ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐH CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

LÊ VĂN SƠN

NGHIÊN CỨU TẬP MỤC THƯỜNG XUYÊN VÀ LUẬT KẾT HỢP

LUẬN VĂN THẠC SỸ KHOA HỌC MÁY TÍNH

Chuyên ngành : Khoa học máy tính

Mã số : 60 48 01

Thái Nguyên, năm 2011

MỤC LỤC

Trang

MỞ ĐẦU....1
Chương 1: TỔNG QUAN VỀ PHÁ DỮ LIỆU....3
1.1. Khai phá dữ liệu. 3
1.2. Các phương pháp chính trong khai phá dữ liệu 4
1.3. Các cơ sở dữ liệu có thể khai phá 5
1.4. Quá trình khai phá dữ liệu....6
1.5. Một số ứng dụng của khai phá dữ liệu....7
1.6. Khai phá dữ liệu và các lĩnh vực có liên quan 8
1.7. Những khó khăn, thách thức trong khai phá dữ liệu.....9
Chương 2: TẬP MỤC THƯỜNG XUYÊN VÀ LUẬT KẾT HỢP TRONG KHAI PHÁ DỮ LIỆU....12
2.1. Các khái niệm cơ bản 12
2.1.1. Cơ sở dữ liệu giao tác....12
2.1.2. Tập mục và độ hỗ trợ 14
2.1.3. Tập mục thường xuyên (Frequent intemset)....14
2.1.4. Luật kết hợp (Association Rule) 15
2.2. Khai phá tập mục thường xuyên và mở rộng....16
2.2.1. Khai phá tập mục thường xuyên 16
2.2.2. Mở rộng bài toán khai phá tập mục thường xuyên 17
2.3. Khai phá luật kết hợp 18
2.4. Một số tính chất của tập mục thường xuyên và luật kết hợp 20
2.4.1. Một số tính chất của tập mục thường xuyên 20
2.4.2. Một số tính chất của luật kết hợp. 20
2.5. Một số hướng tiếp cận trong khai phá luật kết hợp 21

CHƯƠNG 3: MỘT SỐ PHƯƠNG PHÁP KHAI PHÁ DỮ LIỆU BẰNG LUẬT KẾT HỢP.....23
3.1. Mở đầu 23
3.2. Thuật toán APRIORI khai phá tập mục thường xuyên..... 24
3.3. Khai phá tập mục thường xuyên theo hướng tiếp cận không sinh ứng cử. 30
3.3.1. Thuật toán tạo cây FP-tree [4]..... 31
3.3.2. Duyệt cây FP-tree để sinh các tập mục thường xuyên..... 40
3.4. Khai phá tập mục cổ phần cao 43
3.5. Thuật toán FSM..... 47
3.5.1. Cơ sở lý thuyết của thuật toán FSM..... 47
3.5.2 Thuật toán FSM..... 48
3.6. Thuật toán AFSM..... 52
3.6.1 Cơ sở lý thuyết của thuật toán AFSM..... 52
3.6.2 Thuật toán AFSM..... 55
Chương 4: XÂY DỰNG ỨNG DỤNG KHAI PHÁ TẬP MỤC CỔ PHẦN CAO - THỬ NGHIỆM TRÊN CSDL BÁN HÀNG.....59
4.1. Đặt bài toán 59
4.2. Thiết kế modul chương trình và giải thuật..... 59
4.3. Giao diện sử dụng và chức năng chương trình 60
4.4. Đánh giá kết quả và hướng phát triển của chương trình. 63
KẾT LUẬN.....64
TÀI LIỆU THAM KHẢO.....65

DANH MỤC CÁC KÝ HIỆU VÀ CÁC TỪ VIẾT TẮT

Từ hoặc cụm từ	Từ viết tắt	Tiếng Anh
Khai phá tri thức	KDD	Knowledge Discovery in Database
Khai phá dữ liệu	KPDL	Data Mining
Cơ sở dữ liệu	CSDL	Database

DANH MỤC CÁC BẢNG

Bảng 2.1: Biểu diễn ngang của cơ sở dữ liệu giao tác.....15
Bảng 2.2: Biểu diễn dọc của cơ sử dữ liệu giao tác.....15
Bảng 2.3: Ma trận giao tác của cơ sở dữ liệu bảng 2.1.....16
Bảng 2.4: Các tập mục thường xuyên của CSDL bảng 2.3 với *minsup*=50%....17
Bảng 2.5: Các luật kết hợp sinh ra từ tập mục thường xuyên BCE.....21
Bảng 3.1: Ký hiệu mô tả trong thuật toán Apriori.....26
Bảng 3.2: CSDL minh hoạ thuật toán Apriori.....29
Bảng 3.3: Danh sách các tập mục thường xuyên của CSDL bảng 3.2.....31
Bảng 3.4: CSDL giao tác minh hoạ xây dựng cây FP-tree.....35
Bảng 3.5: Thống kê tần xuất của các mục trong CSDL.....35
Bảng 3.6: CSDL giao tác sau khi loại bỏ các mục không thường xuyên và sắp xếp các mục theo thứ tự giảm dần của tần xuất.....37
Bảng 3.7: Cơ sở dữ liệu ví dụ.....46
Bảng 3.8: Giá trị *lmv* và cổ phần của các mục dữ liệu trong CSDL bảng 3.7.....47
Bảng 3.9: Các tập mục cổ phần cao của CSDL bảng 3.7.....47
Bảng 3.10: Cơ sở dữ liệu minh hoạ thuật toán FSM.....52
Bảng 3.11: Giá trị *lmv* và *CF* với k =1.....52
Bảng 3.12: Giá trị *lmv* và *CF* với k = 2.....52
Bảng 3.13: Giá trị *lmv* và *CF* với k = 3.....53
Bảng 3.14: Giá trị *lmv* và *CF* với k=4.....53
Bảng 3.15: Các giá trị *lmv* và hàm tới hạn với k=2.....58
Bảng 3.16: Các giá trị *lmv* và hàm tới hạn với k=3.....58

