

Tài liệu giảng dạy môn Khai thác dữ liệu Web

# CHUẨN BỊ DỮ LIỆU WEB

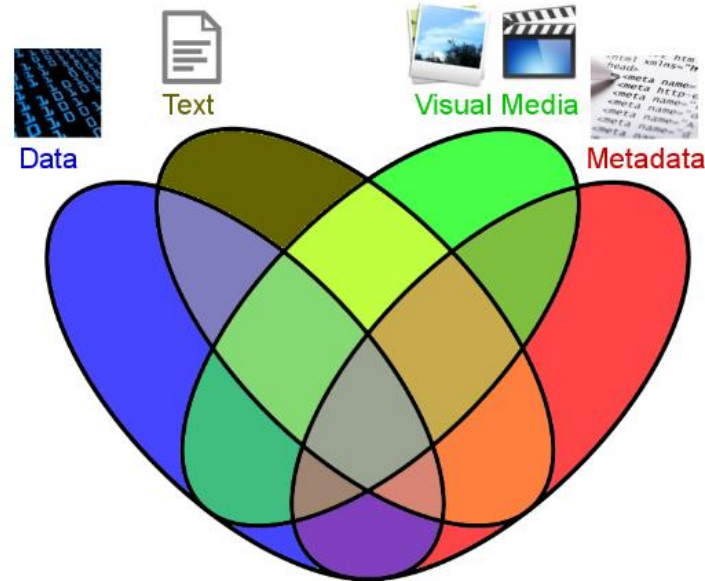
**TS. Nguyễn Ngọc Thảo – ThS. Lê Ngọc Thành**  
Bộ môn Khoa học Máy tính, FIT HCMUS, VNUHCM

Thành phố Hồ Chí Minh, 02/2019

# Nội dung bài giảng

---

- Giới thiệu Khai thác nội dung Web
- Tiền xử lý tài liệu Web
  - Tiền xử lý văn bản
  - Tiền xử lý trang HTML
  - Phát hiện trùng
- Chỉ mục đảo và nén chỉ mục
  - Tìm kiếm với chỉ mục đảo
  - Xây dựng chỉ mục và nén chỉ mục
- Chỉ mục ngữ nghĩa tiềm ẩn



# Khai thác nội dung Web

# Khai thác nội dung Web

---

- Rút trích tri thức hữu ích từ nội dung của các tài liệu Web
- Dữ liệu Web đa dạng về mặt hình thức
  - Dữ liệu có cấu trúc (danh sách, bảng tính,...), bán cấu trúc (tài liệu HTML, XML,...), văn bản phi cấu trúc
  - Dữ liệu đa phương tiện (ảnh, âm thanh, video)
- Bài giảng tập trung vào bài toán **khai thác nội dung văn bản** từ tài liệu Web.
  - Kỹ thuật truy vấn thông tin (mô hình truy vấn, đánh giá mô hình, v.v.)
  - Xử lý ngôn ngữ tự nhiên (chuyển thể ngôn ngữ, định từ loại, v.v.).

# Ứng dụng của Khai thác nội dung Web

- Nhận diện chủ đề của tài liệu Web → phân loại/gom nhóm trang Web theo chủ đề
  - Xác định danh sách từ phổ biến trong nội dung chính của trang Web

The image shows a screenshot of the CNN International website. The top navigation bar includes links for 'Edition: International', 'U.S.', 'Mexico', and 'Arabic'. Below this, there are links for 'Home', 'Video', 'World', 'U.S.', 'Africa', 'Asia', 'Europe', 'Latin America', 'Middle East', 'Business', 'World Sport', 'Entertainment', 'Tech', and 'Travel'. The main content area features a large article titled 'Beyond Seoul: 19 reasons to explore Korea' with a sub-headline 'Where to find a ginseng spa, film a Korean drama and have coffee in a cave.' Below this, there are several smaller articles and a 'More Business' section. The 'More Business' section includes articles like 'Is oil the new Greece?', 'Riding in Berlin's driverless car', and 'Deutsche Börse to sue over NYSE block'. There is also a 'FedEx Express' advertisement and a 'Analysis' section with the headline 'Goldman Sachs uproar is off-base'.



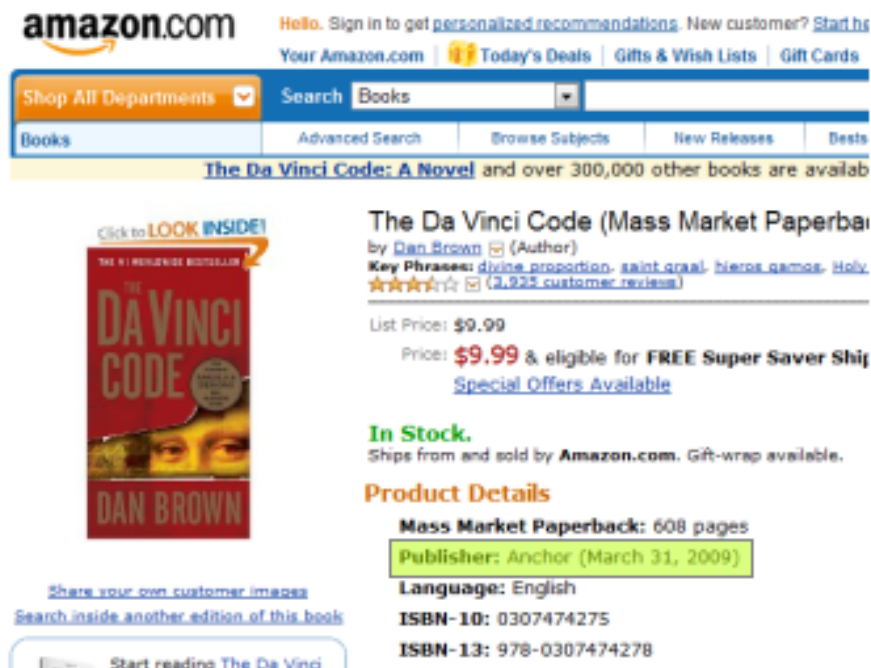
- Phát hiện trang sao chép (hợp lệ hoặc không hợp lệ)



# Ứng dụng của Khai thác nội dung Web

- Truy vấn thông tin: người dùng cung cấp truy vấn, hệ thống trả về thông tin liên quan trong trang Web
  - Hỗ trợ tìm kiếm nội bộ trong website

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 | Bests

[The Da Vinci Code: A Novel](#) and over 300,000 other books are available

Click to **LOOK INSIDE!**

**THE DA VINCI CODE**

DAN BROWN

[Share your own customer images](#)

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

[Start reading The Da Vinci](#)

**The Da Vinci Code (Mass Market Paperback)**  
by [Dan Brown](#) (Author)  
**Key Phrases:** [divine proportion](#), [saint graal](#), [hieros gamos](#), [Holy](#)  
★★★★☆ (2,922 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.

