BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG

ISO 9001:2008

TRẦN THỊ HẰNG NGA

LUẬN VĂN THẠC SĨ
NGÀNH HỆ THỐNG THÔNG TIN

HẢI PHÒNG, 2016

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG

TRẦN THỊ HẰNG NGA

XÂY DỰNG HỆ THỐNG HỖ TRỢ LỰA CHỌN ĐỊA ĐIỂM ĐẶT MÁY ATM TẠI THÀNH PHỐ HẢI PHÒNG BẰNG KỸ THUẬT PHÂN CỤM KHÔNG GIAN

LUẬN VĂN THẠC SĨ
NGÀNH CÔNG NGHỆ THÔNG TIN

CHUYÊN NGÀNH: HỆ THỐNG THÔNG TIN
MÃ SỐ: 60 48 01 04

NGƯỜI HƯỚNG DẪN KHOA HỌC:
PGS.TS. ĐẶNG VĂN ĐỨC

MỤC LỤC

MỘT SỐ THUẬT NGỮ VIẾT TẮT

CSDL	Cơ sở dữ liệu
GIS	Hệ thông tin địa lý
KDD	Khám phá tri thức từ cơ sở dữ liệu
KPDL	Khai phá dữ liệu
OLAP	Xử lý phân tích dữ liệu trực tuyến

DANH MỤC HÌNH VẼ

DANH MỤC BẢNG

LỜI CẢM ƠN

Lời đầu tiên, em xin được gửi lời cảm ơn chân thành và sâu sắc tới PGS.TS Đặng Văn Đức, người thầy đã cho em những định hướng và ý kiến quý báu trong suốt quá trình hoàn thành luận văn.

Em xin chân thành cảm ơn các thầy, cô trong trường Đại học Dân lập Hải Phòng và Viện Công nghệ Thông tin - Viện Hàn lâm Khoa học Việt Nam đã giảng dạy, truyền đạt cho em những kiến thức quý báu trong thời gian qua.

Tôi xin được gửi lời cảm ơn sâu sắc tới gia đình, bạn bè và đồng nghiệp những người luôn kịp thời động viên, khích lệ giúp đỡ tôi vượt qua những khó khăn để tôi có thể hoàn thành nhiệm vụ của mình.

Do còn hạn chế về nhiều mặt nên luận văn không thể tránh khỏi những hạn chế, thiếu sót. Rất mong nhận được sự chỉ dẫn, góp ý của Thầy, cô và các bạn./.

Xin trân trọng cảm ơn!

Hải Phòng, tháng 11 năm 2016

Học viên

Phú Thị Quyên

LỜI CAM ĐOAN

Tôi xin cam đoan toàn bộ nội dung bản luận văn ***"Xây dựng hệ thống tìm kiếm âm thanh theo nội dung dựa trên các đặc trưng miền tần số"*** là do tôi tự sưu tầm, tra cứu và tìm hiểu theo tài liệu tham khảo và làm theo hướng dẫn của người hướng dẫn khoa học.

Nội dung bản luận văn chưa từng được công bố hay xuất bản dưới bất kỳ hình thức nào và cũng không được sao chép từ bất kỳ một công trình nghiên cứu nào. Các nguồn lấy từ tài liệu tham khảo đều được chú thích rõ ràng, đúng quy định.

Xin trân trọng cảm ơn!

Hải Phòng, tháng 11 năm 2016

Học viên

Phú Thị Quyên

MỞ ĐẦU

Thông tin địa lý bao gồm dữ liệu về bề mặt Trái đất và các diễn giải dữ liệu để con người dễ hiểu. Thông tin địa lý gồm hai loại dữ liệu: không gian (spatial data) và phi không gian (non-spatial data).

Hệ thống thông tin Địa lý (Geograpgic Information System) đã bắt đầu được sử dụng rộng rãi ở các nước phát triển từ nhiều thập niên qua, đây là một dạng ứng dụng công nghệ tin học (Information Technology) nhằm mô tả thế giới thực (Real world) mà loài người đang sống-tìm hiểu-khai thác. Với những tính năng ưu việt, kỹ thuật GIS ngày nay đang được ứng dụng trong nhiều lãnh vực nghiên cứu và quản lý, đặc biệt trong quản lý và quy hoạch sử dụng-khai thác các nguồn tài nguyên một cách bền vững và hợp lý.

Sự phát triển không ngừng của công nghệ thông tin đã đưa tin học thâm nhập sâu vào nhiều lĩnh vực khoa học và đời sống, mở ra một giai đoạn mới trong quá trình phát triển khoa học. Hệ thống thông tin địa lý là một trong những ứng dụng rất có giá trị của công nghệ tin học trong ngành địa lý, điều tra cơ bản, quy hoạch đô thị và cảnh báo môi trường.

Khai phá dữ liệu không gian hay còn gọi là khai phá tri thức từ dữ liệu không gian là một lĩnh vực được áp dụng rộng rãi. Từ dữ liệu đầu vào bao gồm một khối lượng dữ liệu không gian khổng lồ được thu thập từ nhiều ứng dụng khác nhau, chẳng hạn từ thiết bị viễn thám đến hệ thống thông tin địa lý, từ bản đồ số, từ các hệ thống quản lý và đánh giá môi trường, …Việc phân tích và khai thác lượng thông tin khổng lồ này ngày càng thách thức và khó khăn, đòi hỏi phải có các nghiên cứu sâu hơn để tìm ra các kỹ thuật khai phá dữ liệu hiệu quả hơn.

Khai phá dữ liệu không gian được sử dụng nhiều trong các hệ thống thông tin địa lý (GIS), viễn thám, khai phá dữ liệu ảnh chẳng hạn ảnh y học, rô bốt dẫn đường, … Khám phá tri thức từ dữ liệu không gian có thể được thực hiện dưới nhiều hình thức khác nhau như sử dụng các quy tắc đặc trưng và quyết định, trích rút và mô tả các cấu trúc hoặc cụm nổi bật, kết hợp không gian, …

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

- Những con phố nào dẫn đến siêu thị Big C Hải Phòng ?

- Những căn nhà nào nằm trong vùng quy hoạch mở rộng tại thành phố Hải Phòng?

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, siêu thị…?

Một trong những bài toán có ý nghĩa thực tế cao là bài toán xác định vị trí tối ưu cho việc đặt các máy ATM của các ngân hàng. Trong những năm gần đây, cùng với sự phát triển của xã hội, việc sử dụng thẻ ATM tại Việt Nam rất phổ biến. Thẻ ATM thực chất như một loại ví điện tử cho phép người sử dụng chỉ cần mang theo một chiếc thẻ gọn nhẹ, thay vì rất nhiều tiền mặt. Thẻ ATM không những cho phép người dùng rút tiền khi cần tiền mặt, còn cho phép thực hiện nhiều giao dịch khác tại máy ATM hoặc điện thoại, chẳng hạn chuyển khoản, thanh toán tàu xe ... Thẻ ATM còn có thể dùng để thanh toán tại các nhà hàng, siêu thị, trung tâm mua sắm, các điểm bán hàng có đặt ATM. Ngoài việc tiện lợi trong sử dụng ra, chủ thẻ còn được hưởng lãi suất từ tài khoản tiền gửi.

Xuất phát từ nhu cầu thực tế đó, luận văn giới thiệu tổng quan về GIS và phân cụm dữ liệu, giới thiệu một số thuật toán phân cụm dữ liệu không gian và thuật toán xếp chồng bản đồ được sử dụng hiện nay. Trên cơ sở đó cài đặt thử nghiệm một ứng dụng sử dụng kỹ thuật phân cụm dữ liệu địa lý và xếp chồng bản đồ, trong đó khai thác thông tin địa lý của các đối tượng địa lý có tầm ảnh hưởng quan trọng đến vị trí đặt các máy ATM như: các siêu thị, trung tâm mua sắm, nhà hàng, khách sạn, bệnh viện, trường học, ... để hỗ trợ giải quyết bài toán hỗ trợ tìm vị trí tối ưu đặt các máy ATM trong khu vực nội thành thành phố Hải Phòng.

Luận văn được chia thành các chương mục sau:

- Mở đầu
- Chương 1: Tổng quan về Hệ thông tin Địa lý (GIS) và phân cụm dữ liệu.
- Chương 2: Một số thuật toán liên quan
- Chương 3: Xây dựng chương trình thử nghiệm
- Kết luận

CHƯƠNG 1. TỔNG QUAN VỀ HỆ THỐNG THÔNG TIN ĐỊA LÝ (GIS) VÀ PHÂN CỤM DỮ LIỆU

1.1 Một số vấn đề cơ bản của Hệ thông tin địa lý (GIS)

Địa lý (geography) được hình thành từ hai khái niệm: trái đất (geo-earth) và tiến trình mô tả (graphy). Như vậy, địa lý được xem như tiến trình mô tả trái đất. Là lĩnh vực khoa học nghiên cứu về các vùng đất, địa hình, dân cư và các hiện tượng trên Trái Đất .

Khi mô tả Trái đất, các nhà địa lý luôn đề cập đến quan hệ không gian (*spatial relationship*) của các đối tượng trong thế giới thực. Mối quan hệ này được thể hiện thông qua các bản đồ (*map*) trong đó biểu diễn đồ họa của tập các đặc trưng trừu tượng và quan hệ không gian tương ứng trên bề mặt trái đất, ví dụ: bản đồ dân số biểu diễn dân số tại từng vùng địa lý.

Dữ liệu bản đồ còn là loại dữ liệu có thể được số hóa. Để lưu trữ và phân tích các số liệu thu thập được, cần có sự trợ giúp của hệ thông tin địa lý (*Geographic Information System-GIS*).

1.1.1 Một số định nghĩa về hệ thông tin địa lý

Có nhiều định nghĩa khác nhau về GIS, Các cách định nghĩa này đều mô tả việc nghiên cứu các thông tin địa lý và các khía cạnh khác liên quan.

GIS cũng giống như các hệ thống thông tin khác, có khả năng nhập, tìm kiếm và quản lý các dữ liệu lưu trữ, để từ đó đưa ra các thông tin cần thiết cho người sử dụng. Ngoài ra, GIS còn cho phép lập bản đồ với sự trợ giúp của máy tính, giúp cho việc biểu diễn dữ liệu bản đồ tốt hơn so với cách truyền thống. Dưới đây là một số định nghĩa GIS hay dùng [1]:

Định nghĩa của dự án The Geographer's Craft, Khoa Địa lý, Trường Đại học Texas

GIS là cơ sở dữ liệu số chuyên dụng trong đó hệ trục tọa độ không gian là phương tiện tham chiếu chính. GIS bao gồm các công cụ để thực hiện những công việc sau:

- Nhập dữ liệu từ bản đồ giấy, ảnh vệ tinh, ảnh máy bay, số liệu điều tra và các nguồn khác.

- Lưu trữ dữ liệu, khai thác, truy vấn cơ sở dữ liệu.

- Biến đổi dữ liệu, phân tích, mô hình hóa, bao gồm cả dữ liệu thống kê và dữ liệu không gian.

- Lập báo cáo, bao gồm bản đồ chuyên đề, bảng biểu, biểu đồ và kế hoạch.

Từ định nghĩa trên, ta thấy: *Thứ nhất*, GIS có quan hệ với ứng dụng cơ sở dữ liệu. Thông tin trong GIS đều liên kết với tham chiếu không gian và GIS sử dụng tham chiếu không gian như phương tiện chính để lưu trữ và truy nhập thông tin. *Thứ hai*, GIS là công nghệ tích hợp, cung cấp các khả năng phân tích như phân tích ảnh máy bay, ảnh vệ tinh hay tạo lập mô hình thống kê, vẽ bản đồ... Cuối cùng, GIS có thể được xem như một hệ thống cho phép trợ giúp quyết định. Cách thức nhập, lưu trữ, phân tích dữ liệu trong GIS phải phản ánh đúng cách thức thông tin sẽ được sử dụng trong công việc lập quyết định hay nghiên cứu cụ thể.

Định nghĩa của David Cowen, NCGIA, Mỹ

GIS là hệ thống phần cứng, phần mềm và các thủ tục được thiết kế để thu thập, quản lý, xử lý, phân tích, mô hình hóa và hiển thị các dữ liệu qui chiếu không gian để giải quyết các vấn đề quản lý và lập kế hoạch phức tạp.

Một cách đơn giản, có thể hiểu GIS như một sự kết hợp giữa bản đồ (*map*) và cơ sở dữ liệu (*database*).

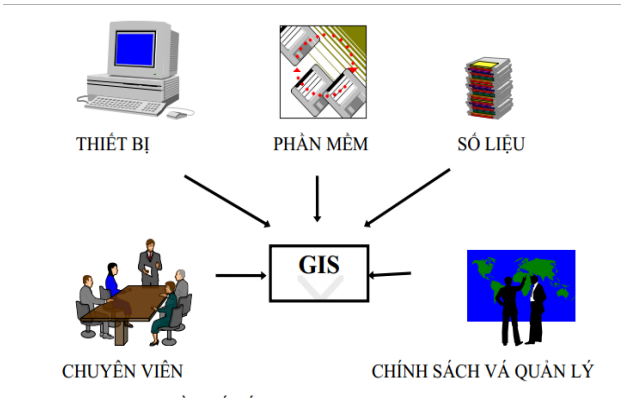
GIS = Bản đồ + Cơ sở dữ liệu

Bản đồ trong GIS là một công cụ hữu ích cho phép chỉ ra vị trí của từng địa điểm. Với sự kết hợp giữa bản đồ và cơ sở dữ liệu, người dùng có thể xem thông tin chi tiết về từng đối tượng/thành phần tương ứng với địa điểm trên bản đồ thông qua các dữ liệu đã được lưu trữ trong cơ sở dữ liệu. Ví dụ, khi xem bản đồ về các thành phố, người dùng có thể chọn một thành phố để xem thông tin về thành phố đó như diện tích, số dân, thu nhập bình quân, số quận/huyện của thành phố, …

1.1.2 Các thành phần của hệ thông tin địa lý.

Một hệ thông tin địa lý bao gồm 5 thành phần:

* Thiết bị (hardware)
* Phần mềm (software)
* Số liệu (Geographic data)
* Chuyên gia (Expertise)
* Chính sách và cách thức quản lý (Policy and management)



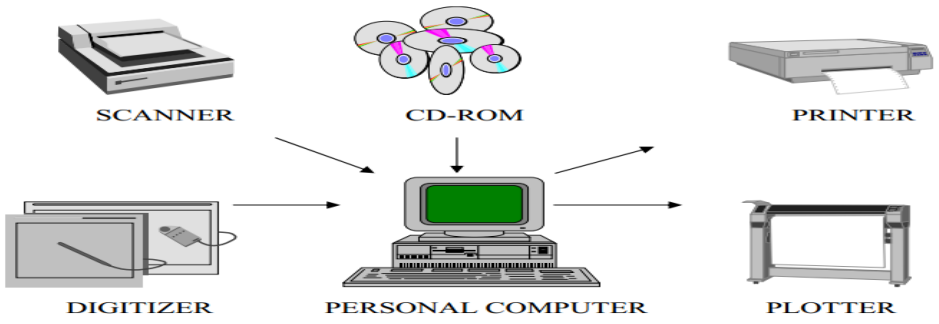
Hình 1.1: Các thành tố của Gis

· **Thiết bị phần cứng:**

Phần cứng là hệ thống máy tính trên đó một ứng dụng GIS hoạt động. Ngày nay, phần mềm GIS có khả năng chạy trên rất nhiều dạng phần cứng, từ máy chủ trung tâm đến các máy trạm hoạt động độc lập hoặc liên kết mạng.

Thiết bị bao gồm máy vi tính (computer), máy vẽ (plotters), máy in (printer), bàn số hoá (digitizer), thiết bị quét ảnh (scanners), các phương tiện lưu trữ số liệu (Floppy diskettes, optical cartridges, C.D ROM v.v...).

.



Hình 1.2: Các thành phần thiết bị cơ bản của GIS

Phần mềm

Là tập hợp các câu lệnh, chỉ thị nhằm điều khiển phần cứng của máy tính thực hiện một nhiệm vụ xác định, phần mềm hệ thống thông tin địa lý có thể là một hoặc tổ hợp các phần mềm máy tính. Phần mềm được sử dụng trong kỹ thuật GIS phải bao gồm các tính năng cơ bản sau:

- Nhập và kiểm tra dữ liệu (Data input)
- Lưu trữ và quản lý cơ sở dữ liệu (Geographic database).
- Xuất dữ liệu (Display and reporting)
- Biến đổi dữ liệu (Data transformation)
- Tương tác với người dùng (Query input)

· **Dữ liệu**

Có thể coi thành phần quan trọng nhất trong một ứng dụng GIS là dữ liệu. Một hệ thống thông tin không thể thiếu dữ liệu, dữ liệu là nguồn đầu vào, là nguyên liệu để hệ thống thực hiện phân tích, xử lý và cho ra kết quả phục vụ nhu cầu khai thác thông tin của người dùng. Một cách tổng quát, người ta chia dữ liệu địa lý gồm 2 thành phần (component):

* Thành phần dữ liệu không gian (spatial) cho ta biết kích thước vật lý, hình dạng và vị trí địa lý của các đối tượng trên bề mặt trái đất.

* Thành phần dữ liệu thuộc tính (non-spatial) là các dữ liệu ở dạng văn bản cho ta biết thêm thông tin thuộc tính của đối tượng.

Các dữ liệu không gian (hình học) và dữ liệu thuộc tính liên quan có thể được người sử dụng tự tập hợp hoặc được mua từ nhà cung cấp dữ liệu thương mại. Hệ GIS sẽ kết hợp dữ liệu không gian với các nguồn dữ liệu khác, thậm chí có thể sử dụng hệ quản trị cơ sở dữ liệu để tổ chức lưu giữ và quản lý dữ liệu.

Nhân lực

Đây là một trong những hợp phần quan trọng của công nghệ GIS, đòi hỏi những chuyên viên hướng dẫn sử dụng hệ thống để thực hiện các chức năng phân tích và xử lý các số liệu. Đòi hỏi phải thông thạo về việc lựa chọn các công cụ GIS để sử dụng, có kiến thức về các số liệu đang được sử dụng và thông hiểu các tiến trình đang và sẽ thực hiện.

Nhân lực tham gia vào hệ thông tin địa lý với một hoặc nhiều vai trò sau:

* Người dùng GIS là những người sử dụng các phần mềm GIS để giải quyết các bài toán không gian theo mục đích của họ. Họ thường là những người được đào tạo tốt về lĩnh vực GIS hay là các chuyên gia.

* Người xây dựng bản đồ: sử dụng các lớp bản đồ được lấy từ nhiều nguồn khác nhau, chỉnh sửa dữ liệu để tạo ra các bản đồ theo yêu cầu.

* Người phân tích: giải quyết các vấn đề như tìm kiếm, xác định vị trí…

* Người thiết kế CSDL: xây dựng các mô hình dữ liệu lôgic và vật lý.

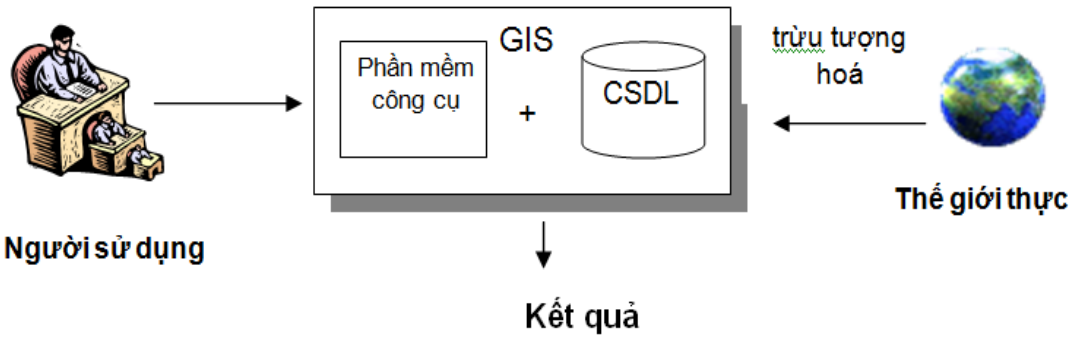
* Người phát triển: xây dựng hoặc cải tạo các phần mềm GIS để đáp ứng các nhu cầu cụ thể.

· **Chính sách và quản lý**

Phần này rất quan trọng trong việc đảm bảo khả năng hoạt động có hiệu quả của hệ thống, là yếu tố quyết định sự thành công của việc phát triển công nghệ GIS. Hệ thống GIS cần được điều hành bởi một bộ phận quản lý, bộ phận này phải được đào tạo chuyên nghiệp để tổ chức hoạt động hệ thống GIS một cách có hiệu quả và phục vụ người sử dụng thông tin.

Như vậy, trong 5 hợp phần của GIS, hợp phần chính sách và quản lý đóng vai trò rất quan trọng để đảm bảo khả năng hoạt động của hệ thống, đây là yếu tố quyết định sự thành công của việc phát triển hệ thông tin địa lý.

Các thành phần này kết hợp với nhau nhằm tự động quản lý và phân phối thông tin thông qua biểu diễn địa lý.



Hình 1.3 : Mối quan hệ giữa các thành phần của GIS

1.1.3 Biểu diễn dữ liệu địa lý

Các thành phần của dữ liệu địa lý

Một cơ sở dữ liệu của hệ thống thông tin địa lý có thể chia ra làm 2 loại dữ liệu cơ bản: dữ liệu không gian và phi không gian. Mỗi loại có những đặc điểm

riêng và chúng khác nhau về yêu cầu lưu giữ số liệu, hiệu quả, xử lý và hiển thị.

Thành phần dữ liệu không gian

Thành phần dữ liệu không gian hay thường gọi là dữ liệu hình học hay dữ liệu bản đồ, là dữ liệu về đối tượng mà vị trí của nó được xác định trên bề mặt trái đất. Dữ liệu không gian sử dụng trong hệ thống địa lý luôn được xây dựng trên một hệ thống tọa độ, bao gồm tọa độ, quy luật và các ký hiệu dùng để xác định một hình ảnh bản đồ cụ thể trên mỗi bản đồ.

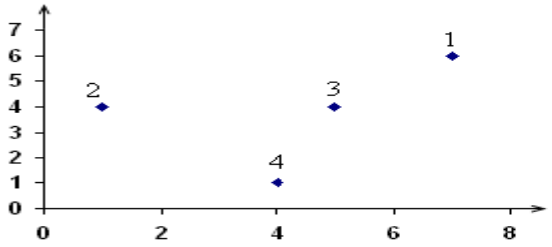
Hệ thống GIS dùng thành phần dữ liệu không gian để tạo ra bản đồ hay hình ảnh bản đồ trên màn hình hoặc trên giấy thông qua thiết bị ngoại vi. Mỗi hệ thống GIS có thể dùng các mô hình khác nhau để mô hình hóa thế giới thực sao cho giảm thiểu sự phức tạp của không gian nhưng không mất đi các dữ liệu cần thiết để mô tả chính xác các đối tượng trong không gian. Hệ thống GIS hai chiều 2D dùng **ba kiểu dữ liệu cơ sở** sau để mô tả hay thể hiện các đối tượng trên bản đồ vector (sẽ làm rõ hơn ở phần sau), đó là:

- **Điểm (*Point*)**

Điểm được xác định bởi cặp giá trị tọa độ (x, y). Các đối tượng đơn với thông tin về địa lý chỉ bao gồm vị trí thường được mô tả bằng đối tượng điểm.

Các đối tượng biểu diễn bằng kiểu điểm thường mang đặc tính chỉ có tọa độ đơn (x, y) và không cần thể hiện chiều dài và diện tích. Ví dụ, trên bản đồ, các vị trí của bệnh viện, các trạm rút tiền tự động ATM, các cây xăng, … có thể được biểu diễn bởi các điểm.

Hình 1.4 là ví dụ về vị trí nước bị ô nhiễm. Mỗi vị trí được biểu diễn bởi 1 điểm gồm cặp tọa độ (x, y) và tương ứng với mỗi vị trí đó có thuộc tính độ sâu và tổng số nước bị nhiễm bẩn. Các vị trí này được biểu diễn trên bản đồ và lưu trữ trong các bảng dữ liệu.



DiemID	X	Y
1	7	6
2	1	4
3	5	4
4	4	1

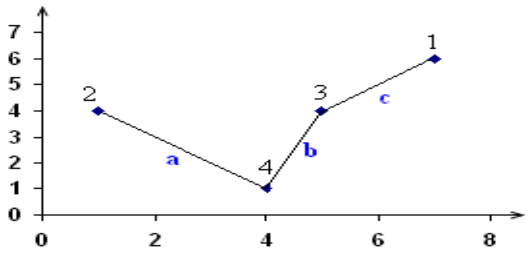
DiemID	Độ sâu(m)	Lượng nước(m^3)
1	12.5	63.5
2	26	80
3	3.5	24.34
4	17	69.23

Hình 1.4: Ví dụ biểu diễn vị trí nước bị ô nhiễm

- **Đường – Cung (*Line - Arc*)**

Đường được xác định bởi dãy các điểm hoặc bởi 2 điểm đầu và điểm cuối. Đường dùng để mô tả các đối tượng địa lý dạng tuyến như đường giao thông, sông ngòi, tuyến cấp điện, cấp nước…

Các đối tượng được biểu diễn bằng kiểu đường thường mang đặc điểm là có dãy các cặp tọa độ, các đường bắt đầu và kết thúc hoặc cắt nhau bởi điểm, độ dài đường bằng chính khoảng cách của các điểm. Ví dụ, bản đồ hệ thống đường bộ, sông, đường biên giới hành chính, … thường được biểu diễn bởi đường và trên đường có các điểm (*vertex*) để xác định vị trí và hình dáng của đường đó.



DiemID	X	Y
1	7	6
2	1	4
3	5	4
4	4	1

DuongID	Đầu	Cuối
a	12.5	63.5
b	26	80
c	3.5	24.34

DuongID	Lưu lượng	Dung lượng
a	960	2220
b	1250	2000
c	1100	1950

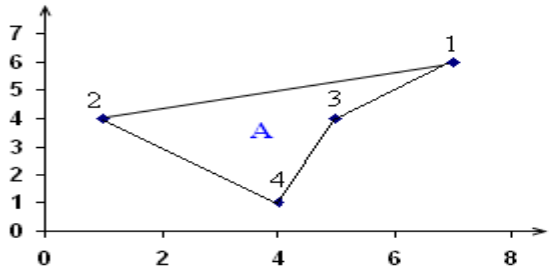
Hình 1.5: Ví dụ biểu diễn đường

- **Vùng (*Polygon*)**

Vùng được xác định bởi ranh giới các đường, có điểm đầu trùng với điểm cuối. Các đối tượng địa lý có diện tích và được bao quanh bởi đường thường được biểu diễn bởi vùng.

Các đối tượng biểu diễn bởi vùng có đặc điểm là được mô tả bằng tập các đường bao quanh vùng và điểm nhãn (*label point*) thuộc vùng để mô tả, xác định

cho mỗi vùng. Ví dụ, các khu vực hành chính, hình dạng các công viên,… được mô tả bởi kiểu dữ liệu vùng. Hình 1.6 mô tả ví dụ cách lưu trữ một đối tượng vùng.



