Data Integration

Nguồn:

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

I. Tổng quan

Khai phá dữ liệu thường đòi hỏi tích hợp dữ liệu - quá trình hợp nhất thông tin từ nhiều nguồn lưu trữ khác nhau. Việc tích hợp (Integration) cẩn thận giúp giảm thiểu tình trạng dư thừa và mâu thuẫn trong bộ dữ liệu cuối cùng, từ đó nâng cao độ chính xác và tốc độ của quá trình khai phá tiếp theo.

II. Vấn đề nhận diện thực thể (The Entity Identification Problem)

1. Giới thiệu  
Trong quá trình phân tích dữ liệu, nhiệm vụ tích hợp dữ liệu là điều không thể tránh khỏi, khi các nguồn dữ liệu khác nhau được kết hợp lại thành một kho dữ liệu thống nhất như trong trường hợp xây dựng dữ liệu cho kho dữ liệu (data warehousing). Các nguồn dữ liệu này có thể bao gồm nhiều cơ sở dữ liệu, các khối dữ liệu (data cubes) hoặc các tệp dữ liệu dạng phẳng.

2. Vấn đề nhận diện thực thể (Entity Identification Problem)

Một trong những vấn đề quan trọng trong quá trình tích hợp dữ liệu là làm sao để ghép nối các thực thể cùng loại từ các nguồn dữ liệu khác nhau, hay nói cách khác, xác định rằng các thực thể trong thế giới thực được trình bày dưới các tên gọi khác nhau trong các nguồn dữ liệu khác nhau đều đề cập đến cùng một đối tượng.

* Ví dụ:
  + Làm sao để nhà phân tích dữ liệu hoặc hệ thống tự động khẳng định rằng “customer id” trong một cơ sở dữ liệu và “cust number” trong cơ sở dữ liệu khác thực sự chỉ về cùng một thuộc tính.

3. Vai trò của metadata trong xác định thực thể

Để giải quyết vấn đề nhận diện thực thể, metadata đóng vai trò rất quan trọng. Các thông tin metadata của mỗi thuộc tính bao gồm:

* Tên thuộc tính
* Ý nghĩa của thuộc tính
* Kiểu dữ liệu
* Phạm vi giá trị cho phép
* Quy tắc xử lý giá trị trống, số 0 hoặc giá trị null

Những thông tin này giúp xác định và tránh được các lỗi trong quá trình tích hợp sơ đồ (schema integration). Metadata cũng được sử dụng để hỗ trợ quá trình biến đổi dữ liệu, ví dụ: chuyển đổi các mã cho loại hình thanh toán (trong một cơ sở dữ liệu mã có thể là “H” và “S”, trong khi ở cơ sở dữ liệu khác là 1 và 2).

4. Chú ý đặc biệt khi ghép nối thuộc tính giữa các cơ sở dữ liệu  
Khi tiến hành ghép nối thuộc tính từ một cơ sở dữ liệu này sang cơ sở dữ liệu khác, cần lưu ý đến cấu trúc của dữ liệu, nhằm đảm bảo:

* Các phụ thuộc hàm của các thuộc tính (functional dependencies)
* Các ràng buộc tham chiếu (referential constraints)

Ví dụ, trong một hệ thống, chiết khấu có thể được áp dụng cho toàn bộ đơn hàng, trong khi ở hệ thống khác, chiết khấu lại được áp dụng cho từng dòng sản phẩm trong đơn hàng. Nếu không phát hiện kịp thời sự khác biệt này trước khi tích hợp, các mục trong hệ thống đích có thể bị áp dụng chiết khấu không chính xác.

III. Redundancy and Correlation Analysis

1. Giới thiệu   
Dư thừa là một vấn đề quan trọng trong quá trình tích hợp dữ liệu. Một thuộc tính (chẳng hạn “doanh thu hằng năm”) có thể bị coi là dư thừa nếu nó có thể được suy ra từ một hoặc một tập các thuộc tính khác. Bên cạnh đó, sự không nhất quán trong đặt tên thuộc tính hoặc chiều dữ liệu cũng có thể dẫn đến tình trạng dư thừa trong tập dữ liệu thu được sau quá trình tích hợp.

