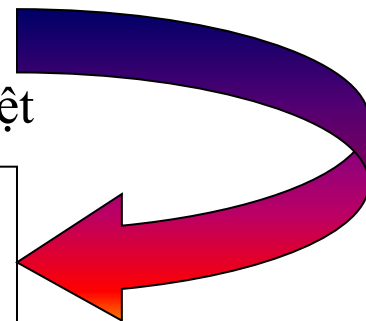
Quảng bá nội dung bằng SEO

- Động cơ
 - Thương mại, chính trị, tôn giáo, vận động, v.v..
 - Nội dung quảng bá có thể được tài trợ bởi quỹ quảng cáo
- Thực hiện
 - Theo hợp đồng cho các công ty, khách hàng
 - Người sở hữu trang Web
 - v.v...
- Diễn đàn
 - Các diễn đàn SEO tiếng việt
 - Các nghiên cứu, giải mã cơ chế xếp hạng 🧐

Hình thức đơn giản nhất

- Máy tìm kiếm thế hệ đầu tiên sử dụng tf/idf như một chỉ số xếp hạng quan trọng
 - Kết quả tìm kiếm được xếp hạng cao nhất cho truy vấn Du thuyền có thể chỉ chứa rất nhiều từ Du thuyền
- SPAM khai thác giải thuật xếp hạng này bằng cách lặp lại các từ khóa được lựa chọn với mật độ dày đặc
 - Ví dụ, Tôm Hùm, Tôm Hùm, Tôm Hùm, Tôm Hùm ...
 - Các từ được lặp lại có thể có màu trùng màu nền:
 - Được đưa vào chỉ mục bởi bộ thu thập
 - Nhưng vô hình với người dùng trên trình duyệt

Tần suất từ không đủ tin cậy để xếp hạng trong Web



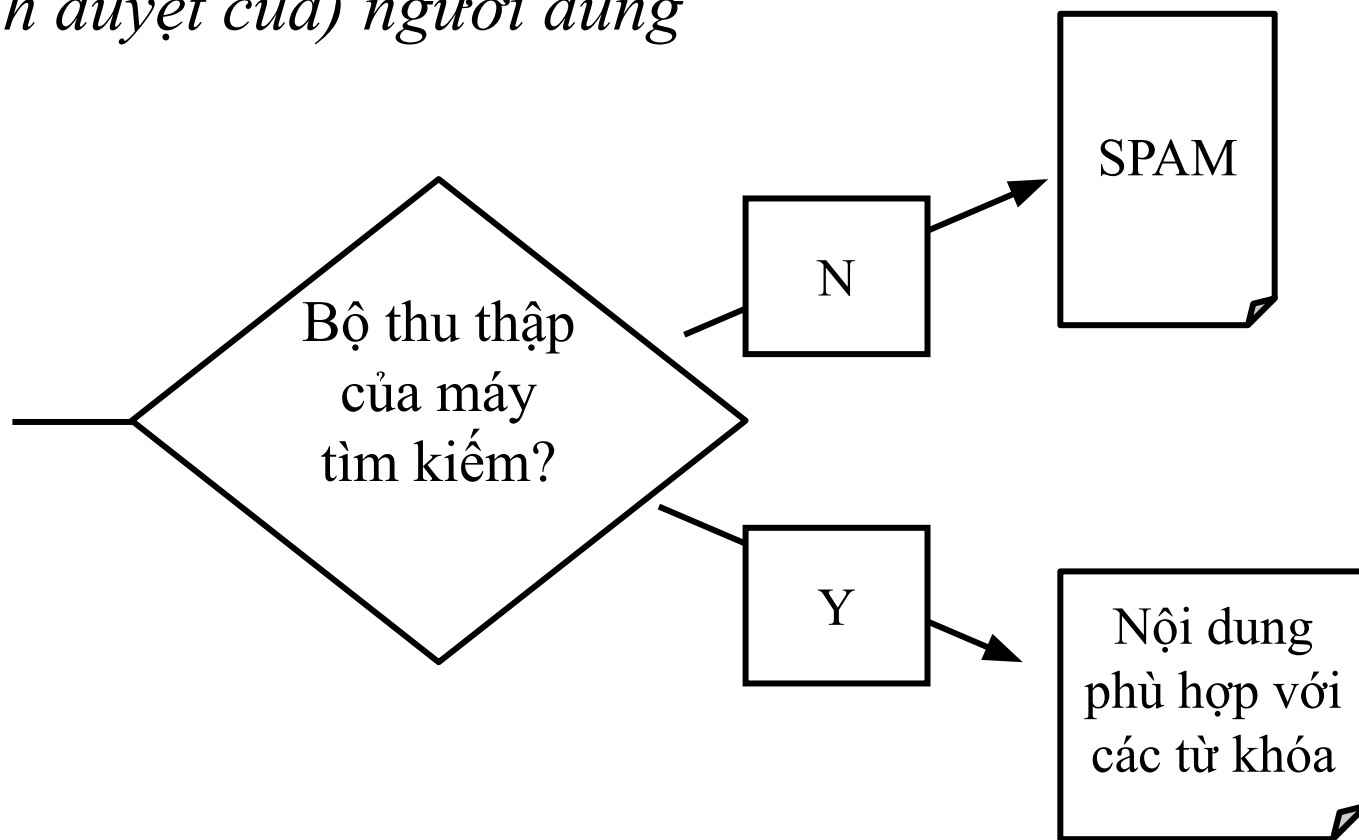
Các tùy chỉnh từ khóa

- Sử dụng thẻ Meta
- Ấn văn bản với cỡ chữ rất nhỏ, màu văn bản trùng với màu nền (thủ thuật CSS), v.v.

```
<meta name="keywords" contents="tôm hùm, tôm hùm, tôm hùm, tôm hùm, du thuyền">
```

Tàng hình

Cung cấp nội dung phù hợp với các từ khóa được lựa chọn cho bộ thu thập, nhưng cung cấp nội dung được quảng bá (SPAM) cho (trình duyệt của) người dùng



Những thủ thuật gian lận khác

- Các trang tổng hợp
 - Các trang được tối ưu hóa cho một từ khóa cụ thể và điều hướng người dùng tới các trang đích
- Gian lận liên kết
 - Cộng đồng liên kết, liên kết ẩn
 - Tham chiếu ảo (Domain flooding): Nhiều miền Web cùng chứa liên kết tới hoặc điều hướng tới trang đích
- Lưu lượng ảo (Robots)
 - Luồng truy vấn ảo
 - Hàng triệu lượt truy cập tự động
- V.V..

Ngăn chặn các hình thức gian lận

- Ưu tiên các tín hiệu chất lượng dựa trên:
 - Bình chọn từ những trang khác (các tín hiệu liên kết)
 - Bình chọn từ người dùng (các tín hiệu sử dụng)
- Giới hạn từ khóa trong các thẻ meta
- Kiểm duyệt URL:
 - Bỏ qua các liên kết có nghi vấn
 - Sử dụng phân tích liên kết để phát hiện nguồn gian lận
- V.V..

Ngăn chặn các hình thức gian lận₍₂₎

- Phát hiện SPAM bằng học máy
 - Huấn luyện bộ phân lớp với những SPAM đã biết
- Lọc nội dung theo giới hạn độ tuổi
 - Phân tích ngôn ngữ, các kỹ thuật phân loại tổng quát, v.v.
 - Phát hiện các nội dung bạo lực v.v.
- Người giám sát nội dung
 - Danh sách đen
 - Kiểm tra những truy vấn thường xuyên nhất
 - Các than phiền từ người dùng
 - Phát hiện các mẫu khả nghi
 - v.v..

Tìm hiểu thêm về SEO

- Máy tìm kiếm Web có các quy định về những hoạt động SEO mà họ ủng hộ hoặc ngăn chặn
 - Webmaster
- Đối kháng trong TKTT: "Cuộc chiến" (kỹ thuật) không hồi kết giữa SEO và máy tìm kiếm
- Tham khảo thêm <http://airWeb.cse.lehigh.edu/> (AIRWeb)

Nội dung

1. Một số đặc điểm cơ bản của Web
2. Biểu diễn đồ thị của Web
3. Gian lận nội dung trong môi trường Web
4. Quảng cáo như mô hình kinh tế
5. Giao diện tìm kiếm
6. Đặc điểm nhu cầu thông tin
7. Ước lượng kích thước chỉ mục
8. Phát hiện nội dung trùng lặp

Mô hình Goto

www.goto.com/d/search?sessionid=1A04214AAA0R50FIEF30PUQ?type=home&tr=1&Keywords=Wilmington+

Wilmington real estate.