DiemID	X	Y
1	7	6
2	1	4
3	5	4
4	4	1

VungID	Thứ tự điểm
A	1, 2, 4, 3, 1

VungID	Diện tích	Số dân
A	3 Km^2	2.5nghìn

Hình 1.6: Ví dụ biểu diễn khu vực hành chính

Một đối tượng có thể biểu diễn bởi các kiểu khác nhau tùy thuộc vào tỷ lệ của bản đồ đó. Ví dụ, đối tượng công viên có thể được biểu diễn bởi điểm trong bản đồ có tỷ lệ nhỏ, và bởi vùng trong bản đồ có tỷ lệ lớn.

Thành phần phi không gian

Thành phần dữ liệu phi không gian hay còn gọi là dữ liệu thuộc tính, là những diễn tả đặc tính, số lượng, mối quan hệ của các hình ảnh bản đồ với vị trí địa lý của chúng thông qua một cơ chế thống nhất. Hệ thống GIS có cơ chế liên kết dữ liệu không gian và phi không gian của cùng một đối tượng với nhau. Có thể nói, một trong những chức năng đặc biệt của công nghệ GIS chính là khả năng liên kết và xử lý đồng thời dữ liệu bản đồ và dữ liệu thuộc tính. Dữ liệu thuộc tính trong hệ thống GIS bất kỳ thường phân thành 4 loại sau:

Bộ xác định: có thể là một số duy nhất, liên tục, ngẫu nhiên hoặc chỉ báo địa lý, số liệu xác định vị trí lưu trữ chung. Bộ xác định cho một thực thể chứa tọa độ phân bố của nó, số hiệu mảnh bản đồ, mô tả khu vực hay con trỏ đến vị trí lưu trữ của số liệu liên quan. Bộ xác định thường lưu trữ với các bản ghi tọa độ hay mô tả khác của hình ảnh không gian và các bản ghi số liệu thuộc tính liên quan.

Số liệu hiện tượng, tham khảo địa lý: miêu tả thông tin danh mục, các hoạt động liên quan đến các vị trí địa lý xác định (ví dụ như: cho phép xây dựng, báo cáo tai nạn, nghiên cứu y tế,…) Thông tin này được lưu trữ và quản lý trong các tệp/ bảng độc lập, trong đó mỗi bản ghi chứa yếu tố xác định vị trí của sự kiện hay hiện

tượng quản lý.

Chỉ số địa lý: bao gồm tên, địa chỉ, khối, phương hướng định vị, … liên quan đến các đối tượng địa lý. Một chỉ số có thể bao gồm nhiều bộ xác định cho thực thể địa lý. Ví dụ: chỉ số địa lý về đường phố và địa chỉ địa lý liên quan
đến phố đó.

Quan hệ giữa các đối tượng tại một vị trí địa lý cụ thể trong không gian. Đây là thông tin quan trọng cho các chức năng xử lý của hệ thống thông tin địa lý. Các mối quan hệ không gian có thể là mối quan hệ đơn giản hay lôgic, ví dụ tiếp theo số nhà 37 phải là số nhà 38.

1.1.4 Mô hình biểu diễn dữ liệu không gian.

Dữ liệu của GIS có được thông qua việc mô hình hóa các thực thể địa lý. Mô hình biểu diễn dữ liệu địa lý là cách thức chúng ta biểu diễn trừu tượng các thực thể địa lý. Mô hình biểu diễn dữ liệu địa lý đóng vai trò quan trọng vì cách thức biểu diễn thông tin sẽ ảnh hưởng tới khả năng thực hiện phân tích dữ liệu và khả năng hiển thị đồ họa của một hệ thống thông tin địa lý.

Các mức trừu tượng của dữ liệu được thể hiện qua 3 mức mô hình, bao gồm [1]:

- Mô hình khái niệm
- Mô hình logic
- Mô hình vật lý

Mô hình khái niệm

Đây là mức trừu tượng đầu tiên trong tiến trình biểu diễn các thực thể địa lý. Là tập ***các thành phần và các quan hệ*** giữa chúng liên quan đến hiện tượng tự nhiên nào đó. Mô hình này độc lập lập với hệ thống, độc lập với cấu trúc, tổ chức và quản lý dữ liệu. Một số mô hình quan niệm thường được sử dụng trong GIS là:

- *Mô hình không gian trên cơ sở đối tượng*:

Mô hình này tập trung vào các hiện tượng, thực thể riêng rẽ được xem xét độc lập hay cùng với quan hệ của chúng với thực thể khác. Bất kỳ thực thể lớn hay nhỏ đều được xem như một đối tượng và có thể độc lập với các thực thể láng giềng. Đối

tượng này lại có thể bao gồm các đối tượng khác và chúng cũng có thể có quan hệ với các đối tượng khác. Ví dụ các đối tượng kiểu thửa đất và hồ sơ là tách biệt với các đối tượng khác về không gian và thuộc tính.

Mô hình hướng đối tượng phù hợp với các thực thể do con người tạo ra như nhà cửa, đường quốc lộ, các điểm tiện ích hay các vùng hành chính. Một số thực thể tự nhiên như sông hồ, đảo… cũng thường được biểu diễn bằng mô hình đối tượng do chúng cần được xử lý như các đối tượng rời rạc. Mô hình dữ liệu kiểu vector (sẽ đề cập đến ở phần sau) là một ví dụ của mô hình không gian trên cơ sở đối tượng.

- *Mô hình không gian trên cơ sở mạng*:

Mô hình này có một vài khía cạnh tương đồng với mô hình hướng đối tượng, nhưng mở rộng xem xét cả mối quan hệ tương tác giữa các đối tượng không gian. Mô hình này thường quan tâm đến tính liên thông, hay đường đi giữa các đối tượng không gian, ví dụ mô hình mạng lưới giao thông, mạng lưới cấp điện, cấp thoát nước…Trong mô hình này, hình dạng chính xác của đối tượng thường không được quan tâm nhiều. Mô hình topo là một ví dụ về mô hình không gian trên cơ sở mạng.

- *Mô hình quan sát trên cơ sở nền*:

Mô hình này quan tâm đến tính liên tục, trải dài về mặt không gian của thực thể địa lý, ví dụ các thực thể như thảm thực vật, vùng mây bao phủ, vùng ô nhiễm khí quyển, nhiệt độ bề mặt đại dương…thích hợp khi sử dụng mô hình này. Mô hình dữ liệu kiểu raster (sẽ đề cập ở phần sau) là một ví dụ về mô hình quan sát trên cơ sở nền.

Mô hình logic

Sau khi biểu diễn các thực thể ở mức mô hình quan niệm, bước tiếp theo là cụ thể hóa mô hình quan niệm của các thực thể địa lý thành các cách thức tổ chức hay còn gọi là ***cấu trúc dữ liệu*** cụ thể để có thể được xử lý bởi hệ thông tin địa lý. Ở mô hình logic, các thành phần biểu diễn thực thể và quan hệ giữa chúng được chỉ rõ dưới dạng các cấu trúc dữ liệu. Một số cấu trúc dữ liệu được sử dụng trong GIS là:

- *Cấu trúc dữ liệu toàn đa giác*:

Mỗi tầng trong cơ sở dữ liệu của cấu trúc này được chia thành tập các đa giác. Mỗi đa giác được mã hóa thành trật tự các vị trí hình thành đường biên của vùng khép kín theo hệ trục tọa độ nào đó. Mỗi đa giác được lưu trữ như một đặc trưng độc lập, do vậy không thể biết được đối tượng kề của một đối tượng địa lý. Như vậy quan hệ topo (thể hiện mối quan hệ không gian giữa các đối tượng địa lý như quan hệ kề nhau, bao hàm nhau, giao cắt nhau…) không thể hiện được trong cấu trúc dữ liệu này. Nhược điểm của cấu trúc dữ liệu này là một số đường biên chung giữa hai đa giác kề nhau sẽ được lưu hai lần, và như vậy, việc cập nhật, sửa đổi dữ liệu thường gặp nhiều khó khăn.

- *Cấu trúc dữ liệu cung nút*:

Cấu trúc dữ liệu cung nút mô tả các thực thể địa lý dưới dạng các điểm (nút) và các đường (cung). Như vậy, có thể biểu diễn được quan hệ topo giữa các đối tượng địa lý. Trong cấu trúc dữ liệu này, các phần đối tượng không gian kề nhau sẽ được lưu trữ một lần, ngoài ra, các đối tượng lân cận của một đối tượng địa lý cũng được chỉ rõ, điều này giúp dễ dàng thực hiện các phép phân tích không gian, đồng thời cũng tối ưu được dung lượng lưu trữ dữ liệu.

- *Cấu trúc dữ liệu dạng cây*:

Trong một số mô hình dữ liệu như mô hình raster, dữ liệu có thể được phân hoạch thành các đối tượng nhỏ hơn với nhiều mức khác nhau để giảm thiểu dung lượng lưu trữ và tăng tốc độ truy vấn. Ví dụ cấu trúc cây tứ phân chia một vùng dữ liệu làm 4 phần, trong mỗi phần này lại có thể được chia tiếp thành 4 phần con.

Mô hình dữ liệu vật lý

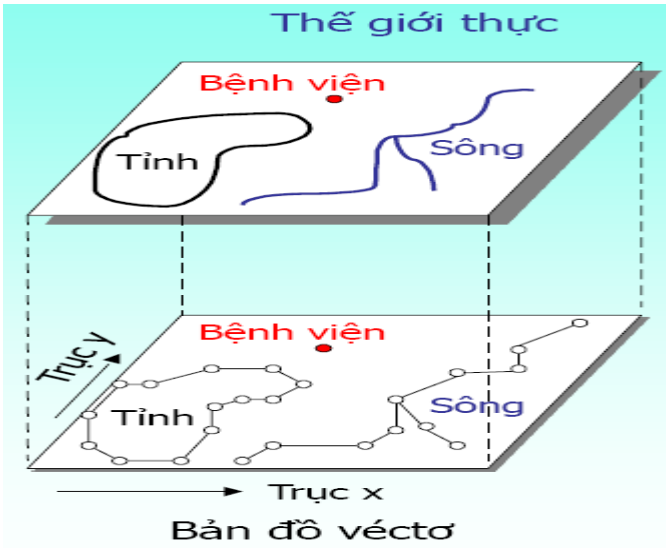
Dữ liệu địa lý cần được lưu trữ vật lý trên máy tính theo một cách thức nhất định, tùy theo các hệ thống thông tin địa lý cụ thể mà ***cách thức lưu trữ, cài đặt dữ liệu*** khác nhau. Mô hình dữ liệu vật lý thường khá khác nhau đối với từng hệ thống GIS cụ thể. Một số hệ GIS thương mại có thể kể đến như: Arc/Info, ERDAS, Geovision, Grass, Caris, Intergres, Oracle, Postgres…

Vậy, từ một thực thể địa lý, thông qua 3 mức mô hình biểu diễn mà được cụ thể hóa thành dữ liệu trên máy tính sẽ có dạng thể hiện rất khác nhau đối với từng hệ GIS cụ thể. Mỗi hệ thông tin địa lý đều sử dụng mô hình dữ liệu quan niệm riêng để biểu diễn mô hình dữ liệu vật lý duy nhất. Hệ thông tin địa lý cung cấp các phương pháp để người sử dụng làm theo các mô hình quan niệm tương tự ba lớp mô hình mô tả trên.

Hai nhóm mô hình dữ liệu không gian thường gặp trong các hệ GIS thương mại là mô hình dữ liệu vector và mô hình dữ liệu raster.

Mô hình vector

Mô hình vector sử dụng tọa độ 2 chiều (x, y) để lưu trữ hình khối của các thực thể không gian trên bản đồ 2D. Mô hình này sử dụng các đặc tính rời rạc như điểm, đường, vùng để mô tả không gian, đồng thời cấu trúc topo của các đối tượng cũng cần được mô tả chính xác và lưu trữ trong hệ thống.

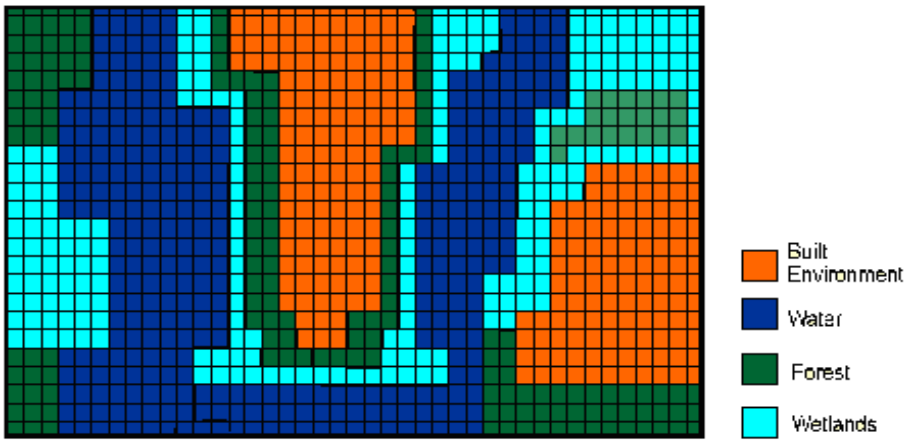


Hình 1.7: Biểu diễn vector của đối tượng địa lý

Theo Hình 1.7 các đối tượng không gian được lưu trữ dưới dạng vertor, đồng thời các thuộc tính liên quan đến lĩnh vực cần quản lý (dữ liệu chuyên đề - *thematic data*) của đối tượng đó cũng cần kết hợp với dữ liệu trên. Các nhân tố chỉ ra sự tác động qua lại lẫn nhau giữa các đối tượng cũng được quản lý, các nhân tố đó có thể là quan hệ topo (giao/ không giao nhau, phủ, tiếp xúc, bằng nhau, chứa, …), khoảng cách và hướng (láng giềng về hướng nào).

Mô hình raster

Mô hình raster hay còn gọi mô hình dạng ảnh (*image*) biểu diễn các đặc tính dữ liệu bởi ma trận các ô (*cell*) trong không gian liên tục. Mỗi ô có chỉ số tọa độ (*coordinate*) và các thuộc tính liên quan. Mỗi vùng được chia thành các hàng và cột, mỗi ô có thể là hình vuông hoặc hình chữ nhật và chỉ có duy nhất một giá trị.



Hình 1.8: Biểu diễn thế giới bằng mô hình raster

Trên thực tế, chọn kiểu mô hình nào để biểu diễn bản đồ là câu hỏi luôn đặt ra với người sử dụng. Việc lưu trữ kiểu đối tượng nào sẽ quyết định mô hình sử dụng. Ví dụ nếu lưu vị trí của các khách hàng, các trạm rút tiền hoặc dữ liệu cần tổng hợp theo từng vùng như vùng theo mã bưu điện, các hồ chứa nước, … thì sử dụng mô hình vector. Nếu đối tượng quản lý được phân loại liên tục như loại đất, mức nước hay độ cao của núi, … thì thường dùng mô hình raster. Đồng thời, nếu dữ liệu thu thập từ các nguồn khác nhau được dùng một mô hình nào đó thì có thể chuyển đổi từ mô hình này sang mô hình khác để phục vụ tốt cho việc xử lý của người dùng.

Mỗi mô hình có ưu điểm và nhược điểm khác nhau. Về mặt lưu trữ, việc lưu trữ giá trị của tất cả các ô/điểm ảnh trong mô hình raster đòi hỏi không gian nhớ lớn hơn so với việc chỉ lưu các giá trị khi cần trong mô hình vector. Cấu trúc dữ liệu lưu trữ của raster đơn giản, trong khi vector dùng các cấu trúc phức tạp hơn. Dung lượng lưu trữ trong mô hình raster có thể lớn hơn gấp 10 đến 100 lần so với mô hình vector. Đối với thao tác chồng phủ (xem mục 1.1.5.4), mô hình raster cho phép

thực hiện một cách dễ dàng, trong khi mô hình vector lại phức tạp và khó khăn hơn. Về mặt hiển thị, mô hình vector có thể hiển thị đồ họa vector giống như bản đồ truyền thống, còn mô hình raster chỉ hiển thị ảnh nên có thể xuất hiện hình răng cưa tại đường biên của các đối tượng tùy theo độ phân giải của tệp raster. Với dữ liệu vector, người dùng có thể bổ sung, co dãn hoặc chiếu bản đồ, thậm chí có thể kết hợp với các tầng bản đồ khác thuộc các nguồn khác nhau. Hiện nay, mô hình vector được sử dụng nhiều trong các hệ thống GIS bởi các lý do trên, ngoài ra mô hình này cho phép cập nhật và duy trì đơn giản, dễ truy vấn dữ liệu.

Tuy nhiên trong đề tài này mô hình được luận văn đề cập đến là mô hình véc tơ.

1.1.5 Tìm kiếm và các kỹ thuật phân tích dữ liệu không gian trong GIS:

Các phép phân tích và xử lý dữ liệu không gian là một trong 5 yếu tố cấu thành nên một hệ thông tin địa lý (xem mục 1.1.2). Mục này đề cập đến một số phép phân tích xử lý dữ liệu cơ bản nhất của một hệ GIS. Các thao tác trên dữ liệu không gian thường chia làm hai lớp bài toán cơ bản là các bài toán về tìm kiếm và phân tích không gian và các bài toán về xử lý dữ liệu không gian.

Lớp bài toán tìm kiếm và phân tích không gian: bao gồm các bài toán liên quan đến việc khai thác thông tin và tri thức từ dữ liệu không gian. Ví dụ như bài toán tìm kiếm đối tượng trên bản đồ theo thuộc tính, bài toán phân tích đường đi, tìm đường…

Lớp bài toán xử lý dữ liệu không gian: bao gồm các bài toán thao tác trực tiếp tới khuôn dạng, giá trị của dữ liệu không gian, làm thay đổi dữ liệu không gian. Ví dụ như các thao tác nắn chỉnh dữ liệu, tổng quát hóa dữ liệu, chuyển đổi hệ tọa độ, chuyển đổi khuôn dạng dữ liệu…Dưới đây đề cập khái quát một số phép phân tích và xử lý dữ liệu không gian chính.

1.1.5.1 Tìm kiếm theo vùng

Là phép phân tích không gian đơn giản nhất, phép phân tích này thực hiện tìm kiếm đối tượng bản đồ trong một vùng không gian cho trước. Vùng này có thể là một cửa sổ hình chữ nhật. Đây là phép truy vấn không gian cơ bản trong GIS, tuy

nhiên mức độ phức tạp của nó cao hơn truy vấn query trong cơ sở dữ liệu cổ điển bởi khả năng cắt xén đối tượng nếu đối tượng đó chỉ nằm một phần trong cửa sổ truy vấn.

1.1.5.2 Tìm kiếm lân cận

Phép phân tích này thực hiện tìm kiếm các đối tượng địa lý trong vùng cận kề với một hoặc một tập đối tượng địa lý biết trước. Có một vài kiểu tìm kiếm cận kề như:

- Tìm kiếm trong vùng mở rộng (vùng đệm) của một đối tượng: Ví dụ: Tìm các trạm thu phát sóng điện thoại di động BTS nằm trong vùng phủ sóng của một trạm BTS nào đó.
- Tìm kiếm liền kề: Ví dụ như tìm các thửa đất liền kề với thửa đất X nào đó.

1.1.5.3 Phân tích đường đi và dẫn đường

Phân tích đường đi là tiến trình tìm đường đi ngắn nhất, giá rẻ nhất giữa hai vị trí trên bản đồ. Giải pháp cho bài toán này dựa trên việc sử dụng mô hình dữ liệu mạng hay mô hình dữ liệu raster trên cơ sở lưới vùng. Mô hình dữ liệu mạng lưu trữ đối tượng đường đi dưới dạng cung và giao của chúng dưới dạng nút, việc tìm đường bao gồm việc duyệt qua các đường đi từ điểm đầu tới điểm cuối qua các cung nút và chỉ ra cung đường nào ngắn nhất. Trong mô hình raster, việc tìm đường thực hiện bởi sự dịch chuyển từ một tế bào sang tế bào lân cận của nó.

1.1.5.4 Tìm kiếm hiện tượng và bài toán chồng phủ

Việc tìm kiếm hiện tượng trong GIS bao gồm tìm kiếm hiện tượng độc lập hoặc tìm kiếm tổ hợp các hiện tượng.

Tìm kiếm hiện tượng độc lập là bài toán đơn giản, chỉ bao hàm tìm kiếm một hiện tượng, thực thể mà không quan tâm đến một hiện tượng, thực thể khác. Việc tìm kiếm đơn giản chỉ là truy nhập dữ liệu không gian dựa trên thuộc tính đã xác định trước. Ví dụ như tìm các tỉnh, thành phố có dân số lớn hơn 2 triệu người…

Tìm kiếm tổ hợp thực thể là bài toán phức tạp hơn, nhưng lại là bài toán hấp dẫn và là thế mạnh của GIS, việc tìm kiếm liên quan đến nhiều thực thể hay lớp

thực thể, ví dụ, tính diện tích đất nông nghiệp của huyện Vĩnh Bảo, thành phố Hải Phòng. Bài toán này đòi hỏi phải tổ hợp 2 lớp thực thể địa lý là lớp đất nông nghiệp của thành phố Hải Phòng và lớp ranh giới hành chính thành phố Hải Phòng. Kiểu bài toán này trong GIS gọi là bài toán chồng phủ bản đồ.

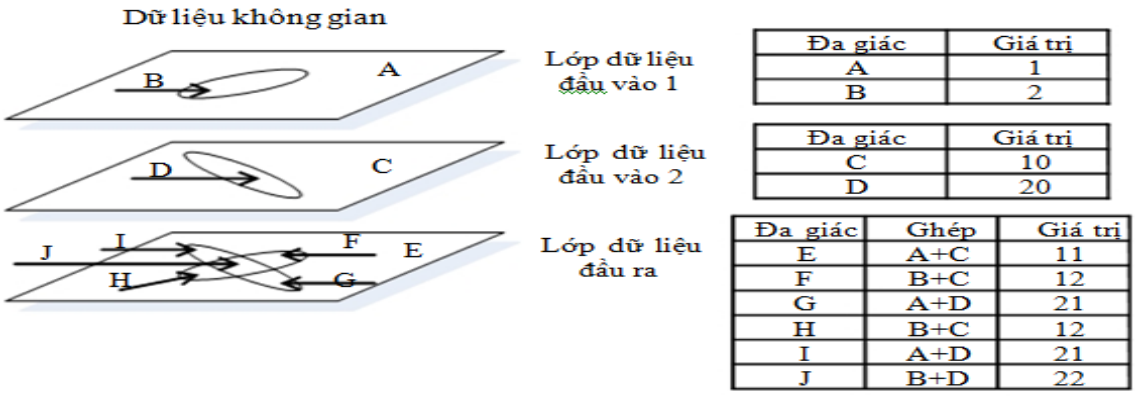
Bài toán chồng phủ bản đồ

Nhiều vấn đề trong GIS đòi hỏi sử dụng lớp chồng xếp của các lớp dữ liệu chuyên đề khác nhau. Ví dụ chúng ta muốn biết vị trí của các quán ăn ngon nằm trong khu vực gần trường học; hoặc là vị trí các siêu thị, nhà hàng, trường học,… tại thành phố Hải Phòng. Trong ví dụ liên quan đến đất xói mòn trên, một lớp dữ liệu đất đai có thể được sử dụng để nhận biết các khu vực đất đai bị xói mòn, đồng thời lớp dữ liệu về hiện trạng sử dụng đất cũng được sử dụng để nhận biết vị trí các vùng đất sử dụng cho mục đích nông nghiệp. Thông thường thì các đường ranh giới của vùng đất bị xói mòn sẽ không trùng với các đường ranh giới của các vùng đất nông nghiệp, do đó, dữ liệu về loại đất và sử dụng đất sẽ phải được kết hợp lại với nhau theo một cách nào đó. Chồng phủ bản đồ chính là phương tiện hàng đầu hỗ trợ việc thực hiện phép kết hợp dữ liệu đó

Theo mô hình vector, các đối tượng địa lý được biểu diễn dưới dạng các điểm, đường và vùng. Vị trí của chúng được xác định bởi các cặp tọa độ và thuộc tính của chúng được ghi trong các bảng thuộc tính.

Với từng kiểu bản đồ, người ta phân biệt ba loại chồng phủ bản đồ vector sau:

+ ***Chồng phủ đa giác trên đa giác***: Chồng phủ đa giác là một thao tác không gian trong đó một lớp bản đồ chuyên đề dạng vùng chứa các đa giác được chồng xếp lên một lớp khác để hình thành một lớp chuyên đề mới với các đa giác mới. Mỗi đa giác mới là một đối tượng mới được biểu diễn bằng một dòng trong bảng thuộc tính. Mỗi đối tượng có một thuộc tính mới được biểu diễn bằng một cột trong bảng thuộc tính.



Đa giác	Giá trị
A	1
B	2

Đa giác	Giá trị
C	10
D	20

Đa giác	Ghép	Giá trị
E	A+C	11
F	B+C	12
G	A+D	21
H	B+C	12
I	A+D	21
J	B+D	22

Hình 1.9: Chồng phủ đa giác

Việc chồng phủ và so sánh hai bộ dữ liệu hình học có nguồn gốc và độ chính xác khác nhau thường sinh ra một số các đa giác nhỏ. Các đa giác này có thể được loại bỏ theo diện tích, hình dạng và các tiêu chuẩn khác. Tuy nhiên, trong thực tế, khó đặt ra các giới hạn để giảm được số đa giác nhỏ không mong muốn đồng thời giữ lại các đa giác khác có thể nhỏ hơn nhưng hữu ích.

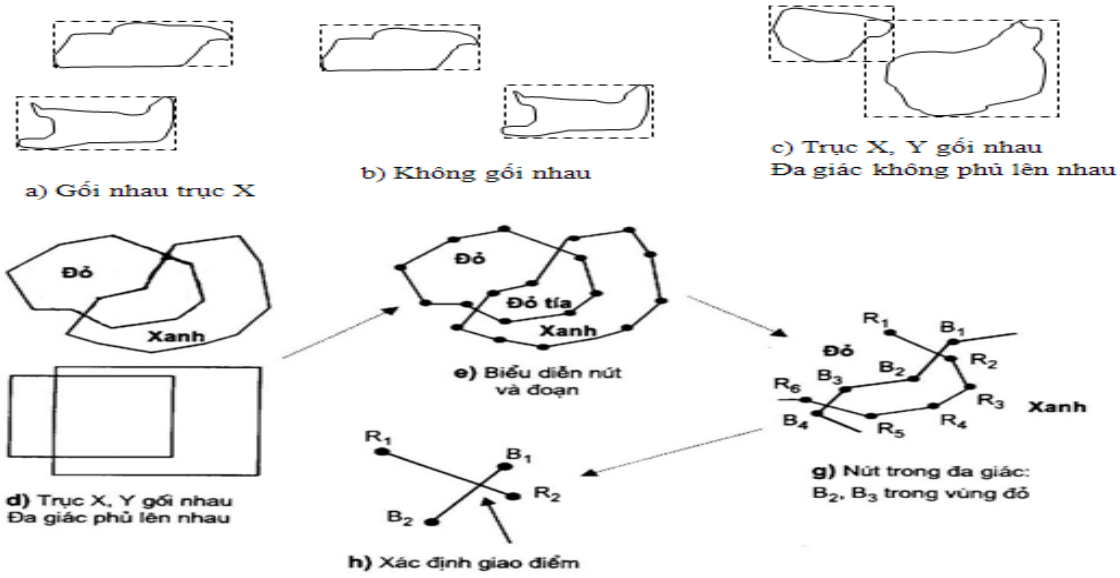
+ Chồng phủ điểm trên đa giác

Các đối tượng điểm cũng có thể được chồng xếp trên các đa giác. Các điểm sẽ được gán các thuộc tính của đa giác mà trên đó chúng được chồng lên. Các bảng thuộc tính sẽ được cập nhật sau khi tất cả các điểm được kết hợp với đa giác.

