

TÀI LIỆU LÝ THUYẾT KHAI THÁC WEB

Chủ đề 3

**KHAI THÁC NỘI DUNG WEB
(PHẦN 1)**

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

NỘI DUNG

- Định nghĩa Khai thác Nội dung Web
- Tiền xử lý trang Web và văn bản
 - Một số kỹ thuật tiền xử lý
 - Đánh chỉ mục trong cơ sở dữ liệu lớn

ĐỊNH NGHĨA KHAI THÁC NỘI DUNG WEB

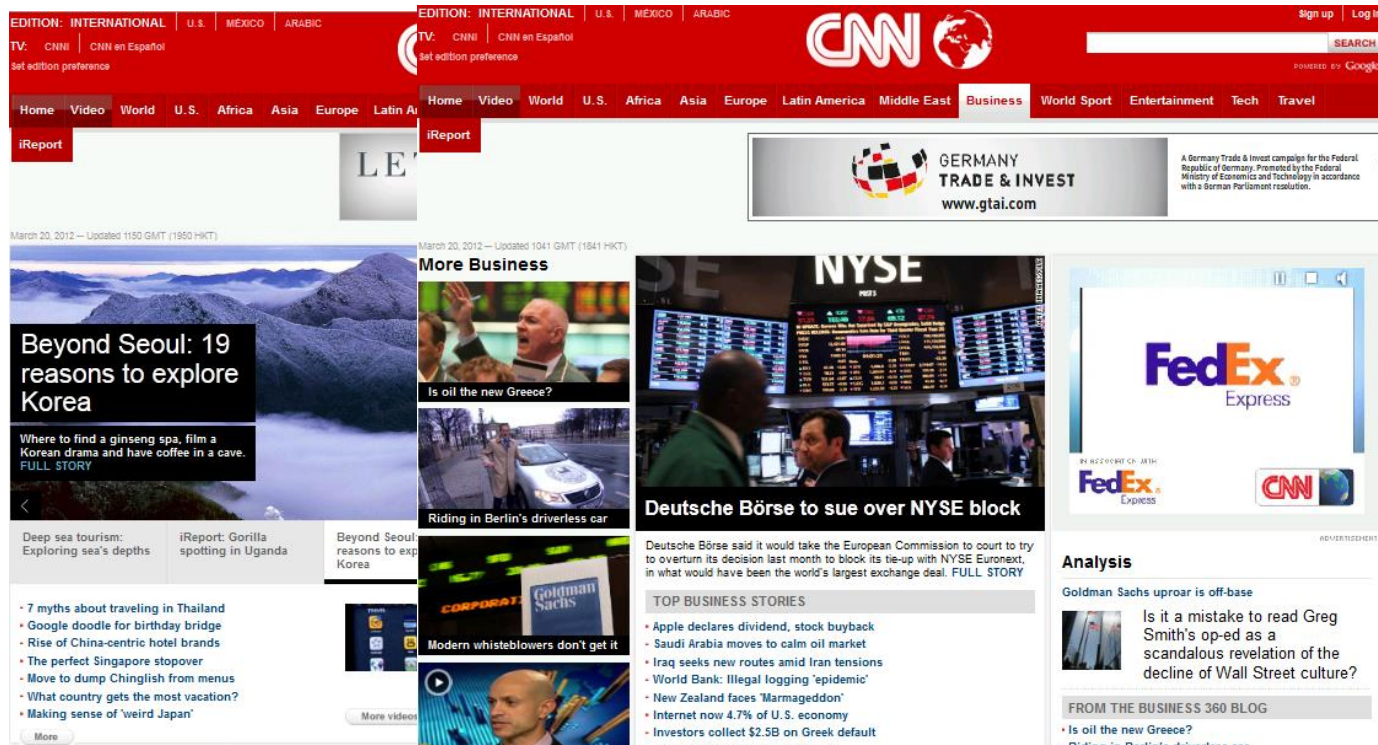


KHAI THÁC NỘI DUNG WEB

- **Mục tiêu:** Rút trích tri thức hữu ích từ nội dung của các tài liệu Web.
 - Dữ liệu Web: 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.
 - Các mô hình truy vấn thông tin và độ đo đánh giá
 - Kỹ thuật xử lý văn bản: stemming, pos-tagging,...

KHAI THÁC NỘI DUNG WEB

- Ứng dụng của khai thác nội dung Web
 - Nhận diện các chủ đề trong tài liệu Web, phân loại/gom nhóm trang Web theo chủ đề.



KHAI THÁC NỘI DUNG WEB

• Ứng dụng của

6wzxp0.joulottospel.com

FAKE

DOMAIN

www.eonet.com

e1Net
Internet Marketing Coach Asia No.1
Making You No.1 on the Net!

Internet Marketing Coach | Be Our Affiliate | Contact Us | Site Map
eOneNet.com, top Internet marketing company Asia and its customers are still growing their business online globally, selling B2B, B2C and even digital products on the Internet without any advertising costs."

HOME e1SEMINARS e1WEB e1PROMOTE e1LISTING e1SHOPPING e1WEBTOOL e1CLUB e1MAIL

Intro Join e1Club Seminars Free e-Cards Success Interview Make Money Online
Login Site Builder

Internet Marketing Coach says, "No traffic and no sales for your website despite reading internet marketing ebooks and attending Internet marketing seminars? Discover how Internet marketing coach, eOneNet.com, top Internet marketing company Asia and its customers are still growing their business online globally, selling B2B, B2C and even digital products on the Internet without any advertising costs."

#1 Asia Internet Marketing Company Warning : Don't attend another Internet marketing seminar until you've READ THIS!

Ranked #1 Internet Marketing Coaching
More about Internet Marketing Coach at FioneTan.com
"Discover how we make millions of dollars each year from the Internet and you can do the same." Fione Tan, President & CEO, eOneNet.com

No.1 Internet Marketing Coaching
Want to start making money online in 60 days from zero, even on low budget and with no products?
Discover how our graduates are making money by selling products, services, information online, and how you can do the same. **CLICK HERE**

No.1 Internet Marketing Video
Want to make your First Million online? Apply the secret strategies of the World's No.1 Internet Marketing Coach. **CLICK HERE**

H1N1 Mask-Swine Flu Mask
Buy original N95 swine flu mask
Guaranteed original 3M N95 mask
Worldwide shipping direct from N95 mask wholesaler.

Hong Kong CEF
Hong Kong CEF approved eOneNet's Certificate for Practical Internet Marketing up to HK\$10,000 reimbursable

Internet Marketing Seminars
Malaysia Internet Marketing Seminar
New Web 3.0 Half Day

Download Fione Tan's New eBook

Name: Country: Select

Download Fione Tan's New eBook

Name: Country: Select Gender: Select

KHAI THÁC NỘI DUNG WEB

- Ứng dụng của khai thác nội dung Web
 - Truy vấn thông tin: đưa vào câu truy vấn, hệ thống trả về nội dung trang Web có liên quan.

