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

NGUYỄN XUÂN TRƯỜNG

NGHIÊN CỨU CÁC PHẦN TỬ NGOẠI LAI TRONG CSDL & ỨNG DỤNG

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYÊN – 2014

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

NGUYỄN XUÂN TRƯỜNG

NGHIÊN CỨU CÁC PHẦN TỬ NGOẠI LAI TRONG CSDL & ỨNG DỤNG

Chuyên ngành: 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 – 2014

LỜI CAM ĐOAN

Luận văn thạc sỹ này tôi nghiên cứu và thực hiện dưới sự hướng dẫn của Thầy giáo GS.TS Vũ Đức Thi . Để hoàn thành bản luận văn này, ngoài các tài liệu đã liệt kê, tôi cam đoan không sao chép các công trình hoặc thiết kế tốt nghiệp của người khác.

Thái Nguyên, ngày 18 tháng 04 năm 2014

Học viên

Nguyễn Xuân Trường

LỜI CẢM ƠN

Trước hết, tôi vô cùng biết ơn sâu sắc đến Thầy giáo GS.TS Vũ Đức Thi, người thầy đã trực tiếp dành nhiều thời gian tận tình hướng dẫn, cung cấp những thông tin, tài liệu quý báu giúp đỡ tôi hoàn thành bản luận văn này.

Sau cùng tôi xin bày tỏ lòng biết ơn đến người thân, cùng bạn bè, đồng nghiệp cơ quan, những người luôn cổ vũ động viên tôi hoàn thành bản luận văn tốt nghiệp này.

Thái Nguyên, ngày 18 tháng 04 năm 2014

Học viên

Nguyễn Xuân Trường

MỤC LỤC

DANH MỤC THUẬT NGỮ

Từ viết tắt	Nghĩa của từ
Box_Cox	Tên phép biến đổi thành dạng xấp xỉ chuẩn
DB (Distance Based)	Dựa theo khoảng cách
DSE (Donoho Stahel)	Tên bộ ước lượng mạnh đa biến
KDD (Know ledgement Discovery in Database)	Khám phá tri thức trong cơ sở dữ liệu
LOF (Local Outlier Factor)	Yếu tố ngoại lai cục bộ
MAD (Median Absolute Deviation)	Là tên một bộ ước lượng mạnh đơn biến
NL (Nested Loop)	Tên một thuật toán phát hiện phần tử ngoại lai
Shorth (Shortest half)	Là tên một bộ ước lượng mạnh đơn biến

DANH MỤC HÌNH VẼ

MỞ ĐẦU

Thế kỷ XXI được xem là một kỷ nguyên của nền kinh tế tri thức. Các công nghệ khám phá tri thức được áp dụng rộng rãi trong nhiều lĩnh vực và đã đem lại những thành tựu to lớn. Nhưng các công nghệ khám phá tri thức thường nhằm mục đích tìm kiếm, khám phá, các dạng mẫu thường gặp. Chủ yếu tập trung vào các hướng: Tìm kiếm các luật kết hợp, nhận dạng và phân lớp mẫu…Còn lĩnh vực khám phá phần tử ngoại lai mới bước đầu được sự quan tâm nghiên cứu.

Mặc dù nó được ứng dụng trong nhiều lĩnh vực trong cuộc sống: như phát hiện những thẻ bất thường trong hệ thống ngân hàng, những tuyến đường bất ổn không hợp lý trong giao thong, ứng dụng trong hệ thống an ninh, dự báo thời tiết, trong thị trường chứng khoán, trong lĩnh vực thể thao ... Tuy nhiên, với số lượng dữ liệu được tập trung và lưu trữ trong cơ sở dữ liệu ngày càng lớn thì việc tìm kiếm các ngoại lệ hoặc các phần tử ngoại lai trở nên cấp thiết hơn nhiều.

CHƯƠNG I: KHÁM PHÁ TRI THỨC TRONG CƠ SỞ DỮ LIỆU VÀ PHẦN TỬ NGOẠI LAI

Nội dung của chương này giới thiệu quá trình khám phá tri thức, khai thác dữ liệu và các ứng dụng thực tế có sự hỗ trợ của các kỹ thuật khai thác dữ liệu. Đồng thời trình bày khái niệm về phần tử ngoại lai và mối quan hệ giữa các lĩnh vực khám phá phần tử ngoại lai và lĩnh vực khai thác dữ liệu**.**

1.1 Khám phá tri thức.

Với sự tiến bộ của khoa học kỹ thuật và nhu cầu con người ngày càng tăng đã tạo nên một thời đại bùng nổ thông tin trong mọi lĩnh vực của đời sống. Với lượng thông tin " khổng lồ" đó thì cần có các kỹ thuật khai thác dữ liệu hiệu quả để lấy ra những thông tin hữu ích. Một số các ngôn ngữ chuy vấn được sử dụng nhằm lấy ra những thông tin yêu cầu của người sử dụng, nhưng hầu hết các ngôn ngữ này chỉ lấy ra được dữ liệu theo những yêu cầu đơn giản. Các kiểu dữ liệu đa phương tiện được một số các hệ thống cơ sở dữ liệu hỗ trợ như: Dữ liệu âm thanh, hình ảnh…không thể đáp ứng được khi các yêu cầu của người sử dụng ngày càng cao và phức tạp. Do đó, với nhu cầu tìm kiếm tri thức trong cơ sở dữ liệu đã hình thành một lĩnh vực mới: Khám phá tri thức trong cơ sở dữ liệu. Khám phá tri thức là toàn bộ quá trình tìm kiếm tri thức dữ liệu, bao gồm các bước sau:

- *Chuẩn bị dữ liệu* : Dữ liệu được tập chung vào trong các cơ sở dữ liệu, các kho dữ liệu. Dữ liệu có thể là chưa sạch tức là có cả dữ liệu sai sót, không phù hợp, nhiễu, và các dữ liệu không đủ thông tin. Do đó, trong bước này dữ liệu được làm sạch để loại bỏ các dữ liệu không phù hợp, dữ liệu không liên quan. Công việc này có thể được tiến hành trước hoặc sau khi phát hiện dữ liệu không sạch. Đồng thời, sau khi được làm sạch, dữ liệu được làm