+ Chồng phủ đường trên đa giác

Các đối tượng đường cũng có thể được chồng xếp trên các đa giác để tạo ra một bộ các đường mới chứa các thuộc tính của các đường ban đầu và của các đa giác. Cũng như trong chồng xếp đa giác, các điểm cắt được tính toán, các nút và các liên kết được hình thành, topo được thiết lập và cuối cùng là các bảng thuộc tính được cập nhật.

Minh họa cụ thể cho vấn đề chồng xếp bản đồ chúng ta sẽ xét tới tiến trình phủ đa giác. Tiến trình này được minh họa bởi hình sau:



Hình 1.10: Tiến trình phủ đa giác

Tiến trình tổng quát của phủ đa giác là tạo ra các đa giác mới từ các đa giác cho trước bao gồm các bước nhỏ sau:

- Nhận dạng các đoạn thẳng
- Lập chữ nhật bao tối thiểu đa giác
- Khẳng định các đoạn thẳng của một đa giác thuộc lớp bản đồ này ở trong đa giác của lớp bản đồ khác (phủ) bằng tiến trình "điểm trong đa giác".
- Tìm giao của các đoạn thẳng là cạnh đa giác
- Lập các bản ghi cho đoạn thẳng mới và lập quan hệ topo của chúng
- Lập các đa giác mới từ các đoạn thẳng phù hợp
- Gán lại nhãn và các dữ liệu thuộc tính nếu có cho đa giác

1.1.5.5 Nắn chỉnh dữ liệu không gian

Dữ liệu bản đồ ngoài việc được kiểm tra độ chính xác về mặt hình học còn cần được kiểm tra hiệu chỉnh về độ chính xác không gian. Các sai lệch về mặt không gian thường phát sinh trong quá trình đo đạc hoặc số hoá bản đồ giấy, dẫn đến việc toạ độ các điểm trên bản đồ không trùng khớp với toạ độ đo thực địa, do đó cần có thao tác nắn chỉnh toạ độ bản đồ.

Có nhiều phương pháp nắn chỉnh bản đồ, một phương pháp phổ biến là phương pháp sử dụng điểm điều khiển mặt đất, hay còn gọi là *phương pháp tấm cao*

su[1] . Phương pháp này dựa trên ý tưởng là chọn một số điểm thực tế trên mặt đất, đo đạc chính xác tọa độ của điểm đó, dùng các điểm này làm điểm khống chế. Đối chiếu với bản đồ để tìm ra các điểm tương ứng với các điểm khống chế, thường chọn các điểm khống chế là những điểm dễ đánh dấu mốc, ví dụ như các ngã tư, giao lộ, sân bay, bờ biển...để có thể dễ dàng tìm thấy điểm tương ứng trên bản đồ. Lúc này, việc nắn chỉnh bản đồ tương đương với việc làm biến dạng bản đồ để đưa các điểm tương ứng về trùng với các điểm khống chế. Ta có thể tưởng tượng cả bản đồ giống như một tấm cao su, sử dụng các đinh ghim cắm tại các điểm tương ứng với điểm khống chế, sau đó dịch chuyển các đinh ghim này về đúng vị trí của các điểm khống chế, khi đó, cả bản đồ sẽ như một tấm cao su bị co kéo bởi các đinh ghim để về đúng tọa độ thực tế. Như vậy, cần có một hàm số để biến đổi toàn bộ các giá trị của các điểm bản đồ sang giá trị mới sao cho các điểm tương ứng với điểm khống chế trở về gần điểm khống chế nhất.

1.1.6 Ứng dụng của GIS:

1.1.6.1 Các lĩnh vực liên quan với hệ thông tin địa lý

Công nghệ GIS được sử dụng trong nhiều lĩnh vực khác nhau như bản đồ học, đầu tư, quản lý nguồn tài nguyên, quản lý tài sản, khảo cổ học (*archaeology*), phân tích điều tra dân số, đánh giá sự tác động lên môi trường, kế hoạch đô thị, nghiên cứu tội phạm,… Việc trích rút thông tin từ dữ liệu địa lý thông qua hệ thống GIS bao gồm các câu hỏi cơ bản sau [1]:

*** Nhận diện** (*identification*): Nhận biết tên hay các thông tin khác của đối tượng bằng việc chỉ ra vị trí trên bản đồ. Ví dụ, có cái gì tại tọa độ (X, Y).

*** Vịtrí** (*location*): Câu hỏi này đưa ra một hoặc nhiều vị trí thỏa mãn yêu cầu. Nó có thể là tập tọa độ hay bản đồ chỉ ra vị trí của một đối tượng cụ thể, hay toàn bộ đối tượng. Ví dụ, cho biết vị trí các văn phòng của công ty nào đó trong thành phố.

*** Xu thế** (*trend*): Câu hỏi này liên quan đến các dữ liệu không gian tạm thời. Ví dụ, câu hỏi liên quan đến xu hướng phát triển thành thị dẫn tới chức năng hiển thị bản đồ của GIS để chỉ ra các vùng lân cận được xây dựng từ 1990 đến 2000.

*** Tìm đường đi tối ưu** (*optimal path*): Trên cơ sở mạng lưới đường đi (hệ thống đường bộ, đường thủy...), câu hỏi là cho biết đường đi nào là tối ưu nhất (rẻ nhất, ngắn nhất,...) giữa 2 vị trí cho trước.

*** Mẫu** (*pattern*): Câu hỏi này khá phức tạp, tác động trên nhiều tập dữ liệu. Ví dụ, cho biết *quan hệ giữa* khí hậu địa phương *và* vị trí của các nhà máy, công trình công cộng trong vùng lân cận.

*** Mô hình** (*model*): Câu hỏi này liên quan đến các hoạt động lập kế hoạch và dự báo.Ví dụ, cần phải nâng cấp, xây dựng hệ thống mạng lưới giao thông, điện như thế nào nếu phát triển khu dân cư về phía bắc thành phố.

1.1.6.2 Những bài toán của GIS

Một số ứng dụng cụ thể của GIS thường gặp trong thực tế bao gồm:

* Quản lý và lập kế hoạch mạng lưới giao thông đường bộ: giải quyết các nhu cầu như tìm kiếm địa chỉ, chỉ dẫn đường đi, phân tích không gian, chọn địa điểm xây dựng, lập kế hoạch phát triển mạng lưới giao thông…

* Giám sát tài nguyên thiên nhiên, môi trường: giúp quản lý hệ thống sông ngòi, vùng đất nông nghiệp, thảm thực vật, vùng ngập nước, phân tích tác động môi trường…

* Quản lý đất đai: giám sát, lập kế hoạch sử dụng đất, quy hoạch…

* Quản lý và lập kế hoạch các dịch vụ công cộng: tìm địa điểm phù hợp cho việc bố trí các công trình công cộng, cân đối tải điện, phân luồng giao thông…

* Phân tích, điều tra dân số, lập bản đồ y tế, bản đồ vùng dịch bệnh…

Trong địa lý vị trí đặt cây ATM tạo thành các lớp địa lý. Các địa điểm nhà hàng, khách sạn, siêu thị, bệnh viện, ngân hàng, trường học,... cũng tạo thành các lớp địa lý. Làm thế nào để tìm ra vị trí đặt cây ATM tối ưu. Việc đặt cây ATM ở vị trí được coi là tối ưu nếu như vị trí đó ở gần những nơi có nhu cầu sử dụng thẻ ATM nhiều nhất chẳng hạn như ở gần các siêu thị, nhà hàng, khách sạn...Vậy để tìm ra vị trí tối ưu để đặt cây ATM cần phải tiến hành phân cụm các vị trí nhà hàng, khách sạn, siêu thị thành các cụm dữ liệu. Sau đó xếp chồng các cụm để tìm giao của vùng xếp chồng - đó là nơi vị trí thích hợp nhất để đặt cây ATM. Vậy khai

phá dữ liệu là gì? phân cụm dữ liệu là gì?. Nội dung của phần 2 sẽ đề cập về khai phá dữ liệu và phân cụm dữ liệu.

1.2 Khái quát về khai phá dữ liệu và phân cụm dữ liệu

1.2.1 Khái quát về khai phá dữ liệu:

Có nhiều định nghĩa về Khai phá dữ liệu (Data Mining) được đưa ra, nhìn chung, có thể hiểu khai phá dữ liệu là quá trình tìm ra các quy luật, các mối quan hệ và các thông tin có ích tiềm ẩn giữa các mẫu dữ liệu trong một cơ sở dữ liệu. Các thông tin có ích này không hoặc khó có thể được tìm ra bởi các hệ cơ sở dữ liệu giao dịch truyền thống. Các tri thức mà khai phá dữ liệu mang lại là công cụ hữu hiệu đối với tổ chức trong việc hoạch định chiến lược và ra quyết định kinh doanh.

Khác với các câu hỏi mà hệ cơ sở dữ liệu truyền thống có thể trả lời như:

* Hãy hiển thị số tiền của bà A trong ngày 21 tháng Tám? ghi nhận riêng lẻ do xử lý giao dịch trực tuyến (on-line transaction processing – OLTP).

* Có bao nhiêu nhà đầu tư nước ngoài mua cổ phiếu X trong tháng trước ? ghi nhận thống kê do hệ thống hỗ trợ quyết định thống kê (stastical decision suppport system - DSS)

* Hiển thị mọi cổ phiếu trong CSDL với mệnh giá tăng ? ghi nhận dữ liệu đa chiều do xử lý phân tích trực tuyến (on-line analytic processing - OLAP).

Khai phá dữ liệu giúp trả lời các câu hỏi mang tính trừu tượng, tổng quát hơn như:

- Các cổ phiếu tăng giá có đặc trưng gì ?
- Tỷ giá US$ - DMark có đặc trưng gì ?
- Hy vọng gì về cổ phiếu X trong tuần tiếp theo ?
- Trong tháng tiếp theo, sẽ có bao nhiêu đoàn viên công đoàn không trả được nợ của họ ?
- Những người mua sản phẩm Y có đặc trưng gì ?

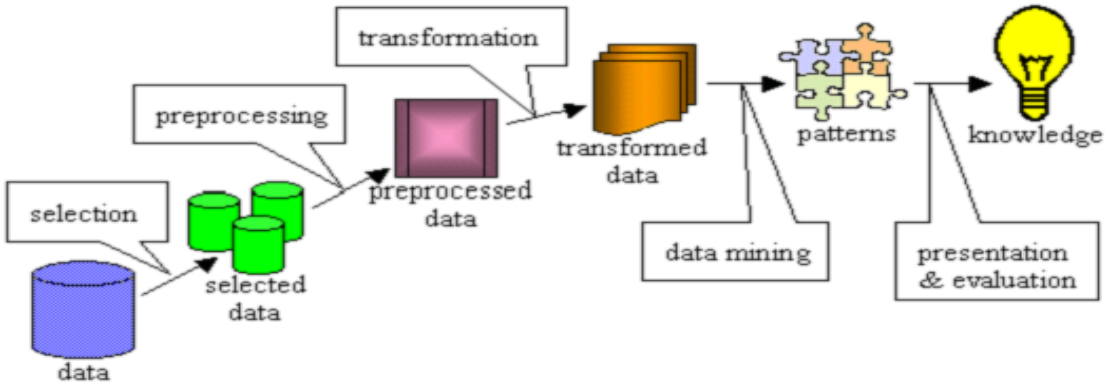
Khai phá dữ liệu là sự kết hợp của nhiều chuyên ngành như cơ sở dữ liệu, học máy, trí tuệ nhân tạo, lý thuyết thông tin, xác suất thống kê, tính toán hiệu năng cao và các phương pháp tính toán mềm…

1.2.1.1 Tiến trình khai phá dữ liệu

Một số nhà khoa học xem khai phá dữ liệu là một cách gọi khác của một thuật ngữ rất thông dụng: *Khám phá tri thức từ cơ sở dữ liệu* (Knowledge Discovery in Database- KDD). Mặt khác, khi chia các bước trong quá trình khám phá tri thức, một số nhà nghiên cứu lại cho rằng, KPDL chỉ là một bước trong quá trình khám phá tri thức[4].

Như vậy, khi xét ở mức tổng quan thì hai thuật ngữ này là tương đương nhau, nhưng khi xét cụ thể thì KPDL được xem là một bước trong quá trình khám phá tri thức.

Nhìn chung, khai phá dữ liệu hay khám phá tri thức từ cơ sở dữ liệu bao gồm các bước sau [6]:



Hình 1.11: Tiến trình khám phá tri thức từ cơ sở dữ liệu

Trích chọn dữ liệu: Là quá trình trích lọc một lượng dữ liệu phù hợp, cần thiết từ tập dữ liệu lớn (cơ sở dữ liệu tác nghiệp, kho dữ liệu)…

Tiền xử lý dữ liệu: Là bước làm sạch dữ liệu (xử lý dữ liệu không đầy đủ, dữ liệu nhiễu, ngoại lai, dữ liệu không nhất quán…), rút gọn dữ liệu (lấy mẫu dữ liệu, lượng tử hóa…), rời rạc hóa dữ liệu. Kết quả sau bước này là dữ liệu có tính nhất quán, đầy đủ, được rút gọn và được rời rạc hóa.

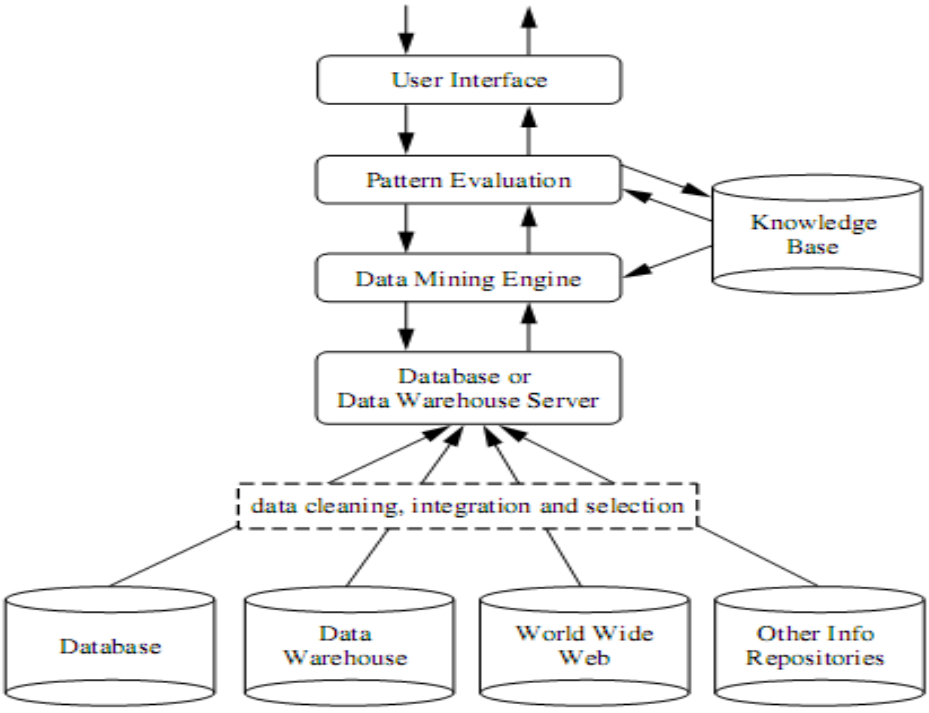
Chuyển đổi dữ liệu: Là bước chuẩn hóa khuôn dạng và làm mịn dữ liệu, nhằm đưa dữ liệu về dạng thuận lợi nhất để phục vụ cho việc áp dụng các giải thuật khai phá dữ liệu ở bước sau.

Khai phá dữ liệu: Sử dụng các phương pháp, kỹ thuật, các thuật toán để trích lọc ra mẫu có ý nghĩa cùng với các tri thức, quy luật, biểu thức mô tả mối quan hệ của dữ liệu trong một khía cạnh nào đó. Đây là bước quan trọng và tốn nhiều thời gian nhất của toàn bộ tiến trình KDD.

Đánh giá và biểu diễn tri thức: Trình bày các tri thức, quy luật, biểu thức có ý nghĩa đã tìm được ở bước trước dưới các dạng thức gần gũi, dễ hiểu đối với người sử dụng như đồ thị, biểu đồ, cây, bảng biểu, luật…Đồng thời đưa ra những đánh giá về tri thức khám phá được theo những tiêu chí nhất định.

Trong giai đoạn khai phá dữ liệu, có thể cần sự tương tác của con người để điều chỉnh cách thức và kỹ thuật sử dụng trong khai phá, nhằm thu được tri thức phù hợp nhất.

Dựa trên các bước của quá trình khai phá dữ liệu như trên, kiến trúc điển hình của một hệ khai phá dữ liệu có thể bao gồm các thành phần như sau:



Hình 1.12: Kiến trúc điển hình của một hệ khai phá dữ liệu

1.2.1.2 Các mô hình khai phá dữ liệu

Mô hình khai phá dữ liệu là mô tả về phương pháp, cách thức khai phá thông tin từ dữ liệu và định hướng kiểu tri thức cần khai phá.

Một mô hình khai phá dữ liệu có thể được mô tả ở 2 mức:

* Mức chức năng (Function level): Mô tả mô hình bằng những thuật ngữ về dự định sử dụng. Ví dụ: Phân lớp, phân cụm…

* Mức biểu diễn (Representation level): Biểu diễn cụ thể một mô hình. Ví dụ: Mô hình log-linear, cây phân lớp, phương pháp láng giềng gần nhất…

Các mô hình khai phá dữ liệu dựa trên 2 kiểu học: có giám sát và không giám sát (đôi khi được nói đến như là học trực tiếp và không trực tiếp -directed and undirected learning) [7]

* Các hàm học có giám sát (Supervised learning functions) được sử dụng để dự đoán giá trị. Một ví dụ của thuật toán học có giám sát bao gồm Naive Bayes cho phân lớp (classification).

* Các hàm học không giám sát được dùng để tìm ra cấu trúc bên trong, các quan hệ hoặc tính giống nhau trong nội dung dữ liệu nhưng không có lớp hay nhãn nào được gán ưu tiên. Ví dụ của các thuật toán học không giám sát gồm phân nhóm k-mean (k-mean clustering) và các luật kết hợp Apriori.

Tương ứng có 2 loại mô hình khai phá dữ liệu:

*** Các mô hình dự báo** (học có giám sát):

- Phân lớp: nhóm các đối tượng thành các lớp riêng biệt và dự đoán một đối tượng sẽ thuộc vào lớp nào.

- Hồi qui (Regression): xấp xỉ hàm và dự báo các giá trị liên tục

*** Các mô hình mô tả** (học không giám sát):

- Phân cụm (Clustering): Tìm các nhóm tự nhiên trong dữ liệu

- Các mô hình kết hợp (Association models): Phân tích "giỏ hàng"

- Trích chọn đặc trưng (Feature extraction): Tạo các thuộc tính (đặc trưng) mới như là kết hợp của các thuộc tính ban đầu

1.2.1.3 Các hướng tiếp cận và kỹ thuật sử dụng trong khai phá dữ liệu

Xuất phát từ hai mô hình khai phá dữ liệu chủ yếu như đã đề cập ở trên, các bài toán (hay chức năng) khai phá dữ liệu giải quyết thường được phân chia thành các dạng sau [6]:

*** Mô tả khái niệm** (concept description & summarization): . Tổng quát, tóm tắt các đặc trưng dữ liệu, Ví dụ: tóm tắt văn bản…

*** Phân lớp và dự đoán** (classification & prediction): Xây dựng các mô hình (chức năng) để mô tả và phân biệt khái niệm cho các lớp hoặc khái niệm để dự đoán trong tương lai, xếp một đối tượng vào một trong những lớp đã biết trước.

Ví dụ: phân lớp vùng địa lý theo dữ liệu thời tiết. Hướng tiếp cận này thường sử dụng một số kỹ thuật của machine learning như cây quyết định (decision tree), mạng nơ ron nhân tạo (neural network), .v.v. Phân lớp còn được gọi là học có giám sát (học có thầy – supervised learning).

*** Luật kết hợp** (association rules): Biểu diễn mối tương quan nhân quả giữa dữ liệu và xu hướng của dữ liệu dưới dạng luật biểu diễn tri thức ở dạng khá đơn giản.

Ví dụ: "60 % nam giới vào siêu thị nếu mua bia thì có tới 80% trong số họ sẽ mua thêm thịt bò khô". Luật kết hợp được ứng dụng nhiều trong lĩnh vực kinh doanh, y học, tin-sinh, tài chính & thị trường chứng khoán, .v.v.

*** Khai phá chuỗi theo thời gian** (sequential/temporal patterns): tương tự như khai phá 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 vì nó có tính dự báo cao.

*** Phân cụm** (clustering/segmentation): xếp các đối tượng theo từng cụm (số lượng cũng như tên của cụm chưa được biết trước. Phân cụm còn được gọi là học không giám sát (học không có thầy – unsupervised learning).

*** Phân tích bất thường (ngoại lê)**: Phát hiện sự bất thường của dữ liệu: đối tượng dữ liệu không tuân theo hành vi chung của toàn bộ dữ liệu nhằm phát hiện gian lận hoặc phân tích các sự kiện hiếm…

1.2.1.4 Các dạng dữ liệu có thể khai phá

Khai phá dữ liệu là kết hợp của nhiều lĩnh vực khoa học, xử lý nhiều kiểu dữ liệu khác nhau [6]. Sau đây là một số kiểu dữ liệu điển hình:

* CSDL quan hệ (relational databases)

* CSDL đa chiều (multidimensional structures, data warehouses)

* CSDL dạng giao dịch (transactional databases)

* CSDL quan hệ - hướng đối tượng (object-relational databases)

* Dữ liệu không gian và thời gian (spatial and temporal data)

* Dữ liệu chuỗi thời gian (time-series data)

* CSDL đa phương tiện (multimedia databases) như âm thanh (audio), hình ảnh (image), phim ảnh (video), .v.v.

* Dữ liệu Text và Web (text database & www)

1.2.1.5 Các ứng dụng của khai phá dữ liệu

Khai phá dữ liệu được vận dụng để giải quyết các vấn đề thuộc nhiều lĩnh vực khác nhau. Chẳng hạn như giải quyết các bài toán phức tạp trong các ngành đòi hỏi kỹ thuật cao, như tìm kiếm mỏ dầu, từ ảnh viễn thám, cảnh báo hỏng hóc trong các hệ thống sản xuất; Được ứng dụng cho việc quy hoạch và phát triển các hệ thống quản lý và sản xuất trong thực tế như dự đoán tải sử dụng điện, mức độ tiêu thụ sản phẩm, phân nhóm khách hàng; Áp dụng cho các vấn đề xã hội như phát hiện tội phạm, tăng cường an ninh… Có thể liệt kê ra đây một số ứng dụng điển hình như:

* Phân tích dữ liệu và hỗ trợ ra quyết định (data analysis & decision support)

* Điều trị y học (medical treatment): mối liên hệ giữa triệu chứng, chẩn đoán và phương pháp điều trị (chế độ dinh dưỡng, thuốc men, phẫu thuật, …).

* Text mining & Web mining: phân lớp văn bản và các trang web, tóm tắt văn bản, .v.v.

* Tin-sinh (bio-informatics): tìm kiếm, đối sánh các hệ gene và thông tin di truyền, mối liên hệ giữa một số hệ gene và một số bệnh di truyền, .v.v.

* Tài chính và thị trường chứng khoán (finance & stock market): phân tích tình hình tài chính và dự báo giá của các loại cổ phiếu trong thị trường chứng khoán,...

* Bảo hiểm (insurance)

* ...

1.2.2 Phân cụm dữ liệu:

Phân cụm dữ liệu là quá trình nhóm một tập các đối tượng thực thể hay trừu tượng thành lớp các đối tượng tương tự nhau theo một hoặc nhiều tiêu chí nào đó. Một cụm là một tập hợp các đối tượng dữ liệu mà các phần tử của nó tương tự nhau cùng trong một cụm và phi tương tự với các đối tượng trong các cụm khác. Một cụm các đối tượng dữ liệu có thể xem như là một nhóm trong nhiều ứng dụng.

Cho tới nay, một số lượng lớn các giải thuật phân cụm đã được đề xuất. Việc lựa chọn giải thuật phân cụm tuỳ thuộc vào kiểu dữ liệu cho sẵn, mục đích riêng và ứng dụng. Nếu như phép phân tích cụm được dùng như một công cụ mô tả hay thăm dò thì có thể thử một vài giải thuật trên cùng dữ liệu để xem xem dữ liệu có thể thể hiện được điều gì.

Nhìn chung, các phương pháp phân cụm được phân thành các loại chính như sau:

* Phân cụm phân hoạch
* Phân cụm phân cấp
* Phân cụm dựa trên mật độ
* Phân cụm dựa trên lưới

Phần tiếp theo sẽ khảo sát một số phương pháp phân cụm và xem xét chi tiết một vài giải thuật phân cụm đã được cài đặt trong chương trình ứng dụng của học viên.

1.2.2.1 Phân cụm phân hoạch

Cho trước một cơ sở dữ liệu với n đối tượng hay các bộ dữ liệu, một phương pháp phân chia được xây dựng để chia dữ liệu thành k phần, mỗi phần đại diện cho một cụm, $k \leq n$. Đó là phân loại dữ liệu vào trong k nhóm, chúng thoả các yêu cầu sau: Mỗi nhóm phải chứa ít nhất một đối tượng, Mỗi đối tượng phải thuộc về chính xác một nhóm.

Cho trước k là số lượng các phần chia cần xây dựng, phương pháp phân chia tạo lập phép phân chia ban đầu. Sau đó nó dùng kỹ thuật lặp lại việc định vị, kỹ thuật này cố gắng cải thiện sự phân chia bằng cách gỡ bỏ các đối tượng từ nhóm

này sang nhóm khác. Tiêu chuẩn chung của một phân chia tốt là các đối tượng trong cùng cụm là "gần" hay có quan hệ với nhau, ngược lại, các đối tượng của các cụm khác nhau lại "tách xa" hay rất khác nhau. Có nhiều tiêu chuẩn khác nhau để đánh giá chất lượng các phép phân chia.

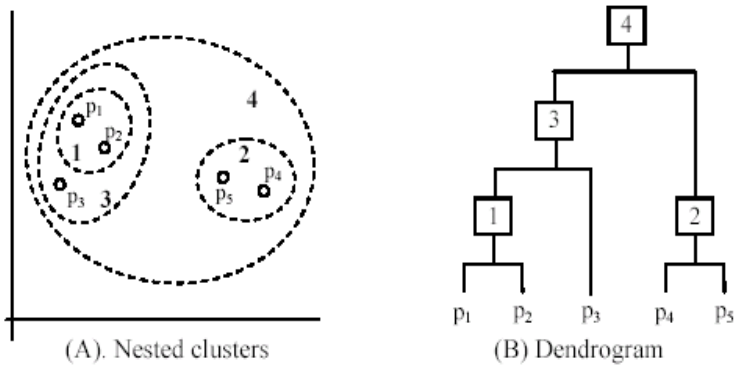
Trong phân cụm dựa trên phép phân chia, hầu hết các ứng dụng làm theo một trong hai phương pháp heuristic phổ biến: Giải thuật k-means với mỗi cụm được đại diện bởi giá trị trung bình của các đối tượng trong cụm; Giải thuật k-medoids với mỗi cụm được đại diện bởi một trong số các đối tượng định vị gần tâm của cụm. Các phương pháp phân cụm heuristic này làm việc tốt khi tìm kiếm các cụm có hình cầu trong các cơ sở dữ liệu có kích thước từ nhỏ tới trung bình. Để tìm ra các cụm với các hình dạng phức tạp và phân cụm cho các tập dữ liệu rất lớn, các phương pháp dựa trên phân chia cần được mở rộng.