Query: publisher



amazon.com Hello. Sign in to get [personalized recommendations](#). New customer? [Start here](#)

Your Amazon.com | [Today's Deals](#) | [Gifts & Wish Lists](#) | [Gift Cards](#)

Shop All Departments Search Books

Books Advanced Search Browse Subjects New Releases Deals

The Da Vinci Code: A Novel and over 300,000 other books are available

Click to **LOOK INSIDE!**

THE DA VINCI CODE

DAN BROWN

The Da Vinci Code (Mass Market Paperback)
by [Dan Brown](#) (Author)
Key Phrases: [divine proportion](#), [saint graal](#), [hieros gamos](#), [Holy](#)
★★★★☆ (3,925 customer reviews)

List Price: \$9.99

Price: **\$9.99** & eligible for **FREE Super Saver Shipping**
[Special Offers Available](#)

In Stock.
Ships from and sold by Amazon.com. Gift-wrap available.

Product Details

Mass Market Paperback: 608 pages

Publisher: Anchor (March 31, 2009)

Language: English

ISBN-10: 0307474275

ISBN-13: 978-0307474278

[Share your own customer images](#)

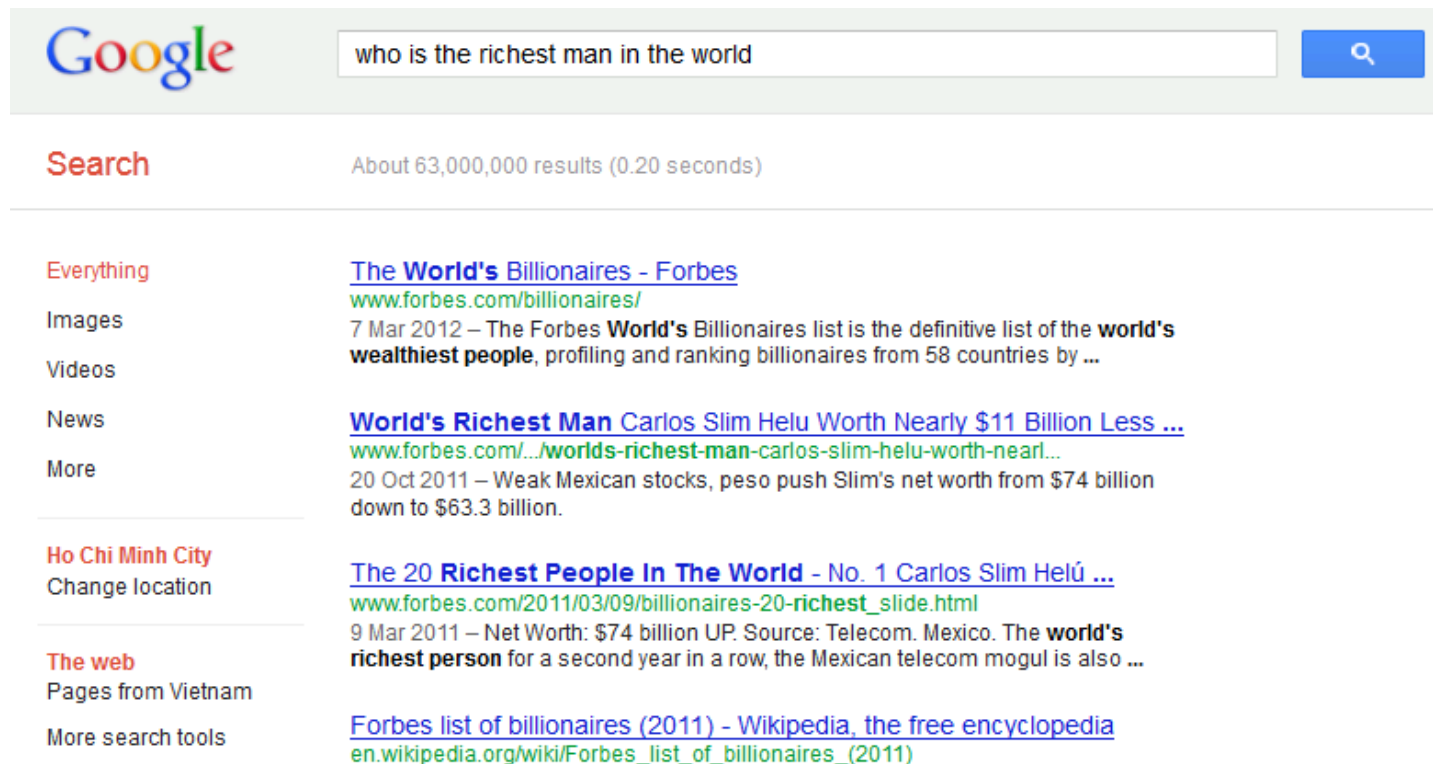
[Search inside another edition of this book](#)

Start reading The Da Vinci

Publisher: Anchor (March 31, 2009)

KHAI THÁC NỘI DUNG WEB

- Ứng dụng của khai thác nội dung Web
 - **Truy vấn thông tin:** là một bộ phận cần thiết cho các công cụ tìm kiếm.



The screenshot shows a Google search interface. The search bar contains the text "who is the richest man in the world" and a blue search button with a magnifying glass icon. Below the search bar, the word "Search" is displayed in red, followed by the text "About 63,000,000 results (0.20 seconds)".

On the left side, there is a vertical menu with the following options: "Everything", "Images", "Videos", "News", "More", "Ho Chi Minh City", "Change location", "The web", "Pages from Vietnam", and "More search tools".

The main content area displays three search results:

- The World's Billionaires - Forbes**
www.forbes.com/billionaires/
7 Mar 2012 – The Forbes **World's Billionaires** list is the definitive list of the **world's wealthiest people**, profiling and ranking billionaires from 58 countries by ...
- World's Richest Man Carlos Slim Helu Worth Nearly \$11 Billion Less ...**
www.forbes.com/.../worlds-richest-man-carlos-slim-helu-worth-nearl...
20 Oct 2011 – Weak Mexican stocks, peso push Slim's net worth from \$74 billion down to \$63.3 billion.
- The 20 Richest People In The World - No. 1 Carlos Slim Helú ...**
www.forbes.com/2011/03/09/billionaires-20-richest_slide.html
9 Mar 2011 – Net Worth: \$74 billion UP. Source: Telecom. Mexico. The **world's richest person** for a second year in a row, the Mexican telecom mogul is also ...
- Forbes list of billionaires (2011) - Wikipedia, the free encyclopedia**
[en.wikipedia.org/wiki/Forbes_list_of_billionaires_\(2011\)](http://en.wikipedia.org/wiki/Forbes_list_of_billionaires_(2011))

KHAI THÁC NỘI DUNG WEB

- Kết quả khai thác nội dung Web có thể là dữ liệu cho các bài toán Web khác.
 - Tìm kiếm các nhận xét về một sản phẩm (từ trang bán hàng, blog, forum,...): phân tích và cải tiến sản phẩm, tìm hiểu sở thích người dùng,...



☆☆☆☆☆ First impression **FAIL FAIL FAIL**

I am so angry right now First thing I tryd to do is set up a custom ring tone off of one of my songs. 3 hours of FAIL. Be warned.

