ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

TRẦN KHÁNH

KHAI PHÁ PHỤ THUỘC HÀM XẤP XỈ SỬ DỤNG PHỦ TỐI THIỂU VÀ LỚP TƯƠNG ĐƯƠNG

Chuyên ngành: Khoa học máy tính
Mã số: 60 48 01

TÓM TẮT LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

Thái Nguyên - 2015

MỤC LỤC

DANH MỤC VIẾT TẮT VÀ KÍ HIỆU SỬ DỤNG TRONG LUẬN VĂN

Ký hiệu	Diễn giải
$R(U)$	Quan hệ trên tập thuộc *tính U*
$U = \{A_1, ..., A_m\}$	Tập m thuộc tính.
$S = (U, F)$	Lược đồ quan hệ với U là tập thuộc tính , F là tập các phụ thuộc hàm trên U
LĐQH	Lược đồ quan hệ
CSDL	Cơ sở dữ liệu
PTH	Phụ thuộc hàm

DANH MỤC CÁC BẢNG BIỂU

DANH MỤC CÁC HÌNH VẼ

MỞ ĐẦU

1. Đặt vấn đề

Trong những năm gần đây, Công nghệ thông tin (CNTT) phát triển mạnh mẽ đã tác động đến mọi mặt của xã hội, những thành tựu của công nghệ lưu trữ đã cho phép tạo ra những nguồn dữ liệu khổng lồ. Việc khai thác các nguồn dữ liệu này ngày càng cấp thiết, đặt ra những thách thức lớn cho ngành CNTT, đặc biệt là lĩnh vực khai phá dữ liệu. Với nguồn dữ liệu lớn như vậy thì việc tìm kiếm, phân tích, xử lý và đưa ra các thông tin cần thiết, phù hợp với thời gian và yêu cầu là điều không dễ dàng.

Các phương pháp khai thác cơ sở dữ liệu truyền thống ngày càng không đáp ứng được nhu cầu thực tế này. Vì vậy các phương pháp nghiên cứu, tiếp cận với những công cụ cho phép phân tích, tổng hợp, khai phá tri thức từ dữ liệu một cách thông minh, hiệu quả đã được nhiều nhà khoa học quan tâm nghiên cứu.

Khái niệm phụ thuộc hàm đóng một vai trò rất quan trọng trong lý thuyết cơ sở dữ liệu quan hệ. Các phụ thuộc hàm rất hữu ích trong việc phân tích và thiết kế cơ sở dữ liệu quan hệ như xác định khóa, xác định các dạng chuẩn, các vấn đề về nhất quán dữ liệu ... Tuy nhiên trong thực tế do có một số giá trị dữ liệu không chính xác hoặc một số ngoại lệ nào đó làm cho các phụ thuộc hàm không thỏa. Sự phụ thuộc tuyệt đối này dường như quá nghiêm ngặt khi ta hình dung tới một quan hệ có hàng nghìn bộ, trong khi đó chỉ có khoảng vài bộ vi phạm phụ thuộc hàm. Bỏ qua các phụ thuộc hàm này sẽ làm mất tính chất phụ thuộc vốn có giữa các thuộc tính. Vì vậy các nhà nghiên cứu đã mở rộng khái niệm phụ thuộc hàm thành phụ thuộc hàm xấp xỉ theo một cách thức, một nghĩa nào đó, các phụ thuộc hàm xấp xỉ (Approximate Functional Dependencies - AFDs) này cho phép có một số lượng lỗi nhất định của các bộ dữ liệu đối với phụ thuộc hàm.

Phụ thuộc hàm xấp xỉ được khai phá từ CSDL quan hệ biểu diễn các mối

quan hệ có ý nghĩa, có nhiều ứng dụng khác nhau như: Dự đoán giá trị thiếu thuộc tính trong bảng quan hệ bằng cách sử dụng các giá trị của các thuộc tính trong việc xác định tập hợp các AFDs, tối ưu hóa truy vấn, viết lại câu truy vấn, chuẩn hóa cơ sở dữ liệu để cho hiệu suất tốt hơn và thiết kế lưu trữ hiệu quả hơn,…

Luận văn sẽ tìm hiểu về phụ thuộc hàm xấp xỉ và nghiên cứu thuật toán AFDMCEC, một thuật toán mới tìm các phụ thuộc hàm xấp xỉ trong các CSDL lớn dựa trên độ đo xấp xỉ. Thuật toán này sử dụng một số khái niệm trong lý thuyết thiết kế CSDL quan hệ, đặc biệt là các khái niệm phủ tối thiểu và lớp tương đương.

2. Đối tượng và phạm vi nghiên cứu

Luận văn tìm hiểu tổng quan về khai phá dữ liệu, đi sâu tìm hiểu khái niệm phụ thuộc hàm, phụ thuộc hàm xấp xỉ và các tính chất, độ đo lỗi của phụ thuộc hàm xấp xỉ, từ đó nghiên cứu thuật toán TANE sửa đổi và thuật toán AFDMCEC tìm phụ thuộc hàm xấp xỉ.

3. Hướng nghiên cứu của đề tài

- Tìm hiểu về phụ thuộc hàm, phụ thuộc hàm xấp xỉ và các độ đo lỗi của chúng.