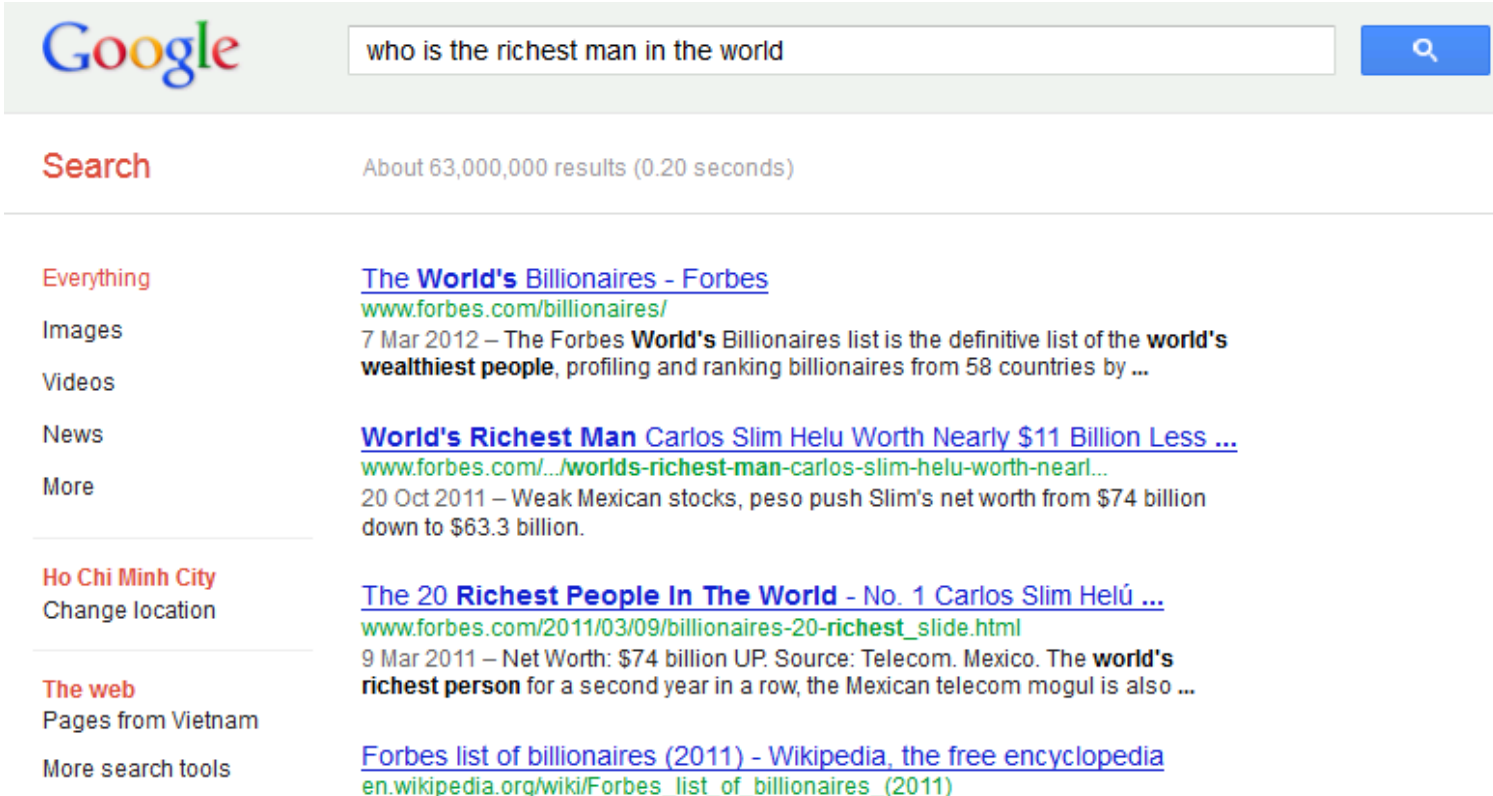
**Product Details**

**Mass Market Paperback:** 608 pages  
**Publisher:** Anchor (March 31, 2009)  
**Language:** English  
**ISBN-10:** 0307474275  
**ISBN-13:** 978-0307474278

Publisher: Anchor (March 31, 2009)

# Ứng dụng của Khai thác nội dung Web

- Truy vấn thông tin: người dùng cung cấp truy vấn, cỗ máy tìm kiếm trả về danh sách xếp hạng trang Web có liên quan



The screenshot shows a Google search interface. The search bar contains the text "who is the richest man in the world". 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 are links for "Everything", "Images", "Videos", "News", and "More". Below these, there is a section for "Ho Chi Minh City" with a link to "Change location". Further down, there is a section for "The web" with links for "Pages from Vietnam" and "More search tools". The main search results are listed on the right. The first result is "The World's Billionaires - Forbes" with a link to "www.forbes.com/billionaires/". The second result is "World's Richest Man Carlos Slim Helu Worth Nearly \$11 Billion Less ..." with a link to "www.forbes.com/.../worlds-richest-man-carlos-slim-helu-worth-nearl...". The third result is "The 20 Richest People In The World - No. 1 Carlos Slim Helú ..." with a link to "www.forbes.com/2011/03/09/billionaires-20-richest\_slide.html". The fourth result is "Forbes list of billionaires (2011) - Wikipedia, the free encyclopedia" with a link to "en.wikipedia.org/wiki/Forbes\_list\_of\_billionaires\_(2011)".

Google

who is the richest man in the world

Search

About 63,000,000 results (0.20 seconds)

Everything

Images

Videos

News

More

Ho Chi Minh City

Change location

The web

Pages from Vietnam

More search tools

[The World's Billionaires - Forbes](#)  
[www.forbes.com/billionaires/](http://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...](http://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](http://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))



# Ứng dụng của Khai thác nội dung Web

- Kết quả khai thác nội dung Web là đầu vào cho bài toán khai thác dữ liệu khác.
  - Ví dụ, thu thập nhận xét về sản phẩm (từ trang bán hàng, blog, forum,...) để phân tích sản phẩm, tìm hiểu sở thích người dùng, v.v.



## ☆☆☆☆☆ **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?

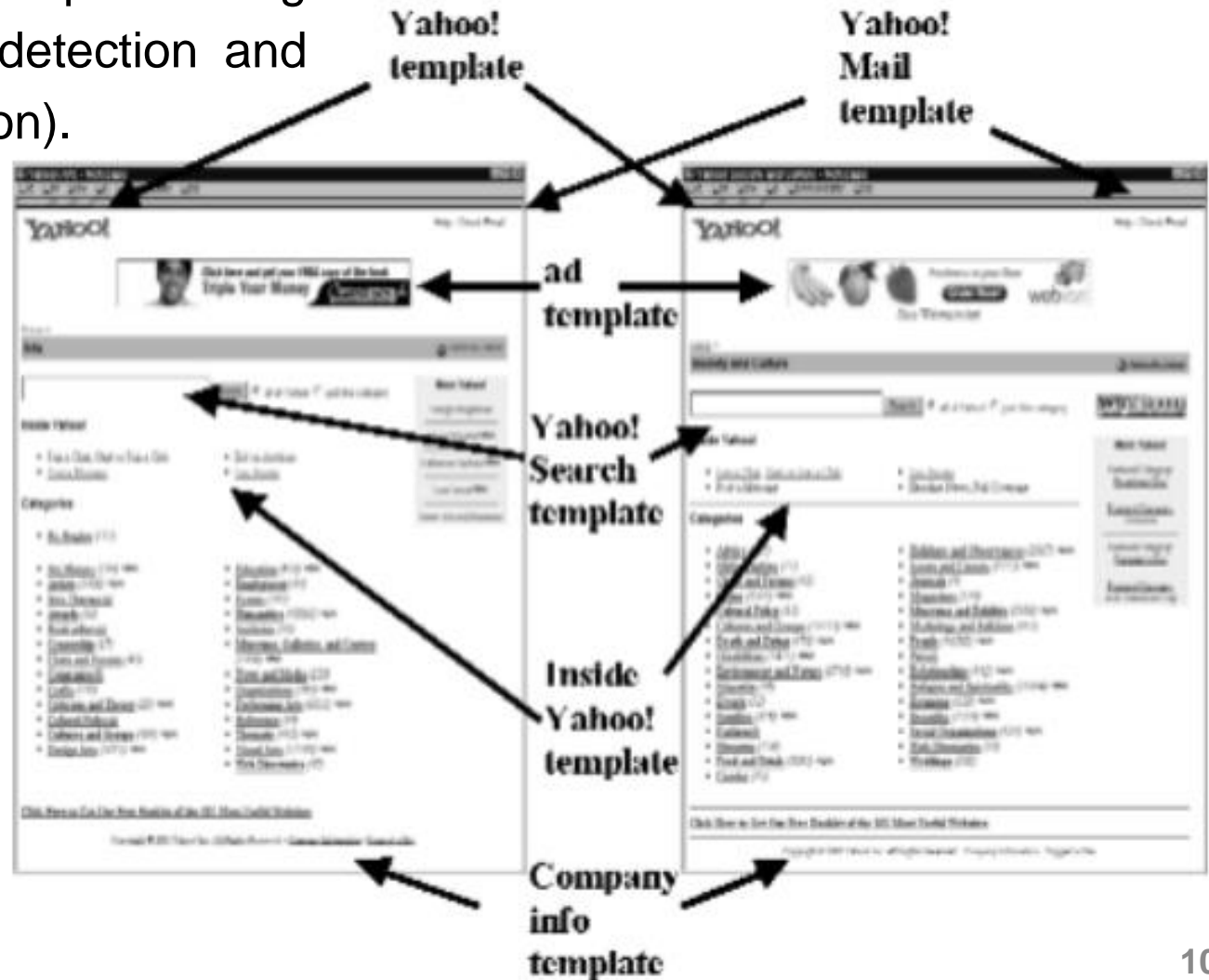
I want to buy it, but I need to know it first.

