ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

TRẦN THỊ THU TRANG

KHAI PHÁ LUẬT KẾT HỢP TỪ DỮ LIỆU CHUỖI THỜI GIAN

LUẬN VĂN THẠC SĨ: KHOA HỌC MÁY TÍNH

Thái Nguyên - 2012

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn "Khai phá luật kết hợp từ dữ liệu chuỗi thời gian" là công trình nghiên cứu của riêng tôi dưới sự hướng dẫn của PGS.TS. Bùi Thế Hồng. Toàn bộ phần mềm do chính tôi lập trình và kiểm thử. Tôi xin chịu trách nhiệm về lời cam đoan của mình.

Các số liệu và thông tin sử dụng trong luận văn này hoàn toàn là trung thực.

Tác giả

Trần Thị Thu Trang

MỤC LỤC

DANH MỤC HÌNH VẼ

DANH MỤC CÁC BẢNG

DANH MỤC CÁC KÝ HIỆU, CÁC TỪ VIẾT TẮT

Các từ viết tắt	Nghĩa tiếng anh	Nghĩa tiếng việt
FI	Frequent Itemset	Tập mục thường xuyên
FCI	Frequent Closed Itemset	Tập mục thường xuyên đóng
MFI	Maximally Frequent Itemset	Tập mục thường xuyên lớn nhất
CSDL		Cơ sở dữ liệu
FP-Tree	Frequent Pattern Tree	Cây mẫu thường xuyên
CFP-Tree	Compressed Frequent Pattern Tree	Cây mẫu thường xuyên nén
ITARM	Inter-Transaction Association Rules Mining	Khai phá luật kết hợp liên giao dịch

MỞ ĐẦU

Ngày nay, cuộc cách mạng của kỹ thuật số cho phép số hóa thông tin dễ dàng và chi phí lưu trữ thấp.Với sự phát triển của phần mềm, phần cứng và trang bị nhanh hệ thống máy tính trong kinh doanh. Số lượng dữ liệu khổng lồ được tập trung và lưu trữ trong cơ sở dữ liệu. Dữ liệu sau khi phục vụ cho một mục đích nào đó được lưu lại trong kho dữ liệu và theo ngày tháng khối lượng dữ liệu được lưu trữ ngày càng lớn. Trong khối lượng dữ liệu to lớn này có rất nhiều thông tin có ích mang tính tổng quát, thông tin có tính quy luật vẫn còn đang tiềm ẩn mà chúng ta chưa biết. Từ khối lượng dữ liệu rất lớn cần có những công cụ tự động rút các thông tin và kiến thức có ích. Một hướng tiếp cận có khả năng giúp các công ty khai thác các thông tin có nhiều ý nghĩa từ các tập dữ liệu lớn đó là khai phá dữ liệu.

Với sự bùng nổ và phát triển của công nghệ thông tin đã mang lại nhiều hiệu quả đối với khoa học cũng như các hoạt động thực tế, trong đó khai phá dữ liệu là một trong những lĩnh vực mang lại hiệu quả thiết thực cho con người. Khai phá dữ liệu đã giúp người sử dụng thu được những tri thức hữu ích từ những cơ sở dữ liệu hoặc các kho dữ liệu khổng lồ khác. Luận văn đề cập đến các khái niệm và vấn đề cơ bản trong khai phá luật kết hợp từ dữ liệu chuỗi thời gian được áp dụng trong cơ sở dữ liệu bán hàng.

Luận văn cấu trúc gồm 3 chương:

Chương 1:

Trong chương 1 tìm hiểu khái quát về khai phá dữ liệu và dữ liệu chuỗi thời gian và phương pháp tiền xử lý dữ liệu chuỗi thời gian.

Chương 2:

Trong chương 2 sẽ tìm hiểu phương pháp khai phá dữ liệu từ chuỗi thời gian qua thuật toán ITARM dựa trên cấu trúc cây CFPTree.

Chương 3:

Trong chương 3 tiến hành cài đặt thuật toán ở chương 2 và cài đặt ứng dụng của thuật toán trên cơ sở dữ liệu bán hàng.

Luận văn này được hoàn thành dưới sự hướng dẫn tận tình của PGS.TS **Bùi Thế Hồng**, em xin bày tỏ lòng biết ơn chân thành của mình đối với thầy. Em xin chân thành cảm ơn các thầy, cô giáo Viện Công nghệ thông tin, Trường Đại học Công nghệ thông tin và Truyền thông - Đại học Thái Nguyên đã tham gia giảng dạy, giúp đỡ em trong suốt qúa trình học tập nâng cao trình độ kiến thức. Tuy nhiên vì điều kiện thời gian và khả năng có hạn nên luận văn không thể tránh khỏi những thiếu sót. Em kính mong các thầy cô giáo và các bạn đóng góp ý kiến để đề tài được hoàn thiện hơn.

CHƯƠNG 1: KHÁI QUÁT VỀ KHAI PHÁ DỮ LIỆU VÀ CHUỖI THỜI GIAN

1.1. Khai phá dữ liệu

1.1.1. Khai phá dữ liệu là gì?

Khai phá dữ liệu là một khái niệm ra đời vào những năm cuối của thập kỷ 80. Nó bao hàm một loạt các kỹ thuật nhằm phát hiện ra các thông tin có giá trị tiềm ẩn trong các tập dữ liệu lớn(các kho dữ liệu). Về bản chất, khai phá dữ liệu liên quan đến việc phân tích các dữ liệu và sử dụng các kỹ thuật để tìm ra các mẫu hình có tính chính quy trong tập dữ liệu.

Năm 1989, Fayyad, Piatestsky-Shapiro và Smyth đã dùng khái niệm *Phát hiện tri thức trong cơ sở dữ liệu* để chỉ toàn bộ quá trình phát hiện các tri thức có ích từ các tập dữ liệu lớn. Trong đó, *khai phá dữ liệu* là một bước đặc biệt trong toàn bộ quá trình, sử dụng các giải thuật đặc biệt để chiết xuất ra các mẫu (hay các mô hình) từ dữ liệu.

Ở một mức độ trừu tượng nhất định có thể định nghĩa về khai phá dữ liệu: *Khai phá dữ liệu* là một quá trình tìm kiếm, phát hiện các tri thức mới, tiềm ẩn, hữu dụng trong cơ sở dữ liệu lớn.

Khám phá tri thức là mục tiêu chính của khai phá dữ liệu, do vậy hai khái niệm đó được xem như hai lĩnh vực tương đương nhau. Nhưng, nếu phân chia một cách tách bạch thì khai phá dữ liệu là một bước chính trong quá trình khám phá tri thức.

1.1.2. Nhiệm vụ của khai phá dữ liệu

Các bài toán liên quan đến khai phá dữ liệu về bản chất là các bài toán thống kê. Điểm khác biệt giữa các kỹ thuật khai phá dữ liệu và các công cụ phục vụ tính toán thống kê mà chúng ta đã biết là ở khối lượng cần tính toán. Một khi dữ liệu đã trở nên khổng lồ thì những khâu như: thu thập dữ liệu, tiền xử lý và xử lý dữ liệu đều đòi hỏi phải được tự động hóa. Tuy