ĐẠI HỌC THÁI NGUYÊN
KHOA CÔNG NGHỆ THÔNG TIN

LÊ THỊ VIỆT HOA

KHAI PHÁ DỮ LIỆU VÀ THUẬT TOÁN KHAI PHÁ LUẬT KẾT HỢP SONG SONG

Chuyên ngành: KHOA HỌC MÁY TÍNH
Mã số ***:*** 60.48.01

LUẬN VĂN THẠC SĨ CÔNG NGHỆ THÔNG TIN

Hướng dẫn khoa học: PGS.TS ĐOÀN VĂN BAN

THÁI NGUYÊN 2008

LỜI CẢM ƠN

Xin chân thành cảm ơn Thầy giáo PGS.TS Đoàn Văn Ban đã tận tình chỉ dạy và hướng dẫn tôi trong suốt thời gian học tập và làm luận văn.

Tôi cũng xin xin lời biết ơn chân thành đến quý Thầy giáo, cô giáo Viện Công nghệ Thông đã tận tình giảng dạy, trang bị cho tôi những kiến thức quý báu trong suốt quá trình học tập tại Khoa.

Xin cảm ơn tất cả các anh chị em học viên Cao học khóa 5, cám ơn cán bộ công chức, giảng viên – Khoa Công nghệ Thông tin - Đại học Thái Nguyên đã tạo điều kiện giúp đỡ tôi trong suốt quá trình học tập và làm luận văn.

Cuối cùng xin cảm ơn gia đình, bạn bè, đồng nghiệp đã giúp đỡ tôi trong suốt thời gian học tập và hoàn thành luận văn này.

Thái Nguyên, tháng 9 năm 2008

Tác giả

Lê Thị Việt Hoa

LỜI CAM ĐOAN

Tôi xin cam đoan đề tài khoa học "***Khai phá dữ liệu và thuật toán khai phá luật kết hợp song song***" này là công trình nghiên cứu của bản thân tôi. Các số liệu và kết quả nghiên cứu nêu trong luận văn này là trung thực, được các tác giả cho phép sử dụng và các tài liệu tham khảo như đã trình bày trong luận văn. Tôi xin chịu trách nhiệm về luận văn của mình.

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU VIẾT TẮT

Ký hiệu	Diễn giải
C_k	Tập các k-itemset ứng viên
$\overline{C_k}$	Tập các k-itemset ứng viên mà TID của giao dịch sinh ra liên kết với tập mục ứng viên
Conf	Độ tin cậy (Confidence)
CFPT	FP-Tree điều kiện cơ sở (Fisst conditional FP-Tree)
D	Cơ sở dữ liệu giao dịch
D_i	Phần thứ i của cơ sở dữ liệu D
Item	Mục
Itemset	Tập mục
I	Tập các mục
KDD	Phát hiện tri thức trong cơ sở dữ liệu (Knowledge Discovery in Database)
CSDL	Cơ sở dữ liệu (Database)
k-itemset	Tập mục gồm k mục
L_k	Tập các k-itemset phổ biến
MPI	Truyền thông điệp
minconf	Ngưỡng tin cậy tối thiểu
minsup	Ngưỡng hỗ trợ tối thiểu
OLAP	Phân tích trực tuyến
OLTP	Xử lý giao dịch trực tuyến
SC	Số đếm hỗ trợ (support count)
sup	Độ hỗ trợ (support)
T	Giao dịch (transaction)
Tid	Định danh của giao dịch
Tid-List	Danh sách các định danh của giao dịch
$X \Rightarrow Y$	Luật kết hợp (với X là tiền đề, Y là hệ quả)

DANH MỤC HÌNH VẼ VÀ BẢNG

Trang

MỞ ĐẦU

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 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.

Cơ sở dữ liệu trong các đơn vị, tổ chức kinh doanh, quản lý khoa học chứa đựng nhiều thông tin tiềm ẩn, phong phú và đa dạng, đòi hỏi phải có những phương pháp nhanh, phù hợp, chính xác, hiệu quả để lấy được những thông tin bổ ích. Những "*tri thức*" chiết suất từ nguồn cơ sở dữ liệu trên sẽ là nguồn thông tin hỗ trợ cho lãnh đạo trong việc lên kế hoạch hoạt động hoặc trong việc ra quyết định sản xuất kinh doanh. Tiến hành công việc như vậy chính là thực hiện quá trình phát hiện tri thức trong cơ sở dữ liệu (Knowledge Discovery in Database) mà trong đó kỹ thuật khai phá dữ liệu (Data Mining) cho phép phát hiện những tri thức tiềm ẩn. Để lấy được thông tin mang tính tri thức trong khối dữ liệu khổng lồ, cần thiết phải phát triển các kỹ thuật có khả năng tích hợp các dữ liệu từ các hệ thống giao dịch khác nhau, chuyển chúng thành một tập hợp các cơ sở dữ liệu ổn định có chất lượng. Các kỹ thuật như vậy được gọi là kỹ thuật tạo kho dữ liệu và môi trường các dữ liệu nhận được khi áp dụng các kỹ thuật tạo kho dữ liệu nói trên được gọi là kho dữ liệu (Data Warehouse) [19, 24].

Một trong các nội dung cơ bản nhất trong khai phá dữ liệu và rất phổ biến là phát hiện các luật kết hợp. Phương pháp này nhằm tìm ra các tập thuộc tính thường xuất hiện đồng thời trong cơ sở dữ liệu và rút ra các luật về ảnh hưởng của một tập thuộc tính dẫn đến sự xuất hiện của một (hoặc một tập) thuộc tính khác như thế nào. Bên cạnh đó, nhu cầu song song hóa và xử lý phân tán là rất cần thiết hiện nay bởi kích thước lưu trữ dữ liệu ngày càng nhiều nên đòi hỏi tốc độ xử lý cũng như dung lượng bộ nhớ hệ thống phải đảm bảo. Vì thế, yêu cầu cần có những thuật toán song song hiệu quả cho việc phát hiện luật kết hợp.

