



Nội dung

- WWW là gì?
- Khai phá dữ liệu là gì?
- Khai phá web là gì?

WWW là gì?

- WWW ảnh hưởng đến hầu hết các mặt của đời sống
 - Nguồn dữ liệu, thông tin và tri thức lớn nhất: phổ biến nhiều nhất, dễ dàng truy cập và tìm kiếm
 - Chứa hàng tỉ trang web liên kết với nhau được viết bởi hàng triệu tác giả khác nhau
- Web thay đổi nhận thức của con người về tìm kiếm thông tin
 - **Trước:** tham khảo ý kiến của bạn bè/người quen, mượn/mua các sản phẩm
 - **Với Internet,** giao dịch chỉ đơn giản ngay tại tại nhà.
- Web là một kênh giao dịch quan trọng
 - Có thể mua được gần như mọi thứ trên mạng mà không phải trực tiếp đi đến cửa hàng
 - Dễ dàng kết nối với bạn bè, thảo luận, chia sẻ quan điểm, ý kiến với bất cứ ai trên thế giới
 - Web là một thế giới ảo phản ánh chân thực xã hội loài người.

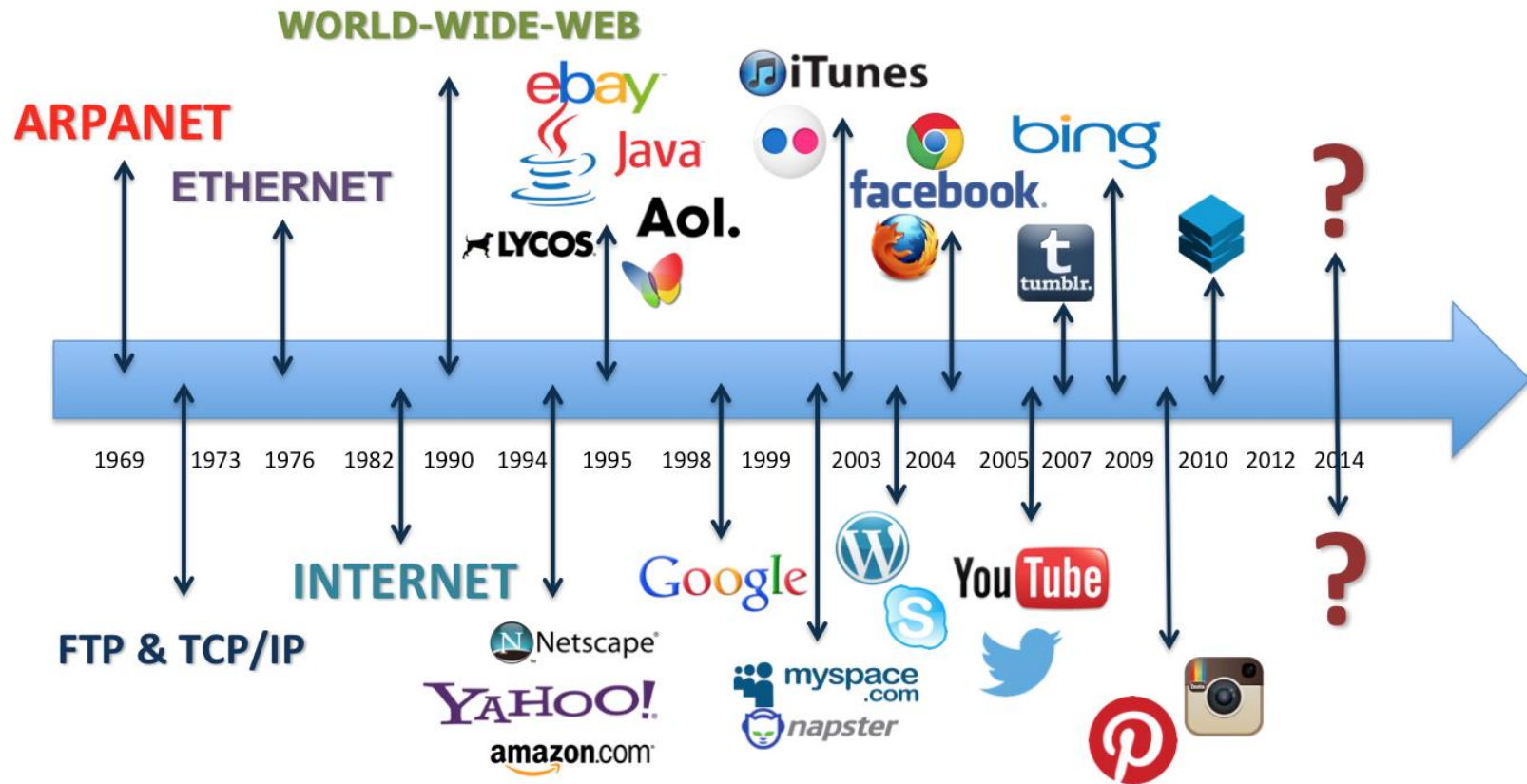
Định nghĩa WWW

- ❑ “Web là mạng máy tính cho phép người dùng (user) ở một máy tính truy cập đến thông tin lưu trữ trên một máy khác thông qua mạng Internet”
- ❑ Được xây dựng dựa trên kiến trúc client-server
 - Người dùng sử dụng một chương trình (khách) để kết nối với một máy từ xa (chủ) chứa dữ liệu
 - Việc duyệt web được dựa trên trình duyệt (VD: Firefox, Chrome):
 1. Gửi yêu cầu thông tin (request) tới máy chủ
 2. Nhận hồi đáp (response) từ máy chủ
 3. Biên dịch hồi đáp dưới dạng HTML
 4. Trình bày nội dung dưới dạng đồ họa trên màn hình
- ❑ Siêu văn bản (hypertext) cho phép liên kết tài liệu đến bất kỳ tài liệu khác thông qua siêu liên kết (hyperlink)
 - Để xem các tài liệu liên kết, người dùng cần click vào siêu liên kết
 - Siêu văn bản được phát minh bởi Ted Nelson vào năm 1965
- ❑ Siêu văn bản cho phép nhúng các nội dung đa phương tiện vào văn bản (ảnh, video, audio)