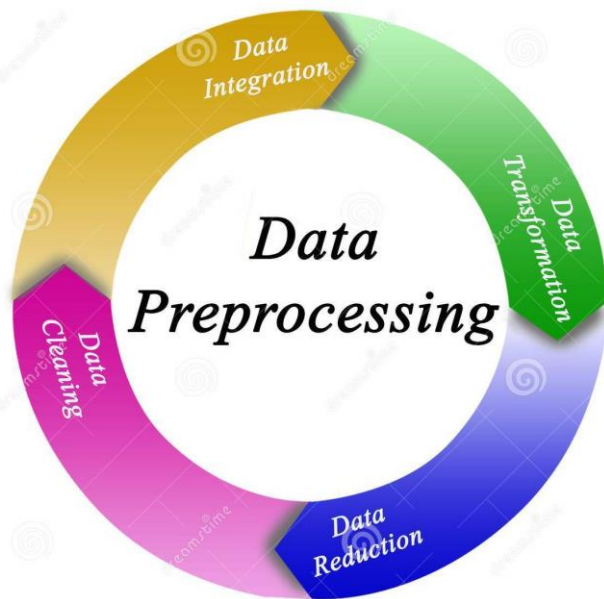
Published 1 month ago by Renan

Từ thể hiện  
cảm nghĩ của  
người dùng  
hoặc đặc tính  
của sản phẩm

# Ứng dụng của Khai thác nội dung Web

- Phát hiện mẫu và phân vùng trang (Template detection and Page segmentation).





---

# Tiền xử lý tài liệu Web

---

# Tiền xử lý văn bản

---

- Văn bản truyền thống không chứa các HTML tag.
- Rút trích từ (word extraction): đơn giản
  - Ví dụ, I am a student.  $\Rightarrow$  {i, am, a, student}.
  - Tiếng Anh: không cần thực hiện, tiếng Việt: đạt kết quả cao ~99%
- Loại bỏ stopwords
  - Từ xuất hiện thường xuyên trong câu nhưng không quan trọng (a, an, the, will, with, v.v.)
- Stemming
  - Chuyển biến thể của từ về hình thái gốc (going  $\rightarrow$  go, went  $\rightarrow$  go)
- Xử lý chữ số, dấu nối, dấu câu (. , ; ...) và chữ hoa/thường

# Loại bỏ stopwords

- **Stopword** là từ xuất hiện thường xuyên và không quan trọng trong một ngôn ngữ, giúp xây dựng câu nhưng không biểu đạt nội dung của tài liệu.

- ~ 400 – 500 từ loại article, preposition, conjunction và pronouns

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 cụ thể theo ứng dụng.
  - Độ quan trọng của từ phụ thuộc ngữ cảnh.
  - Ví dụ, appear, best, example được xem là stopwords chỉ trong một số trường hợp.



# Loại bỏ stopwords

---

- Stopword cần được loại bỏ trước khi đánh chỉ mục và lưu trữ tài liệu.
- Stopword trong truy vấn cũng cần được loại bỏ.
- *Tại sao 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 của truy vấn
    - Stopword gây nhiễu cho các hệ thống truy vấn

# Stemming

---

- Từ có nhiều thể cú pháp khác nhau tùy thuộc vào ngữ cảnh sử dụng từ.
  - Ví dụ, thể số nhiều của danh từ (apple – apples), thể tiếp diễn của động từ (eat – eating), và chia thì động từ (eat – ate – eaten).
- Những biến thể này khiến cho hệ thống truy vấn có độ phủ (recall) thấp.
- **Stemming** là quá trình biến đổi một từ về thể gốc.
  - **Stem** (root) 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”

# Giải thuật Stemming

- Có nhiều thuật toán Stemming, gọi là bộ **stemmers**
- **Giải thuật Martin Porter's:** sử dụng bộ luật cú pháp được định nghĩa trước

Ví dụ Step 1

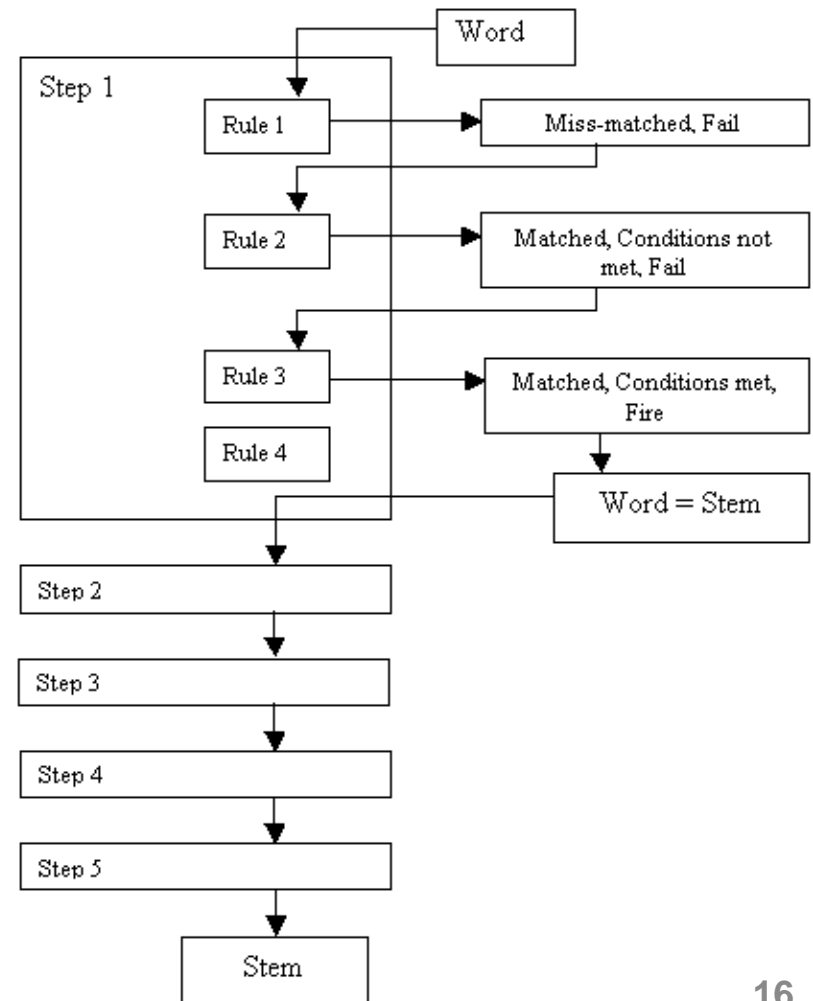
**SS**ES → **SS**

caresses → caress

**I**ES → **I**

ponies → poni

ties → ti



# Stemming

---

- Stemming làm giảm kích thước cấu trúc chỉ mục.
  - Kết hợp từ cùng gốc có thể giảm 40-50% kích thước chỉ mục.
- Stemming làm tăng độ phủ (recall) của hệ thống truy vấn nhưng có thể làm giảm độ chính xác (precision).
  - Truy vấn xét nhiều biến thể của từ cùng lúc khiến cho nhiều tài liệu không liên quan (ngữ nghĩa) bị cho là liên quan.
  - Ví dụ, “cop” and “cope” → “cop”, tài liệu chỉ chứa “cope” không liên quan đến police.

# Các tác vụ tiền xử lý văn bản khác

## • Ký tự số

- Thường được loại bỏ trong hệ thống truy vấn thông tin truyền thống
- Vẫn được giữ lại khi đánh chỉ mục cho cỗ máy tìm kiếm.
- Ngoại lệ: ngày tháng, thời gian, mẫu regular expression định nghĩa trước

- Welcome
- Introduction
- Overview:
  1. Basics
  2. Theory
  3. Practice
  4. Advanced topics:
    - Dissertations
    - Police reports
- Test
- Feedback|

## • Ký tự hoa/thường

- Đồng loạt chuyển thành một thể
- Ví dụ, từ “CapTalizE” chuyển thành “capitalize” hay “CAPITALIZE”



# Các tác vụ tiền xử lý văn bản khác

- **Dấu nối** được loại bỏ theo một số luật tổng quát nhưng vẫn cần xét ngoại lệ.

- Hai hình thức loại bỏ dấu nối thường gặp
  - state-of-the-art → state of the art
  - state-of-the-art → stateoftheart

- Cả hai hình thức đều được duy trì trong chỉ mục khi khó xác định cái nào hoàn toàn đúng (ví dụ, pre-processing và preprocessing)
- Ngoại lệ: dấu nối có thể tích hợp vào trong từ, ví dụ, “Y-21”
- **Dấu câu** được xử lý tương tự như dấu nối.

