

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

CHU THỊ HẢO

**KỸ THUẬT PHÂN CỤM DỮ LIỆU
KHÔNG GIAN CÓ RÀNG BUỘC**

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYÊN, 2017

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

CHU THỊ HẢO

**KỸ THUẬT PHÂN CỤM DỮ LIỆU
KHÔNG GIAN CÓ RÀNG BUỘC**

Chuyên ngành: Khoa học máy tính
Mã số: 60 48 01 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Người hướng dẫn khoa học: PGS.TS. ĐẶNG VĂN ĐỨC

THÁI NGUYÊN, 2017

MỤC LỤC

MỞ ĐẦU	1
Chương 1. TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU VÀ DỮ LIỆU KHÔNG GIAN	4
1.1. Khai phá dữ liệu	4
1.1.1. Một số khái niệm.....	4
1.1.2. Quá trình khai phá dữ liệu.....	4
1.1.3. Các kỹ thuật khai phá dữ liệu.....	7
1.2. Dữ liệu không gian địa lý.....	9
1.3. Hệ thống thông tin địa lý và ứng dụng.....	10
1.3.1. Một số định nghĩa về hệ thống tin địa lý	11
1.3.2. Mô hình biểu diễn dữ liệu địa lý không gian	14
1.3.3. Quan hệ không gian giữa các đối tượng địa lý	20
1.4. Khái niệm và mục tiêu của Phân cụm dữ liệu	20
1.5. Kết luận	23
Chương 2. MỘT SỐ KỸ THUẬT PHÂN CỤM DỮ LIỆU KHÔNG GIAN...	24
2.1. Phương pháp phân cụm theo phân hoạch	24
2.2. Phương pháp phân cụm dựa trên mật độ	26
2.3. Phương pháp phân cụm dựa trên lưới.....	32
2.4. Phương pháp phân cụm dữ liệu ràng buộc.....	35
2.4.1. Thuật toán phân cụm dữ liệu không gian.....	37
2.4.2. Thuật toán.....	45
2.5. Kết luận	48
Chương 3. CÀI ĐẶT VÀ THỬ NGHIỆM	49
3.1. Phân tích bài toán	49
3.1.1. Nguồn dữ liệu đầu vào và phạm vi bài toán	49
3.1.2. Phương pháp kỹ thuật giải quyết bài toán.....	50

3.2. Xây dựng chương trình ứng dụng	51
3.2.1. Phân tích thiết kế hệ thống	51
3.2.2. Cài đặt chương trình.....	52
3.3. Thử nghiệm và đánh giá các thuật toán phân cụm.....	54
KẾT LUẬN VÀ KIẾN NGHỊ	61
TÀI LIỆU THAM KHẢO	62

DANH MỤC CÁC BẢNG

- Bảng 3.1: So sánh tổng quan các thuật toán K-means, DBSCAN và DBRS.....54
- Bảng 3.2: Kết quả so sánh thời gian thực hiện phân cụm của các thuật toán K-means, DBSCAN và DBRS với cùng một tập dữ liệu đầu vào.....56
- Bảng 3.3: Kết quả so sánh thời gian thực hiện phân cụm của các thuật toán K-means, DBSCAN và DBRS trên các tập dữ liệu khác nhau57