Lịch sử Web

- ❑ Được phát minh bởi **Tim Berners-Lee (CERN)** vào năm 1989 thông qua đề xuất về hệ thống siêu văn bản phân tán:
 - Cơ chế tổ chức thông tin phân cấp bậc lộ nhiều hạn chế
 - Đề xuất giao thức (protocol) có khả năng yêu cầu thông tin được lưu trữ trên một máy tính từ xa trên mạng
 - Đề xuất định dạng chung của các văn bản cho phép một văn bản có thể liên kết đến các văn bản khác
- ❑ Các thành phần cơ bản đầu tiên của web:
 - Máy chủ (server)
 - Trình duyệt (browser)
 - Giao thức liên lạc giữa máy chủ và máy khách (HTTP)
 - Ngôn ngữ đánh dấu siêu văn bản để soạn thảo văn bản (HTML)
 - Định dạng tài nguyên tổng quát (URL) để định danh văn bản

Lịch sử Web (tiếp)



Sự phát triển của web (Nguồn ảnh: Medium)

Lịch sử Web (tiếp)

- ❑ Mosaic ra đời năm 1993 tại Đại học Illinois
 - Trình duyệt đầu tiên cho phép sử dụng giao diện đồ họa và thao tác click chuột để duyệt web
 - Chạy trên ba hệ điều hành phổ biến là UNIX, Macintosh và Windows
- ❑ Năm 1994, Mosaic được công bố ra công chúng dưới cái tên Netscape
- ❑ Năm 1995, Internet Explorer của Microsoft ra đời



Lịch sử Web (tiếp)

- ❑ ARPANET (1969) được phát triển bởi ARPA, Bộ quốc phòng Mỹ
- ❑ Giao thức TCP/IP (1973) cho phép nhiều mạng máy tính kết nối và liên lạc với nhau
- ❑ Mạng Internet ra đời năm 1982 dựa trên giao thức TCP/IP



Lịch sử Web (tiếp)

- ❑ Thông tin được chia sẻ trên Web đã làm xuất hiện nhu cầu tìm kiếm thông tin một cách hiệu quả
- ❑ Máy tìm kiếm Excite được giới thiệu năm 1993
- ❑ Yahoo! được thành lập năm 1994
- ❑ Google được thành lập năm 1998
- ❑ Microsoft ra mắt MSN năm 2003 (Bing)
- ❑ W3C (The World Wide Web Consortium) được thành lập năm 1994 bởi MIT và CERN
 - Mục tiêu dẫn dắt sự phát triển của Web
 - Xây dựng các tiêu chuẩn cho Web
 - Thiết lập các đặc tả và tham chiếu để hỗ trợ sự tương tác giữa các sản phẩm trên Web
- ❑ Hội nghị WWW được tổ chức lần đầu tiên năm 1994
- ❑ 1995– 2001, Web được đầu tư phát triển và mở rộng
- ❑ 2001: bong bóng dot com

Khai phá dữ liệu là gì?

- Định nghĩa khai phá dữ liệu (KPD L)
- Lịch sử khai phá dữ liệu
- Các loại dữ liệu
- Các mẫu có thể khai thác
- Các kỹ thuật sử dụng trong KPD L
- Các ứng dụng của KPD L
- Cách thử thách trong KPD L

2.1. Định nghĩa khai phá dữ liệu

- ❑ Còn được gọi là quá trình khám phá tri thức trong CSDL (Knowledge Discovery in Databases)
- ❑ “là quá trình khám phá các mẫu (pattern) hoặc tri thức (knowledge) hữu ích từ các nguồn dữ liệu”
- ❑ Các mẫu phải đảm bảo các tính chất: đúng đắn, hữu ích, và dễ hiểu
- ❑ Các nguồn dữ liệu: CSDL, văn bản, ảnh, Web v.v.
- ❑ Khai phá dữ liệu là lĩnh vực liên ngành bao gồm học máy, thống kê, CSDL, trí tuệ nhân tạo, truy hồi thông tin, và trực quan hóa
- ❑ Các tác vụ chính trong khai phá dữ liệu: học có giám sát (phân loại), học không giám sát (phân cụm), khai phá luật kết hợp, khai phá mẫu tuần tự

2.1. Định nghĩa KPD L (tiếp)

❑ **Tiền xử lý:**

- Dữ liệu thô thường không phù hợp để khai phá
- Cần làm sạch để loại bỏ nhiễu hoặc bất thường
- Trong trường hợp dữ liệu quá lớn hoặc chứa nhiều thuộc tính không liên quan, cần thực hiện lấy mẫu hoặc trích chọn đặc trưng/thuộc tính (feature/attribute)

❑ **Khai phá dữ liệu:** Áp dụng các kỹ thuật khai phá trên dữ liệu đã tiền xử lý để tạo ra các mẫu hay tri thức

❑ **Hậu xử lý:** Lựa chọn các mẫu/tri thức hữu ích thông qua các kỹ thuật đánh giá hoặc/và trực quan hóa

❑ Quá trình khai phá dữ liệu được thực hiện lặp lại cho đến khi đạt được kết quả mong muốn