1.2.2.2 Phân cụm phân cấp

Phương pháp phân cấp tạo ra một phân rã của tập đối tượng dữ liệu dưới dạng cây (*dendrogram*, theo Hy Lạp thì *dendron* là "*cây*", *gramma* là "*vẽ*"), trong đó chia đệ quy cơ sở dữ liệu thành các tập con nhỏ hơn, để minh họa trật tự các cụm được sinh ra. Cây có thể biểu diễn dưới 2 dạng là *bottom-up* và *top-down*.

Tiếp cận *bottom-up* hay còn gọi là tiếp cận hội tụ (*agglomerative*), bắt đầu với mỗi đối tượng thành lập một cụm riêng biệt. Sau đó tiến hành hợp hoặc nhóm các đối tượng theo một vài tiêu chí đo như khoảng cách giữa trung tâm của 2 nhóm. Thuật toán kết thúc khi tất cả các nhóm được hợp thành một nhóm (nút gốc của cây) hoặc thỏa mãn điều kiện dừng.

Còn tiếp cận *top-down* được gọi là tiếp cận phân chia (*divisive*), bắt đầu coi tất cả các đối tượng trong một cụm. Tại mỗi bước lặp thì cụm được phân chia thành cụm nhỏ hơn theo tiêu chí nào đó. Việc phân chia dừng khi mỗi đối tượng là một cụm hoặc thỏa mãn điều kiện dừng.



Hình 1.13: Phân cụm phân cấp

Ưu điểm của phương pháp này là kết hợp linh hoạt vào mức độ chi tiết, dễ dàng xử lý với bất kỳ kiểu đo độ tương tự/khoảng cách nào, thích hợp với mọi kiểu dữ liệu thuộc tính. Tuy nhiên, phương pháp tồn tại nhược điểm là điều kiện để dừng vòng lặp rất mơ hồ, không cụ thể. Mặt khác, phương pháp không duyệt lại các mức trước khi xây dựng để cải tiến chất lượng các cụm.

Thuật toán xuất hiện sớm nhất của phương pháp phân cấp là thuật toán AGNES (*Agglomerative NEsting*) và DIANA (*DIvisia ANAlysic*) được Kaufman L. và Rousseeuw P. J giới thiệu vào năm 1990. Hai thuật toán này sử dụng độ đo đơn giản trong quá trình hợp/phân chia cụm, do vậy kết quả đưa ra đôi khi không chính xác [8]. Ngoài ra, phương pháp phân cấp thực hiện trên cơ sở dữ liệu không gian còn có các thuật toán CURE (*Clustering Using Representatives*), BIRCH (*Balance Iterative Reducing and Clustering using Hierarchies*), CHAMELEON.

1.2.2.3 Phân cụm dựa trên mật độ

Hầu hết các phương pháp phân chia cụm các đối tượng dựa trên khoảng cách giữa các đối tượng. Các phương pháp như vậy có thể chỉ tìm được các cụm có hình cầu và sẽ gặp khó khăn khi các cụm đang khám phá lại có hình dạng tuỳ ý. Các phương pháp phân cụm được phát triển dựa trên khái niệm mật độ. Ý tưởng chung đó là tiếp tục phát triển cụm cho trước với điều kiện là mật độ (số các đối tượng hay các điểm dữ liệu) trong "lân cận" vượt quá ngưỡng, tức là đối với mỗi điểm dữ liệu trong phạm vi một cụm cho trước thì lân cận trong vòng bán kính đã cho chứa ít nhất một số lượng điểm tối thiểu. Một phương pháp như vậy có thể được dùng để lọc ra các giá trị ngoại lai (outlier) và khám phá ra các cụm có hình dạng bất kỳ.

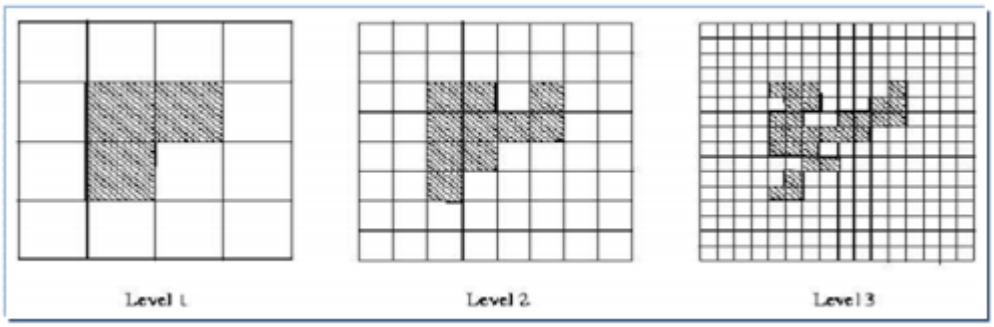
Một số khái niệm được sử dụng trong phân cụm dựa trên mật độ bao gồm:

*** Eps**: bán kính của vùng lân cận của một đối tượng, gọi là**ε**-neighborhood.

*** MinPts**: số lượng đối tượng tối thiểu được yêu cầu trong vùng lân cận Eps của một đối tượng.

1.2.2.4 Phân cụm dựa trên lưới

Phương pháp này lượng tử hoá không gian đối tượng vào trong một số hữu hạn các ô hình thành nên một cấu trúc lưới. Sau đó nó thực hiện tất cả các thao tác phân cụm trên cấu trúc lưới (tức là trên không gian đã lượng tử hoá). Thuận lợi chính của tiếp cận này là thời gian xử lý nhanh chóng của nó độc lập với số các đối tượng dữ liệu và chỉ tuỳ thuộc vào số lượng các ô trong mỗi chiều của không gian lượng tử.



Hình 1.14: Phân cụm dựa theo lưới vùng

Cách tiếp cận dựa trên lưới hiệu quả hơn so với phương pháp dựa trên mật độ và phân cấp, vì chỉ làm việc với từng đối tượng trong từng ô mà không phải đối tượng dữ liệu, mặt khác phương pháp này không trộn/hòa nhập các ô như phân cấp.

Một số thuật toán điển hình theo phương pháp dựa trên lưới như: STING (*STatistical INformation Grid*), WaveCluster, CLIQUE (*CLustering In QUEst*).

Ngoài ra còn có các tiếp cận phân cụm dữ liệu khác như phân cụm dựa trên mô hình (*model-based clustering*), phân cụm dựa trên các ràng buộc (*constraints-based clustering*)…

Nhiều giải thuật phân cụm tích hợp các ý tưởng của một vài phương pháp phân cụm, bởi vậy việc phân loại giải thuật đó không dễ như loại giải thuật chỉ phụ

thuộc vào duy nhất một loại phương pháp phân cụm. Hơn nữa, nhiều ứng dụng có thể có giới hạn phân cụm với yêu cầu tích hợp một số kỹ thuật phân cụm

1.3 Tổng kết chương

Chương 1 luận văn đã chỉ ra hệ thống thông tin địa lý được định nghĩa với nhiều cách khác nhau và đều mô tả việc nghiên cứu các thông tin địa lý và các khía cạnh khác liên quan. Một hệ thống thông tin địa lý gồm 5 thành phần: Thiết bị, phần mềm, số liệu, chuyên viên, chính sách và cách quản lý. Cả 5 thành phần kết hợp với nhau nhằm tự động quản lý và phân phối thông tin thông qua biểu diễn địa lý. Dữ liệu địa lý gồm 2 thành phần là dữ liệu không gian và dữ liệu phi không gian. Có 3 mô hình biểu diễn dữ liệu không gian: mô hình khái niệm, mô hình logic, mô hình vật lý. Các phép phân tích và xử lý dữ liệu không gian gồm: tìm kiếm theo vùng, tìm kiếm lân cận, phân tích đường đi và dẫn đường, tìm kiếm hiện tượng và bài toán chồng phủ, nắn chỉnh không gian. Ứng dụng của GIS gồm các lĩnh vực liên quan với hệ thông tin địa lý như: nhận diện, vị trí, xu thế, tìm đường đi tối ưu, mẫu và mô hình. Những bài toán của GIS thường gặp trong thực tế chẳng hạn: Quản lý và lập kế hoạch mạng lưới giao thông đường bộ, giám sát tài nguyên thiên nhiên, môi trường, quản lý đất đai, quản lý và lập kế hoạch các dịch vụ công cộng, phân tích, điều tra dân số, lập bản đồ y tế, bản đồ vùng dịch bệnh...

Tìm vị trí tối ưu đặt cây ATM tại Hải Phòng là một bài toán cần phải kết hợp hệ thống thông tin địa lý và phân cụm dữ liệu. Nội dung chương I đã khái quát về khai phá dữ liệu và phân cụm dữ liệu. Khai phá dữ liệu là quá trình tìm ra các quy luật, các mối quan hệ và các thông tin có ích tiềm ẩn giữa các mẫu dữ liệu trong một cơ sở dữ liệu. Tiến trình khai phá dữ liệu gồm: trích chọn dữ liệu, tiền xử lý dữ liệu, chuyển đổi dữ liệu, khai phá dữ liệu, đánh giá và biểu diễn tri thức. Mô hình khai phá dữ liệu gồm: Các mô hình dự báo, các mô hình mô tả. Các hướng tiếp cận và kỹ thuật sử dụng trong khai phá dữ liệu: Mô tả khái niệm, phân lớp và dự đoán, luật kết hợp, khai phá chuỗi theo thời gian, phân cụm, phân tích bất thường. Các dạng dữ liệu có thể khai phá: CSDL quan hệ, CSDL đa chiều, CSDL dạng giao dịch, CSDL quan hệ - hướng đối tượng, dữ liệu không gian và thời gian, dữ liệu chuỗi

thời gian, CSDL đa phương tiện, dữ liệu Text và Web. Các ứng dụng của khai phá dữ liệu: Phân tích dữ liệu và hỗ trợ ra quyết định, điều trị y học, Tài chính và thị trường chứng khoán, bảo hiểm... Phân cụm dữ liệu là quá trình nhóm một tập các đối tượng thực thể hay trừu tượng thành lớp các đối tượng tương tự nhau theo một hay nhiều tiêu chí nào đó. Các phương pháp phân cụm được phân thành các loại chính như sau: Phân cụm phân hoạch, phân cụm phân cấp, phân cụm dựa trên mật độ, phân cụm dựa trên lưới

Để xây dựng được chương trình ứng dụng thử nghiệm giải quyết bài toán tìm vị trí tối ưu lắp đặt các máy ATM trong thành phố Hải Phòng cần một vài thuật toán đã được sử dụng trong phân cụm dữ liệu không gian. Do đó nội dung của chương tiếp theo sẽ đề cập đến một số thuật toán liên quan.

CHƯƠNG 2. MỘT SỐ THUẬT TOÁN LIÊN QUAN

2.1 Thuật toán phân cụm dữ liệu không gian

2.1.1 Thuật toán K-means

Đây là thuật toán nổi tiếng và được sử dụng nhiều nhất trong hướng tiếp cận phân nhóm phân hoạch. Thuật toán này có nhiều biến thể khác nhau nhưng được đưa ra đầu tiên bởi J.B MacQueen vào năm 1967. Đầu vào của thuật toán này là một tập gồm n mẫu và một số nguyên K. Cần phân n đối tượng này thành K cluster sao cho sự giống nhau giữa các mẫu trong cùng cluster là cao hơn là giữa các đối tượng khác cluster.

Tư tưởng của thuật toán này như sau: Đầu tiên chọn ngẫu nhiên K mẫu, mỗi mẫu này coi như biểu diễn 1 cluster, như vậy lúc này trong mỗi cluster thì đối mẫu đó cũng là tâm của cluster (hay còn gọi là nhân). Các mẫu còn lại được gán vào một nhóm nào đó trong K nhóm đã có sao cho tổng khoảng cách từ nhóm mẫu đó đến tâm của nhóm là nhỏ nhất. Sau đó tính lại tâm cho các nhóm và lặp lại quá trình đó cho đến khi hàm tiêu chuẩn hội tụ. Hàm tiêu chuẩn hay được dùng nhất là hàm tiêu chuẩn sai-số vuông. Thuật toán này có thể áp dụng được đối với CSDL đa chiều, nhưng để dễ minh họa chúng ta mô tả thuật toán trên dữ liệu hai chiều.

Thuật toán k-means được mô tả cụ thể như sau:

Input: K, và dữ liệu về n mẫu của 1 CSDL.

Output: Một tập gồm K cluster sao cho cực tiểu về tổng sai-số vuông.

Thuật toán:

Bước 1: Chọn ngẫu nhiên K mẫu vào K cluster. Coi tâm của cluster chính là mẫu có trong cluster.

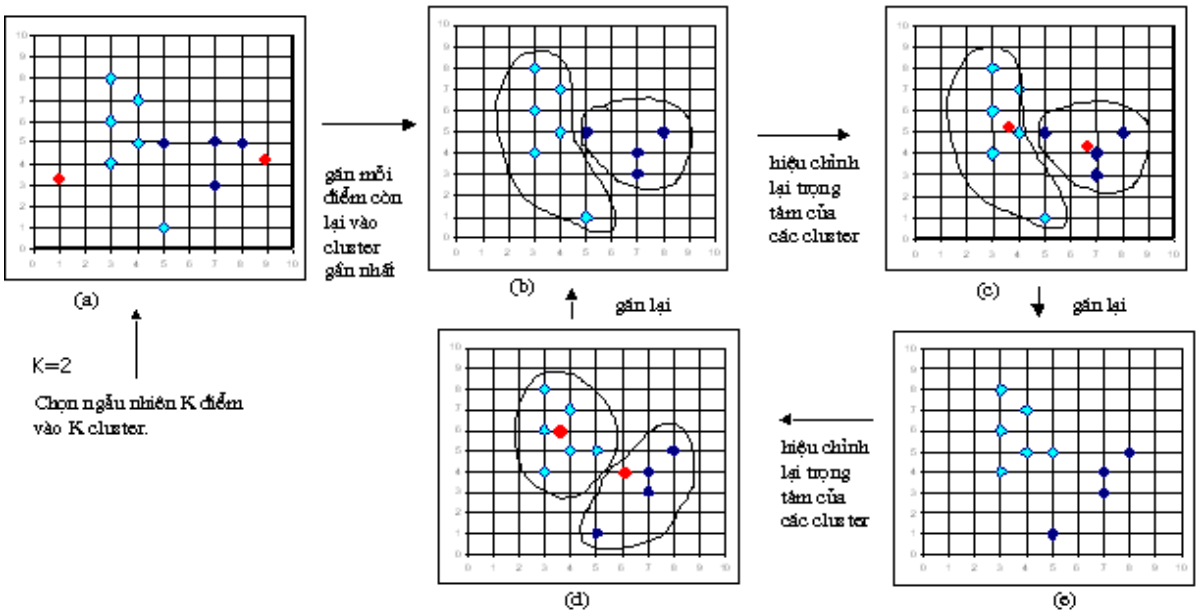
Bước 2: Tìm tâm mới của cluster.

Bước 3: Gán (gán lại) các mẫu vào từng cluster sao cho khoảng cách từ mẫu đó đến tâm của cluster đó là nhỏ nhất.

Bước 4: Nếu các cluster không có sự thay đổi nào sau khi thực hiện bước 3 thì chuyển sang bước 5, ngược lại sang bước 2.

Bước 5: Dừng thuật toán.

Ví dụ: Giả sử trong không gian hai chiều, cho 12 điểm (n = 12) cần phân 12 điểm này thành hai cluster (k=2).Đầu tiên chọn hai điểm ngẫu nhiên vào hai cluster, giả sử chọn điểm (1,3) và điểm (9,4) (điểm có màu đỏ trên hình 29.a). Coi điểm (1,3) là tâm của cluster 1 và điểm (9,4) là tâm của cluster hai. Tính toán khoảng cách từ các điểm khác đến hai điểm này và ta gán được các điểm còn lại này vào một trong hai cluster, những điểm có màu xanh lơ vào cluster 1, những điểm có màu xanh đậm vào cluster 2 (hình 2.1b). Hiệu chỉnh lại tâm của hai cluster, điểm màu đỏ trên hình 2.1c là tâm mới của hai cluster. Tính lại các khoảng cách các điểm đến tâm mới và gán lại các điểm này, hình 2.1d. Tiếp tục hiệu chỉnh lại tâm của hai cluster. Cứ như thế lặp lại cho đến khi không còn sự thay đổi nữa thì dừng. Khi đó ta thu được output của bài toán.



Hình 2.1: Minh họa thuật toán k-means

Dễ thấy độ phức tạp cuả thuật toán này là O(tKn). Trong đó n là số mẫu trong CSDL, K là số cluster, t là số lần lặp. Thông thường t, K<< n. Nên thuật toán này có hiệu quả tương đối với các CSDL lớn. Thuật toán này có ưu điểm là rõ ràng, dễ dàng cài đặt. Nhưng nhược điểm của thuật toán này là phải chỉ ra số lượng cụm và yêu cầu CSDL cần phân nhóm phải xác định được tâm. Thuật toán này cũng không phù hợp với việc khai phá các dữ liệu gồm các cluster có hình dạng không lồi

(non-convex). Có thể đưa thêm nhiều cải tiến vào k-mean để được thuật toán hiệu quả hơn, như thay đổi cách chọn các mẫu khởi đầu, cách tính tiêu chuẩn,...

Các thuật toán sau này như k-medoids, CLARANS,..đều là sự cải tiến của thuật toán k-means.

2.1.2. Thuật toán toán phân cụm dựa trên mật độ

Các thuật toán phân cụm dựa trên mật độ bao gồm: DBSCAN, DENCLUE, OPTICS…

DBSCAN là một phương pháp dựa trên mật độ điển hình, nó tăng trưởng các cụm theo một ngưỡng mật độ. DBRS là thuật toán thừa hưởng tư tưởng của DBSCAN nhưng cải tiến về tốc độ do sử dụng việc duyệt dữ liệu trên một số mẫu ngẫu nhiên chứ không duyệt toàn bộ cơ sở dữ liệu, do đó giảm được đáng kể số lần truy vấn không gian. Ngoài ra, DBRS còn mở rộng cho cả dữ liệu phi không gian. OPTICS là một phương pháp dựa trên mật độ, nó tính toán một thứ tự phân cụm tăng dần cho phép phân tích cụm tự động và tương tác.

Thuật toán DBSCAN (Density Based Spatial Clustering of Applications with Noise)

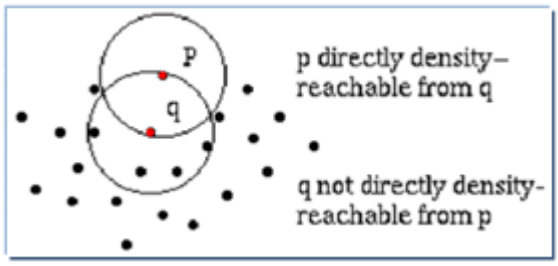
Thuật toán DBSCAN được Ester giới thiệu vào năm 1996, khi nghiên cứu các thuật toán phân cụm dữ liệu không gian. DBSCAN được khẳng định qua thực nghiệm là tốt hơn các thuật toán khác. Cụ thể so với thuật toán CLARANS thì DBSCAN phát hiện ra các cụm bất kì nhiều hơn và thực hiện tốt trên 100 tiêu chuẩn đánh giá hiệu quả thuật toán [3].

Các tác giả của thuật toán [3] đưa vào một số định nghĩa sau:

Định nghĩa 1: Lân cận Eps của đối tượng p, ký hiệu $N_{Eps(p)}$ là tập hợp các đối tượng q sao cho khoảng cách giữa p và q: dist(p,q) nhỏ hơn Eps.

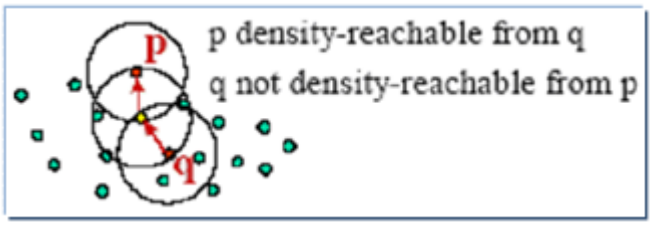
$N_{Eps(q)} = \{q \in D \mid dist(p,q) \leq Eps\}$

Định nghĩa 2: Đối tượng p là kề mật độ trực tiếp (directly density reachable) từ đối tượng q nếu p $\in N_{Eps(q)}$ và q là đối tượng lõi ($|N_{Eps(q)}| \geq$ Min Pts.



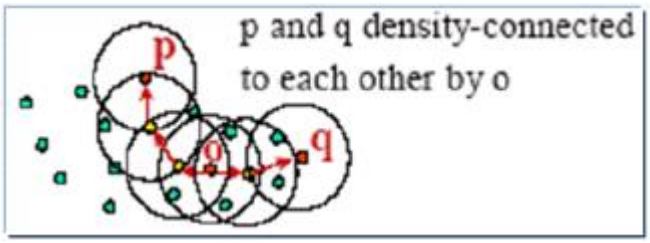
Hình 2.2: Kề mật độ trực tiếp

Định nghĩa 3: Đối tượng p là "kề mật độ" (density-reachable) từ đối tượng q nếu tồn tại một dãy p1, p2, ..., pn (p1 =q, pn= p) sao cho pi+1 là kề mật độ trực tiếp từ pi.



Hình 2.3: Kề mật độ

Định nghĩa 4: Đối tượng p là "kết nối theo mật độ" (density-connected) với đối tượng q nếu tồn tại đối tượng o sao cho cả p và q đều kề mật độ từ o.



Hình 2.4: Kết nối theo mật độ

Định nghĩa 5 (cụm): Cho cơ sở dữ liệu D, cụm C thỏa Eps và MinPts là tập con khác rỗng của D thỏa 2 điều kiện sau:

1. ∀p,q: nếu pЄC và q là kề mật độ từ p thì pЄC
2. ∀p,qЄC: nếu pЄC và p kết nối theo mật độ với q

Định nghĩa 6 (nhiễu): Cho các cụm C_1 ,. . ., C_k của cơ sở dữ liệu D với các tham số Eps_i và $MinPts_i$, (i = 1, . . ., k). Tập nhiễu là tập các đối tượng thuộc D nhưng không thuộc bất kỳ cụm Ci nào: Noise = {pЄD| ∀i:p $\notin C_i$ }

Trong phần này chúng ta mô tả giải thuật DBSCAN để phát hiện ra các cụm và nhiễu theo định nghĩa 5 và 6. Như đã biết, các thuật toán theo hướng mật độ đều có hai tham số là Eps và MinPts, việc xác định giá trị của hai tham số này có ảnh hưởng lớn đến các cụm được phát hiện. Thực tế cho thấy là không có cách nào xác định được giá trị của hai tham số này một cách chính xác, tất cả các thuật toán đều xác định dựa trên một hàm xấp xỉ. Lý tưởng nhất là xác định được cho mỗi cụm một cặp giá trị Eps và MinPts. Nhưng điều đó là khó thực hiện vì trước khi phân cụm thì chúng ta không thu nhận được thông tin về các cụm trong CSDL. DBSCAN đưa ra một heuristic hiệu quả và khá đơn giản để xác định các tham số Eps và MinPts của cụm "mỏng nhất" (thinnest) trong CSDL [3]. Vì thế DBSCAN sử dụng giá trị toàn cục Eps và MinPts đối với tất cả các cụm.

Ngoài việc sử dụng các định nghĩa từ 1 đến 6, DBSCAN còn có hai bổ đề để quan trọng để kiểm tra tính đúng đắn của thuật toán. Với các tham số đã Eps và MinPts cho trước chúng ta có thể phát hiện một cụm theo phương pháp mật độ theo hai bước sau: Thứ nhất chọn một điểm bất kì trong CSDL thỏa mãn điều kiện điểm nhân, coi điểm này là hạt giống (seed). Thứ hai tìm tất cả các điểm kề mật độ với điểm hạt giống trong cụm.

Bổ đề 1: Gọi p là một điểm trong D và $|N_{Eps}(p)| \geq$MinPts. Tập $O = \{o \mid o \in D$ và o là điểm kề mật độ với p$\}$ là một cụm.

Không chắc chắn rằng cụm C được xác định bởi duy nhất một điểm nhân nào của nó. Tuy nhiên, mỗi điểm trong C là điểm kề mật độ với một điểm nhân bất kì của C và theo đó C chính xác chứa các điểm mà kề mật độ với một điểm nhân bất kì của C.

Bổ đề 2: Cho C là một cụm và p là điểm bất kỳ trong C với $|N_{Eps}(p)| \geq$MinPts. Thì C bằng với tập $O = \{o \mid o$ là điểm kề mật độ với p$\}$.

Thuật toán

Để xác định các cụm, DBSCAN bắt đầu với một điểm p bất kỳ và tìm tất cả các điểm lân cận mật độ với p. Nếu p là điểm nhân, thủ tục này sẽ tạo ra một cụm

(xem bổ đề 2). Nếu p là điểm biên, thì không có điểm nào kề mật độ với p và DBSCAN thăm điểm kế tiếp trong CSDL.

Khi chúng ta sử dụng các giá trị toàn cục cho Eps và MinPts, DBSCAN có thể trộn 2 cụm theo định nghĩa 5 thành một cụm, nếu hai cụm này "gần" với nhau. Khoảng cách giữa hai tập điểm S_1 và S_2 được định nghĩa là $dist(S_1, S_2) = min \{dist(p,q) \mid p \in S1, q \in S2\}$. Hai tập điêm có các điểm tối thiểu của một cụm mỏng sẽ tách rời nhau chỉ nếu khoảng cách giữa hai tập lớn hơn Eps. Do đó, lời gọi đệ quy của DBSCAN có thể là cần thiết đối với các cụm được phát hiện với giá trị cao hơn MinPts. Tuy nhiên, đây không phải là một nhược điểm vì ứng dụng đệ quy của DBSCAN tạo ra một thuật toán cơ bản, khá hiệu quả.

Hơn nữa, phân cụm đệ quy các điểm của một cụm chỉ cần thiết trong trường hợp có thể được phát hiện dễ dàng.

