

BỘ GIÁO DỤC VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

VIỆN CÔNG NGHỆ THÔNG TIN

CAO TÙNG ANH

**KHAI THÁC DỮ LIỆU PHÂN TÁN
BẢO TOÀN TÍNH RIÊNG TƯ**

LUẬN ÁN TIẾN SĨ TOÁN HỌC

HÀ NỘI- 2014

BỘ GIÁO DỤC VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

VIỆN CÔNG NGHỆ THÔNG TIN

CAO TÙNG ANH

**KHAI THÁC DỮ LIỆU PHÂN TÁN
BẢO TOÀN TÍNH RIÊNG TƯ**

**Chuyên ngành: BẢO ĐẢM TOÁN HỌC CHO MÁY TÍNH
VÀ HỆ THỐNG TÍNH TOÁN**

Mã số: 62.46.35.01

LUẬN ÁN TIẾN SĨ TOÁN HỌC

NGƯỜI HƯỚNG DẪN KHOA HỌC:

1. PGS.TSKH. NGUYỄN XUÂN HUY
2. PGS.TS. NGUYỄN MẬU HÂN

HÀ NỘI - 2014

LỜI CAM ĐOAN

Tôi cam đoan đây là công trình nghiên cứu của riêng tôi. Các số liệu, kết quả trong luận án là trung thực và chưa từng công bố trong bất kỳ công trình nào khác.

Tác giả luận án

Cao Tùng Anh

LỜI CẢM ƠN

Luận án được thực hiện và hoàn thành dưới sự hướng dẫn của PGS.TSKH. Nguyễn Xuân Huy và PGS.TS. Nguyễn Mậu Hân. Trong thời gian thực hiện luận án, tác giả đã nhận được sự giúp đỡ và chỉ dẫn khoa học rất tận tình từ hai người thầy của mình để có thể hoàn thành luận án này. Nhân dịp này tác giả xin được gửi đến hai thầy: PGS.TSKH. Nguyễn Xuân Huy và PGS.TS. Nguyễn Mậu Hân lòng biết ơn sâu sắc và lời cảm ơn chân thành nhất.

Tác giả cũng xin được trân trọng cảm ơn PGS.TS. Thái Quang Vinh, GS.TS. Vũ Đức Thi, PGS.TS. Đoàn Văn Ban, PGS.TS. Đặng Văn Đức, PGS.TS. Ngô Quốc Tạo, PGS.TS. Đỗ Năng Toàn, PGS.TS. Lương Chi Mai, PGS.TS. Nguyễn Thanh Tùng là những thầy (cô) của Viện Công Nghệ Thông Tin đã quan tâm chỉ bảo, động viên và giúp đỡ tác giả trong suốt quá trình học tập, nghiên cứu và hoàn thiện luận án.

Tác giả cũng xin trân trọng cảm ơn PGS.TS. Lê Hoài Bắc và các bạn đồng nghiệp trong nhóm nghiên cứu tại TP.Hồ Chí Minh đã đọc và cho những ý kiến đóng góp quý báu cho nội dung luận án.

Cuối cùng xin chân thành cảm ơn các bạn đồng nghiệp tại khoa CNTT, trường Đại học Công nghệ TP.Hồ Chí Minh đã cổ vũ, động viên, giúp đỡ về nhiều mặt cho tác giả trong thời gian thực hiện luận án.

