

**ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**



LÊ THỊ TUYẾT NHUNG

Tên đề tài:

**KHAI PHÁ TẬP MỤC THƯỜNG XUYÊN ĐÓNG
TRONG CƠ SỞ DỮ LIỆU VÀ ỨNG DỤNG**

Chuyên ngành: KHOA HỌC MÁY TÍNH

Mã số : 60.48.01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Hướng dẫn khoa học: TS. NGUYỄN HUY ĐỨC

Thái Nguyên - 2013

LỜI CẢM ƠN

Trước hết em xin gửi lời cảm ơn đến TS. Nguyễn Huy Đức, người thầy đã hướng dẫn em rất nhiều trong suốt quá trình tìm hiểu, nghiên cứu và hoàn thành luận văn tốt nghiệp từ lý thuyết đến ứng dụng. Sự hướng dẫn của thầy đã giúp em có thêm được những hiểu biết khai phá dữ liệu và ứng dụng của nó.

Đồng thời em cũng xin chân thành cảm ơn các thầy cô trong trường cũng như các thầy cô ở Viện Khoa học và công nghệ Việt Nam đã tận tình giảng dạy, trang bị cho em những kiến thức cơ bản cần thiết để em có thể hoàn thành tốt luận văn.

Em xin gửi lời cảm ơn đến gia đình, bạn bè đã tạo mọi điều kiện thuận lợi để em có thể xây dựng thành công luận văn này.

Thái Nguyên, tháng 06 năm 2013

Học viên

Lê Thị Tuyết Nhung

LỜI CAM ĐOAN

Tôi xin cam đoan đề tài “*Khai phá tập mục thường xuyên đóng trong cơ sở dữ liệu và ứng dụng*” là công trình nghiên cứu của bản thân tôi. Các số liệu và kết quả nghiên cứu nêu trong luận văn này là trung thực, được các tác giả cho phép sử dụng và các tài liệu tham khảo như đã trình bày trong luận văn. Tôi xin chịu trách nhiệm về luận văn của mình.

MỤC LỤC

Lời cảm ơn	i
Lời cam đoan	ii
MỤC LỤC	iii
DANH MỤC CÁC KÝ HIỆU VÀ CÁC CHỮ VIẾT TẮT	v
DANH MỤC CÁC BẢNG	vi
DANH MỤC HÌNH VẼ	vii
MỞ ĐẦU	1
Chương 1: Tổng quan về khai phá dữ liệu và khai phá tập mục thường xuyên	2
1.1 Khái niệm về khai phá tri thức và khai phá dữ liệu	2
1.2 Kiến trúc của hệ thống khai phá dữ liệu	4
1.3 Quá trình khai phá dữ liệu	5
1.4 Một số kỹ thuật khai phá dữ liệu	8
1.4.1 Phân lớp và dự đoán (Classification & Prediction)	8
1.4.2 Luật kết hợp (Association Rules)	11
1.4.3 Khai thác mẫu tuần tự (Sequential/ Temporal patterns)	11
1.4.4 Phân nhóm - đoạn (Clustering/ Segmentation)	11
1.4.5 Hồi quy (Regression)	12
1.4.6 Tổng hợp hóa (Summarization)	12
1.4.7 Mô hình hóa sự phụ thuộc (dependency modeling)	12
1.4.8 Phát hiện sự biến đổi và độ lệch (Change and deviation detection)	13
1.5 Các cơ sở dữ liệu phục vụ cho khai phá dữ liệu	13
1.6 Một số ứng dụng của khai phá dữ liệu	14
1.7 Khai phá luật kết hợp	14
1.7.1 Bài toán phát hiện luật kết hợp	14
1.7.2 Các khái niệm	15
1.7.3 Các cách tiếp cận khai phá tập mục thường xuyên	18
1.7.4 Một số thuật toán điển hình tìm tập mục thường xuyên	19