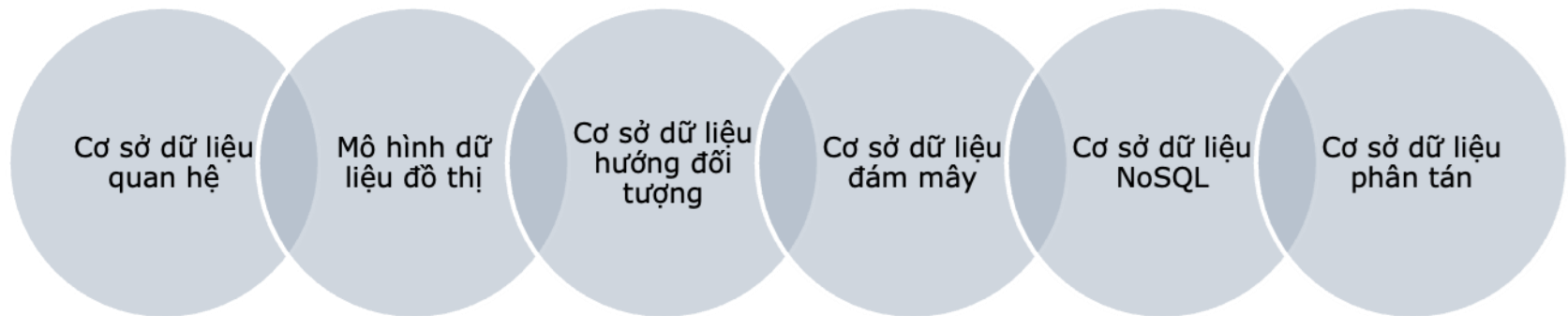
❑ Các kỹ thuật khai phá dữ liệu truyền thống dựa trên các dữ liệu có cấu trúc, với sự phát triển của Web, việc khai phá dữ liệu bán cấu trúc và phi cấu trúc trở nên quan trọng

2.2. Lịch sử KPDL

- ❑ Các hệ quản trị CSDL (70'-80')
 - Hệ quản trị CSDL phân cấp
 - Hệ quản trị CSDL mạng
 - Mô hình hóa dữ liệu: Mô hình thực thể -quan hệ
 - Các phương pháp đánh chỉ mục và truy cập Các ngôn ngữ truy vấn: SQL
 - Giao diện người dùng, form, báo cáo
 - Xử lý truy vấn và tối ưu hóa
 - Giao dịch, kiểm soát xung đột, khôi phục
 - Xử lý giao dịch trực tuyến (OLTP)

2.2. Lịch sử KPDL (tiếp)

- ❑ Các hệ quản trị CSDL tiên tiến (80'-nay)
 - Mô hình quan hệ mở rộng, mô hình quan hệ đối tượng
 - Quản lý dữ liệu phức tạp: dữ liệu không gian, thời gian, đa phương tiện, chuỗi; các đối tượng có cấu trúc, các đối tượng động
 - Các CSDL web (XML, web ngữ nghĩa)
 - Quản lý dữ liệu không chắc chắn và làm sạch dữ liệu
 - Tích hợp các nguồn không đồng nhất
 - Các hệ thống CSDL văn bản và tích hợp với tìm kiếm thông tin
 - Quản lý dữ liệu lớn
 - Tinh chỉnh hệ thống CSDL và các hệ thống tùy biến
 - Truy vấn nâng cao: xếp hạng
 - Điện toán đám mây và xử lý dữ liệu song song



2.2. Lịch sử KPDL (tiếp)

- ❑ Phân tích dữ liệu nâng cao (80'-nay)
 - Data warehouse và OLAP
 - Khai phá dữ liệu và khám phá tri thức: phân loại, phân cụm, phân tích ngoại lai, kết hợp và tương quan, tóm tắt so sánh,...
 - Khai phá dữ liệu phức tạp: dòng, chuỗi, văn bản, không gian, thời gian, đa phương tiện, web, mạng lưới
 - Ứng dụng của khai phá dữ liệu: kinh doanh, xã hội, buôn bán, ngân hàng, viễn thông, khoa học và công nghệ, mạng xã hội



2.3. Các loại dữ liệu

□ Dữ liệu từ CSDL

- Hệ quản trị CSDL gồm một tập hợp các dữ liệu có quan hệ với nhau và các phần mềm để quản lý và truy cập dữ liệu.
- Các phần mềm cung cấp cơ chế
 - Định nghĩa cấu trúc CSDL và lưu trữ dữ liệu
 - Mô tả và quản lý xung đột, chia sẻ, phân tán
 - Đảm bảo tính nhất quán và bảo mật
- Một CSDL quan hệ bao gồm các bảng:
 - Mỗi bảng bao gồm một tập các thuộc tính (cột, trường)
 - Các bản ghi (hàng) trong một bảng thể hiện một đối tượng được định danh bởi một khóa **duy nhất** và được miêu tả bởi các thuộc tính
- CSDL được truy cập dựa trên các câu truy vấn
 - Câu truy vấn được chuyển đổi thành một tập hợp các thao tác quan hệ như kết hợp, lựa chọn và sau đó được tối ưu hóa
 - Một câu truy vấn cho phép lấy về một phần cụ thể của dữ liệu
- Trong khai phá CSDL quan hệ, các tác vụ chủ yếu là **tìm kiếm xu hướng, mẫu dữ liệu hoặc phân tích độ lệch**

2.3. Các loại dữ liệu

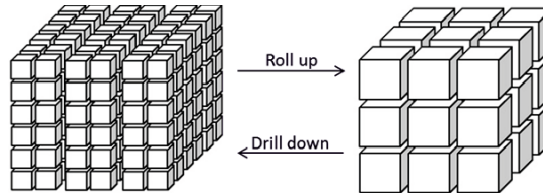
□ Kho Dữ liệu

- Là một kho thông tin được thu thập từ nhiều nguồn, được lưu trữ dưới một lược đồ thống nhất
- Được xây dựng thông qua một quy trình gồm làm sạch, tích hợp, chuyển đổi, dung nạp, và bảo trì định kỳ dữ liệu.
- Dữ liệu trong kho thường được tổ chức hướng đối tượng.
- Dữ liệu được *tóm tắt* và được lưu trữ để cung cấp thông tin theo góc nhìn lịch sử nhằm hỗ trợ ra quyết định (cho tổ chức).
- Kho DL thường được mô hình hóa bởi một cấu trúc dữ liệu đa chiều, được gọi là khối dữ liệu
 - Mỗi chiều là một hoặc một tập hợp các thuộc tính trong lược đồ
 - Mỗi ô (cell) lưu trữ một giá trị tổng hợp như số lượng (count) hoặc tổng (sum)
 - Một khối dữ liệu cung cấp một góc nhìn đa chiều của dữ liệu và cho phép tính toán trước và truy cập nhanh dữ liệu đã được tóm tắt