MỤC LỤC

LỜI CAM ĐOAN	1
LỜI CẢM ƠN.....	2
MỤC LỤC.....	3
DANH MỤC CÁC HÌNH	5
DANH MỤC CÁC BẢNG	7
DANH MỤC TỪ VIẾT TẮT.....	8
PHẦN MỞ ĐẦU	9
CHƯƠNG 1 Một số khái niệm về cơ sở dữ liệu phân tán, khai thác dữ liệu và bảo toàn tính riêng tư	19
1.1. Cơ sở dữ liệu phân tán.....	19
1.1.1. Khái niệm cơ sở dữ liệu phân tán	19
1.1.2. Cơ sở dữ liệu phân tán ngang	19
1.1.3. Cơ sở dữ liệu phân tán dọc	21
1.2. Khai thác dữ liệu.....	23
1.2.1. Khái niệm khai thác dữ liệu	23
1.2.2. Một số thuật toán khai thác dữ liệu.....	24
1.3. Bảo đảm tính riêng tư	31
1.3.1. Khái niệm.....	31
1.3.2. Phân loại các phương pháp PPDM	32
1.3.3. Đánh giá một thuật toán PPDM.....	34
1.4. Một số phương pháp giấu dữ liệu.....	35
1.4.1. Xáo trộn	35
1.4.2. Ngăn chặn	36
1.4.3. Gom / trộn.....	36
1.4.4. Đổi chỗ.....	36
1.4.5. Lấy mẫu	37
1.4.6. Ứng dụng lý thuyết giàn giao	41
1.5. Một số kỹ thuật khai thác dữ liệu bảo đảm tính riêng tư.....	49
1.5.1. Kỹ thuật chỉnh sửa dữ liệu trong cơ sở dữ liệu nhị phân	49
1.5.2. Kỹ thuật thay giá trị dữ liệu thật bằng giá trị không xác định	53
1.5.3. Phương pháp tái tạo	56
1.6. Kết chương	58
CHƯƠNG 2 Khai thác dữ liệu trên CSDL phân tán.....	60
2.1. Giới thiệu	60

2.2. Khai thác trên cơ sở dữ liệu phân tán dọc	60
2.2.1. Cách thực hiện	60
2.2.2. Thuật toán khai thác trên CSDL phân tán dọc với phép kết ngoại	62
2.2.3. Thuật toán khai thác CSDLPT dọc với phép kết ngoại hai chiều.....	66
2.2.4. Thuật toán khai thác CSDLPT dọc bằng phép kết tự nhiên	68
2.3. Khai thác trên cơ sở dữ liệu phân tán ngang	73
2.3.1. Cách thực hiện	73
2.3.2. Nhận xét phương pháp.....	75
2.4. Khai thác song song tập phổ biến trên CSDL phân tán	75
2.4.1. Đặt vấn đề	75
2.4.2. Mô hình khai thác	76
2.4.3. Thuật toán khai thác tập phổ biến trên Master.....	78
2.5. Khai thác tập mục có lợi ích cao	81
2.5.1. Đặt vấn đề	81
2.5.2. Khai thác tập mục có lợi ích cao.....	81
2.6. Kết chương	86
CHƯƠNG 3 Khai thác dữ liệu phân tán bảo đảm tính riêng tư	87
3.1. Giới thiệu chương.....	87
3.2. Khai thác CSDL phân tán dọc bảo đảm tính riêng tư	87
3.2.1. Đặt vấn đề	87
3.2.2. Thuật toán	88
3.2.3. Minh họa thuật toán:	89
3.3. Khai thác CSDL phân tán ngang bảo đảm tính riêng tư	94
3.3.1. Đặt vấn đề	94
3.3.2. Một số công cụ tính toán đa bên an toàn.	95
3.3.3. Giải thuật khai thác tập phổ biến đảm bảo riêng tư và chống thông đồng trên dữ liệu phân tán ngang.	96
3.4. Giao thức khai thác CSDL phân tán ngang bảo đảm tính riêng tư	107
3.4.1. Đặt vấn đề	107
3.4.2. Cơ sở lý thuyết.....	108
3.4.3. Giao thức khai thác	109
3.4.4. Đánh giá giao thức	113
3.4.5. Thực nghiệm giao thức	113
3.5. Kết chương	114
PHẦN KẾT LUẬN.....	116
DANH MỤC CÔNG TRÌNH ĐÃ CÔNG BỐ.....	120
TÀI LIỆU THAM KHẢO	121