Access 75% of all users now!
Premium Listings reach 75% of all
Internet users. [Sign up](#) for Premium
Listings today!

1. [Wilmington Real Estate - Buddy Blake](#)
Wilmington's information and real estate guide. This is your on
anything to do with Wilmington.
[www.buddyblake.com](#) (Cost to advertiser: **10.28**)
2. [Coldwell Banker Sea Coast Realty](#)
Wilmington's number one real estate company.
[www.cbseacoast.com](#) (Cost to advertiser: **10.37**)
3. [Wilmington, NC Real Estate Becky Bullard](#)
Everything you need to know about buying or selling a home c
on my Web site!
[www.iwwc.net](#) (Cost to advertiser: **10.35**)

*xếp hạng theo giá,
không sử dụng tính
phù hợp*

Mô hình Google

Web Images Maps News Shopping Gmail more

Google discount broker Search Advanced Search Preferences

Web Results 1 - 10 of about 807,000 for discount broker [definition]. (0.12 seconds)

Discount Broker Reviews
Information on online discount brokers emphasizing rates, charges, and customer comments and complaints.
[www.broker-reviews.us/](#) - 94k - Cached - Similar pages

Discount Broker Rankings (2008 Broker Survey) at SmartMoney.com
Discount Brokers. Rank/ Brokerage/ Minimum to Open Account, Comments, Standard Commission*, Reduced Commission, Account Fee Per Year (How to Avoid), Avg. ...
[www.smartmoney.com/brokers/index.cfm?story=2004-discount-table](#) - 121k - Cached - Similar pages

Stock Brokers | Discount Brokers | Online Brokers
Most Recommended. Top 5 Brokers headlines. 10. Don't Pay Your Broker for Free Funds May 15 at 3:39 PM. 5. Don't Discount the Discounters Apr 18 at 2:41 PM ...
[www.fool.com/investing/brokers/index.aspx](#) - 44k - Cached - Similar pages

Discount Broker
Discount Broker - Definition of Discount Broker on Investopedia - A stockbroker who carries out buy and sell orders at a reduced commission compared to a ...
[www.investopedia.com/terms/d/discountbroker.asp](#) - 31k - Cached - Similar pages

Discount Brokerage and Online Trading for Smart Stock Market ...
Online stock broker SogoTrade offers the best in discount brokerage investing. Get stock market quotes from this Internet stock trading company.
[www.sgotrade.com/](#) - 39k - Cached - Similar pages

15 questions to ask discount brokers - MSN Money
Jan 11, 2004 ... If you're not big on hand-holding when it comes to investing, a discount broker can be an economical way to go. Just be sure to ask these ...
[moneycentral.msn.com/content/Investing/StartInvesting/P66171.asp](#) - 34k - Cached - Similar pages

Sponsored Links

Rated #1 Online Broker
No Minimums. No Inactivity Fee.
Transfer to Firstrate for Free!
[www.firstrate.com](#)

Discount Broker
Commission free trades for 30 days.
No maintenance fees. Sign up now.
[TDAMERITRADE.com](#)

TradeKing - Online Broker
\$4.95 per Trade, Market or Limit
SmartMoney Top Discount Broker 2001
[www.TradeKing.com](#)

Scottrade Brokerage
\$7 Trades, No Share Limit. In-Depth Research. Start Trading Online Now!
[www.Scottrade.com](#)

Stock trades \$1 - \$3
100 free trades, up to \$100 back for transfer costs, \$500 minimum
[www.sgotrade.com](#)

\$3.95 Online Stock Trades
Market/Limit Orders, No Share Limit and No Inactivity Fees
[www.Marsco.com](#)

INGDIRECT | ShareBuilder
Real-time Quotes, 100% Free

SogoTrade xuất hiện trong kết quả tìm kiếm

SogoTrade trong mục quảng cáo

Công cụ tìm kiếm có ưu tiên nội dung được quảng cáo trong xếp hạng?

Đa phần đều tuyên bố không.

Xếp hạng quảng cáo

- **Đấu giá:**
 - Nhà quảng cáo đặt giá cho từ khóa;
 - Bất kỳ ai cũng có thể đặt giá cho bất kỳ từ khóa nào.
- **Các mô hình tính phí: CPC, CPM.**

Xếp hạng quảng cáo₍₂₎

- Vấn đề: Nếu xếp hạng theo giá như trong Goto, thì truy vấn và quảng cáo có thể không khớp về nội dung;
- Thay đổi trong mô hình của Google: Kết hợp giá và tính phù hợp
 - Tính phù hợp được đánh giá chủ yếu dựa trên tỉ lệ mở xem quảng cáo
 - Các tiêu chí xếp hạng khác: địa điểm, thời gian, chất lượng và tốc độ tải nội dung, v.v..
- Kết quả: Quảng cáo kém phù hợp sẽ khó được xếp hạng cao
 - Người dùng muốn tìm được thông tin hữu ích.
 - Sự hài lòng của người dùng là cực đại => Doanh thu cực đại cho máy tìm kiếm.

Đấu giá với giá thứ hai (Google)

advertiser	bid	CTR	ad rank	rank	paid
A	\$4.00	0.01	0.04	4	(minimum)
B	\$3.00	0.03	0.09	2	\$2.68
C	\$2.00	0.06	0.12	1	\$1.51
D	\$1.00	0.08	0.08	3	\$0.51

Trong đó:

- **bid**: Giá được đưa ra bởi người quảng cáo;
- **CTR**: Tỷ lệ mở quảng cáo: (số lần người xem mở quảng cáo / số lần quảng cáo được hiển thị), **thể hiện tính phù hợp**.
- **ad rank**: Trọng số xếp hạng, $\text{bid} \times \text{CTR}$
- **rank**: Thứ hạng, kết quả xếp hạng
- **paid**: Chi phí thực tế mà người quảng cáo phải chi trả.

Đấu giá với giá thứ hai (Google) (2)

advertiser	bid	CTR	ad rank	rank	paid
A	\$4.00	0.01	0.04	4	(minimum)
B	\$3.00	0.03	0.09	2	\$2.68
C	\$2.00	0.06	0.12	1	\$1.51
D	\$1.00	0.08	0.08	3	\$0.51

Giá thứ hai (chi phí phải chi trả thực tế) bằng **khoản tiền tối thiểu để duy trì vị trí của họ trong đấu giá** cộng thêm 0.01\$.

$$\text{price}_1 \times \text{CTR}_1 = \text{bid}_2 \times \text{CTR}_2 \text{ (để } \text{rank}_1 = \text{rank}_2 \text{)}$$

$$\text{price}_1 = \text{bid}_2 \times \text{CTR}_2 / \text{CTR}_1$$

$$\text{price}_1 = \text{bid}_2 \times \text{CTR}_2 / \text{CTR}_1 = 3.00 \times 0.03 / 0.06 = 1.50; \text{paid}_1 = \$1.51$$

$$\text{price}_2 = \text{bid}_3 \times \text{CTR}_3 / \text{CTR}_2 = 1.00 \times 0.08 / 0.03 = 2.67; \text{paid}_2 = \$2.68$$

$$\text{price}_3 = \text{bid}_4 \times \text{CTR}_4 / \text{CTR}_3 = 4.00 \times 0.01 / 0.08 = 0.50; \text{paid}_3 = \$0.51$$

$$\text{paid} = \text{price} + 0.01$$

Đấu giá theo giá thứ hai của (Google)₍₃₎

- Cộng thêm một xu:
 - Cộng thêm **một xu** từ mỗi quảng cáo sẽ **đem lại** khoản lợi nhuận rất lớn.
- Mô hình quảng cáo có thể là lĩnh vực nghiên cứu quan trọng bậc nhất đối với máy tìm kiếm trên Web
 - Computational advertising

Giá cao nhất được trả cho 1 lần mở = ?