Tài liệu [3] giới thiệu phiên bản cơ bản của DBSCAN bỏ qua các chi tiết về kiểu dữ liệu và việc tạo ra các thông tin bổ sung về các cụm như sau:

```
DBSCAN (SetOfPoints, Eps, MinPts)
    // SetOfPoints là tập chưa phân lớp.
    ClusterId := nextId(Noise);
    For i From 1 To SetOfPoints.size Do
        Point := SetOfPoints.get(i);
        IF Point.ClId = Unclassified Then
    IF ExpandCluster(SetOfPoints, Point,ClusterId, Eps, MinPts)
 Then
                    ClusterId := nextId(ClusterId)
        End If
     End If
   End For
End; // DBSCAN
```

Trong đó, **SetOfPoints** là toàn bộ CSDL hoặc cụm được phát hiện từ lần chạy trước. **Eps** và **MinPts** là các tham số mật độ toàn cục được xác định bằng tay hoặc

dựa trên heuristic. Hàm **SetOfPoints.get(i)** trả về phần tử thứ i của SetOfPoints. Hàm quan trọng nhất được sử dụng bởi DBSCAN là ExpandCluster được minh hoạ như sau:

```
ExpandCluster(SetOfPoints, Point, ClId, Eps,MinPts) : Boolean;
seeds:=SetOfPoints.regionQuery(Point,Eps);
If seeds.size<MinPts Then // không phải là điểm nhân
        SetOfPoint.changeClId(Point,Noise);
        Return False;
Else // tất cả các điểm trong seeds là kề mật độ với Point
        SetOfPoints.changeClIds(seeds,ClId);
        seeds.delete(Point);
        While seeds <> Empty Do
                currentP := seeds.first();
                result := SetOfPoints.regionQuery(currentP,Eps);
                If result.size >= MinPts Then
                        For i From 1 To result.size Do
                                resultP := result.get(i);
                                If resultP.ClId
                                        In {Unclassified, Noise} Then
                                                IF resultP.ClId = Unclassified THEN
                                                seeds.append(resultP);
                                        END IF;
                                SetOfPoints.changeClId(resultP,ClId);
                                End If; // Unclassified hoặc Noise
                        End For;
                End If; // result.size >= MinPts
                seeds.delete(currentP);
        End While; // seeds <> Empty
        Return True;
```

End If

End; // ExpandCluster

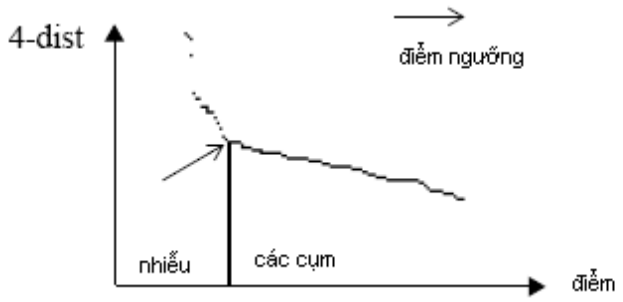
Lời gọi hàm **SetOfPoints.regionQuery(Point,Eps)** trả về Eps-Neighborhood của **Point** trong **SetOfPoints** như một danh sách các điểm. Các truy vấn vùng có thể được hỗ trợ một cách hiệu quả bởi các phương thức truy nhập không gian như cây R* [11] được cho trong hệ thống CSDL không gian (SDBS) để xử lý một cách hiệu quả cho một số kiểu truy vấn không gian.Chiều cao của cây R* là O(logn) với CSDL có n điểm trong trường hợp tồi nhất và một truy vấn với vùng truy vấn nhỏ phải duyệt trên một số giới hạn các đường đi trên cây R*. Khi lân cận Epsilon được trông đợi là nhỏ khi so sánh với kích thước của toàn bộ không gian dữ liệu, độ phức tạp thời gian tính trung bình của một truy vấn vùng đơn giản là O(log n). Với mỗi n điểm của CSDL, chúng ta có nhiều nhất một truy cấn vùng. Do đó, độ phức về thời gian tính trung bình của DBSCAN là **O(n * logn)**. ClId (ID của cụm) của các điểm mà được đánh dấu là NOISE có thể bị thay đổi, nếu chúng là điểm tới được từ một số điểm khác trong CSDL. Điều này xảy ra đối với một số điểm biên của cụm. Các điểm này không được bổ sung vào seeds-list vì chúng ta biết rằng một điểm với ClId của NOISE thì không phải là điểm hạt nhân. Bổ sung các điểm này vào seeds sẽ chỉ có kết quả trong các truy vấn vùng bổ sung mà không mang lại câu trả lời mới. Khi hai cụm C_1 và C_2 gần nhau, có thể xảy ra trường hợp có một số điểm p thuộc về cả hai cụm C_1 và C_2. Khi đó, p phải là điểm biên của cả hai cụm vì nếu không C_1 và C_2 sẽ bằng nhau khi chúng ta sử dụng các tham số toàn cục. Trong trường hợp này, điểm p sẽ được gán với cụm được phát hiện ra trước. Trừ các trường hợp hiếm khi xảy ra như trên, kết quả của DBSCAN độc lập với thứ tự mà trong đó các điểm của CSDL đã được thăm theo bổ đề 2.

Xác định tham số mật độ Eps và MinPts

Trong tài liệu[3] đã xây dựng một heuristic đơn giản nhưng hiệu quả để xác định các tham số Eps và MinPts của cụm "mỏng nhất" trong CSDL. Heuristic này được dựa trên các quan sát sau: Gọi d là khoảng cách từ điểm p tới k láng giềng gần nhất, thì d láng giềng của p chứa đúng k+1 điểm đối với hầu hết các điểm p. d láng

giềng của p chứa nhiều hơn k+1 điểm chỉ khi một số điểm có cùng khoảng cách d tới điểm p. Hơn nữa, thay đổi hệ số k đối với một điểm trong cụm không tạo ra một sự thay đổi lớn nào đối với d. Điều này chỉ xảy ra khi k láng giềng gần nhất của p với k=1, 2, 3… được xác định xấp xỉ trên một đường thẳng.

Với k cho trước, định nghĩa một hàm k-dist từ CSDL D tới các số thực, ánh xạ mỗi điểm thông qua hàm khoảng cách tới k láng giềng gần nhất của nó. Khi các điểm trong CSDL được sắp xếp theo thứ tự giảm dần giá trị k-dist của nó, đồ thị của hàm này đưa ra một số dấu hiệu có liên quan đến phân bố mật độ trong CSDL. Gọi đồ thị này là đồ thị k-dist đã sắp xếp. Nếu chọn một điểm p bất kỳ, gán tham số Eps với k-dist(p) và gán tham số MinPts với k, tất cả các điểm bằng hoặc nhỏ hơn giá trị k-dist sẽ là những điểm hạt nhân. Nếu chúng ta có thể xác định điểm ngưỡng với giá trị k-dist lớn nhất trong cụm "mỏng nhất" của D, thì chúng ta sẽ có các giá trị tham số mật độ. Điểm ngưỡng (threshold point) là điểm đầu tiên trong vùng đầu tiên của đồ thị k-dist đã được sắp xếp. Tất cả các điểm có giá trị cao hơn k-dist(ở bên trái điểm ngưỡng) có thể coi là nhiễu, tất cả các điểm còn lại (ở bên phải điểm ngưỡng) được gán cho một số cụm nào đó.



Hình 2.5: Đồ thị đã sắp xếp 4-dist đối với CSDL mẫu 3

Nói chung, việc xác định ra vùng đầu tiên một cách tự động khá khó khăn, nhưng lại rất dễ dàng đối với người sử dụng tự xác định thấy vùng này trên đồ thị. Vì vậy, nhóm tác giả trên đề xuất một phương pháp tương tác để xác định điểm ngưỡng như sau:

DBSCAN cần có hai tham số, Eps và MinPts. Tuy nhiên, các thí nghiệm đã chỉ ra rằng đồ thị k-dist với k>4 không có sai khác gì nhiều so với đồ thị 4-dist, hơn nữa, lại phải tính toán nhiều hơn. Vì vậy, gán tham số MinPts bằng 4 với tất cả

CSDL (dữ liệu 2 chiều). Nhóm tác giả trên đã đề xuất phương pháp tương tác sau để xác định tham số Eps của DBSCAN:

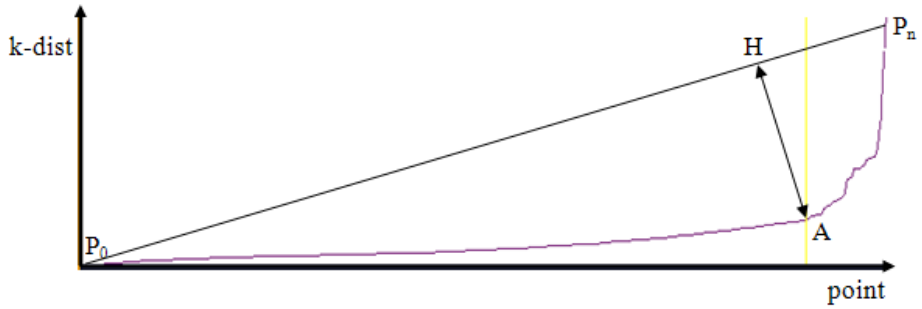
* Hệ thống tính toán và hiển thị đồ thị 4-disp cho CSDL.

* Nếu người sử dụng có thể đánh giá tỷ lệ phần trăm của nhiễu, thì tỷ lệ này được đưa vào và hệ thống lấy đánh giá này để xác định điểm ngưỡng.

* Người sử dụng chấp nhận điểm ngưỡng này hoặc tự lựa chọn điểm ngưỡng khác. Giá trị 4-dist của điểm ngưỡng được sử dụng như giá trị Eps trong DBSCAN.

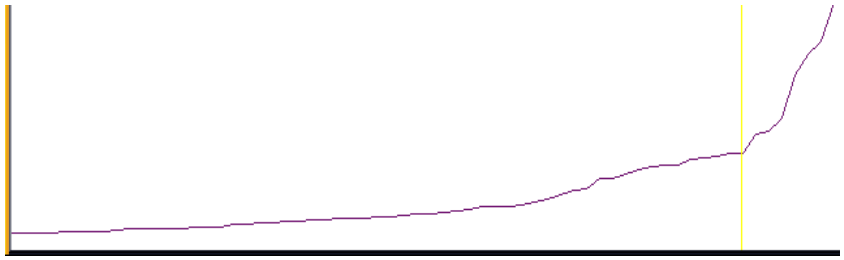
Học viên đã đề xuất một phương pháp đơn giản nhưng khá hiệu quả (theo khảo sát thực nghiệm) để xác định tham số Eps một cách tự động như sau:

Thay vì yêu cầu người dùng lựa chọn điểm ngưỡng A, chương trình tự động tính toán khoảng cách lớn nhất AH tới đường thẳng đi qua giá trị đầu và cuối P_0 và P_n trong đồ thị khoảng cách k-dist như hình vẽ, giá trị k-dist tại điểm A sẽ là giá trị epsilon cần ước lượng.

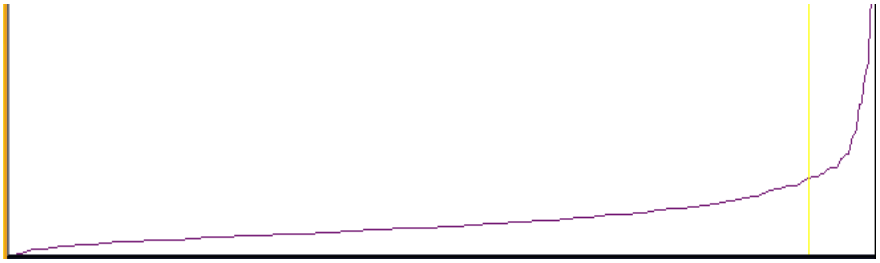


Hình 2.6: Đồ thị k-dist và một phương pháp ước lượng tham số Eps

Thử nghiệm cho thấy kết quả ước lượng tự động này khá hiệu quả đối với hầu hết các tập dữ liệu.



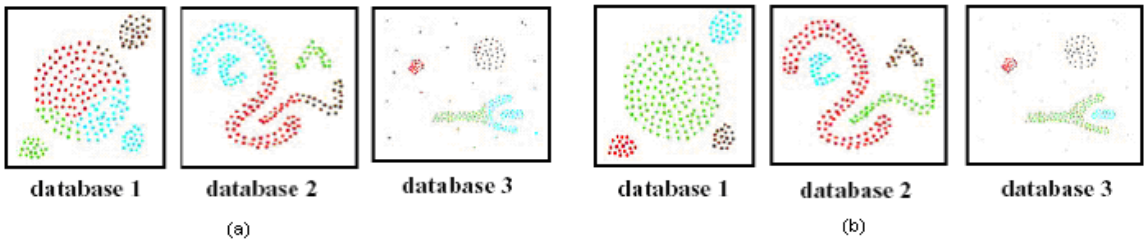
Hình 2.7: Đồ thị K-dist của lớp bản đồ "Hệ thống siêu thị"



Hình 2.8: Đồ thị K-dist của lớp bản đồ "Ngân hàng"

Hiệu quả của thuật toán

Để đánh giá hiệu quả của DBSCAN, [3] đã so sánh với thuật toán khá nổi tiếng là CLARANS. Kết quả thực nghiệm cho thấy thời gian chạy của DBSCAN cao hơn tổng số lượng các điểm. Thời gian chạy của CLARANS, xấp xỉ với bình phương tổng số điểm. Kết quả cũng chỉ ra rằng DBSCAN tốt hơn CLARANS gấp 250 đến 1900 lần, hệ số này sẽ tăng lên khi kích thước của CSDL lớn dần lên. Với cùng CSDL mẫu trong hình vẽ dưới thì CLARANS và DBSCAN cho kết quả lần lượt như minh họa ở hình dưới(các điểm được phát hiện cùng một cụm được minh họa bằng màu giống nhau).

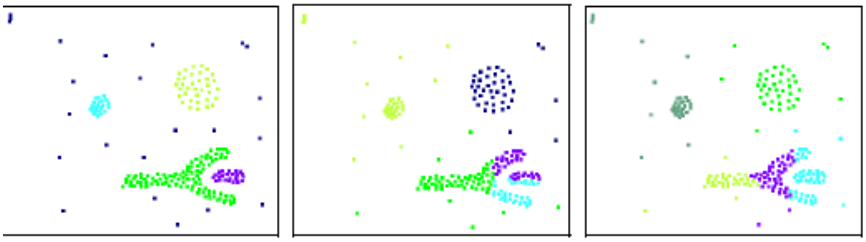


Hình 2.9: Các cụm phát hiện được bởi CLARANS (a) và DBSCAN (b)

Như vậy, DBSCAN hiệu quả hơn trong việc phát hiện ra các cụm với hình dạng bất kì, kể cả hình không lồi. Và khử nhiễu tốt hơn CLARANS. Tuy nhiên vẫn có một số trường hợp mà DBSCAN không phù hợp:

* Nếu các cụm có mật độ khác nhau nhiều thì DBSCAN sẽ không giữ được tính hiệu quả. Trên những dữ liệu như thế ta phải áp dụng mật độ của cụm có mật độ thấp nhất cho tất cả các cụm khác. Với các cụm có mật độ rất cao thì DBSCAN tốn nhiều thời gian để xác định lân cận của các điểm một cách không cần thiết.

* Nếu có quan tâm đến các thuộc tính phi không gian (non-spatial) thì sử dụng DBSCAN không thích hợp vì DBSCAN không chú ý đến các thuộc tính đó.



(b) DBSCAN (c) k-means (d) CLARANS

Hình 2.10: Các cụm được phát hiện bởi DBSCAN(b), K-Means(c), CLARANS(d)

2.2 Thuật toán xếp chồng bản đồ.

2.2.1. Khái quát về xếp chồng bản đồ

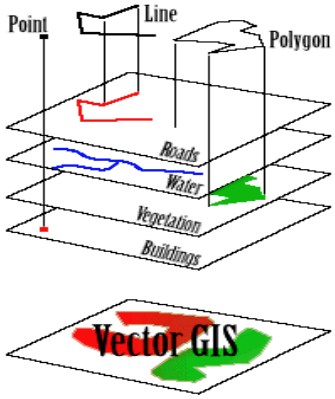
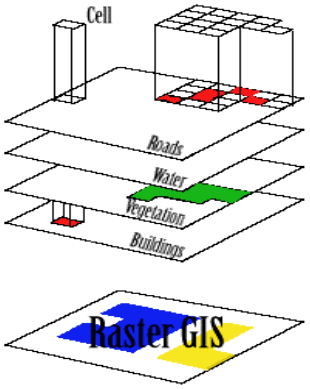
Bản đồ là tập hợp các lớp layer khác nhau.

Layer: Chứa các điểm tiện ích,

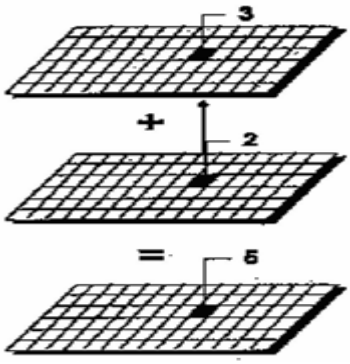
Chứa các đường thủy hệ, giao thông.

Chứa đa giác (vùng biên hành chính)

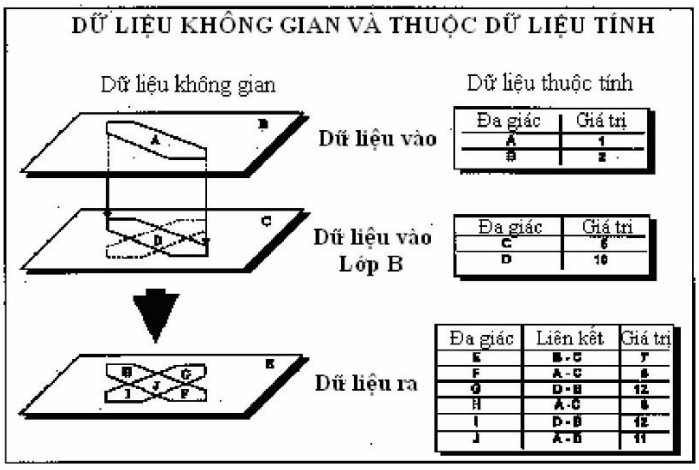
Trong tài liệu [10], xếp chồng các bản đồ trong kỹ thuật GIS là một khả năng ưu việt của GIS trong việc phân tích các số liệu thuộc về không gian, để có thể xây dựng thành một bản đồ mới mang các đặc tính hoàn toàn khác với bản đồ trước đây.

Dạng Vector	Dạng Raster	Xếp chồng
		**A** **B** **C** **D** **U=f(A,B,C,D..)**

Hình 2.11 Nguyên lý khi xếp chồng các bản đồ

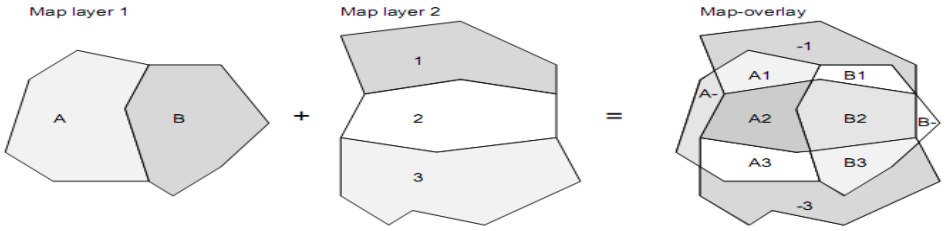
Chồng lắp cơ bản dữ liệu	
	Dữ liệu vào (lớp A) Dữ liệu vào (lớp B) Dữ liệu Ra

Hình 2.12: Việc xếp chồng các bản đồ theo phương pháp cộng



Hình 2.13: Một thí dụ trong việc xếp chồng các bản đồ.

Xếp chồng bản đồ có thể được định nghĩa là một hoạt động không gian, kết hợp các lớp địa lý khác nhau để tạo ra lớp thông tin mới. Xếp chồng bản đồ được thực hiện bằng cách sử dụng số học, logic, các toán tử quan hệ và được thực hiện trong cả hai loại dữ liệu Vector và Raster.



Hình 2.14 Xếp chồng 2 lớp bản đồ

Quá trình thực hiện Overlay bản đồ qua 2 bước:

1. Xác định tọa độ các giao điểm và tiến hành chồng khít 2 lớp bản đồ tại giao điểm này.

2. Kết hợp dữ liệu không gian và thuộc tính của hai lớp bản đồ.

2.2.2. Các phương pháp trong xếp chồng bản đồ

2.2.2.1. Phương pháp Raster Overlay

Phương pháp Raster Overlay sử dụng số học và các toán tử Boolean để kết hợp các điểm ảnh hoặc giá trị tế bào trong mỗi bản đồ tạo ra một giá trị mới trong bản

đồ kết hợp. Các bản đồ có thể được coi là các biến số học và thực hiện các chức năng đại số phức tạp.

Có nhiều phương pháp xếp chồng khác nhau thực hiện trên những vector địa lý. Phương pháp raster overlay dựa trên ý tưởng bản đồ đại số. Sử dụng bản đồ đại số dữ liệu đầu vào có thể được cộng, trừ, nhân, chia để tạo dữ liệu ra. Hoạt động của thuật toán là thực hiện trên giá trị của các ô tương ứng của hai hoặc nhiều tầng dữ liệu vào để cho ra một giá trị mới.

Bản đồ đại số chức năng sử dụng các biểu thức tóan học để tạo ra các lớp raster mới bằng cách so sánh chúng.

7	5	5	3		5	1	2	1		2	4	3	2
1	1	1	1		1	1	1	1		0	0	0	0
1	1	1	1	-	1	1	1	2	=	0	0	0	-1
1	1	1	2		1	2	2	2		0	-1	-1	0

(A) **(B)** **(C)**

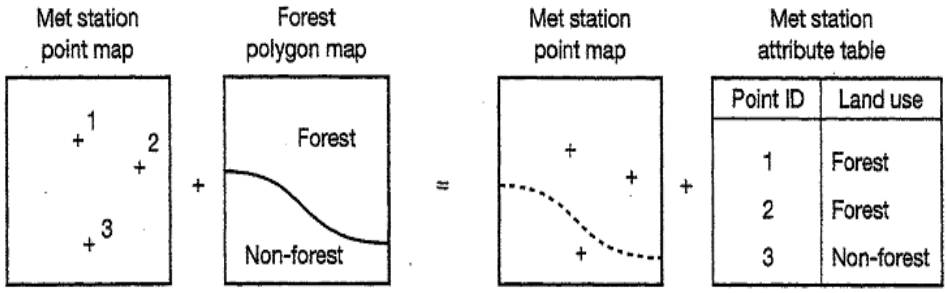
Hình 2.15 Minh họa Raster Overlay

Trong đề tài luận văn này phương pháp được đề cập đến là phương pháp Vector Overlay.

2.2.2.2. Phương pháp Vector Overlay

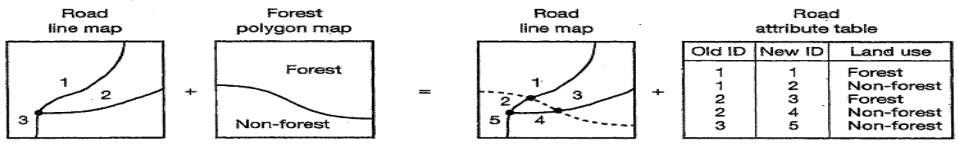
Trong Vector Overlay, các tính năng và thuộc tính của bản đồ được tích hợp để cho ra một bản đồ mới. Vector overlay có thể được thực hiện trên các kiểu chức năng của bản đồ như: Điểm và đường (Point in Line), đoạn và đa giác (Line in Polygon), đa giác và đa giác (Polygon in Polygon). Các phép xếp chồng bản đồ trên dữ liệu Vector được chia thành 3 loại. Dưới đây là 3 ví dụ minh họa cho 3 phép xếp chồng bản đồ trên dữ liệu vector.

- Điểm và đa giác: xếp chồng hai lớp điểm và đa giác để tạo ra lớp điểm mới.



Hình 2.16. Xếp chồng điểm và đa giác

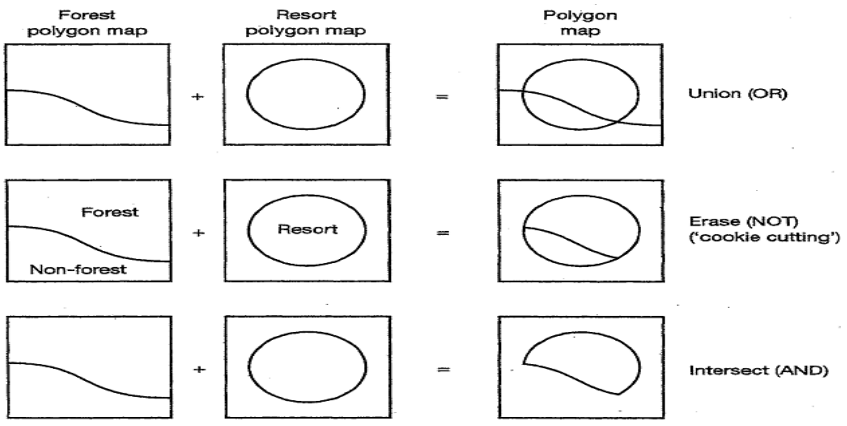
- ***Đoạn và đa giác***: Chồng khít lớp đường và đa giác để tạo ra lớp đường mới.



Hình 2.17. Xếp chồng đoạn và đa giác

- ***Đa giác và đa giác***: Chồng khít đa giác và đa giác để tạo ra lớp đa giác mới.

Khi chồng khít 2 lớp đa giác có thể có 3 trường hợp xảy ra.



Hình 2.18. Xếp chồng đa giác và đa giác

2.2.3. Một số phép toán cơ bản trong Overlay

Các phép toán trong overlay bao gồm: Phép hợp (Union), phép giao (Intersect) và phép đồng nhất (Indentity).

2.2.3.1. Phép hợp (Union)

Phép hợp hoạt động như toán tử Or.

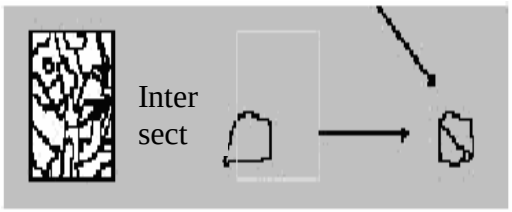
Đầu vào là hai lớp bản đồ là kiểu đa giác (polygon), đầu ra là một lớp bản đồ mới bằng cách xếp chồng hai miền dữ liệu đầu vào và dữ liệu thuộc tính của chúng.

Điều kiện: miền dữ liệu phải là polygon.

Hình 2.19. Phép hợp trong Overlay

2.2.3.2. Phép giao (Intersect)

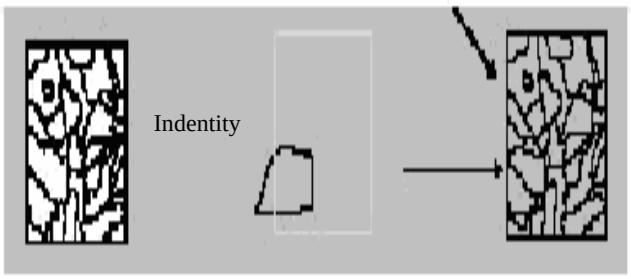
Phép giao hoạt động như toán tử And. Tạo ra một vùng bao phủ mới bằng cách xếp chồng hai tập dữ liệu đầu. Kết quả đầu ra bao gồm phần dữ liệu thuộc vào cả hai tập dữ liệu đầu vào.



Hình 2.20. Phép giao trong Overlay

2.2.3.3. Phép đồng nhất (Indentity)

Tạo ra một vùng bao phủ mới bằng cách xếp chồng hai tập dữ liệu đầu vào. Kết quả đầu ra bao gồm toàn bộ phần dữ liệu của lớp đầu tiên và chỉ những phần nào của lớp thứ hai được chồng khít.



Hình 2.21. Phép đồng nhất trong Overlay

2.2.4. Một số thuật toán cơ bản xếp chồng bản đồ

2.2.4.1. Thuật toán giao hai đoạn thẳng (Bentley – Ottmann)

Thuật toán BO đã được trình bày tại tài liệu [9], trong hình học tính toán thuật toán Bentley – Ottmann (BO) là một thuật toán quét dòng để liệt kê tất cả các đoạn thẳng giao nhau trong mặt phẳng. Tương tự như các thuật toán khác để kiểm tra có hay không các đoạn thẳng giao nhau, với đầu vào là n đoạn thẳng và k điểm cắt nhau BO có độ phức tạp là O(n+k)logn.

