ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CNTT&TT

ĐỖ THỊ HẢI YẾN

KHAI PHÁ TẬP MỤC LỢI ÍCH CAO TRONG CƠ SỞ DỮ LIỆU LỚN

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYÊN - 2011

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CNTT&TT

ĐỖ THỊ HẢI YẾN

KHAI PHÁ TẬP MỤC LỢI ÍCH CAO TRONG CƠ SỞ DỮ LIỆU LỚN

CHUYÊN NGÀNH: KHOA HỌC MÁY TÍNH
MÃ SỐ: 60 48 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

HƯỚNG DẪN KHOA HỌC: PGS.TS. NGUYỄN THANH TÙNG

THÁI NGUYÊN - 2011

LỜI CAM ĐOAN

Tôi xin cam đoan Luận văn "Khai phá tập mục lợi ích cao trong cơ sở dữ liệu lớn" là công trình nghiên cứu của riêng tôi dưới sự hướng dẫn của PGS.TS Nguyễn Thanh Tùng. Kết quả đạt được trong luận văn là sản phẩm của riêng cá nhân tôi, không sao chép lại của người khác. Trong toàn bộ luận văn, những điều được trình bày là của cá nhân hoặc 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 30 tháng 9 năm 2011

Người cam đoan

Đỗ Thị Hải Yến

LỜI CẢM ƠN

Lời đầu tiên tôi xin gửi lời cảm ơn chân thành và biết ơn sâu sắc tới PGS.TS. Nguyễn Thanh Tùng - Viện Công nghệ thông tin, người thầy đã chỉ bảo và hướng dẫn tận tình cho tôi trong suốt quá trình nghiên cứu khoa học 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 30 tháng 9 năm 2011

Tác giả

Đỗ Thị Hải Yến

MỤC LỤC

Trang

DANH MỤC CÁC TỪ VIẾT TẮT

STT	Cụm từ viết tắt	Nghĩa của cụm từ viết tắt
1	CNTT	Công nghệ thông tin
2	CSDL	Cơ sở dữ liệu
3	KDD	Khám phá tri thức trong cơ sở dữ liệu

DANH MỤC CÁC BẢNG

Trang

DANH MỤC CÁC HÌNH

Trang

LỜI MỞ ĐẦU

Trong những năm gần đây, cùng với sự phát triển vượt bậc của công nghệ thông tin, truyền thông, khả năng thu thập và lưu trữ thông tin của các hệ thống thông tin không ngừng được nâng cao. Với lượng dữ liệu khổng lồ và luôn gia tăng theo thời gian, rõ ràng các phương pháp phân tích dữ liệu truyền thống sẽ không còn hiệu quả, gây tốn kém và dễ dẫn đến những kết quả sai lệch. Để có thể khai thác hiệu quả các cơ sở dữ liệu lớn, một lĩnh vực khoa học mới đã ra đời: Khám phá tri thức trong cơ sở dữ liệu (Knowledge Discovery in Databases – KDD). Khai phá dữ liệu (Data Mining) là một công đoạn chính trong qúa trình khám phá tri thức, nhằm *tìm kiếm, phát hiện các tri thức mới, hữu ích tiềm ẩn trong các cơ sở dữ liệu lớn.*

Khai phá luật kết hợp là một nhiệm vụ quan trọng của khai phá dữ liệu. Bài toán truyền thống (hay còn gọi bài toán nhị phân) khai phá luật kết hợp do R. Agrawal, T. Imielinski và A. N. Swami đề xuất và nghiên cứu lần đầu tiên vào năm 1993 khi phân tích các cơ sở dữ liệu của các siêu thị. Mục tiêu của nó là phát hiện các *tập mục thường xuyên*, từ đó tạo các *luật kết hợp* phản ánh hành vi mua hàng của khách hàng. Những thông tin như vậy giúp nhà quản lý có thể lựa chọn phương án tiếp thị, kinh doanh hiệu quả hơn.

Cho đến nay, bài toán khai phá luật kết hợp truyền thống có nhiều ứng dụng, tuy vậy do tập mục thường xuyên chỉ mang ngữ nghĩa thống kê nên mô hình bài toán truyền thống chỉ đáp ứng được phần nào nhu cầu ứng dụng thực tiễn. Thật ra, trong kinh doanh, điều mà người quản lý quan tâm hơn là phát hiện những khách VIP, đem lại lợi nhuận cao. Trong thực hành, có những tập mục thường xuyên nhưng chỉ đóng góp phần rất nhỏ, ngược lại có những tập mục không thường xuyên lại đóng góp phần đáng kể vào lợi nhuận chung của công ty.

Gần đây, nhằm khắc phục hạn chế của bài toán truyền thống khai phá luật kết hợp, các nhà nghiên cứu đã mở rộng nó theo nhiều hướng khác nhau, trong đó có vấn đề khai phá tập mục lợi ích cao. Lợi ích của một tập mục là số đo lợi nhuận mà nó có thể mang lại trong kinh doanh, được tính toán dựa trên giá trị khách quan và