**Khai khoáng dữ liệu là gì?**

Trích xuất từ dữ liệu những thông tin hữu ích nhưng tiềm ẩn, chưa được biết.

**Những gì khai khoáng dữ liệu có thể làm?**

Mô tả (Description), Ước lượng (Estimation), Dự đoán (Prediction), Phân lớp (Classification), Gom nhóm (Clustering), Kết hợp (Association)

**Mô tả chi tiết quá trình khám phá tri thức/phát hiện tri thức/khai phá dữ liệu?**

Quy trình khai phá dữ liệu là một chuỗi lặp (iterative) và tương tác (interactive) gồm các bước (giai đoạn) bắt đầu với dữ liệu thô (raw data) và kết thúc với tri thức (knowledge of interest) đáp ứng được sự quan tâm của người sử dụng.

Bao gồm các bước cơ bản sau đây:

* Chọn lọc dữ liệu (Selection)
* Tiền xử lý dữ liệu (Preprocessing)
* Chuyển đổi dữ liệu (Transformation)
* Khai phá dữ liệu (Data mining)
* Đánh giá kết quả mẫu (Interpretation evaluation)

**Tại sao phải tiền xử lý dữ liệu?**

No quality data, no quality mining results

* Để đưa ra quyết định hiệu quả cần phải dựa trên dữ liệu chất lượng
* Kho dữ liệu cần phải được tích hợp bởi các dữ liệu chất lượng
* Trùng lắp hay thiếu dữ liệu sẽ dẫn đến việc thống kê sai hay hiểu nhầm đặc điểm của dữ liệu

Data preparation, cleaning, and transformation comprises the majority of the work in a data mining application (90%)

Data in the real world is dirty

* Không đầy đủ (incomplete) thiếu giá của thuộc tính, thiếu thông tin thuộc tính quan tâm, … Ví dụ: nghề nghiệp = “”
* Nhiễu (noisy) chứa sai sót hoặc ngoại lệ. Ví dụ: dung lượng = “-40”
* Không nhất quán (inconsistent) có sự sai biệt trong mã hoặc tên. Ví dụ: tuổi = 42, ngày sinh = “03/07/1997” hay đánh giá “1, 2, 3” và đánh giá “A, B, C”, …

**Công việc/ nhiệm vụ của Tiền xử lý dữ liệu:**

* Làm sạch (Data Cleaning): điền các giá trị còn thiếu, làm trơn các dữ liệu nhiễu (smooth noisy data), xác định hay loại bỏ các ngoại lệ, giải quyết dữ liệu không nhất quán.
* Tích hợp (Data Integration): tích hợp nhiều cơ sở dữ liệu, khối dữ liệu, tệp tin hoặc ghi chú
* Chuyển đổi (Data Transformation): Chuẩn hoá dữ liệu (scaling to a specific range), Kết hợp dữ liệu (aggregation)
* Giảm thiểu (Data reduction): Obtains reduced representation in volume but produces the same or similar analytical results; Data discretization: with particular importance, especially for numerical data; Data aggregation, dimensionally reduction, data compresstion, generalization

**Data cleaning – làm sạch dữ liệu**

* Fill in missing values – Bổ sung dữ liệu bị thiếu;
* Identify outliers and smooth out noisy data: nhận diện phần tử biên và giảm thiểu nhiễu;
* Correct inconsistent data – Xử lý dữ liệu không nhất quán;
* Fill in missing values – Xử lý dữ liệu bị thiếu;

**Nguyên nhân gây thiếu dữ liệu:**

* Sự cố thiết bị.
* Không tương thích với dữ liệu trước đó nen giá trị (mới) bị xoá đi.
* Dữ liệu không được nhập vào (lỗi người nhập liệu).
* Không lưu trữ lịch sử hay sự thay đổi của dữ liệu (thông tin thuyên chuyển của cán bộ trong 1 đơn vị).