2.3. Các loại dữ liệu

❑ Kho Dữ liệu

- Hỗ trợ thực hiện các thao tác xử lý phân tích trực tuyến (OLAP)
- OLAP dựa trên nền tảng tri thức của miền dữ liệu để trình diễn dữ liệu ở các mức trừu tượng khác nhau.
- Hai thao tác cơ bản của OLAP là **drill-down** và **roll-up** cho phép người dùng quan sát dữ liệu ở các mức độ tóm tắt khác nhau. VD:
 - **Drill-down** cho phép quan sát dữ liệu tổng hợp mức tháng từ dữ liệu mức quý
 - **Roll-up** cho phép quan sát dữ liệu của đất nước dựa trên dữ liệu các tỉnh thành
- Các kĩ thuật khai phá dữ liệu đa chiều tổng quát cho phép kết hợp nhiều chiều ở các mức độ chi tiết khác nhau. Từ đó cho phép khám phá ra các mẫu biểu diễn các tri thức quan trọng. Là một kho thông tin được thu thập từ nhiều nguồn, được lưu trữ dưới một lược đồ thống nhất



2.3. Các loại dữ liệu

□ Dữ liệu giao dịch

- Mỗi bản ghi trong CSDL giao dịch thể hiện một giao dịch, bao gồm một định danh duy nhất và các thành phần tham gia vào giao dịch
- Các loại giao dịch phổ biến bao gồm chuyển khoản, thanh toán, mua hàng, đặt vé, click chuột
- CSDL giao dịch có thể chứa các bảng bổ sung như thông tin về người bán hay thông tin chi nhánh
- Tập trung vào việc phát hiện các tập phổ biến. Vd, trả lời câu hỏi *"Các sản phẩm nào hay được (khách hàng) mua cùng nhau?"*

2.3. Các loại dữ liệu

❑ Các loại dữ liệu khác

- ❑ Dữ liệu theo thời gian (chứng khoán), chuỗi (sinh học), không gian (bản đồ), thiết kế công nghiệp (thiết kế xây dựng, thành phần hệ thống, bo mạch), siêu văn bản và đa phương tiện, đồ thị và mạng lưới
- ❑ Các thách thức của cấu trúc dữ liệu (chuỗi, cây, đồ thị, mạng lưới) và ngữ nghĩa của dữ liệu (thứ tự, ngữ nghĩa của dữ liệu đa phương tiện, tính liên thông)
- ❑ Một số ứng dụng:
 - ❑ **Dữ liệu thời gian:** Xác định xu hướng giao dịch; Xác định xu hướng chứng khoán; Xác định bất thường
 - ❑ **Dữ liệu không gian:** Xác định tỉ lệ nghèo; Xác định các cộng đồng
 - ❑ **Dữ liệu văn bản:** Xác định mức độ hài lòng của khách hàng đối với sản phẩm dựa trên các nội dung đánh giá
 - ❑ **Dữ liệu đa phương tiện:** Nhận diện và phân loại đối tượng trong ảnh, xác định đoạn video có bàn thắng trong trận đấu

2.4. Các mẫu có thể khai thác

- ❑ Các chức năng trong khai phá dữ liệu:
 - Mô tả và phân biệt dữ liệu,
 - Khai phá các mẫu phổ biến,
 - Kết hợp và tương quan,
 - Phân loại và hồi quy,
 - Phân tích cụm và phân tích ngoại lai
- ❑ Các tác vụ trong khai phá dữ liệu chia làm hai loại: mô tả và dự đoán



2.4.1. Mô tả và phân biệt dữ liệu

- ❑ Mô tả DL là tóm tắt các tính chất/ đặc trưng chung của một lớp dữ liệu
 - ❑ Thống kê dữ liệu
 - ❑ Thao tác roll-up trong OLAP mô tả dữ liệu theo một chiều nhất định.
VD: Tổng kết các đặc điểm chung của những khách hàng tiêu dùng trên 100 triệu/năm → 40-50 tuổi, có nghề nghiệp (có thể drill-down tiếp theo chiều nghề nghiệp)
 - ❑ Kỹ thuật suy diễn hướng thuộc tính cho phép tổng quát hóa và mô tả dữ liệu mà không cần thực hiện tương tác người dùng theo từng bước
 - ❑ Phân biệt dữ liệu là so sánh các đặc trưng tổng quát của một lớp với các đặc trưng tổng quát của một hoặc vài lớp tương phản.
 - ❑ Các lớp tương phản được cung cấp bởi người dùng. Dữ liệu của các lớp này có thể được thu thập từ CSDL
 - ❑ Các mô tả so sánh có thể được thể hiện dưới dạng các luật
 - ❑ **VD:** Phân biệt những khách hàng thường xuyên vs hiếm khi mua đồ công nghệ
- **80%** khách hàng **thường xuyên** mua đồ công nghệ có độ tuổi **20-40** và có **bằng đại học**, **60%** khách hàng **hiếm khi** mua đồ công nghệ **không có bằng đại học**, drill-down theo chiều **trình độ học vấn** hoặc **thu nhập**.
-