Published 1 day ago by Doppleganger

☆☆☆☆☆ iPhone

Way better than my only Sony I had. better reception, better sounds and pictures. Only thing its a bit more expensive but to me is worth it.

Published 1 month ago by Jill "333"

☆☆☆☆☆ Doubt

Does anybody knows if this Iphone is CDMA or GSM?

I want to buy it, but I need to know it first.

Published 1 month ago by Renan

Từ mạng
cảm nghĩ



CÁC KỸ THUẬT TIỀN XỬ LÝ

TIỀN XỬ LÝ VĂN BẢN

- Rút trích từ: dễ dàng đối với tiếng Anh
- Loại bỏ Stopword
 - Ví dụ: a, an, the, will, with...
- Stemming: chuyển biến thể từ về thể gốc
 - Ví dụ: going → go, went → go
- Đếm tần số xuất hiện từ và tính trọng số từ TF-IDF
- Kỹ thuật khác: loại bỏ dấu câu, xử lý chữ hoa/thường

STOPWORD

- **Stopword** là các từ xuất hiện thường xuyên trong câu, giúp xây dựng câu nhưng không biểu đạt nội dung của tài liệu.
 - Article, preposition, conjunction và một số pronouns. Có khoảng 400 đến 500 từ.
 - Ví dụ: a, about, an, are, as, at, be, by, for, from, how, in, is, of, on, or, that, the, these, this, to, was, what, when, where, who, will, with,...
- Danh sách stopwords được xây dựng đặc trưng cho mỗi ứng dụng.

LOẠI BỎ STOPWORD

- Stopword cần được loại bỏ trước khi câu được đánh chỉ mục và lưu trữ.
- Câu truy vấn cũng cần loại bỏ stopwords.
- **Tại sao ta cần phải bỏ stopwords?**
 - Giảm kích thước tập tin chỉ mục (hoặc dữ liệu)
 - Stopword chiếm 20-30% tổng số từ trong tài liệu.
 - Tăng hiệu suất và hiệu quả
 - Stopword gây nhiễu cho quá trình tìm kiếm và khai thác văn bản

STEMMING

- Trong ngôn ngữ, một từ có nhiều thể cú pháp khác nhau phụ thuộc vào ngữ cảnh sử dụng.
- Ví dụ xét ngôn ngữ tiếng Anh
 - Danh từ có thể số ít và số nhiều: **apple** và **apples**
 - Động từ có thể nguyên bản và tiếp diễn (+ing): **eat** và **eating**.
 - Động từ ở các thì khác nhau: **eat**, **ate** và **eaten**

ẢNH HƯỞNG STEMMING

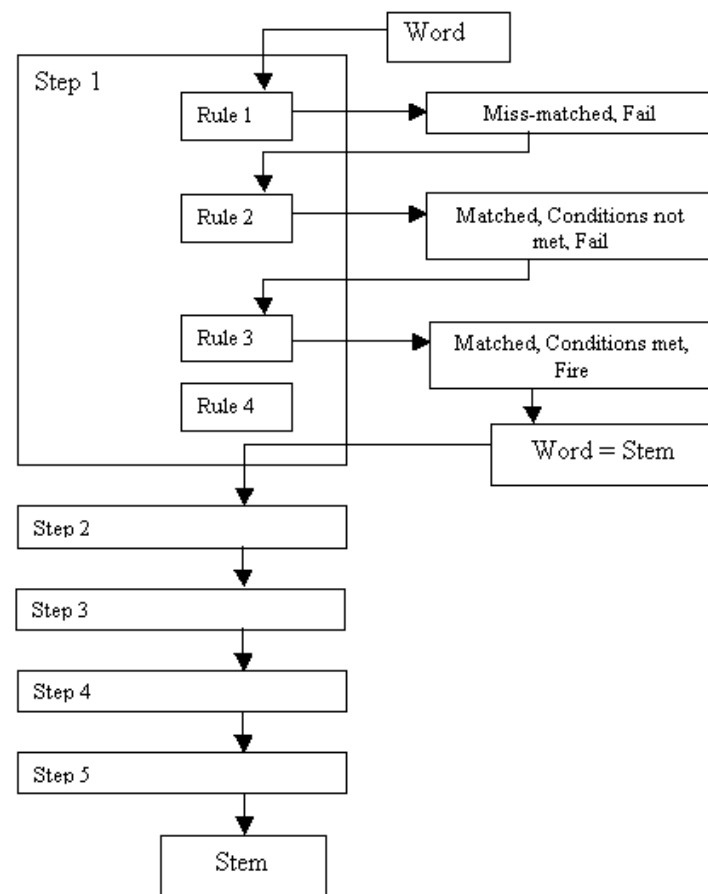
- Các thể cú pháp của một từ được xem như những biến thể của cùng một thể gốc.
- Những biến thể này làm giảm độ bao phủ (recall) của một hệ thống truy vấn.
 - Tài liệu liên quan có thể chứa biến thể của từ truy vấn nhưng không phải từ gốc.
- Vấn đề này có thể giải quyết một phần bằng phương pháp **Stemming**.

PHƯƠNG PHÁP STEMMING

- **Stemming** là quá trình biến đổi một từ về thể gốc (gọi là **stem** hay **root**).
- **Stem** là phần còn lại của từ sau khi loại bỏ tiền tố và hậu tố.
 - Ví dụ: “computer”, “computing”, “compute” → “comput.”; “walks”, “walking”, “walker” → “walk”
- Có nhiều thuật toán stemming khác nhau, gọi là **stemmers**.

PORTER STEMMER

- Do Martin F. Porter đề xuất vào năm 1980.
- Thuật toán hoạt động theo cơ chế kiểm tra tính thỏa của từ đối với một bộ luật.
- Ví dụ bước 1
 - **SS**ES → **SS**
caresses → caress
 - **I**ES → **I**
ponies → poni
ties → ti



<http://www.comp.lancs.ac.uk/computing/research/stemming/general/porter.htm>

LỢI ÍCH PP STEMMING

- Stemming giúp tăng hiệu quả truy vấn và khai thác văn bản.
 - So khớp các từ tương tự, cải thiện độ bao phủ
- Stemming làm giảm kích thước của cấu trúc đánh chỉ mục.
 - Kết hợp các từ có chung thể gốc có thể giảm kích thước chỉ mục đến 40-50%.

BẤT LỢI PP STEMMING

- Tuy nhiên, Stemming có thể ảnh hưởng đến độ chính xác vì tài liệu không liên quan cũng bị xem là liên quan.
 - Ví dụ: “cop” và “cope” → “cop”, nhưng tài liệu chỉ chứa “cope” không liên quan đến chủ đề cảnh sát.
- Khi áp dụng, cần thực nghiệm trước hiệu quả stemming trên tập dữ liệu.

TẦN SỐ TỪ (TF-IDF)

- Tần số từ (term frequency): số lần xuất hiện của một từ trong một tài liệu.
 - Sử dụng tần số xuất hiện để chỉ độ quan trọng tương đối của từ trong tài liệu.
 - Nếu một từ xuất hiện thường xuyên trong văn bản thì văn bản có liên hệ với chủ đề mà từ biểu diễn.
- Tần số tài liệu (document frequency): số tài liệu trong ngữ liệu chứa một từ xác định.