- Nghiên cứu về thuật toán khai phá phụ thuộc hàm xấp xỉ từ bảng quan hệ.

4. Phương pháp nghiên cứu

Phương pháp nghiên cứu chính của luận văn là nghiên cứu lý thuyết kết hợp với đánh giá thực nghiệm, cụ thể là: Phân tích, tổng hợp các kết quả nghiên cứu về phụ thuộc hàm, phụ thuộc hàm xấp xỉ, … đã công bố trên các bài báo khoa học, hội thảo chuyên ngành trong và ngoài nước. Từ đó, trình bày làm rõ vấn đề khai phá phụ thuộc hàm xấp xỉ sử dụng phủ tối thiểu và lớp tương đương.

5. Ý nghĩa khoa học và thực tiễn

Phụ thuộc hàm đóng vai trò quan trọng trong lý thuyết CSDL quan hệ. Tuy nhiên, trong thực tế do có một số giá trị dữ liệu không chính xác hoặc một số ngoại lệ nào đó, làm cho các phụ thuộc hàm không thỏa mãn. Sự phụ thuộc tuyệt đối này dường như quá nghiêm ngặt khi ta hình dung một quan hệ có hàng nghìn bộ, trong khi đó chỉ có vài bộ vi phạm phụ thuộc hàm. Do vậy, mở rộng khái niệm phụ thuộc hàm thành phụ thuộc hàm xấp xỉ, cho phép có một số lỗi nhất định của các bộ dữ liệu, là rất cần thiết và có ý nghĩa cả về mặt lý thuyết cũng như thực tiễn.

Các phụ thuộc hàm xấp xỉ không những giúp chúng ta thấy được mối quan hệ tiềm ẩn giữa các thuộc tính mà còn giúp ta thuận tiện hơn trong việc phân tích dữ liệu, đánh giá thông tin.

Phát hiện phụ thuộc hàm xấp xỉ trong CSDL là một vấn đề nghiên cứu hấp dẫn và cũng là một trong những mục tiêu của phát hiện tri thức. Tiếp cận phụ thuộc hàm xấp xỉ sử dụng phủ tối thiểu và lớp tương đương của khai phá dữ liệu là một hướng đi thú vị, hứa hẹn nhiều kết quả và ứng dụng hiệu quả trong thực tiễn.

6. Cấu trúc luận văn:

Luận văn được trình bày trong 3 chương:

Chương 1: Tổng quan về khai phá dữ liệu và khai phá phụ thuộc hàm, phụ thuộc hàm xấp xỉ.

Chương 2: Thuật toán khai phá phụ thuộc hàm xấp xỉ sử dụng phủ tối thiểu và lớp tương đương.

Chương 3: Thực nghiệm khai phá phụ thuộc hàm xấp xỉ.

Cuối cùng là kết luận của luận văn và tài liệu tham khảo.

CHƯƠNG 1
TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU VÀ KHAI PHÁ PHỤ THUỘC HÀM, PHỤ THUỘC HÀM XẤP XỈ

1.1. Khai phá dữ liệu

1.1.1. Khám phá tri thức và khai phá dữ liệu

Khai phá dữ liệu (KPDL) là việc rút trích tri thức một cách tự động và hiệu quả từ một khối dữ liệu lớn. Tri thức đó thường ở dạng các mẫu có tính chất không tầm thường, không tường minh (ẩn), chưa được biết đến và có tiềm năng mang lại lợi ích. Có một số nhà nghiên cứu còn gọi KPDL là phát hiện tri thức từ cơ sở dữ liệu (Knowledge Discovery in Database – KDD). Ở đây chúng ta có thể coi KPDL là cốt lõi của quá trình phát hiện tri thức.

Quá trình phát hiện tri thức gồm các bước:

Bước 1: ***Trích chọn dữ liệu (data selection)***: Là bước trích chọn những tập dữ liệu cần được khai phá từ các tập dữ liệu lớn (databases, data ware houses).

Bước 2: ***Tiền xử lý dữ liệu (data preprocessing)*****:** Là bước làm sạch dữ liệu (xử lý dữ liệu không đầy đủ, dữ liệu nhiễu, dữ liệu không nhất quán,…v.v), rút gọn dữ liệu (sử dụng các phương pháp thu gọn dữ liệu, histograms, lấy mẫu…v.v), rời rạc hóa dữ liệu (dựa vào histograms, entropy, phân khoảng,...v.v). Sau bước này, dữ liệu sẽ nhất quán, đầy đủ, được rút gọn và được rời rạc hóa.

Bước 3: ***Biến đổi dữ liệu (data transformation)***: Là bước chuẩn hóa và làm mịn dữ liệu để đưa dữ liệu về dạng thuận lợi nhất nhằm phục vụ cho các kỹ thuật khai thác ở bước sau.

Bước 4: ***Khai phá dữ liệu (data mining)*****:** Đây là bước quan trọng và tốn nhiều thời gian nhất của quá trình khám phá tri thức, áp dụng các kỹ thuật khai phá (phần lớn là các kỹ thuật của machine learning) để khai phá, trích chọn được các mẫu (pattern) thông tin, các mối liên hệ đặc biệt trong dữ liệu.