2.4.2. Các mẫu phổ biến, kết hợp, tương quan

- ❑ Tập các đối tượng phổ biến bao gồm các đối tượng thường xuất hiện cùng nhau trong một CSDL giao dịch (VD: sữa và bánh mì thường được nhiều khách hàng mua cùng nhau); mẫu chuỗi phổ biến (vd: khách hàng thường mua lần lượt máy tính, máy ảnh và thẻ nhớ); mẫu cấu trúc phổ biến
- ❑ Phân tích kết hợp: Những sản phẩm nào thường được mua cùng nhau trong cùng một giao dịch?

mua(X, máy tính) → mua(X, phần mềm) [support=1%, confidence=50%)

- **X:** khách hàng
- **Độ tự tin (confidence)** thể hiện khả năng khách hàng mua phần mềm nếu biết khách hàng mua máy tính
- **Độ hỗ trợ (support)** thể hiện tỉ lệ giao dịch mà máy tính và phần mềm được mua cùng nhau trên tổng số giao dịch được phân tích
- Luật kết hợp theo một chiều mua: **máy tính → phần mềm [1%,50%]**
- Luật kết hợp đa chiều **tuổi(X, 20..29)^thunhập(X, 20..30tr) → mua(X, laptop)[2%, 60%]**
- Có **2%** có tuổi **20-29**, thu nhập **20-30tr** đã mua laptop trong tổng số khách hàng được phân tích (thu thập từ CSDL quan hệ)
- Có **60%** khả năng những người trong độ tuổi **20-29** và có thu nhập **20-30tr** sẽ mua laptop
- Luật kết hợp **không** thỏa mãn nếu ở dưới ngưỡng độ hỗ trợ tối thiểu (minimum support threshold) và độ tự tin tối thiểu (minimum confidence threshold)

2.4.3. Phân loại và hồi quy

- ❑ Phân loại là quá trình tìm kiếm một mô hình mô tả và phân biệt các lớp dữ liệu.
 - Mô hình được suy ra từ dữ liệu huấn luyện (các đối tượng đã biết nhãn)
 - Mô hình được sử dụng để dự đoán nhãn của các đối tượng chưa biết
 - Mô hình có thể được biểu diễn dưới dạng các luật, cây quyết định, công thức toán học/xác suất, hoặc mạng nơ-ron
- ❑ Hồi quy mô hình hóa các hàm có giá trị liên tục.



Hồi quy

Ngày mai nhiệt độ là bao nhiêu?

Dự đoán

84°



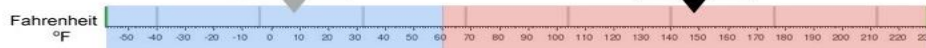
Phân loại

Ngày mai trời Nóng hay Lạnh?

Dự đoán

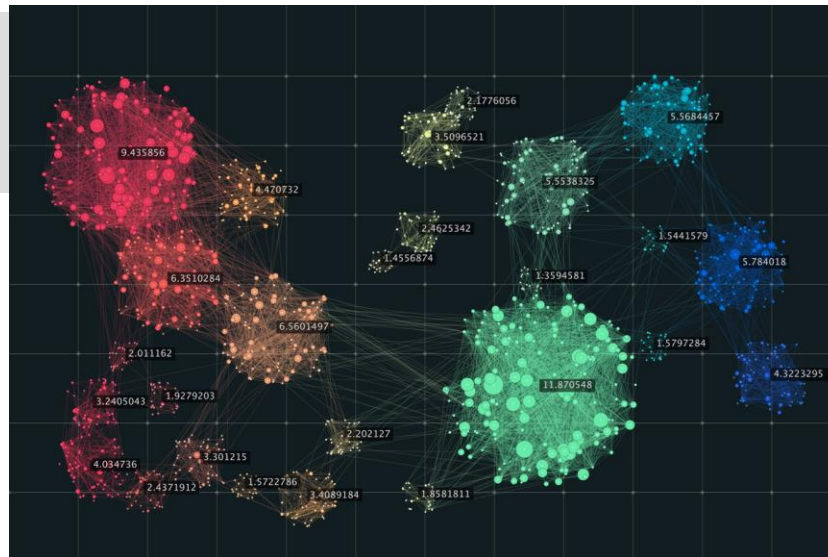
COLD

HOT



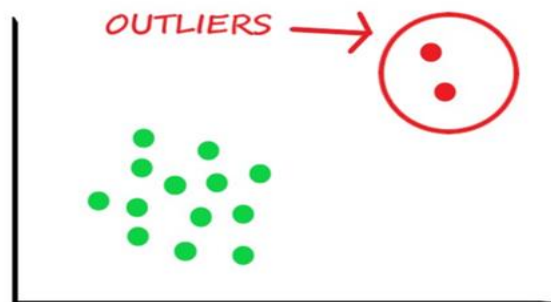
2.4.4. Phân cụm

- ❑ Các đối tượng được phân cụm dựa trên các tiêu chí cực đại hóa độ tương tự bên trong cụm cực tiểu hóa độ tương tự giữa các cụm
- ❑ Mỗi cụm có thể coi là một lớp đối tượng để từ đó rút ra các quy luật
- ❑ Phân cụm cũng được áp dụng vào xây dựng cây phân cấp (taxonomy)