TẦN SỐ TỪ – TF-IDF

- TF-IDF (term frequency – inverse document frequency) cho biết độ quan trọng của một từ đối với một tài liệu trong ngữ liệu.
- Công thức: $tf * idf(t, d, D) = tf(t, d) \times idf(t, D)$
 - $tf(t, d)$: tần số xuất hiện từ t trong tài liệu d
 - $idf(t, D) = \log \frac{|D|}{|\{d \in D: t \in d\}|}$, $|D|$ là tổng số ngữ liệu trong tài liệu.

XỬ LÝ CHỮ SỐ

- Đối với hệ thống truy vấn truyền thống, số và chữ số cần được loại bỏ, ngoại trừ một số loại đặc biệt (ví dụ ngày tháng, thời gian, mẫu định trước,...).
- Tuy nhiên, trong các công cụ tìm kiếm, chúng thường được đánh chỉ mục.

DẤU NỔI, DẤU NGẮT

- Loại bỏ dấu nổi thường được áp dụng để giải quyết vấn đề không nhất quán.
 - Ví dụ: “state-of-the-art” và “state of the art”
- Hệ thống thường tuân theo một luật tổng quát (ví dụ loại bỏ mọi dấu nổi) và một vài ngoại lệ.
 - Ngoại lệ: dấu nổi là một phần của từ, ví dụ “Y-21”

XỬ LÝ DẤU NỔI, DẤU NGẮT

- Có hai cách loại bỏ dấu nổi
 - Mỗi dấu nổi được thế bằng một khoảng trắng
 - Ví dụ: “state-of-the-art” → “state of the art”
 - Bỏ dấu nổi nhưng không thế bằng khoảng trắng.
 - Ví dụ: “state-of-the-art” → “stateoftheart”
- Một số hệ thống đánh chỉ mục cho cả hai dạng này.
- **Dấu ngắt** cũng có thể được xử lý theo cách tương tự.

XỬ LÝ CHỮ HOA/THƯỜNG

- **Giải pháp:** Mọi ký tự được chuyển thành chữ hoa hay chữ thường.
 - Ví dụ: từ “CapTalizE” chuyển thành “capitalize” hay “CAPITALIZE”

TIỀN XỬ LÝ TRANG WEB

- Nhận diện các vùng văn bản
 - Văn bản trong vùng title, metada, body.
- Nhận diện anchor text
 - Ví dụ: YouTube là trang Web chia sẻ video clip.
- Loại bỏ HTML Tag
- Nhận diện khối nội dung chính

NHẬN DIỆN VÙNG VĂN BẢN

- Trong HTML, có nhiều vùng văn bản khác nhau (title, metadata, và body).
- Các hệ thống truy vấn đặt độ ưu tiên cho từ ngữ ở những vùng này khác biệt nhau.
 - Từ trong title quan trọng hơn trong các vùng văn bản khác
 - Từ nhấn mạnh (thuộc header tag <h1>, <h2>, ..., bold tag , ...) được gán trọng số cao hơn.

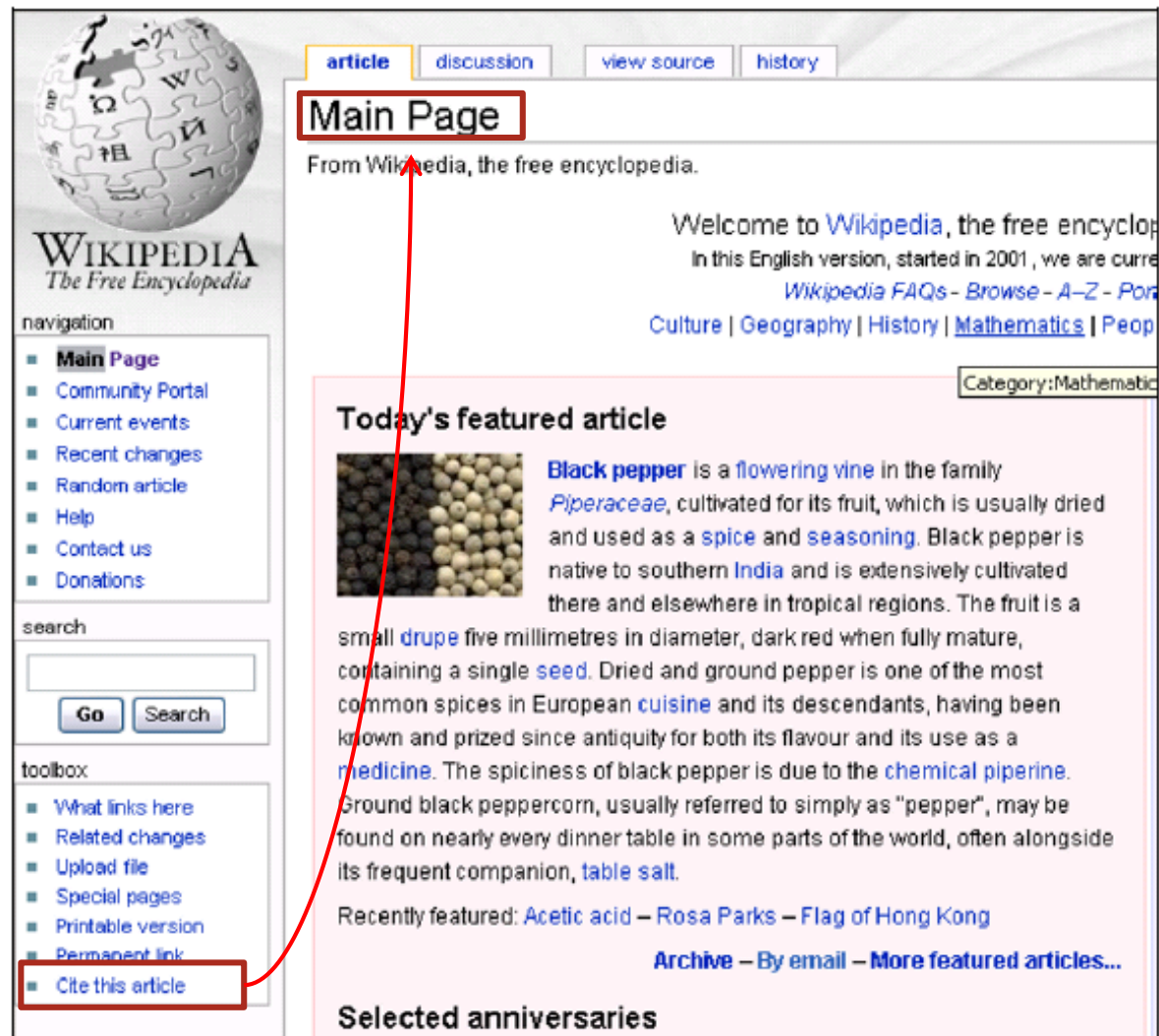
NHẬN DIỆN ANCHOR TEXT

- **Anchor text** là đoạn văn bản dùng để hiển thị link và người dùng nhấp vào để tới một site khác.
 - Ví dụ: a hyperlink to the [English-language Wikipedia's homepage](#)
- Được xử lý đặc biệt trong các công cụ tìm kiếm vì chứa đựng mô tả chính xác về thông tin của trang được trở tới.