DANH MỤC CÁC HÌNH

Hình 1.1:	Khai phá dữ liệu trong tập dữ liệu	4
Hình 1.2:	Tiến trình khám phá tri thức từ cơ sở dữ liệu	5
Hình 1.3:	Kiến trúc điển hình của một hệ khai phá dữ liệu	6
Hình 1.4:	Ví dụ biểu diễn vị trí trước bị ô nhiễm	13
Hình 1.5:	Ví dụ biểu diễn đường xác định bởi ranh giới các đường, có điểm đầu trùng với điểm cuối.....	13
Hình 1.6:	Ví dụ biểu diễn khu vực hành chính	14
Hình 1.7:	Biểu diễn vector của đối tượng địa lý.....	18
Hình 1.8:	Biểu diễn thế giới bằng mô hình raster.....	19
Hình 1.9:	Mô tả tập dữ liệu được phân thành 3 cụm	21
Hình 2.1:	Minh họa thuật toán k-means	25
Hình 2.2:	Kề mật độ.....	27
Hình 2.3:	Kết nối theo mật độ.....	27
Hình 2.4:	Hình dạng các cụm được khám phá bởi thuật toán DBSCAN.....	28
Hình 2.5:	Cấu trúc phân cấp.....	32
Hình 2.3:	Các cách mà các cụm có thể đưa ra	36
Hình 2.6:	Phân cụm các đối tượng dữ liệu ràng buộc.	37
Hình 2.7:	Phân cụm các đối tượng dữ liệu ràng buộc.....	40
Hình 2.8:	Các đa giác đơn giản và tạo ra các đường cản trở	44
Hình 2.9:	Thuật toán 1: phân cụm có các ràng buộc.....	47
Hình 2.10:	Thuật toán 2: Mở rộng một cụm.....	47
Hình 2.11:	Tìm các điểm lúng giềng	47
Hình 3.1:	Phân cụm lớp dữ liệu "Khách sạn-Trường học trong nội thành Hà Nội, các vùng màu vàng là các cụm tìm được.....	53

Hình 3.2:	Hình ảnh chồng phủ (vùng màu vàng) của các cụm “Siêu thị” (màu xanh) và các cụm “Khách sạn- Trường học” (màu đỏ). Vùng màu vàng có thể coi là vị trí tối ưu cho việc đặt địa điểm Nhà hàng.	53
Hình 3.3:	Kết quả phân cụm DBSCAN đối với dữ liệu thử nghiệm tự tạo.....	54
Hình 3.4:	Khả năng phát hiện nhiễu và cụm có hình dạng bất kỳ của K-means (trái) và DBSCAN (phải), đường bao màu xanh là đường biên cụm	55
Hình 3.5:	Khả năng phân cụm theo thuộc tính của DBSCAN (trái) và DBRS (phải)	55
Hình 3.5:	Đồ thị so thời gian thực hiện phân cụm của các thuật toán K-measn, DBSCAN và DBRS với cùng một tập dữ liệu đầu vào.	57
Hình 3.6:	Phân cụm tập dữ liệu DS1	59
Hình 3.7:	Phân cụm DS2.....	60

MỞ ĐẦU

Hệ thống thông tin địa lý (GIS) được ứng dụng ngày càng phổ biến, không chỉ trong lĩnh vực giám sát, quản lý, lập kế hoạch về tài nguyên môi trường mà còn trong nhiều bài toán kinh tế xã hội khác. Kết quả là, khối lượng dữ liệu liên quan đến địa lý, còn gọi là dữ liệu không gian thu thập được tăng lên nhanh chóng. Một câu hỏi đặt ra là làm thế nào để tận dụng, khai thác, khám phá, phát hiện những tri thức hữu ích từ kho dữ liệu này?

Khai phá dữ liệu là áp dụng các kỹ thuật và công cụ để trích rút các tri thức có ích từ nguồn dữ liệu về một lĩnh vực nào đó mà ta quan tâm. Khai phá dữ liệu với GIS hay còn gọi là khai phá dữ liệu không gian, mở rộng khai phá dữ liệu trong các CSDL quan hệ, xét thêm các thuộc tính của dữ liệu không gian được phản ánh trong hệ thống tin địa lý, ví dụ khoảng cách (gần kề hay cách xa), điều kiện môi trường tự nhiên hay kinh tế xã hội (rừng núi, đồng bằng, ven biển, đô thị, v.v...).

Các bài toán truyền thống của một hệ thống tin địa lý có thể trả lời các câu hỏi kiểu như:

- Những con phố nào dẫn đến sân bay Tân Sơn Nhất ?
- Những căn nhà nào nằm trong vùng quy hoạch mở rộng phố?

