ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

NGUYỄN THỊ HẢI LÝ

KHAI PHÁ LUẬT KẾT HỢP HIẾM TRÊN CƠ SỞ DỮ LIỆU VÀ ỨNG DỤNG

LUẬN VĂN THẠC SĨ

CHUYÊN NGÀNH KHOA HỌC MÁY TÍNH

Thái Nguyên - 2015

MỤC LỤC

LỜI CAM ĐOAN i
LỜI CẢM ƠN ii
MỤC LỤC iii
DANH MỤC CÁC KÝ HIỆU, VIẾT TẮT iv
DANH MỤC CÁC BẢNG BIỂU v
DANH MỤC CÁC HÌNH VẼ vi
MỞ ĐẦU.....1
CHƯƠNG 1 3
KHAI PHÁ DỮ LIỆU VÀ BÀI TOÁN KHAI PHÁ LUẬT KẾT HỢP 3
1.1. Khai phá dữ liệu 3
1.1.1. Quá trình phát hiện tri thức từ cơ sở dữ liệu.....3
1.1.2. Kiến trúc của hệ thống khai phá dữ liệu..**Error! Bookmark not defined.**
1.1.3. Quá trình khai phá dữ liệu 5
1.1.4. Nhiệm vụ của khai phá dữ liệu. .....**Error! Bookmark not defined.**
1.1.5. Các ứng dụng của khai phá dữ liệu 6
1.2. Khai phá luật kết hợp trong cơ sở dữ liệu 7
1.2.1. Bài toán mở đầu.....**Error! Bookmark not defined.**
1.2.2. Các khái niệm cơ sở.....7
1.2.2.1. Cơ sở dữ liệu giao tác.....7
1.2.2.2. Tập mục phổ biến.....8
1.2.2.3. Luật kết hợp 8
1.2.3. Khai phá luật kết hợp.....9
1.2.4. Các cách tiếp cận khai phá tập mục phổ biến..... **Error! Bookmark not defined.**
1.2.5. Các thuật toán điển hình khai phá tập mục phổ biến.....10
1.2.5.1 Thuật toán Apriori.....10
1.2.5.2.Thuật toán FP_growth.....13
1.2.6. Thuật toán sinh luật kết hợp: 14

1.2.7. Một số mở rộng khai phá luật kết hợp 17
Kết luận chương 1 18
Chương 2: LUẬT KẾT HỢP HIẾM 19
2.1. Giới thiệu chung về luật kết hợp hiếm. 19
2.2. Một số hướng nghiên cứu chính phát hiện luật kết hợp hiếm 20
2.2.1. Sử dụng ràng buộc phần hệ quả của luật 20
2.2.2. Thiết lập đường biên phân chia các tập phổ biến và không phổ biến 21
2.2.3. Phát hiện luật kết hợp hiếm từ các CSDL định lượng 22
2.3. Khuynh hướng nghiên cứu về luật hiếm 23
2.4. Phát hiện luật kết hợp hiếm Sporadic trên CSDL giao tác 24
2.4.1. Khái niệm về luật hiếm Sporadic 24
2.4.2.Thuật toán Apriori-Inverse 27
2.4.3. Thuật toán tìm tập Sporadic tuyệt đối hai ngưỡng đóng 32
2.4.3.1. Tập Sporadic tuyệt đối hai ngưỡng 33
2.4.3.2. Thuật toán MCPSI tìm tập Sporadic tuyệt đối hai ngưỡng đóng 35
Kết luận chương 2 38
Chương 3 38
THỰC NGHIỆM TÌM LUẬT HIẾM SPORADIC TUYỆT ĐỐI 38
3.1. Giới thiệu bài toán 39
3.2. Dữ liệu thực nghiệm 40
3.3. Xây dựng chương trình 42
3.4. Thực nghiệm khai phá 43
3.5. Kết quả thực nghiệm 47
KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 49
TÀI LIỆU THAM KHẢO 50

LỜI CAM ĐOAN

Tôi xin cam đoan Luận văn "KHAI PHÁ LUẬT KẾT HỢP HIẾM TRÊN CƠ SỞ DỮ LIỆU 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 trong luận văn 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 trính 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 tháng năm 2015

Người cam đoan

Nguyễn Thị Hải Lý

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, Thầy 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 cám 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 khuyết khích tôi trong cuộc sống và trong công việc.

