**TIỀN XỬ LÝ DỮ LIỆU TRONG KHAI THÁC DỮ LIỆU**

Nguồn:

Data Mining: Concepts and Techniques - Jiawei Han, Micheline Kamber, Jian Pei

I. Giới thiệu

Tiền xử lý dữ liệu là bước quan trọng trong quá trình khai thác dữ liệu (data mining), giúp cải thiện chất lượng dữ liệu và nâng cao hiệu quả phân tích. Trong thực tế, dữ liệu thường chứa nhiều vấn đề như thiếu sót, nhiễu, không nhất quán hoặc dư thừa.

II. Các yếu tố ảnh hưởng đến chất lượng dữ liệu

1. Độ chính xác (Accuracy)  
   Dữ liệu chất lượng phải đảm bảo tính chính xác tuyệt đối, không chứa sai sót từ khâu nhập liệu, truyền tải hay vận hành thiết bị.
2. Độ đầy đủ (Completeness)  
   Dữ liệu lý tưởng cần bao gồm đầy đủ tất cả các thuộc tính quan trọng, không để xảy ra tình trạng thiếu giá trị hoặc thông tin chưa hoàn thiện.
3. Tính nhất quán (Consistency)  
   Hệ thống dữ liệu chuẩn phải duy trì sự đồng bộ cao trong cách đặt tên, định dạng và phương thức mã hóa xuyên suốt các nguồn dữ liệu khác nhau.
4. Tính kịp thời (Timeliness)  
   Dữ liệu chất lượng luôn được cập nhật liên tục và kịp thời, đảm bảo phản ánh chính xác nhất trạng thái hiện tại của hệ thống.
5. Độ tin cậy (Believability)  
   Dữ liệu đáng tin cậy phải tạo được niềm tin vững chắc từ người dùng thông qua quá trình kiểm chứng nghiêm ngặt và lịch sử chính xác.
6. Khả năng diễn giải (Interpretability)  
   Dữ liệu chuẩn mực cần được trình bày rõ ràng, dễ hiểu với hệ thống mã hóa thân thiện, giúp người dùng dễ dàng nắm bắt thông tin.

III. Các kỹ thuật tiền xử lý dữ liệu chính

1. Làm sạch dữ liệu (Data Cleaning)

Data Cleaning là một quy trình quan trọng để "làm sạch" dữ liệu bằng cách xử lý các vấn đề như:

* Điền giá trị thiếu: Đảm bảo rằng các giá trị bị thiếu được điền đầy đủ.
* Làm mịn dữ liệu bị nhiễu: Xử lý các dữ liệu không đồng nhất hoặc bị sai lệch để không ảnh hưởng đến quá trình phân tích.
* Loại bỏ giá trị ngoại lệ: Phát hiện và loại bỏ các dữ liệu không hợp lý hoặc ngoài phạm vi thông thường.
* Giải quyết sự không nhất quán: Xử lý các vấn đề về sự không nhất quán trong dữ liệu, như các giá trị có tên gọi khác nhau cho cùng một khái niệm.

Dữ liệu bẩn có thể làm cho quá trình khai thác dữ liệu trở nên khó khăn và không đáng tin cậy. Do đó, việc làm sạch dữ liệu là rất cần thiết trước khi thực hiện các bước khai thác dữ liệu tiếp theo.

2. Tích hợp dữ liệu (Data Integration)

Trong nhiều trường hợp, dữ liệu được thu thập từ nhiều nguồn khác nhau, chẳng hạn như cơ sở dữ liệu, tệp dữ liệu, hoặc các khối dữ liệu. Việc tích hợp dữ liệu từ các nguồn khác nhau có thể gặp phải một số vấn đề như:

* Khác biệt trong tên gọi thuộc tính: Các thuộc tính đại diện cho một khái niệm có thể có tên gọi khác nhau trong các cơ sở dữ liệu khác nhau, gây ra sự không nhất quán và dư thừa.
* Sự không nhất quán trong giá trị thuộc tính: Các giá trị thuộc tính có thể khác nhau mặc dù chúng thực sự đại diện cho cùng một đối tượng. Ví dụ, tên “Bill” có thể xuất hiện trong cơ sở dữ liệu này, nhưng lại là “William” hoặc “B.” trong cơ sở dữ liệu khác.

Việc xử lý sự không nhất quán và dư thừa trong quá trình tích hợp dữ liệu là rất quan trọng. Ngoài việc làm sạch dữ liệu, cần phải tránh sự dư thừa khi tích hợp dữ liệu để không làm chậm quá trình khai thác tri thức.

3. Giảm kích thước dữ liệu (Data Reduction)

Khi xử lý dữ liệu có kích thước quá lớn, việc giảm dung lượng của dữ liệu mà không làm mất đi giá trị phân tích là rất cần thiết. Các chiến lược giảm dữ liệu bao gồm:

* Giảm chiều dữ liệu (Dimensionality Reduction): Áp dụng các phương pháp mã hóa dữ liệu để giảm kích thước dữ liệu mà không làm mất đi thông tin quan trọng. Các ví dụ bao gồm phân tích thành phần chính (PCA) hoặc biến đổi sóng.
* Giảm số lượng dữ liệu (Numerosity Reduction): Sử dụng các mô hình tham số hoặc không tham số để thay thế dữ liệu gốc bằng các đại diện nhỏ hơn, chẳng hạn như thông qua phân cụm hoặc mẫu.

Giảm dữ liệu giúp làm giảm khối lượng công việc tính toán, đồng thời giữ lại các thông tin cần thiết cho quá trình khai thác dữ liệu.

4. Biến đổi dữ liệu (Data Transformation)

Chuyển đổi dữ liệu bao gồm các quy trình để thay đổi hoặc tái cấu trúc dữ liệu sao cho phù hợp với các thuật toán khai thác dữ liệu. Một số kỹ thuật chuyển đổi dữ liệu bao gồm:

* Chuẩn hóa dữ liệu (Normalization): Các thuộc tính có phạm vi giá trị khác nhau, như tuổi và lương hàng năm, có thể gây ra sự không công bằng trong việc tính toán khoảng cách. Chuẩn hóa giúp các giá trị này nằm trong một phạm vi đồng nhất, ví dụ như [0.0, 1.0].
* Phân biệt hóa dữ liệu (Discretization và Concept Hierarchy Generation): Các giá trị thô của thuộc tính có thể được thay thế bằng các phạm vi hoặc cấp độ khái niệm cao hơn, ví dụ như thay thế giá trị tuổi bằng các nhóm như "thanh niên", "người trưởng thành" hoặc "người cao tuổi".
* Tạo cấp độ khái niệm (Concept Hierarchy Generation): Các thuộc tính có thể được nhóm thành các cấp độ khái niệm cao hơn để giúp phân tích dữ liệu ở các mức độ trừu tượng khác nhau.

Các phương pháp này giúp dữ liệu phù hợp hơn với các thuật toán khai thác dữ liệu như mạng nơ-ron, phân loại gần nhất hoặc phân cụm.

A diagram of data processing

AI-generated content may be incorrect.