

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

NGUYỄN THỊ LUYẾN

KHAI PHÁ TẬP MỤC LỢI ÍCH CAO
DỰA TRÊN CẤU TRÚC CÂY TIỀN TỔ

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYÊN - NĂM 2014

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

NGUYỄN THỊ LUYẾN

**KHAI PHÁ TẬP MỤC LỢI ÍCH CAO
DỰA TRÊN CẤU TRÚC CÂY TIỀN TỔ**

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Chuyên ngành: KHOA HỌC MÁY TÍNH

Mã số: 60 48 01

Người hướng dẫn khoa học: TS. LÊ VĂN PHÙNG

Thái Nguyên, 2014

LỜI CAM ĐOAN

Tôi xin cam đoan Luận văn "*Khai phá tập mục lợi ích cao dựa trên cấu trúc cây tiền tố*" đã được thực hiện theo đúng mục tiêu đề ra dưới sự hướng dẫn của TS. Lê Văn Phùng. Kết quả đạt được trong luận văn là sản phẩm của cá nhân tôi. Trong toàn bộ luận văn, những điều được trình bày là của cá nhân và 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 29 tháng 9 năm 2014

Người cam đoan

Nguyễn Thị Luyện

LỜI CẢM ƠN

Lời đầu tiên tôi xin gửi lời cảm ơn chân thành và biết ơn sâu sắc tới TS. Lê Văn Phùng – Trường Đại học công nghệ Thông tin và Truyền thông, Thầy đã chỉ bảo và hướng dẫn tận tình cho tôi trong suốt quá trình làm việc 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 29 tháng 9 năm 2014

Tác giả

Nguyễn Thị Luyện

MỤC LỤC

LỜI CAM ĐOAN	i
LỜI CẢM ƠN	iv
DANH MỤC CÁC HÌNH VẼ.....	vii
DANH MỤC CÁC BẢNG.....	viii
DANH MỤC CÁC KÝ HIỆU.....	ix
DANH MỤC CHỮ VIẾT TẮT.....	x
MỞ ĐẦU.....	1
CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU	3
1.1 Quá trình khám phá tri thức.....	3
1.1.1 Khái niệm về quá trình khám phá tri thức và khai phá dữ liệu	3
1.1.2 Kiến trúc về một số hệ thống khai phá dữ liệu	5
1.1.3. Một số ứng dụng của khai phá dữ liệu	6
1.2 Một số phương pháp khai phá dữ liệu thông dụng.....	7
1.2.1 Phương pháp luật kết hợp	7
1.2.2 Phương pháp cây quyết định	8
1.3 Kết luận chương 1	12
CHƯƠNG 2: KHAI PHÁ TẬP MỤC THƯỜNG XUYÊN VÀ TẬP MỤC LỢI ÍCH CAO	13
2.1 Khai phá tập mục thường xuyên.....	13
2.1.1 Cơ sở dữ liệu giao tác	13
2.1.2 Tập mục thường xuyên và luật kết hợp	15
2.1.3 Bài toán khai phá luật kết hợp và một số thuật toán về khai phá tập mục thường xuyên	17
2.2 Bài toán Khai phá tập mục lợi ích cao	29
2.2.1 Khái niệm về tập mục lợi ích cao	29
2.2.2 Một số bài toán khai phá tập mục lợi ích cao	29
2.3 Khai phá tập mục lợi ích cao dựa trên cây tiền tố	34
2.3.1 Định nghĩa cây tiền tố.....	34

