**ĐẠI HỌC THÁI NGUYÊN**
**TRƯỜNG ĐẠI HỌC CNTT VÀ TRUYỀN THÔNG**

ĐÀO THỊ THÚY QUỲNH

# NGHIÊN CỨU PHƯƠNG PHÁP TÌM TẬP THƯỜNG XUYÊN SỬ DỤNG CÂY TIỀN TỐ NÉN

THÁI NGUYÊN 2015

# LỜI CẢM ƠN

Luận văn này được hoàn thành với sự hướng dẫn tận tình của PGS.TS Ngô Quốc Tạo – Viên Công nghệ thông tin - Viện Hàn Lâm Khoa học Việt Nam. Trước tiên tôi xin chân thành bày tỏ lòng biết ơn sâu sắc tới PGS.TS Ngô Quốc Tạo người đã tận tình hướng dẫn, động viên giúp đỡ tôi trong suốt thời gian thực hiện luận văn. Tôi cũng xin chân thành cảm ơn các thầy cô trong trường Công Nghệ thông tin và Truyền thông – Đại học Thái Nguyên, tạo điều kiện thuận lợi cho tôi hoàn thành tốt khóa học.

Xin chân thành cảm ơn các anh, các chị và các bạn học viên lớp Cao học CHK12A đã luôn động viên, giúp đỡ và nhiệt tình chia sẻ với tôi những kinh nghiệm học tập, công tác trong suốt khoá học.

Cuối cùng, tôi xin gửi lời cảm ơn sâu sắc đến gia đình, người thân, bạn bè đã động viên, khuyến khích và hỗ trợ cần thiết để tôi hoàn thành luận văn này.

Mặc dù rất cố gắng, song luận văn này không thể tránh khỏi những thiếu sót, kính mong được sự chỉ dẫn của các quý thầy cô và các bạn.

*Thái Nguyên, ngày 16 tháng 05 năm 2015*

**Người viết**

**Đào Thị Thúy Quỳnh**

# LỜI CAM ĐOAN

Tôi xin cam đoan rằng số liệu và kết quả nghiên cứu trong luận văn này là trung thực và không trùng lặp với các đề tài khác. Tôi cũng xin cam đoan rằng mọi sự giúp đỡ cho việc thực hiện luận văn này đã được cảm ơn và các thông tin trích dẫn trong luận văn đã được chỉ rõ nguồn gốc.

*Thái Nguyên, ngày 16 tháng 05 năm 2015*

**Người cam đoan**

**Đào Thị Thúy Quỳnh**

# BẢNG KÝ HIỆU CHỮ VIẾT TẮT

| TT | Ký hiệu viết tắt | Giải thích |
|---|---|---|
| 1 | CNTT | Công nghệ thông tin |
| 2 | KPDL | Khai phá dữ liệu |
| 3 | CSDL | Cơ sở dữ liệu |
| 4 | KDD | Khám phá tri thức trong cơ sở dữ liệu (Knowledge Discovery in Databases) |
| 5 | ITL | Item - TransLink |
| 6 | CT-ITL | Compressed Tree - Item TransLink |
| 7 | CFP | Compressed FP - Tree |
| 8 | FP - Tree | Frequent pattern Tree |
| 9 | $D = \{T_1, T_2, \ldots, T_n\}$. | Tập hợp n giao dịch |
| 10 | $I= \{i_1, i_2, \ldots, i_m\}$ | Tập hợp m phầntử trong CSDL |
| 11 | *Minsup* | Ngưỡng độ hỗ trợ |
| 12 | *Minconf* | Ngưỡng độ tin cậy tối thiểu |
| 13 | Conditional pattern - base | Cơ sở mẫu có điều kiện |
| 14 | Conditional FP-Tree | Cây FP có điều kiện |

# DANH MỤC CÁC BẢNG

# DANH MỤC CÁC HÌNH

# MỤC LỤC

# MỞ ĐẦU

Ngày nay, cùng với sự phát triển của công nghệ thông tin (CNTT) là khả năng thu thập và lưu trữ thông tin của các hệ thống thông tin tăng một cách chóng mặt. Bên cạnh đó, việc tin học hóa nhanh chóng trong nhiều lĩnh vực đời sống văn hóa xã hội, quản lý kinh tế, khoa học kỹ thuật cũng như nhiều lĩnh vực khác đã tạo cho chúng ta một lượng dữ liệu khổng lồ cần lưu trữ.