DANH MỤC CÁC HÌNH

Hình 1.1: Quá trình khai phá dữ liệu.....9
Hình 1.2: Khai phá dữ liệu và các lĩnh vực có liên quan.....11
Hình 2.1: Phân loại các thuật toán khai phá tập mục thường xuyên.....19
Hình 3.1: Bảng Header của cây FP-tree của CSDL bảng 3.5.....36
Hình 3.2: Cây FP-tree sau khi xét giao tác với TID = T01.....37
Hình 3.3: Cây FP-tree sau khi xét giao tác với TID = T02.....37
Hình 3.4: Cây FP-tree sau khi xét giao tác với TID = T03.....38
Hình 3.5: Cây FP-tree sau khi xét giao tác với TID = T04.....38
Hình 3.6: Cây FP-tree sau khi xét giao tác với TID = T05.....39
Hình 3.7: Cây FP-tree sau khi xét giao tác với TID = T06.....39
Hình 3.8: Cây FP-tree sau khi xét giao tác với TID = T07.....40
Hình 3.9: Cây FP-tree sau khi xét giao tác với TID = T08.....40
Hình 3.10: Cây FP-tree sau khi xét giao tác với TID = T09.....41
Hình 3.11: Cây FP-tree CSDL bảng 3.4.....42
Hình 3.12: Không gian tìm kiếm tập mục cổ phần cao theo thuật toán AFSM..59
Hình 4.1: Cửa sổ giao diện chương trình.....61
Hình 4.2: Cửa sổ thực hiện nhập CSDL.....62
Hình 4.3: Nhập ngưỡng cổ phần *minShare*.....63
Hình 4.4: Cửa sổ thể hiện các bước tìm tập mục cổ phần cao.....64
Hình 4.5: Cửa sổ hiển thị kết quả tìm tập mục cổ phần cao.....64

MỞ ĐẦU

Sự phát triển nhanh chóng của các ứng dụng Công nghệ thông tin và Internet vào nhiều lĩnh vực đời sống xã hội như quản lý kinh tế, quản lý nhân sự, khoa học kỹ thuật…đã mở ra nhiều cơ hội cho các tổ chức, doanh nghiệp trong việc thu thập và xử lý thông tin. Hơn nữa các công nghệ lưu trữ, phục hồi dữ liệu phát triển một cách nhanh chóng, làm xuất hiện nhiều cơ sở dữ liệu khổng lồ. Để khai thác có hiệu quả nguồn thông tin từ các cơ sở dữ liệu khổng lồ trên, yêu cầu cấp thiết đặt ra là cần phải có những kỹ thuật, công cụ để chuyển đổi kho dữ liệu khổng lồ thành những tri thức có ích. Từ đó các kỹ thuật Khai phá dữ liệu trở thành một lĩnh vực được đặc biệt quan tâm trong ngành Công nghệ thông tin.

Khai phá dữ liệu là một khái niệm được ra đời vào những năm cuối của thập kỷ 1980, là quá trình khám phá thông tin ẩn được tìm thấy trong các cơ sở dữ liệu và được ứng dụng một cách rộng rãi trong nhiều lĩnh vực khác nhau như: marketing, tài chính, ngân hàng và bảo hiểm, khoa học, y tế, an ninh, internet…Rất nhiều tổ chức và công ty lớn trên thế giới đã áp dụng kỹ thuật khai phá dữ liệu vào các hoạt động sản xuất kinh doanh của mình và đã thu được những lợi ích to lớn.

Một trong các nội dung cơ bản và phổ biến nhất trong khai phá dữ liệu là tìm các tập mục thường xuyên từ đó 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. Đồng thời mở rộng khai phá tập mục thường xuyên là khai phá tập mục cổ phần cao để có thể đánh giá được sự đóng góp của tập mục trong tổng số các mục dữ liệu của cơ sở dữ liệu.

Từ những lý do trên tôi chọn đề tài ***"Nghiên cứu tập mục thường xuyên và luật kết hợp"***. Luận văn được xây dựng dựa trên một số nghiên cứu trong lĩnh vực khai phá tập mục thường xuyên và luật kết hợp trong những năm gần đây.