Một số dư thừa có thể được nhận diện nhờ phân tích tương quan (correlation analysis). Khi xét hai thuộc tính, phương pháp này đo lường mức độ một thuộc tính hàm ý giá trị của thuộc tính còn lại, dựa trên dữ liệu sẵn có.

* Nominal data
* Là dữ liệu định tính dùng để đặt tên hoặc gán nhãn cho các biến mà không cung cấp giá trị số (Ví dụ: Màu sắc: Đỏ, Xanh, Vàng.)
* Có thể áp dụng phép kiểm định chi bình phương để đánh giá.
* Thuộc tính số (Numeric attributes):
* Là dữ liệu định lượng, có thể đo lường bằng số
* Sử dụng hệ số tương quan (correlation coefficient) và hiệp phương sai (covariance) để đánh giá mức độ biến thiên của một thuộc tính so với thuộc tính còn lại.

2.Kiểm định tương quan cho nominal data ( Correlation Test for Nominal Data)

a. Tổng quan

Đối với nominal data, việc xác định mối tương quan (correlation) giữa hai thuộc tính có thể được thực hiện thông qua kiểm định tương quan .

Ví dụ, ta có hai thuộc tính và . Giả sử thuộc tính có giá trị phân biệt, ký hiệu và thuộc tính có giá trị phân biệt, ký ​.

Khi đó, các bộ dữ liệu (hay các “tuple”) được mô tả bởi hai thuộc tính và có thể trình bày dưới dạng *contingency table*, trong đó:

* Các giá trị của tạo thành cột,
* Các giá trị của tạo thành hàng.

Mỗi hàng ứng với một giá trị của , mỗi cột ứng với một giá trị của . Bên trong bảng, ô thể hiện cặp kết hợp .

b. Công thức tính giá trị

Giá trị (còn gọi là thống kê Pearson ) được xác định bằng công thức sau:

Trong đó:

* ​ là tần suất quan sát được (observed frequency) của cặp
* là tần suất kỳ vọng (expected frequency) của cặp khi giả định hai thuộc tính độc lập, và được tính như sau:

với

* là tổng số bộ dữ liệu.
* là số lượng bộ dữ liệu có giá trị ​ cho thuộc tính .
* là số lượng bộ dữ liệu có giá trị ​ cho thuộc tính .

Thống kê được dùng để kiểm định giả thuyết rằng và là độc lập, tức là không có mối tương quan giữa chúng. Bài kiểm định này dựa trên một mức ý nghĩa xác định, với bậc tự do. Nếu giả thuyết độc lập bị bác bỏ, điều đó cho thấy A và B có **tương quan thống kê**.

Nếu hai thuộc tính độc lập, ​ (tần suất quan sát được) sẽ xấp xỉ giá trị ​. Nhưng nếu quan sát thực tế khác nhiều so với giá trị kỳ vọng, mỗi sự khác biệt đó sẽ đóng góp vào tổng , làm tăng lên.

c. Ví dụ minh họa

**A table with numbers and a few letters

AI-generated content may be incorrect.**

Giả sử một nhóm gồm 1.500 người được khảo sát. Giới tính của từng người được ghi lại. Mỗi người cũng được hỏi về loại sách họ thích đọc: hư cấu (fiction) hay phi hư cấu (non-fiction). Như vậy, chúng ta có hai thuộc tính: giới tính (gender) và loại sách ưa thích (preferred\_reading).

* Giới tính: male, female,
* Thể loại đọc ưa thích: fiction, non-fiction.

Tần suất quan sát (hoặc số lần xuất hiện) của từng cặp kết hợp có thể xảy ra được tổng hợp trong *contingency table* như minh họa.