Tôi xin chân thành cảm ơn!

Thái Nguyên, ngày tháng năm 2015

Tác giả

Nguyễn Thị Hải Lý

DANH MỤC CÁC KÝ HIỆU VÀ CÁC CHỮ VIẾT TẮT

Ký hiệu	Diễn giải
KPDL	Khai phá dữ liệu
CSDL	Cơ sở dữ liệu
DB	Cơ sở dữ liệu giao tác
TID	Định danh của giao tác
I	Tập các mục dữ liệu
T	Giao tác (transaction)
C_k	Tập các ứng viên là tập mục có k mục dữ liệu
L_k	Tập các tập mục phổ biến có k mục dữ liệu
k-itemset	Tập mục gồm k mục
BFS	Breadth First Search (Duyệt theo chiều rộng)
DFS	Depth First Search (Duyệt theo chiều sâu)
FP-growth	Frequent-Pattern Growth
FP-tree	Frequent pattern tree
Sup	Độ hỗ trợ (support)
Conf	Độ tin cậy (Confiden)
Minsup	Ngưỡng hỗ trợ tối thiểu
Minconf	Ngưỡng tin cậy tối thiểu

DANH MỤC CÁC BẢNG BIỂU

Bảng 1.1: Danh mục các tập mục trong CSDL**Error! Bookmark not defined.**
Bảng 1.2: Biểu diễn ngang của CSDL giao tác.**Error! Bookmark not defined.**
Bảng 1.3: Biểu diễn dọc của CSDL giao tác.. **Error! Bookmark not defined.**
Bảng 1.4: Ma trận giao tác của CSDL bảng 1.2**Error! Bookmark not defined.**
Bảng 1.5: Cơ sở dữ liệu DB ..... **Error! Bookmark not defined.**
Bảng 1.6 : Độ hỗ trợ của các mục ..... **Error! Bookmark not defined.**
Bảng 1.7: Độ hỗ trợ của các tập mục ..... **Error! Bookmark not defined.**
Bảng 1.8: Độ tin cậy của các luật..... **Error! Bookmark not defined.**
Bảng 1.10: Cơ sở dữ liệu minh họa thực hiện thuật toán COFI-tree.**Error! Bookmark n**
Bảng 1.11 : Các mục dữ liệu và độ hỗ trợ..... **Error! Bookmark not defined.**
Bảng 1.12 : Các mục dữ liệu phổ biến đã sắp thứ tự.**Error! Bookmark not defined.**
Bảng 1.13 : Các mục dữ liệu trong giao tác giảm dần theo độ hỗ trợ.**Error! Bookmark**
Bảng 2.1 : Ví dụ CSDL giao tác D cho thuật toán Apriori-Inverse 29
Bảng 2.2 : Biểu diễn dọc của CSDL D trong bảng 2.1 29
Bảng 2.3 : Độ hỗ trợ của từng mục dữ liệu của CSDL D 30
Bảng 2.4 : Các mục sporadic và độ hỗ trợ 30
Bảng 2.5 : Các 2- tập mục ứng viên 31
Bảng 2.6 : Các tập mục sporadic tuyệt đối..... 31
Bảng 2.7 : Các luật sporadic tuyệt đối..... 32
Bảng 2.8: CSDL giao tác minh họa thuật toán MCPSI..... 37
Bảng 3.1: Dữ liệu đã trích chọn để khai phá 40
Bảng 3.2: Mã hóa các mặt hàng 40