**Fill in missing values – Xử lý dữ liệu bị thiếu**

Cách xử lý dữ liệu bị thiếu

* Bỏ qua các bản ghi có dữ liệu bị thiếu
* Bổ sung dữ liệu bị thiếu bằng tay
* Bổ sung dữ liệu bị thiếu tự động:
  + Giá trị trung bình của thuộc tính, của thuộc tính cùng lớp
  + Giá trị hằng số nhất định
  + Giá trị có thể xảy ra nhất

**Identify outliers and smooth out noisy data – Xử lý dữ liệu bị nhiễu**

* Dữ liệu nhiễu là dữ liệu (đối tượng) không tuân theo đặc tính/ hành vi chung của tập dữ liệu.
* Giá trị không chính xác do:
  + Lỗi do thiết bị thu thập dữ liệu
  + Vấn đề nhập dữ liệu: người dùng hoặc máy có thể sai
  + Vấn đề truyền dữ liệu: sai từ thiết bị gửi/nhận/truyền
  + Hạn chế của công nghệ: ví dụ, phần mềm có thể xử lý không đúng
  + Thiết nhất quán khi đặt tên: cũng một tên song cách viết khác nhau

**Identify outliers and smooth out noisy data – Xử lý dữ liệu bị nhiễu (giải pháp)**

* Phân khoảng (Bining):
  + Sắp dữ liệu tăng và chia “đều” vào các thùng (bin).
  + Làm trơn: theo trung bình, theo trung vị,…
  + Hồi quy (Regression): Gắn dữ liệu với một hàm hồi quy (regression function), …
* Phân cụm (Clustering): Phát hiện liệu và loại bỏ các ngoại lai (sau khi đã xác định các cụm).
* Kết hợp giữa máy tính (phát hiện) và kiểm tra của con người (hiệu chỉnh).

**Handle noisy data - Nắm bắt dữ liệu nhiễu**

* Phương pháp đóng thùng (Bining):
  + Sắp xếp dữ liệu tăng và chia “đều” vào các thùng.
  + Làm trơn: theo trung bình, theo trung tuyến, theo biên…
* Phân cụm (Clustering):
  + Phát hiện và loại bỏ ngoại lai (outliers).
* Kết hợp kiểm tra máy tính và con người:
  + Phát hiện giá trị nghi ngờ để con người kiểm tra (chẳng hạn, đối phó với ngoại lai có thể).
* Hồi quy:
  + Làm trơn: ghép dữ liệu theo các hàm hồi quy.

**Correct inconsistent data – Xử lý dữ liệu không nhất quán**

* Dữ liệu được ghi nhận khác nhau, ví dụ: 9/3/2018 và 3/9/2018, …
* Nguyên nhân gây ra sự không nhất quán:
  + Sự không nhất quán trong các qui ước đặt tên hay mã dữ liệu.
  + Định dạng không nhất quán của các vùng nhập liệu.
  + Thiết bị ghi nhận dữ liệu.
* Giải pháp:
  + Tạo các ràng buộc khi nhập liệu.
  + Điều chỉnh dữ liệu không nhất quán bằng tay sau khi nhập liệu.
  + Viết các giải thuật điều chỉnh, chuyển đồi tự động.