# Bài tập 1: Tiền xử lý văn bản

---

- Cho ba tài liệu  $d_1$ ,  $d_2$  và  $d_3$  như bên dưới.
- $d_1$ : An information retrieval model governs how a document and a query are represented and how the relevance of a document to a user query is defined.
- $d_2$ : Information retrieval is the area of study concerned with searching for documents, for information within documents, and for metadata about documents, as well as that of searching structured storage, relational databases, and the World Wide Web.
- $d_3$ : Web search has become very important in the information age. Increased exposure of pages on the Web can result in significant financial gains and/or fames for organizations and individuals.

# Bài tập 1: Tiền xử lý văn bản

- Loại bỏ stopwords theo danh sách stopwords được cung cấp

how, and, or, an, a, the, there, that, of, for, to, is, are, can, has, with, within, in, on, about, as, well, very

- Stemming theo các luật như sau
  - Chuyển danh từ thành thể số ít
    - E.g., books → book, classes → class, ponies → pony
  - Chuyển động từ về thể nguyên bản
    - E.g., running → run, used → use, written → write
- Viết thường, không xét ký tự ngoài chữ cái
- Từ được sắp xếp theo thứ tự từ điển

# Tiền xử lý trang Web

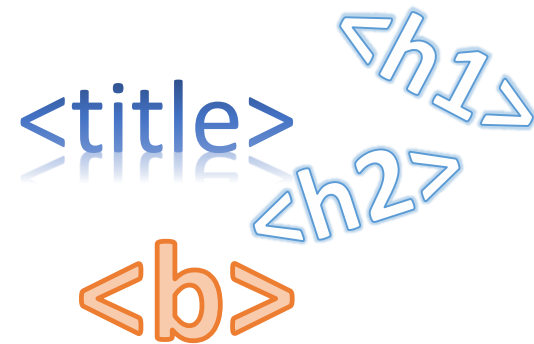
---

- Tài liệu Web cần thêm các tác vụ tiền xử lý khác, bên cạnh những thao tác đã áp dụng cho văn bản.
- Nhận diện trường văn bản
  - Ví dụ, văn bản trong title, metadata, body, v.v.
- Nhận diện anchor text
  - Ví dụ: [YouTube](#) là trang Web chia sẻ video clip
- Loại bỏ HTML tags
- Nhận diện khối nội dung chính

# Nhận diện trường văn bản

---

- Trang HTML thường chứa nhiều trường văn bản khác nhau.
  - Ví dụ, văn bản trong title, metadata, body, v.v.



- Cổ máy tìm kiếm đặt độ ưu tiên khác nhau cho từ ở những trường khác nhau.
  - Ví dụ, từ trong trường title hay từ được nhấn mạnh (thuộc header tags <h1>, <h2>, ..., bold tag <b>, v.v.) có ý nghĩa quan trọng hơn.



# Nhận diện anchor text

---

- **Anchor text** là đoạn văn bản kết hợp với hyperlink, khi người dùng nhấp vào sẽ di chuyển tới một trang/site khác.
  - Chứa mô tả chính xác về thông tin của trang được trở tới
  - Ví dụ, a hyperlink to the [English-language Wikipedia's homepage](#)
- Anchor text của hyperlink nói các trang có giá trị đặc biệt.
  - Thông tin do nguồn khác cung cấp thay vì chủ của trang Web → đáng tin cậy hơn.

# Loại bỏ HTML tags

- Thực hiện tương tự như đối với dấu ngắt.
- Loại bỏ HTML tags có thể ảnh hưởng đến loại truy vấn theo cụm hoặc theo tính lân cận.
  - Không được nối văn bản của những khối khác nhau

Cụm từ “cite this article” ở cuối cột bên trái sẽ bị nối (không mong muốn) với cụm từ “Main Page” ở khối bên phải.



# Nhận diện khối nội dung chính

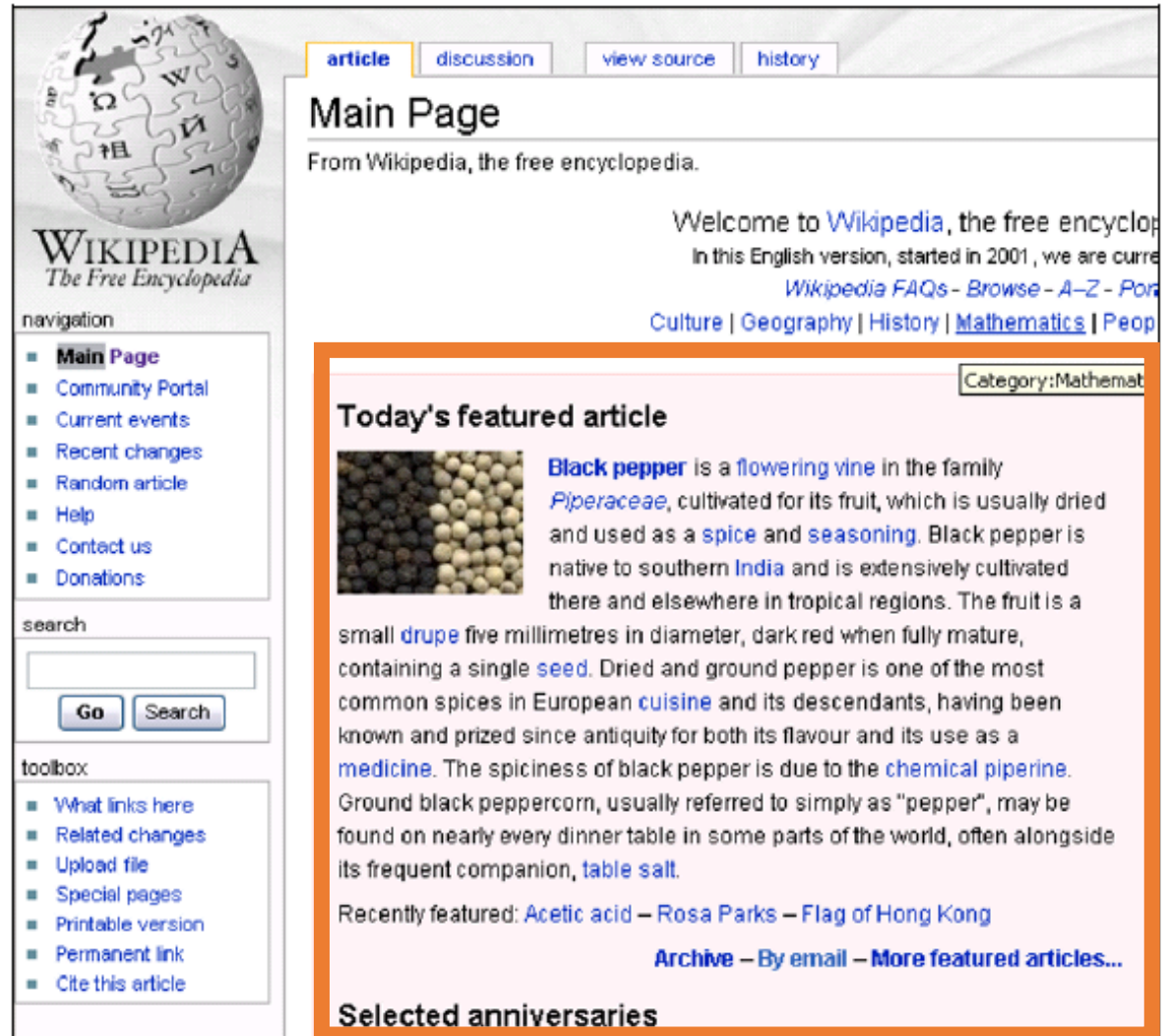
---

- Một lượng lớn thông tin trong trang Web không phải là nội dung chính (banner ads, navigation, copyright notices, v.v.)
- Kết quả tìm kiếm và khai thác dữ liệu có thể được cải thiện đáng kể nếu chỉ xét khối nội dung chính
- Hai kỹ thuật nhận diện khối nội dung chính
  - Phân vùng theo dấu hiệu thị giác (partitioning based on visual cues)
  - So khớp cây (tree matching)

# Nhận diện khối nội dung chính

Khối nội dung chính là khối có chứa “Today’s featured article.”

Không nên xét anchor text của những navigation link vào nội dung của trang.