Theo <http://www.cwire.org/highest-paying-search-terms/>

\$69.1	mesothelioma treatment options
\$65.9	personal injury lawyer michigan
\$62.6	student loans consolidation
\$61.4	car accident attorney los angeles
\$59.4	online car insurance quotes
\$59.4	arizona dui lawyer
\$46.4	asbestos cancer
\$40.1	home equity line of credit
\$39.8	life insurance quotes
\$39.2	refinancing
\$38.7	equity line of credit
\$38.0	lasik eye surgery new york city
\$37.0	2nd mortgage
\$35.9	free car insurance quote

Quảng cáo trong tìm kiếm

Cả ba bên đều có lợi?

- **Công cụ tìm kiếm** thu phí mỗi khi người dùng mở quảng cáo.
- **Người dùng** chỉ mở quảng cáo nếu họ thực sự quan tâm.
 - Công cụ tìm kiếm sẽ *phạt* những quảng cáo không phù hợp.
 - ... Ưu tiên hiển thị những nội dung quảng cáo phù hợp.
- **Người quảng cáo** tiếp cận được khách hàng mới

Đánh lừa hệ thống?: Đầu cơ từ khóa

- Mua một từ khóa từ mạng quảng cáo, ví dụ Google AdSense
- Sau đó chuyển hướng lưu lượng tới bên thứ ba
 - Đối tác trả giá cao hơn;
 - Ví dụ, chuyển tới một trang chứa rất nhiều quảng cáo.
- V.V.

Đánh lừa hệ thống? Vi phạm thương hiệu

- Mua từ khóa tìm kiếm là tên thương hiệu của đối thủ cạnh tranh?
 - Mục đích?
- Tham khảo Trademark Complaint
- <https://support.google.com/adspolicy/answer/2562124?hl=en>

Nội dung

1. Một số đặc điểm cơ bản của Web
2. Biểu diễn đồ thị của Web
3. Gian lận nội dung trong môi trường Web
4. Quảng cáo như mô hình kinh tế
5. Giao diện tìm kiếm
6. Đặc điểm nhu cầu thông tin
7. Ước lượng kích thước chỉ mục
8. Phát hiện nội dung trùng lặp

Các đối tượng người dùng

- Trong môi trường Web có nhiều đối tượng người dùng khác nhau
- Người dùng chuyên nghiệp có thể sử dụng các cấu trúc tìm kiếm phức tạp, cung cấp cho máy tìm kiếm nhiều thông tin hơn, và thường tìm thấy kết quả phù hợp nhanh hơn.
- Người dùng phổ thông thường ưa thích giao diện đơn giản, ít sử dụng các cấu trúc truy vấn phức tạp.

Cung cấp nhiều giao diện tìm kiếm?

Các vấn đề giao diện

- Lựa chọn từ khóa tìm kiếm:
 - Khi không nhớ từ khóa, người dùng vẫn có thể chọn lựa nếu thấy.
 - Các công cụ hỗ trợ người dùng viết câu truy vấn,
 - Gợi ý từ khóa,
 - Sửa lỗi cú pháp,
 - v.v..
- Đáp ứng nhu cầu sử dụng:
 - Giao diện đơn giản cho người dùng phổ thông;
 - Giao diện nâng cao với nhiều thông tin và cú pháp phức tạp hơn.

Người dùng tự chuyển đổi giao diện theo nhu cầu

Người dùng đánh giá kết quả tìm kiếm

- Trải nghiệm người dùng được hình thành từ nhiều yếu tố:
 - Riêng độ chính xác và tính phù hợp không đủ phản ánh thực tế;
 - Các yếu tố khác (ngoài tính phù hợp)
 - Nội dung: Đáng tin cậy, phong phú, không trùng lặp, quản lý tốt
 - Hình thức: Hiển thị đúng, ưa nhìn, dễ đọc, nhanh
 - Các nội dung khác: Các pop-ups, nội dung gây phân tán v.v.
- Độ chính xác vs. độ đầy đủ
 - Trong môi trường Web, độ chính xác quan trọng hơn
- Yếu tố quan trọng:
 - Độ chính xác tại 1? Độ chính xác ở top k?
 - Tính toàn diện - Khả năng xử lý nhiều câu hỏi nhập nhằng, đa nghĩa
 - Độ đầy đủ quan trọng khi số lượng kết quả tìm được là rất nhỏ

Nhận định của người dùng có tính chủ quan vì vậy cần được tổng hợp trên quy mô lớn

Người dùng đánh giá kết quả tìm kiếm₍₂₎

- Giao diện người dùng: Đơn giản, ổn định
- Tin tưởng - Các kết quả tìm kiếm mang tính khách quan
 - Máy tìm kiếm có đang thực sự trợ giúp tìm thông tin phù hợp?
- Khả năng bao phủ các chủ đề đối với các truy vấn đa nghĩa
- Các công cụ tiền/hậu xử lý hỗ trợ:
 - Giảm ảnh hưởng do lỗi thao tác của người dùng (tự động kiểm tra lỗi cú pháp, trợ lý tìm kiếm, v.v...)
 - Tiếp tục tìm kiếm: Tìm kiếm trong phạm vi kết quả, tìm kết quả tương tự, định nghĩa lại truy vấn
 - Gợi ý: Các chủ đề liên quan
- Tương thích với những đặc điểm riêng
 - Từ khóa riêng của Web
 - Ảnh hưởng đến tiền xử lý từ khóa, kiểm tra cú pháp, v.v.
 - Các địa chỉ Web được nhập trong hộp tìm kiếm

Các giải pháp tìm kiếm phổ biến

- Tìm kiếm bằng từ khóa (Google)
- Tìm kiếm theo danh mục (Yahoo!)
- Mô phỏng vấn đáp (Ask Jeeves)

Một số điểm khác biệt của Google

- Tập trung vào tính phù hợp
 - Tiết kiệm thời gian của người dùng
- Giao diện đơn giản
 - Dễ sử dụng
 - Ổn định & phản hồi nhanh
 - ☐ tạo trải nghiệm tốt cho người dùng

Nội dung

1. Một số đặc điểm cơ bản của Web
2. Biểu diễn đồ thị của Web
3. Gian lận nội dung trong môi trường Web
4. Quảng cáo như mô hình kinh tế
5. Giao diện tìm kiếm
6. Đặc điểm nhu cầu thông tin
7. Ước lượng kích thước chỉ mục
8. Phát hiện nội dung trùng lặp

Nhu cầu thông tin của người dùng

Các kết quả phân tích lịch sử truy vấn cho thấy

- Người dùng Web thường sử dụng những câu truy vấn ngắn
- Phần lớn chỉ xem một vài trang kết quả đầu tiên
- Hiệu chỉnh lại câu truy vấn nếu chưa được đáp ứng nhu cầu thông tin

Độ dài câu truy vấn

- Các câu truy vấn tìm kiếm trên Web thường ngắn
 - Độ dài trung bình trong khoảng 2-3 từ truy vấn

Table 2.4. Comparative statistics for Excite web queries [208].

