**ĐẠI HỌC THÁI NGUYÊN**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG**

----------

**PHẠM ĐỨC QUANG**

# KHAI PHÁ LUẬT KẾT HỢP CÓ TRỌNG SỐ TRONG CƠ SỞ DỮ LIỆU LỚN

**LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH**

**THÁI NGUYÊN 2012**

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

----------

**PHẠM ĐỨC QUANG**

# KHAI PHÁ LUẬT KẾT HỢP CÓ TRỌNG SỐ TRONG CƠ SỞ DỮ LIỆU LỚN

**Chuyên ngành:** KHOA HỌC MÁY TÍNH
**Mã số:** 60.48.01

## LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

**Hướng dẫn khoa học: PGS.TS. NGUYỄN THANH TÙNG**

**THÁI NGUYÊN 2012**

# LỜI CÁM ƠN

Trước hết em xin gửi lời cám ơn chân thành đến toàn thể các thầy cô giáo Viện Công nghệ thông tin - Viện Khoa học và Công nghệ Việt Nam và Trường Đại học Công nghệ thông tin và Truyền thông - Đại học Thái nguyên đã dạy dỗ chúng em trong suốt quá trình học tập chương trình cao học tại trường.

Đặc biệt em xin bày tỏ lòng biết ơn sâu sắc tới PGS.TS. Nguyễn Thanh Tùng đã quan tâm, định hướng, đưa ra những gợi ý, góp ý và chỉnh sửa vô cùng quí báu cho em trong quá trình thực hiện luận văn này.

Cuối cùng, tôi xin chân thành cám ơn các bạn bè đồng nghiệp, gia đình và người thân đã quan tâm, giúp đỡ và chia sẻ với tôi trong suốt quá trình làm luận văn tốt nghiệp.

*Thái Nguyên, ngày 10 tháng 9 năm 2012*

Học viên

**Phạm Đức Quang**

# LỜI CAM ĐOAN

Tôi xin cam đoan:

Những nội dung trong luận văn này là do tôi thực hiện dưới sự hướng dẫn của PGS.TS. Nguyễn Thanh Tùng.

Mọi tham khảo sử dụng trong luận văn đều được trích dẫn rõ ràng tác giả, tên công trình, thời gian, địa điểm công bố.

Tôi xin chịu trách nhiệm với lời cam đoan này.

*Thái Nguyên, ngày 10 tháng 9 năm 2012*

Học viên

**Phạm Đức Quang**

# MỤC LỤC

*Trang*

Trang bìa phụ
Lời cảm ơn
Lời cam đoan
Mục lục ..... i
Danh mục các từ, các ký hiệu viết tắt ..... iv
Danh mục các bảng ..... v
**LỜI MỞ ĐẦU ..... 1**

**Chương 1. KHÁI QUÁT VỀ KHAI PHÁ DỮ LIỆU VÀ BÀI TOÁN KHAI PHÁ TẬP MỤC THƯỜNG XUYÊN ..... 3**

1.1. Khai phá dữ liệu ..... 3

1.2. Khai phá luật kết hợp ..... 8

1.2.1. Cơ sở dữ liệu giao tác ..... 8

1.2.2. Phát biểu bài toán khai phá luật kết hợp ..... 10

1.2.3. Thuật toán Apriori khám phá tập mục thường xuyên ..... 12

1.3. Mở rộng bài toán khai phá tập mục thường xuyên ..... 18

1.4. Kết luận chương ..... 19

**Chương 2. KHAI PHÁ LUẬT KẾT HỢP CÓ TRỌNG SỐ ..... 20**

2.1. Mở đầu ..... 20

2.2. Khai phá luật kết hợp có trọng số không chuẩn hóa ..... 21

2.2.1. Mô hình bài toán ..... 21

2.2.2. Thuật toán MINWAL(O) khai phá tập mục thường xuyên có trọng số ..... 24

2.2.2.1. Cơ sở toán học ..... 24

2.2.2.2. Thuật toán MINWAL(O) ..... 27

2.3. Khai phá luật kết hợp có trọng số chuẩn hóa ..... 34

2.3.1. Mô hình bài toán ..... 34

2.3.2. Thuật toán MINWAL(W) khai phá tập mục thường xuyên có trọng số chuẩn hóa ..... 37

2.3.2.1. Cơ sở toán học ..... 37
2.3.2.2. Thuật toán MINWAL(W) ..... 37
2.2.3. Lập trình và tính toán thử nghiệm ..... 45
2.4. Kết luận chương ..... 46
**Chương 3. KHAI PHÁ LUẬT KẾT HỢP CÓ TRỌNG SỐ BẰNG PHƯƠNG PHÁP CHỌN MẪU ..... 47**
3.1 Tổng thể và mẫu trong thống kê toán học ..... 47
3.2. Thuật toán khai phá luật kết hợp có trọng số dựa vào chọn mẫu ..... 50
3.2.1. Xác định cỡ mẫu ..... 50
3.2.2. Thuật toán ..... 53
3.3. Lập trình và tính toán thử nghiệm ..... 54
3.4. Kết luận chương ..... 55
**KẾT LUẬN ..... 57**
**TÀI LIỆU THAM KHẢO ..... 58**
**PHỤ LỤC 1: Chương trình nguồn thuật toán MINWAL(O) ..... 61**
**PHỤ LỤC 2: Chương trình nguồn thuật toán MINWAL(O) ..... 73**
**PHỤ LỤC 3: Chương trình nguồn thuật toán SRS ..... 81**

# DANH MỤC CÁC KÝ HIỆU VÀ CHỮ VIẾT TẮT

**Các ký hiệu:**

$I = \{i_1,...,i_M\}$: Tập tất cả *M* mục dữ liệu của cơ sở dữ liệu giao tác.