Trong đó các con số trong ngoặc là giá trị kỳ vọng, các giá trị kỳ vọng này được tính dựa trên phân phối dữ liệu của cả hai thuộc tính, sử dụng phương trình.

- Tính số lượng bộ dữ liệu cho mỗi thuộc tính

* Tổng quan sát

- Tính tần suất kỳ vọng

Để tiến hành kiểm định , trước hết ta tính tần suất kỳ vọng cho từng ô trong bảng, giả sử hai thuộc tính là độc lập. Công thức như sau:

Cụ thể

* Ô (male, fiction)
* Ô (male, non-fiction)
* Ô (female, fiction)
* Ô (female, non-fiction)

- Tính giá trị

Giá trị (còn gọi là thống kê Pearson ) được tính bằng

Với là tần suất quan sát được

* Ô (male, fiction)
* Ô (male, fiction)
* Ô (male, fiction)
* Ô (male, fiction)
* Tính tổng tất cả các ô

- Xác định bậc tự do và so sánh với giá trị tới hạn

* Bậc tự do (Degree of Freedom): Ở đây và , do đó bậc tự do .
* So sánh với giá trị tới hạn:
* Ở mức ý nghĩa , tra Distribution Table, giá trị tới hạn xấp xỉ.
* Nếu tính được lớn hơn , ta bác bỏ giả thuyết độc lập; ngược lại, nếu nhỏ hơn , ta chưa đủ bằng chứng để bác bỏ giả thuyết

Trong ví dụ này (theo số liệu gốc của sách), giá trị tính được lớn hơn , vì vậy ta có thể kết luận rằng gender và preferred\_reading là **không độc lập** và do đó **tương quan** với nhau ở mức ý nghĩa 0.001.

3. Hệ số tương quan cho dữ liệu số (Correlation Coefficient for Numeric Data)

a. Tổng quan

Khi phân tích dữ liệu số (numeric data), một trong những mục tiêu quan trọng là xác định xem hai thuộc tính có xu hướng biến đổi cùng nhau hay không. Để đo lường mức độ liên hệ đó, chúng ta có thể sử dụng **hệ số tương quan** (Correlation Coefficient), còn được gọi là **hệ số tương quan tích moment****của Pearson** (Pearson’s product moment coefficient), do Karl Pearson phát triển.

b. Công thức tính hệ số tương quan

Giả sử chúng ta có hai thuộc tính số và . Xem xét một tập dữ liệu gồm bộ, với là giá trị của và trong bộ dữ liệu thứ . Khi đó, **hệ số tương quan Pearson** giữa và được xác định bởi công thức:

Trong đó:

* và lần lượt là giá trị trung bình (mean) của và .
* ​ và ​ là độ lệch chuẩn (standard deviation) của và .
* đại diện cho tổng tất cả các tích trên các bộ dữ liệu.
* là tổng số bộ dữ liệu (tuples).

c. Ý nghĩa của hệ số tương quan

* Miền giá trị: Hệ số tương quan ​ nằm trong khoảng .
* Tương quan dương: Nếu , giá trị của tăng sẽ đi kèm với việc cũng tăng; ​ càng gần , mối liên hệ tuyến tính càng mạnh.
* Tương quan âm: Nếu , tăng thì giảm; ​ càng gần , mối liên hệ tuyến tính ngược chiều càng mạnh.
* Gần 0: Nếu , ta có thể kết luận rằng và độc lập và không có mối tương quan giữa chúng.

d. Chú ý

Lưu ý rằng tương quan không đồng nghĩa với quan hệ nhân quả. Nghĩa là, nếu A và B có tương quan, điều đó không có nghĩa là A gây ra B hoặc B gây ra A.

Ví dụ, khi phân tích một cơ sở dữ liệu dân số, ta có thể thấy rằng số lượng bệnh viện và số vụ trộm ô tô trong một khu vực có tương quan với nhau. Tuy nhiên, điều này không có nghĩa là một yếu tố gây ra yếu tố kia. Thực chất, cả hai đều liên quan nhân quả đến một thuộc tính thứ ba, đó là dân số.