DANH MỤC CÁC HÌNH

Hình 1.1. Thuật toán IT-Tree phát sinh tập phổ biến thỏa ngưỡng minsup.....	30
Hình 1.2. Kết quả khai thác với ngưỡng minsup=50%	31
Hình 1.3. Đồ thị dàn các tập mục thường xuyên	43
Hình 1.4. Gian giao đầy đủ của tập Poset (ABE)	44
Hình 1.5. Thuật toán Itemhide- Ẩn tập mục nhạy cảm.....	46
Hình 1.6. Thuật toán 1a.....	51
Hình 1.7. Thuật toán 1b.....	51
Hình 1.8. Thuật toán 2a.....	52
Hình 1.9. Thuật toán 2b.....	52
Hình 2.1. Mô hình hoạt động khai thác luật trên CSDL phân tán	60
Hình 2.2. Thuật toán Eclat_Distribute_Left_Join.....	63
Hình 2.3. Biểu diễn các mục đơn của DB ₁	65
Hình 2.4. Biểu diễn các mục đơn của DB ₁ và DB ₂	65
Hình 2.5. Kết quả khai thác trên CSDL phân tán với phép kết Left-join	66
Hình 2.6. Thuật toán Eclat_Distribut_Full_Join.....	67
Hình 2.7. Kết quả khai thác trên CSDLPT dọc với phép kết ngoại hai chiều	68
Hình 2.8. Thuật toán phát sinh tập phổ biến thỏa ngưỡng minsup	69
Hình 2.9. Cây biểu diễn các mục đơn của DB ₁ và DB ₂	71
Hình 2.10. Cây biểu diễn khai thác tập phổ biến trên CSDL phân tán.....	71
Hình 2.11. Mô hình tổng quát khai thác trên CSDL phân tán ngang	77
Hình 2.12. Trao đổi thông tin và khai thác tập phổ biến giữa Master và Slaver ...	77
Hình 2.13. Kết quả khai thác từ Slave 1 theo thuật toán Eclat	78
Hình 2.14. Kết quả khai thác từ Slave 2 theo thuật toán Eclat	78
Hình 2.15. Thuật toán PEclat	79
Hình 2.16. Kết quả của PEclat với minsup=50%	80
Hình 2.17. Cây WIT-Tree	82
Hình 2.18. Thuật toán TWU-Mining	83
Hình 2.19. Minh họa thuật toán TWU-Mining	84
Hình 3.1. Thuật toán phát sinh tập phổ biến	88
Hình 3.2. Sơ đồ hoạt động của thuật toán	89
Hình 3.3. Kết quả tạo ra lớp tương đương [∅]	91
Hình 3.4. Kết quả khai thác trên CSDL phân tán dọc.....	91
Hình 3.5. Thủ tục Create_Fitree.....	98
Hình 3.6. Thủ tục Secure_Support(X)	98
Hình 3.7. Thủ tục Extend_Fitree& Upper_Bound.....	99

Hình 3.8. Thủ tục Upper_Bound.....	100
Hình 3.9. Kết quả FITree sau khi xử lý nút gốc.....	101
Hình 3.10. Kết quả FITree sau khi xử lý nút A.....	102
Hình 3.11. Sự phụ thuộc thời gian vào số lượng máy trên CSDL Accident	107
Hình 3.12. Sự phụ thuộc thời gian vào số lượng máy trên CSDL bảo hiểm	107
Hình 3.13. Giao thức đảm bảo tính riêng tư	110
Hình 3.14. CSDL tập trung và CSDL phân tán	112
Hình 3.15. Các bên tính độ hỗ trợ cục bộ	112
Hình 3.16. Tính độ hỗ trợ toàn cục và tập phổ biến toàn cục	112
Hình 3.17. So sánh tổng chi phí của GTDX và GT M.Hussein.....	114

