

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

VY ĐẠI NGHĨA

**PHÁT HIỆN MỐI QUAN HỆ TRONG
CƠ SỞ DỮ LIỆU VÀ ỨNG DỤNG TRONG Y
HỌC**

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên - 2015

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

VY ĐẠI NGHĨA

**PHÁT HIỆN MỐI QUAN HỆ TRONG
CƠ SỞ DỮ LIỆU VÀ ỨNG DỤNG TRONG Y
HỌC**

Chuyên ngành: Khoa học máy tính

Mã số: 60 48 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS.TS. Đỗ Trung Tuấn

Thái Nguyên - 2015

Lời cảm ơn

Trước tiên, tôi xin được gửi lời cảm ơn đến tất cả quý thầy cô đã giảng dạy trong chương trình Cao học do Trường Đại học Công nghệ thông tin và truyền thông tổ chức, những người đã truyền đạt cho tôi những kiến thức hữu ích về khoa học máy tính làm cơ sở cho tôi thực hiện tốt luận văn này.

Tôi xin chân thành cảm ơn PGS. TS. Đỗ Trung Tuấn đã tận tình hướng dẫn cho tôi trong thời gian thực hiện luận văn. Mặc dù trong quá trình thực hiện luận văn có giai đoạn không được thuận lợi nhưng những gì Thầy đã hướng dẫn, chỉ bảo đã cho tôi nhiều kinh nghiệm trong thời gian thực hiện đề tài.

Tôi cũng xin gửi lời cảm ơn đến tất cả các Thầy Cô đang làm việc tại Phòng khám đa khoa trường Cao đẳng Y tế Phú Thọ đã tận tình giúp đỡ trong việc thu thập thông tin, lấy số liệu về bệnh và thuốc làm cơ sở dữ liệu cho luận văn.

Sau cùng tôi xin gửi lời biết ơn sâu sắc đến các anh chị trong lớp và gia đình đã luôn tạo điều kiện tốt nhất cho tôi trong suốt quá trình học cũng như thực hiện luận văn.

Do thời gian có hạn và kinh nghiệm nghiên cứu khoa học chưa nhiều nên luận văn còn nhiều thiếu sót, rất mong nhận được ý kiến góp ý của Thầy/Cô và các anh chị học viên.

Phú Thọ, tháng 7 năm 2015

Học viên

Vy Đại Nghĩa

Lời cam đoan

Tôi cam đoan những kết quả trong luận văn là của việc tìm hiểu, có trích dẫn và tham chiếu đến các nguồn tư liệu tin cậy. Nội dung luận văn không sao chép từ các kết quả của các luận văn, luận án khác.

MỤC LỤC

Lời cảm ơn	i
Lời cam đoan	iii
MỤC LỤC	iv
DANH MỤC CÁC TỪ VIẾT TẮT	vi
DANH MỤC CÁC BẢNG, HÌNH VẼ	vii
MỞ ĐẦU	1
CHƯƠNG 1	6
TỔNG QUAN VỀ PHÁT HIỆN MỐI QUAN HỆ GIỮA CÁC DỮ LIỆU TRONG CƠ SỞ DỮ LIỆU	6
1. 1. Mục tiêu của việc phát hiện mối quan hệ giữa các dữ liệu	6
1. 2. Các bước chính của quá trình khai phá tri thức	6
1. 3. Các dạng dữ liệu có thể khai phá	7
1. 4. Các hướng tiếp cận chính trong khai phá dữ liệu	8
1. 5. Phân loại và ứng dụng các hệ thống khai phá dữ liệu	11
1. 5. 1. Phân loại các hệ thống khai phá dữ liệu	11
1. 5. 2. Ứng dụng của khai phá dữ liệu	12
1. 6. Kết luận chương	12
CHƯƠNG 2	13
MỘT SỐ MỐI QUAN HỆ DỮ LIỆU ĐƯỢC PHÁT HIỆN THÔNG QUA NGÔN NGỮ TRUY VẤN	13
2. 1. Luật kết hợp	13
2. 1. 1. Các khái niệm cơ bản	13
2. 1. 2. Bài toán khai phá luật kết hợp	16
2. 2. Khai thác tập phổ biến dựa trên ngôn ngữ truy vấn	17
2. 2. 1. Ngôn ngữ truy vấn	17
2. 2. 2. Tìm tập phổ biến bằng K-way join	20

2. 2. 3. Kết quả thử nghiệm 3 phương pháp đếm độ hỗ trợ.....	27
2. 2. 4. Phân tích các cải tiến của thuật toán k-way join	32
2. 2. 5. Phát sinh luật kết hợp.....	38
2. 2. 6. Rút ngọn luật kết hợp.....	42
2. 3. Kết luận chương	49
CHƯƠNG 3.....	51
ỨNG DỤNG TRONG TÍNH TOÁN THỬ NGHIỆM	51
3. 1. Các bài toán.....	51
3. 1. 1. Bài toán tìm luật kết hợp dạng $X \rightarrow Y$	51
3. 1. 2. Bài toán tìm độ hỗ trợ và độ tin cậy của luật	52
3. 1. 3. Bài toán đánh giá độ tin cậy của luật theo ngưỡng	53
3. 1. 5. Giải pháp giúp thực hiện các bài toán	54
3. 2. Chương trình thử nghiệm	56
3. 2. 1. Cơ sở dữ liệu của bài toán.....	57
3. 2. 2. Kết quả khai phá dữ liệu khi thực hiện các bài toán	58
3. 3. Kết luận chương	65
KẾT LUẬN	67
PHỤ LỤC.....	68
TÀI LIỆU THAM KHẢO	76