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

Hình 1.1. Quá trình khám phá tri thức 4
Hình 1.2. Kiến trúc của hệ thống khai phá dữ liệu**Error! Bookmark not defined.**
Hình 1.3: Quá trình khai phá dữ liệu..... 6
Hình 1.4 : Phân loại các thuật toán khai phá tập mục phổ biến.**Error! Bookmark not de**
Hình 1.5: Cây FP-tree của CSDL bảng 1.10... **Error! Bookmark not defined.**
Hình 1.6: Cây COFI-tree của mục D. ..... **Error! Bookmark not defined.**
Hình 1.7: Các bước khai phá cây D-COFI-tree.**Error! Bookmark not defined.**
Hình 2.1: Minh họa tìm các tập Sporadic tuyệt đối hai ngưỡng đóng..... 38
Hình 3.1: Dữ liệu đã mã hóa chuẩn bị cho khai phá..... 42
Hình 3.2: Giao diện chương trình 43
Hình 3.3: Giao diện chương trình tìm tập Sporadic tuyệt đối 44
Hình 3.4: Kết quả tìm tập Sporadic tuyệt đối 45
Hình 3.5: Giao diện chương trình tìm luật Sporadic tuyệt đối 46
Hình 3.6: Kết quả tìm luật Sporadic tuyệt đối 47

MỞ ĐẦU

1. Đặt vấn đề

Trong lĩnh vực khai phá dữ liệu (data mining), luật kết hợp (association rule) được dùng để chỉ mối quan hệ kiểu "điều kiện→hệ quả" giữa các phần tử dữ liệu (chẳng hạn, sự xuất hiện của tập mặt hàng này "kéo theo" sự xuất hiện của tập mặt hàng khác) trong một tập bao gồm nhiều đối tượng dữ liệu (chẳng hạn, các giao dịch mua hàng)…. Phát hiện luật kết hợp là phát hiện các mối quan hệ đó trong phạm vi của một tập dữ liệu đã cho. Bài toán phát hiện luật kết hợp được Rakesh Agrawal và cộng sự giới thiệu lần đầu tiên vào năm 1993[4] và nhanh chóng trở thành một trong những hướng nghiên cứu quan trọng của khai phá dữ liệu, đặc biệt trong những năm gần đây.

Phát hiện luật kết hợp [5, 10] đã được ứng dụng thành công trong nhiều lĩnh vực kinh tế-xã hội khác nhau như: thương mại, y tế, sinh học, tài chính - ngân hàng,…Hiện tại, nhiều khuynh hướng nghiên cứu và ứng dụng liên quan đến phát hiện luật kết hợp đã và đang tiếp tục được hình thành.

Một trong những vấn đề về phát hiện luật kết hợphiện đang nhận được nhiều quan tâm của các nhà nghiên cứu là phát hiện luật kết hợp hiếm. Luật kết hợp hiếm là những luật kết hợp ít xảy ra. Mặc dù tần suất xảy ra thấp, nhưng trong nhiều trường hợp, các luật này lại rất có giá trị.

Từ những yêu cầu, thực tế trên, em đã chọn đề tài **"*Khai phá luật kết hợp hiếm trên cơ sở dữ liệu và ứng dụng*".**

2. Đối tượng và phạm vi nghiên cứu

Nghiên cứu các phương pháp, thuật toán khai phá luật kết hợp, đi sâu vào bài toán phát hiện luật kết hợp hiếm thuộc lĩnh vực phát hiện tri thức từ dữ liệu và ứng dụng.

Luận văn tìm hiểu luật kết hợp hiếm Sporadic, trong hai loại của luật hiếm Sporadic là luật hiếm Sporadic tuyệt đốivà luật hiếm Sporadic không tuyệt đối, luận văn đi sâu tìm hiểu luật hiếm Sporadic tuyệt đối trên cơ sở dữ liệu giao tác.

3. Hướng nghiên cứu của đề tài