2.4.5. Phân tích ngoại lai

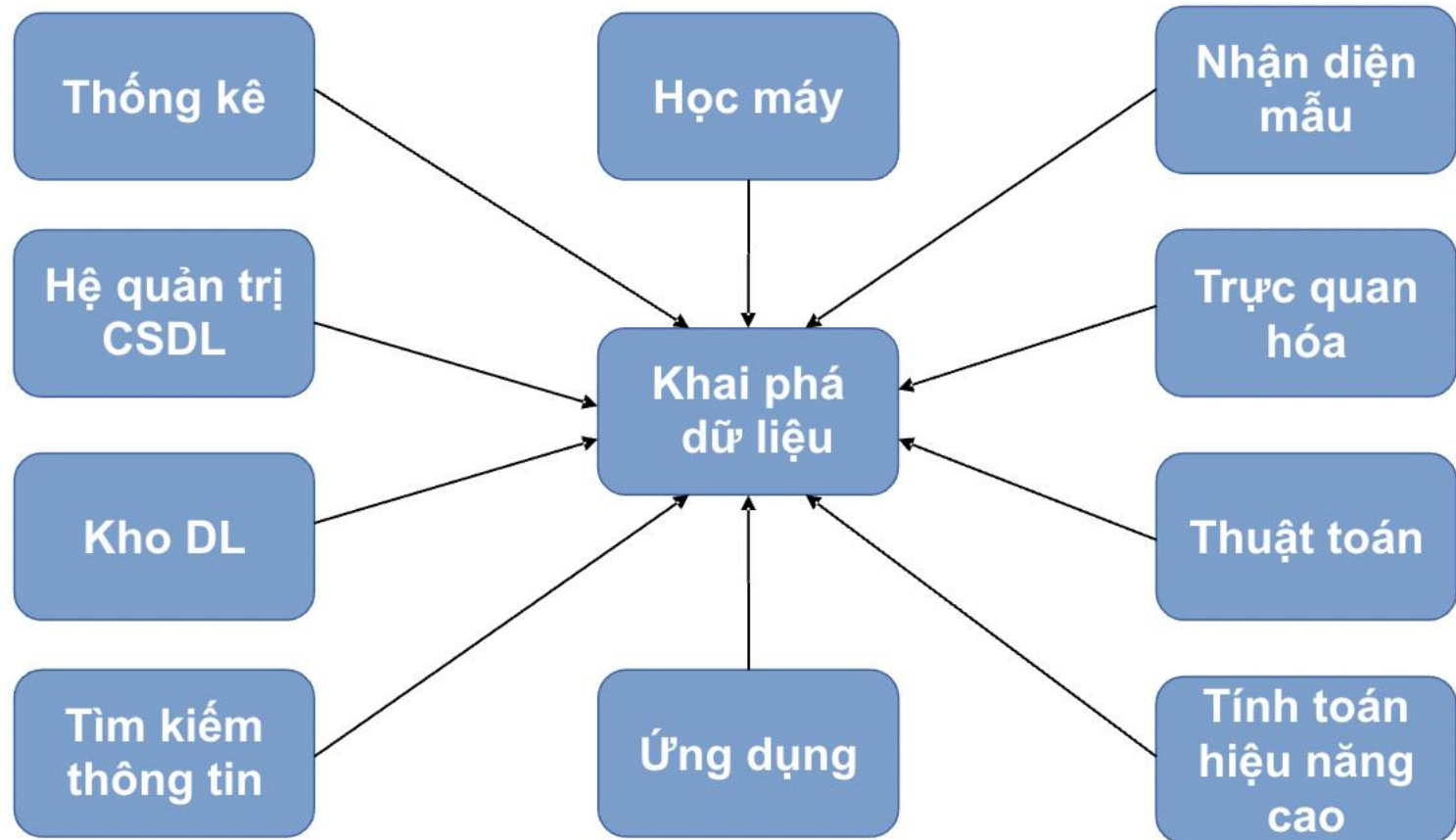
- ❑ Là các đối tượng không phù hợp với hành vi chung/mô hình của tập dữ liệu
- ❑ Phần lớn các kĩ thuật KPD L coi phần tử ngoại lai là nhiễu hoặc ngoại lệ.
- ❑ Trong một số ứng dụng như phát hiện giao dịch phạm pháp thì các phần tử này được quan tâm hơn là các phần tử thông thường
- ❑ Các kiểm tra thông kê thông thường có thể phát hiện các phần tử ngoại lai dựa trên việc mô hình hóa dữ liệu theo một phân bố/mô hình xác suất
- ❑ Phương pháp dựa trên khoảng cách phát hiện ra các phần tử bất thường nằm cách xa các cụm dữ liệu
- ❑ Phương pháp dựa trên mật độ có thể phát hiện phần tử bất thường trong một khu vực dù các phân phối thống kê tổng quát không phân biệt được
- ❑ **VD:** Các giao dịch thẻ phạm pháp có thể được phát hiện dựa trên giá trị giao dịch khi so sánh với mức giao dịch thông thường của cùng tài khoản; dựa trên các thông tin về địa điểm, loại giao dịch hoặc tần suất giao dịch



2.4.6. Các mẫu đáng quan tâm

- ❑ **Mẫu đáng quan tâm:** i) dễ hiểu; ii) đúng đắn trên dữ liệu mới hoặc dữ liệu thử nghiệm với một mức độ chắc chắn; iii) có (tiềm năng) hiệu quả; iv) mới hoặc xác nhận một giả thiết mà người dùng đặt ra
- ❑ Một mẫu đáng quan tâm thể hiện một tri thức
- ❑ Các phép đo khách quan dựa trên cấu trúc và thống kê của các mẫu
 - Trong phân tích kết hợp, độ hỗ trợ của một luật kết hợp $X \Rightarrow Y$ thể hiện tỉ lệ đáp ứng được luật kết hợp đó. Độ tự tin thể hiện xác suất $P(Y|X)$
 - Trong phân loại, độ chính xác (accuracy): tỉ lệ dữ liệu được phân loại chính xác; độ phủ (coverage): tỉ lệ dữ liệu đáp ứng luật đó
- ❑ Các phép đo chủ quan dựa trên niềm tin của người dùng vào dữ liệu
 - ❑ Mẫu đáng quan tâm nếu không được người dùng kỳ vọng (ngược lại với niềm tin)
 - ❑ Mẫu đáng quan tâm nếu cung cấp thông tin chiến lược dựa trên đó người dùng có thể ra quyết định (**VD:** “một trận động đất thường kèm theo nhiều dư chấn nhỏ”)
 - ❑ Mẫu đáng quan tâm nếu đúng như kỳ vọng nhằm xác nhận một giả thuyết của người dùng
- ❑ Một hệ thống khai phá dữ liệu có thể sinh ra tất cả các mẫu đáng quan tâm hay không?
- ❑ Một hệ thống khai phá dữ liệu có thể chỉ sinh ra các mẫu đáng quan tâm hay không?

