ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP

KHỔNG MINH TỰ

NGHIÊN CỨU, TÌM HIỂU MỘT SỐ THUẬT TOÁN CƠ BẢN VỀ PHÂN NHÓM DỮ LIỆU TRÊN CƠ SỞ DỮ LIỆU KHÔNG GIAN

LUẬN VĂN THẠC SĨ KỸ THUẬT ĐIỆN TỬ

THÁI NGUYÊN - 2014

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP

.....

KHỔNG MINH TỰ

NGHIÊN CỨU, TÌM HIỂU MỘT SỐ THUẬT TOÁN CƠ BẢN VỀ PHÂN NHÓM DỮ LIỆU TRÊN CƠ SỞ DỮ LIỆU KHÔNG GIAN

Chuyên ngành: KỸ THUẬT ĐIỆN TỬ
Mã số: 60. 52. 02. 03

LUẬN VĂN THẠC SĨ KỸ THUẬT

PHÒNG QUẢN LÝ ĐÀO TẠO SAU ĐẠI HỌC

NGƯỜI HƯỚNG DẪN KHOA HỌC

PGS.TS. LƯƠNG CHI MAI

KHOA ĐIỆN TỬ
TRƯỞNG KHOA

THÁI NGUYÊN - 2014

LỜI CAM ĐOAN

Tôi xin cam đoan đây là công trình nghiên cứu của riêng tôi, các số liệu, kết quả nêu trong luận văn này là trung thực và là công trình nghiên cứu của riêng tôi, luận văn này không giống hoàn toàn bất cứ luận văn hoặc các công trình đã có trước đó.

Thái Nguyên, ngày 24 tháng 02 năm 2014

Tác giả luận văn

Khổng Minh Tự

LỜI CẢM ƠN

Trong suốt quá trình học tập và tốt nghiệp, tôi đã nhận được sự giúp đỡ tận tình của các thầy cô trong Khoa Điện tử - Trường Đại học Kỹ thuật Công nghiệp - Đại học Thái Nguyên. Tôi xin bày tỏ lòng biết ơn đối với các thầy cô giáo và Phòng Đào tạo sau đại học vì sự giúp đỡ tận tình này. Tôi đặc biệt muốn cảm ơn **PGS.TS. Lương Chi Mai** đã tận tình giúp đỡ, hướng dẫn tôi trong thời gian thực hiện đề tài, cảm ơn sự giúp đỡ của gia đình, bạn bè và các đồng nghiệp trong thời gian qua.

Mặc dù đã cố gắng, song do điều kiện thời gian và kinh nghiệm thực tế còn nhiều hạn chế nên không thể tránh khỏi thiếu sót. Vì vậy, tôi rất mong nhận được sự đóng góp ý kiến của các thầy cô cũng như của các bạn bè, đồng nghiệp.

Tôi xin chân thành cảm ơn!

Tác giả luận văn

Khổng Minh Tự

LỜI NÓI ĐẦU

Trong thời đại bùng nổ Công nghệ thông tin, các công nghệ lưu trữ dữ liệu ngày càng phát triển nhanh chóng tạo điều kiện cho các đơn vị thu thập dữ liệu nhiều hơn và tốt hơn. Đặc biệt trong lĩnh vực quản lý, kinh doanh, các doanh nghiệp đã nhận thức được tầm quan trọng của việc nắm bắt và xử lí thông tin. Tất cả lí do đó khiến cho các cơ quan, đơn vị và các doanh nghiệp đã tạo ra một lượng dữ liệu khổng lồ cỡ Gigabyte thậm chí là Terabyte cho riêng mình. Các kho dữ liệu ngày càng lớn và tiềm ẩn nhiều thông tin có ích. Sự bùng nổ đó dẫn tới một yêu cầu cấp thiết là phải có những kĩ thuật và công cụ mới để biến kho dữ liệu khổng lồ kia thành những thông tin (tri thức) cô đọng và có ích.

Tuy nhiên ngay cả khi đã có những công cụ phù hợp để lưu trữ và quản lý các dạng thông tin nói trên, thì để nhận được những thông tin có ích đối với dạng CSDL loại này, các biện pháp phân tích dữ liệu thông thường cũng gặp rất nhiều khó khăn, đôi khi là không thể giải quyết được. Đó chính là cơ sở cho sự xuất hiện của kỹ thuật khai phá dữ liệu.

Tác giả xin bày tỏ lòng biết ơn chân thành đến các thầy cô giáo, đặc biệt là cô giáo hướng dẫn: PGS.TS. Lương Chi Mai đã tận tình giúp đỡ để hoàn thành luận văn này.

Trong khuôn khổ giới hạn của luận văn cùng khả năng kiến thức và thời gian nghiên cứu còn hạn chế, nên mặc dù đã có nhiều cố gắng song luận văn chắc chắn không tránh khỏi những thiếu sót. Tác giả mong nhận được sự đóng góp ý kiến của các thầy giáo, cô giáo để đề tài được hoàn thiện hơn.

Xin trân trọng cảm ơn!

