ĐẠI HỌC THÁI NGUYÊN

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

**NGUYỄN TÚ NAM**

# KHAI PHÁ TẬP MỤC THƯỜNG XUYÊN CÓ TRỌNG SỐ TRÊN CƠ SỞ DỮ LIỆU GIAO TÁC

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

**THÁI NGUYÊN - 2015**

# LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của tôi dưới sự hướng dẫn của TS Nguyễn Long Giang.

Các số liệu, kết quả nghiên cứu trong luận văn là trung thực và mọi trích dẫn trong báo cáo đều được ghi rõ nguồn gốc. Nếu có sử dụng bất hợp pháp kết quả công trình nghiên cứu của người khác trong báo cáo tôi xin hoàn toàn chịu trách nhiệm.

Tác giả

**Nguyễn Tú Nam**

# LỜI CẢM ƠN

Lời đầu tiên tôi muốn bày tỏ lòng biết ơn sâu sắc và kính trọng của mình tới thầy giáo, TS Nguyễn Long Giang. Trong quá trình tìm hiểu nghiên cứu để hoàn thành luận văn tôi gặp không ít khó khăn, nhưng những lúc như vậy tôi luôn nhận được sự động viên khích lệ của thầy. Thầy đã giúp đỡ tôi rất nhiều trong quá trình nghiên cứu, hướng dẫn tận tình trong cách thức và phương pháp nghiên cứu khoa học cũng như hỗ trợ tôi trong việc tìm tài liệu.

Để có được những kết quả trong luận văn này, tôi xin gửi lời cảm ơn sâu sắc đến Thầy, Cô Trường Đại học Công nghệ thông tin và Truyền thông Thái Nguyên đã tạo điều kiện cho tôi được học hỏi kiến thức thông qua các môn học cũng như hoàn thành khóa học.

Cuối cùng tôi xin bày tỏ lòng cảm ơn chân thành đến gia đình, người thân và bạn bè đồng nghiệp đã khích lệ và động viên tôi hoàn thành luận văn này.!

# MỤC LỤC

# Danh mục các ký hiệu, các chữ viết tắt

| **Ký hiệu, chữ viết tắt** | **Diễn giải** |
|---|---|
| *CSDL* | *Cơ sở dữ liệu* |
| *TID* | *Transction Identifcation* |
| *W* | *Tập các trọng số của các mục* |
| *L* | *Tập tất cả các mục thường xuyên* |
| $C_k$ | *Tập các k-tập mục ứng viên* |
| $L_k$ | *Tập các k-tập mục thường xuyên* |
| *SC(X)* | *Số đếm hỗ trợ của các tập mục X* |
| $WFI_k$ | *Tập các k-tập mục thường xuyên có trọng số* |
| *WFI* | *Tập tất cả các tập mục thường xuyên có trọng số* |
| *MaxW* | *Trọng số có giá trị lớn nhất trong CSDL giao tác* |
| *MinW* | *Trọng số có giá trị nhỏ nhất trong tập mục điều kiện* |
| *min_weight* | *Ngưỡng trọng số tối thiểu* |
| *min_sup* | *Ngưỡng hỗ trợ tối thiểu* |
| *support* | *Độ hỗ trợ của các tập mục* |
| *conf* | *Độ tin cậy* |
| *minconf* | *Độ tin cậy cực tiểu* |
| *BFS* | *Breadth First Search* |
| *DFS* | *Depth First Search* |
| *WFIM* | *Weighted Frequent Itemset Mining* |

# Danh mục các bảng

# Danh mục các hình

# MỞ ĐẦU

## Lý do chọn đề tài

Khai phá dữ liệu và khám phá tri thức (Data mining and Knowledge discovery) là một lĩnh vực quan trọng của ngành Công nghệ thông tin. Đây là lĩnh vực đã thu hút đông đảo các nhà khoa học trên thế giới và trong nước tham gia nghiên cứu. Khai phá luật kết hợp (Mining association rules) là bài toán có vai trò quan trọng trong nhiều nhiệm vụ khai phá dữ liệu và có nhiều ứng dụng thực tiễn trong các lĩnh vực khác nhau của đời sống, đặc biệt là trong lĩnh vực kinh doanh.

Khai phá luật kết hợp được giới thiệu bởi Agrawal vào năm 1993 khi phân tích cơ sở dữ liệu bán hàng của siêu thị, phân tích sở thích mua của khách hàng bằng cách tìm ra những mặt hàng khác nhau được khách hàng mua cùng trong một lần mua. Những thông tin như vậy sẽ giúp người quản lý kinh doanh tiếp thị chọn lọc và thu xếp không gian bày hàng hợp lý hơn, giúp cho kinh doanh hiệu quả hơn. Bài toán khai phá luật kết hợp bao gồm hai bài toán con. Bài toán thứ nhất là tìm các *tập mục thường xuyên* (Frequent itemset) thỏa mãn ngưỡng hỗ trợ tối thiểu cho trước, bài toán thứ hai là sinh ra các *luật kết hợp* (Association rule) thỏa mãn ngưỡng tin cậy cho trước từ tập mục thường xuyên tìm được. Mọi khó khăn của bài toán khai phá luật kết hợp tập trung ở bài toán thứ nhất, đó là khai phá tất cả các tập mục thường xuyên thỏa mãn ngưỡng độ hỗ trợ cho trước, và các nghiên cứu về khai phá luật kết hợp tập trung vào bài toán khai phá tập mục thường xuyên.

Xuất phát từ những lợi ích thực tế trên tác giả đã mạnh dạn chọn đề tài *"***KHAI PHÁ TẬP MỤC THƯỜNG XUYÊN CÓ TRỌNG SỐ TRÊN CƠ SỞ DỮ LIỆU GIAO TÁC***"* làm đề tài nghiên cứu cho luận văn tốt nghiệp của mình.

**Mục tiêu đề tài** tiếp tục nghiên cứu và đề xuất các thuật toán khai phá tập mục thường xuyên có trọng số trong CSDL giao tác

Xây dựng và đề xuất một số giải thuật khai phá tập mục thường xuyên có trọng số.

Lập trình, thử nghiệm các giải thuật khai phá tập mục thường xuyên có trọng số.

**Đối tượng nghiên cứu** các cơ sở dữ liệu giao tác được cập nhật từ kho dữ liệu mẫu UCI

**Phạm vi nghiên cứu** nghiên cứu và thử nghiệm bài toán khai phá tập mục thường xuyên có trọng số trên cơ sở dữ liệu giao tác.

**Phương pháp nghiên cứu** luận văn là nghiên cứu lý thuyết và nghiên cứu thực nghiệm. Về nghiên cứu lý thuyết: các định lý, mệnh đề trong luận văn được chứng minh dựa vào các kiến thức cơ bản và các kết quả nghiên cứu đã công bố. Về nghiên cứu thực nghiệm luận văn thực hiện cài đặt các thuật toán, chạy thử nghiệm thuật toán.

**Bố cục luận văn**

Luận văn được chia làm 3 chương:

*Chương 1*: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

*Chương 2*: KHAI PHÁ TẬP MỤC THƯỜNG XUYÊN CÓ TRỌNG SỐ

*Chương 3*: ĐÁNH GIÁ CÁC THUẬT TOÁN VÀ ỨNG DỤNG