Characteristic	1997	1999	2001
Mean terms per query	2.4	2.4	2.6
Terms per query			
1 term	26.3%	29.8%	26.9%
2 term	31.5%	33.8%	30.5%
3+ term	43.1%	36.4%	42.6%
Mean queries per user	2.5	1.9	2.3

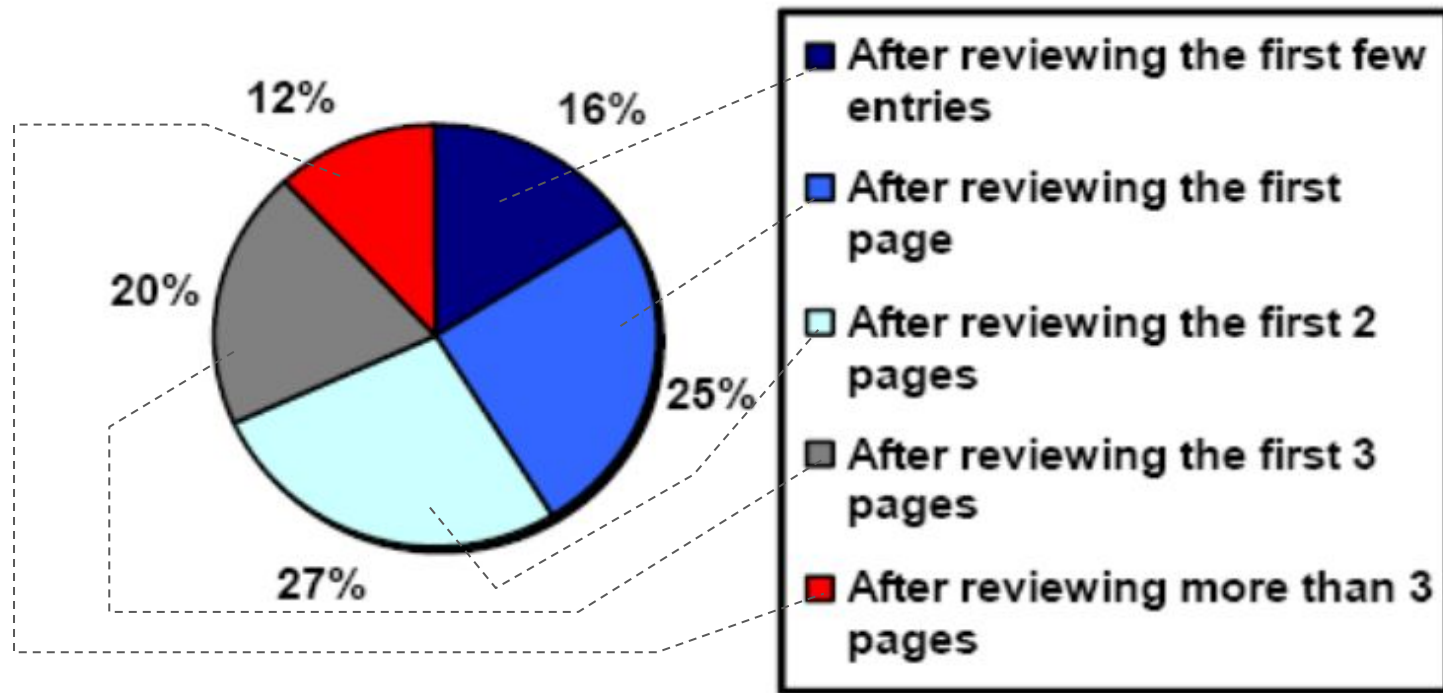
Đặc điểm nhu cầu thông tin người dùng

Need [Brod02, RL04]

- Vấn tin (Informational): Muốn tìm hiểu về một điều gì đó (~40%/65%)
Sinh vật đơn bào
- Định hướng (Navigational): Muốn tìm chỉ một trang Web (~25%/15%)
Website ĐHBKHN
- Giao dịch (Transactional): Muốn làm gì đó qua Web, (~35%/20%)
 - Sử dụng dịch vụ
Tur vấn thiết kế
 - Tải về
Bài giảng TKTT pdf
 - Mua sắm
Máy chủ cấu hình cao
- Còn lại (Gray areas)
 - Trải nghiệm, thử nghiệm các tính năng
Trang thông tin tổng hợp

Người dùng xem những kết quả nào?

“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)”



(Source: iprospect.com WhitePaper_2006_SearchEngineUserBehavior.pdf)

88% người dùng không hài lòng nếu không thấy kết quả phù hợp trong phạm vi 3 trang đầu tiên.

Tỉ lệ phiên tìm kiếm một truy vấn

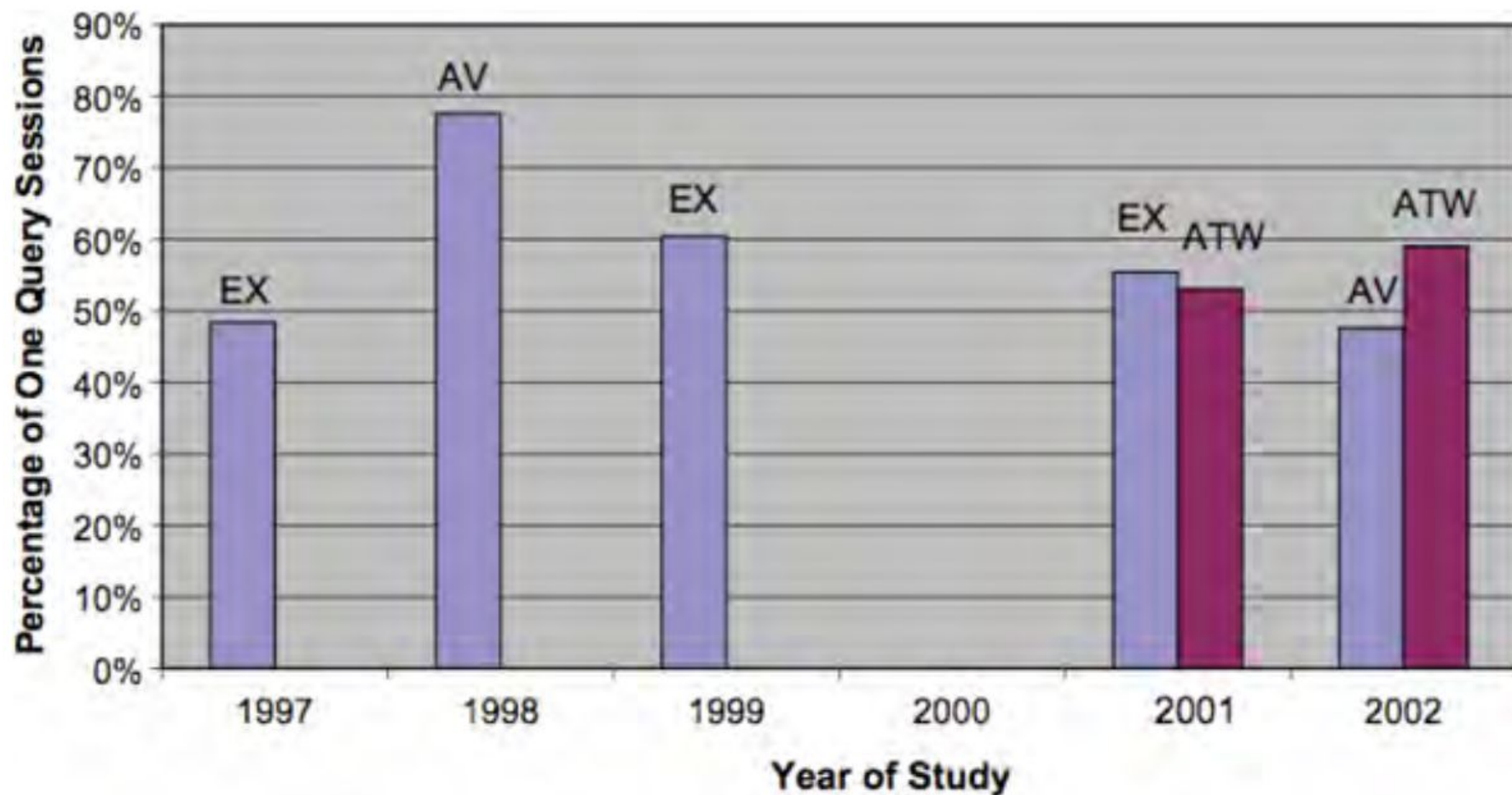


Fig. 3.1 Percentage of single query sessions. From [107].