DANH MỤC CÁC TỪ VIẾT TẮT

ADO	Active X Data Object
ANSI	Chuẩn quốc gia Hoa Kỳ
Client/ server	Khách/ chủ
confidence	Độ tin cậy
CSDL	Cơ sở dữ liệu
DB2	Tên hệ quản trị cơ sở dữ liệu của IBM
DBMS	Hệ quản trị cơ sở dữ liệu
HQTCSDL	Hệ quản trị cơ sở dữ liệu
ISO	Tổ chức tiêu chuẩn hóa quốc tế
MOLAP	multidimensional OLAP
OLAP	Online Analysis Processing
ORACLE	Tên công ty ORACLE, tên hệ quản trị cơ sở dữ liệu
ROLAP	Relational OLAP
SQL	Ngôn ngữ truy vấn
support	Độ hỗ trợ, trợ giúp

DANH MỤC CÁC BẢNG, HÌNH VẼ

Hình. Thí dụ về xử lí dữ liệu y tế tại trường Cao đẳng Y tế Phú Thọ.....	2
Hình 1. 1: Các bước trong quá trình khai phá tri thức.....	6
Hình 1. 2: Các kiến trúc khai phá tích hợp với cơ sở dữ liệu	9
Hình 1. 3: Kiến trúc gắn kết lỏng	9
Hình 1. 4: Kiến trúc thủ tục nội và hàm do người dùng định nghĩa	10
Hình 1. 5: Kiến trúc dựa trên truy vấn SQL	10
Hình 2. 1: Minh họa luật kết hợp	16
Bảng 2. 1: Cấu trúc bảng ban đầu	20
Bảng 2. 2: Cấu trúc bảng dùng để khai khác	21
Hình 2. 2: Tiến trình phát sinh tập ứng viên C_k	23
Hình 2. 2: Đếm độ hỗ trợ bằng cách tiếp cận K-way Join.....	24
Hình 2. 3: Biểu đồ hình cây cho Sub Query Q_i	26
Hình 2. 4: Đồ thị thời gian thực thi của 3 thuật toán khi minsup=10% và D=100000 .	28
Hình 2. 5: Đồ thị thời gian thực thi 3 thuật toán khi minsup=10% và D=50000	29
Hình 2. 7: Đồ thị thời gian thực thi của 3 thuật toán khi minsup=10% và D=10000 ...	29
Hình 2. 6: Đồ thị tổng hợp thời gian thực thi của 3 thuật toán khi minsup lớn.....	29
Hình 2. 7: Đồ thị thời gian thực thi 3 thuật toán khi minsup=5% và D=100000	30
Hình 2. 8: Đồ thị thời gian thực thi 3 thuật toán khi minsup=5% và D=50000	30
Hình 2. 9: Đồ thị thời gian thực thi 3 thuật toán khi minsup=5% và D=10000	30
Hình 2. 10: Đồ thị tổng hợp thời gian thực thi 3 thuật toán khi minsup trung bình	31
Hình 2. 11: Đồ thị thời gian thực thi 3 thuật toán khi minsup = 1% và D = 100000....	31
Hình 2. 12: Đồ thị thời gian thực thi 3 thuật toán khi minsup = 1% và D= 50000	32
Hình 2. 13: Đồ thị thời gian thực thi của 3 thuật toán khi minsup =1% và D=10000 ..	32

Hình 2. 154: Đồ thị tổng hợp thời gian thực thi của 3 thuật toán khi minsup nhỏ	32
Bảng 2. 3: Cơ sở dữ liệu ban đầu D	44
Bảng 2. 4: Cơ sở dữ liệu sau khi chuyển đổi	44
Bảng 2. 5: Kết quả F_1	45
Bảng 2. 6: Kết quả F_2	46
Bảng 2. 7: Kết quả C_3	46
Bảng 2. 8: Kết quả $Comb_3$	47
Bảng 2. 9: Kết quả F_3	47
Bảng 2. 10: Kết quả C_4	48
Bảng 2. 11: Kết quả $Comb_4$	49
Bảng 2. 12: Kết quả F_4	49
Bảng 2. 13. Kết quả	49
Bảng 3. 1. Cấu trúc bảng dữ liệu ban đầu.....	55
Bảng 3. 2. Cấu trúc bảng dùng để khai phá dữ liệu	56
Hình 3. 1. Mẫu đơn thuốc của Phòng khám đa khoa Trường cao đẳng Y Phú Thọ	57
Hình 3. 2. Minh họa cấu trúc dữ liệu ban đầu.....	58
Hình 3. 3. Cấu trúc dữ liệu dùng để khai phá	58
Hình 3. 4. Tính độ hỗ trợ và độ tin cậy của luật $\{Cefalecin\} \Rightarrow \{Paracetamol\}$	61
Hình 3. 5. Tính độ hỗ trợ và độ tin cậy của một luật $\{Decolgen\} \Rightarrow \{Vitamin C\}$	61
Hình 3.6. Đánh giá độ tin cậy của luật $\{Decolgen\} \Rightarrow \{Vitamin B1\}$	65
Hình 3.7. Đánh giá độ tin cậy của luật $\{Cefalecin\} \Rightarrow \{Vitamin C\}$	65
Hình PL1: Minh họa dữ liệu đầu vào	68