$DT = \{T_1,T_2,...,T_N\}$ : Cơ sở dữ liệu *DT* gồm *N* giao tác

*X*, *Y*, ... : Các tập con của tập tất cả các mục trong cơ sở dữ liệu giao tác.

*X* = *abc* thay cho $X = \{a,b,c\}$ trong các ví dụ.

$SC(X)$: Số đếm hỗ trợ tập mục *X* (hay số giao tác chứa tập mục *X*).

sup(*X*) : Độ hỗ trợ của tập mục *X*.

Wsup(*X*) : Độ hỗ trợ có trọng số của tập mục *X*.

NWsup(*X*) : Độ hỗ trợ có trọng số chuẩn hóa của tập mục *X*.

*minsup* : Ngưỡng độ hỗ trợ tối thiểu.

*wminsup* : Ngưỡng độ hỗ trợ có trọng số tối thiểu.

*nwminsup* : Ngưỡng độ hỗ trợ có trọng số chuẩn hóa tối thiểu.

$\sup(X \to Y)$: Độ hỗ trợ của luật kết hợp $X \to Y$.

$conf(X \to Y)$: Độ tin cậy của luật kết hợp $X \to Y$.

$|A|$ : Lực lượng (bản số) của tập hợp *A*.

$\lceil r \rceil$ : Cận trên nguyên nhỏ nhất của số thực *r*.

Pr(*E*) : Xác suất xủa biến cố ngẫu nhiên *E*.

N(0,1) : Phân phối chuẩn chuẩn tắc.

$z_{1-\alpha/2}$ : Phân vị mức $1-\alpha/2$ của phân phối chuẩn chuẩn tắc.

**Viết tắt:**

CNTT: Công nghệ Thông tin.

CSDL: Cơ sở dữ liệu.

# DANH MỤC CÁC BẢNG BIỂU

*Trang*

Bảng 1.1. Biểu diễn ngang của cơ sở dữ liệu giao tác ..... 9
Bảng 1.2. Biểu diễn dọc của cơ sở dữ liệu giao tác ..... 9
Bảng 1.3. Ma trận giao tác của cơ sở dữ liệu bảng 1.1 ..... 9
Bảng 1.4. Cơ sở dữ liệu giao tác minh hoạ thực hiện thuật toán Apriori ..... 16
Bảng 2.1. Cơ sở dữ liệu giao tác ví dụ ..... 26
Bảng 2.2. Trọng số của các mục trong CSDL giao tác 2.1 ..... 26

# LỜI MỞ ĐẦU

*Khai phá luật kết hợp* là một kỹ thuật quan trọng, có nhiều ứng dụng của khai phá dữ liệu. Mô hình đầu tiên (mô hình nhị phân) của bài toán khai phá luật kết hợp được đề xuất bởi Agrawal và cộng sự vào năm 1993, trong công trình nghiên cứu phát hiện các mối quan hệ (luật kết hợp) giữa các mặt hàng (mục dữ liệu - items) trong cơ sở dữ liệu giao tác của các siêu thị [4, 5]. Sau công trình kinh điển này, vấn đề khai phá luật kết hợp trong cơ sở dữ liệu (CSDL) giao tác được rất nhiều nhà nghiên cứu lý thuyết và ứng dụng quan tâm. Nhiều thuật toán mới, hiệu quả khai phá luật kết hợp, cũng như mô hình mở rộng bài toán đã được các nhà nghiên cứu đề xuất [8, 9].

Mô hình nhị phân của bài toán khai phá luật kết hợp có một số hạn chế, không đáp ứng được những đòi hỏi khác nhau của người sử dụng. Một trong những hạn chế là trong mô hình này tất cả các mục dữ liệu được xử lý như nhau (xuất hiện hay không xuất hiện trong một giao tác), nhưng trên thực tế chúng có tầm quan trọng khác nhau. Nhằm khắc phục hạn chế này người ta đã đề xuất mô hình bài toán *khai phá luật kết hợp có trọng số*, trong đó các mục dữ liệu được gán cho các trọng số khác nhau tùy theo mức độ quan trọng của chúng trong việc mang lại lợi nhuận kinh doanh [3, 7, 8, 18].

Những năm gần đây, khai phá luật kết hợp có trọng số đã trở thành một đề tài hấp dẫn, một nội dung quan trọng của khai phá dữ liệu, thu hút sự quan tâm của nhiều nhà nghiên cứu và ứng dụng.

Đề tài luận văn của học viên nhằm nghiên cứu bài toán, các thuật toán và tìm hiểu khả năng ứng dụng kỹ thuật khai phá luật kết hợp có trọng số từ các CSDL lớn.

Nội dung chính của luận văn gồm 3 chương:

Chương 1 trình bày khái quát về khai phá dữ liệu, tóm tắt quá trình khai phá, các kỹ thuật, các ứng dụng và những thách thức; bài toán khai phá luật kết hợp nhị phân và thuật toán cơ bản Apriori.

Chương 2 trình bày hai mô hình mở rộng bài toán khai phá luật kết hợp nhị phân: Khai phá luật kết hợp có trọng số và khai phá luật kết hợp có trọng số chuẩn hóa, cùng với các giải thuật tương ứng.

01) Chương 3 trình bày cách tiếp cận bài toán khai phá luật kết hợp có trọng số bằng phương pháp lấy mẫu ngẫu nhiên từ CSDL ban đầu.

*Thái Nguyên, tháng 09 năm 2012.*

Học viên

**Phạm Đức Quang**