Hiệu chỉnh trong phiên tìm kiếm

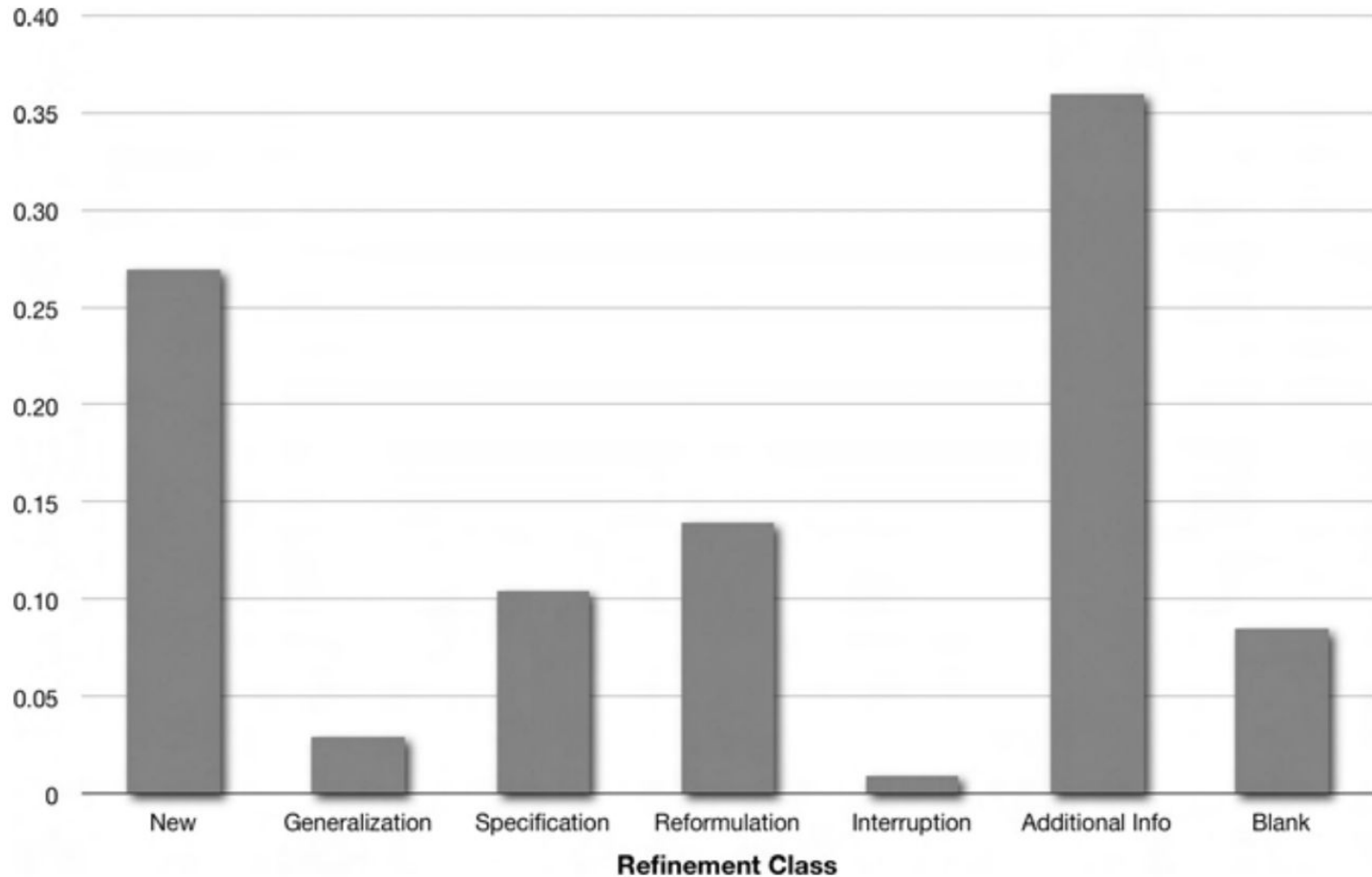
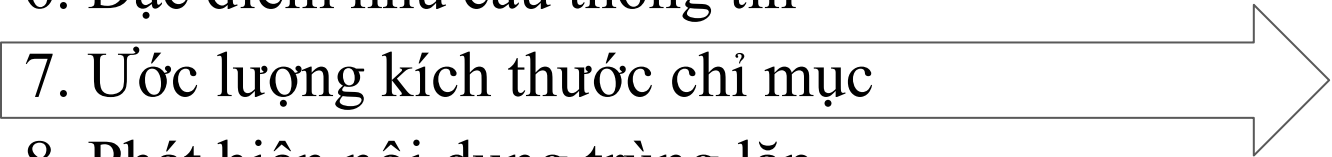


Fig. 3.3 Breakdown of the 4,960 queries analyzed in [127] into the different query modification categories defined.

Nội dung

1. Một số đặc điểm cơ bản của Web
 2. Biểu diễn đồ thị của Web
 3. Gian lận nội dung trong môi trường Web
 4. Quảng cáo như mô hình kinh tế
 5. Giao diện tìm kiếm
 6. Đặc điểm nhu cầu thông tin
 7. Ước lượng kích thước chỉ mục
 8. Phát hiện nội dung trùng lặp
- 

Chỉ mục của các máy tìm kiếm trên Web

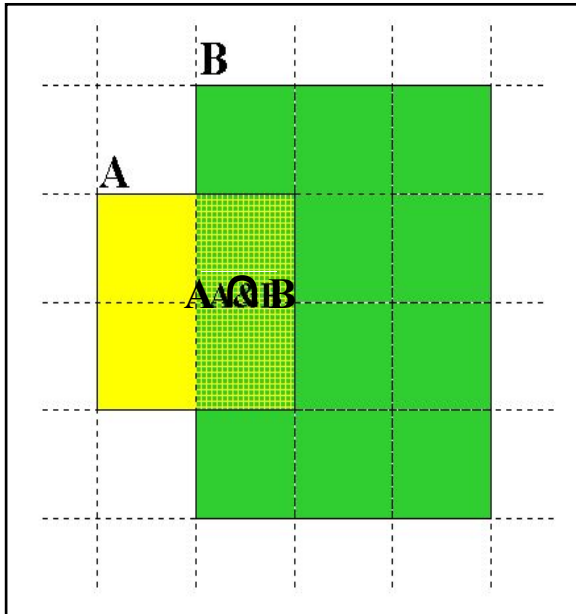
- Các máy tìm kiếm có thể sử dụng các tùy chỉnh khác nhau
 - Độ sâu url, luật phát hiện nội dung gian lận, quy định ưu tiên, v.v.
- Các máy tìm kiếm có thể đánh chỉ mục những nội dung khác nhau từ cùng một URL
 - Các khung, các thẻ meta, chọn lọc nội dung, tài liệu mở rộng v.v..
- Trang Web có thể được đánh chỉ mục 1 phần:
 - Các tài liệu mở rộng: Máy tìm kiếm đánh chỉ mục các trang chưa được thu thập bằng với nội dung được tổng hợp từ văn bản liên kết.
 - Giới hạn kích thước văn bản: Máy tìm kiếm giới hạn nội dung được đánh chỉ mục (n từ đầu tiên, các từ đặc trưng, v.v..)

Có thể ước lượng tỷ lệ chỉ mục của các máy tìm kiếm

Tỷ lệ kích thước dựa trên tỷ lệ chồng lấn

Cho hai máy tìm kiếm A và B

Lấy mẫu ngẫu nhiên URLs từ A, **kiểm tra** xem có trong B hay không và ngược lại



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

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

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

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

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

Mỗi phép thử bao gồm: (i) Lấy mẫu (ii) Kiểm tra

Lấy mẫu URLs

- Trong điều kiện lý tưởng: Lấy ngẫu nhiên URL và kiểm tra sự tồn tại trong chỉ mục.
 - Tuy nhiên: Khó lấy ngẫu nhiên URL vì không có toàn bộ URLs.
- Cách tiếp cận 1: Lấy ngẫu nhiên URL có trong chỉ mục của máy tìm kiếm.
 - Phù hợp để ước lượng kích thước tương đối.
- Cách tiếp cận 2: Lấy ngẫu nhiên URL trong Web
 - *(Có thể ước lượng kích thước Web).*

Các phương pháp thống kê

- Cách tiếp cận truy vấn
 - Truy vấn ngẫu nhiên
 - Phiên tìm kiếm ngẫu nhiên
- Cách tiếp cận duyệt
 - Địa chỉ IP ngẫu nhiên
 - Di chuyển ngẫu nhiên

Các truy vấn ngẫu nhiên

Không phải từ
điển phổ thông

- Sinh một truy vấn ngẫu nhiên: Bằng cách nào?
 - Bộ từ vựng: Tổng hợp từ kết quả thu thập dữ liệu, 400 000+ từ
 - Các truy vấn toàn AND, ví dụ, Lập trình AND Ngôn ngữ AND Công cụ
- Lấy 100 URLs từ các kết quả của máy tìm kiếm A
- Chọn ngẫu nhiên một URLs để kiểm tra sự tồn tại trong máy tìm kiếm B. Bằng cách nào?