IV. Hiệp phương sai (Covariance of Numeric Data)

1. Giới thiệu

Trong lĩnh vực xác suất và thống kê, **hệ số tương quan (correlation)** và **hiệp phương sai (covariance)** là hai thước đo quan trọng được sử dụng để đánh giá xem hai thuộc tính số thay đổi cùng nhau nhiều hay ít.

Để làm rõ ý tưởng này, chúng ta sẽ xem xét hai thuộc tính số và trên một tập quan sát có bộ dữ liệu . Giá trị trung bình của và được ký hiệu lần lượt là và , hay còn gọi là kỳ vọng của và .

2.Định nghĩa hiệp phương sai (Covariance)

Hiệp phương sai giữa và được định nghĩa như sau:

Trong đó:

* và là giá trị trung bình (kỳ vọng) của và .

Hiệp phương sai giữa và cũng có thể được tính bằng công thức

Phương trình này sẽ làm đơn giản việc tính toán

Ý nghĩa của hiệp phương sai:

* Nếu , khi tăng thì thường có xu hướng tăng.
* Nếu , khi tăng thì thường có xu hướng giảm (ngược chiều).
* Nếu và là độc lập, tức là chúng không có tương quan với nhau, thì:

Khi đó, hiệp phương sai:

Tuy nhiên, **điều ngược lại không đúng**. Một số cặp biến ngẫu nhiên có thể có hiệp phương sai bằng 0 nhưng **vẫn không độc lập**. Chỉ khi có thêm giả định nhất định (ví dụ: dữ liệu tuân theo phân phối chuẩn đa biến), thì hiệp phương sai bằng 0 mới đồng nghĩa với tính độc lập.

3. Ví dụ minh họa

Bảng dưới đây minh họa một ví dụ thực tế về giá cổ phiếu của hai công ty AllElectronics và HighTech tại 5 thời điểm khác nhau.

Ta đặt câu hỏi: “Liệu giá cổ phiếu của hai công ty này có bị ảnh hưởng bởi cùng một xu hướng ngành (nghĩa là tăng giảm theo cùng chiều) hay không?”

A table with numbers and text

AI-generated content may be incorrect.

- Tính giá trị trung bình (kỳ vọng)

* Giả sử giá cổ phiếu (theo thời điểm) của AllElectronics là 6, 5, 4, 3, 2, thì:
* Tương tự, giả sử giá cổ phiếu HighTech là 20, 10, 14, 5, 5, thì:

- Tính hiệp phương sai

* Dùng Phương trình ở trên:
* Khi đó

Vì vậy, dựa vào hiệp phương sai dương, ta có thể kết luận rằng giá cổ phiếu của cả hai công ty có xu hướng tăng **cùng nhau**.

4. Mối liên hệ giữa hiệp phương sai và hệ số tương quan

Hệ số tương quan (correlation coefficient) giữa và có thể được tính từ hiệp phương sai, bằng công thức:

Trong đó ​ và ​ lần lượt là độ lệch chuẩn của và .

V. Trùng lặp dữ liệu (Tuple Duplication)

1. Giới thiệu

Trong quá trình phát hiện các dạng dư thừa dữ liệu, bên cạnh việc xem xét trùng lặp ở mức độ thuộc tính, ta cũng cần chú ý đến hiện tượng trùng lặp dữ liệu (tuple duplication). Hiện tượng này xảy ra khi có từ hai bộ dữ liệu trở lên giống hệt nhau (hay gần như giống hệt nhau) cho cùng một trường hợp nhập liệu duy nhất.

Việc sử dụng bảng phi chuẩn hoá (denormalized tables)—thường nhằm tối ưu hóa hiệu năng bằng cách tránh sử dụng các phép nối (joins)—là một trong những nguyên nhân gây nên dư thừa dữ liệu.

