ĐẠI HỌC THÁI NGUYÊN

**TR- ỜNG 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 TỐ

## LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

**THÁI NGUYÊN - NĂM 2014**

ĐẠI HỌC THÁI NGUYÊN

**TR- ỜNG 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 TỐ

**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 Luyến**

# 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 Luyến**

# MỤC LỤC

# DANH MỤC CÁC HÌNH VẼ

# DANH MỤC CÁC BẢNG

# 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.

Conf(X->Y): Độ tin cậy của một luật X->Y

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

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

I = {$i_1$, $i_2$,…, $i_n$}: 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->X)

P(Y/X): Xác suất có điều kiện (độ tin cậy của luật kết hợp X->Y)

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 |