2.3.2 Một số thuật toán khai phá tập mục lợi ích cao dựa trên cây tiền tố	35
2.3.3 Các cấu trúc cây tiền tố cho khai phá lợi ích cao	56
2.3.4 Thuật toán UP-Growth	59
2.4 Kết luận chương 2	62
CHƯƠNG 3: THỰC NGHIỆM KHAI PHÁ TẬP MỤC LỢI ÍCH CAO DỰA TRÊN CẤU TRÚC CÂY TIỀN TỐ	63
3.1. Bài toán phát hiện nhóm các mặt hàng có lợi nhuận cao	63
3.2. Mô tả dữ liệu	63
3.3 Xây dựng chương trình.....	70
3.4 Thực nghiệm khai phá tìm tập mục lợi ích cao	71
3.5 Kết luận chương 3	72
KẾT LUẬN	73
1. Những kết quả chính của luận văn	73
2. Hướng nghiên cứu tiếp theo	73
TÀI LIỆU THAM KHẢO.....	74
A. Tiếng việt.....	74
B. Tiếng Anh	74

DANH MỤC CÁC HÌNH VẼ

Hình 1.1. Các bước trong Data Mining và KDD	5
Hình 1.2. Kiến trúc của một hệ thống khai phá dữ liệu	5
Hình 1.3. Luồng thông tin được sử dụng theo cách kết hợp.....	8
Hình 1.4 Cây quyết định về khái niệm mua máy tính	9
Hình 1.5. Cây quyết định phân lớp (bad/good) mức lương	11
Hình 1.6 Các bước thực hiện thuật toán K-Mean	12
Hình 2.1. Cây FP-tree của CSDL bảng 2.5.....	27
Hình 2.2. Cây COFI-tree của mục D	27
Hình 2.3. Minh họa các bước khai phá cây D-COFI-tree.....	28
Hình 2.4. Cây TWUI-tree sau khi lưu giao tác T_1	39
Hình 2.5. Cây TWUI-tree sau khi lưu giao tác T_1 và T_2	39
Hình 2.6. Cây TWUI-tree của CSDL bảng 2.9 và bảng 2.10	40
Hình 2.7. Cây C-COUI-tree sau khi lưu mẫu CBE.....	42
Hình 2.8. Cây C-COUI-tree sau khi lưu mẫu CBE và CE.....	43
Hình 2.9. Cây C-COUI-tree sau khi xây dựng xong.....	43
Hình 2.10. Cây D-COUI-tree	43
Hình 2.11. Cây B-COUI-tree	44
Hình 2.12. Các bước khai phá cây D-COUI-Tree	45
Hình 2.13. Không gian tìm kiếm tập mục lợi ích cao theo thuật toán Hai pha.....	56
Hình 2.14. Cây TWUI-tree có các mục dữ liệu sắp tăng dần theo trật tự từ điển của cơ sở dữ liệu bảng 2.9 và bảng 2.10	57
Hình 2.15. Cây TWUI-tree có các mục dữ liệu sắp giảm dần theo số lần xuất hiện của chúng trong cơ sở dữ liệu bảng 2.9 và bảng 2.10.....	57
Hình 2.16. Cây TWUI-tree có các mục dữ liệu sắp giảm dần theo TWU của chúng trong cơ sở dữ liệu bảng 2.9 và bảng 2.10	58
Hình 2.17. Cây TWUI-tree của CSDL bảng 2.8 với $minutil = 40$	62
Hình 2.18. Cây UP-tree của CSDL bảng 2.8 với $minutil = 40$	62
Hình 3.1. Tập CSDL.txt biểu diễn dữ liệu đầu vào	70
Hình 3.2. Giao diện chính của chương trình.....	71
Hình 3.3. Tập các mục lợi ích cao	72