2.5. Các kỹ thuật trong KPDL

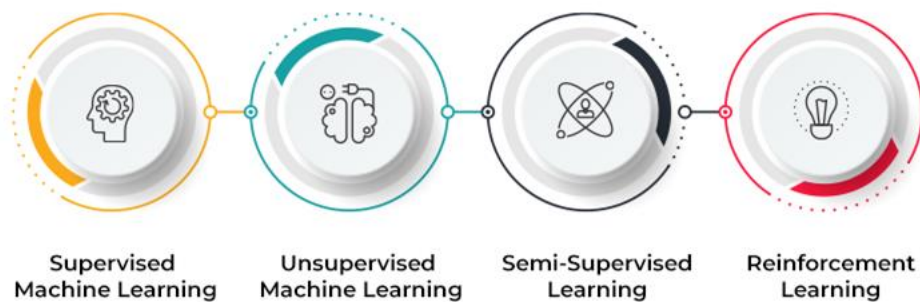


2.5.1. Thống kê

- ❑ Lĩnh vực thống kê nghiên cứu việc thu thập, phân tích, diễn giải, giải thích và biểu diễn dữ liệu.
- ❑ Mô hình thống kê là một tập hợp các hàm toán học mô tả hành vi của các đối tượng trong một lớp mục tiêu dưới dạng các biến ngẫu nhiên và các phân phối xác suất liên quan.
- ❑ Thống kê suy diễn (inferential statistics) mô hình hóa dữ liệu dựa trên quá trình ngẫu nhiên và sự không chắc chắn của các quan sát
- ❑ Mô tả và phân loại DL có đầu ra có thể là các mô hình thống kê
- ❑ Việc khai phá mẫu có thể sử dụng các mô hình thống kê để nhận diện và xử lý nhiễu và khuyết thiếu trong dữ liệu dựa trên các mô hình thống kê
- ❑ Kết quả của khai phá dữ liệu có thể được xác nhận bởi kiểm định giả thuyết thống kê (statistical hypothesis test).
- ❑ Kết quả của mô hình đáng kể về mặt thống kê (statistically significant) nếu nó ít có khả năng được sinh ra một cách vô tình
- ❑ Phần lớn các phương pháp thống kê có độ phức tạp tính toán lớn

2.5.2. Học máy

- ❑ Học máy nghiên cứu cách thức để máy có thể học (hoặc cải thiện hiệu quả học) dựa trên dữ liệu
- ❑ Một hướng nghiên cứu chính là làm cho các chương trình máy tính có thể tự động nhận diện các mẫu phức tạp trong dữ liệu và đưa ra các quyết định thông minh dựa trên dữ liệu
- ❑ Học có giám sát (phân loại) dựa trên dữ liệu có nhãn trong tập huấn luyện
- ❑ Học không giám sát (phân cụm) sử dụng dữ liệu không có nhãn
- ❑ Học bán giám sát dựa trên cả dữ liệu có nhãn và không có nhãn.
- ❑ Học chủ động: Cho phép người dùng tham gia vào quá trình học, mục tiêu là tối ưu hóa chất lượng của mô hình



2.5.2. Hệ quản trị CSDL và Kho dữ liệu

- ❑ Các hệ thống CSDL tạo ra, vận hành, và sử dụng các CSDL cho tổ chức và người dùng cuối.
- ❑ Hệ thống CSDL thiết lập các nguyên lý về mô hình dữ liệu, ngôn ngữ truy vấn, xử lý truy vấn, các phương pháp tối ưu, lưu trữ dữ liệu, đánh chỉ mục và các phương pháp truy cập.
- ❑ Các hệ thống CSDL có thể xử lý những tập dữ liệu có cấu trúc lớn
- ❑ Khai phá dữ liệu cần xử lý lượng dữ liệu lớn, thời gian thực, dữ liệu dòng.
- ❑ Khai phá dữ liệu có thể ứng dụng các kĩ thuật trong hệ thống CSDL và tích hợp các chức năng khai phá dữ liệu và hệ thống CSDL
- ❑ Kho DL tích hợp dữ liệu từ nhiều nguồn khác nhau, hỗ trợ các thao tác OLAP và khai phá dữ liệu đa chiều

2.5.3. Tìm kiếm thông tin

- ❑ Tìm kiếm thông tin là khoa học tìm kiếm các tài liệu hoặc thông tin trong các tài liệu
- ❑ Các tài liệu có dạng văn bản hoặc đa phương tiện (hình ảnh, âm thanh, video)
- ❑ Hai giả thiết:
 - Dữ liệu ở dạng phi cấu trúc
 - Câu truy vấn bao gồm các từ khóa và không có cấu trúc phức tạp
- ❑ Mô hình chủ đề cho phép phát hiện ra các chủ đề chính trong một tập văn bản cũng như trong từng văn bản



2.6. Các ứng dụng của KPDL

- ❑ Các doanh nghiệp cần có hiểu biết về bối cảnh kinh doanh của họ: khách hàng, thị trường, nguồn cung, tài nguyên và đối thủ.
- ❑ Thông minh doanh nghiệp (Business Intelligence) cung cấp góc nhìn lịch sử, hiện tại, và dự đoán các hoạt động
- ❑ Bgồm báo cáo, OLAP, quản lý hiệu năng doanh nghiệp, thông minh cạnh tranh, xác lập tiêu chuẩn và phân tích dự đoán
- ❑ Giúp doanh nghiệp phân tích thị trường một cách hiệu quả, so sánh phản hồi của khách hàng, phân tích điểm mạnh và điểm yếu của đối thủ, và đưa ra những quyết định khôn ngoan
- ❑ Các công cụ OLAP dựa trên kho DL và khai phá dữ liệu đa chiều; phân tích thị trường, nguồn cung và bán hàng; quản lý quan hệ khách hàng; đưa ra dịch vụ phù hợp