HỌC VIÊN

Khổng Minh Tự

MỤC LỤC

LỜI CAM ĐOAN i
LỜI CẢM ƠN ii
LỜI NÓI ĐẦU iii
MỤC LỤC iv
BẢNG THUẬT NGỮ VIẾT TẮT vii
DANH MỤC CÁC HÌNH viii
MỞ ĐẦU 1
Chương 1. TỔNG QUAN VỀ KHAI PHÁ TRI THỨC VÀ CƠ SỞ DỮ LIỆU KHÔNG GIAN 6
1.1. Khai phá tri thức trong cơ sở dữ liệu (Knowledge Discovery in Databases - DD) 6
1.1.1. Sự ra đời của khai phá tri thức trong cơ sở dữ liệu 6
1.1.2. Khái niệm khai phá dữ liệu 7
1.1.3. Quá trình khai phá tri thức trong cơ sở dữ liệu 7
1.1.4. Các nhiệm vụ của khai phá dữ liệu 8
1.2. Phân nhóm (Clustering) và các cách tiếp cận chính 9
1.2.1. Phân nhóm và các ứng dụng 9
1.2.2. Các cách tiếp cận chính 11
1.3. Hệ quản trị cơ sở dữ liệu không gian 16
1.3.1. Cơ sở dữ liệu không gian 16
1.3.2. Hệ quản trị cơ sở dữ liệu không gian 17
1.3.3. Phương pháp truy nhập không gian 18
1.4. Kết luận 20
Chương 2. CÁC CÁCH TIẾP CẬN CỦA KỸ THUẬT PHÂN NHÓM 21
2.1. Thuật toán DBSCAN 21
2.1.1. Các định nghĩa và bổ đề được sử dụng trong thuật toán DBSCAN 22
2.1.2. Thuật toán DBSCAN 25

2.2. Thuật toán DBCLASD 27
2.2.1. Một số định nghĩa 27
2.2.2. Thuật toán DBCLASD 30
2.3. Thuật toán DENCLUE 34
2.3.1. Một số định nghĩa 35
2.3.2. Những tính chất của phương pháp DENCLUE 37
2.3.3. Thuật toán DENCLUE 38
2.4. Kết luận 43
Chương 3. CÁC GIẢI THUẬT PHÂN NHÓM TRÊN CƠ SỞ DỮ LIỆU KHÔNG GIAN LỚN 44
3.1. Một số khái niệm cần thiết khi tiếp cận phân nhóm dữ liệu 44
3.1.1. Phân loại các kiểu dữ liệu 44
3.1.2. Độ đo tương tự và phi tương tự 45
3.2. Thuật toán K-MEANS 49
3.3. Giải thuật DBSCAN 53
3.4. Kết luận 55
Chương 4. XÁC ĐỊNH THAM SỐ, CÀI ĐẶT THỬ NGHIỆM VÀ ĐÁNH GIÁ KẾT QUẢ 56
4.1. Môi trường thử nghiệm 56
4.2. Công cụ thử nghiệm 56
4.3. Xác định tham số 56
4.3.1. Xác định tham số cho thuật toán DBSCAN 56
4.3.2. Tối ưu hoá việc lựa chọn các tham số σ và ξ cho thuật toán DENCLUE 62
4.4. Cài đặt thử nghiệm và đánh giá kết quả 63
4.4.1. Xây dựng chương trình cài đặt thuật toán phân nhóm 63
4.4.2. Tạo lập dữ liệu 64
4.4.3. Cài đặt thuật toán phân nhóm 65
4.4.4. Lưu trữ và hiển thị kết quả 73
4.5. Đánh giá kết quả trên một số tập dữ liệu 74

4.5.1. Tập dữ liệu.....74
4.5.2. Đánh giá kết quả.....75
4.5.3. Nhận xét.....79
4.6. Kết luận.....81
KẾT LUẬN.....82
TÀI LIỆU THAM KHẢO.....84

BẢNG THUẬT NGỮ VIẾT TẮT

Từ hoặc nhóm từ	Từ viết tắt	Từ tiếng anh
Cơ sở dữ liệu	CSDL	DataBase
Khai phá dữ liệu	KPDL	Data Mining
Khai phá tri thức	KPTT	Knowledge Discovery
Khai phá tri thức trong cơ sở dữ liệu	KDD	Knowledge Discovery in Databases
Phân nhóm dữ liệu	PNDL	Data Clustering

DANH MỤC CÁC HÌNH

Hình 1.1: Các bước trong quá trình khám phá tri thức KDD....8
Hình 1.2: Biểu đồ Hertzsprung-Russell....10
Hình 1.3: Mô tả cách phân nhóm theo phương pháp từ dưới lên và từ trên xuống....14
Hình 1.4: Những điểm nằm trong miền tô sẫm mới được xét đến khi tìm điểm gần nhất cho điểm x. Những điểm ngoài miền không cần xét đến....17
Hình 1.5: Một cách chia lưới. Những ô mầu sẫm là những ô chứa dữ liệu và được lưu trữ. Những ô màu trắng là những ô không chứa dữ liệu....19
Hình 1.6: Mô phỏng một R*-tree gồm 3 mức....20
Hình 2.1: Lân cận của P với ngưỡng Eps....22
Hình 2.2: Mật độ - đến được trực tiếp....23
Hình 2.3: Mật độ đến được....23
Hình 2.4: Mật độ liên thông....24
Hình 2.5: Nhóm và nhiễu....24
Hình 2.6: Mô phỏng thuật toán DBSCAN....25
Hình 2.7: Thủ tục ExpandCluster....26
Hình 2.8: Ví dụ dữ liệu tập các điểm được chia thành 2 lớp....27
Hình 2.9: Ảnh hưởng của độ rộng ô lưới đến việc xác định vùng xấp xỉ....29
Hình 2.11: Ví dụ một cách chia và đánh số trong không gian hai chiều....40
Hình 3.1: Minh họa số đo chiều rộng, chiều cao một đối tượng....46
Hình 3.2: Khoảng cách Euclidean....48
Hình 3.3: Các thiết lập để xác định ranh giới các nhóm ban đầu....49
Hình 3.4: Tính toán trọng tâm của các nhóm mới....50
Hình 3.5: Ví dụ các bước của thuật toán K-means....52
Hình 3.6: Một số hình dạng khám phá bởi phân nhóm dưa trên mật độ....54
Hình 3.7: Thuật toán DBSCAN....54
Hình 4.1: Môi trường thử nghiệm....56