LOẠI BỎ HTML TAG

- Việc loại bỏ HTML Tag có thể thực hiện tương tự như đối với dấu ngắt.
- **Vấn đề khó khăn:** những đoạn văn bản không được phép nối với nhau sẽ bị nối sau khi gỡ bỏ các tag.
⇒ ảnh hưởng đến các phép truy vấn.
- Đến nay vẫn chưa có giải pháp thỏa đáng.

LOẠI BỎ HTML TAG



The screenshot shows the Wikipedia Main Page. A red box highlights the 'Main Page' link in the left sidebar under the 'navigation' section. A red arrow points from this box to the 'Main Page' heading at the top of the main content area. Another red box highlights the 'Cite this article' link in the 'toolbox' section of the sidebar.

article | discussion | view source | history


Main Page

From Wikipedia, the free encyclopedia.

Welcome to [Wikipedia](#), the free encyclopedia in this English version, started in 2001, we are currently [Wikipedia FAQs](#) - [Browse](#) - [A-Z](#) - [Portals](#) - [Culture](#) | [Geography](#) | [History](#) | [Mathematics](#) | [People](#)

Category:Mathematics

Today's featured article



Black pepper is a [flowering vine](#) in the family [Piperaceae](#), cultivated for its fruit, which is usually dried and used as a [spice](#) and [seasoning](#). Black pepper is native to southern [India](#) and is extensively cultivated there and elsewhere in tropical regions. The fruit is a small [drupe](#) five millimetres in diameter, dark red when fully mature, containing a single [seed](#). Dried and ground pepper is one of the most common spices in European [cuisine](#) and its descendants, having been known and prized since antiquity for both its flavour and its use as a [medicine](#). The spiciness of black pepper is due to the [chemical](#) [piperine](#). Ground black peppercorn, usually referred to simply as "pepper", may be found on nearly every dinner table in some parts of the world, often alongside its frequent companion, [table salt](#).

Recently featured: [Acetic acid](#) - [Rosa Parks](#) - [Flag of Hong Kong](#)


[Archive](#) - [By email](#) - [More featured articles...](#)

Selected anniversaries

NHẬN DIỆN KHỎI NỘI DUNG

- Các trang Web thông thường, đặc biệt là trang thương mại, chứa một lượng lớn thông tin không phải là nội dung chính.
 - Ví dụ: banner ads, navigation bar, copyright notices,...
- Những nội dung bên lề làm cho quá trình tìm kiếm và khai thác có kết quả thấp.

NHẬN DIỆN KHỎI NỘI DUNG



WIKIPEDIA
The Free Encyclopedia

navigation

- **Main Page**
- Community Portal
- Current events
- Recent changes
- Random article
- Help
- Contact us
- Donations

search

toolbox

- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link
- Cite this article

[article](#) [discussion](#) [view source](#) [history](#)

Main Page


From Wikipedia, the free encyclopedia.

Welcome to [Wikipedia](#), the free encyclopedia.
In this English version, started in 2001, we are currently at [Wikipedia:Welcome](#).
[Wikipedia FAQs](#) - [Browse](#) - [A-Z](#) - [Portals](#) - [Recent changes](#) - [Random article](#) - [Help](#) - [Contact us](#) - [Donations](#)

[Culture](#) | [Geography](#) | [History](#) | [Mathematics](#) | [People](#) | [Science](#) | [Society](#) | [Sports](#) | [Technology](#) | [Arts](#) | [Health](#) | [Education](#) | [Business](#) | [Law](#) | [Politics](#) | [Religion](#) | [Environment](#) | [Travel](#) | [Hobbies](#) | [Games](#) | [Animals](#) | [Plants](#) | [Fiction](#) | [Non-fiction](#) | [Reference](#) | [List of all pages](#)

Category:Mathematics

Today's featured article



Black pepper is a [flowering vine](#) in the family [Piperaceae](#), cultivated for its fruit, which is usually dried and used as a [spice](#) and [seasoning](#). Black pepper is native to southern [India](#) and is extensively cultivated there and elsewhere in tropical regions. The fruit is a small [drupe](#) five millimetres in diameter, dark red when fully mature, containing a single [seed](#). Dried and ground pepper is one of the most common spices in European [cuisine](#) and its descendants, having been known and prized since antiquity for both its flavour and its use as a [medicine](#). The spiciness of black pepper is due to the [chemical](#) [piperine](#). Ground black peppercorn, usually referred to simply as "pepper", may be found on nearly every dinner table in some parts of the world, often alongside its frequent companion, [table salt](#).

Recently featured: [Acetic acid](#) - [Rosa Parks](#) - [Flag of Hong Kong](#)

[Archive](#) - [By email](#) - [More featured articles...](#)

Selected anniversaries

PHÂN VÙNG TRANG WEB

- Phân vùng theo dấu hiệu thị giác: sử dụng thông tin thị giác để tìm khối nội dung chính trong một trang.
- Thông tin có thể lấy từ trình duyệt Web.
 - Ví dụ: API của Internet Explorer cung cấp tọa độ X và Y của mỗi phần tử HTML.
- Mô hình học máy sử dụng đặc trưng tọa độ và biểu hiện để nhận diện các khối nội dung chính.

SO KHỚP CÂY

- Phương pháp so khớp cây tìm những mẫu cấu trúc tiềm ẩn của trang Web.
 - Hầu hết Website thương mại được phát sinh bằng một số mẫu cố định.
- Vì HTML có cấu trúc lồng, ta dễ dàng xây dựng tag tree cho mỗi trang.
- Thực hiện so khớp cây trên nhiều trang của cùng một site để tìm mẫu tiềm ẩn.

SO KHỚP CÂY

- Khi tìm thấy mẫu, ta có thể nhận diện khối nội dung chính bằng nhận xét sau:
 - Văn bản trong khối nội dung chính thường khác nhau xa giữa các trang có cùng mẫu.
 - Các khối nội dung phụ thường giống nhau qua nhiều trang.
- Để xác định độ tương tự văn bản giữa các khối (cây con), áp dụng phương pháp **shingle**.

PHÁT HIỆN TRÙNG

- Trùng lặp văn bản không phải là vấn đề trong truy vấn văn bản truyền thống.
- Tuy nhiên, trong **ngữ cảnh Web**, đây là **vấn đề nghiêm trọng**.



- Phát hiện trùng giúp giảm kích thước chỉ mục và cải thiện kết quả tìm kiếm.

PHÁT HIỆN TRÙNG

- Sao chép một trang gọi là **duplication** hay **replication**, sao chép toàn site là **mirroring**.
- Các trang hay site bản sao thường được dùng để tăng hiệu quả tìm kiếm và tải tài liệu toàn cầu.
 - Do giới hạn băng thông, chất lượng mạng xấu.
- Bên cạnh đó, một số trang bản sao là ăn cắp bản quyền.