Sự bùng nổ này dẫn tới một yêu cầu cấp thiết là cần có những kỹ thuật và công cụ mới để tự động chuyển đổi lượng dữ liệu khổng lồ thành các tri thức có ích. Từ đó, bên cạnh những phương pháp khai thác thông tin truyền thống xuất hiện một khuynh hướng kỹ thuật mới ra đời đó là Khai phá dữ liệu (Datamining) một lĩnh vực quan trọng của ngành CNTT. Khai phá dữ liệu (KPDL) đang được áp dụng một cách rộng rãi trong nhiều lĩnh vực đời sống như : marketing, tài chính – ngân hàng, bảo hiểm, khoa học, y tế, an ninh, internet…Rất nhiều tổ chức và công ty lớn trên thế giới đã áp dụng kỹ thuật KPDL vào các hoạt động sản xuất kinh doanh của mình và thu được lợi ích to lớn.

Khai phá tập thường xuyên đóng vai trò thiết yếu trong KPDL, nó là nền tảng cho các nhiệm vụ KPDL khác như khai phá luật kết hợp, phân lớp, phân cụm dữ liệu, tìm kiếm mối tương quan, và các mối quan hệ trong cơ sở dữ liệu. Do vậy, khai phá tập thường xuyên đã trở thành nhiệm vụ quan trọng trong KPDL.Có rất nhiều thuật toán được đề xuất với mục đích khai phá tập thường xuyên nhanh và chính xác. Tuy nhiên với cơ sở dữ liệu lớn rất cần một cấu trúc dữ liệu nhỏ gọn lưu trữ trên bộ nhớ và hiệu quả trong khai phá tập thường xuyên.Từ những nhận định trên và được sự gợi ý của giáo viên hướng dẫn, tôi quyết định chọn đề tài: **"Nghiên cứu phương pháp tìm tập thường xuyên sử dụng cây tiền tố nén".**

Nhiệm vụ chính của luận văn là nắm vững kiến thức tổng quan của lĩnh vực KPDL, nghiên cứu một số thuật toán khai phá tập thường xuyên, nghiên cứu thuật toán khai phá tập thường xuyên trên cây tiền tố nén lấy điển hình trên cấu trúc CFP (Compressed FP-Tree) và cấu trúc CT-ITL (Compressed Tree - Item TransLink) sau đó cài đặt chương trình thử nghiệm, đánh giá, so sánh hiệu quả của thuật toán khai

phá tập thường xuyên trên cây CFP với thuật toán Apriori và FP-Growth ( Những thuật toán điển hình trong khai phá tập thường xuyên).

**Mục tiêu của luận văn:**

- Nắm vững kiến thức tổng quan của lĩnh vực Khai phá dữ liệu.
- Nghiên cứu một số thuật toán khai phá tập thường xuyên.
- Nghiên cứu thuật toán khai phá tập thường xuyên sử dụng cấu trúc dữ liệu và thuật toán CT-ITL.
- Nghiên cứu thuật toán khai phá tập thường xuyên trên cây FP nén sử dụng thuật toán CT-PRO và cài đặt chương trình thực nghiệm đánh giá, so sánh hiệu quả của thuật toán này với một số thuật toán khác trong khai phá tập thường xuyên.

**Phương pháp nghiên cứu:**

- Kết hợp lý thuyết với đánh giá thực nghiệm
- Sưu tầm và tổng hợp các kết quả nghiên cứu về tập mục thường xuyên, Khai phá tập mục thường xuyên từ nguồn sách và các bài báo khoa học, hội thảo chuyên ngành trong nước và ngoài nước.

**Một số kết quả nghiên cứu đạt được:**

- Tổng kết kiến thức cơ bản về khai phá dữ liệu và khai phá tập thường xuyên. Trình bày hai thuật toán cơ bản trong khai phá tập thường xuyên: thuật toán Apriori, thuật toán tăng trưởng mẫu FP-Growth.
- Trình bày chi tiết hai thuật toán khai phá tập thường xuyên trên cây tiền tố nén là cấu trúc cây FP nén và cấu trúc CT-ITL.
- Luận văn đã tiến hành cài đặt ba thuật toán Apriori, FP-Growth và thuật toán CT-PRO sau đó đánh giá, so sánh tốc độ thực hiện của ba thuật toán này trên nhiều CSDL lớn.

**Ý nghĩa khoa học của đề tài:**

- Làm rõ tầm quan trọng của khai phá tập thường xuyên.
- Để có cái nhìn tổng quan, chi tiết về mỗi thuật toán và thảo luận về ý tưởng tối ưu hóa của mỗi thuật toán.