Khai phá dữ liệu không gian có thể giúp trả lời cho các câu hỏi dạng:

- Xu hướng của các dòng chảy, các đứt gãy địa tầng ?
- Nên bố trí các trạm tiếp sóng điện thoại di động như thế nào?
- Những vị trí nào là tối ưu để đặt các máy ATM, xăng dầu, nhà hàng,...?

Một trong những bài toán liên quan đến dữ liệu không gian, cụ thể là dữ liệu địa lý có ý nghĩa thực tế cao là bài toán xác định vị trí tối ưu cho việc đặt các cây xăng. Cả nước hiện có 374 tổng đại lý và hơn 14.000 cửa hàng bán lẻ xăng dầu. Để xác định được vị trí đặt các trạm bán lẻ xăng dầu cần

phải tuân theo các quy định của Bộ Công thương, nhất là các quy định về an toàn, phòng chống cháy nổ. Ngoài ra, cây xăng cũng phải đặt ở vị trí thuận lợi cho việc kinh doanh đạt doanh số cao. Hoặc một bài toán khác cũng có ý nghĩa thực tiễn rất lớn đó là xác định vị trí tối ưu để mở một nhà hàng. Hiện nay trên địa bàn thành phố Hà Nội cũng đã có rất nhiều nhà hàng, quán ăn đã được mở ra. Nhưng không phải tất cả các nhà hàng, quán ăn đó đều có thể cho doanh thu tốt. Có khi có nhà hàng mới mở ra được một thời gian ngắn đã phải đóng cửa vì không có khách dẫn đến chủ đầu tư phải chịu thua lỗ nặng. Một trong những nguyên nhân chính dẫn đến thất bại đó là địa điểm kinh doanh chưa hợp lý. Một vị trí tối ưu cho việc mở nhà hàng, quán ăn thì vị trí đó phải thỏa mãn một số yếu tố sau: nằm trong khu vực đông dân cư, gần nhiều cơ quan công sở hay trường học, có khu vực để xe, có quang cảnh xung quanh thoáng mát...các vấn đề này đã được rất nhiều các đề tài nghiên cứu tuy nhiên với những vị trí phức tạp có các ngăn cách con sông hay cây cầu v.v... thì cần phải có những đánh giá chính xác hơn nữa.

Xuất phát từ nhu cầu thực tế đó và do đặc thù, khả năng ứng dụng rất phong phú của kỹ thuật phân cụm dữ liệu trong không gian nên em đã chọn nghiên cứu đề tài kỹ thuật phân cụm dữ liệu không gian có ràng buộc làm luận văn tốt nghiệp cao học.

Trên cơ sở đó cài đặt thử nghiệm một ứng dụng sử dụng kỹ thuật phân cụm dữ liệu không gian, trong đó khai thác thông tin địa lý của các đối tượng để hỗ trợ giải quyết bài toán ví dụ như tìm vị trí tối ưu đặt nhà hàng.

Luận văn được chia thành các chương mục sau

- Chương 1: Tổng quan về khai phá dữ liệu và dữ liệu không gian
- Chương 2: Một số kỹ thuật phân cụm dữ liệu không gian
- Chương 3: Xây dựng chương trình thử nghiệm, kết luận, đánh giá

Luận văn này được hoàn thành dưới sự hướng dẫn tận tình của PGS.TS Đặng Văn Đức, em xin bày tỏ lòng biết ơn chân thành của mình đối với thầy. Em xin chân thành cảm ơn các thầy, cô giáo Viện Công nghệ thông tin,

Trường Đại học Công nghệ thông tin và Truyền thông - Đại học Thái Nguyên đã tham gia giảng dạy, giúp đỡ em trong suốt quá trình học tập nâng cao trình độ kiến thức. Tuy nhiên vì điều kiện thời gian và khả năng có hạn nên luận văn không thể tránh khỏi những thiếu sót. Em kính mong các thầy cô giáo và các bạn đóng góp ý kiến để đề tài được hoàn thiện hơn