1.7.4.1 Thuật toán Apriori (Phương pháp sinh ứng viên)	19
1.7.4.2 Thuật toán FP-Growth	23
1.8 Kết luận chương 1	31
Chương 2: Khai phá tập mục thường xuyên đóng trong cơ sở dữ liệu	32
2.1 Cơ sở toán học của tập mục thường xuyên đóng	32
2.1.1 Ánh xạ đóng	32
2.1.2 Tập đóng	32
2.1.3 Kết nối Galois	32
2.1.4 Bao đóng của tập mục dữ liệu	33
2.2 Khái niệm, tính chất tập mục thường xuyên đóng	34
2.3 Một số thuật toán điển hình khai phá tập mục thường xuyên đóng	35
2.3.1 Thuật toán CHARM (Phương pháp dựa trên cây IT-Tree)	35
2.3.1.1 Giới thiệu thuật toán CHARM	35
2.3.1.2. Cây tìm kiếm và lớp tương đương	35
2.3.1.3 Các tính chất cơ bản của cặp tập mục – tập định danh	36
2.3.1.4 Thiết kế thuật toán	37
2.3.2 Thuật toán Closet +	41
2.4 Kết luận chương 2	45
Chương 3: Chương trình thực nghiệm ứng dụng trong lĩnh vực y tế	46
3.1 Bài toán phát hiện luật kết hợp trong dữ liệu y tế	46
3.2 Xây dựng chương trình	50
3.3 Kết quả thực nghiệm	57
3.4 Nhận xét	58
KẾT LUẬN	59
TÀI LIỆU THAM KHẢO	60
PHỤ LỤC	62

DANH MỤC CÁC KÝ HIỆU VÀ CÁC CHỮ VIẾT TẮT

Ký hiệu	Diễn giải
C_k	Tập các k tập mục ứng viên
BFS	Breadth First Search
CSDL	Cơ sở dữ liệu
CHARM	Closed Association Rule Mining
DB	Cơ sở dữ liệu giao tác
DFS	Depth First Search
FP -growth	Frequent -Pattern Growth
FP -tree	Frequent pattern tree
IT-tree	Itemset-Tidset tree
I	Tập các mục dữ liệu
k-itemset	Tập mục gồm k mục
KPDL	Khai phá dữ liệu
Minsup	Ngưỡng hỗ trợ tối thiểu
L_k	Tập các k-tập mục thường xuyên
Supp	Độ hỗ trợ (support)
TID	Định danh của giao tác
T	Giao tác (transaction)
DL	Dữ liệu
TX	Thường xuyên
TTHN	Tình trạng hôn nhân

DANH MỤC CÁC BẢNG

Bảng 1.1 Cơ sở dữ liệu giao tác minh họa thực hiện thuật toán Apriori

Bảng 1.2: CSDL giao tác minh họa cho thuật toán FP-Growth

Bảng 2.1: a) CSDL giao tác biểu diễn ngang b) CSDL giao tác biểu diễn dọc

Bảng 3.1 : Dữ liệu bệnh hen suyễn

Bảng 3.2: Lựa chọn thuộc tính

Bảng 3.3: Thuộc tính “Tuổi” sau khi phân hoạch

Bảng 3.4: Dữ liệu tìm kiếm sau khi thực hiện phân loại dữ liệu

Bảng 3.5: Chuyển đổi dữ liệu

Bảng 3.6: Dữ liệu cho khai phá

DANH MỤC HÌNH VẼ

Hình 1.1: Kiến trúc của một hệ thống khai phá dữ liệu

Hình 1.2: Quá trình khám phá tri thức

Hình 1.3: Quá trình khai phá dữ liệu

Hình 1.4: Cây FP-tree được xây dựng dần khi thêm các giao tác $t_1 \div t_6$

Hình 1.5: Cây FP-tree của CSDL DB trong bảng 1.5

Hình 2.1: Kết nối Galois

Hình 2.2: Cây IT-tree dùng Tidset với $\text{minSup} = 3$

Hình 2.3: Cây IT-tree tìm tập mục thường xuyên đóng thỏa mãn ngưỡng $\text{minsup} = 50\%$

Hình 2.4: Áp dụng tính chất của tập thường xuyên đóng

Hình 2.5: Minh họa xây dựng cây kết quả

Hình 3.1: Mô hình khai phá cho dữ liệu y tế

Hình 3.2: Giao diện chính chứa dữ liệu gốc

Hình 3.3: Hiển thị dữ liệu chuyển đổi

Hình 3.4: Giao diện thêm mới bản ghi

MỞ ĐẦU

Khai phá dữ liệu (*Data Mining*), hiện nay đang được rất nhiều người chú ý. Nó thực sự đã đem lại những lợi ích đáng kể trong việc cung cấp những thông tin tiềm ẩn trong các cơ sở dữ liệu lớn, giúp người sử dụng thu được những tri thức hữu ích từ những cơ sở dữ liệu hoặc các kho dữ liệu khổng lồ khác. Những “tri thức” chiết xuất từ nguồn cơ sở dữ liệu đó phục vụ các yêu cầu trợ giúp quyết định ngày càng có ý nghĩa quan trọng và là nhu cầu to lớn trong mọi lĩnh vực hoạt động kinh doanh, quản lý. Tiến hành công việc như vậy chính là thực hiện quá trình phát triển tri thức trong cơ sở dữ liệu (*Knowledge Discovery in Database*) mà trong đó kỹ thuật khai phá dữ liệu (*Data Mining*) cho phép phát hiện những tri thức tiềm ẩn.