Các điểm không nhất quán (inconsistencies) thường nảy sinh do:

* Nhập dữ liệu không chính xác hoặc sai sót trong quá trình thu thập thông tin.
* Cập nhật dữ liệu không đồng bộ, chỉ cập nhật một vài vị trí mà bỏ qua những vị trí khác có cùng bản sao dữ liệu.

2. Ví dụ

Giả sử một cơ sở dữ liệu đơn đặt hàng (purchase order) chứa các thuộc tính về tên và địa chỉ của người mua. Thay vì chỉ lưu một khoá (key) trỏ đến thông tin này trong một cơ sở dữ liệu khách hàng riêng, người thiết kế lại đưa trực tiếp tên và địa chỉ của người mua vào nhiều bản ghi đơn đặt hàng khác nhau.

Hậu quả:

* Nếu cùng một khách hàng thay đổi địa chỉ, và việc cập nhật chỉ được thực hiện ở một số đơn đặt hàng (nhưng không phải tất cả), sự không nhất quán xuất hiện.
* Một khách hàng có thể “xuất hiện” với nhiều địa chỉ khác nhau trong cùng một cơ sở dữ liệu đơn đặt hàng, dù thực tế chỉ nên có một địa chỉ hiện tại.

VI. Phát hiện và giải quyết xung đột giá trị dữ liệu (Detection and Resolution of Data Value Conflicts)

1. Giới thiệu

Khi tiến hành tích hợp dữ liệu từ nhiều nguồn, ta không chỉ đối mặt với vấn đề nhận diện thực thể hay dư thừa dữ liệu, mà còn phải xử lý xung đột giá trị dữ liệu (data value conflicts). Đây là những trường hợp mà, đối với cùng một thực thể trong thế giới thực, các giá trị của thuộc tính từ những nguồn khác nhau có thể không thống nhất.

2. Xung đột giá trị dữ liệu

Trong thực tế, cùng một thực thể trong thế giới thực có thể có các giá trị thuộc tính khác nhau khi đến từ những nguồn dữ liệu khác nhau. Các xung đột này có thể xuất phát từ khác biệt trong cách biểu diễn, thang đo hoặc mã hóa.

Ví dụ:

* Thuộc tính cân nặng (weight) có thể được lưu dưới đơn vị hệ mét trong một hệ thống, nhưng lại được lưu dưới đơn vị hệ Anh trong một hệ thống khác.
* Giá phòng khách sạn trong cùng một chuỗi khách sạn nhưng ở các thành phố khác nhau có thể không chỉ khác nhau về đơn vị tiền tệ, mà còn về dịch vụ kèm theo (ví dụ bữa sáng miễn phí) và thuế suất áp dụng.

3. Khác biệt giữa các hệ thống tổ chức

Khi trao đổi thông tin giữa các trường học, mỗi trường có thể có chương trình đào tạo và hệ thống đánh giá riêng biệt:

* Một trường đại học có thể sử dụng hệ thống theo quý, giảng dạy ba môn học về cơ sở dữ liệu và cho điểm theo thang từ A+ đến F.
* Một trường khác có thể dùng hệ thống học kỳ, chỉ có hai môn học về cơ sở dữ liệu và cho điểm theo thang điểm từ 1 đến 10.

Việc xây dựng quy tắc chuyển đổi giữa các khóa học và hệ thống điểm số giữa hai trường là rất khó, dẫn đến khó khăn trong việc trao đổi thông tin.

4. Khác biệt về mức độ trừu tượng

Các thuộc tính cũng có thể khác nhau về mức độ trừu tượng. Ví dụ:

* Trong một cơ sở dữ liệu, doanh số bán hàng tổng có thể ám chỉ một chi nhánh cụ thể của chuỗi All Electronics.
* Trong khi đó, cùng tên thuộc tính trong một cơ sở dữ liệu khác có thể đại diện cho tổng doanh số của tất cả cửa hàng trong khu vực.

Những khác biệt này tạo ra sự không thống nhất trong biểu diễn dữ liệu, làm phức tạp thêm quá trình tích hợp.