ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CNTT&TT

TÌM HIỂU MỘT SỐ THUẬT TOÁN KHAI PHÁ TẬP MỤC LỢI ÍCH CAO VÀ ỨNG DỤNG

Vũ Anh Đức

Người hướng dẫn: TS Nguyễn Huy Đức

Thái Nguyên- năm 2016

MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG BIỂU

LỜI CẢM ƠN

Lời đầu tiên tôi xin gửi lời cảm ơn chân thành và biết ơn sâu sắc tới TS. Nguyễn Huy Đức – Trường Cao đẳng Sư phạm Trung ương, người đã chỉ bảo và hướng dẫn tận tình cho tôi trong suốt quá trình nghiên cứu khoa học và thực hiện luận văn này.

Tôi xin chân thành cảm ơn sự dạy bảo, giúp đỡ, tạo điều kiện và khuyến khích tôi trong quá trình học tập và nghiên cứu của các thầy cô giáo của Viện Công nghệ Thông tin, Trường Đại học Công nghệ Thông tin và Truyền thông – Đại học Thái Nguyên.

Và cuối cùng, tôi xin gửi lời cảm ơn tới gia đình, người thân và bạn bè – những người luôn ở bên tôi những lúc khó khăn nhất, luôn động viên tôi, khuyến khích tôi trong cuộc sống và trong công việc. Tôi xin chân thành cảm ơn!

Thái Nguyên, ngày 10 tháng 07 năm 2016

Tác giả

Vũ Anh Đức

LỜI CAM ĐOAN

Tôi xin cam đoan Luận văn "***Tìm hiểu một số thuật toán khai phá tập mục lợi ích cao và ứng dụng***" là công trình nghiên cứu của riêng tôi dưới sự hướng dẫn của **TS. Nguyễn Huy Đức**. Kết quả đạt được trong luận văn là sản phẩm của riêng cá nhân tôi, không sao chép lại của người khác. Trong toàn bộ luận văn, những điều được trình bày là của cá nhân hoặc là được tổng hợp từ nhiều nguồn tài liệu. Tất cả các tài liệu tham khảo đều có xuất xứ rõ ràng và được trích dẫn hợp pháp.

Tôi xin chịu hoàn toàn trách nhiệm và chịu mọi hình thức kỷ luật theo quy định cho lời cam đoan của mình.

Thái Nguyên, ngày 10 tháng 07 năm 2016

Người cam đoan

Vũ Anh Đức

LỜI MỞ ĐẦU

Khai phá tập mục phổ biến có vai trò quan trọng trong nhiều nhiệm vụ khai phá dữ liệu. Khai phá tập mục phổ biến xuất hiện như là bài toán con của nhiều lĩnh vực khai phá dữ liệu như khám phá luật kết hợp, khám phá mẫu tuần tự… Bài toán khai phá luật kết hợp do Agrawal, T.Imielinski và A. N. Swami [3] đề xuất và nghiên cứu lần đầu vào năm 1993 với mục tiêu là phát hiện các tập mục phổ biến, từ đó tạo các luật kết hợp. Trong mô hình của bài toán này, giá trị của mỗi mục dữ liệu trong một giao tác là 0 hoặc 1, tức là chỉ quan tâm mục dữ liệu có xuất hiện trong giao tác hay không. Bài toán cơ bản này có nhiều ứng dụng, tuy vậy, do tập mục phổ biến chỉ mang ngữ nghĩa thống kê nên nó chỉ đáp ứng được phần nào nhu cầu của thực tiễn.

Nhằm khắc phục hạn chế của bài toán cơ bản khai phá luật kết hợp, nhiều nhà nghiên cứu đã mở rộng bài toán theo nhiều hướng khác nhau. Năm 1997, Hilderman và các cộng sự đề xuất bài toán khai phá tập mục cổ phần cao. Trong mô hình này, giá trị của mục dữ liệu trong giao tác là một số. Năm 2004, nhóm các nhà nghiên cứu H. Yao, Hamilton và Butz [9], mở rộng tiếp bài toán, đề xuất mô hình khai phá tập mục lợi ích cao.

Trong mô hình khai phá tập mục lợi ích cao, giá trị của mục dữ liệu trong giao tác là một số (như số lượng đã bán của mặt hàng, gọi là giá trị khách quan), ngoài ra còn có bảng lợi ích cho biết lợi ích mang lại khi bán một đơn vị hàng đó (gọi là giá trị chủ quan). Lợi ích của tập mục là số đo lợi nhuận mà tập mục đó mang lại. Khai phá tập mục lợi ích cao là khám phá tất cả các tập mục có lợi ích không nhỏ hơn ngưỡng lợi ích tối thiểu của người sử dụng.

Trong những năm gần đây, bài toán này đã và đang thu hút sự quan tâm của nhiều nhà nghiên cứu trong và ngoài nước. Với mục đích tìm hiểu bài toán tìm tập mục lợi ích cao và các thuật toán khai phá hiệu quả gần đây, em đã quyết định lựa chọn đề tài "**Tìm hiểu một số thuật toán khai phá tập mục lợi ích cao và ứng dụng**".

Nội dung luận văn gồm 3 chương:

Chương 1: Tổng quan về khai phá dữ liệu và khai phá tập mục phổ biến

Chương 2: Một số thuật toán hiệu quả khai phá tập mục lợi ích cao.

Chương 3: Chương trình thực nghiệm.

Chương I: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU VÀ KHAI PHÁ TẬP MỤC PHỔ BIẾN

1.1 Khái niệm về khai phá tri thức và 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 trong 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 [4]:

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.

Bước 5: Đánh giá và biểu diễn tri thức (knowledge representation & evaluation): Dùng các kỹ thuật hiển thị dữ liệu để trình bày các mẫu thông tin