Luận văn được tổ chức thành 04 chương

Chương 1. Tổng quan về khai phá dữ liệu

Chương 2. Tập mục thường xuyên và luật kết hợp trong khai phá dữ liệu

Chương 3: Một số phương pháp khai phá dữ liệu bằng luật kết hợp

Chương 4. Xây dựng ứng dụng khai phá tập mục cổ phần cao-ứng dụng thử nghiệm trên CSDL bán hàng

Kết luận

Chương 1: TỔNG QUAN VỀ PHÁ DỮ LIỆU

1.1. Khai phá dữ liệu.

Trong kỷ nguyên bùng nổ công nghệ thông tin, các công nghệ lưu trữ dữ liệu ngày càng phát triển tạo điều kiện cho việc thu thập, lưu trữ dữ liệu tốt hơn. Đặc biệt trong lĩnh vực kinh doanh, các doanh nghiệp đã nhận thức được tầm quan trọng của việc nắm bắt và xử lý thông tin, nhằm giúp các chủ doanh nghiệp trong việc vạch ra các chiến lược kinh doanh, kịp thời mang lại lợi nhuận to lớn. Từ đó khiến các tổ chức, doanh nghiệp tạo ra một lượng dữ liệu khổng lồ cho riêng mình. Các kho dữ liệu ngày càng lớn trong đó tiềm ẩn nhiều thông tin có ích. Để khai thác có hiệu quả nguồn thông tin từ các kho dữ liệu khổng lồ dẫn tới một yêu cầu cấp thiết là phải có những kỹ thuật và công cụ mới để biến kho dữ liệu khổng lồ thành những thông tin cô đọng và có ích. Kỹ thuật Khai phá dữ liệu (Data mining) ra đời như một kết quả tất yếu đáp ứng yêu cầu đó.

Khai phá dữ liệu (Data mining) là quá trình trích xuất các thông tin có giá trị tiềm ẩn bên trong lượng lớn dữ liệu được lưu trữ trong các cơ sở dữ liệu, kho dữ liệu. Hiện nay, ngoài thuật ngữ khai phá dữ liệu người ta còn dùng một số thuật ngữ khác có ý nghĩa tương tự như: Khai phá tri thức từ CSDL (Knowledge mining from databases), trích lọc dữ liệu (Knowledge extraction), phân tích dữ liệu/mẫu (data/pattern analysis), khảo cổ dữ liệu (data archaeology), nạo vét dữ liệu (data dredging). Nhiều người coi khai phá dữ liệu và một thuật ngữ thông dụng khác là khám phá tri thức trong CSDL (Knowledge Discovery in Databases –KDD) là như nhau. Thực tế khai phá dữ liệu chỉ là một bước thiết yếu trong quá trình khám phá tri thức trong CSDL.

Theo Giáo sư Tom Mitchell [11] định nghĩa KPDL: "KPDL là việc sử dụng dữ liệu lịch sử để khám phá những qui tắc và cải thiện những quyết định trong tương lai"

Tóm lại: Khai phá dữ liệu là một quá trình tìm kiếm, phát hiện ra các tri thức mới và các tri thức có ích ở dạng tiềm ẩn trong các cơ sở dữ liệu lớn.

1.2. Các phương pháp chính trong khai phá dữ liệu

Các phương pháp chính trong KPDL có thể được phân chia theo chức năng hay lớp các bài toán khác nhau [7] như sau:

- Phân lớp và dự đoán (classicfication and prediction): Xếp một đối tượng vào một trong những lớp đã biết trước. Ví dụ: Phân lớp các bệnh nhân theo dữ liệu trong hồ sơ bệnh án. Hướng tiếp cận này thường sử dụng các kỹ thuật như học máy (Machine learning), cây quyết định (Decision tree), mạng nơ ron nhân tạo (Neural network)...Với phương pháp này còn được gọi là học có giám sát.

- Luật kết hợp (Association rules): Là dạng luật biểu diễn tri thức ở dạng khá đơn giản. Mục tiêu của phương pháp này là phát hiện và đưa ra các mối liên hệ giữa các giá trị dữ liệu trong cơ sở dữ liệu.

Luật kết hợp được ứng dụng nhiều trong lĩnh vực kinh doanh, y học, tin – sinh học, tài chính, thị trường chứng khoán...

- Khai phá chuỗi theo thời gian (Sequential temporal patterns): Tương tự như khai phá dữ liệu bằng luật kết hợp nhưng có thêm tính thứ tự và tính thời gian. Hướng tiếp cận này được ứng dụng nhiều trong lĩnh vực tài chính và thị trường chứng khoán bởi vì chúng có tính dự báo cao.

- Phân cụm và phân đoạn (Clusterring and Segmentation): Sắp xếp các đối tượng theo từng cụm dữ liệu tự nhiên (số lượng và tên của cụm chưa được biết trước). Các đối tượng được gom cụm sao cho mức độ tương tự giữa các đối tượng trong cùng một cụm là lớn nhất và mức độ tương tự giữa các