ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

NGUYỄN HỒNG SÂM

KHAI PHÁ TẬP MỤC LỢI ÍCH CAO SỬ DỤNG CẤU TRÚC CÂY TIỀN TỐ

Chuyên ngành: Khoa học máy tính

Mã số: 60.48.01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC

TS. NGUYỄN HUY ĐỨC

Thái Nguyên - 2014

LỜI CẢM ƠN

Lời đầu tiên tôi xin gửi lời cảm ơn chân thành và biết ơn sâu sắc tới TS. Nguyễn Huy Đức – Trường Cao đẳng Sư phạm Trung ương, người 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 12 tháng 03 năm 2014

Tác giả

Nguyễn Hồng Sâm

LỜI CAM ĐOAN

Tôi xin cam đoan Luận văn ***"Khai phá tập mục lợi ích cao sử dụng cấu trúc cây tiền tố"*** được thực hiện theo đúng mục tiêu đề ra dưới sự hướng dẫn của TS. Nguyễn Huy Đức. 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 loại tài liệu đều có xuất xứ rõ ràng và được trích dẫn hợp pháp. Tôi xin chịu trách nhiệm về luận văn của mình.

Thái Nguyên, ngày 12 tháng 03 năm 2014

Tác giả

Nguyễn Hồng Sâm

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU VÀ CÁC CHỮ VIẾT TẮT

Trong luận văn này, dùng thống nhất các ký hiệu và chữ viết tắt sau:

Các ký hiệu:

$I = \{i_1, i_2, \ldots, i_n\}$: Tập n mục dữ liệu.

$DB = \{T_1, T_2, \ldots, T_m\}$: 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 một cơ sở dữ liệu giao tác.

A, B, C,…: Tên các mục dữ liệu trong cơ sở dữ liệu giao tác ví dụ.

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 ví dụ.

Nếu X $\subseteq$ Y thì X gọi là tập con của tập Y, Y gọi là tập cha của tập X.

minsup: Ngưỡng độ hỗ trợ tối thiểu.

minutil: Giá trị lợi ích tối thiểu.

$|X|$: Số phần tử của tập hợp X.

Viết tắt:

KPDL: Khai phá dữ liệu

CSDL: Cơ sở dữ liệu.

CNTT: Công nghệ thông tin.

CNTT và TT: Công nghệ Thông tin và Truyền thông.

DANH MỤC CÁC BẢNG

DANH MỤC HÌNH VẼ

LỜI MỞ ĐẦU

Khai phá tập mục thường xuyên đóng vai trò quan trọng trong nhiều nhiệm vụ khai phá dữ liệu. Khai phá tập mục thường xuyên xuất hiện như là bài toán con của nhiều lĩnh vực khai phá dữ liệu như khám phá luật kết hợp, khám phá mẫu tuần tự,… Bài toán khai phá luật kết hợp do Agrawal, T.Imielinski và A. N. Swami đề xuất và nghiên cứu lần đầu vào năm 1993 với mục tiêu là phát hiện các tập mục thường xuyên, từ đó tạo các luật kết hợp. Trong mô hình của bài toán này, giá trị của mỗi mục dữ liệu trong một giao tác là 0 hoặc 1, tức là chỉ quan tâm mục dữ liệu có xuất hiện trong giao tác hay không. Bài toán cơ bản này có nhiều ứng dụng, tuy vậy, do tập mục thường xuyên chỉ mang ngữ nghĩa thống kê nên nó chỉ đáp ứng được phần nào nhu cầu của thực tiễn.

Nhằm khắc phục hạn chế của bài toán cơ bản khai phá luật kết hợp, nhiều nhà nghiên cứu đã mở rộng bài toán theo nhiều hướng khác nhau. Năm 1997, Hilderman và các cộng sự đề xuất bài toán khai phá tập mục cổ phần cao. Trong mô hình này, giá trị của mục dữ liệu trong giao tác là một số. Năm 2004, nhóm các nhà nghiên cứu H. Yao, Hamilton và Butz, mở rộng tiếp bài toán, đề xuất mô hình khai phá tập mục lợi ích cao.

Trong mô hình khai phá tập mục lợi ích cao, giá trị của mục dữ liệu trong giao tác là một số (như số lượng đã bán của mặt hàng, gọi là giá trị khách quan), ngoài ra còn có bảng lợi ích cho biết lợi ích mang lại khi bán một đơn vị hàng đó (gọi là giá trị chủ quan). Lợi ích của tập mục là số đo lợi nhuận mà tập mục đó mang lại. Khai phá tập mục lợi ích cao là khám phá tất cả các tập mục có lợi ích không nhỏ hơn ngưỡng lợi ích tối thiểu của người sử dụng.

Trong những năm gần đây, bài toán này đã và đang thu hút sự quan tâm của nhiều nhà nghiên cứu trong và ngoài nước. Đã có một số thuật toán phát hiện tập mục lợi ích cao được đề xuất. Các thuật toán này có thể phân thành hai loại:

- Thuật toán kiểu Apriori (Apriori-like), sinh ra các tập mục ứng viên, duyệt theo chiều rộng.

- Thuật toán không sinh ứng viên, chuyển đổi cơ sở dữ liệu thành cấu trúc cây, duyệt theo chiều sâu để phát hiện các tập mục lợi ích cao. Các thuật toán loại này hiệu quả hơn và tránh được khối lượng tính toán lớn.

Với ý nghĩa đó và mục đích tìm hiểu bài toán tìm tập mục lợi ích cao và các thuật toán khai phá sử dụng cấu trúc cây tiền tố, em đã quyết định lựa chọn đề tài luận văn: **" KHAI PHÁ TẬP MỤC LỢI ÍCH CAO SỬ DỤNG CẤU TRÚC CÂY TIỀN TỐ"**

Nội dung luận văn gồm 3 chương:

Chương 1: Tổng quan về khai phá dữ liệu và khai phá tập mục thường xuyên.

Chương 2: Khai phá tập mục lợi ích cao sử dụng cấu trúc cây tiền tố.

Chương 3: Chương trình thực nghiệm và ứng dụng.