Kiểm tra dựa trên truy vấn

- Sử dụng truy vấn để kiểm tra liệu một máy tìm kiếm B có chứa một văn bản d hay không?:
 - Tải d về và tách lấy nội dung
 - Sử dụng 8 từ có tần suất xuất hiện thấp nhất để tạo một truy vấn AND và gửi đến B
 - Kiểm tra xem d có trong tập kết quả hay không
 - *(Một số máy tìm kiếm cho phép tìm kiếm theo URL)*
- Các vấn đề:
 - Gần-trùng lặp
 - Các khung
 - Điều hướng
 - Giới hạn thời gian
 - 8 từ truy vấn liệu có đủ?

Truy vấn ngẫu nhiên: Các yếu tố sai lệch

- Sai lệch truy vấn: Các trang chứa nhiều từ trong bộ từ vựng được lựa chọn thường xuyên
- Sai lệch xếp hạng: Các kết quả được xếp hạng cao có khả năng được lựa chọn nhiều hơn
- Sai lệch kiểm tra: Trùng lặp, bỏ qua kết quả ở cuối danh sách
- Sai lệch giới hạn: Máy tìm kiếm có thể không xử lý chính xác truy vấn toàn AND với 8 từ truy vấn
- Sai lệch nội dung: Các gian lận nội dung
- Các vấn đề vận hành: Giới hạn thời gian truy cập, lỗi dịch vụ, điều chỉnh chỉ mục

Phiên tìm kiếm ngẫu nhiên

- Lấy các truy vấn (500-1000) từ dữ liệu lịch sử trong mạng nội bộ
- Thực thi:
 - Giới hạn chỉ sử dụng các truy vấn có < 600 kết quả
 - Đếm các URLs từ mỗi máy tìm kiếm
 - Tính tỉ lệ kích thước & chồng lấn cho từng truy vấn
 - Ước lượng tỉ lệ kích thước chỉ mục & chồng lấn bằng cách lấy trung bình trên toàn bộ truy vấn.

Phiên tìm kiếm ngẫu nhiên₍₂₎

- Ưu điểm

- Có thể mô phỏng tốt sự tiếp nhận của con người về tính bao phủ

- Nhược điểm

- Các mẫu mô phỏng theo nguồn của dữ liệu lịch sử có thể khác với dữ liệu ở thời điểm tiến hành thực nghiệm
- Các vấn đề thống kê: Không có cơ sở lý thuyết thống kê cho đại lượng trung bình của các tỷ lệ

Các truy vấn trong Lawrence và Giles [97]

- *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*

Địa chỉ IP ngẫu nhiên

- Sinh ngẫu nhiên một địa chỉ IP
- Tìm một máy chủ Web ở địa chỉ IP đó
- Thu thập các trang Web từ máy chủ (nếu có)
 - Chọn một trang ngẫu nhiên trong phạm vi thu thập được

Địa chỉ IP ngẫu nhiên₍₂₎

- Các yêu cầu HTTP gửi đến địa chỉ IP ngẫu nhiên
 - Bỏ qua: Phản hồi rỗng hoặc yêu cầu đăng nhập hoặc bị từ chối
- [Lawr99] Ước lượng 2.8 triệu địa chỉ IP đang có các máy chủ Web
 - Thu thập toàn bộ 2500 máy chủ trong số những máy chủ được phát hiện và có thể thu thập được
 - Ước lượng kích thước Web là 800 triệu trang
 - Tỷ lệ sử dụng các thẻ mô tả:
 - Thẻ meta (từ khóa, mô tả) trong 34% các trang chủ, siêu dữ liệu Dublin core là 0.3%.
- OCLC tìm thấy 8.7 triệu máy chủ năm 2001
 - Netcraft[Netc02] truy cập 37.2 triệu máy chủ năm 2002.

Địa chỉ IP ngẫu nhiên₍₃₎

- Ưu điểm
 - Không yêu cầu tập URLs mầm để khởi tạo
- Nhược điểm
 - Nhiều máy chủ có thể chia sẻ cùng một địa chỉ IP, hoặc không tiếp nhận các yêu cầu theo IP
 - Không có sự đảm bảo tất cả các trang đều liên kết với trang chủ
 - Có thể chịu ảnh hưởng từ sự gian lận nội dung (nhiều IP cho cùng một máy chủ để tránh bị chặn IP).

Di chuyển ngẫu nhiên

- Cõi Web như một đồ thị có hướng
- Xây dựng giải thuật di chuyển ngẫu nhiên trên đồ thị Web
 - Bao gồm cả luật nhảy ngẫu nhiên tới các trang đã biết
 - Để tránh bị mắc kẹt trong các bẫy thu thập
 - Có thể đi theo tất cả các liên kết
 - Hội tụ về một phân bố ổn định
 - Không phụ thuộc vào các khởi tạo ban đầu,
 - ... tuy nhiên không biết chắc chắn khi nào thì hội tụ
 - Lấy mẫu từ phân bố ở trạng thái ổn định

Di chuyển ngẫu nhiên₍₂₎

- Ưu điểm
 - Không giới hạn quy mô Web
- Nhược điểm
 - Vấn đề lựa chọn tập mẫu
 - Các khó khăn liên quan đến mô phỏng thực tế
 - Có thể bị ảnh hưởng bởi gian lận liên kết
 - Không rõ thời gian hội tụ

Tổng hợp về vấn đề lấy mẫu URL

- Mỗi giải pháp đều có những nhược điểm
- Có nhiều ý tưởng mới
- tuy nhiên các vấn đề ngày càng khó hơn

Các nghiên cứu định lượng là những chủ đề nghiên cứu thú vị

Nội dung

1. Một số đặc điểm cơ bản của Web
2. Biểu diễn đồ thị của Web
3. Gian lận nội dung trong môi trường Web
4. Quảng cáo như mô hình kinh tế
5. Giao diện tìm kiếm
6. Đặc điểm nhu cầu thông tin
7. Ước lượng kích thước chỉ mục
8. Phát hiện nội dung trùng lặp

Vấn đề trùng lặp nội dung

Trong môi trường Web có rất nhiều nội dung trùng lặp:

- Một nội dung Web có thể được đăng lại y nguyên ở một địa chỉ khác: Trùng lặp tuyệt đối
- Một trang Web được hiệu chỉnh lại một chút rồi được đăng lại ở một địa chỉ khác: Gần-trùng lặp
- Các trang có nội dung giống nhau về nghĩa nhưng được biên soạn độc lập, có cách trình bày khác nhau thì không phải trùng lặp
 - Ví dụ, các tin bài về cùng một sự kiện được soạn bởi các báo khác nhau

Vấn đề trùng lặp nội dung₍₂₎

- Người dùng không muốn nhận những kết quả trùng lặp
 - Một kết quả dù phù hợp nhưng có thể bị coi như không phù hợp nếu trùng lặp với kết quả đã được trả về trước.
- Lưu nội dung trùng lặp làm lãng phí tài nguyên hệ thống

Cần loại bỏ những tài liệu trùng lặp, chỉ cần giữ một tài liệu cho một nhóm tài liệu trùng lặp!

Phát hiện trùng lặp

- Trùng lặp tuyệt đối: Có thể được phát hiện bằng các tổng đại diện
 - Tổng đại diện bằng tổng mã ký tự của các ký tự trong văn bản:
Một hàm băm đơn giản
- Gần-trùng lặp:
 - Khó phát hiện
 - Tính độ tương đồng dựa trên từ vựng
 - Sử dụng ngưỡng tương đồng để kết luận
 - Ví dụ, độ tương đồng $> 80\%$ \Rightarrow Các văn bản là gần-trùng lặp

Ví dụ 9.1. Trùng lặp gần

Apple M1: Dòng chip khiến Mac trở nên tuyệt vời hơn - Google Chrome

thegioididong.com/tin-tuc/apple-m1-la-gi-1305904

Những điều cần biết về Apple M1: Dòng chip mới từ nhà Táo sẽ giúp dòng máy tính Mac trở nên tuyệt vời hơn

Nguyễn Anh Tuấn • 22/11

Trong sự kiện ngày 10/11, Apple đã chính thức giới thiệu dòng chip Apple Silicon dựa trên cấu trúc ARM với tên gọi Apple M1. Đây chính là dấu mốc đánh dấu cho sự khởi đầu một kỷ nguyên mới của nhà Táo khi làm chủ hoàn toàn cấu trúc phần cứng của dòng Macbook như những gì họ đã làm với iPhone, iPad, Apple Watch và các sản phẩm khác của mình.