**Data integration – tích hợp dữ liệu**

* Tích hợp dữ liệu (Data integration):
  + Kết hợp dữ liệu từ nhiều nguồn vào một kho dữ liệu thống nhất
* Tích hợp ở mức mô hình (Schema integration):
  + Tích hợp metadata từ các nguồn khác nhau
  + Ví dụ: A.cust-id B.customID
* Vấn đề xác định thực thể (để tránh dư thừa dữ liệu):
  + Cần xác định các thực thể (identities) trên thực tế từ nhiều nguồn dữ liệu
  + Ví dụ: Bill Clinton B. Clinton
* Phát hiện và xử lý các mâu thuẫn đối với giá trị dữ liệu:
  + Đối với cùng một thực thể trên thực tế, nhưng các giá trị thuộc tính từ nhiều nguồn khác nhau lại khác nhau. Các lý do thực thể:
    - Các cách biểu diễn khác nhau.
    - Mức đánh giá, độ đo (scales) khác nhau. Ví dụ: hệ đo lường mét với hệ đo lường của Anh.
* Dư thừa dữ liệu (redundant data) thường xuyên xảy ra, khi tích hợp dữ liệu từ nhiều nguồn (ví dụ: từ nhiều CSDL):
  + Định danh đối tượng: Cùng một thuộc tính (hay cùng một đối tượng) có thể mang các tên (định danh) khác nhau trong các CSDL khác nhau.
  + Dữ liệu suy ra được: Một thuộc tính trong một bảng có thể là một thuộc tính được suy ra (derived attribute) trong một bảng khác. Ví dụ: “Annual Revenue” và “Monthly Revenue”.
* Các thuộc tính dư thừa có thể được phát hiện bằng phân tích tương quan (Correlation analys): Pearson, Cosine, chi-square, …
* Yêu cầu chung đối với quá trình tích dữ liệu: Giảm thiểu (tránh được là tốt nhất) các dư thừa và các mâu thuẫn
  + Giúp cải thiện tốc độ của quá trình khai phá dữ liệu, và nâng cao chất lượng của các kết quả (tri thức) thu được.

**Data transformation – chuyển đổi dữ liệu**

* Biến đồi dữ liệu: quá trình biến đổi hay kết hợp dữ liệu vào những dạng thích hợp cho quá trình khai phá dữ liệu
* Làm trơn dữ liệu (smoothing): ước lượng Laplace
* Kết hợp dữ liệu (aggregation).
* Chuẩn hoá (normalization):
  + min-max normalization
    - Giá trị cũ: v [minA, maxA]
    - Giá trị mới: v’ [new\_minA, new\_maxA]
      * Ví dụ: chuẩn hoá điểm số từ 0-4.0 sang 0-10.0
      * Đặc điểm của phép chuẩn hoá min-max?



* + z-score normalization
    - Giá trị cũ: v tương ứng với mean và standard deviation A
    - Giá trị mới: v’
      * Đặc điểm của chuẩn hoá z-score?

A close-up of a calculator

Description automatically generated with low confidence

**Data reduction – Thu giảm dữ liệu**

* Kho dữ liệu chứa tới hàng TB:
  + Phân tích/khai phá dữ liệu phức mất thời gian rất dài khi chạy trên tập toàn bộ dữ liệu.
* Rút gọn/ thu giảm dữ liệu:
  + Có được trình bày gọn của tập dữ liệu mà nhỏ hơn nhiều về khối lượng mà sinh ra cùng (hoặc hầu như cùng) kết quả.
* Chiến lược rút gọn dữ liệu:
  + Rút gọn đặc trưng – loại bỏ thuộc tính không quan trọng.
  + Kết hợp khối dữ liệu.
  + Thu giảm chiều: PCA (phân tích thành phần chính).
* Rút gọn đặc trưng (như., lựa chọn tập con thuộc tính):
  + Lựa chọn tập nhỏ nhất các đặc trưng mà phân bố xác suất của các lớp khác nhau cho giá trị khi cho giá trị của các lớp này gần như phân bổ vốn có đã cho giá trị của các đặc trưng.
  + Rút gọn # của các mẫu trong tập mẫu dễ dàng hơn để hiểu dữ liệu.
* Kết hợp khối dữ liệu (data cube aggregation):
  + Dạng dữ liệu: additive, semi-additive (numerical).
  + Kết hợp dữ liệu bằng các hàm nhóm: average, min, max, sum, count, …
    - Dữ liệu ở các mức trừu tượng khác nhau.
    - Mức trừu tượng càng cao giúp thu giảm lượng dữ liệu càng nhiều.