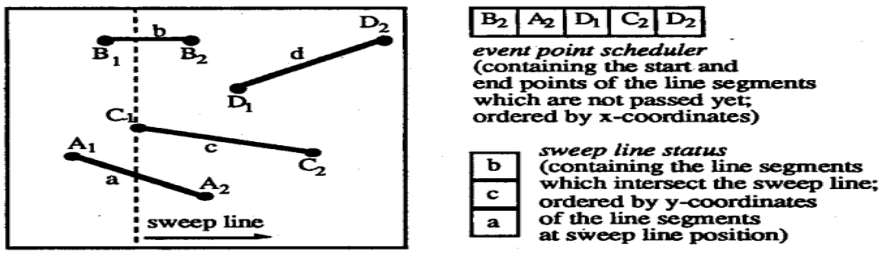
Thuật toán này được phát triển ban đầu bởi Jon Bentleyand Thomas Ottmann (1979) [9]. Mặc dù không phải là thuật toán tốt nhất nhưng nó được lựa chọn để thực hành bởi sự đơn giản và chiếm ít bộ nhớ.

Giả thiết đầu vào của thuật toán này như sau:

- Không có đoạn thẳng nào thẳng đứng.
- Các điểm mút của đoạn thẳng này không nằm trên đoạn khác.
- Điểm giao nhau chỉ là điểm giao của 2 đoạn thẳng.
- Không có quá hai điểm mút và điểm cắt nhau có cùng độ độ x.

2.2.4.1.1. Ý tưởng của thuật toán

Ý tưởng chính của thuật toán BO là sử dụng tiếp cận quét dòng, trong đó một dòng thẳng đứng L chuyển động từ trái sang phải trong mặt phẳng, những đoạn giao nhau sẽ được lưu lại trên đường mà nó di chuyển. Điều đó thật đơn giản để mô tả thuật toán trong trường hợp nhập vào ở vị trí nói chung, tức là không có đoạn thẳng kết thúc hoặc điểm giao nhau trên cùng trục x, không có điểm cuối đoạn thẳng trên phân khúc khác và không có 3 đoạn thẳng giao nhau tại 1 điểm. Trong trường hợp này L sẽ luôn luôn giao nhau những đoạn đường vào trong tập hợp các điểm và chỉ thay đổi theo chiều dọc tại một tập hữu hạn các sự kiện rời rạc. Do đó, chuyển động liên tục của L có thể được chia thành mỗi chuỗi hữu hạn các bước và mô phỏng bằng một thuật toán chạy trong một khoảng thời gian hữu hạn.



Hình 2.22. Minh hoạ thuật toán quét dòng

Có hai sự kiện có thể xảy ra trong quá trình mô phỏng này, khi L quét qua một điểm cuối của một đoạn thẳng s, giao điểm của L và s sẽ được thêm vào hoặc gỡ ra từ một tập có thứ tự các điểm giao nhau. Sự kiện này dễ dàng được dự đoán như các điểm đầu mút của đoạn thẳng (đã biết từ đầu vào của thuật toán). Sự kiện còn lại xảy ra khi L quét qua chỗ cắt nhau của 2 đoạn thẳng s và t, sự kiện này cũng được dự đoán trước từ thực tế, ngay từ khi xảy ra sự kiện này, các điểm giao nhau của L với s và t được đặt liền kề trong tập các điểm giao nhau có thứ tự.

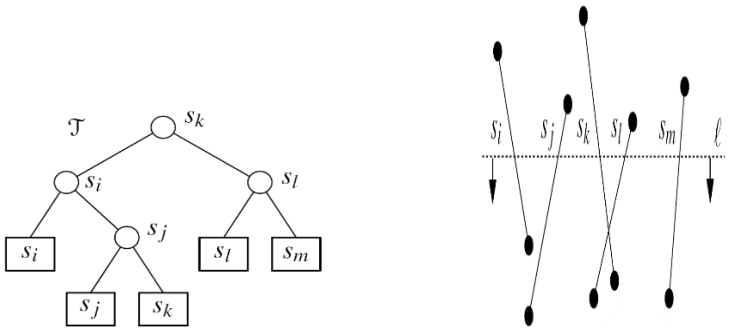
Thuật toán BO sử dụng cấu trúc dữ liệu biểu diễn tập các điểm giao nhau thẳng đứng của dòng quét với các đoạn thẳng đầu vào, và tập hợp các sự kiện có tiềm năng trong tương lai hình thành bởi các cặp liền kề của các điểm giao nhau. Nó xử lý các sự kiện lần lượt cập nhật cấu trúc dữ liệu để biểu diễn tập các điểm giao nhau mới.

2.2.4.1.2. Cấu trúc dữ liệu

Để lưu trữ hiệu quả các giao điểm của đường quét L với các đoạn đường vào và một chuỗi các sự kiện trong tương lai thuật toán BO lưu trữ cấu trúc dữ liệu là:

Một cây tìm kiếm nhị phân chứa tập các đoạn thẳng qua L, theo thứ tự trục Y của các điểm mà các đoạn thẳng qua L. Những điểm cắt không được đại diện một cách rõ ràng trong cây nhị phân tìm kiếm. Thuật toán BO sẽ chèn thêm một đoạn mới s vào cấu trúc dữ liệu khi dòng quét L đi qua điểm cuối P của đoạn này, vị trí chính xác của s trong cây tìm kiếm có thể xác định bởi tìm kiếm nhị phân, mỗi bước kiểm tra p là trên hay dưới các đoạn khác mà L đi qua. Do đó việc chèn sẽ được thực hiện trong thời gian logarits.. Thuật toán BO sẽ xoá các đoạn từ cây nhị phân và sử dụng tìm kiếm nhị phân để xác định đoạn ngay ở dưới hoặc trên các đoạn

khác, các thao tác này có thể được thực hiện bởi cây tự cấu trúc mà không quan tâm đến hình học cơ bản của đoạn thẳng.



Hình 2.23. Cấu trúc cây nhị phân

Thứ tự sắp xếp: r và s là 2 đoạn thẳng thì r < s nếu và chỉ nếu

- $r_{x1} < s_{x1}$ hoặc
- $r_{x1} = s_{x1}$ và $r_{x2} < s_{x2}$.

- Một hàng đợi ưu tiên để duy trì một chuỗi các sự kiện có tiềm năng trong tương lai của thuật toán BO. Mỗi sự kiện được liên kết với một điểm p trong mặt phẳng,, điểm đầu cuối, các điểm cắt. Sự kiện này xảy ra khi dòng L cắt qua p. Do đó các sự kiện có thể được đánh số ưu tiên theo trục x của điểm liên kết các sự kiện. Trong thuật toán BO các sự kiện tương lai tiềm năng bao gồm các đầu mút đoạn thẳng mà chưa được quét qua và các điểm giao nhau của các phân đoạn của đường mà ở ngay bên dưới hoặc bên trên đoạn khác.

Thuật toán không cần phải lưu trữ một cách rõ ràng đại diện của dòng quét L hoặc vị trí của nó trong mặt phẳng. Thay vào đó, vị trí của L được thể hiện gián tiếp: đó là đường thẳng đứng qua điểm gắn với sự kiện gần đây nhất là xử lý.

2.2.4.1.3. Chi tiết thuật toán BO

Thuật toán BO thực hiện qua những bước sau đây:

1. Khởi tạo một hàng đợi ưu tiên Q các sự kiện có tiềm năng trong tương lai, mỗi liên kiết với một điểm trong mặt phăng ưu tiên theo trục x. Ban đầu, Q chứa danh sách các điểm đầu mút của các đoạn thẳng.

2. Khởi tạo một tìm kiếm nhị phân T của các phân đoạn thẳng qua L quét đường, theo thứ tự trên trục y của các điểm giao nhau. Ban đầu, T rỗng.

3. Trong khi Q là không rỗng, tìm và loại bỏ các sự kiện từ Q liên kết với một điểm p có toạ độ x thấp nhất. Xác định loại sự kiện này là sự kiện gì và quá trình đó theo các trường hợp sau đây:

* Nếu p là điểm cuối bên trái của đoạn s, chèn s vào T. Tìm các đoạn r và t ở bên dưới hay trên s trong T (nếu có) nếu chúng giao nhau bởi một điểm tiềm năng trong hàng đợi các sự kiện thì loại bỏ nó. Nếu s giao r hoặc t thì thêm điểm cắt này vào hàng đợi.

* Nếu p là điểm cuối bên phải của một đoạn s, loại bỏ s từ T. Tìm đoạn r và t ngay và dưới s trong T (trước khi loại bỏ s) . Nếu r và t giao nhau thì thêm điểm này vào trong danh sách hàng đợi.

* Nếu p là điểm giao nhau của s và t (với s ở dưới t theo hướng từ trái qua) , hoán đổi vị trí của s và t trong T. Tìm các đoạn r và u (nếu có) bên dưới và trên s t tương ứng. Huỷ bỏ các điểm cắt rs và tu từ hàng đợi sự kiện, và, nếu r cắt t hay s cắt u, thêm điểm này vào hàng đợi các sự kiện.

2.2.4.1.4. Phân tích thuật toán

Thuật toán xử lý sự kiện mỗi điểm mút của đoạn hoặc điểm gia nhau. Được sắp xếp theo thứ tự của toạ độ theo trục x. Khi một sự kiện thứ i được xử lý, sự kiện tiếp theo (nếu nó là điểm cắt) phải là một điểm giao nhau của 2 đoạn thẳng liền kề biểu diễn trong T, bởi vì thuật toán lưu trữ tất cả các điểm giao nhau của các đoạn thẳng như là các sự kiện tiềm năng trong tương lai, và do vậy sự kiện tiếp theo luôn luôn có mặt trong hàng đợi sự kiện. Kết quả là tìm được chính xác các điểm cắt của các đoạn thẳng.

Thuật toán BO xử lý chuỗi 2n+k sự kiện, trong đó n là số đoạn thẳng và k là số điểm cắt, mỗi sự kiện được xử lý bởi một số cố định phép toán trong cây tìm kiếm nhị phân và hàng đợi sự kiện, và bởi vì nó chỉ chứa các điểm mút và điểm cắt giữa 2 đoạn liền kề nên hàng đơi sự kiện chứa không quá 3n sự kiện. Do đó tất cả các thao tác mất thời gian là O (logn) và tổng thời gian của thuật toán là O((n+k)logn).

Nếu điểm cắt tìm thấy bởi thuật toán không cần phải lưu trữ một khi tìm thấy thì không gian lưu trữ của thuật toán là O(n).

2.2.4.1.5. Kết luận thuật toán

Kỹ thuật xếp chồng bản đồ là kỹ thuật rất khó trong quá trình phân tích thông tin, đòi hỏi phải có những giải pháp tối ưu về thời gian và không gian lưu trữ. Trong phần trên học viên đã trình bày thuật toán quét dòng để xác định sự giao nhau của hai đoạn thẳng. Tuy thuật toán đơn giản nhưng nó được sử dụng nhiều trong quá trình thực hành vì đơn giản và chiếm ít bộ nhớ.

2.2.4.2. Thuật toán giao của hai đa giác

Đã có nhiều thuật toán tìm giao của hai đa giác được công bố. Phần lớn các thuật toán này xuất phát từ tìm giao của hai đa giác lồi. Do vậy, nếu vùng nghiên cứu là đa giác bất kỳ thì chúng phải được tách ra thành các đa giác lồi trước khi thực hiện thuật toán. Trong luận văn học viên đề xuất thuật toán tìm trực tiếp giao hai đa giác, có thể áp dụng trong hệ GIS véc tơ.

Phát biểu bài toán: Hãy tìm phần giao của hai đa giác phẳng không tự cắt A và B. Cho biết $A=a_1a_2...\ a_n$ và $B=b_1b_2...\ b_m$.

2.2.4.2.1. Chi tiết thuật toán

Dễ thấy rằng: phần giao của A và B có thể là tập rỗng hay là tập các đa giác không giao nhau. Để đơn giản ta gọi phần giao của A và B là tập đa giác giao, và gọi một cạnh là cạnh của tập đa giác giao với ý nghĩa nó là cạnh của một đa giác trong tập đa giác giao.

Với P là một đa giác thì ta gọi I(P) và O(P) lần lượt là miền trong và miền ngoài của P.

Tư tưởng của thuật toán là tìm tất cả các cạnh của tập đa giác giao, nếu tập cạnh này khác rỗng thì bằng cách ghép chúng lại sẽ được tập các đa giác là giao của A và B.

Thuật toán bao gồm hai bước chính như sau:

Bước 1: Trường hợp hai đa giác không có cặp cạnh nào song song và giao nhau

Với mỗi cạnh $v=a_ia_{i+1} \in A$ (i=1,2,..,n), ta tìm mọi giao điểm của v với tất cả các cạnh $u= b_kb_{k+1} \in B$ (k=1,2,..,m), trong đó a_{n+1} và b_{m+1} tương ứng được gán là a_1 và b_1.

Đặt X_v= {x| x là giao điểm của cạnh v với các cạnh $u \in B\} \cup \{a_i,a_{i+1}\}$(nếu trong X_v có nhiều điểm trùng nhau thì chỉ giữ lại một điểm trong số các điểm trùng nhau đó).

Sắp xếp các điểm trong X_v theo chiều tăng dần về khoảng cách từ mỗi điểm đến a_i, ta được X_v $=\{x_1=a_i,x_2,..,x_{lv-1},x_{lv}=a_{i+1}\}$, với $|X_v|$=lv. Khi đó, *cạnh* x_ix_{i+1} *(i=1,2,..,lv-1) là một cạnh của tập đa giác giao nếu trung điểm của nó thuộc I(B).*

Xử lý tương tự cho các cạnh của đa giác B.

Bước 2: Trường hợp hai đa giác có cặp cạnh song song và giao nhau

Trước hết, ta chèn thêm những điểm mới vào các đa giác để nếu có trường hợp tồn tại cặp cạnh song song và giao nhau thì tạo ra cặp cạnh trùng nhau.

Giả sử cạnh a_ia_{i+1} và cạnh b_kb_{k+1} song song và giao nhau (nhưng không trùng nhau). Ta xử lý như sau:

-Nếu $a_i \in b_kb_{k+1}$(a_i nằm trong đoạn b_kb_{k+1}), thì chèn a_i vào giữa b_k và b_{k+1}, tức là coi b_ka_i và a_ib_{k+1} là hai cạnh mới của đa giác B.

- Xử lý tương tự cho ba đỉnh: a_{i+1}, b_k và b_{k+1}.

Sau đó, xét mỗi cặp cạnh trùng nhau $v=a_ia_{i+1} \in A$ và $u=b_kb_{k+1} \in B$ (giả sử $a_i \equiv b_k$ và $a_{i+1} \equiv b_{k+1}$), thực hiện thao tác tìm giao điểm và sắp xếp như bước 1 ở trên với hai cạnh $a_{i+1}a_{i+2}$ và $b_{k+1}b_{k+2}$ ta được hai tập hợp:

$$X_v=\{x_1=a_{i+1},x_2,..,x_{lv-1},x_{lv}=a_{i+2}\},$$

$$Y_u=\{y_1=b_{k+1},y_2,..,y_{lu-1},y_{lu}=b_{k+2}\}.$$

Để kiểm tra xem cạnh a_ia_{i+1} (hoặc b_kb_{k+1}) có là một cạnh của tập đa giác giao hay không, ta dựa vào tính chất sau:

Gọi N và M lần lượt là trung điểm các cạnh x_1x_2 và y_1y_2. Khi đó, cạnh a_ia_{i+1} (hoặc b_kb_{k+1}) là một cạnh của tập đa giác giao nếu một trong hai điều kiện sau thoả mãn:

1. $N \in I(B)$ và $M \in O(A)$.

2. $N \in O(B)$ và $M \in I(A)$.

Các thuật toán liên quan

Thuật toán trình bày trên có sử dụng hai thuật toán khác để cài đặt, đó là kiểm tra điểm trong đa giác và tìm giao của hai đoạn thẳng. Để kiểm tra một điểm có nằm trong đa giác hay không ta có thể sử dụng thuật toán sau:

Đầu vào: Cho trước đa giác P và điểm p

Đầu ra: p nằm trong hay ngoài P.

Begin

If (p nằm trên cạnh P), p trong P

Else

đếm=0

l= tia song song trục X vẽ từ p

F**or** (i=1 **to** n)

B**egin**

I**f** (nếu cạnh (i) cắt l) **and not** cạnh (i) không trùng với l) **then**

B**egin**

I**f** (một đầu cuối của cạnh (i) nằm phía trên tia l) đếm=đếm+1

End

End For

I**f** (đếm là lẻ), p nằm trong P

End If

End

Để tìm giao của hai đoạn thẳng ta sử dụng thuật toán biểu diễn đoạn thẳng bằng phương trình tham số như sau:

Phương trình đoạn thẳng là cạnh đa giác được xác định từ hai toạ độ đỉnh liên tiếp. Giả sử ta có tham số t thay đổi từ 0 đến 1 cho phần đoạn thẳng AB giữa hai đỉnh đa giác và có giá trị 0 tại một đầu, giá trị 1 tại đầu cuối kia. Vậy với $0 \le t \le 1$, ta có:

$$x = x_A + t(x_B - x_A)$$
$$y = y_A + t(y_B - y_A) \quad (1)$$

Tương tự, cạnh CD của đa giác thứ hai sẽ được biểu diễn bởi tham số s (2)

$$x = x_C + s(x_D - x_C)$$
$$y = y_C + s(y_D - y_C)$$

và phương trình sau:

$$t = \frac{(x_C - x_A)(y_C - y_D) - (x_C - x_A)(y_C - y_A)}{(x_B - x_A)(y_C - y_D) - (x_C - x_D)(y_B - y_A)}$$
$$s = \frac{(x_B - x_A)(y_C - y_A) - (x_C - x_A)(y_B - y_A)}{(x_B - x_A)(y_C - y_D) - (x_C - x_D)(y_B - y_A)} \quad (3)$$

Từ các công thức sau đây ta tính được giá trị t và s:

trong đó, nếu $0 \le t \le 1$ và $0 \le s \le 1$ thì hai đoạn thẳng cắt nhau tại một điểm và giao điểm này được tính từ (1) và (2).

2.2.4.2.2. Phân tích và cài đặt thuật toán

Phần này trình bày tóm tắt các bước chính cài đặt chức năng xếp chồng các chủ đề bản đồ trong hệ thống GIS véc tơ. Giả sử ta phải thực hiện tính toán phần phủ của vùng địa lý được biểu diễn bởi đa giác P trong chủ đề T_1 với các vùng của chủ đề T_2.

Bước 1. Xác định xem đa giác P của chủ đề T_1 giao với các đa giác nào của chủ đề T_2 . Một bản đồ chủ đề chứa vô số đa giác (thí dụ bản đồ hành chính Việt nam chia đến cấp xã có đến 10511 xã), đa giác biểu diễn xã lại có vô số cạnh. Để

tăng tốc độ xử lý của máy tính ta sẽ không so sánh đa giác P của T_1 với mọi đa giác của T_2. Cấu trúc CSDL địa lý thường lưu trữ chữ nhật bao của các đa giác. Trước khi kiểm tra hai đa giác có giao nhau hay không thì cần kiểm tra chữ nhật bao của chúng có giao nhau hay không vì hai đa giác giao nhau chỉ khi hai chữ nhật bao của chúng giao nhau. Giải pháp này làm giảm đáng kể số lần tính toán. Việc xác định chính xác hai đa giác P, Q có giao nhau hay không được thực hiện theo thuật toán sau:

Đầu vào: Đa giác P, Q

Đầu ra: P và Q có giao nhau?

Begin

If (không có cạnh nào của P cắt cạnh nào đó của Q) **do**

Begin

p = điểm bất kỳ nào trên biên của P

If (p nằm trong Q)

$P \subset Q$

Else

Begin

q= điểm bất kỳ trên biên của Q

If (q nằm trong Q)

$Q \subset P$

Else

P và Q không giao nhau

End If

End If

Else P giao với Q

End

Độ phức tạp của thuật toán tìm giao các cạnh hai đa giác sẽ là O ($n$log$n$).

Bước 2. Phân lớp các đỉnh đa giác P và Q. Mỗi đỉnh đa giác được gán bởi giá trị I (trong), O (ngoài) hay B (biên) so với đa giác kia. Các giá trị này được thực hiện nhờ thuật toán điểm trong đa giác trình bày trên. Các giá trị của đỉnh được lưu vào danh sách P_v cho đa giác P và Q_v cho đa giác Q như sau:

$$P_v=<(p_1, O), (p_2, I) .... (p_n, B)>$$

$$Q_v=<(q_1, I), (q_2, B) .... (q_m, O)>$$

trong đó, n là tổng số đỉnh của đa giác P, m là tổng số đỉnh của đa giác Q.

Bước 3. Tìm giao điểm của các cạnh đa giác P và Q. Giao của các cạnh đa giác được tính theo công thức (3) trình bày trên. Phải xét lần lượt các cạnh của P có cắt các cạnh của Q hay không. Nếu có điểm cắt thì xen chúng vào danh sách P_v và Q_v với giá trị B (biên).

Kết quả cho lại các cạnh của đa giác được chia thành đoạn nhỏ. Mỗi đoạn của đa giác này sẽ nằm toàn bộ trong hay toàn bộ ngoài đa giác kia. Ta có thể sử dụng tính chất mô tả ở phần trên để xác định các đoạn thẳng nằm trong hay ngoài đa giác: nếu điểm giữa đoạn thẳng nằm trong hay nằm ngoài đa giác thì đoạn thẳng đó sẽ nằm trong hay nằm ngoài đa giác.

Bước 4. Lập các đa giác kết quả. Từ danh sách P_v và Q_v ta lọc ra các đoạn thẳng có giá trị I hay B để lập các đa giác mới.

Quan sát 4 bước trên đây thì bước 2 có độ phức tạp lớn nhất. Nó đòi hỏi tìm giao điểm của mọi cạnh hai đa giác cho nên có độ phức tạp $O(n^2)$, việc xen giá trị giao điểm vào danh sách sẽ có độ phức tạp $O(k^2)$, trong đó k là số phần tử trong danh sách. Trường hợp xấu nhất sẽ là $k=n^2$. Do vậy, độ phức tạp của bước này sẽ là $O(n^4)$.

2.2.4.2.3. Kết luận thuật toán

Thuật toán tìm giao của hai đa giác là nền tảng của việc xây dựng chức năng xếp chồng của hệ thống GIS véc tơ. Thuật toán trên đây được cài đặt trong một số

hệ thống GIS chuyên dụng đang được sử dụng.

2.3. Tổng kết chương

Để xây dựng được chương trình thử nghiệm giải quyết tìm vị trí tối ưu để đặt cây ATM tại Hải Phòng cần phải sử dụng các thuật toán phân cụm không gian và thuật toán xếp chồng bản đồ. Thuật toán phân cụm không gian mà nội dung luận văn đề cập đến đó là thuật toán Kmean. Đây là thuật toán được sử dụng nhiều nhất trong hướng tiếp cận phân nhóm phân hoạch. Thuật toán gồm 5 bước, thuật toán có ưu điểm rõ ràng, dễ cài đặt. Nhưng nhược điểm là phải chỉ ra số lượng cụm và yêu cầu CSDL cần phân nhóm phải xác định được tâm. Thuật toán này cũng không phù hợp với việc khai phá dữ liệu gồm các Cluster có hình dạng không lồi. Thuật toán phân cụm dựa trên mật độ DBSCAN được giới thiệu vào năm 1996. DBSCAN được khẳng định qua thực nghiệm là tốt hơn các thuật toán khác trong việc phát hiện ra các cụm với hình dạng bất kì, kể cả hình không lồi. Tuy nhiên nhược điểm của thuật toán là: nếu các cụm có mật độ khác nhau nhiều thì DBSCAN sẽ không giữ được tính hiệu quả, tốn nhiều thời gian để xác định lâ cận của các điểm một cách không cần thiết.. DBSCAN không chú ý đến thuộc tính phi không gian. Thuật toán xếp chồng bản đồ gồm thuật toán giao hai đoạn thẳng. Đây là thuật toán quyét dòng để liệt kê tất cả các đoạn thẳng giao nhau trong mặt phẳng, ưu điểm của thuật toán thực hiện đơn giản và chiếm ít bộ nhớ. Thuật toán giao của hai đa giác được sử dụng để tìm giao của hai đa giác. Thuật toán này là nền tảng của việc xây dựng chức năng xếp chồng của hệ thống GIS véc tơ.

Dựa vào các thuật toán phân cụm không gian và xếp chồng bản đồ để xây dựng chương trình thử nghiệm tìm vị trí tối ưu đặt cây ATM tại thành phố Hải Phòng.

CHƯƠNG 3. XÂY DỰNG ỨNG DỤNG THỬ NGHIỆM.

3.1. Giới thiệu về bài toán xác định vị trí đặt máy ATM tại thành phố Hải Phòng.

Trong vài năm trở lại đây, cùng với sự phát triển của xã hội, việc sử dụng thẻ ATM tại Việt Nam rất phổ biến. Thẻ ATM thực chất như một loại ví điện tử cho phép người sử dụng chỉ cần mang theo một chiếc thẻ gọn nhẹ, thay vì rất nhiều tiền mặt. Thẻ ATM không những cho phép người dùng rút tiền khi cần tiền mặt, còn cho phép thực hiện nhiều giao dịch khác tại máy ATM hoặc điện thoại, chẳng hạn chuyển khoản, thanh tàu, xe… Thẻ ATM còn có thể dùng để thanh toán tại các nhà hàng, siêu thị, trung tâm mua sắm, các điểm bán hàng có đặt máy ATM. Ngoài việc tiện lợi trong sử dụng ra, chủ thẻ còn được hưởng lãi suất từ tài khoản tiền gửi.

Hải Phòng là một trong ba thành phố lớn của Việt Nam, là thành phố Cảng quan trọng, trung tâm công nghiệp lớn nhất phía Bắc Việt Nam, đồng thời cũng là trung tâm kinh tế, văn hóa, y tế, giáo dục, khoa học, thương mại và công nghệ của Vùng duyên hải Bắc Bộ. Tính đến tháng 12/2015, có diện tích là 1.523,9 km², dân số Hải Phòng là 2.103.500 người, là thành phố đông dân thứ 3 ở Việt Nam. Theo khảo sát số người dân sử dụng thẻ ATM ở Hải Phòng với số lượng tương đối lớn, trong đó 100% các cơ quan nhà nước, các doanh nghiệp trả lương cho người lao động qua thẻ ATM. Để thực hiện các giao dịch thẻ ATM cần phải có hệ thống máy ATM. Việc lựa chọn địa điểm đặt máy ATM thích hợp là rất quan trọng để sao cho người sử dụng thẻ ATM thấy tiện lợi nhất chẳng hạn không phải đi quá xa hay phải đợi một thời gian dài để thực hiện giao dịch.