PHƯƠNG PHÁP BĂM

- “Băm” (hash) toàn bộ tài liệu.
 - Ví dụ: thuật toán MD5, tính toán số tích hợp (checksum).
- Những phương pháp này chỉ có thể phát hiện những bản sao chính xác.
 - Trên Web, rất hiếm khi gặp bản sao chính xác.
 - Ngay cả mirror site cũng có URL, Web master, thông tin liên lạc, mục quảng cáo khác nhau tùy theo nhu cầu.

PHƯƠNG PHÁP SHINGLE

- Đây là kỹ thuật phát hiện trùng hiệu quả dựa trên ***n*-grams** (còn gọi là ***shingles***).
- *n*-gram là một chuỗi các từ liên tiếp nhau có kích thước cửa sổ *n*.
 - Ví dụ: câu “John went to school with his brother” có thể biểu diễn bằng 5 cụm từ 3-gram “John went to”, “went to school”, “to school with”, “school with his”, và “with his brother”.
 - 1-gram biểu diễn các từ đơn.

PHƯƠNG PHÁP SHINGLE

- Gọi $S_n(d)$ là tập hợp các n -gram (shingle) phân biệt nhau chứa trong một tài liệu d .
- Mỗi n -gram được mã hóa bằng một con số hoặc giá trị băm MD5 (hệ thập lục phân 32 chữ số).

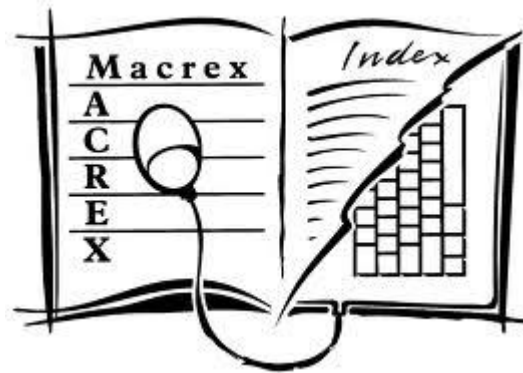
PHƯƠNG PHÁP SHINGLE

- Cho trước biểu diễn n -gram của hai tài liệu d_1 và d_2 , $S_n(d_1)$ và $S_n(d_2)$, **hệ số Jaccard** tính độ tương tự:

$$\text{sim}(d_1, d_2) = \frac{|S_n(d_1) \cap S_n(d_2)|}{|S_n(d_1) \cup S_n(d_2)|}$$

- So sánh $\text{sim}(d_1, d_2)$ với **ngưỡng** để xác định d_1 và d_2 có phải là bản sao của nhau.
- Kích thước cửa sổ n và ngưỡng tương tự được chọn bằng thực nghiệm.

ĐÁNH CHỈ MỤC TRONG CƠ SỞ DỮ LIỆU LỚN



GIỚI THIỆU VẤN ĐỀ

- Phương pháp cơ bản của tìm kiếm Web và truy vấn thông tin là duyệt cơ sở dữ liệu tuần tự để tìm tài liệu chứa từ truy vấn.
⇒ không khả thi cho dữ liệu lớn như Web
- Một số cấu trúc dữ liệu, gọi là **chỉ mục (index)**, có thể giúp tăng tốc việc tìm kiếm.

CHỈ MỤC ĐẢO

- Chỉ mục đảo (*inverted index*) là cấu trúc dữ liệu tổ chức cho mỗi từ một danh sách các tài liệu chứa từ đó.
 - Thao tác tìm kiếm từ truy vấn tốn thời gian hằng.
 - Có thể tìm tài liệu chứa nhiều từ khóa cùng lúc.
- Là phương pháp đánh chỉ mục phổ biến trong các công cụ tìm kiếm.
 - Truy tìm tài liệu chứa từ khóa hữu hiệu
 - Dễ dàng xây dựng.

CHỈ MỤC ĐẢO

- Cho tập tài liệu $D = \{d_1, d_2, \dots, d_N\}$, mỗi tài liệu có ID đơn nhất.
- Chỉ mục đảo bao gồm hai bộ phận:
 - Ngữ vựng V chứa mọi từ phân biệt trong dữ liệu văn bản.
 - Với từ phân biệt t_i , danh sách đảo các posting.

CHỈ MỤC ĐẢO

- Posting lưu ID (kí hiệu id_i) của tài liệu d_i chứa từ t_i và những thông tin khác về t_i trong d_i , tùy ứng dụng.
 - Ví dụ: $\langle id_i, f_{ij}, [o_1, o_2, \dots, o_{|f_{ij}|}]$
 - id_i là ID của văn bản d_i chứa từ t_i , f_i là tần số của t_i trong d_i , O_k là vị trí của t_i trong d_i
- Posting của một từ được sắp xếp theo thứ tự tăng dần của ID và offset trong mỗi posting cũng vậy.

VÍ DỤ CHỈ MỤC ĐÀO

- Ta có ba tài liệu id_1 , id_2 , and id_3 :

id_1 : Web mining is useful.

1 2 3 4

id_2 : Usage mining applications.

1 2 3

id_3 : Web structure mining studies the Web hyperlink structure.

1 2 3 4 5 6 7 8

- Con số dưới mỗi tài liệu là offset của từ.
- Ngữ vựng $V = \{\text{Web, mining, useful, applications, usage, structure, studies, hyperlink}\}$
- Loại bỏ stopwords (“is”, “the”). Chưa stemming.

VÍ DỤ CHỈ MỤC ĐẢO

- Chỉ mục đảo đơn giản (A) và chỉ mục đảo phức tạp (B).
 - (B) chứa thêm thông tin về tần số từ, vị trí từ trong tài liệu.

Applications: id_2
Hyperlink: id_3
Mining: id_1, id_2, id_3
Structure: id_3
Studies: id_3
Usage: id_2
Useful: id_1
Web: id_1, id_3

(A)

Applications: $\langle id_2, 1, [3] \rangle$
Hyperlink: $\langle id_3, 1, [7] \rangle$
Mining: $\langle id_1, 1, [2] \rangle, \langle id_2, 1, [2] \rangle, \langle id_3, 1, [3] \rangle$
Structure: $\langle id_3, 2, [2, 8] \rangle$
Studies: $\langle id_3, 1, [4] \rangle$
Usage: $\langle id_2, 1, [1] \rangle$
Useful: $\langle id_1, 1, [4] \rangle$
Web: $\langle id_1, 1, [1] \rangle, \langle id_3, 2, [1, 6] \rangle$

(B)

TÌM KIẾM VỚI CHỈ MỤC ĐẢO

- Cho trước từ truy vấn, quá trình tìm tài liệu liên quan trong chỉ mục đảo gồm ba bước:
 1. **Tìm ngữ vựng**: tìm từ truy vấn trong tập ngữ vựng để xác định danh sách đảo.
 - Thực hiện trên bộ nhớ chính, áp dụng bảng băm, trie hay B-tree để tăng tốc tìm kiếm.
 - Cũng có thể áp dụng thứ tự từ điển kết hợp với tìm kiếm nhị phân.
 - Độ phức tạp $O(\log|V|)$, với $|V|$ là kích thước ngữ vựng.
 - Nếu câu truy vấn chỉ có một từ, sang bước 3. Nếu câu chứa nhiều từ, sang bước 2.

