ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

BẾ QUANG HUẤN

NGHIÊN CỨU MỘT SỐ THUẬT TOÁN KHAI PHÁ TẬP MỤC THƯỜNG XUYÊN VÀ TẬP MỤC CỔ PHẦN CAO TRONG CƠ SỞ DỮ LIỆU

***Chuyên nghành*: Khoa học máy tính**

***Mã số*: 60.48.01**

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Người hướng dẫn khoa học: GS. TS Vũ Đức Thi

THÁI NGUYÊN 2012

LỜI CAM ĐOAN

Tôi xin cam đoan toàn bộ nội dung trong Luận văn hoàn toàn theo đúng nội dung đề cương cũng như nội dung mà cán bộ hướng dẫn giao cho. Nội dung luận văn, các phần trích lục các tài liệu hoàn toàn chính xác. Nếu có sai sót tôi hoàn toàn chịu trách nhiệm.

Tác giả luận văn

Bế Quang Huấn

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU VÀ CÁC CHỮ VIẾT TẮT

Ký hiệu	Diễn giải
$I=\{i_1,i_2,...,i_n\}$	Tập n mục dữ liệu
$DB=\{T_1,T_2,...,T_m\}$	Cơ sở dữ liệu có m giao tác
db	Cơ sở dữ liệu giao tác con của *DB*, $db \subseteq DB$
i_p	Mục dữ liệu thứ p
T_q	Giao tác thứ q
n	Số mục dữ liệu một cơ sở dữ liệu giao tác
m	Số giao tác của một cơ sở dữ liệu giao tác
A, B, C,…	Tên các mục dữ liệu trong cơ sở dữ liệu giao tác
X, Y,…	Tập con của tập mục dữ liệu I, X, $Y \subseteq I$
X=ABC	Thay cho X={A,B,C} trong các cơ sở dữ liệu giao tác
minsup	Ngưỡng độ hỗ trợ
minShare	Ngưỡng cổ phần tối thiểu
minconf	Ngưỡng độ tin cậy tối thiểu
$\|X\|$	Số phần tử của tập hợp X
CSDL	Cở sở dữ liệu
CNTT	Công nghệ thông tin

DANH MỤC CÁC BẢNG BIỂU

DANH MỤC CÁC HÌNH VẼ, ĐỒ THỊ

MỞ ĐẦU

Một trong những ứng dụng quan trọng nhất của công nghệ thông tin trong đời sống là giúp giải quyết các bài toán quản lý. Kể từ khi máy tính điện tử trở thành một công cụ lao động quan trọng thì một trong những nhu cầu đầu tiên là lưu trữ, tìm kiếm và xử lý số liệu thống kê. Đến nay, các cơ sở dữ liệu đã trở nên khổng lồ và người ta mong muốn kho dữ liệu đó cần được khai thác hiệu quả hơn trên nhiều bình diện. Trong những năm gần đây, khai phá dữ liệu (Data mining) đã trở thành một trong những hướng nghiên cứu lớn nhất của lĩnh vực khoa học máy tính và công nghệ thông tin. Khai phá dữ liệu đang được áp dụng một cách rộng rãi trong nhiều lĩnh vực kinh doanh và đời sống khác nhau: marketing, tài chính, ngân hàng và bảo hiểm, khoa học, y tế, an ninh, internet…

Khai phá dữ liệu và khám phá tri thức (Data Mining and Knowledge Discovery) 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 tập mục thường xuyên là bài toán có vai trò quan trọng trong nhiều nhiệm vụ khai phá dữ liệu. Khai phá tập mục thường xuyên được biết đến ban đầu là bài toán con của bài toán 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 trọn lọc và thu xếp không gian bày hàng hợp lý hơn, giúp cho kinh doanh hiệu quả hơn.

Mô hình khai phá tập mục thường xuyên cơ bản có nhiều ứng dụng trong thực tế nhưng có những hạn chế, không đáp ứng đầy đủ yêu cầu của người sử dụng.