Tóm tắt nhanh về Apple M1:

- M1 là bộ vi xử lý 8 lõi, kiến trúc 5 nm.
- M1 đem đến hiệu suất nhanh hơn trên máy tính Macbook.
- M1 cho phép chạy các ứng dụng iPhone và iPad trên Macbook.

M1 là con chip mạnh nhất mà Apple từng tạo ra cho Macbook

Chúng ta có thể nói đùa rằng vì M1 là con chip đầu tiên mà Apple tạo ra cho Macbook nên vì thế con chip này cũng là mạnh nhất mà Apple thiết kế cho Macbook. Con chip này được chế tạo trên tiến trình 5 nm mới nhất, được trang bị tới 16 tỉ bóng bán dẫn. Do đó, M1 mang đến hiệu suất CPU nhanh hơn 3.5 lần, hiệu suất GPU nhanh hơn 6 lần và máy học nhanh hơn tới 15 lần. Đồng thời cho phép thời lượng pin nhiều hơn tới 2 lần so với máy Macbook thế hệ trước.

Những điều Cần Biết Về Apple M1: Dòng Chip Mới Từ Nhà Táo Sẽ Giúp Dòng Máy Tính Mac Trở Nên Tuyệt Vời Hơn - Google Chrome

itsystems.vn/tin-tuc-it/nhung-dieu-can-biet-ve-apple-m1/

Những điều cần biết về Apple M1: Dòng chip mới từ nhà Táo sẽ giúp dòng máy tính Mac trở nên tuyệt vời hơn

IT SYSTEMS

Home Dịch Vụ Data Center Thiết Bị IT Khuyến Mãi Hướng Nghiệp IT Kiến Thức IT Liên Hệ

Những điều cần biết về Apple M1: Dòng chip mới từ nhà Táo sẽ giúp dòng máy tính Mac trở nên tuyệt vời hơn

Trong sự kiện ngày 10/11, Apple đã chính thức giới thiệu dòng chip Apple Silicon dựa trên cấu trúc ARM với tên gọi Apple M1. Đây chính là dấu mốc đánh dấu cho sự khởi đầu một kỷ nguyên mới của nhà Táo khi làm chủ hoàn toàn cấu trúc phần cứng của dòng Macbook như những gì họ đã làm với iPhone, iPad, Apple Watch và các sản phẩm khác của mình.

Tóm tắt nhanh về Apple M1:

- M1 là bộ vi xử lý 8 lõi, kiến trúc 5 nm.
- M1 đem đến hiệu suất nhanh hơn trên máy tính Macbook.
- M1 cho phép chạy các ứng dụng iPhone và iPad trên Macbook.

M1 là con chip mạnh nhất mà Apple từng tạo ra cho Macbook

Chúng ta có thể nói đùa rằng vì M1 là con chip đầu tiên mà Apple tạo ra cho Macbook nên vì thế con chip này cũng là mạnh nhất mà Apple thiết kế cho Macbook. Con chip này được chế tạo trên tiến trình 5 nm mới nhất, được trang bị tới 16 tỉ bóng bán dẫn. Do đó, M1 mang đến hiệu suất CPU nhanh hơn 3.5 lần, hiệu suất GPU nhanh hơn 6 lần và máy học nhanh hơn tới 15 lần. Đồng thời cho phép thời lượng pin nhiều hơn tới 2 lần so với máy Macbook thế hệ trước.

0909.10.45.79

Biểu diễn văn bản: Mô hình tập chuỗi n-từ

- Chuỗi n-từ là một **n-gram trên từ** (n từ viết liền thành một chuỗi, shingles).
- Ví dụ, với $n = 3$, "Everything is itself and at the same time not itself" có mô hình tập chuỗi n-từ như sau:
 - { Everything-is-itself, is-itself-and, itself-and-at, and-at-the, at-the-same, the-same-time, same-time-not, time-not-itself }

Độ tương đồng của hai tài liệu được đánh giá theo hệ số Jaccard của hai tập chuỗi n-từ.

Hệ số Jaccard

- Cho hai tập A và B, trong đó $A \neq \emptyset$ hoặc $B \neq \emptyset$

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

- $Jaccard(A, A) = 1$
- $Jaccard(A, \bar{A}) = 0$
- $Jaccard(A, B)$ có miền giá trị là $[0, 1]$

Ví dụ 9.2. Tính độ tương đồng

- Cho ba tài liệu:

d_1 : "Jack London traveled to Oakland"

d_2 : "Jack London traveled to the city of Oakland"

d_3 : "Jack traveled from Oakland to London"

- Hãy tính hệ số Jaccard của tập chuỗi 2-từ?
 - $J(d_1, d_2)$ và $J(d_1, d_3)$

Ví dụ 9.2. Tính độ tương đồng₍₂₎

$d_1 = \{\text{Jack-London, London-traveled, traveled-to, to-Oakland}\}$

$d_2 = \{\text{Jack-London, London-traveled, traveled-to, to-the, the-city, city-of, of-Oakland}\}$

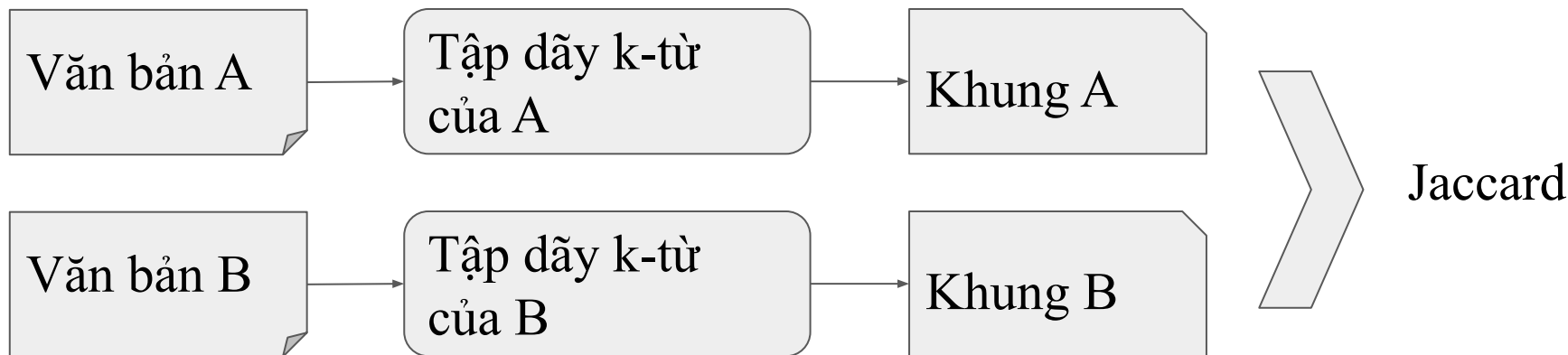
$d_3 = \{\text{Jack-traveled, traveled-from, from-Oakland, Oakland-to, to-London}\}$

$$J(d_1, d_2) = 3/8 = 0.375; J(d_1, d_3) = 0$$

Hệ số Jaccard trên tập chuỗi n-từ rất nhạy với trật tự từ

Chuỗi n-từ và các phép toán trên tập hợp

- Lấy giao của các tập chuỗi n-từ của tất cả các cặp văn bản đòi hỏi khối lượng tính toán rất lớn. Vì vậy cần:
 - Ước lượng sử dụng các tập con với các phần tử được lựa chọn khéo léo (biểu diễn khung, sketch)
 - Thực hiện ước lượng sử dụng các biểu diễn khung ngắn gọn