BUSINESS INTELLIGENCE



2.6. Các ứng dụng của KPD L

- ❑ Các kĩ thuật khai phá dữ liệu được áp dụng trong máy tìm kiếm bao gồm crawling (crawl trang nào và tần suất crawl), đánh chỉ mục (đánh chỉ mục trang nào và những phần nào trong trang), tìm kiếm (xếp hạng trang web, hiển thị quảng cáo, cá nhân hóa và phụ thuộc ngữ cảnh)
- ❑ Máy tìm kiếm sử dụng hạ tầng điện toán đám mây với hàng nghìn thậm chí hàng trăm nghìn nút
- ❑ Các mô hình tìm kiếm và bộ phân loại truy vấn được xây dựng offline; câu truy vấn được xử lý online; các mô hình và bộ phân loại truy vấn cần được cập nhật theo thời gian do sự thay đổi của các câu truy vấn và dữ liệu web
- ❑ Cá nhân hóa kết quả tìm kiếm đưa ra câu trả lời dựa trên hồ sơ người dùng và lịch sử tìm kiếm; việc xử lý các câu truy vấn hiếm là một thách thức lớn

2.7. Các thách thức trong KPD L

- ❑ **Khai phá tri thức mới và đa dạng:** Kết hợp phân cụm và xếp hạng tạo ra các cụm chất lượng cao và xếp hạng các đối tượng trên mạng lưới lớn.
- ❑ **Khai phá trong không gian đa chiều** bằng cách kết hợp các chiều (thuộc tính) dữ liệu
- ❑ **Khai phá trong môi trường liên kết ngữ nghĩa:** tri thức khai phá từ một tập các đối tượng thúc đẩy việc khai phá tri thức từ các đối tượng liên quan
- ❑ **Xử lý dữ liệu nhiễu, không chắc chắn, khuyết thiếu.** Các loại dữ liệu này có thể làm quá trình khai phá dữ liệu bị ảnh hưởng và sinh ra mẫu sai. Các kĩ thuật làm sạch dữ liệu, tiền xử lý dữ liệu, phát hiện và loại bỏ bất thường, suy diễn không chắc chắn cần được áp dụng.
- ❑ **Đánh giá mẫu dựa trên các tiêu chí chủ quan;** định hướng quá trình khai phá theo các mẫu đáng quan tâm hoặc theo người dùng để tăng chất lượng mẫu khai phá được và giới hạn không gian tìm kiếm

2.7. Các thách thức trong KPDL (tiếp)

- ❑ Các thuật toán cần hiệu quả và khả mở để trích rút thông tin từ lượng dữ liệu lớn và trong môi trường động. Thời gian thực thi cần dự đoán được, ngắn và chấp nhận được.
- ❑ **Tính toán phân tán và song song:** Dữ liệu được chia thành các phần (piece) và được xử lý song song trên các tiến trình; Các tiến trình này có thể tương tác với nhau để trao đổi thông tin và dữ liệu; Sau đó, kết quả của các tiến trình được kết hợp với nhau.
- ❑ Khai phá cải thiện cho phép cập nhật dữ liệu mới mà không phải thực hiện quá trình khai phá từ đầu trên toàn bộ dữ liệu chính để cập nhật và làm giàu tri thức đã được khai phá.

2.8. Tác động của KPDL đến xã hội

- ❑ Hạn chế tác hại, đem lại lợi ích cho xã hội
- ❑ Đánh giá tính nhạy cảm của dữ liệu, đảm bảo quyền riêng tư
- ❑ Tích hợp khai phá dữ liệu vào các hệ thống sẵn có nhằm nâng cao hiệu quả phục vụ người dùng mà không đòi hỏi hiểu biết của người dùng về các kỹ thuật khai phá dữ liệu



3. Khai phá Web là gì?

- ❑ Web là nguồn dữ liệu có thể truy cập công cộng lớn nhất trên thế giới, được phát triển bởi rất nhiều người dùng (user).
- ❑ Các tính chất đặc thù của dữ liệu Web:
 - Lượng dữ liệu/thông tin trên Web **rất lớn** và **vẫn đang phát triển**. Mức độ bao phủ thông tin rộng và đa dạng. Người dùng có nhu cầu tìm thấy thông tin về mọi thứ trên Web
 - **Dữ liệu đa dạng về cấu trúc:** có cấu trúc (structured), bán cấu trúc (semi-structured), phi cấu trúc (unstructured)
 - Thông tin trên Web **hỗn tạp** (heterogeneous).
 - Một số lượng đáng kể thông tin trên Web được **liên kết**.
 - Thông tin trên Web nhiều loạn:
 - Một trang web thông thường chứa nhiều mẫu thông tin
 - Web về cơ bản không có chính sách kiểm duyệt nội dung.
 - Web là một kênh kinh doanh và thương mại.
 - Web có tính động. Nội dung trên Web thay đổi liên tục.
 - Web là một xã hội ảo.

3. Khai phá Web là gì? (tiếp)

- ❑ Xác định sự phân bố của thông tin trên web
- ❑ Xác định tính chất và phân loại các trang web
- ❑ Theo dõi sự vận động của web
- ❑ Xác định các mối quan hệ giữa các thực thể web: trang web, người dùng, cộng đồng và các hoạt động trên web



Q&A

Tài liệu tham khảo

- ❑ Zdravko Markov and Daniel T. Larose. **Data Mining the Web: Uncovering Patterns in Web Content, Structure, and Usage**, Wiley, 2007, ISBN: 978-0-471-66655-4.
- ❑ Web Mining, ĐHBK Hà Nội.

Thank you!