Các ngân hàng tại thành phố Hải Phòng đang không ngừng đầu tư, nâng cấp thiết bị máy móc, nâng cao chất lượng dịch vụ để phục vụ khách hàng và thực thi chính sách của Nhà nước. Tuy nhiên một hệ thống ATM có chi phí rất lớn. Giả sử, một ngân hàng muốn đầu tư 1000 máy ATM, ngân hàng đó phải bỏ ra khoảng 600 tỷ đồng tiền mua máy, đó là chưa kể đến chi phí lắp đặt, thuê địa điểm, thuê người giám sát. Vì vậy các Ngân hàng đầu tư máy ATM trên toàn thành phố Hải Phòng là chưa nhiều. Theo nguồn thông tin trên các trang web cho biết địa điểm đặt máy

ATM trong toàn thành phố Hải Phòng của Ngân hàng Quân đội có 21 điểm [16], điểm đặt ATM ngân hàng Techcombank có 16 điểm [15], điểm đặt ATM ngân hàng Vietinbank - Ngân hàng công thương có 12 điểm [18], điểm đặt ATM ngân hàng Đông Á có 13 điểm[13], điểm giao dịch ATM ngân hàng BIDV - Ngân hàng đầu tư và phát triển Việt Nam có 11 điểm [17], điểm đặt ATM Ngân hàng Agribank- Ngân hàng nông nghiệp Việt Nam có 13 điểm [14]. Theo khảo sát vị trí đặt máy ATM chủ yếu tập trung ở một số quận Ngô Quyền, Hồng Bàng của thành phố, có những địa điểm ở gần nhau mà có nhiều máy ATM cùng đặt, cũng có nhiều địa điểm không có cây ATM . Do đó việc phân bố địa điểm ATM của các ngân hàng là chưa hợp lý, chồng chéo nhau. Có những cây ATM luôn ở trong tình trạng quá tải, lỗi, hết tiền làm bất lợi cho người chủ thẻ, có những máy ATM ít được sử dụng vì khu vực đấy có nhiều máy ATM được đặt gần nhau.

Xuất phát từ nhu cầu thực tế việc xây dựng hệ thống hỗ trợ lựa chọn địa điểm đặt máy ATM tại thành phố Hải Phòng bằng kỹ thuật phân cụm không gian là rất thích hợp. Qua hệ thống hỗ trợ giúp cho các ngân hàng lựa chọn địa điểm ATM tại thành phố Hải Phòng một cách tối ưu và tiện lợi cho người sử dụng. Và đó cũng là lý do mà Học viên chọn đề tài "Xây dựng hệ thống hỗ trợ lựa chọn địa điểm đặt máy ATM tại thành phố Hải Phòng bằng kỹ thuật phân cụm không gian". Luận văn giới thiệu tổng quan về GIS và phân cụm dữ liệu, giới thiệu một số thuật toán phân cụm dữ liệu không gian và thuật toán xếp chồng bản đồ được sử dụng hiện nay. Trên cơ sở đó cài đặt thử nghiệm một ứng dụng sử dụng kỹ thuật phân cụm dữ liệu địa lý và xếp chồng bản đồ, trong đó khai thác thông tin địa lý của các đối tượng địa lý có tầm ảnh hưởng quan trọng đến vị trí đặt các máy ATM như: các siêu thị, trung tâm mua sắm, nhà hàng, khách sạn, bệnh viện, trường học, ... để hỗ trợ giải quyết bài toán tìm vị trí tối ưu đặt các máy ATM trong khu vực nội thành thành phố Hải Phòng.

Trong khuôn khổ của luận văn, cơ sở để giải quyết bài toán này là dựa trên việc khai phá dữ liệu không gian, cụ thể là thực hiện phân cụm các điểm tiện ích để tìm ra phân bố không gian của chúng một cách tự động, từ đó hỗ trợ việc ra quyết

định lựa chọn các vị trí gần nhất tới các cụm điểm tiện ích. Phần này làm rõ bài toán cũng như tính khả thi của nó thông qua việc xem xét các yếu tố quyết định đến kết quả của bài toán, các yếu tố này bao gồm:

- Dữ liệu đầu vào.
- Phạm vi bài toán.
- Phương pháp kỹ thuật sử dụng để giải quyết bài toán.

3.2. Nguồn dữ liệu đầu vào và phạm vi bài toán

Dữ liệu bản đồ số các tỉnh thành của Việt Nam nhìn chung còn hạn chế về số lượng, chất lượng và thường không đồng bộ về tỷ lệ bản đồ, nội dung hiển thị. Đa phần bản đồ số chỉ bao gồm các lớp về ranh giới hành chính, địa hình. Các thông tin về các điểm tiện ích như khu du lịch, khách sạn, nhà hàng, siêu thị, đền chùa, khu vui chơi giải trí, vân vân, đều rất hạn chế. Đặc điểm của bài toán khai phá dữ liệu nói chung và phân cụm dữ liệu nói riêng là tính hiệu quả phụ thuộc nhiều vào khối lượng dữ liệu đầu vào.

Phạm vi bài toán gói gọn trong việc phân tích dữ liệu địa lý của Thành phố Hải Phòng, học viên đã thu thập và xây dựng được bản đồ số Thành phố Hải Phòng bao gồm 10 lớp thông tin bản đồ, được chia thành 9 nhóm thông tin như: nhóm lớp thông tin bản đồ nền, nhóm Văn hóa, giáo dục, Y tế, Kinh doanh và dịch vụ, Du lịch, hành chính...

Trong các lớp thông tin trên, lượng thông tin chủ yếu tập trung ở khu vực nội thành và các lớp thông tin có ý nghĩa đối với yêu cầu đặt ra của bài toán chủ yếu là các lớp thông tin dạng điểm tiện ích như:

- Khách sạn, nhà khách
- Cơ quan
- Siêu thị
- Ngân hàng
- Nhà hàng, quán bia, cà phê
- Trường học
- Bệnh viện

Như vậy, dữ liệu đầu vào sử dụng được cho bài toán chủ yếu là các lớp thông tin dạng điểm, phạm vi chủ yếu trong nội thành Thành phố Hải Phòng.

3.3. Phương pháp kỹ thuật giải quyết bài toán

Để giải quyết bài toán, cần thực hiện phân cụm các lớp dữ liệu điểm tiện ích, có hai cách tiếp cận:

- Sử dụng phân cụm đa chiều.
- Phân cụm đơn chiều trên các lớp dữ liệu và thực hiện tổng hợp các kết quả phân cụm.

Học viên lựa chọn tiếp cận thứ hai với lý do mong muốn áp dụng một bài toán quan trọng của hệ thông tin địa lý là bài toán chồng phủ bản đồ. Theo cách tiếp cận này, từng lớp dữ liệu điểm tiện ích sau khi phân cụm sẽ thực hiện xây dựng đường bao của các cụm, tạo thành một lớp bản đồ dạng vùng bao phủ của các cụm. Tiến hành chồng phủ các lớp bản đồ dạng vùng này sẽ thu được vùng giao cắt là vùng thỏa mãn tiêu chí: khoảng cách địa lý tới các cụm điểm tiện ích là nhỏ nhất, là vị trí có thể coi là tối ưu để đặt các điểm tiện ích mới, chẳng hạn các máy ATM, các trung tâm mua sắm…

Lựa chọn phương pháp phân cụm

Với đặc điểm của dữ liệu đầu vào như đã đề cập ở trên, chúng ta lựa chọn phương pháp phân cụm theo mật độ bởi:

- Đối tượng phân cụm chủ yếu là các điểm tiện ích, tức là các đối tượng dạng điểm. Kiểu đối tượng này khá phù hợp với phương pháp phân cụm theo mật độ.
- Không cần thiết phải biết trước số cụm điểm tiện ích phân hoạch được, do đó không sử dụng tiếp cận phân hoạch.
- Không cần lưu trữ thông tin các mức trung gian trong quá trình phân cụm, do đó không sử dụng tiếp cận theo lưới.

Lựa chọn độ đo sử dụng trong phân cụm

Chúng ta quan tâm đến tính liên tục về mặt không gian của các điểm tiện ích trong cụm và khoảng cách giữa các điểm này chứ không quan tâm đến hướng của chúng. Hơn nữa với các đối tượng dạng điểm, quan hệ về topology mang ít ý nghĩa

ngoại trừ các đối tượng này mang thông tin về mạng lưới liên thông như: mạng lưới cột điện, mạng lưới cấp nước…Do vậy ta sử dụng độ đo **khoảng cách** trong bài toán phân cụm đã đề ra (các độ đo đã được đề cập trong mục 3.3 chương 3).

3.4. Công nghệ sử dụng

Chương trình thử nghiệm được cài đặt bằng ngôn ngữ C#, trong đó có sử dụng thư viện mã nguồn mở **SharpMap** do tác giả Morten Nielsen (www.iter.dk) và cộng đồng mã nguồn mở phát triển để hỗ trợ hiển thị bản đồ. Một số chức năng chính đã được cài đặt trong chương trình:

- Duyệt bản đồ: hiển thị bản đồ, phóng to, thu nhỏ, trượt bản đồ
- Phân cụm dữ liệu bản đồ
- Chồng phủ bản đồ
- Lưu bản đồ

Học viên đã tiến hành cài đặt thử nghiệm thuật toán phân cụm dựa trên mật độ là thuật toán DBSCAN, ngoài ra cũng cài đặt thêm thuật toán phân cụm dựa trên phân hoạch K-means để so sánh và đánh giá.

3.5. Phân tích thiết kế hệ thống.

Hệ thống phải đảm bảo cung cấp các chức năng tối thiểu của một hệ thông tin địa lý như:

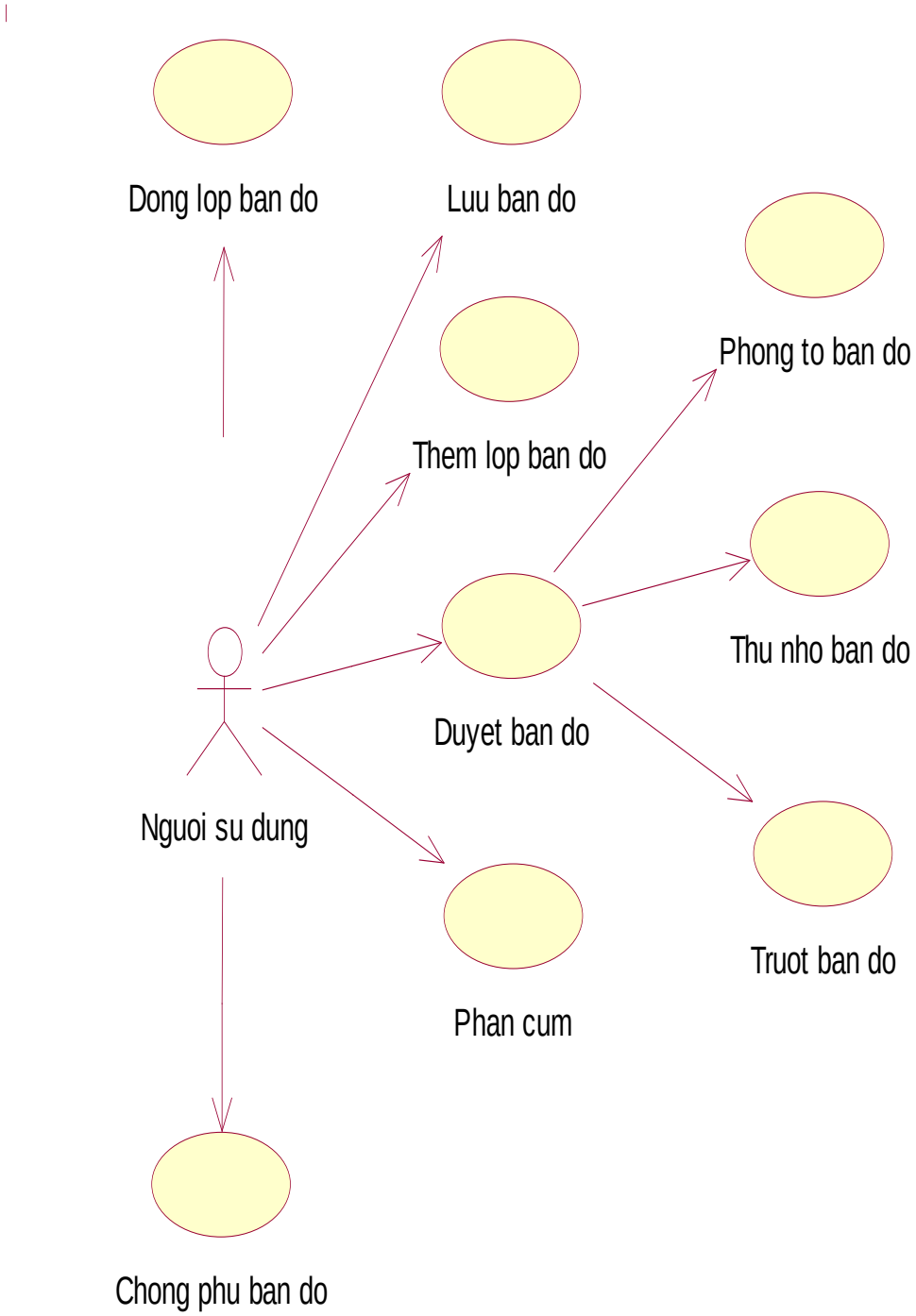
- Duyệt bản đồ
- Phóng to
- Thu nhỏ
- Trượt bản đồ
- Xem thông tin bản đồ

Ngoài ra, phục vụ yêu cầu của bài toán đã đề ra, hệ thống cần có thêm các chức năng:

- Phân cụm dữ liệu
- Chồng phủ bản đồ
- Lưu kết quả chồng phủ.

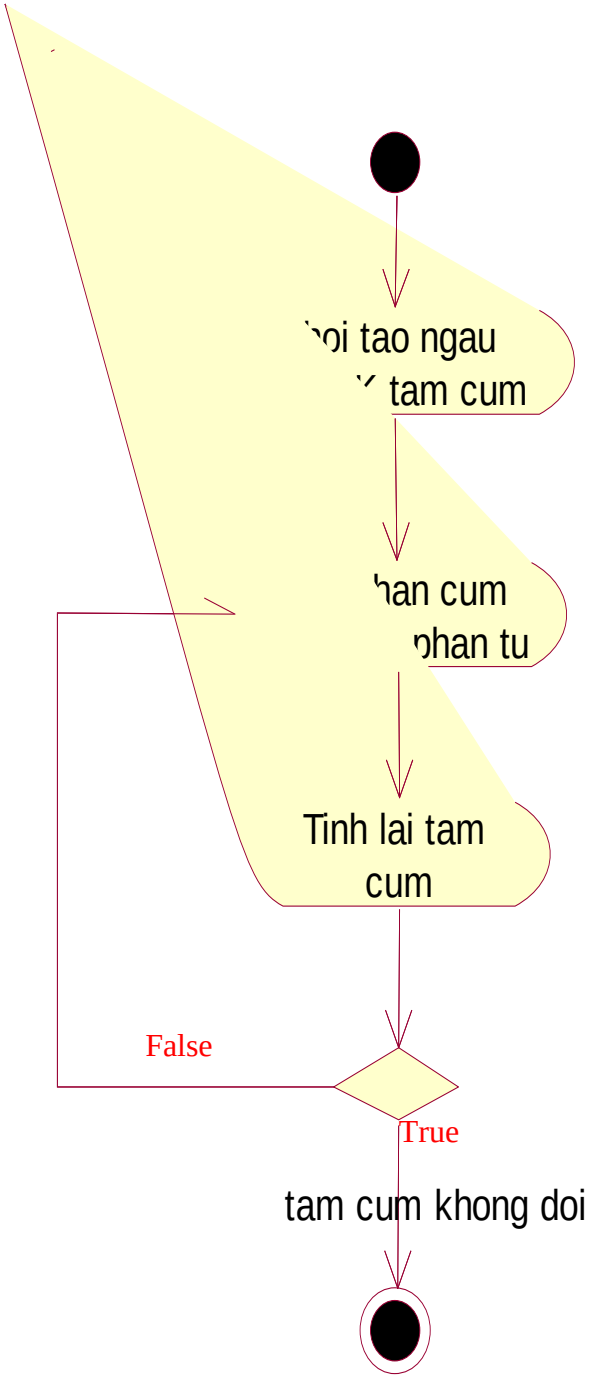
Trên cơ sở phân tích các chức năng của hệ thống như trên, chúng ta xây dựng được biểu đồ Use case thể hiện các chức năng chính của hệ thống như sau:

Biểu đồ ca sử dụng

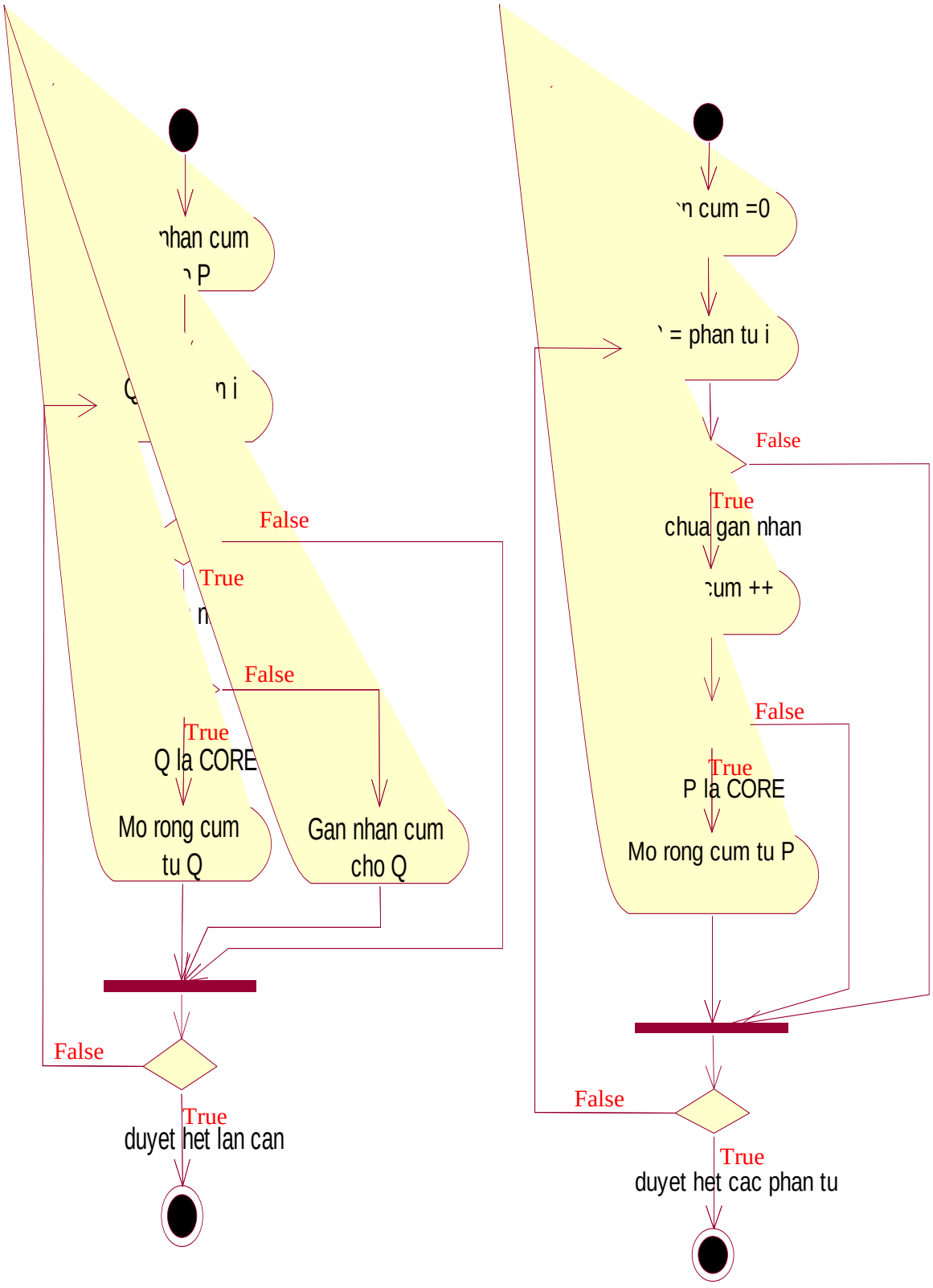


Biểu đồ hoạt động của một số thuật toán phân cụm đã cài đặt

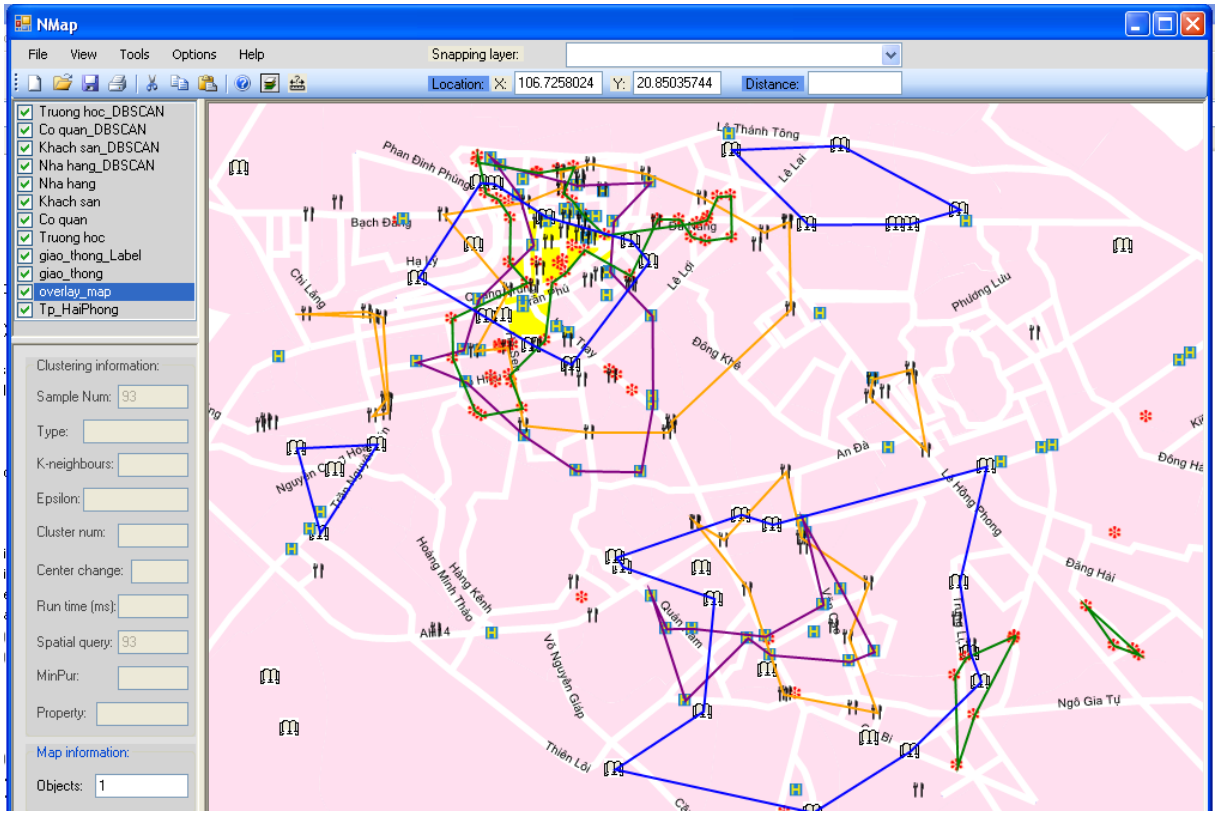
K-means



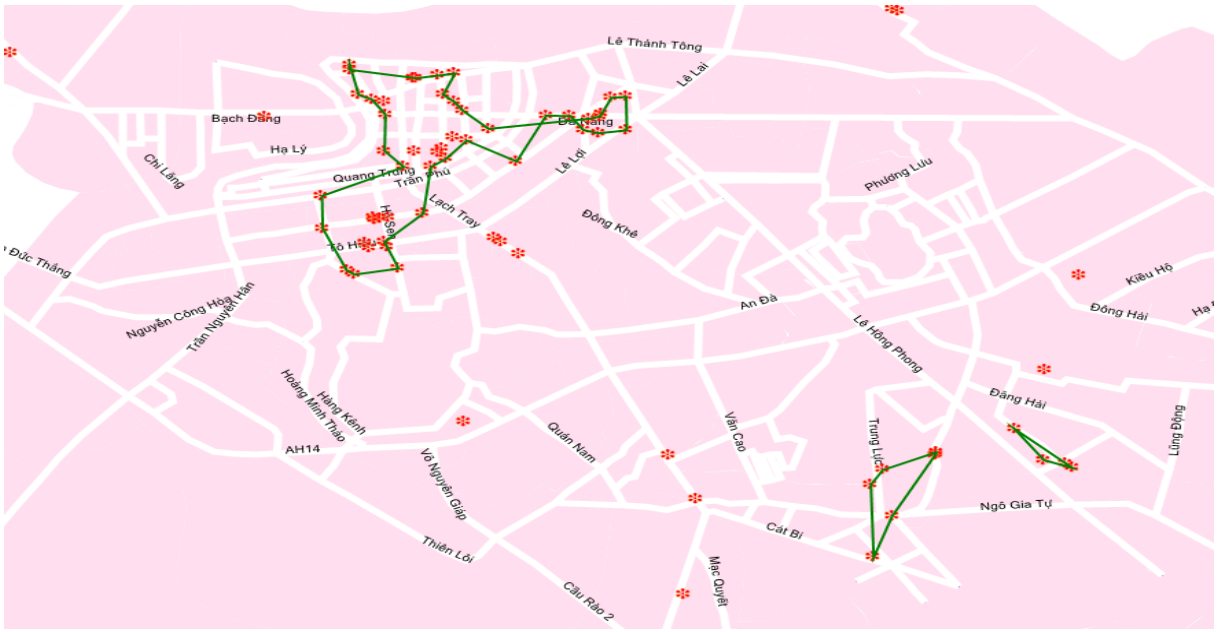
DBSCAN


nhan cum
P
n i
False
True
n
False
True
Q la CORE
Mo rong cum
tu Q
Gan nhan cum
cho Q
False
True
duyet het lan can
n cum =0
= phan tu i
False
True
chua gan nhan
cum ++
False
True
P la CORE
Mo rong cum tu P
False
True
duyet het cac phan tu