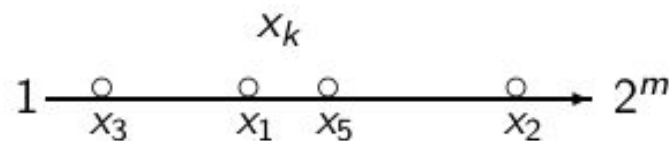
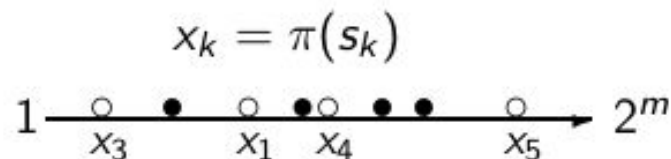
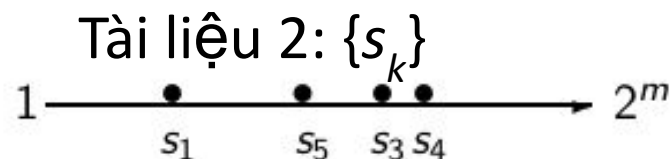
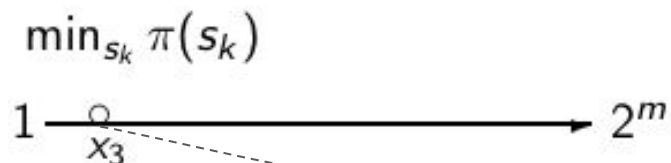
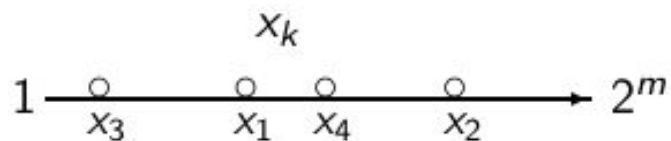
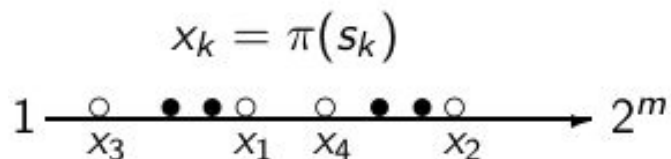
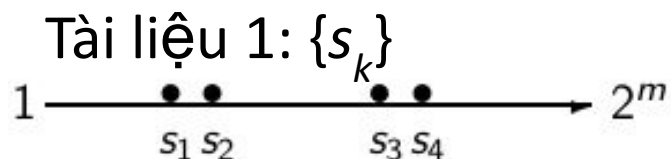


Biểu diễn khung của văn bản

- Tính vec-tơ khung (kích thước ~ 200) cho văn bản
 - Tính tổng đại diện cho các chuỗi n-từ
 - Ký hiệu s_i là tổng đại diện của chuỗi n-từ thứ i , $1 \leq s_i \leq 2^m$.
 - Sử dụng các phép trộn $\pi_1 \dots \pi_K$ trên tập tổng đại diện
 - Thành phần trong vec-tơ khung tương ứng với một phép trộn là cực tiểu của kết quả trộn.
 - Vec-tơ khung của một văn bản d là:
$$\langle \min_{s \in d} \pi_1(s), \min_{s \in d} \pi_2(s), \dots, \min_{s \in d} \pi_K(s) \rangle$$

Ước lượng độ tương đồng bằng tỉ lệ thành phần chung của các vec-tơ khung.

Khái niệm phép trộn thành công



Bằng nhau?

Phép trộn cho các thành phần cực tiểu bằng nhau được gọi là phép trộn thành công

Ước lượng độ tương đồng

- Ước lượng tỉ lệ phép trộn thành công
 - 1. Sử dụng K phép trộn, v.d., $K = 200$
 - 2. Đếm số lượng phép trộn thành công, ký hiệu là k
 - 3. Ước lượng: k/K là giá trị gần đúng của $J(d_1, d_2)$.
- *Cơ sở lý thuyết: Với số lượng phép trộn đủ lớn, hệ số Jaccard của hai tập chuỗi n -tử bằng tỉ lệ phép trộn thành công trên các tập tổng đại diện.*

Chứng minh?

Độ tương đồng của hai văn bản

- Biểu diễn các tập tổng đại diện như các cột của một ma trận O (occurrences), mỗi dòng cho một giá trị trong miền giá trị của tổng đại diện, $o_{ij} = 1$ nếu giá trị i xuất hiện trong tập j , $o_{ij} = 0$ nếu ngược lại
- Ví dụ

C_1	C_2	
0	1	
1	0	
1	1	$Jaccard(C_1, C_2) = 2/5 = 0.4$
0	0	
1	1	
0	1	

Độ tương đồng của hai văn bản₍₂₎

- Với 2 cột C_i và C_j chúng ta có bốn loại dòng

	C_i	C_j
A	1	1
B	1	0
C	0	1
D	0	0

- Ký hiệu: $A = \# \text{dòng kiểu A, ...}$
- Chúng ta có

$$\text{Jaccard}(C_i, C_j) = A / (A + B + C)$$

Ước lượng độ tương đồng

- Phép trộn có hiệu ứng tương tự trộn ngẫu nhiên các dòng của ma trận O
- Ký hiệu m là dòng đầu tiên trong ma trận O có ít nhất một giá trị bằng 1
- Ký hiệu hàm băm $h(C_i)$ là chỉ số dòng đầu tiên có giá trị 1 ở cột C_i
 - Đồng thời là cực tiểu của tập tổng đại diện
 - $h(C_i)$ còn được gọi là hàm băm cực tiểu
- Chúng ta có
 - $h(C_i) = h(C_j) \Leftrightarrow$ dòng m thuộc nhóm A (1 1)
 - $p(h(C_i) = h(C_j)) = P(A) = A/(A+B+C) = \text{Jaccard}(C_i, C_j)$

Ví dụ 9.3. Ước lượng độ tương đồng

	C_1	C_2	C_3
R_1	1	0	1
R_2	0	1	1
R_3	1	0	0
R_4	1	0	1
R_5	0	1	0

Các phép trộn

$$\pi_1 = (12345)$$

$$\pi_2 = (54321)$$

$$\pi_3 = (34512)$$

Vec-tơ khung

S_1	S_2	S_3
1	2	1
4	5	4
3	5	4

Độ tương đồng

Giá trị thực:

Ước lượng:

1-2	1-3	2-3
0.00	0.50	0.25
0.00	0.67	0.00

Các giải pháp phát hiện trùng lặp gần khác

- Vấn đề: Cần phải tính $N*(N-1)/2$ giá trị tương đồng.
 - Khối lượng tính toán lớn vẫn còn lớn.
- Các giải pháp cải thiện hiệu năng:
 - Hàm băm cục bộ (LSH)

Andoni, Alexandr, Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab Mirrokni. 2006. Locality-sensitive hashing using stable distributions. In Nearest Neighbor Methods in Learning and Vision: Theory and Practice. MIT Press. 314, 519, 522, 524, 527

- Giải pháp dựa trên sắp xếp (Henzinger 2006)

Henzinger, Monika R., Allan Heydon, Michael Mitzenmacher, and Marc Najork. 2000. On near-uniform URL sampling. In Proc. WWW, pp. 295–308. North-Holland. DOI: [dx.doi.org/10.1016/S1389-1286\(00\)00055-4](https://doi.org/10.1016/S1389-1286(00)00055-4). 442, 524, 527, 528

Bài tập 9.1

Cho các tập tổng đại diện của văn bản

	C_1	C_2	C_3
R_1	0	1	1
R_2	1	0	1
R_3	1	1	0
R_4	1	0	1
R_5	0	1	1
R_6	1	0	1

Sử dụng các phép trộn

$$\pi_1 = (1, 3, 5, 2, 4, 6), \pi_2 = (6, 5, 4, 3, 2, 1), \pi_3 = (2, 1, 5, 3, 6, 4)$$

Hãy ước lượng các giá trị độ tương đồng và so sánh với hệ số Jaccard
 $J(d_1, d_2), J(d_1, d_3), J(d_2, d_3)$