The screenshot shows the Wikipedia Main Page. At the top left is the Wikipedia logo, a globe made of puzzle pieces with various characters and symbols. Below it is the text "WIKIPEDIA The Free Encyclopedia". To the right of the logo are navigation links: "article", "discussion", "view source", and "history". Below the logo is a "navigation" section with links: "Main Page", "Community Portal", "Current events", "Recent changes", "Random article", "Help", "Contact us", and "Donations". Below the navigation section is a "search" section with a text input field and "Go" and "Search" buttons. Below the search section is a "toolbox" section with links: "What links here", "Related changes", "Upload file", "Special pages", "Printable version", "Permanent link", and "Cite this article". The main content area is titled "Main Page" and includes the text "From Wikipedia, the free encyclopedia." Below this is a welcome message: "Welcome to Wikipedia, the free encyclopedia. In this English version, started in 2001, we are currently..." followed by links: "Wikipedia FAQs - Browse - A-Z - Portal" and "Culture | Geography | History | Mathematics | People". The "Today's featured article" section is highlighted with an orange border. It features a small image of black and white peppercorns. The text describes black pepper as a flowering vine in the family Piperaceae, cultivated for its fruit, which is usually dried and used as a spice and seasoning. It mentions that 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. Below the article text are links: "Recently featured: Acetic acid - Rosa Parks - Flag of Hong Kong" and "Archive - By email - More featured articles...". At the bottom of the featured article section is the heading "Selected anniversaries".

# Phân vùng theo dấu hiệu thị giác

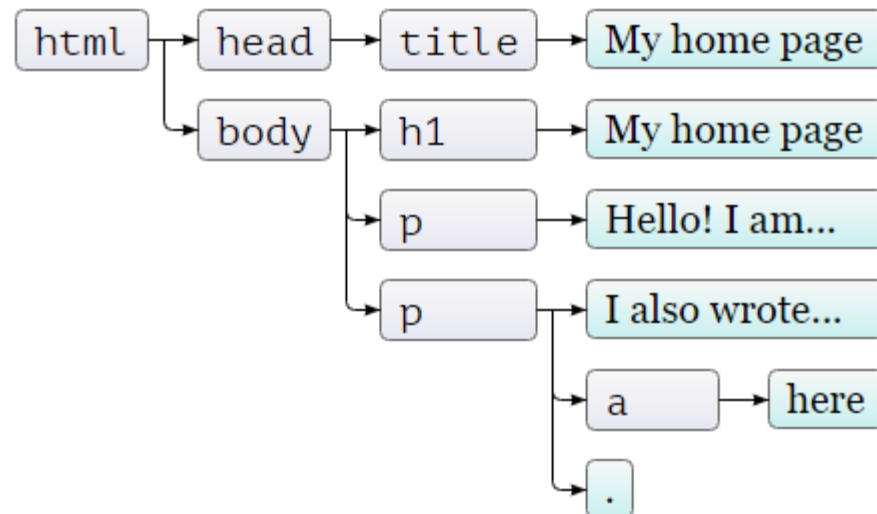
---

- Sử dụng thông tin thị giác của mỗi HTML element trong trang để tìm khối nội dung chính.
- Thông tin rendering của HTML element từ trình duyệt Web
  - Ví dụ, Internet Explorer hỗ trợ API xuất ra tọa độ (X, Y) của element
- Mô hình học máy sử dụng đặc trưng tọa độ và biểu hiện
  - Cần gán nhãn thủ công cho một lượng lớn mẫu huấn luyện.



# So khớp cây

- Hầu hết trang Web thương mại được phát sinh từ một số template cố định.



- Văn bản trong khối nội dung chính thường khác xa nhau trong những trang có cùng template.
  - Trong khi các nội dung “phụ” thường rất giống nhau qua nhiều trang.
- Đánh giá độ tương tự về nội dung giữa hai trang Web bằng **phương pháp Shingle**.

# Phát hiện trùng lặp

---

- Trùng lặp văn bản là **vấn đề nghiêm trọng đối với Web**
  - Không phải là vấn đề quan tâm trong hệ thống truy vấn truyền thống
- Phát hiện trùng lặp giúp giảm kích thước chỉ mục và cải thiện kết quả tìm kiếm.
- Các thể loại trùng lặp trang và nội dung trên Web bao gồm
  - **Duplication** (hoặc **replication**): sao chép một trang
  - **Mirroring**: sao chép toàn bộ site

# Phát hiện trùng lặp

---

- Trang trùng lặp giúp tăng hiệu suất tìm kiếm và tải tài liệu trên phạm vi toàn cầu.
  - Giới hạn băng thông giữa các vùng địa lý, mạng có chất lượng xấu hoặc gặp sự cố
- Bên cạnh đó, một số bản sao được tạo ra với mục đích xấu
  - Lừa đảo người dùng, vi phạm bản quyền.

# Phát hiện trùng: Ví dụ

Secure | [https://www.google.com.hk/#q=donald+trump&\\*>](https://www.google.com.hk/#q=donald+trump&*>)


gle donald trump

Gmail Images Sign


All Images News Videos Maps More Settings Tools

About 454,000,000 results (0.72 seconds)


Top stories





George W. Bush opens up on Trump's war with the media, Russia and travel



Trump to propose \$54 billion in cuts to 'most federal agencies'



With Donald Trump scaring allies, Australia has never been so popular



Donald Trump  
45th U.S. President


More Images

gle donald trump


All Images News Videos Books More Settings Tools

About 453,000,000 results (0.80 seconds)


Top stories





The infuriating silence of Donald Trump over an Indian engineer's murder in Kansas



George W. Bush opens up on Trump's war with the media, Russia and travel ban



Donald Trump promises 'historic' increase in US military budget



Donald Trump  
45th U.S. President

More Images

donaldjtrump.com

Google tạo mirror site đặc trưng cho từng địa phương

# Phát hiện trùng: Ví dụ



**Fake Facebook URL:**  
**www.facelook.cixx6.com**

**Facebook Login**

You must log in to see this page.

Email address:

Password:

☐ Keep me logged in

or [Sign up for Facebook](#)

[Forgotten your password?](#)

English (US) Español Português (Brasil) Français (France) Deutsch Italiano العربية हिन्दी 中文(简体)  
日本語 »

Trang đăng nhập Facebook login giả mạo có giao diện giống hệt, chỉ khác URL

# Băm toàn bộ văn bản

---

- Giải pháp đơn giản nhất, sử dụng giải thuật MD5 hay tính toán con số tích hợp (ví dụ, checksum)
- Hữu ích đối với phát hiện trang trùng lặp hoàn toàn → không phổ biến do trùng hoàn toàn hiếm gặp trên Web
  - Ngay cả các mirror site cũng khác nhau về URL, Web master, thông tin liên lạc, và mục quảng cáo,...để phù hợp với nhu cầu cục bộ.

# Phương pháp Shingle

---

- Kỹ thuật phát hiện trùng hiệu quả bằng  $n$ -gram
- Một  **$n$ -gram** (hay **shingle**) là chuỗi liên tiếp các từ trong một cửa sổ kích thước cố định  $n$ .
  - Ví dụ, “John went to school with his brother,” được biểu diễn thành 5 cụm từ 3-gram “John went to”, “went to school”, “to school with”, “school with his”, và “with his brother”.

# Phương pháp Shingle

- Gọi  $S_n(\mathbf{d})$  là tập hợp các  $n$ -gram phân biệt trong tài liệu  $\mathbf{d}$ .
  - Mỗi  $n$ -gram có thể được mã hóa bằng một số hay giá trị băm MD5.
- Gọi  $S_n(\mathbf{d}_1)$  và  $S_n(\mathbf{d}_2)$  là tập hợp  $n$ -gram biểu diễn  $\mathbf{d}_1$  và  $\mathbf{d}_2$ .
- **Hệ số Jaccard** ước lượng độ tương tự giữa hai tài liệu

$$\text{sim}(\mathbf{d}_1, \mathbf{d}_2) = \frac{|S_n(\mathbf{d}_1) \cap S_n(\mathbf{d}_2)|}{|S_n(\mathbf{d}_1) \cup S_n(\mathbf{d}_2)|}$$

- So sánh  $\text{sim}(\mathbf{d}_1, \mathbf{d}_2)$  với ngưỡng để xác định khả năng  $\mathbf{d}_1$  và  $\mathbf{d}_2$  là bản sao của nhau.
- Kích thước cửa sổ  $n$  và ngưỡng tương tự được xác định theo thực nghiệm.