TÌM KIẾM VỚI CHỈ MỤC ĐẢO

- Cho trước từ truy vấn, quá trình tìm tài liệu liên quan trong chỉ mục đảo gồm ba bước:
 2. **Trộn kết quả**: trộn các danh sách đảo để tìm phần giao.
 - Heuristic: dùng danh sách ngắn nhất để trộn với những danh sách khác dài hơn. Với mỗi posting, áp dụng tìm kiếm nhị phân trên danh sách dài để tăng tốc.
 - Tìm kiếm từng phần cũng có thể thực hiện theo cách tương tự.
 - Mở rộng: phân tích độ phổ biến của từ truy vấn để chỉ cần tải một phần của chỉ mục đảo vào bộ nhớ.

TÌM KIẾM VỚI CHỈ MỤC ĐẠO

- Cho trước từ truy vấn, quá trình tìm tài liệu liên quan trong chỉ mục đảo gồm ba bước:
 3. **Tính số điểm hạng**: tính điểm số liên quan (hạng) cho mỗi tài liệu
 - Dựa trên hàm liên quan (ví dụ cosine hay okapi) xét độ gần của từ hay cụm từ.
 - Các điểm số này dùng cho việc xếp hạng tài liệu.

VÍ DỤ TÌM KIẾM CHỈ MỤC ĐẢO

- Sử dụng chỉ mục đảo (B) tại slide 46. Ta cần tìm “web mining”.
- Bước 1: tìm thấy hai danh sách đảo
 - Mining: $\langle id_1, 1, [2] \rangle, \langle id_2, 1, [2] \rangle, \langle id_3, 1, [3] \rangle$
 - Web: $\langle id_1, 1, [1] \rangle, \langle id_3, 2, [1, 6] \rangle$
- Bước 2: duyệt hai danh sách để tìm tài liệu chứa mọi từ $\Rightarrow id_1$ và id_3 .
- Bước 3: id_1 xếp hạng cao hơn id_3

XÂY DỰNG CHỈ MỤC ĐẢO

- Có thể thực hiện một cách đơn giản và hiệu quả bằng cấu trúc dữ liệu trie.
- Độ phức tạp thời gian là $O(T)$
 - T là số lượng từ (kể cả từ trùng) trong dữ liệu đã tiền xử lý.
- **Mô tả:** duyệt tuần tự từng tài liệu và với mỗi từ, tìm từ đó trong trie.
 - Nếu tìm thấy, thêm ID và thông tin khác (ví dụ vị trí từ) vào danh sách đảo của từ.
 - Nếu không tìm thấy, tạo nút lá mới cho từ

VÍ DỤ CẤU TRÚC TRIE

- Dữ liệu văn bản và cây trie tương ứng.

id_1 : Web mining is useful.

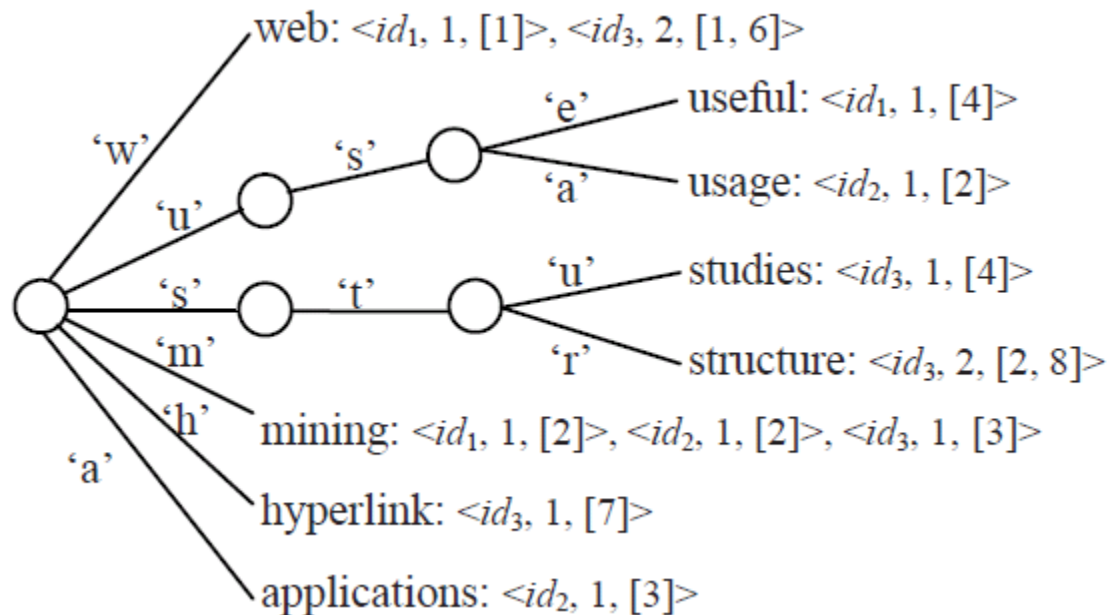
1 2 3 4

id_2 : Usage mining applications.

1 2 3

id_3 : Web structure mining studies the Web hyperlink structure

1 2 3 4 5 6 7 8



XÂY DỰNG CHỈ MỤC ĐẢO

- Thực hiện thuật toán sau để xây dựng chỉ mục từng phần cho đến khi bộ nhớ đầy.
 - Giả sử đã xây dựng chỉ mục từng phần I_1 và đang lưu trên đĩa. Xử lý các tài liệu tiếp theo và xây dựng chỉ mục I_2 trong bộ nhớ.
 - Cuối cùng, ta có k chỉ mục từng phần I_1, I_2, \dots, I_k trên đĩa.
 - Trộn các chỉ mục từng phần một cách phân cấp.
 - Trộn I_1 và I_2, I_3 và I_4, \dots tạo thành I_{1-2}, I_{3-4}, \dots . Trộn I_{1-2} và I_{3-4}, I_{5-6} và I_{7-8}, \dots
 - Tiếp tục cho đến khi tạo thành một chỉ mục đơn.

XÂY DỰNG CHỈ MỤC ĐẢO

- Đối với Web, vấn đề thêm, chỉnh sửa, hoặc xóa trang trên chỉ mục đảo rất quan trọng.
 - Một thay đổi nhỏ có thể cần cập nhật rất nhiều phần trong chỉ mục.

XÂY DỰNG CHỈ MỤC ĐẢO

- **Giải pháp:** tạo thêm hai chỉ mục, dành cho trang thêm vào và trang bị xóa.
 - Chỉnh sửa gồm phép xóa kể đến là phép thêm.
 - Việc tìm kiếm từ truy vấn thực hiện trên chỉ mục chính và cả hai chỉ mục phụ.
 - Gọi các trang trả về từ phép tìm kiếm trên chỉ mục chính là D_0 , từ chỉ mục trang thêm là D_+ và từ chỉ mục trang xóa là D_- .
 - Kết quả trả về cho người dùng là $(D_0 \cup D_+) - D_-$.