Một trong các nội dung cơ bản trong khai phá dữ liệu là khai phá luật kết hợp. Khai phá luật kết hợp gồm hai bước: Bước thứ nhất, tìm tất cả các tập mục thường xuyên, đòi hỏi sự tính toán lớn. Bước thứ hai, dựa vào các tập mục thường xuyên tìm các luật kết hợp, đòi hỏi tính toán ít hơn, song gặp phải một vấn đề là có thể sinh ra quá nhiều luật, vượt khỏi sự kiểm soát của người khai phá hoặc người dùng, trong đó có nhiều luật không cần thiết. Để giải quyết vấn đề đó, trong bước thứ nhất, không cần thiết khai phá tất cả các tập mục thường xuyên mà chỉ cần khai phá các tập mục thường xuyên đóng. Khai phá luật kết hợp dựa trên tập mục thường xuyên đóng cho hiệu quả cao hơn, nó đảm bảo không tìm ra các tập mục thường xuyên không cần thiết, không sinh ra các luật dư thừa.

Với ý nghĩa đó và mục đích tìm hiểu về bài toán tìm tập mục thường xuyên trong cơ sở dữ liệu lớn, em đã quyết định lựa chọn đề tài “*Khai phá tập mục thường xuyên đóng trong cơ sở dữ liệu và ứng dụng*”.

Nội dung luận văn gồm 3 chương:

Chương 1: Tổng quan về khai phá dữ liệu và khai phá tập mục thường xuyên

Chương 2: Khai phá tập mục thường xuyên đóng trong cơ sở dữ liệu

Chương 3: Chương trình thực nghiệm ứng dụng trong lĩnh vực y tế

CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU VÀ KHAI PHÁ TẬP MỤC THƯỜNG XUYỀN

1.1 Khái niệm về khai phá tri thức và khai phá dữ liệu

"Khám phá tri thức là quá trình tìm ra những tri thức, đó là những mẫu tìm ẩn, trước đó chưa biết và là thông tin hữu ích đáng tin cậy". Còn khai phá dữ liệu (KPDL) là một bước quan trọng trong quá trình khám phá tri thức, sử dụng các thuật toán KPDL chuyên dùng với một số qui định về hiệu quả tính toán chấp nhận được để chiết xuất ra các mẫu hoặc các mô hình có ích trong dữ liệu. Nói một cách khác, mục đích của khám phá tri thức và KPDL chính là tìm ra các mẫu hoặc mô hình đang tồn tại trong các cơ sở dữ liệu (CSDL) nhưng vẫn còn bị che khuất bởi hàng núi dữ liệu.

Khám phá tri thức từ CSDL là một quá trình sử dụng các phương pháp và công cụ tin học, trong đó con người là trung tâm của quá trình. Do đó, con người cần phải có kiến thức cơ bản về lĩnh vực cần khám phá để có thể chọn được tập con dữ liệu tốt, từ đó phát hiện các mẫu phù hợp với mục tiêu đề ra. Đó chính là tri thức, được rút ra từ CSDL, thường để phục vụ cho việc giải quyết một loạt nhiệm vụ nhất định trong một lĩnh vực nhất định. Tuy vậy, quá trình khám phá tri thức mang tính chất hướng nhiệm vụ vì không phải là mọi tri thức tìm được đều áp dụng vào thực tế được.

Để có được những thông tin quý báu chúng ta phải tìm ra các mẫu có trong tập CSDL trước. Việc đánh giá các mẫu được tìm thấy cũng là một điều thú vị và tất yếu có tính chất quyết định đến sự sử dụng hay không sử dụng chúng. Đầu ra của một chương trình là khám phá những mẫu có ích được gọi là tri thức. Tri thức được khám phá có các đặc điểm chính:

- Kiến thức cao cấp: Ngày càng có nhiều câu hỏi mang tính chất định tính cần phải trả lời dựa trên một khối lượng dữ liệu khổng lồ đã có. Quá trình để tìm ra kiến thức như vậy không phải từ những phương pháp thống kê cổ điển mà mà nó được được đúc kết từ các kinh nghiệm đã có, được thể hiện trong dữ liệu, những kết quả đó có thể lĩnh hội được.