# Phương pháp Shingle: Ví dụ

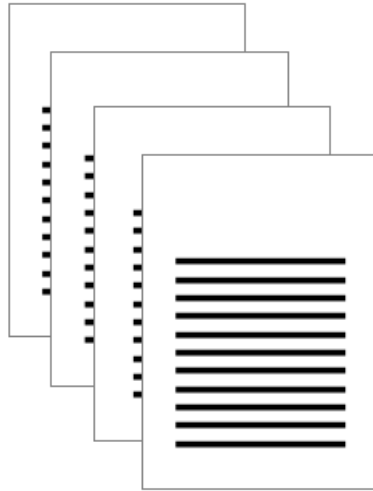
- Cho ba tài liệu  $\mathbf{d}_1$ ,  $\mathbf{d}_2$  và  $\mathbf{d}_3$  như bên dưới
  - $\mathbf{d}_1$ : Jack London traveled to Oakland
  - $\mathbf{d}_2$ : Jack London traveled to the city of Oakland
  - $\mathbf{d}_3$ : Jack traveled from Oakland to London
- Các tài liệu sau khi tiền xử lý sẽ chứa các từ khóa
  - $\mathbf{d}_1$ : jack, london, travel, oakland
  - $\mathbf{d}_2$ : jack, london, travel, citi, oakland
  - $\mathbf{d}_3$ : jack, travel, oakland, london
- Do đó,  $sim(\mathbf{d}_1, \mathbf{d}_2) = \frac{|S_2(\mathbf{d}_1) \cap S_2(\mathbf{d}_2)|}{|S_2(\mathbf{d}_1) \cup S_2(\mathbf{d}_2)|} = \frac{2}{5} = 0.4$

# Bài tập 2: Phương pháp Shingle

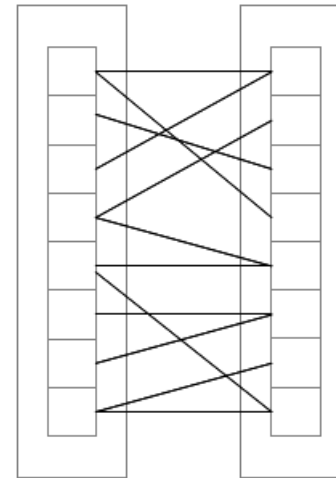
---

- Cho hai tài liệu  $\mathbf{d}_1$  và  $\mathbf{d}_2$  như bên dưới
  - $\mathbf{d}_1$ : I do not like green eggs and ham.
  - $\mathbf{d}_2$ : I do not like them, Sam I am.
- Hãy xác định  $S_n(\mathbf{d}_1)$  và  $S_n(\mathbf{d}_2)$  với  $n = 2$  (2-grams)
- Áp dụng hệ số Jaccard để tính độ tương tự giữa hai tài liệu,  $sim(\mathbf{d}_1, \mathbf{d}_2)$

DOCUMENT  
COLLECTION



INDEX



docs

words

# Chỉ mục đảo

# Tìm kiếm với chỉ mục đảo

---

- Tìm các tài liệu chứa từ trong truy vấn của người dùng
  - Phương pháp cơ bản của tìm kiếm Web và truy vấn thông tin
- Duyệt tài liệu tuần tự hiển nhiên không khả thi đối với tập tài liệu lớn như Web → sử dụng chỉ mục
- **Chỉ mục đảo** (inverted index) là lựa chọn phổ biến nhất trong cỗ máy tìm kiếm, tốt hơn hầu hết các chiến lược khác.
  - Hiệu suất truy vấn tốt, xây dựng nhanh chóng

# Chỉ mục đảo

---

- Chỉ mục đảo của một tập tài liệu là cấu trúc dữ liệu liên kết mỗi từ với một danh sách tài liệu chứa từ đó.
  - Việc truy vấn tài liệu chứa từ cần tìm tốn thời gian hằng, có thể tìm tài liệu chứa nhiều từ khóa cùng lúc.
- Cho tập tài liệu  $D = \{d_1, d_2, \dots, d_N\}$ , mỗi tài liệu có định danh (ID) duy nhất.
- Một chỉ mục đảo gồm có hai thành phần
  - Tập ngữ vựng  $V$  chứa mọi từ phân biệt trong  $D$ , và
  - Với mỗi từ  $t_i$ , lưu một danh sách đảo gồm các posting.

# Chỉ mục đảo: Postings

- Mỗi **posting** lưu ID (kí hiệu  $id_j$ ) của tài liệu  $d_j$  chứa từ  $t_i$  và những thông tin khác về  $t_i$  trong  $d_j$ .
  - Tùy theo thuật toán xếp hạng hay truy vấn nào được sử dụng, posting có thể tích hợp thêm nhiều loại thông tin.
  - Ví dụ, để hỗ trợ truy vấn kiểu lân cận và kiểu cụm từ, một posting cho từ  $t_i$  có thể bao gồm  $\langle id_j, f_{ij}, [o_1, o_2, \dots, o_{|f_{ij}|}] \rangle$ 
    - $id_j$  là ID của tài liệu  $d_j$  chứa từ  $t_i$ ,  $f_{ij}$  là tần số xuất hiện của  $t_i$  trong  $d_j$ ,  $o_k$  là vị trí của  $t_i$  trong  $d_j$ .
- Các posting của một từ được sắp xếp tăng dần theo  $id_j$
- Tương tự cho tập offset trong mỗi posting

# Chỉ mục đảo: Ví dụ

---

- Cho ba tài liệu có định danh là  $id_1$ ,  $id_2$  và  $id_3$ .

$id_1$ : 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

- $V = \{\text{Web, mining, useful, applications, usage, structure, studies, hyperlink}\}$
- Loại bỏ stopwords: “is” và “the”. Không áp dụng stemming

# Chỉ mục đảo: Ví dụ

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)

- Chỉ mục đảo đơn giản (A): mỗi từ chỉ liên kết với danh sách đảo gồm ID của các tài liệu chứa từ.
- Chỉ mục đảo phức tạp (B) chứa thêm thông tin về tần số xuất hiện của từ và vị trí xuất hiện tương ứng trong tài liệu.



# Tìm kiếm với chỉ mục đảo

---

- Truy vấn được đánh giá bằng cách, đầu tiên tìm các danh sách đảo của từ trong truy vấn, và sau đó xử lý chúng để tìm những tài liệu chứa tất cả (hoặc một vài) từ.
- Tìm kiếm tài liệu liên quan trong chỉ mục đảo gồm ba bước chính như sau
  1. Tìm ngữ vựng (vocabulary search)
  2. Trộn kết quả (results merging)
  3. Tính điểm xếp hạng (ranking score computation)

# Bước 1: Tìm ngữ vựng

---

- Tìm từ truy vấn trong tập ngữ vựng để xác định danh sách đảo tương ứng
- Giải pháp tăng tốc tìm kiếm
  - Tập ngữ vựng nằm trên bộ nhớ chính, nhiều phương pháp đánh chỉ mục khác nhau (hashing hay B-tree) được sử dụng
  - Thứ tự từ điển cũng có thể được sử dụng do tính hiệu quả về mặt không gian, sau đó áp dụng tìm kiếm nhị phân.
    - Độ phức tạp là  $O(\log|V|)$ , với  $|V|$  là kích thước ngữ vựng.
- Truy vấn đơn từ: đã tìm thấy mọi tài liệu liên quan, sang bước 3
- Truy vấn đa từ: sang bước 2