NÉN CHỈ MỤC

- Giảm kích thước chỉ mục là vấn đề quan trọng vì chỉ mục có thể rất lớn.
- Nén chỉ mục biểu diễn cùng lượng thông tin với số bit hoặc byte ít hơn.
 - **Nén không mất thông tin**: chỉ mục gốc có thể được tái tạo từ nội dung nén.
 - **Nén có mất thông tin**: chỉ mục gốc được tái tạo không hoàn chỉnh.

NÉN CHỈ MỤC

- Chỉ mục đảo lưu trữ chủ yếu là ID của tài liệu và vị trí từ.
 - Các thông tin được biểu diễn bằng số nguyên.
⇒ Áp dụng kỹ thuật nén số nguyên
- Mỗi số nguyên thường được biểu diễn bằng 4 byte (32 bit).
 - Hầu hết số nguyên không dùng hết 4 byte và do đó việc nén là khả thi.

NÉN SỐ NGUYÊN

- Các mô hình nén tổng quát dành cho danh sách đảo:
 1. **Mô hình variable-bit**: mỗi số nguyên được lưu bằng một lượng bit tích hợp.
 - Unary coding, Elias gamma coding, delta coding và Golomb coding
 2. **Mô hình variable-byte**: mỗi số nguyên được lưu bằng một lượng byte tích hợp.
 - Variable-byte coding.

LƯU TRỮ HIỆU SỐ ID

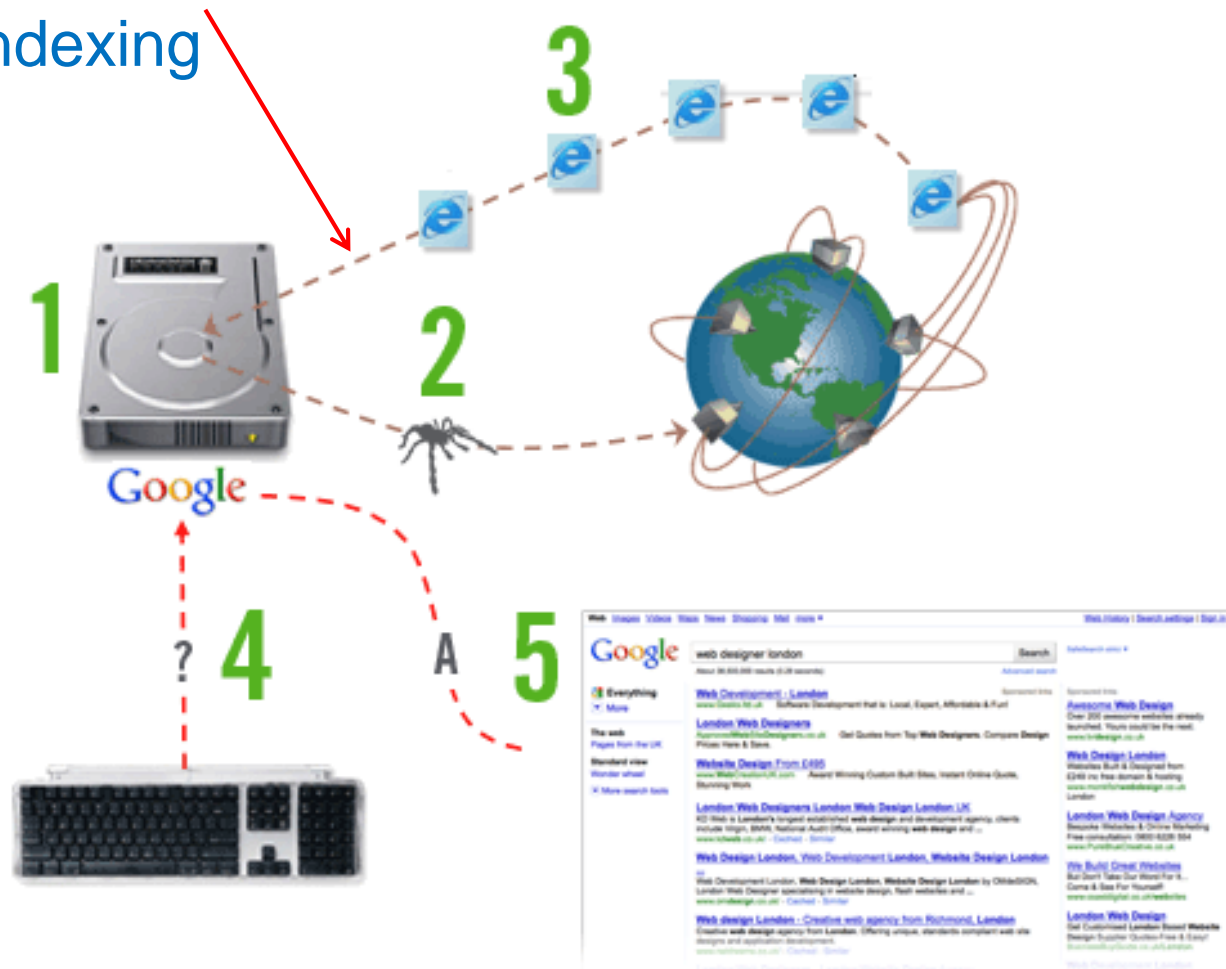
- Ta có thể lưu trữ hiệu số giữa hai ID liên kề, id_i và id_{i+1} ($id_{i+1} > id_i$), thay vì ID thật sự.
 - Do các ID trong danh sách đảo đã sắp tăng dần.
- Hiệu số này gọi là **khoảng cách (gap)** giữa id_i và id_{i+1} . Nó nhỏ hơn id_{i+1} và cần ít bit biểu diễn hơn.
- Ví dụ:
 - Các ID tài liệu đã sắp xếp: 4, 10, 300, và 305.
 - Biểu diễn bằng hiệu số ID: 4, 6, 290 và 5.

ĐÁNH CHỈ MỤC NGỮ NGHĨA

- Khái niệm và đối tượng có thể được biểu diễn theo nhiều cách khác nhau.
 - Tùy ngữ cảnh và thói quen ngôn ngữ.
 - Nếu truy vấn dùng từ khác với từ trong văn bản, tài liệu liên quan sẽ không được nhận diện.
 - Ví dụ: “picture”, “image” và “photo” là đồng nghĩa trong ngữ cảnh camera kỹ thuật số.
- Latent semantic indexing (Deerwester et al.) giải quyết vấn đề này bằng cách nhận diện liên kết thống kê giữa các từ.

TỔNG KẾT

1. Parsing
2. Indexing



TỔNG KẾT



1. Tiền xử lý câu truy vấn
2. Tìm trang liên quan trong chỉ mục đảo
3. Sắp hạng trang và trả cho người dùng



TÀI LIỆU THAM KHẢO

- Tài liệu bài giảng môn học
- **Chapter 6.** B. Liu, *Web Data Mining- Exploring Hyperlinks, Contents, and Usage Data*, Springer Series on Data-Centric Systems and Applications, 2007.

KẾT THÚC PHẦN I