DANH MỤC CÁC BẢNG

Bảng 1.1. Quan hệ dự án (DA)	19
Bảng 1.2. Kết quả phân tán ngang nguyên thủy	20
Bảng 1.3. Quan hệ chi trả.....	20
Bảng 1.4. Quan hệ nhân viên	20
Bảng 1.5. Kết quả phân mảnh ngang dẫn xuất quan hệ NV	21
Bảng 1.6. Quan hệ nhân viên	21
Bảng 1.7. Kết quả phân tán dọc từ bảng 1.6	22
Bảng 1.8. Cơ sở dữ liệu giao dịch.....	30
Bảng 1.9. CSDL T 22 giao tác được viết thành 2 mảnh	42
Bảng 1.10. Tập mục thường xuyên theo ngưỡng $\sigma = 4$	42
Bảng 1.11. So sánh các thuật toán.....	50
Bảng 2.1. Cơ sở dữ liệu của Master	61
Bảng 2.2. Cơ sở dữ liệu của Slave	61
Bảng 2.3. Cơ sở dữ liệu sau khi kết	61
Bảng 2.4. Cơ sở dữ liệu của 2 bên tham gia khai thác.....	64
Bảng 2.5. Cơ sở dữ liệu kết ngoại (Left Join).....	64
Bảng 2.6. CSDL với phép kết ngoại “hai chiều”	66
Bảng 2.7. Cơ sở dữ liệu của 2 bên tham gia khai thác.....	69
Bảng 2.8. Cơ sở dữ liệu của bên A kết với bên B.....	70
Bảng 2.9. Kết quả thực nghiệm trên CSDL CO-OP Mark TP.HCM.....	73
Bảng 2.10. Cơ sở dữ liệu của Master	74
Bảng 2.11. Cơ sở dữ liệu của Slave	74
Bảng 2.12. Cơ sở dữ liệu sau khi hội Master và Slave	74
Bảng 2.13. CSDL mẫu	76
Bảng 2.14. Cơ sở dữ liệu phân tán của bảng 2.13.....	76
Bảng 2.15. Bảng giá trị khách quan	82
Bảng 2.16. Bảng giá trị chủ quan.....	82
Bảng 2.17. Bảng CSDL thực nghiệm.....	85
Bảng 2.18. Bảng thực nghiệm 2 thuật toán trong CSDL BMS-POS	85
Bảng 2.19. Bảng thực nghiệm 2 thuật toán trong CSDL Retail.....	86
Bảng 3.1. CSDL thực của hai bên Master và Slave.....	89
Bảng 3.2. CSDL giả của hai bên Master và Slave	90
Bảng 3.3. Kết quả thực nghiệm trên CSDL CO-OP Mart TP.HCM.....	93
Bảng 3.4. Minh họa hệ thống gồm 2 bên S_1, S_2	101
Bảng 3.5. Thời gian chạy trên CSDL Accidents.....	106
Bảng 3.6. Thời gian chạy trên CSDL bảo hiểm.....	106
Bảng 3.7. Thông tin về các CSDL thực nghiệm	114

DANH MỤC TỪ VIẾT TẮT

STT	Từ viết tắt	Diễn giải tiếng Anh	Diễn giải tiếng Việt
1	CSDL	Database	Cơ sở dữ liệu
2	CSDLPT	Database distributed	Cơ sở dữ liệu phân tán
3	GTDX	Proposed protocol	Giao thức đề xuất
4	WIT-Tree	Weighted Itemset-Tidset tree	Cây tập mục-tập giao dịch có trọng số
5	TWU	Tree Weighted Utility	Cây lợi ích có trọng số
6	FI	Frequent Itemsets	Tập phổ biến
7	FP-tree	Fast Parallel tree	Cây khai thác song song nhanh
8	FDM	Fast Distributed Mining	Khai thác phân tán nhanh
9	SVM	Support Vector machines	Sử dụng vectơ trong hỗ trợ phân lớp
10	PPDM	Privacy Preserving Data Mining	Khai thác dữ liệu bảo toàn tính riêng tư
11	RSA	Revest-Shamir-Adleman	Hệ mã hóa RSA
12	SM	Safety margin	Ngưỡng an toàn
13	MST	Min support	Độ hỗ trợ tối thiểu
14	MFI	Maximal Frequent Itemset	Tập phổ biến tối đại
15	MCT	Min Confident	Ngưỡng độ tin cậy
16	TID	Transaction index	Chỉ mục của giao dịch
17	IT-Tree	Itemset Tidset tree	Cây tập mục -tập giao dịch
18	HUIs	High Utility Itemsets	Tập tiện ích cao
19	DBS	Dynamic Bit String	Chuỗi bit động
20	SH	Semi Honest	Trung thực một nửa