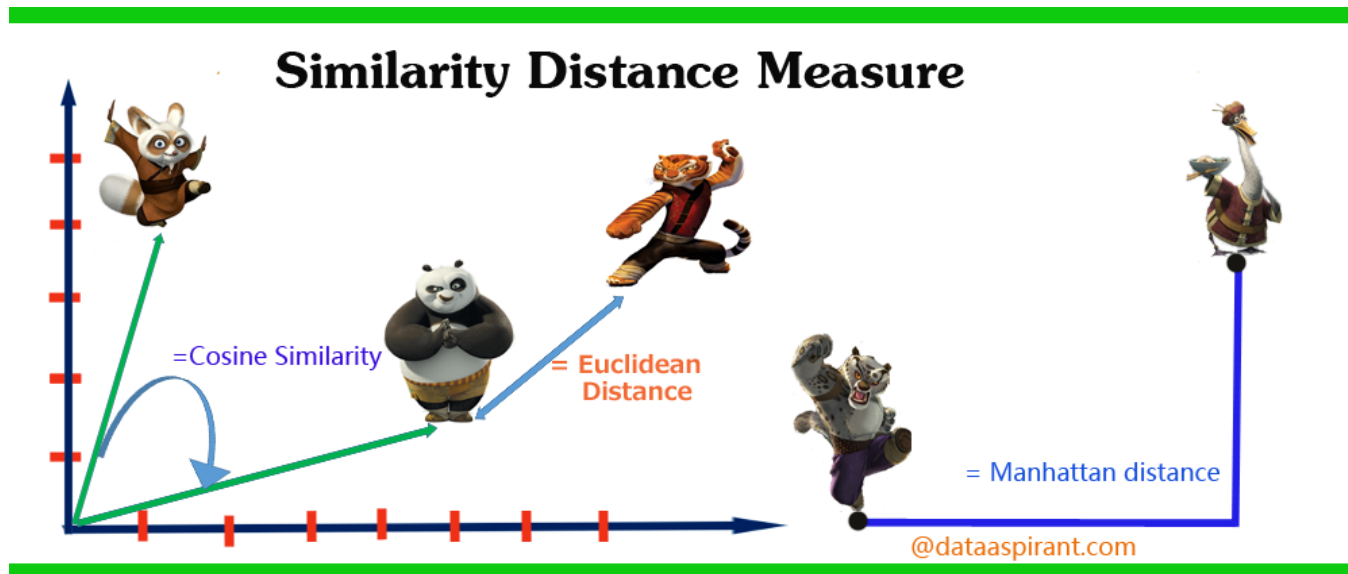
# Bước 2: Trộn kết quả

---

- Trộn các danh sách đảo tìm được từ Bước 1 để xác định phần giao
  - Heuristic để tăng tốc: dùng danh sách ngắn nhất làm cơ sở để trộn với những danh sách khác dài hơn; áp dụng tìm kiếm nhị phân để tìm posting của danh sách ngắn trên danh sách dài hơn.
- Tìm kiếm từng phần có thể thực hiện theo cách tương tự.
- Toàn bộ chỉ mục đảo thường không vừa bộ nhớ trong, do đó chỉ một phần chỉ mục được tải vào.
  - Phân tích nhật ký truy vấn để tìm những từ truy vấn phổ biến → danh sách đảo của những từ này sẽ được lưu trong bộ nhớ.

# Bước 3: Tính điểm xếp hạng

- Tính điểm xếp hạng (hay độ tương quan) cho mỗi tài liệu dựa theo hàm liên quan (okapi hay cosine)
- Có thể xét thêm tính cụm và tính lân cận của từ
- Điểm số này được dùng để xếp hạng kết quả trả về



# Tìm với chỉ mục đảo: Ví dụ

- Xét chỉ mục đảo phức tạp
  - 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:  $\langle id_1, 1, [1] \rangle, \langle id_3, 2, [1, 6] \rangle$
- Truy vấn: “**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 và tìm tài liệu chứa cả hai từ  $\rightarrow id_1$  và  $id_3$ .
- Bước 3:  $id_1$  được xếp hạng cao hơn  $id_3$  do tính gần và thứ tự của các từ

# Bài tập 3: Chỉ mục đảo

---

- Cho 4 tài liệu  $d_1$ ,  $d_2$ ,  $d_3$  và  $d_4$  như bên dưới
  - $d_1$ : breakthrough drug for schizophrenia
  - $d_2$ : new schizophrenia drug
  - $d_3$ : new approach for treatment of schizophrenia
  - $d_4$ : new hopes for schizophrenia patients
- Vẽ biểu diễn chỉ mục đảo cho các tài liệu trên
- Hệ thống trả về kết quả gì cho những truy vấn dưới đây?
  - schizophrenia AND drug
  - for AND NOT(drug OR approach)

# 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 để xây dựng chỉ mục là  $O(T)$ , với  $T$  là số lượng từ (kể cả từ trùng) trong tập tài liệu đã tiền xử lý.
  - Lựa chọn khác: in-memory hash table, hoặc CTDL tương đương
- Duyệt tuần tự từng tài liệu, với mỗi từ trong tài liệu, tìm từ đó trong trie
  - Nếu tìm thấy, thêm ID và thông tin khác (ví dụ, offset của từ) vào danh sách đảo của từ
  - Nếu không tìm thấy, tạo nút lá mới biểu diễn từ

# Ví dụ chỉ mục đảo bằng cây trie

$id_1$ : Web mining is useful.

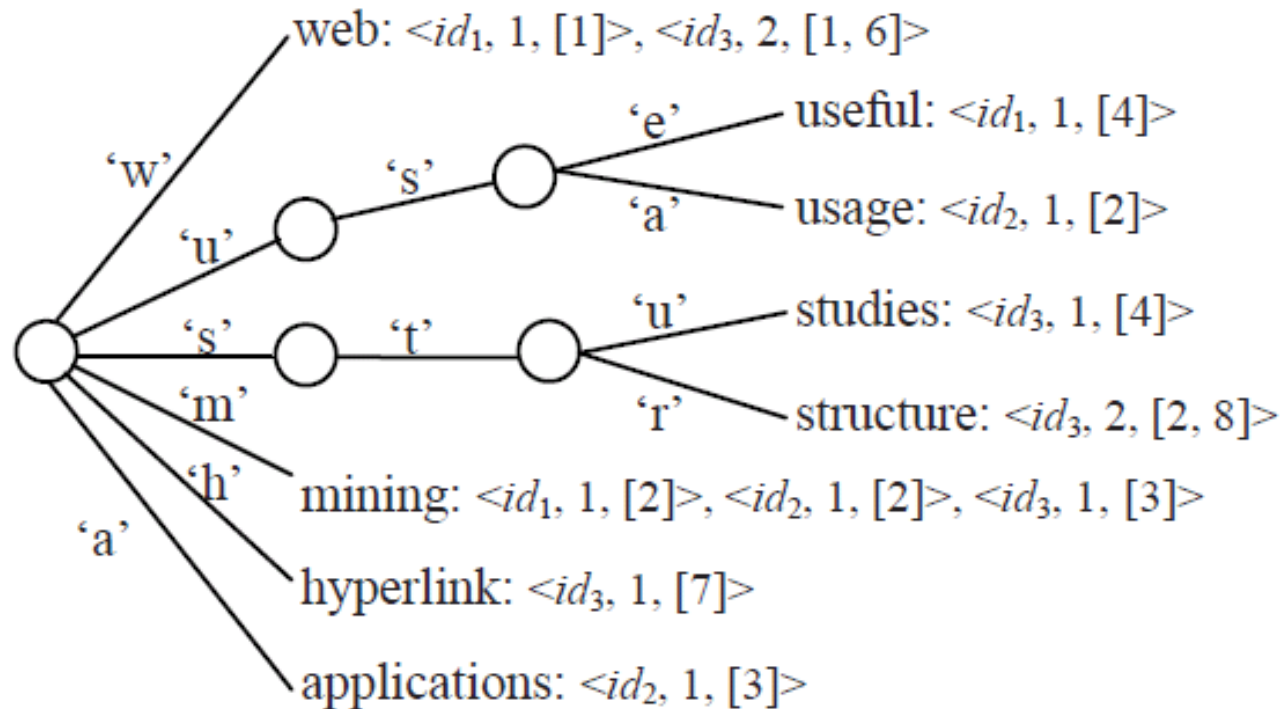
1 2 3 4

$id_2$ : 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

---

- Trie cần được lưu trong bộ nhớ để xây dựng chỉ mục một cách hiệu quả → không khả thi trong ngữ cảnh Web
- Xây dựng chỉ mục (một phần)  $I_1$  đến khi đầy bộ nhớ, và ghi  $I_1$  ra đĩa.
- Thực hiện tương tự cho  $k - 1$  chỉ mục khác,  $I_2, \dots, I_k$
- Trộn theo cặp:  $I_1 + I_2 \rightarrow I_{1-2}$ ,  $I_3 + I_4 \rightarrow I_{3-4}, \dots$ , và rồi đến  $I_{1-2} + I_{3-4}$ ,  $I_{5-6} + I_{7-8}$ , v.v. cho đến khi mọi chỉ mục một phần được trộn thành một chỉ mục đơn