Để đáp ứng nhu yêu cầu của thực tiễn, một số hướng mở rộng bài toán đã được quan tâm nghiên cứu. Một hướng mở rộng bài toán có rất nhiều ứng dụng là quan tâm đến cấu trúc dữ liệu và mức độ quan trọng khác nhau của các mục dữ liệu, các thuộc tính trong cơ sở dữ liệu. Theo hướng này, từ bài toán khai phá tập mục thường xuyên ban đầu, nhiều nhà nghiên cứu đề xuất các mô hình mở rộng: Khai phá tập mục cổ phần cao, đánh giá sự đóng góp của tập mục dữ liệu trong tổng số các mục dữ liệu của cơ sở dữ liệu.

Trên thế giới, các kết quả nghiên cứu về khai phá tập mục cổ phần cao đã được công bố nhiều từ các nhóm nghiên cứu tại một số trường đại học ở Mỹ, Canada, Úc, Đài Loan, Singapo, Hồng Kông,… Tại Việt Nam, Khai phá luật kết hợp đã được các nhóm nghiên cứu tại Viện Công nghệ Thông tin thuộc Viện Khoa học và Công nghệ Việt Nam, các nhóm nghiên cứu tại một số trường đại học như Đại học Quốc gia Hà Nội, Đại học Bách Khoa Hà Nội, Đại học Quốc gia thành phố Hồ Chí Minh thực hiện và đã có nhiều kết quả được công bố.

Với mục đích đóng góp vào lĩnh vực nghiên cứu này, tôi đã chọn đề tài luận văn: ***" Nghiên cứu một số thuật toán khai phá tập mục thường xuyên và tập mục cổ phần cao trong cơ sở dữ liệu"*** làm chủ đề nghiên cứu của mình.

Mục đích của luận văn là phát triển một số thuật toán khai phá tập mục cổ phần cao trong cơ sở dữ liệu giao tác cỡ lớn. Trên cơ sở đó áp dụng vào một bài toán cụ thể là cài đặt trương trình

Với mục tiêu đó, luận văn được trình bày trong ba chương:

Chương 1: Khai phá tập mục thường xuyên và một số mở rộng

Trình bày bài toán khai phá tập mục thường xuyên: Các khái niệm cơ bản và các mô hình khai phá. Sau khi trình bày khái quát các thuật toán khai phá, trong trương trình bày chi tiết hai thuật toán tiêu biểu cho hai cách tiếp cận khác nhau là thuật toán Apriori và thuật toán FP-growth. Thuật toán Apriori tiêu biểu cho phương pháp sinh ra các tập mục ứng viên rồi duyệt cơ sở dữ liệu để tính độ hỗ trợ của nó. Thuật toán FP-growth là thuật toán đầu tiên giới thiệu cấu trúc cây FP-tree nén toàn bộ các giao tác của cơ sở dữ liệu lên cây với 2 lần duyệt, sau đó khai phá theo phương pháp phát triển dần các mẫu ở trên cây mà không cần duyệt cơ sở dữ liệu nữa. Bên cạnh đó luận văn đã trình bày chi tiết phương pháp COFI-tree khai phá cây FP-tree thay cho phương pháp FP-growth.

Chương 2: Khai phá tập mục cổ phần cao

Trình bày mô hình khai phá cổ phần cao, giới thiệu thuật toán FSM là thuật toán nhanh khai phá tất cả các tập mục cổ phần cao trong cơ sở dữ liệu giao tác. Luận văn đề xuất khái niệm "tập mục cổ phần theo giao tác cao" và chứng minh nó có tính chất phản đơn điệu (Anti Monotone), có thể ứng dụng vào nhiều thuật toán khai phá tập mục thường xuyên đã có để tìm được tập mục cổ phần theo giao tác cao, từ đó tìm ra tập mục cổ phần cao. Sử dụng ý tưởng này, luận văn đề xuất thuật toán AFSM *(Advanced FSM)* dựa trên các bước của thuật toán FSM với phương pháp mới tỉa hiệu quả hơn các tập mục ứng viên.

Chương 3: Thực nghiệm và đánh giá thuật toán

Để có được kết quả này tôi đã nhận được sự quan tâm, động viên, giúp đỡ rất nhiều của các Thầy giáo, Cô giáo trong Khoa Công nghệ thông tin - Đại học