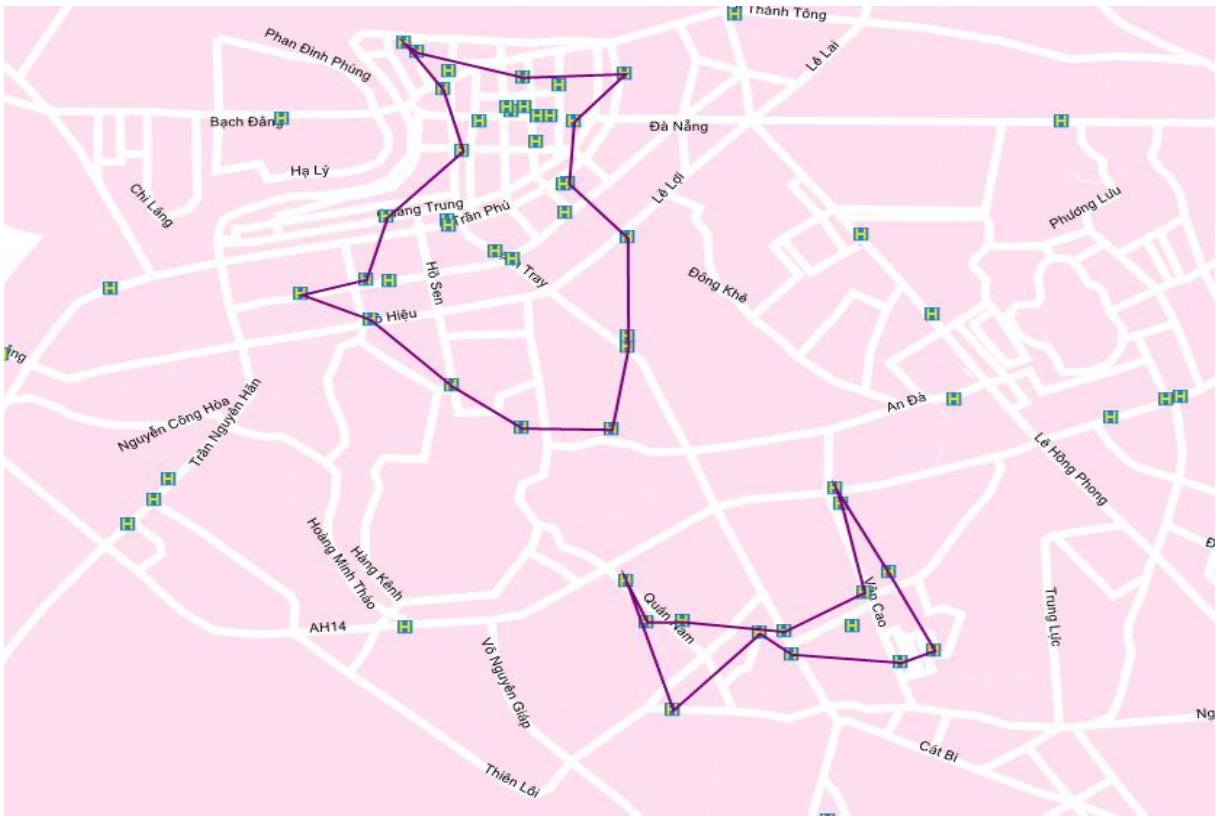
Một số hình ảnh của chương trình



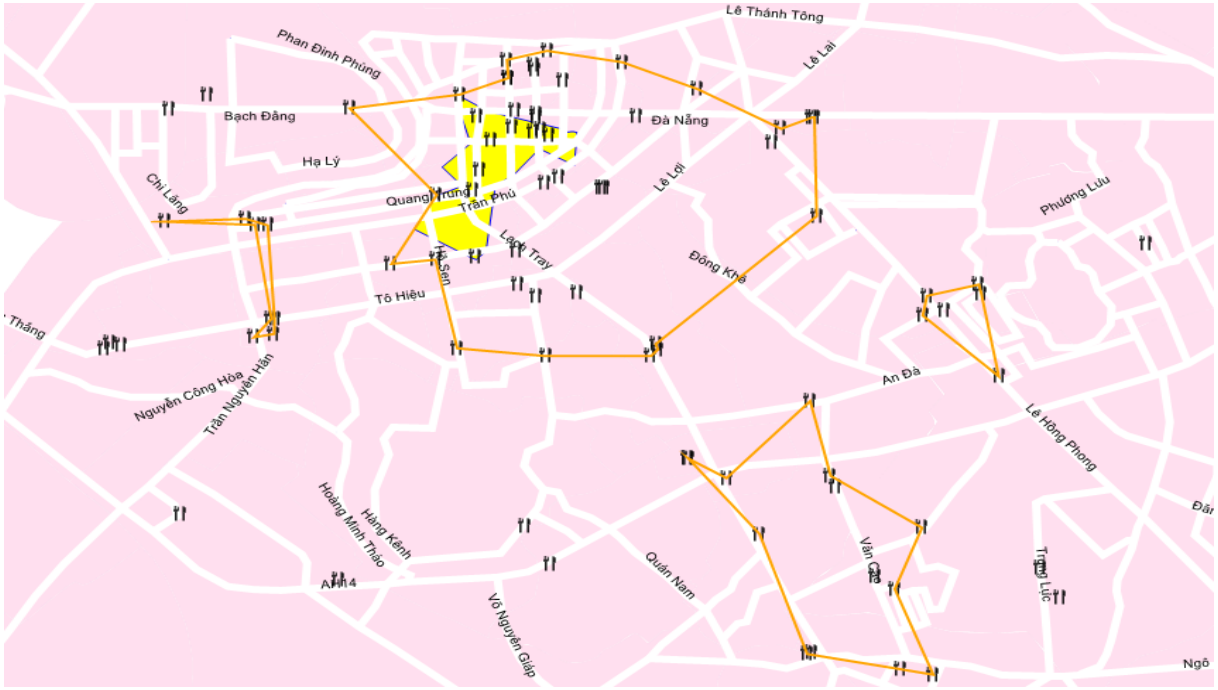
Hình 3.1: Giao diện chương trình



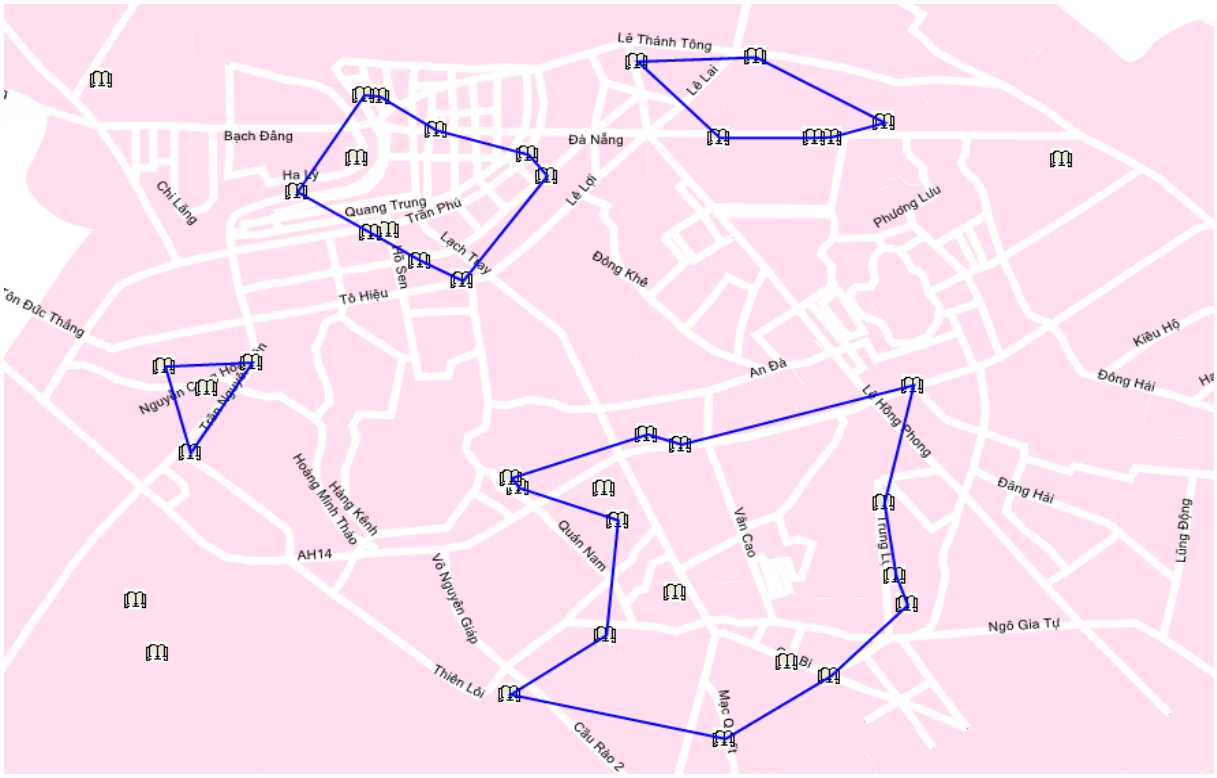
Hình 3.2: Phân cụm lớp dữ liệu "Cơ quan" trong nội thành Thành phố Hải Phòng.



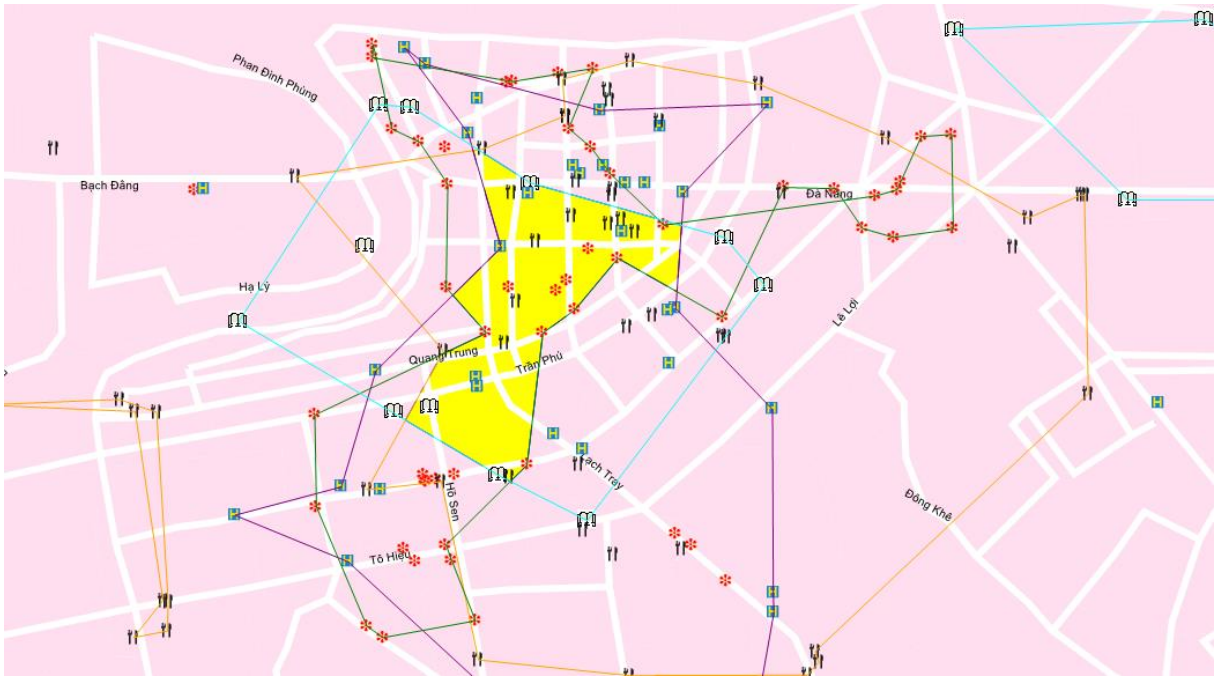
Hình 3.3: Phân cụm lớp dữ liệu "Khách sạn"



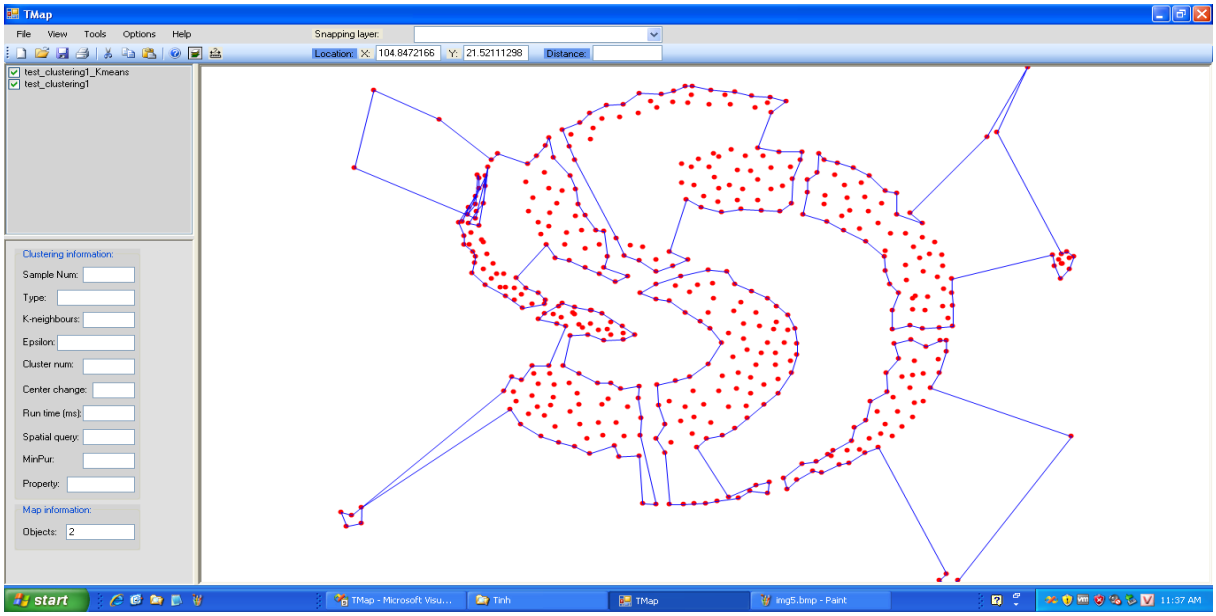
Hình 3.4: Phân cụm lớp dữ liệu "Nhà hàng"



Hình 3.5: Phân cụm lớp dữ liệu "Trường học"



Hình 3.6: Hình ảnh chồng phủ 4 lớp dữ liệu đã phân cụm như mô tả ở 4 hình trước, vùng màu vàng là vùng giao, cho thấy khu vực tập trung nhiều các điểm tiện ích, là khu vực tiềm năng đặt thêm máy ATM.



Hình 3.7: Kết quả phân cụm K-means đối với dữ liệu tự tạo, kết quả cho thấy khả năng phát hiện cụm lõm K-means rất kém chính xác.

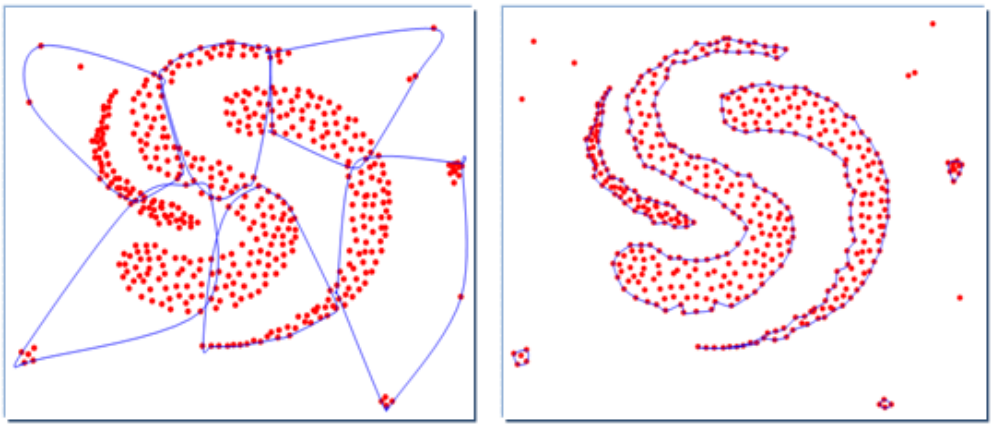
3.6. Đánh giá kết quả thu được

Học viên tiến hành thử nghiệm, so sánh và đánh giá 2 thuật toán đã cài đặt trên hệ thống như sau:

Đánh giá tổng quan 2 thuật toán

Bảng 3.1: So sánh tổng quan các thuật toán K-means, DBSCAN

	K-means	**DBSCAN**
Độ phức tạp	O(tKN)	O(NlogN)
Khả năng phát hiện nhiễu	kém	Tốt
Khả năng phát hiện cụm có hình dạng bất kỳ	không	có
Khả năng phân cụm theo thuộc tính phi không gian	không	không
Kết quả phân cụm	Khác nhau ở mỗi lần chạy	Giống nhau



Hình 3.8 : 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.

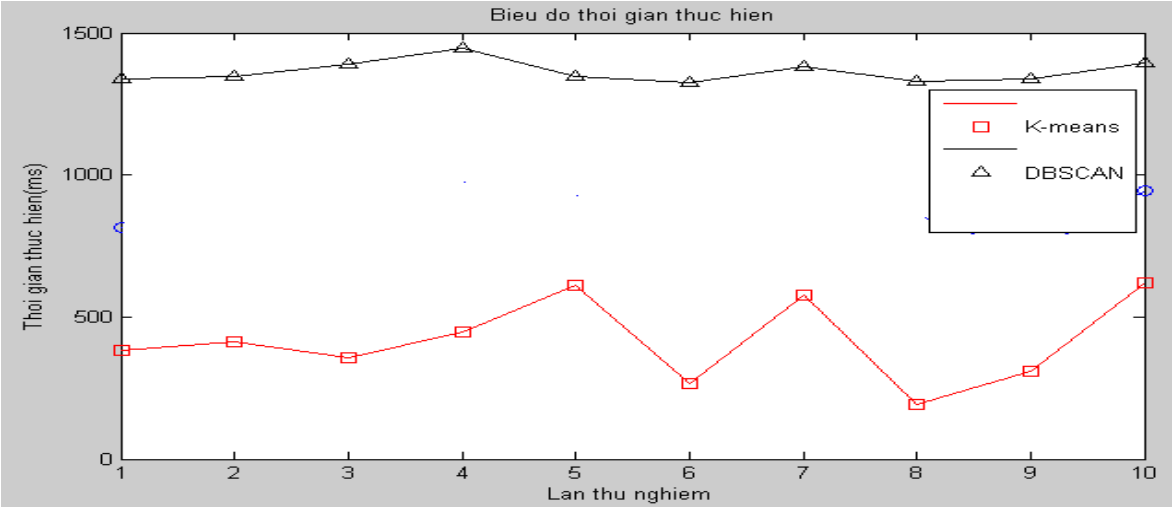
Đánh giá độ phức tạp thuật toán

Thử nghiệm thứ nhất: Thực hiện phân cụm với cùng một tập dữ liệu đầu vào: tệp Cosohatang_KTXH bao gồm 4235 mẫu dữ liệu, thực hiện trên máy tính với CPU 1.6GHz Celeron Mobile đơn lõi, 2GB Ram. Kết quả thu được như sau:

Bảng 3.2: Kết quả so sánh thời gian thực hiện phân cụm của các thuật toán K-means, DBSCAN với cùng một tập dữ liệu đầu vào

Bảng so sánh thời gian thực hiện phân cụm với cùng một tập dữ liệu đầu vào (với cùng một tập dữ liệu đầu vào: tệp Cosohatang_KTXH với 4235 mẫu dữ liệu thực hiện trên máy tính với CPU 1.6GHz Celeron Mobile đơn lõi, RAM 2GB)											
Thuật toán	**Thời gian (ms)**										**Các tham số phân cụm**
	lần 1	**lần 2**	**lần 3**	**lần 4**	**lần 5**	**lần 6**	**lần 7**	**lần 8**	**lần 9**	**lần 10**	
K-means	382	412	356	449	611	266	577	192	311	621	số cụm = 6
DBSCAN	1340	1347	1389	1445	1347	1323	1382	1331	1340	1395	epsilon = 1301.1470 , MinPts=4

Kết quả thể hiện dưới dạng đồ thị như sau:



Hình 3.9: Đồ thị so thời gian thực hiện phân cụm của các thuật toán K-measn, DBSCAN với cùng một tập dữ liệu đầu vào.

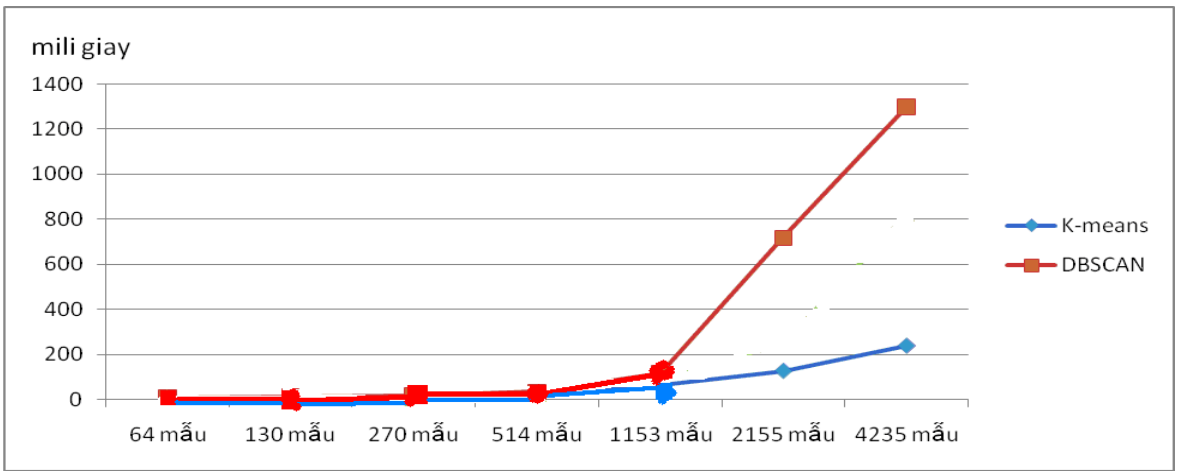
Kết quả cho thấy: với cùng số lượng dữ liệu đầu vào, thời gian thực hiện trung bình của thuật toán K-means thấp nhất, DBSCAN thực hiện lâu nhất. Đồ thị cũng cho thấy sự biến thiên thời gian thực hiện của K-means với mỗi bộ tâm cụm ngẫu nhiên ở mỗi lần chạy.

Thử nghiệm thứ 2: Sử dụng các tập dữ liệu đầu vào khác nhau, với số lượng dữ liệu tăng dần, kết quả thu được như bảng sau:

Bảng 3.3: Kết quả so sánh thời gian thực hiện phân cụm của các thuật toán K-means, DBSCAN trên các tập dữ liệu khác nhau.

Bảng so sánh thời gian thực hiện phân cụm với số lượng mẫu dữ liệu khác nhau (với các tập dữ liệu đầu vào khác nhau, thực hiện trên máy tính với CPU 1.6GHz Celeron Mobile đơn lõi, RAM 2GB)								
Số mẫu dữ liệu	**Thời gian (ms)**							**Các tham số phân cụm**
	64 mẫu	130 mẫu	270 mẫu	514 mẫu	1153 mẫu	2155 mẫu	4235 mẫu	
K-means	2	5	12	19	65	127	238	số cụm = 6
DBSCAN	8	14	19	35	117	717	1298	MinPts=4

Kết quả thể hiện trên đồ thị như sau:



Hình 3.10: Đồ thị thời gian thực hiện phân cụm của các thuật toán K-measn, DBSCAN trên các tập dữ liệu khác nhau.

Kết quả cho thấy, thời gian thực hiện của thuật toán K-means có dạng đường thẳng, phù hợp với độ phức tạp thuật toán O(tKn); thời gian phân cụm của DBSCAN và DBRS có dạng đường cong lên, phù hợp với độ phức tạp thuật toán O(NlogN). Đồ thị cũng cho thấy thuật toán DBRS có thời gian thực hiện thấp hơn DBSCAN do chỉ duyệt một số hữu hạn điểm ngẫu nhiên trong cơ sở dữ liệu.

KẾT LUẬN

Luận văn đã thực hiện được những công việc sau:

- Nghiên cứu tổng quan về Hệ thông tin địa lý và khai phá dữ liệu không gian.
- Nghiên cứu tổng quan về phân cụm dữ liệu và phân cụm dữ liệu không gian.
- Nghiên cứu một vài thuật toán sử dụng trong phân cụm dữ liệu không gian.
- Xây dựng chương trình thử nghiệm phân cụm các lớp dữ liệu điểm tiện ích, sử dụng trong bài toán cụ thể tính toán vị trí tối ưu lắp đặt máy ATM trong nội thành Hải Phòng.
- Đánh giá các thuật toán phân cụm đã cài đặt trên bộ dữ liệu bản đồ nội thành thành phố Hải Phòng.

Tuy nhiên, do bộ dữ liệu sử dụng để đánh giá chưa đủ lớn nên chưa đánh giá hết được hiệu quả và các đặc trưng của từng thuật toán, cũng như tính ổn định của hệ thống thử nghiệm. Trong tương lai, cần phải thử nghiệm và đánh giá trên những bộ dữ liệu lớn hơn.

Những đóng góp chính của luận văn bao gồm:

- Đã thử nghiệm một phương pháp khai phá dữ liệu không gian, trong đó kết hợp việc phân cụm các lớp dữ liệu không gian với các phép phân tích và xử lý dữ liệu không gian, hỗ trợ giải quyết lớp bài toán quản lý và lập kế hoạch dựa trên hệ thông tin địa lý.
- Cài đặt, khảo sát, đánh giá các thuật toán phân cụm K-means, DBSCAN trên dữ liệu không gian.

Hướng phát triển tiếp theo của luận văn:

- Hướng nghiên cứu của luận văn có thể được mở rộng sang lớp dữ liệu không gian dạng đường và dạng vùng, sử dụng trong khai phá dữ liệu liên quan đến các đối tượng địa lý dạng đường và dạng vùng.

- Một số ràng buộc và trọng số có thể được đưa vào bài toán để có thể khai phá dữ liệu một cách mềm dẻo và linh hoạt trong các điều kiện cụ thể của bài toán.
- Vấn đề phân cụm dữ liệu đa chiều có thể được thử nghiệm để so sánh với phương pháp hiện tại là phân cụm đơn chiều kết hợp với phân tích đa chiều dữ liệu không gian.
- Phương pháp tiếp cận sử dụng phân cụm mờ có thể được thử nghiệm bởi tính tương đối cố hữu của bài toán tối ưu.

TÀI LIỆU THAM KHẢO

Tiếng Việt

[1] Đặng Văn Đức, Hệ thống thông tin địa lý, NXB Khoa học và kỹ thuật, Hà Nội, 2001.

Tiếng Anh

[2] David Hand, Heikki Mannila, Padhraic Smyth, Principles of Data Mining, The MIT Press, 2001.

[3] Ester, M., Kriegel, H.-P., Sander, J., & Xu, X., A density-based algorithm for discovering clusters in large spatial databases with noise, Second Int. Conf. on Knowledge Discovery and Data Mining , (pp. 226-231). Portland, Oregon, 1996.

[4] Fayyad M. Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy (1996), *Advances in Knowledge Discovery and Data Mining*. AAAI Press/ The MIT Press.

[5] Harvey J. Miller (Editor), Jiawei Han (Editor), Geographic Data Mining and Knowledge Discovery, Second Edition, Taylor&Francis LLC, 2009.

[6] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques. University of Illinois, Morgan Kaufmann Publishers, 2006.

[7] Oracle,*OracleDataMiningConcepts10gRelease1*(10.1),OracleCorporation, 2003.

[8] Raymond T. Ng, Jiawei Han, CLARANS: A Method for Clustering Objects for Spatial Data Mining, IEEE, 9-10, 2002.

[9] Smid, Michiel (2003), Computing intersections in a set of line segments: the Bentley–Ottmann algorithm.

[10] Satish Puri, Dinesh Agarwal, Map Reduce algorithms for GIS Polygonal Overlay Processing, Georgia State University, USA, 2012.

[11] Tao Y., Papadias D. (2004), "Performance Analysis of R*-trees with Arbitrary Node Extents", *IEEE*.

[12] Wang, X., & Hamilton, H. J., DBRS- A Density-Based Spatial Clustering

Method with Random Sampling, 7th PAKDD, (pp. 563-575). Seoul, Korea, 2003.

Website

[13] http://donga.ngan-hang.com/atm/hai-phong

[14]http://haiphong.gov.vn/Portal/Detail.aspx?Organization=Citizen&MenuID=6774&ContentID=18800

[15] http://www.lukhach24h.com/listing/atm-ngan-hang-techcombank-tai-thanh-pho-hai-phong.html

[16] http://military-bank.ngan-hang.com/atm/hai-phong

[17]http://vayvontieudung.com.vn/index.php?branch_bank=11&branch_province=43&district=0&com=search&ctr=search&act=searchDiemGiaoDich

[18] http://vietinbank.ngan-hang.com/atm/hai-phong