# Xây dựng chỉ mục đảo

---

- Độ phức tạp của mỗi lượt trộn là tuyến tính với số từ ở cả hai chỉ mục một phần.
- Mỗi mức cần một lần xử lý tuyến tính trên toàn chỉ mục
- Toàn bộ quá trình trộn cần độ phức tạp thời gian  $O(k \log_2 k)$
- Để giảm không gian đĩa cần thiết, chỉ mục một phần mới có thể được trộn với một chỉ mục vừa trộn.
  - Ví dụ,  $I_1 + I_2 \rightarrow I_{1-2}$ , một chỉ mục  $I_3$  mới tạo ra được trộn với  $I_{1-2}$  tạo thành  $I_{1-2-3}$ , cứ thế tiếp tục

# Xây dựng chỉ mục đảo

---

- Trang Web thường xuyên được thêm, chỉnh sửa hoặc xóa.
- Hiệu chỉnh trực tiếp trên chỉ mục chính làm giảm hiệu suất
  - Chỉ một trang thay đổi cũng cần cập nhật nhiều record của chỉ mục
- Giải pháp: duy trì thêm hai chỉ mục bổ trợ, một dành cho trang xóa và một dành cho trang thêm.
  - Hiệu chỉnh được xem là một lần xóa rồi một lần thêm.
  - Tìm kiếm từ truy vấn trong cả chỉ mục chính và hai chỉ mục bổ trợ.
- Kết quả trình bày cho người dùng là  $(D_0 \cup D_+) - D_-$ 
  - $D_0$  là các trang trả về từ chỉ mục chính,  $D_+$  và  $D_-$  lần lượt là các trang trả về từ chỉ mục trang thêm và từ trang xóa
- Trộn hai chỉ mục bổ trợ vào chỉ mục chính khi kích thước của chúng trở nên quá lớn.

# Nén chỉ mục

---

- Kích thước chỉ mục đảo có thể rất lớn.
- Để tăng tốc tìm kiếm, chỉ mục cần được tải vào bộ nhớ càng nhiều càng tốt để giảm thao tác đọc/ghi đĩa.
- Do đó, giảm kích thước chỉ mục là vấn đề quan trọng.
- 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: Nén số nguyên

---

- Chỉ mục đảo lưu trữ chủ yếu là các số nguyên biểu diễn ID của tài liệu và vị trí từ → 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.
- **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
- **Mô hình variable-byte**: mỗi số nguyên được lưu bằng một lượng byte tích hợp.
  - Variable-byte coding.

# Elias gamma coding

---

- Một số nguyên dương  $x$  được biểu diễn bằng  $1 + \lfloor \log_2 x \rfloor$  ký số trong hệ nhị phân
- **Mã hóa**
  - Viết  $x$  trong hệ nhị phân. Thêm vào trước  $K$  số 0 với  $K =$  số bit biểu diễn  $x$  trừ đi 1.
  - Ví dụ, 9 được biểu diễn thành 000**1001**
- **Giải mã**
  - Đọc từ stream và đếm số 0 đến khi gặp số 1 đầu tiên, gọi đó là  $K$
  - Ký tự đầu tiên của số nguyên có giá trị  $2^K$ , đọc  $K$  bits còn lại
  - Ví dụ, 000**1001** có  $K = 3$  bit 0, từ bit 1 đọc tiếp 3 bit còn lại  $\rightarrow$  **1001**  
(9)

# Variable-byte coding

---

- **Mã hóa**

- Bảy bit trong mỗi byte được dùng để mã hóa số nguyên.
- Least significant bit là 0 nếu là byte cuối cùng, hoặc là 1 nếu còn có thêm byte phía sau

- Biểu diễn số nguyên nhỏ hiệu quả hơn mô hình variable-bit.

- Ví dụ, **135** được biểu diễn bằng 2 byte vì nằm trong đoạn  $2^7$  và  $2^{14}$ , tức là 000000**11** **00001110**.

- **Giải mã**

- Đọc mọi byte cho đến khi gặp byte có bit cuối cùng bằng 0
- Bỏ least significant bit trong mỗi byte và nối chúng lại với nhau
- Ví dụ, 00000011 00001110 is decoded to 000000**10000111** (**135**)

# Nén chỉ mục bằng hiệu số ID liên kề

---

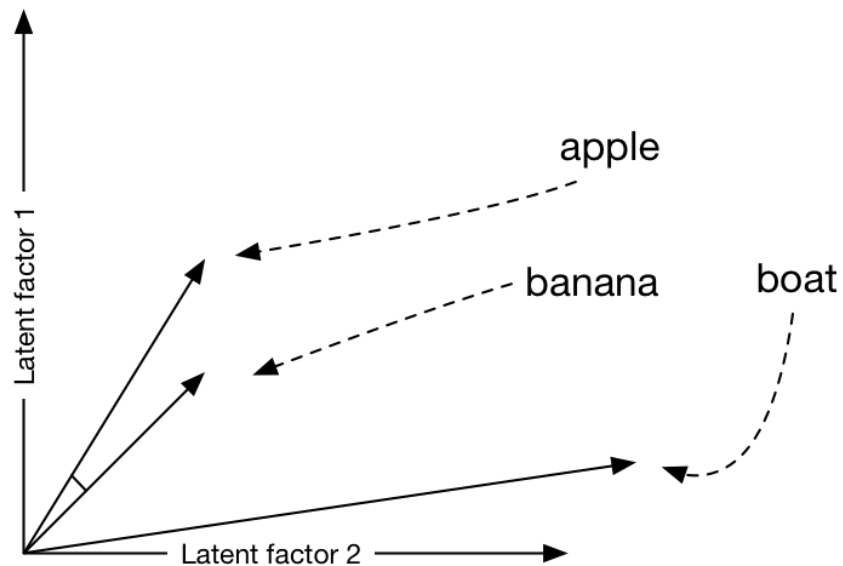
- Lưu trữ hiệu số giữa hai ID liên kề,  $id_i$  và  $id_{i+1}$ , thay cho giá trị ID thật sự.
  - $id_i < id_{i+1}$  vì các ID trong danh sách đảo đã sắp tăng dần.
- Hiệu số được gọi là **khoảng cách** (gap) giữa  $id_i$  và  $id_{i+1}$ .
  - Gap thường nhỏ hơn  $id_{i+1}$  và cần ít bit biểu diễn hơn.
- 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.
- Giá trị offset trong mỗi posting cũng có thể được lưu trữ theo cách tương tự.



# Nén số nguyên: ví dụ

Decimal	Unary	Elias Gamma	Elias Delta	Golomb ( $b = 3$ )	Golomb ( $b = 10$ )	Variable byte
1	1	1	1	1 10	1 001	0000001 0
2	01	0 10	0 100	1 11	1 010	0000010 0
3	001	0 11	0 101	01 0	1 011	0000011 0
4	0001	00 100	0 1100	01 10	1 100	0000100 0
5	00001	00 101	0 1101	01 11	1 101	0000101 0
6	000001	00 110	0 1110	001 0	1 1100	0000110 0
7	0000001	00 111	0 1111	001 10	1 1101	0000111 0
8	00000001	000 1000	00 100000	001 11	1 1110	0001000 0
9	000000001	000 1001	00 100001	0001 0	1 1111	0001001 0
10	0000000001	000 1010	00 100010	0001 10	01 000	0001010 0

Khoảng cách trong các mã phân cách phần tiền tố với phần hậu tố của mã.



---

# Chỉ mục ngữ nghĩa tiềm ẩn

---

# Hạn chế của IR truyền thống

---

- Các hệ thống IR cho đến nay chủ yếu dựa vào so khớp từ.
- Khái niệm và đối tượng có thể được biểu diễn theo nhiều cách khác nhau.
  - Tùy ngữ cảnh và thói quen ngôn ngữ.
  - Ví dụ: “picture”, “image” và “photo” là đồng nghĩa trong ngữ cảnh camera kỹ thuật số.
- Vấn đề này khiến cho hệ thống truy vấn có độ phủ thấp.
  - Nếu truy vấn dùng từ khác với từ đã lưu trữ, hệ thống không nhận diện được tài liệu liên quan.
- Latent semantic indexing (Deerwester et al.) giải quyết vấn đề bằng cách nhận diện mối liên kết thống kê giữa các từ.

# Tài liệu tham khảo

---



- Bing Liu. 2007. *Web Data Mining-Exploring Hyperlinks, Contents, and Usage Data*. Springer Series on Data-Centric Systems and Applications. **Chapter 6**.