DANH MỤC CÁC BẢNG

Bảng 1.1: Tập dữ liệu huấn luyện quyết định phân lớp mức lương.....	10
Bảng 2.1: Biểu diễn ngang của cơ sở dữ liệu giao tác	14
Bảng 2.2: Biểu diễn dọc của cơ sở dữ liệu giao tác	14
Bảng 2.3: Ma trận giao tác của cơ sở dữ liệu cho ở bảng 2.1	15
Bảng 2.4: Cơ sở dữ liệu giao tác minh họa thực hiện thuật toán Apriori	21
Bảng 2.5: CSDL giao tác minh họa thực hiện thuật toán COFI-tree	25
Bảng 2.6: Các mục dữ liệu và độ hỗ trợ.....	25
Bảng 2.7: Các mục dữ liệu thường xuyên đã sắp thứ tự	25
Bảng 2.8: Các mục DL trong giao tác sắp xếp giảm dần theo độ hỗ trợ	26
Bảng 2.9. CSDL giao tác.....	32
Bảng 2.10 Bảng lợi ích.....	32
Bảng 2.11: Lợi ích các giao tác của cơ sở dữ liệu bảng 2.9 và bảng 2.10.....	37
Bảng 2.12: Lợi ích TWU của các mục dữ liệu.....	37
Bảng 2.13: Các mục dữ liệu có lợi ích TWU cao sắp giảm dần theo tw	38
Bảng 2.14. Các mục dữ liệu trong giao tác sắp giảm dần theo lợi ích TWU	38
Bảng 2.15. Kết quả tính lợi ích của các tập mục ứng viên	46
Bảng 2.16: Cơ sở dữ liệu ví dụ cho thuật toán UP-Growth	60
Bảng 2.17: Bảng lợi ích của CSDL bảng 2.15	61
Bảng 2.18: Các giao tác được sắp lại các mục dữ liệu theo TWU giảm dần.....	61
Bảng 3.1 Dữ liệu đã trích chọn để khai phá.....	65
Bảng 3.2. Mã hóa các mặt hàng	68
Bảng 3.3. Bảng lợi ích các mặt hàng.....	69

DANH MỤC CÁC KÝ HIỆU

$|X|$: Số phần tử của tập hợp X.

A, B, C, ...: Tên các mục dữ liệu trong cơ sở dữ liệu giao tác ví dụ.

$\text{Conf}(X \rightarrow Y)$: Độ tin cậy của một luật $X \rightarrow Y$

$db \subseteq DB$: db là cơ sở dữ liệu giao tác con của DB.

$DB = \{T_1, T_2, \dots, T_m\}$: Cơ sở dữ liệu có m giao tác.

$I = \{i_1, i_2, \dots, i_n\}$: Tập n mục dữ liệu.

I_p : Mục dữ liệu thứ p.

m: Số giao tác một cơ sở dữ liệu giao tác.

Minconf: Độ tin cậy tối thiểu

minShare: Ngưỡng cổ phần tối thiểu.

minsup: Ngưỡng độ hỗ trợ tối thiểu.

minutil: Ngưỡng lợi ích tối thiểu

n: Số mục dữ liệu một cơ sở dữ liệu giao tác.

Nếu $X \subseteq Y$ thì X gọi là tập con của tập Y, Y gọi là tập cha của tập X

$P(Y/X)$: Xác suất có điều kiện (độ tin cậy của luật $Y \rightarrow X$)

$P(Y/X)$: Xác suất có điều kiện (độ tin cậy của luật kết hợp $X \rightarrow Y$)

$\text{Sup}(X)$: Tỷ lệ % của giao tác chứa tập X

T_q : Giao tác thứ q.

$U(X)$: Lợi ích của tập mục trong CSDL DB

$X = ABC$ thay cho $X = \{A, B, C\}$ trong các cơ sở dữ liệu giao tác ví dụ.

X, Y, ...: Tập con của tập mục dữ liệu I, $X, Y \subseteq I$.

DANH MỤC CHỮ VIẾT TẮT

AIS		Thuật toán AIS
CHARM		Thuật toán CHAM
CNTT		Công nghệ thông tin
CSDL		Cơ sở dữ liệu
FP-Growth		Thuật toán FP-Growth
SETM		Thuật toán SETM
UP-Growth		Thuật toán UP-Growth
DM	Data Mining	Khai phá dữ liệu
HU	High Utility	Khai phá tập mục lợi ích cao
TWU	Transaction Weighted Utility	Tập mục ràng buộc lợi ích theo giao tác
TWUI-tree	Transaction Weighted Utility tree	Là một cấu trúc cây tiền tố
KDD	Knowledge Discovery from Data	Phát hiện tri thức từ dữ liệu
PT	Prefix-tree	Cây tiền tố