Ứng dụng khai phá dữ liệu đã mang lại những lợi ích to lớn trong việc tổng hợp và cung cấp những thông tin trong các nguồn cơ sở dữ liệu lớn. Hơn nữa hiện nay nhu cầu song song hóa và xử lý phân tán là rất cần thiết bởi kích

thước dữ liệu lưu trữ ngày càng lớn nên đòi hỏi tốc độ xử lý cũng như dung lượng bộ nhớ hệ thống phải đảm bảo Vì thế, yêu cầu cần có những thuật toán song song hiệu quả cho luật kết hợp.

Phương pháp nghiên cứu của luận văn là tổng hợp các kết quả dự a trên các bài báo khoa học trong một số hội thảo quốc tế và các bài báo chuyên ngành, từ đó trình bày các vấn đề khai phá dữ liệu và xây dựng một số thuật toán khai phá luật kết hợp song song.

Nội dung luận văn được trình bày trong 3 chương và phần kết luận

Chương 1: Tổng quan về khai phá dữ liệu: Giới thiệu tổng quan về quá trình khai phá dữ liệu, kho dữ liệu và khai phá dữ liệu; kiến trúc của một hệ thống khai phá dữ liệu; Nhiệm vụ chính và các phương pháp khai phá dữ liệu.

Chương 2: Khai phá luật kết hợp song song: Chương này trình bày tổng quan về luật kết hợp; phát biểu bài toán khai phá dữ liệu, phát hiện luật kết hợp; các khái niệm cơ bản luật kết hợp và các phương pháp khai phá luật kết hợp; khai phá luật kết hợp với một số khái niệm mở rộng.

Chương 3: Một số phương pháp khai phá luật kết hợp song song và phân tích đánh giá các thuật toán song song .

Thái Nguyên 01 tháng 10 năm 2008

Tác giả

Lê Thị Việt Hoa

Chương 1
TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

1.1. Khái niệm

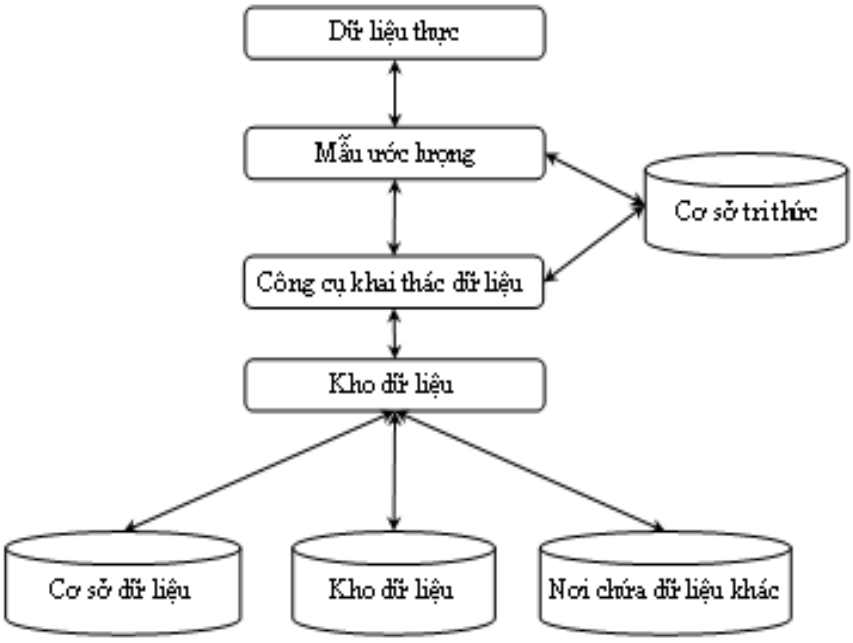
Khai phá dữ liệu là một khái niệm ra đời vào những năm cuối của thập kỷ 80, nó là quá trình tìm kếm, khám phá dưới nhiều góc độ khác nhau nhằm phát hiện các mối liên hệ, quan hệ giữa các dữ liệu, đối tượng bên trong CSDL, kết quả của việc khai phá là xác định các mẫu hay các mô hình tồn tại bên trong nhưng chúng nằm ẩn ở các CSDL [3]. Về bản chất nó là giai đoạn duy nhất rút trích và tìm ra được các mẫu, các mô hình hay thông tin mới, tri thức tiềm ẩn có trong CSDL chủ yếu phục vụ cho mô tả và dự đoán. Đây là giai đoạn quan trọng nhất trong quá trình phát hiện tri thức từ CSDL, các tri thức này hỗ trợ trong việc ra quyết định, điều hành trong khoa học và kinh doanh.

Khai phá dữ liệu là tiến trình khám phá tri thức tiềm ẩn trong các CSDL, cụ thể hơn, đó là tiến trình lọc, sản sinh những tri thức hoặc các mẫu tiềm ẩn, chưa biết những thông tin hữu ích từ các CSDL lớn.

1.2. Kiến trúc của một hệ thống khai phá dữ liệu

Khai phá dữ liệu là quá trình rút trích thông tin bổ ích từ những kho dữ liệu lớn. Khai phá dữ liệu là quá trình chính trong khai phá tri thức từ cơ sở dữ liệu.

Kiến trúc của một hệ thống khai phá dữ liệu có các thành phần như sau: [2]



Hình 1.1. Khám phá tri thức trong cơ sở dữ liệu điển hình